

MADHA ENGINEERING COLLEGE

(Affiliated to Anna University and Approved by AICTE, New
Delhi) Madha Nagar, Kundrathur,
Chennai-600069

DEPARTMENT OF Master of Computer Application



R-2021

Lecture Notes

COURSE OBJECTIVES:

- To gain knowledge on foundations of machine learning and apply suitable dimensionality reduction techniques for an application
- To select the appropriate model and use feature engineering techniques
- To gain knowledge on Probability and Bayesian Learning to solve the given problem
- To design and implement the machine learning techniques for real world problems
- To analyze, learn and classify complex data without predefined models also

UNIT I INTRODUCTION**9**

Human Learning - Types – Machine Learning - Types - Problems not to be solved - Applications - Languages/Tools– Issues. Preparing to Model: Introduction - Machine Learning Activities - Types of data - Exploring structure of data - Data quality and remediation - Data Pre-processing

UNIT II MODEL EVALUATION AND FEATURE ENGINEERING**9**

Model Selection - Training Model - Model Representation and Interpretability - Evaluating Performance of a Model - Improving Performance of a Model - Feature Engineering: Feature Transformation - Feature Subset Selection

UNIT III BAYESIAN LEARNING**9**

Basic Probability Notation- Inference – Independence - Bayes' Rule. Bayesian Learning: Maximum Likelihood and Least Squared error hypothesis-Maximum Likelihood hypotheses for predicting probabilities- Minimum description Length principle -Bayes optimal classifier - Naïve Bayes classifier - Bayesian Belief networks -EM algorithm.

UNIT VI PARAMETRIC MACHINE LEARNING**9**

Logistic Regression: Classification and representation – Cost function – Gradient descent – Advanced optimization – Regularization - Solving the problems on overfitting. Perceptron – Neural Networks – Multi – class Classification - Backpropagation – Non-linearity with activation functions (Tanh, Sigmoid, Relu, PRelu) - Dropout as regularization

UNIT V NON PARAMETRIC MACHINE LEARNING**9**

k- Nearest Neighbors- Decision Trees – Branching – Greedy Algorithm - Multiple Branches – Continuous attributes – Pruning. Random Forests: ensemble learning. Boosting – Adaboost algorithm. Support Vector Machines – Large Margin Intuition – Loss Function - Hinge Loss – SVM Kernels

SUGGESTED ACTIVITIES:

1. Explore the significant steps involved in data preprocessing in Machine Learning
2. Choose a model and train a model in machine learning.
3. Explain the application of Bayes Theorem and how it's useful to predict the future
4. Make the difference between supervised Learning and unsupervised Learning Techniques
5. Differentiate Perceptron, Neural Network, Convolutional Neural Network and Deep Learning

TOTAL:45 PERIODS

COURSE OUTCOMES:

CO1:Understand about Data Preprocessing, Dimensionality reduction

CO2:Apply proper model for the given problem and use feature engineering techniques

CO3:Make use of Probability Technique to solve the given problem.

CO4:Analyze the working model and features of Decision tree

CO5:choose and apply appropriate algorithm to learn and classify the data

REFERENCES

1. Ethem Alpaydin, "Introduction to Machine Learning 3e (Adaptive Computation and Machine Learning Series)", Third Edition, MIT Press, 2014
2. Tom M. Mitchell, "Machine Learning", India Edition, 1st Edition, McGraw-Hill Education Private Limited, 2013
3. Saikat Dutt, Subramanian Chandramouli and Amit Kumar Das, "Machine Learning", 1st Edition, Pearson Education, 2019
4. Christopher M. Bishop, "Pattern Recognition and Machine Learning", Revised Edition, Springer, 2016.
5. Aurelien Geron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow", 2nd Edition, O'Reilly, 2019
6. Stephen Marsland, "Machine Learning – An Algorithmic Perspective", Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

MC4302

INTERNET OF THINGS

L T P C
3 0 0 3

COURSE OBJECTIVES:

- To understand the concepts of IoT and its working models
- To know the various IoT protocols
- To understand about various IoT Physical devices and Endpoints
- To know the security and privacy issues connected with IoT
- To apply the concept of Internet of Things in a real world scenario.

UNIT I FUNDAMENTALS OF IOT

9

Definition and Characteristics of IoT, Sensors, Actuators, Physical Design of IoT – IoT Protocols, IoT communication models, IoT Communication APIs, IoT enabled Technologies – Wireless Sensor Networks, Cloud Computing, Embedded Systems, IoT Levels and Templates, Domain Specific IoTs – Home, City, Environment, Energy, Agriculture and Industry.

UNIT II IOT PROTOCOLS

9

Protocol Standardization for IoT – Efforts – M2M and WSN Protocols – SCADA and RFID Protocols – Issues with IoT Standardization – Unified Data Standards – Protocols – IEEE802.15.4–BACNet Protocol– Modbus – KNX – Zigbee– Network layer – APS layer – Security

UNIT III IOT PHYSICAL DEVICES AND ENDPOINTS

9

Introduction to Arduino and Raspberry Pi- Installation, Interfaces (serial, SPI, I2C), Programming – Python program with Raspberry PI with focus on interfacing external gadgets, controlling output, and reading input from pins.

UNIT - I

Introduction

Session Plan

1. Human learning
2. Human learning types.

1. Human Learning

Human learning process varies from person to person. Once a learning process is set into the minds of people, it is difficult to change it. But, in Machine Learning (ML), it is easy to change the learning method by selecting a different algorithm.

In ML, we have well defined processes to understand and estimate the accuracy of learning. Estimation of human learning is usually done through examinations and it cannot be considered as a measure of intelligence.

Human acquire knowledge through experience either directly or shared by others. Machines acquire knowledge through experience shared in the form of past-data.

Humans have the terms, knowledge, skill and memory being used to define knowledge. intelligence.

	Human	Machine
Cost	Low initial cost and high running cost.	High initial cost (in case of robots) and low running cost. (works 24/7)
Creativity	Creative	Uninspired
Permanency of Intelligence	Human intelligence is perishable. We could not preserve Einstein's intelligence after his death.	Machine intelligence is permanent. It is easy to preserve intelligent tools like Siri and Watson.
Ease of duplication and dissemination of knowledge	Slow language-based communication process and some expertise can never be duplicated.	Knowledge can be copied from a machine and easily moved to another one.
Better is	* fusing data from multiple sources and interpreting the outside world	* faster at performing arithmetic and logical operations
	* distinguishing faces, identifying objects, recognising language sounds.	* dealing with multi-dimensional data

learning from few examples. A kid can differentiate between a man and a tree just by showing him/her one example

discovering complex patterns such as that exist in financial, scientific or product data.

develops new concepts/ imagination and creative reasoning

Operations that require fast, precise, highly repeatable actions.

Working in harsh environments (in case of robots).

2. Human learning types:

There are 3 major types of behaviour at learning. They are, (behavioural psychology described)

1. Learning through association - Classical conditioning
2. Learning through consequences - Operant conditioning
3. Learning through observation - Modeling / Observational learning

Learning:

Learning is a change in behaviour or in potential behavior that occurs as a result of experience.

Learning occurs most rapidly on a schedule of continuous reinforcement.

Types of Learning:

1. Motor Learning:

Our day to day activities like walking, running, driving, etc., must be learnt for ensuring a good life. These activities to a great extent involve muscular coordination.

2. Verbal Learning:

It is related with the language which we use to communicate and various other forms of verbal communication such as: symbols, words, languages, sounds, figures and signs.

3. Concept Learning:

This form of learning is associated with higher order cognitive processes like intelligence, thinking, reasoning, etc. which we learn right from our childhood.

Concept learning involves the processes of abstraction and generalization which is very useful for identifying or recognizing things.

4. Discrimination Learning:

Learning which distinguishes between various stimuli with its appropriate and different responses is regarded as discrimination learning.

5. Learning of Principles:

Learning which is based on

Principles helps in managing the work most effectively. Principles based learning explains the relationship between various concepts.

6. Attitude Learning:

Attitude shapes our behaviour to a very great extent, as our positive or negative behaviour is based on our attitudinal predisposition.

Questions:

1. Differentiate human learning and machine learning.
2. What are the types of human learning? Explain in detail.

Session Plan

1. Machine learning - introduction
2. Types of Machine Learning.

1. Introduction to machine learning:

Machine learning is the Science (and art) of programming computers so they can learn from data.

Arthur Samuel, 1959:

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.

Tom Mitchell, 1997:

A Computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E .

2. Types of Machine Learning:

- They are trained with human supervision (Supervised, unsupervised, Semisupervised and Reinforcement Learning).

- They can learn incrementally on the fly (online versus batch learning).
detect patterns build a predictive model.

1. Supervised Learning :

→ In Supervised learning, the training data you feed to the algorithm includes the desired solution called labels.

→ A typical supervised learning task is classification. eg. Spam, classify new email.

→ Another typical task is to predict a target numeric value.

eg: cars including both their predictors and their labels.

Supervised learning algorithm:

- K-Nearest Neighbours
- Linear Regression
- Logistic Regression
- Support Vector Machines (SVMs)
- Decision Trees and Random Forests.
- Neural Networks

2. Unsupervised Learning:

In unsupervised learning, as you might guess the training data is unlabeled. The system tries to learn without a teacher.

Unsupervised Learning Algorithm:

- Clustering
 - K-means
 - DBSCAN
 - Hierarchical Cluster Analysis (HCA)
- Anomaly detection and novelty detection
 - One-class SVM
 - Isolation Forest

- Visualization and dimensionality Reduction
 - principal Component Analysis (PCA)
 - Kernel PCA
 - Locally-linear Embedding (LLE)
 - t -distributed Stochastic Neighbour Embedding (t -SNE).
- Association Rule Learning
 - Apriori
 - Eclat

For eg: Blog's visitors

→ Cluster algorithm to try to detect groups of similar visitors. 40% of your visitors are males who love comic books and generally read your blog in the evening. 20% are young sci-fi lovers who visit during the weekends.

→ Hierarchical clustering algorithm it may also subdivide each group into smaller groups.

→ Visualization and dimensionality reduction without losing too much information. This is to merge several correlated features into one
eg: car's mileage with its age, represent the car's wear and tear.

→ Anomaly detection automatically removing outliers.

3. SemiSupervised Learning :-

Some algorithm can deal with partially labeled training data, usually a lot of unlabeled data and a little bit of labeled data. This is

called semisupervised learning.

For eg: Google photos, deep belief networks (DBNs) are based on unsupervised components called restricted Boltzmann machines (RBMs)

4. Reinforcement Learning :-

Reinforcement learning is a very different beast. The learning system called an agent in the context, select and perform actions and get rewards in return.

for eg: online shopping delivery time feedback.

AlphaGo program is also a good example May 2017 when it beat the world champion Ke Jie at the game of Go.

Batch and Online Learning :-

The system can learn incrementally from a stream of incoming data.

Batch Learning :-

→ The system is incapable of learning incrementally, so it is typically done offline. without learning anymore this is called offline learning.

→ From scratch on the full dataset (not just the new data, but also the old data).

→ Batch learning system can adapt to change. Update the data and train a new version.

→ Full set of data can take many hours new system only every 24 hours or even just weekly.

→ Training on the full set of data requires a lot of computing resources train from scratch every day.

→ Impossible to use a batch learning algorithm. Able to learn autonomously and it has limited resources.

eg: a smartphone application or a rover on Mars.

Online Learning :-

→ Sequentially, either individually or by small group called mini-batches.

→ Learn about new data on the fly as it arrives, for systems that receive data as a continuous flow (eg: Stock prices)

→ If you have limited computing resources system has learned about new data instances.

→ online learning algorithm can also be used to train systems on huge datasets that cannot fit in one machine's main memory.

Repeat the process until it has run on all of the data.

→ How fast they should adapt to changing data this is called Learning rate.

→ If bad data is fed to the system will gradually decline
for eg: On a Robot.

Instance-Based Versus Model-Based Learning:
Machine Learning systems is by how they generalize.

There are two main approaches to generalization

- i) Instance-based Learning
- ii) Model-based Learning

1. Instance-based Learning:

The system learns the eg by heart, then generalizes to new cases by comparing them to the learned examples (or a subset of them) using a similarity measure.

2. Model-based Learning:

Another way to generalize from a set of examples is to build a model of these eg then use that model to make predictions. This is called model-based Learning.

for eg: if money makes people happy, so you download the Better life Index data from the OECD's website as well as state about GDP per capita from the IMF's website.

Equation - A simple Linear model

$$y = mx + c$$

$$\text{life satisfaction} = \theta_0 + \theta_1 \times \text{GDP-per-capita}$$

→ Before you can use your model you need to define the parameter value θ_0 and θ_1 .

→ You need to specify a performance measure.

→ Define a utility function or fitness function or cost function.

Select a linear model.

```
model = sklearn.linear_model.LinearRegression()
```

Replacing the Linear Regression model with K-Nearest Neighbours regression in the previous code is as simple as replacing these two lines.

```
import sklearn.linear_model
```

```
model = sklearn.linear_model.LinearRegression()
```

With these two:

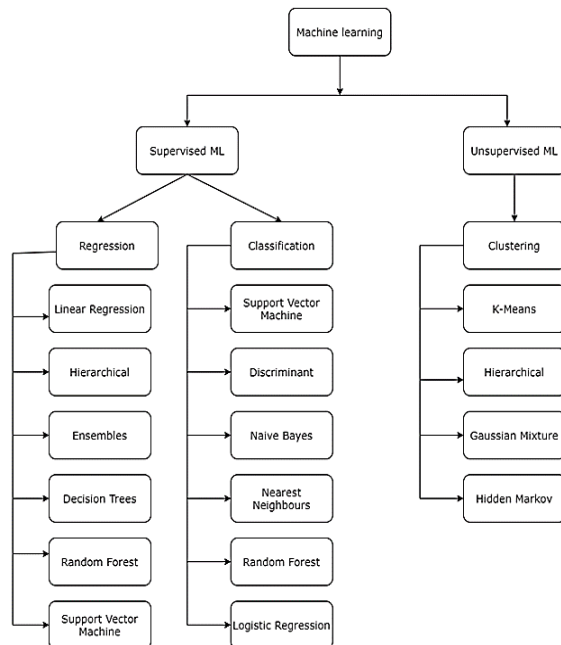
```
import sklearn.neighbors
```

```
model = sklearn.neighbors.KNeighborsRegressor(  
    n_neighbors=3)
```

E.g. A polynomial Regression model

1. Model selection ?

A machine learning model is defined as a mathematical representation of the output of the training process. Machine learning is the study of different algorithms that can improve automatically through experience & old data and build the model. A machine learning model is similar to computer software designed to recognize patterns or behaviors based on previous experience or data. The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on new data.



When solving a Machine Learning problem, we may zero down to several candidate models for the problem. We may further be interested in the selection of

1. The best choice among various ML algorithms (e.g., Logistic regression, support vector machine, neural networks, etc.)
2. Variables for linear regression
3. Basis terms such as polynomials, splines, or wavelets in function estimation
4. Most appropriate parametric family among several alternatives

When we are at it, what we should keep in our minds so that we select the best model?

The two primary criteria for model selection are prediction accuracy and model interpretability, which are listed below

1) Prediction Accuracy – One of the main objectives of Model Selection in Machine Learning is to find a model with the highest prediction accuracy. It can be measured in terms of MSE/Misclassification Error depending upon whether the target variable is quantitative or qualitative, respectively.

2. Model Interpretability – A highly complex model, with too many predictors, not only introduces The Overfitting Problem but also is difficult to interpret. An appropriate model tries to eliminate irrelevant variables from the model to make the model both simpler and accurate.

A good model selection technique will balance between prediction accuracy and simplicity.

Usually, we aim to find the model which works best on the test dataset. But, a designated test set is not available when we are building a predictive model. To address this problem, two conventional approaches are used to find the estimate of the test error.

1. Analytic Methods -We can indirectly estimate test error by making an adjustment to the training error to account for the bias due to overfitting. In these groups of methods, the training error is calculated first and then a penalty is added to the training error to estimate the testing error.

2. Resampling Methods - We can directly estimate the test error, using Resampling Methods. In resampling methods, the model is fit on one dataset and is validated on the complementary dataset and the validation error is recorded for each iteration. This process is repeated multiple times and the mean validation error is taken as an estimate for test error.

The Best Practices for Model Selection

Some general recommendations and best practices that are trendy in the data science community are listed below for reference.

1. Keep in mind the objectives of model selection
2. Cross-Validation is the most attractive method for model selection.
3. 5 or 10-fold cross-validation fares well for the majority of the cases.

In the simple linear models with a large number of predictors(p) and sample size(n), analytic methods perform as good as resampling methods and are computationally inexpensive.

2. What is training model in machine learning?

A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model.

This iterative process is called “model fitting”. The accuracy of the training dataset or the validation dataset is critical for the precision of the model.

Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved. There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning.

Supervised learning is possible when the training data contains both the input and output values. Each set of data that has the inputs and the expected output is called a supervisory signal. The training is done based on the deviation of the processed result from the documented result when the inputs are fed into the model.

Unsupervised learning involves determining patterns in the data. Additional data is then used to fit patterns or clusters. This is also an iterative process that improves the accuracy based on the correlation to the expected patterns or clusters. There is no reference output dataset in this method.

Types of ML Models

Amazon ML supports three types of ML models: binary classification, multiclass classification, and regression. The type of model you should choose depends on the type of target that you want to predict.

Binary Classification Model

ML models for binary classification problems predict a binary outcome (one of two possible classes). To train binary classification models, Amazon ML uses the industry-standard learning algorithm known as logistic regression.

Examples of Binary Classification Problems

- "Is this email spam or not spam?"
- "Will the customer buy this product?"
- "Is this product a book or a farm animal?"
- "Is this review written by a customer or a robot?"

Multiclass Classification Model

ML models for multiclass classification problems allow you to generate predictions for multiple classes (predict one of more than two outcomes). For training multiclass models, Amazon ML uses the industry-standard learning algorithm known as multinomial logistic regression.

Examples of Multiclass Problems

- "Is this product a book, movie, or clothing?"
- "Is this movie a romantic comedy, documentary, or thriller?"
- "Which category of products is most interesting to this customer?"

Regression Model

ML models for regression problems predict a numeric value. For training regression models, Amazon ML uses the industry-standard learning algorithm known as linear regression.

Examples of Regression Problems

- "What will the temperature be in Seattle tomorrow?"
- "For this product, how many units will sell?"
- "What price will this house sell for?"

Training Process

To train an ML model, you need to specify the following:

- Input training datasource
- Name of the data attribute that contains the target to be predicted
- Required data transformation instructions
- Training parameters to control the learning algorithm

During the training process, Amazon ML automatically selects the correct learning algorithm for you, based on the type of target that you specified in the training data source.

Creating a Model in Machine Learning

There are 7 primary steps involved in creating a machine learning model. Here is a brief summarized overview of each of these steps:

1. Defining the Problem

Defining the problem statement is the first step towards identifying what an ML model should achieve. This step also enables recognizing the appropriate inputs and their respective outputs; Questions like “what is the main objective?”, “what is the input data?” and “what is the model trying to predict?” must be answered at this stage.

2. Data Collection

After defining the problem statement, it is necessary to investigate and gather data that can be used to feed the machine. This is an important stage in the process of creating an ML model because the quantity and quality of the data used will decide how effective the model is going to be. Data can be gathered from pre-existing databases or can be built from the scratch

3. Preparing the Data

The data preparation stage is when data is profiled, formatted and structured as needed to make it ready for training the model. This is the stage where the appropriate characteristics and attributes of data are selected. This stage is likely to have a direct impact on the execution time and results. This is also at the stage where data is categorized into two groups – one for training the ML model and the other for evaluating the model. Pre-processing of data by normalizing, eliminating duplicates and making error corrections is also carried out at this stage.

4. Assigning Appropriate Model / Protocols

Picking and assigning a model or protocol has to be done according to the objective that the ML model aims to achieve. There are several models to pick from, like linear regression, k-means and bayesian. The choice of models largely depends on the type of data that is being used. For instance, image processing convolutional neural networks would be the ideal pick and k-means would work best for segmentation.

5. Training the Machine Model or “The Model Training”

This is the stage where the ML algorithm is trained by feeding datasets. This is the stage where the learning takes place. Consistent training can significantly improve the prediction rate of the ML model. The weights of the model must be initialized randomly. This way the algorithm will learn to adjust the weights accordingly.

6. Evaluating and Defining Measure of Success

The machine model will have to be tested against the “validation dataset”. This helps assess the accuracy of the model. Identifying the measures of success based on what the model is intended to achieve is critical for justifying correlation.

7. Parameter Tuning

Selecting the correct parameter that will be modified to influence the ML model is key to attaining accurate correlation. The set of parameters that are selected based on their influence on the model architecture are called hyper parameters. The process of identifying the hyper parameters by tuning the model is called parameter tuning. The parameters for correlation should be clearly defined in a manner in which the point of diminishing returns for validation is as close to 100% accuracy as possible.

3. What is an interpretable model ?

When humans easily understand the decisions a machine learning model makes, we have an “interpretable model”. In short, we want to know what caused a specific decision. If we can tell how a model came to a decision, then that model is interpretable.

For example,

we can train a random forest machine learning model to predict whether a specific passenger survived the sinking of the Titanic in 1912. The model uses all the passenger’s attributes – such as their ticket class, gender, and age – to predict whether they survived.

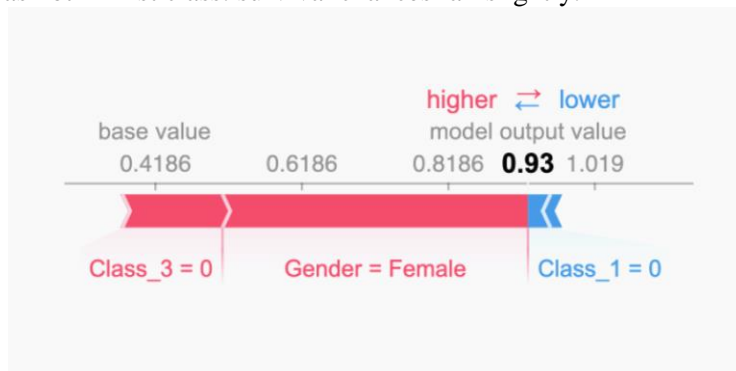
Now let’s say our random forest model predicts a 93% chance of survival for a particular passenger. How did it come to this conclusion?

Random forest models can easily consist of hundreds or thousands of “trees.” This makes it nearly impossible to grasp their reasoning.

But, we can make each individual decision interpretable using an approach borrowed from game theory.

SHAP plots show how the model used each passenger attribute and arrived at a prediction of 93% (or 0.93). In the Shapely plot below, we can see the most important attributes the model factored in.

- the passenger was not in third class: survival chances increase substantially;
- the passenger was female: survival chances increase even more;
- the passenger was not in first class: survival chances fall slightly.



We can see that the model is performing as expected by combining this interpretation with what we know from history: passengers with 1st or 2nd class tickets were prioritized for lifeboats, and women and children abandoned ship before men.

By contrast, many other machine learning models are not currently possible to interpret. As machine learning is increasingly used in medicine and law, understanding why a model makes a specific decision is important.

What do we gain from interpretable machine learning?

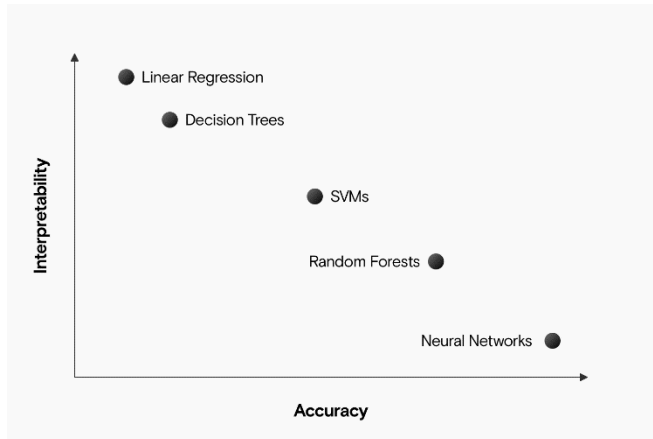
Interpretable models help us reach lots of the common goals for machine learning projects:

- **Fairness:** if we ensure our predictions are unbiased, we prevent discrimination against under-represented groups.
- **Robustness:** we need to be confident the model works in every setting, and that small changes in input don’t cause large or unexpected changes in output.
- **Privacy:** if we understand the information a model uses, we can stop it from accessing sensitive information.
- **Causality:** we need to know the model only considers causal relationships and doesn’t pick up false correlations;
- **Trust:** if people understand how our model reaches its decisions, it’s easier for them to trust it.

Are some algorithms more interpretable than others?

Simpler algorithms like regression and decision trees are usually more interpretable than complex models like neural networks. Having said that, lots of factors affect a model's interpretability, so it's difficult to generalize.

With very large datasets, more complex algorithms often prove more accurate, so there can be a trade-off between interpretability and accuracy.



A chart showing interpretability on the y-axis and accuracy on the x-axis. Linear regression is at the top left (very interpretable, not very accurate) and negative correlation runs through decision trees, SVMs, random forests, and neural networks.

More accurate models are often more difficult to interpret.

Scope of interpretability

By looking at scope, we have another way to compare models' interpretability. We can ask if a model is globally or locally interpretable:

- global interpretability is understanding how the complete model works;
- Local interpretability is understanding how a single decision was reached.

A model is globally interpretable if it's small and simple enough for a human to understand it entirely. A model is locally interpretable if a human can trace back a single decision and understand how the model reached that decision. A model is globally interpretable if we understand each and every rule it factors in. For example, a simple model helping banks decide on home loan approvals might consider:

- The applicant's monthly salary,
- The size of the deposit, and
- The applicant's credit rating.

A human could easily evaluate the same data and reach the same conclusion, but a fully transparent and globally interpretable model can save time.

In contrast, a far more complicated model could consider thousands of factors, like where the applicant lives and where they grew up, their family's debt history, and their daily shopping habits. It might be possible to figure out why a single home loan was denied, if the model made a questionable decision. But because of the model's complexity, we won't fully understand how it comes to decisions in general. This is a locally interpretable model.

Various ways to evaluate a machine learning model's performance?

The performance of our machine learning or deep learning model and why to use one in place of the other. We will discuss terms like:

1. Confusion matrix
2. Accuracy
3. Precision
4. Recall
5. Specificity
6. F1 score
7. Precision-Recall or PR curve
8. ROC (Receiver Operating Characteristics) curve
9. PR vs ROC curve.

For simplicity, we will mostly discuss things in terms of a binary classification problem where let's say we'll have to find if an image is of a cat or a dog. Or a patient is having cancer (positive) or is found healthy (negative). Some common terms to be clear with are:

True positives (TP): Predicted positive and are actually positive.

False positives (FP): Predicted positive and are actually negative.

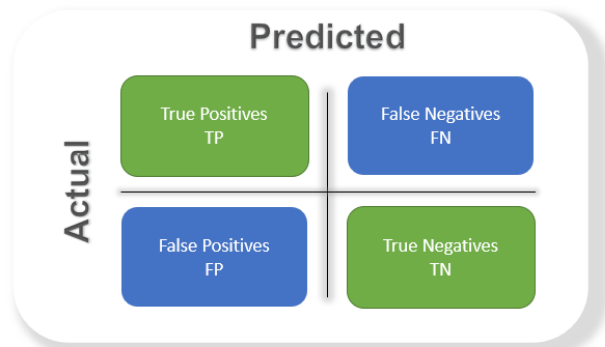
True negatives (TN): Predicted negative and are actually negative.

False negatives (FN): Predicted negative and are actually positive.

So let's get started!

Confusion matrix

It's just a representation of the above parameters in a matrix format. Better visualization is always good :)



Accuracy

The most commonly used metric to judge a model and is actually not a clear indicator of the performance.

The worse happens when classes are imbalanced.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Take for example a cancer detection model. The chances of actually having cancer are very low. Let's say out of 100, 90 of the patients don't have cancer and the remaining 10 actually have it. We don't want to miss on a patient who is having cancer but goes undetected (false negative). Detecting everyone as not having cancer gives an accuracy of 90% straight. The model did nothing here but just gave cancer free for all the 100 predictions.

We surely need better alternatives.

Precision

Percentage of positive instances out of the **total predicted positive** instances. Here denominator is the model prediction done as positive from the whole given dataset. Take it as to find out 'how much the model is right when it says it is right'.

$$\frac{TP}{TP + FP}$$

Recall/Sensitivity/True Positive Rate

Percentage of positive instances out of the **total actual positive** instances. Therefore denominator ($TP + FN$) here is the **actual** number of positive instances present in the dataset. Take it as to find out 'how much extra right ones, the model missed when it showed the right ones'.

$$\frac{TP}{TP + FN}$$

Specificity

Percentage of negative instances out of the **total actual negative** instances. Therefore denominator ($TN + FP$) here is the **actual** number of negative instances present in the dataset. It is similar to recall but the shift is on the negative instances. *Like finding out how many healthy patients were not having cancer and were told they don't have cancer.* Kind of a measure to see how separate the classes are.

$$\frac{TN}{TN + FP}$$

F1 score

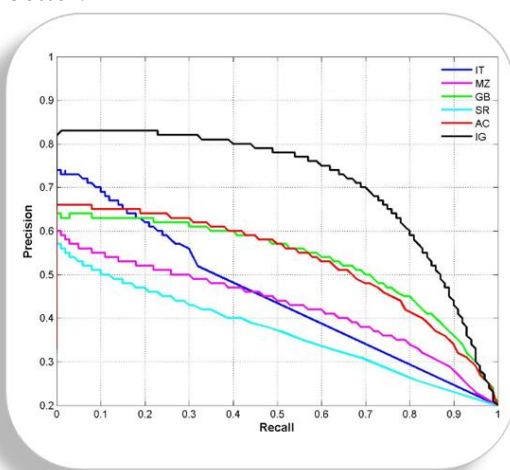
It is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better. See that due to the product in the numerator if one goes low, the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

One drawback is that both precision and recall are given equal importance due to which according to our application we may need one higher than the other and F1 score may not be the exact metric for it. Therefore either weighted-F1 score or seeing the PR or ROC curve can help.

PR curve

It is the curve between precision and recall for various threshold values. In the figure below we have 6 predictors showing their respective precision-recall curve for various threshold values. The top right part of the graph is the ideal space where we get high precision and recall. Based on our application we can choose the predictor and the threshold value. PR AUC is just the area under the curve. The higher its numerical value the better.

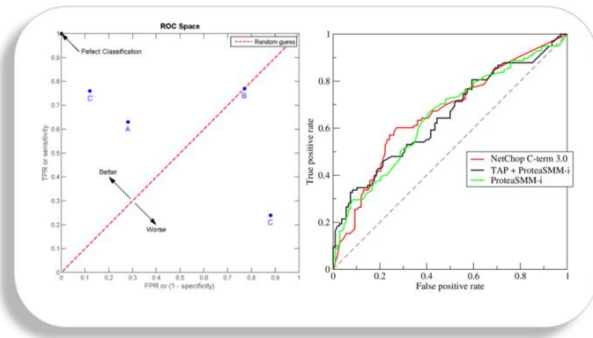


ROC curve

ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. As you can see in the first figure, we have four categories and we want the threshold value that leads us closer to the top left corner. Comparing different predictors (here 3) on a given dataset also becomes easy as you can see in figure 2, one can choose the threshold according to the application at hand. ROC AUC is just the area under the curve, the higher its numerical value the better.

$$\text{True Positive Rate (TPR)} = \text{RECALL} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate (FPR)} = 1 - \text{Specificity} = \frac{FP}{TN+FP}$$



PR vs ROC curve

Both the metrics are widely used to judge a models performance.

Which one to use PR or ROC?



The answer lies in TRUE NEGATIVES.

Due to the absence of TN in the precision-recall equation, they are useful in imbalanced classes. In the case of class imbalance when there is a majority of the negative class. The metric doesn't take much into consideration the high number of TRUE NEGATIVES of the negative class which is in majority, giving better resistance to the imbalance. This is important when the detection of the positive class is very important.

Like to detect cancer patients, which has a high class imbalance because very few have it out of all the diagnosed. We certainly don't want to miss on a person having cancer and going undetected (recall) and be sure the detected one is having it (precision).

Due to the consideration of TN or the negative class in the ROC equation, it is useful when both the classes are important to us. Like the detection of cats and dog. The importance of true negatives makes sure that both the classes are given importance, like the output of a CNN model in determining the image is of a cat or a dog.

Conclusion

The evaluation metric to use depends heavily on the task at hand. For a long time, accuracy was the only measure I used, which is really a vague option. I hope this blog would have been useful for you. That's all from my side. Feel free to suggest corrections and improvements.

4. Improve Performance of ML Models ?

1. Choosing the Right Algorithms

Algorithms are the key factor used to train the ML models. The data feed into this that helps the model to learn from and predict with accurate results. Hence, choosing the right algorithm is important to ensure the performance of your machine learning model.

Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes, kNN, K-Means, Random Forest and Dimensionality Reduction Algorithms and Gradient Boosting are the leading ML algorithms you can choose as per your ML model compatibility.

2. Use the Right Quantity of Data

The next important factor you can consider while developing a machine learning model is choosing the right quantity of data sets. And there are multiple factors and for deep learning-based ML models, a huge quantity of datasets is required for algorithms.

Depending on the complexities of problem and learning algorithms, model skill, data size evaluation and use of statistical heuristic rule are the leading factors determine the quantity and types of training data sets that help in improving the performance of the model.

3. Quality of Training Data Sets

Just like quantity, the quality of machine learning training data set is another key factor, you need to keep in mind while developing an ML model. If the quality of machine learning training data sets is not good or accurate your model will never give accurate results, affecting the overall performance of the model not suitable to use in real-life.

Actually, there are different methods to measure the quality of the training data set. Standard quality-assurance methods and detailed for in-depth quality assessment are the leading two popular methods you can use to ensure the quality of data sets. Quality of data is important to get unbiased decisions from the ML models, so you need to make sure to use the right quality of training data sets to improve the performance of your ML model.

4. Supervised or Unsupervised ML

Moreover, the above-discussed ML algorithms, the performance of such AI-based models are affected by methods or process of machine learning. And supervised, unsupervised and reinforcement learning are the algorithm consist of a target/outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables).

In unsupervised machine learning, a model is given any target or outcome variable to predict/estimate. And, it is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. For supervised ML, labeled or annotated data is required, while for unsupervised ML the approach is different.

Similarly, reinforcement Learning is another important algorithm, used to train the model to make specific decisions. In this training process, the machine learns from previous experiences and tries to store the best suitable knowledge for the right predictions.

5. Model Validation and Testing

Building a machine learning model is not enough to get the right predictions, as you have to check the accuracy and need to validate the same to ensure get the precise results. And validating the model will improve the performance of the ML model.

Actually, there are various types of validation techniques you can follow but you need to make sure choose the best one that is suitable for your ML model validation and help you to improve the overall performance of your ML model and predict in an unbiased manner. Similarly, testing of the model is also important to ensure its accuracy and performance.

Summing-up

Improving machine learning model performance will not only make the model predict in an unbiased manner but make it one of the most reliable and acceptable in the AI world. Hence, a machine learning engineer and data scientist need to make sure all these points while working on such models to improve the overall performance of the AI model.

5. What is Feature Transformation ?

Feature transformation is a mathematical transformation in which we apply a mathematical formula to a particular column (feature) and transform the values, which are useful for our further analysis. It is a technique by which we can boost our model performance. It is also known as Feature Engineering, which creates new features from existing features that may help improve the model performance.

It refers to the algorithm family that creates new features using the existing features. These new features may not have the same interpretation as the original features, but they may have more explanatory power in a different space rather than in the original space. This can also be used for Feature Reduction. It can be done in many ways, by linear combinations of original features or using non-linear functions. It helps machine learning algorithms to converge faster.

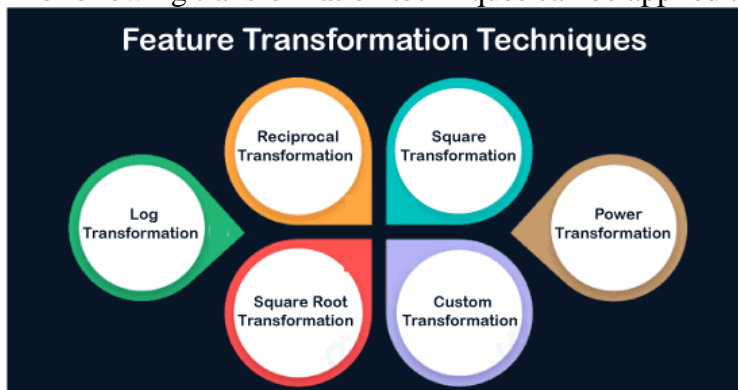
Why do we need Feature Transformations?

Like Linear and Logistic regression, some data science models assume that the variables follow a normal distribution. More likely, variables in real datasets will follow a skewed distribution. By applying some transformations to these skewed variables, we can map this skewed distribution to a normal distribution to increase the performance of our models.

As we know, Normal Distribution is a very important distribution in Statistics, which is key to many statisticians for solving problems in statistics. Usually, the data distribution in Nature follows a Normal distribution like - age, income, height, weight, etc. But the features in the real-life data are not normally distributed. However, it is the best approximation when we are unaware of the underlying distribution pattern.

Feature Transformation Techniques

The following transformation techniques can be applied to data sets, such as:



Feature Transformation in Data Mining

1. **Log Transformation:** Generally, these transformations make our data close to a normal distribution but cannot exactly abide by a normal distribution. This transformation is not applied to those features which have negative values. This transformation is mostly applied to right-skewed data. Convert data from the additive scale to multiplicative scale, i.e., linearly distributed data.
2. **Reciprocal Transformation:** This transformation is not defined for zero. It is a powerful transformation with a radical effect. This transformation reverses the order among values of the same sign, so large values become smaller and vice-versa.
3. **Square Transformation:** This transformation mostly applies to left-skewed data.
4. **Square Root Transformation:** This transformation is defined only for positive numbers. This can be used for reducing the skewness of right-skewed data. This transformation is weaker than Log Transformation.

5. Custom Transformation: A Function Transformer forwards its X (and optionally y) arguments to a user-defined function or function object and returns this function's result. The resulting transformer will not be pickle able if lambda is used as the function. This is useful for stateless transformations such as taking the log of frequencies, doing custom scaling, etc.

6. Power Transformations: Power transforms are a family of parametric, monotonic transformations that make data more Gaussian-like. The optimal parameter for stabilizing variance and minimizing skewness is estimated through maximum likelihood. This is useful for modeling issues related to non-constant variance or other situations where normality is desired. Currently, Power Transformer supports the Box-Cox transform and the Yeo-Johnson transform.

Box-cox requires the input data to be strictly positive (not even zero is acceptable), while Yeo-Johnson supports both positive and negative data.

6. Feature Subset Selection ?

What is Feature Selection?

A feature is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection. Each machine learning process depends on feature engineering, which mainly contains two processes; which are Feature Selection and Feature Extraction. Although feature selection and extraction processes may have the same objective, both are completely different from each other. The main difference between them is that feature selection is about selecting the subset of the original feature set, whereas feature extraction creates new features. Feature selection is a way of reducing the input variable for the model by using only relevant data in order to reduce overfitting in the model.

So, we can define feature Selection as, "It is a process of automatically or manually selecting the subset of most appropriate and relevant features to be used in model building." Feature selection is performed by either including the important features or excluding the irrelevant features in the dataset without changing them.

Need for Feature Selection

Before implementing any technique, it is really important to understand, need for the technique and so for the Feature Selection. As we know, in machine learning, it is necessary to provide a pre-processed and good input dataset in order to get better outcomes. We collect a huge amount of data to train our model and help it to learn better. Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data. Moreover, the huge amount of data also slows down the training process of the model, and with noise and irrelevant data, the model may not predict and perform well. So, it is very necessary to remove such noises and less-important data from the dataset and to do this, and Feature selection techniques are used.

Selecting the best features helps the model to perform well.

For example, suppose we want to create a model that automatically decides which car should be crushed for a spare part, and to do this, we have a dataset. This dataset contains a Model of the car, Year, Owner's name, Miles. So, in this dataset, the name of the owner does not contribute to the model performance as it does not decide if the car should be crushed or not, so we can remove this column and select the rest of the features (column) for the model building.

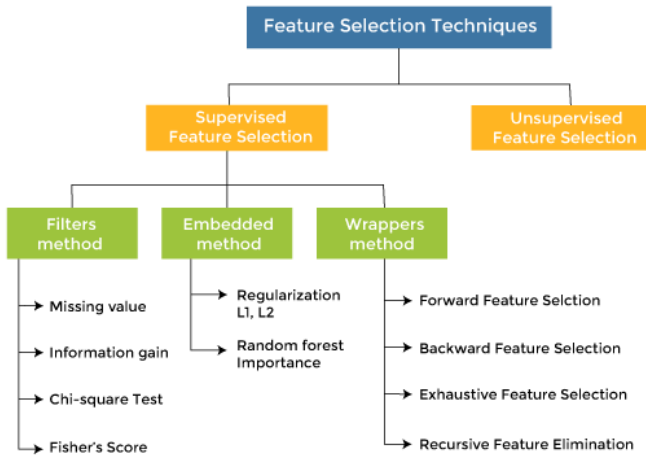
Below are some benefits of using feature selection in machine learning:

- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- It reduces the training time.
- It reduces overfitting hence enhance the generalization.

Feature Selection Techniques

There are mainly two types of Feature Selection techniques, which are:

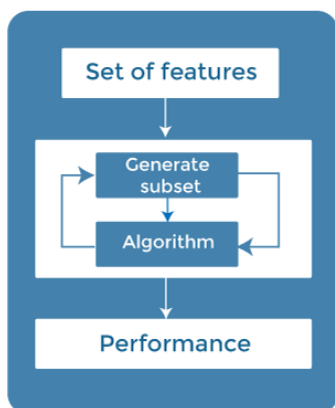
- **Supervised Feature Selection technique**
Supervised Feature selection techniques consider the target variable and can be used for the labelled dataset.
- **Unsupervised Feature Selection technique**
Unsupervised Feature selection techniques ignore the target variable and can be used for the unlabelled dataset.



There are mainly three techniques under supervised feature Selection:

1. Wrapper Methods

In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively.



On the basis of the output of the model, features are added or subtracted, and with this feature set, the model has trained again.

Some techniques of wrapper methods are:

- Forward selection - Forward selection is an iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of a new variable/feature does not improve the performance of the model.
- Backward elimination - Backward elimination is also an iterative approach, but it is the opposite of forward selection. This technique begins the process by considering all the features and removes the least significant feature. This elimination process continues until removing the features does not improve the performance of the model.
- Exhaustive Feature Selection- Exhaustive feature selection is one of the best feature selection methods, which evaluates each feature set as brute-force. It means this method tries & make each possible combination of features and return the best performing feature set.

- Recursive Feature Elimination-

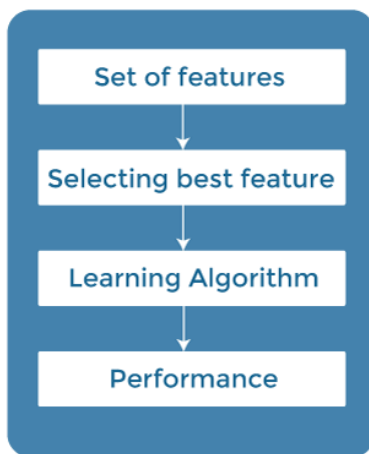
Recursive feature elimination is a recursive greedy optimization approach, where features are selected by recursively taking a smaller and smaller subset of features. Now, an estimator is trained with each set of features, and the importance of each feature is determined using `coef_attribute` or through a `feature_importances_attribute`.

2. Filter Methods

In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.

The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.

The advantage of using filter methods is that it needs low computational time and does not overfit the data.



Some common techniques of Filter methods are as follows:

- Information Gain
- Chi-square Test
- Fisher's Score
- Missing Value Ratio

Information Gain: Information gain determines the reduction in entropy while transforming the dataset. It can be used as a feature selection technique by calculating the information gain of each variable with respect to the target variable.

Chi-square Test: Chi-square test is a technique to determine the relationship between the categorical variables. The chi-square value is calculated between each feature and the target variable, and the desired number of features with the best chi-square value is selected.

Fisher's Score:

Fisher's score is one of the popular supervised technique of features selection. It returns the rank of the variable on the fisher's criteria in descending order. Then we can select the variables with a large fisher's score.

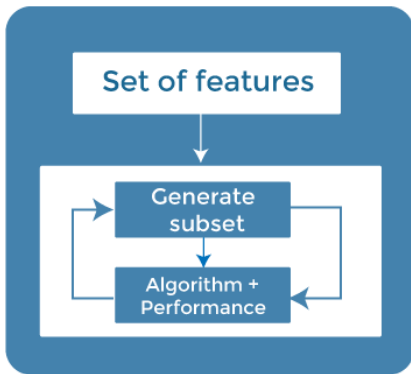
Missing Value Ratio:

The value of the missing value ratio can be used for evaluating the feature set against the threshold value. The formula for obtaining the missing value ratio is the number of missing values in each column divided by the total number of observations. The variable is having more than the threshold value can be dropped.

$$\text{Missing Value Ratio} = \frac{\text{Number of Missing values} \times 100}{\text{Total number of observations}}$$

3. Embedded Methods

Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method.

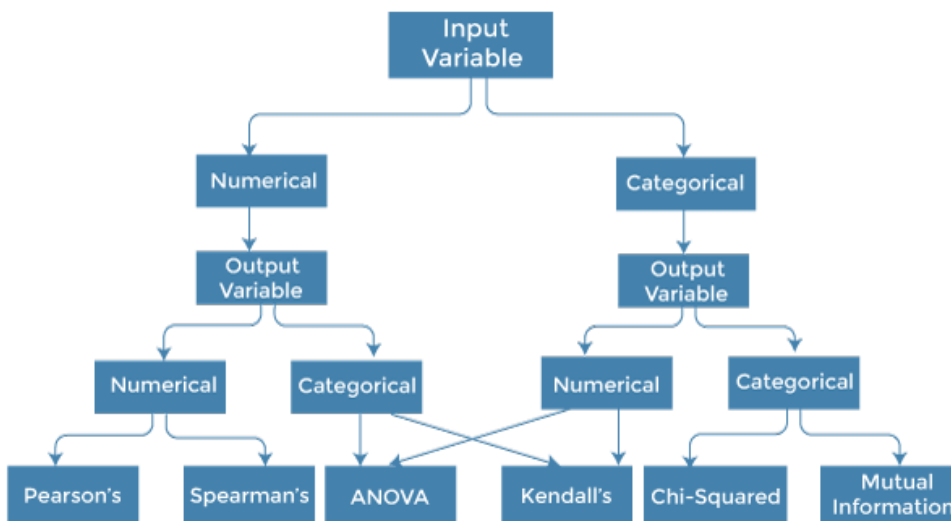


These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration. Some techniques of embedded methods are:

- Regularization- Regularization adds a penalty term to different parameters of the machine learning model for avoiding overfitting in the model. This penalty term is added to the coefficients; hence it shrinks some coefficients to zero. Those features with zero coefficients can be removed from the dataset. The types of regularization techniques are L1 Regularization (Lasso Regularization) or Elastic Nets (L1 and L2 regularization).
- Random Forest Importance - Different tree-based methods of feature selection help us with feature importance to provide a way of selecting features. Here, feature importance specifies which feature has more importance in model building or has a great impact on the target variable. Random Forest is such a tree-based method, which is a type of bagging algorithm that aggregates a different number of decision trees. It automatically ranks the nodes by their performance or decrease in the impurity (Gini impurity) over all the trees. Nodes are arranged as per the impurity values, and thus it allows to pruning of trees below a specific node. The remaining nodes create a subset of the most important features.

How to choose a Feature Selection Method?

For machine learning engineers, it is very important to understand that which feature selection method will work properly for their model. The more we know the datatypes of variables, the easier it is to choose the appropriate statistical measure for feature selection.



To know this, we need to first identify the type of input and output variables. In machine learning, variables are of mainly two types:

- Numerical Variables: Variable with continuous values such as integer, float
- Categorical Variables: Variables with categorical values such as Boolean, ordinal, nominals.

Below are some univariate statistical measures, which can be used for filter-based feature selection:

1. Numerical Input, Numerical Output:

Numerical Input variables are used for predictive regression modelling. The common method to be used for such a case is the Correlation coefficient.

- Pearson's correlation coefficient (For linear Correlation).
- Spearman's rank coefficient (for non-linear correlation).

2. Numerical Input, Categorical Output:

Numerical Input with categorical output is the case for classification predictive modelling problems. In this case, also, correlation-based techniques should be used, but with categorical output.

- ANOVA correlation coefficient (linear).
- Kendall's rank coefficient (nonlinear).

3. Categorical Input, Numerical Output:

This is the case of regression predictive modelling with categorical input. It is a different example of a regression problem. We can use the same measures as discussed in the above case but in reverse order.

4. Categorical Input, Categorical Output:

This is a case of classification predictive modelling with categorical Input variables.

The commonly used technique for such a case is Chi-Squared Test. We can also use Information gain in this case.

Conclusion

Feature selection is a very complicated and vast field of machine learning, and lots of studies are already made to discover the best methods. There is no fixed rule of the best feature selection method. However, choosing the method depend on a machine learning engineer who can combine and innovate approaches to find the best method for a specific problem. One should try a variety of model fits on different subsets of features selected through different statistical Measures.

UNIT V NON-PARAMETRIC MACHINE LEARNING

k- Nearest Neighbors- Decision Trees – Branching – Greedy Algorithm - Multiple Branches – Continuous attributes – Pruning. Random Forests: ensemble learning. Boosting – Adaboost algorithm. Support Vector Machines – Large Margin Intuition – Loss Function - Hinge Loss – SVM Kernels

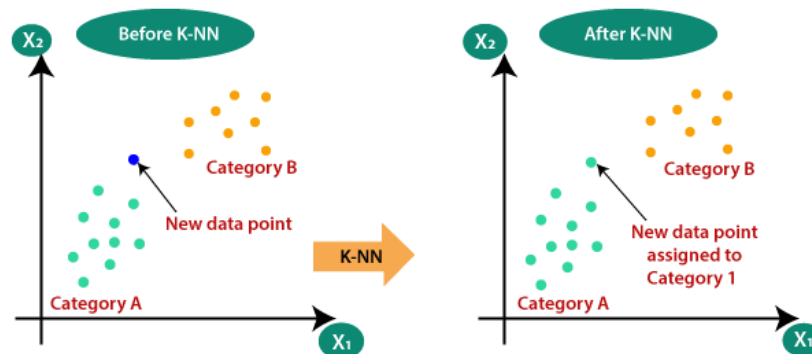
1. K-Nearest Neighbor (KNN) Algorithm for Machine Learning

- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So, for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dog's images and based on the most similar features it will put it in either cat or dog category.



Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x_1 , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

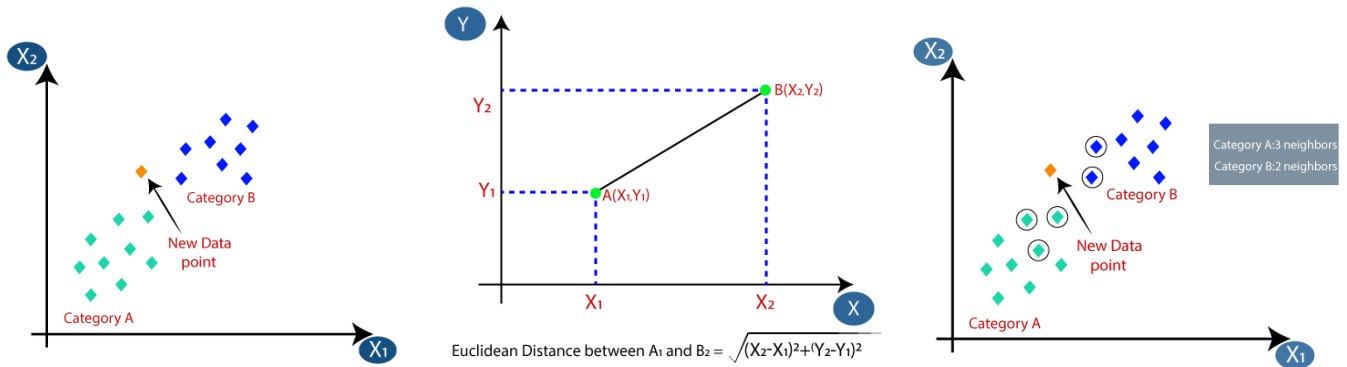


How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:
- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:
- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

How to select the value of K in the K-NN Algorithm?

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Problem for K-NN Algorithm: There is a Car manufacturer company that has manufactured a new SUV car. The company wants to give the ads to the users who are interested in buying that SUV. So for this problem, we have a dataset that contains multiple user's information through the social network. The dataset contains lots of information but the **Estimated Salary** and **Age** we will consider for the independent variable and the **Purchased variable** is for the dependent variable. Below is the dataset:

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

Steps to implement the K-NN algorithm:

- Data Pre-processing step
- Fitting the K-NN algorithm to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

Data Pre-Processing Step:

The Data Pre-processing step will remain exactly the same as Logistic Regression. Below is the code for it: By executing the above code, our dataset is imported to our program and well pre-processed. After feature scaling our test dataset will look like:

From the above output image, we can see that our data is successfully scaled.

Fitting K-NN classifier to the Training data:

Now we will fit the K-NN classifier to the training data. To do this we will import the **KNeighborsClassifier** class of **Sklearn Neighbors** library. After importing the class, we will create the **Classifier** object of the class. The Parameter of this class will be

- **n_neighbors:** To define the required neighbors of the algorithm. Usually, it takes 5.
- **metric='minkowski':** This is the default parameter and it decides the distance between the points.
- **p=2:** It is equivalent to the standard Euclidean metric.

And then we will fit the classifier to the training data. Below is the code for it:

```
from sklearn.neighbors import KNeighborsClassifier #Fitting K-NN classifier to the training set
classifier= KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2 )
classifier.fit(x_train, y_train)
```

Predicting the Test Result: To predict the test set result, we will create a **y_pred** vector as we did in Logistic Regression. Below is the code for it:

```
#Predicting the test set result
y_pred= classifier.predict(x_test)
```

Creating the Confusion Matrix:

Now we will create the Confusion Matrix for our K-NN model to see the accuracy of the classifier. Below is the code for it:

```
from sklearn.metrics import confusion_matrix #Creating the Confusion matrix
cm= confusion_matrix(y_test, y_pred)
```

Visualizing the Training set result:

Now, we will visualize the training set result for K-NN model. The code will remain same as we did in Logistic Regression, except the name of the graph.

The output graph is different from the graph which we have occurred in Logistic Regression. It can be understood in the below points:

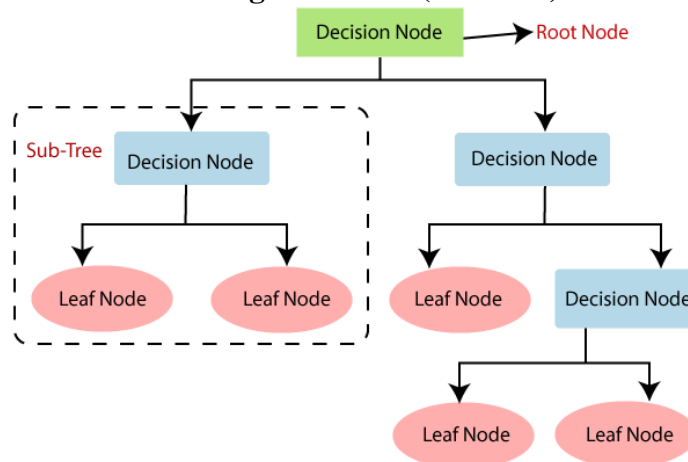
- As we can see the graph is showing the red point and green points. The green points are for Purchased(1) and Red Points for not Purchased(0) variable.
- The graph is showing an irregular boundary instead of showing any straight line or any curve because it is a K-NN algorithm, i.e., finding the nearest neighbor.
- The graph has classified users in the correct categories as most of the users who didn't buy the SUV are in the red region and users who bought the SUV are in the green region.
- The graph is showing good result but still, there are some green points in the red region and red points in the green region. But this is no big issue as by doing this model is prevented from overfitting issues.
- Hence our model is well trained.

Visualizing the Test set result:

After the training of the model, we will now test the result by putting a new dataset, i.e., Test dataset. Code remains the same except some minor changes: such as **x_train** and **y_train** will be replaced by **x_test** and **y_test**.

2. Decision Tree Classification Algorithm:

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome**.
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node**. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- *It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.*
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm**, which stands for **Classification and Regression Tree algorithm**.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
- Below diagram explains the general structure of a decision tree:
A decision tree can contain categorical data (YES/NO) as well as numeric data.



Why use Decision Trees?

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

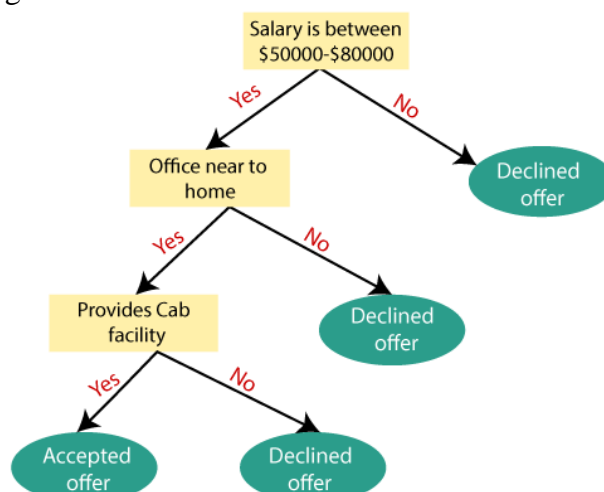
How does the Decision Tree algorithm Work?

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM)**.
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM**. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain**
- **Gini Index**

1. Information Gain:

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

$$\text{Information Gain} = \text{Entropy}(S) - [(\text{Weighted Avg}) * \text{Entropy}(\text{each feature})]$$

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$\text{Entropy}(s) = -P(\text{yes}) \log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{no})$$

Where,

- **S= Total number of samples**
- **P(yes)= probability of yes**
- **P(no)= probability of no**

2. Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

Pruning: Getting an Optimal Decision tree

Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree **pruning** technology used:

- **Cost Complexity Pruning**
- **Reduced Error Pruning.**

Advantages of the Decision Tree

- It is simple to understand as it follows the same process which a human follow while making any decision in real-life.
- It can be very useful for solving decision-related problems.
- It helps to think about all the possible outcomes for a problem.
- There is less requirement of data cleaning compared to other algorithms.

Disadvantages of the Decision Tree

- The decision tree contains lots of layers, which makes it complex.
- It may have an overfitting issue, which can be resolved using the **Random Forest algorithm**.
- For more class labels, the computational complexity of the decision tree may increase.

Python Implementation of Decision Tree

Now we will implement the Decision tree using Python. For this, we will use the dataset "user_data.csv," which we have used in previous classification models. By using the same dataset, we can compare the Decision tree classifier with other classification models such as KNN SVM, LogisticRegression, etc.

- **Data Pre-processing step**
- **Fitting a Decision-Tree algorithm to the Training set**
- **Predicting the test result**
- **Test accuracy of the result(Creation of Confusion matrix)**
- **Visualizing the test set result.**

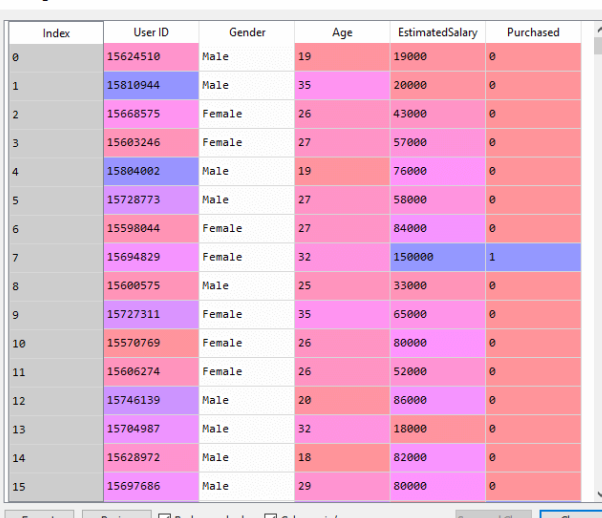
1. Data Pre-Processing Step:

Below is the code for the pre-processing step:

```
import numpy as nm # importing libraries
import matplotlib.pyplot as mtp
import pandas as pd

data_set= pd.read_csv('user_data.csv') #importing datasets
x= data_set.iloc[:, [2,3]].values #Extracting Independent and dependent Variable
y= data_set.iloc[:, 4].values
from sklearn.model_selection import train_test_split # Splitting the dataset into training and test set.
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=0)
from sklearn.preprocessing import StandardScaler #feature Scaling
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
```

In the above code, we have pre-processed the data. Where we have loaded the dataset, which is given as:



Index	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15600575	Male	25	33000	0
9	15727311	Female	35	65000	0
10	15570769	Female	26	80000	0
11	15606274	Female	26	52000	0
12	15746139	Male	20	86000	0
13	15704987	Male	32	18000	0
14	15628972	Male	18	82000	0
15	15697686	Male	29	80000	0

2. Fitting a Decision-Tree algorithm to the Training set

Now we will fit the model to the training set. For this, we will import the **DecisionTreeClassifier** class from **sklearn.tree** library. Below is the code for it:

```
#Fitting Decision Tree classifier to the training set
From sklearn.tree import DecisionTreeClassifier
classifier= DecisionTreeClassifier(criterion='entropy', random_state=0)
classifier.fit(x_train, y_train)
```

In the above code, we have created a classifier object, in which we have passed two main parameters;

- **"criterion='entropy':** Criterion is used to measure the quality of split, which is calculated by information gain given by entropy.
- **random_state=0":** For generating the random states.

3. Predicting the test result

Now we will predict the test set result. We will create a new prediction vector **y_pred**. Below is the code for it:

1. #Predicting the test set result
2. `y_pred= classifier.predict(x_test)`

Output:

In the below output image, the predicted output and real test output are given. We can clearly see that there are some values in the prediction vector, which are different from the real vector values. These are prediction errors.

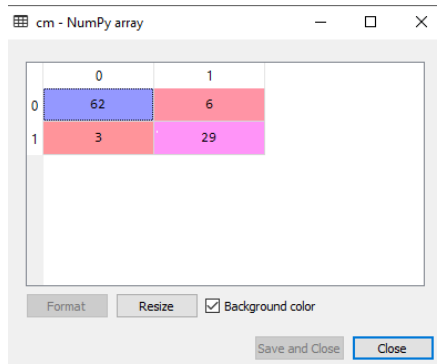


4. Test accuracy of the result (Creation of Confusion matrix)

In the above output, we have seen that there were some incorrect predictions, so if we want to know the number of correct and incorrect predictions, we need to use the confusion matrix. Below is the code for it:

1. #Creating the Confusion matrix
2. `from sklearn.metrics import confusion_matrix`
3. `cm= confusion_matrix(y_test, y_pred)`

Output:



In the above output image, we can see the confusion matrix, which has **6+3= 9 incorrect predictions** and **62+29=91 correct predictions**. Therefore, we can say that compared to other classification models, the Decision Tree classifier made a good prediction.

5. Visualizing the training set result:

Here we will visualize the training set result. To visualize the training set result we will plot a graph for the decision tree classifier. The classifier will predict yes or No for the users who have either Purchased or Not purchased the SUV car as we did in Logistic Regression. Below is the code for it:

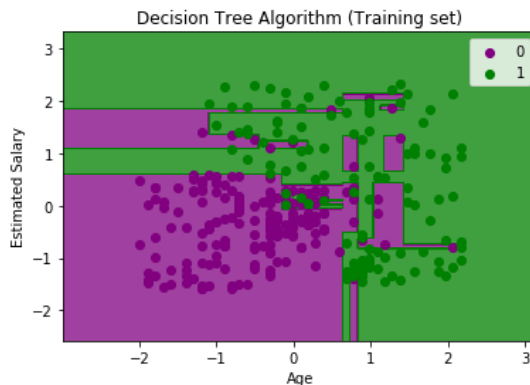
```
#Visualizing the training set result
from matplotlib.colors import ListedColormap
x_set, y_set = x_train, y_train
x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),
nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))
```

```

mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
alpha = 0.75, cmap = ListedColormap(('purple','green' )))
mtp.xlim(x1.min(), x1.max())
mtp.ylim(x2.min(), x2.max())
fori, j in enumerate(nm.unique(y_set)):
mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
            c = ListedColormap(('purple', 'green'))(i), label = j)
mtp.title('Decision Tree Algorithm (Training set)')
mtp.xlabel('Age')
mtp.ylabel('Estimated Salary')
mtp.legend()
mtp.show()

```

Output:



The above output is completely different from the rest classification models. It has both vertical and horizontal lines that are splitting the dataset according to the age and estimated salary variable. As we can see, the tree is trying to capture each dataset, which is the case of overfitting.

6. Visualizing the test set result:

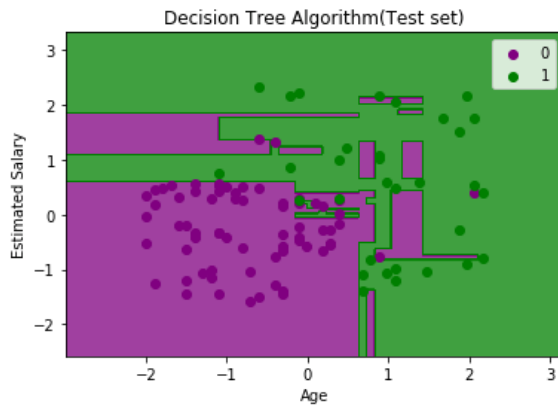
Visualization of test set result will be similar to the visualization of the training set except that the training set will be replaced with the test set.

```

#Visulaizing the test set result
from matplotlib.colors import ListedColormap
x_set, y_set = x_test, y_test
x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),
nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))
mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
alpha = 0.75, cmap = ListedColormap(('purple','green' )))
mtp.xlim(x1.min(), x1.max())
mtp.ylim(x2.min(), x2.max())
fori, j in enumerate(nm.unique(y_set)):
mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
            c = ListedColormap(('purple', 'green'))(i), label = j)
mtp.title('Decision Tree Algorithm(Test set)')
mtp.xlabel('Age')
mtp.ylabel('Estimated Salary')
mtp.legend()
mtp.show()

```

Output:



3. Greedy Algorithm:

The greedy method is one of the strategies like Divide and conquer used to solve the problems. This method is used for solving optimization problems. An optimization problem is a problem that demands either maximum or minimum results.

The Greedy method is the simplest and straightforward approach. It is not an algorithm, but it is a technique. The main function of this approach is that the decision is taken on the basis of the currently available information. Whatever the current information is present, the decision is made without worrying about the effect of the current decision in future.

This technique is basically used to determine the feasible solution that may or may not be optimal. The feasible solution is a subset that satisfies the given criteria. The optimal solution is the solution which is the best and the most favorable solution in the subset. In the case of feasible, if more than one solution satisfies the given criteria then those solutions will be considered as the feasible, whereas the optimal solution is the best solution among all the solutions.

Characteristics of Greedy method

The following are the characteristics of a greedy method:

- To construct the solution in an optimal way, this algorithm creates two sets where one set contains all the chosen items, and another set contains the rejected items.
- A Greedy algorithm makes good local choices in the hope that the solution should be either feasible or optimal.

Components of Greedy Algorithm

The components that can be used in the greedy algorithm are:

- **Candidate set:** A solution that is created from the set is known as a candidate set.
- **Selection function:** This function is used to choose the candidate or subset which can be added in the solution.
- **Feasibility function:** A function that is used to determine whether the candidate or subset can be used to contribute to the solution or not.
- **Objective function:** A function is used to assign the value to the solution or the partial solution.
- **Solution function:** This function is used to intimate whether the complete function has been reached or not.

Applications of Greedy Algorithm

- It is used in finding the shortest path.
- It is used to find the minimum spanning tree using the prim's algorithm or the Kruskal's algorithm.
- It is used in a job sequencing with a deadline.
- This algorithm is also used to solve the fractional knapsack problem.

Pseudo code of Greedy Algorithm

```
Algorithm Greedy (a, n)
{
  Solution := 0;
  for i = 0 to n do
  {
    x := select(a);
    if feasible(solution, x)
    {
      Solution := union(solution, x)
    }
  }
  return solution;
} }
```

The above is the greedy algorithm. Initially, the solution is assigned with zero value. We pass the array and number of elements in the greedy algorithm. Inside the for loop, we select the element one by one and check whether the solution is feasible or not. If the solution is feasible, then we perform the union.

Let's understand through an example.

Suppose there is a problem 'P'. I want to travel from A to B shown as below:

P : A → B

The problem is that we have to travel this journey from A to B. There are various solutions to go from A to B. We can go from A to B by **walk, car, bike, train, aeroplane**, etc. There is a constraint in the journey that we have to travel this journey within 12 hrs. If I go by train or aeroplane then only, I can cover this distance within 12 hrs. There are many solutions to this problem but there are only two solutions that satisfy the constraint.

If we say that we have to cover the journey at the minimum cost. This means that we have to travel this distance as minimum as possible, so this problem is known as a minimization problem. Till now, we have two feasible solutions, i.e., one by train and another one by air. Since travelling by train will lead to the minimum cost so it is an optimal solution. An optimal solution is also the feasible solution, but providing the best result so that solution is the optimal solution with the minimum cost. There would be only one optimal solution.

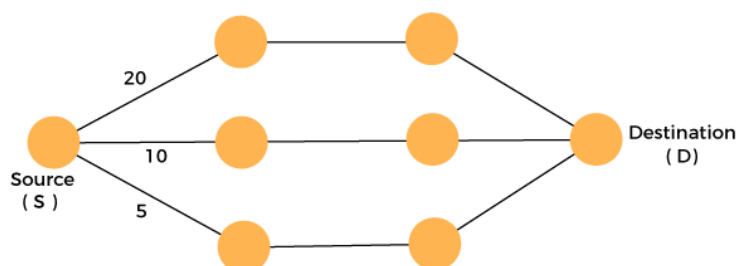
The problem that requires either minimum or maximum result then that problem is known as an optimization problem. Greedy method is one of the strategies used for solving the optimization problems.

Disadvantages of using Greedy algorithm

Greedy algorithm makes decisions based on the information available at each phase without considering the broader problem. So, there might be a possibility that the greedy solution does not give the best solution for every problem.

It follows the local optimum choice at each stage with a intend of finding the global optimum. Let's understand through an example.

Consider the graph which is given below:



We have to travel from the source to the destination at the minimum cost. Since we have three feasible solutions having cost paths as 10, 20, and 5. 5 is the minimum cost path so it is the optimal solution. This is

the local optimum, and in this way, we find the local optimum at each stage in order to calculate the global optimal solution.

Continuous attributes

What are Continuous Variables?

Simply put, if a variable can take any value between its minimum and maximum value, then it is called a continuous variable. By nature, a lot of things we deal with fall in this category: age, weight, height being some of them.

Just to make sure the difference is clear, let me ask you to classify whether a variable is continuous or categorical:

1. Gender of a person
2. Number of siblings of a Person
3. Time on which a laptop runs on battery

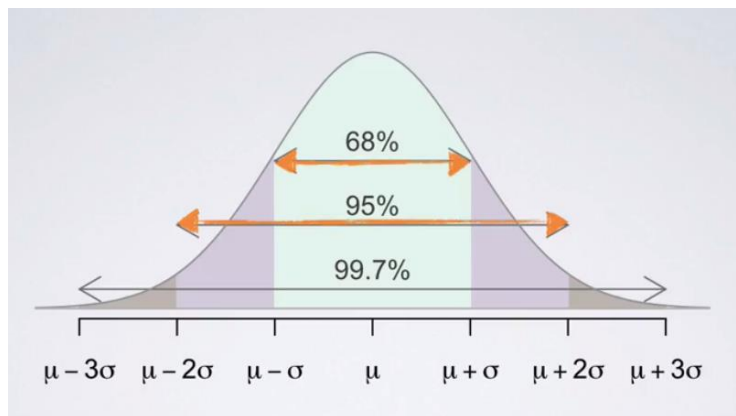
Methods to deal with Continuous Variables

Binning The Variable:

Binning refers to dividing a list of continuous variables into groups. It is done to discover set of patterns in continuous variables, which are difficult to analyze otherwise. Also, bins are easy to analyze and interpret. But, it also leads to loss of information and loss of power. Once the bins are created, the information gets compressed into groups which later affects the final model. Hence, it is advisable to create small bins initially. This would help in minimal loss of information and produces better results.

Normalization:

In simpler words, it is a process of comparing variables at a 'neutral' or 'standard' scale. It helps to obtain same range of values. Normally distributed data is easy to read and interpret. As shown below, in a normally distributed data, 99.7% of the observations lie within 3 standard deviations from the mean. Also, the mean is zero and standard deviation is one. Normalization technique is commonly used in algorithms such as k-means, clustering etc.



A commonly used normalization method is z-scores. Z score of an observation is the number of standard deviations it falls above or below the mean. It's formula is shown below.

$$Z = \frac{x - \mu}{\sigma}$$

x = observation, μ = mean (population), σ = standard deviation (population)

For example:

Randy scored 76 in maths test. Katie score 86 in science test. Maths test has (mean = 70, sd = 2). Science test has (mean = 80, sd = 3)

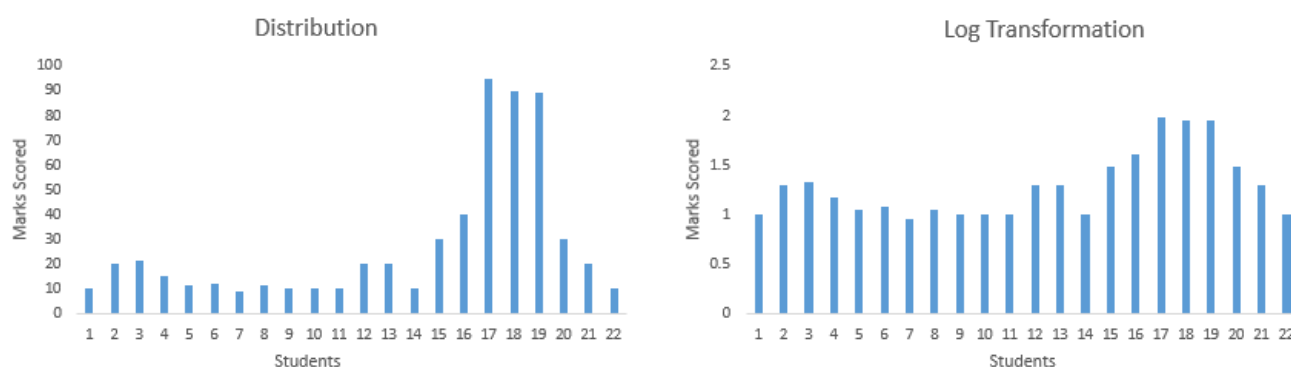
$$z(\text{Randy}) = (76 - 70)/2 = 3$$

$$z(\text{Katie}) = (86 - 80)/3 = 2$$

Transformations for Skewed Distribution:

Transformation is required when we encounter highly skewed data. It is suggested not to work on skewed data in its raw form. Because, it reduces the impact of low frequency values which could be equally significant. At times, skewness is influenced by presence of outliers. Hence, we need to be careful while using this approach. The technique to deal with outliers is explained in next sections.

There are various types of transformation methods. Some are Log, sqrt, exp, Box-cox, power etc. The commonly used method is Log Transformation.



Principal Component Analysis:

Sometime data set has too many variables. May be, 100, 200 variables or even more. In such cases, you can't build a model on all variables. Reason being, 1) It would be time consuming. 2) It might have lots of noise 3) A lot of variables will tell similar information

Hence, to avoid such situation we use PCA a.k.a Principal Component Analysis. It is nothing but, finding out few '**principal**' variables which explain significant amount of variation in dependent variable. Using this technique, a large number of variables are reduced to few significant variables. This technique helps to reduce noise, redundancy and enables quick computations.

Factor Analysis:

Factor Analysis was invented by Charles Spearman (1904). This is a variable reduction technique. It is used to determine factor structure or model. It also explains the maximum amount of variance in the model. Let's say some variables are highly correlated. These variables can be grouped by their correlations i.e., all variables in a particular group can be highly correlated among themselves but have low correlation with variables of other group(s). Here each group represents a single underlying construct or factor. Factor analysis is of two types:

1. EFA (Exploratory Factor Analysis) – It identifies and summarizes the underlying correlation structure in a data set
2. CFA (Confirmatory Factor Analysis) – It attempts to confirm hypothesis using the correlation structure and rate 'goodness of fit'.

Methods to work with Date & Time Variable

Presence of Data Time variable in a data set usually give lots of confidence. Seriously! It does. Because, in data-time variable, you get lots of scope to practice the techniques learnt above. You can create bins, you can create new features, convert its type etc. Date & Time is commonly found in this format:

DD-MM-YYY HH:SS or MM-DD-YYY HH:SS

Pruning

When the size of the features exceeds a certain limit, regression trees become inapplicable due to overfitting. The decision tree's overfitting problem is caused by other factors as well as such as branches sometimes are impacted by noise and outliers of data. Pruning is a critical step in constructing tree based machine learning models that help overcome these issues.

1. A snippet about decision trees
2. About pruning
3. Strategies for pruning
4. Pruning methods

A decision tree is a traditional supervised machine learning technique.

About pruning

Pruning is the process of eliminating weight connections from a network to speed up inference and reduce model storage size. Decision trees and neural networks, in general, are overparameterized. Pruning a network entails deleting unneeded parameters from an overly parameterized network.

Pruning mostly serves as an architectural search inside the tree or network. In fact, because pruning functions as a regularizer, a model will often generalise slightly better at low levels of sparsity. The trimmed model will match the baseline at higher levels. If you push it too far, the model will start to generalise worse than the baseline, but with greater performance.

Need for pruning

Pruning a classifier simplifies it by combining disjuncts that are adjacent in instance space. By removing error-prone components, the classifier's performance may be improved. It also permits additional model analysis for the aim of knowledge gain. Pruning should never be used to remove predicted components of a classifier. As a result, the pruning operation needs a technique for determining if a group of disjuncts is predictive or should be merged into a single, bigger disjunct.

The pruned disjunct represents the "null hypothesis" in a significance test, whereas the unpruned disjuncts represent the "alternative hypothesis." The test determines if the data offer adequate evidence to support the alternative. If this is the case, the unpruned disjuncts are left alone; otherwise, pruning continues.

The obvious rationale for significance tests is that they evaluate whether the apparent correlation between a collection of disjuncts and the data is likely to be attributable to chance alone. They do so by calculating the likelihood of generating a random relationship as least as strong as the observed association if the null hypothesis is confirmed. If the observed relationship is unlikely to be attributable to chance and this likelihood does not exceed a set threshold, the unpruned disjuncts are deemed to be predictive; otherwise, the model is simplified. The aggressiveness of the pruning operation is determined by the "significance level" criterion used in the test.

Strategies for pruning

Pruning is a critical step in developing a decision tree model. Pruning is commonly employed to alleviate the overfitting issue in decision trees. Pre-pruning and post-pruning are two common model tree generating procedures.

Pre pruning

Prepruning is the process of pruning the model by halting the tree's formation in advance. When construction is completed, the leaf nodes inherit the label of the most common class in the subset that is connected to the current node. There are various ways for pre-pruning, including the following

- When the model reaches a specific height, the decision tree's growth is stopped.
- When the eigenvectors of instances associated with a node are identical, the tree model stops developing.
- When the number of occurrences within a node falls below a certain threshold, the tree stops growing. The downside of this strategy is that it is inapplicable not in particular circumstances where the amount of data is tiny.
- An expansion is a process of dividing a node into two child nodes. When the gain value of an expansion falls below a certain threshold, the tree model stops expanding as well.

The major disadvantage of pre-pruning is the narrow viewing field, which implies that the tree's current expansion may not match the standards, but later expansion may. In this situation, the decision tree's development is halted early.

Post-pruning

The decision tree generation is divided into two steps by post-pruning. The first step is the tree-building process, with the termination condition that the fraction of a certain class in the node reaches 100%, and the second phase is pruning the tree structure gained in the first phase.

Post-pruning techniques circumvent the problem of a narrow viewing field in this way. As a result, post-pruning procedures are often more accurate than pre-pruning methods, therefore post-pruning methods are more widely utilised than pre-pruning methods. The pruning procedure identifies the node as a leaf node by using the label of the most common class in the subset associated with the current node, which is the same as in pre-pruning.

Pruning methods

The goal of pruning is to remove sections of a classification model that explain random variation in the training sample rather than actual domain characteristics. This makes the model more understandable to the user and, perhaps, more accurate on fresh data that was not used to train the classifier. An effective approach for differentiating sections of a classifier that are attributable to random effects from parts that describe significant structure is required for pruning. There are different methods for pruning listed in this article used in both strategies.

Reduced Error Pruning (REP)

The aim is to discover the most accurate subtree with the shortest version to the pruning set.

The pruning set is used to evaluate the efficacy of a subtree (branch) of a fully grown tree in this approach, which is conceptually the simplest. It starts with the entire tree and compares the number of classification mistakes made on the pruning set when the subtree is retained to the number of classification errors made when internal nodes are transformed into leaves and assigned to the best class for each internal node of the tree. The simplified tree can sometimes outperform the original tree. It is best to prune the subtree in this scenario. This branch trimming procedure is continued on the simplified tree until the misclassification rate rises. Another restriction limits the pruning condition: the internal node can be pruned only if it includes no subtree with a lower error rate than the internal node itself. This indicates that trimmed nodes are evaluated using a bottom-up traversal technique.

The advantage of this strategy is its linear computing complexity, as each node is only visited once to evaluate the possibility of trimming it. REP, on the other hand, has a proclivity towards over-pruning. This is because all evidence contained in the training set and used to construct a fully grown tree is ignored during the pruning step. This issue is most obvious when the pruning set is significantly smaller than the training set, but it becomes less significant as the percentage of instances in the pruning set grows.

Pessimistic Error Pruning (PEP)

The fact that the same training set is utilised for both growing and trimming a tree distinguishes this pruning strategy. The apparent error rate, that is, the error rate on the training set, is optimistic and cannot be used to select the best-pruned tree. As a result, the continuity correction for the binomial distribution was proposed, which may give "a more realistic error rate."

The distribution of errors at the node is roughly a binomial distribution. The binomial distribution's mean and variance are the likelihood of success and failure; the binomial distribution converges to a normal distribution. The PEP approach is regarded as one of the most accurate decision tree pruning algorithms available today. However, because the mechanism for traversing PEP is similar to pre-pruning, PEP suffers from excessive pruning. Furthermore, due to its top-down nature, each subtree in the tree only has to be consulted once, and the time complexity is in the worst-case linear with the number of non-leaf nodes in the decision tree.

Minimum Error Pruning (MEP)

This method is a bottom-up strategy that seeks a single tree with the lowest “anticipated error rate on an independent data set.” This does not indicate the adoption of a pruning set, but rather that the developer wants to estimate the error rate for unknown scenarios. Indeed, both the original and enhanced versions described exploiting just information from the training set.

In the presence of noisy data, Laplace probability estimation is employed to improve the performance of ID3. Later, the Bayesian technique was employed to enhance this procedure, and the approach is known as an m-probability estimation. There were two modifications:

- Prior probabilities are used in estimate rather than assuming a uniform starting distribution of classes.
- Several trees with differing degrees of pruning may be generated by adjusting the value of the parameter. The degree of pruning is now decided by parameters rather than the number of classes. Furthermore, factors like the degree of noise in the training data may be changed based on domain expertise or the complexity of the problem.

The predicted error rate for each internal node is estimated in the minimal error pruning approach and is referred to as static error. The anticipated error rate of the branch with the node is then estimated as a weighted sum of the expected error rates of the node’s children, where each weight represents the chance that observation in the node would reach the associated child.

Critical Value Pruning (CVP)

This post-pruning approach is quite similar to pre-pruning. Indeed, a crucial value threshold is defined for the node selection measure. Then, if the value returned by the selection measure for each test connected with edges flowing out of that node does not exceed the critical value, an internal node of the tree is pruned. However, a node may meet the pruning criterion but not all of its offspring. The branch is retained in this scenario because it includes significant nodes. This additional check is typical of a bottom-up strategy and distinguishes it from pre-pruning methods that prohibit a tree from developing even if future tests prove to be important.

The degree of pruning changes obviously with the critical value: a greater critical value results in more extreme pruning. The approach is divided into two major steps:

- Prune the mature tree to increase crucial values.
- Choose the best tree from the sequence of trimmed trees by weighing the tree’s overall relevance and forecasting abilities.

Cost-Complexity Pruning (CCP)

The CART pruning algorithm is another name for this approach. It is divided into two steps:

1. Using certain techniques, select a parametric family of subtrees from a fully formed tree.
2. The optimal tree is chosen based on an estimation of the real error rates of the trees in the parametric family.

In terms of the first phase, the primary concept is to prune the branches that exhibit the least increase in apparent error rate per cut leaf to produce the next best tree from the best tree. When a tree is pruned at a node, the apparent error rate increases by a certain amount while the number of leaves reduces by a certain number of units. As a result, the following ratio of the error rate increase to leaf reduction measures the rise in apparent error rate per trimmed leaf. The next best tree in the parametric family is then created by trimming all nodes in the subtree with the lowest value of the above-mentioned ratio.

The best tree in the entire grown tree in terms of predicted accuracy is picked in the second phase. The real error rate of each tree in the family may be estimated in two ways: one using cross-validation sets and the other using an independent pruning set.

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning**, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "**Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.**" Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Why use Random Forest?

Below are some points that explain why we should use the Random Forest algorithm: <="" li="">

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

How does Random Forest algorithm work?

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

The working of the algorithm can be better understood by the below example:

Example: Suppose there is a dataset that contains multiple fruit images. So, this dataset is given to the Random forest classifier. The dataset is divided into subsets and given to each decision tree. During the training phase, each decision tree produces a prediction result, and when a new data point occurs, then based on the majority of results, the Random Forest classifier predicts the final decision. Consider the below image:

Applications of Random Forest

There are mainly four sectors where Random forest mostly used:

1. **Banking:** Banking sector mostly uses this algorithm for the identification of loan risk.
2. **Medicine:** With the help of this algorithm, disease trends and risks of the disease can be identified.
3. **Land Use:** We can identify the areas of similar land use by this algorithm.
4. **Marketing:** Marketing trends can be identified using this algorithm.

Advantages of Random Forest

- Random Forest is capable of performing both Classification and Regression tasks.
- It is capable of handling large datasets with high dimensionality.
- It enhances the accuracy of the model and prevents the overfitting issue.

Disadvantages of Random Forest

- Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

Python Implementation of Random Forest Algorithm

Now we will implement the Random Forest Algorithm tree using Python. For this, we will use the same dataset "user_data.csv", which we have used in previous classification models. By using the same dataset, we can compare the Random Forest classifier with other classification models such as Decision tree Classifier, KNN, SVM, Logistic Regression, etc.

Implementation Steps are given below:

- Data Pre-processing step
- Fitting the Random forest algorithm to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

1.Data Pre-Processing Step:

Below is the code for the pre-processing step:

```
# importing libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd
#importing datasets
data_set= pd.read_csv('user_data.csv')
#Extracting Independent and dependent Variable
x= data_set.iloc[:, [2,3]].values
y= data_set.iloc[:, 4].values
# Splitting the dataset into training and test set.
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=0)
#feature Scaling
from sklearn.preprocessing import StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
```

In the above code, we have pre-processed the data. Where we have loaded the dataset, which is given as:

2. Fitting the Random Forest algorithm to the training set:

Now we will fit the Random forest algorithm to the training set. To fit it, we will import the **RandomForestClassifier** class from the **sklearn.ensemble** library. The code is given below:

1. #Fitting Decision Tree classifier to the training set
2. from sklearn.ensemble import RandomForestClassifier
3. classifier= RandomForestClassifier(n_estimators= 10, criterion="entropy")
4. classifier.fit(x_train, y_train)

In the above code, the classifier object takes below parameters:

- **n_estimators**= The required number of trees in the Random Forest. The default value is 10. We can choose any number but need to take care of the overfitting issue.
- **criterion**= It is a function to analyze the accuracy of the split. Here we have taken "entropy" for the information gain.

3. Predicting the Test Set result

Since our model is fitted to the training set, so now we can predict the test result. For prediction, we will create a new prediction vector `y_pred`. Below is the code for it:

1. `#Predicting the test set result`
2. `y_pred= classifier.predict(x_test)`

Output:

The prediction vector is given as:

By checking the above prediction vector and test set real vector, we can determine the incorrect predictions done by the classifier.

4. Creating the Confusion Matrix

Now we will create the confusion matrix to determine the correct and incorrect predictions. Below is the code for it:

1. `#Creating the Confusion matrix`
2. `from sklearn.metrics import confusion_matrix`
3. `cm= confusion_matrix(y_test, y_pred)`

Output: As we can see in the above matrix, there are **4+4= 8 incorrect predictions** and **64+28= 92 correct predictions**.

5. Visualizing the training Set result

Here we will visualize the training set result. To visualize the training set result we will plot a graph for the Random forest classifier. The classifier will predict yes or No for the users who have either Purchased or Not purchased the SUV car as we did in Logistic Regression. Below is the code for it:

1. `from matplotlib.colors import ListedColormap`
2. `x_set, y_set = x_train, y_train`
3. `x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),`
4. `nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))`
5. `mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),`
6. `alpha = 0.75, cmap = ListedColormap(('purple','green')))`
7. `mtp.xlim(x1.min(), x1.max())`
8. `mtp.ylim(x2.min(), x2.max())`
9. `for i, j in enumerate(nm.unique(y_set)):`
10. `mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],`
11. `c = ListedColormap(('purple', 'green'))(i), label = j)`
12. `mtp.title('Random Forest Algorithm (Training set)')`
13. `mtp.xlabel('Age')`
14. `mtp.ylabel('Estimated Salary')`
15. `mtp.legend()`
16. `mtp.show()`

6. Visualizing the test set result

Now we will visualize the test set result. Below is the code for it:

```
#Visulaizing the test set result
from matplotlib.colors import ListedColormap
x_set, y_set = x_test, y_test
x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),
nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))
mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
alpha = 0.75, cmap = ListedColormap(('purple','green' )))
```

```

mtp.xlim(x1.min(), x1.max())
mtp.ylim(x2.min(), x2.max())
for i, j in enumerate(nm.unique(y_set)):
    mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
               c = ListedColormap(('purple', 'green'))(i), label = j)
mtp.title('Random Forest Algorithm(Test set)')
mtp.xlabel('Age')
mtp.ylabel('Estimated Salary')
mtp.legend()
mtp.show()

```

Ensemble Learning

ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. Basic idea is to learn a set of classifiers (experts) and to allow them to vote.

Advantage: Improvement in predictive accuracy.

Disadvantage : It is difficult to understand an ensemble of classifiers.

Main Challenge for Developing Ensemble Models?

The main challenge is not to obtain highly accurate base models, but rather to obtain base models which make different kinds of errors. For example, if ensembles are used for classification, high accuracies can be accomplished if different base models misclassify different training examples, even if the base classifier accuracy is low.

Methods for Independently Constructing Ensembles –

- Majority Vote
- Bagging and Random Forest
- Randomness Injection
- Feature-Selection Ensembles
- Error-Correcting Output Coding

Methods for Coordinated Construction of Ensembles –

- Boosting
- Stacking

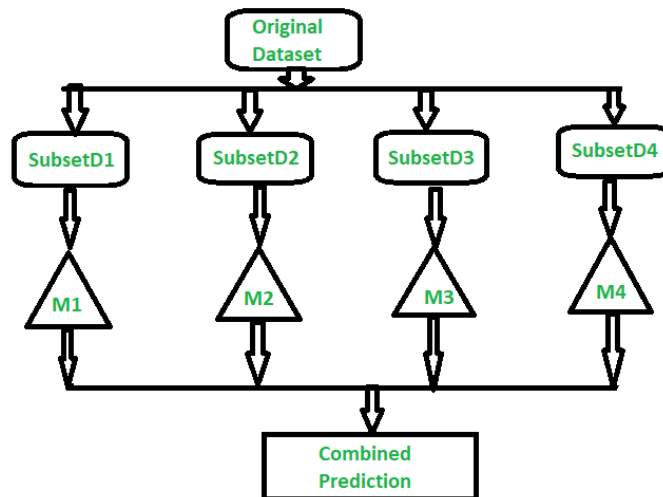
Types of Ensemble Classifier –

Bagging:

Bagging (Bootstrap Aggregation) is used to reduce the variance of a decision tree. Suppose a set D of d tuples, at each iteration i , a training set D_i of d tuples is sampled with replacement from D (i.e., bootstrap). Then a classifier model M_i is learned for each training set $D < i$. Each classifier M_i returns its class prediction. The bagged classifier M^* counts the votes and assigns the class with the most votes to X (unknown sample).

Implementation steps of Bagging –

1. Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.
2. A base model is created on each of these subsets.
3. Each model is learned in parallel from each training set and independent of each other.
4. The final predictions are determined by combining the predictions from all the models.

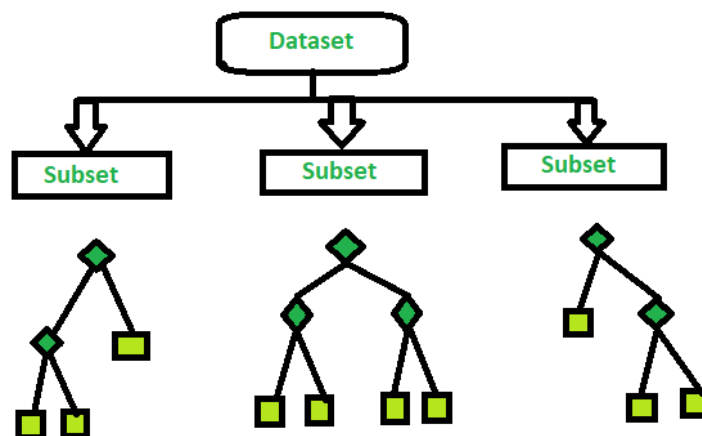


Random Forest:

Random Forest is an extension over bagging. Each classifier in the ensemble is a decision tree classifier and is generated using a random selection of attributes at each node to determine the split. During classification, each tree votes and the most popular class is returned.

Implementation steps of Random Forest –

1. Multiple subsets are created from the original data set, selecting observations with replacement.
2. A subset of features is selected randomly and whichever feature gives the best split is used to split the node iteratively.
3. The tree is grown to the largest.
4. Repeat the above steps and prediction is given based on the aggregation of predictions from n number of trees.



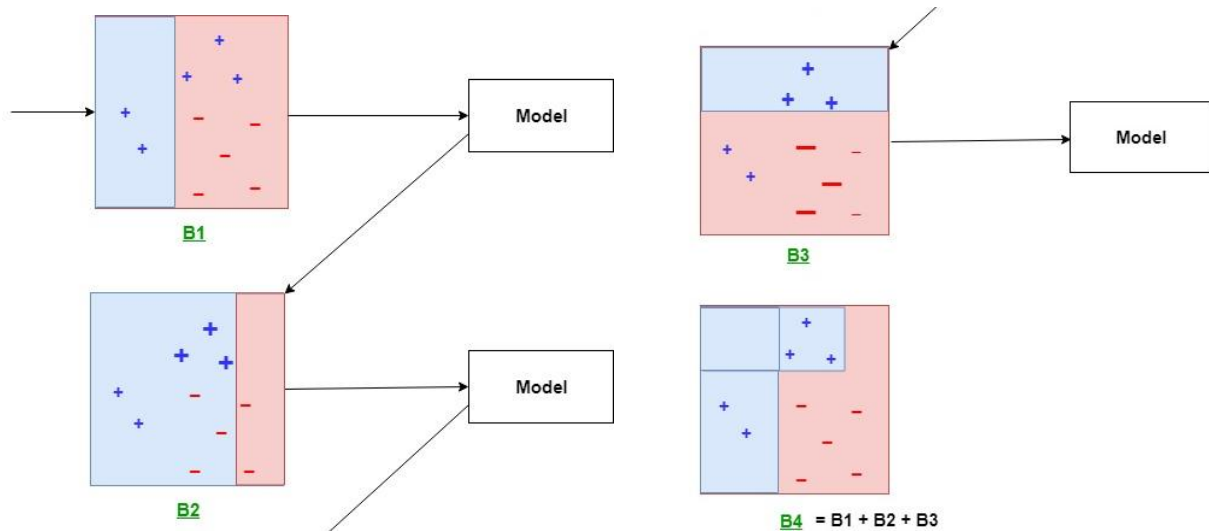
Boosting in Machine Learning - Boosting and AdaBoost

Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers. It is done by building a model by using weak models in series. Firstly, a model is built from the training data. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added.

AdaBoost was the first really successful boosting algorithm developed for the purpose of binary classification. *AdaBoost* is short for *Adaptive Boosting* and is a very popular boosting technique that combines multiple “weak classifiers” into a single “strong classifier”. It was formulated by Yoav Freund and Robert Schapire. They also won the 2003 Gödel Prize for their work.

Algorithm:

1. Initialise the dataset and assign equal weight to each of the data point.
2. Provide this as input to the model and identify the wrongly classified data points.
3. Increase the weight of the wrongly classified data points.
4. if (got required results)
 Goto step 5
 else
 Goto step 2
5. End



Explanation:

The above diagram explains the AdaBoost algorithm in a very simple way. Let's try to understand it in a stepwise process:

- **B1** consists of 10 data points which consist of two types namely plus(+) and minus(-) and 5 of which are plus(+) and the other 5 are minus(-) and each one has been assigned equal weight initially. The first model tries to classify the data points and generates a vertical separator line but it wrongly classifies 3 plus(+) as minus(-).
- **B2** consists of the 10 data points from the previous model in which the 3 wrongly classified plus(+) are weighted more so that the current model tries more to classify these pluses(+) correctly. This model generates a vertical separator line that correctly classifies the previously wrongly classified pluses(+) but in this attempt, it wrongly classifies three minuses(-).
- **B3** consists of the 10 data points from the previous model in which the 3 wrongly classified minus(-) are weighted more so that the current model tries more to classify these minuses(-) correctly. This model generates a horizontal separator line that correctly classifies the previously wrongly classified minuses(-).
- **B4** combines together B1, B2, and B3 in order to build a strong prediction model which is much better than any individual model used.

Making Predictions with AdaBoost

Predictions are made by calculating the weighted average of the weak classifiers.

For a new input instance, each weak learner calculates a predicted value as either +1.0 or -1.0. The predicted values are weighted by each weak learners stage value. The prediction for the ensemble model is taken as the sum of the weighted predictions. If the sum is positive, then the first class is predicted, if negative the second class is predicted.

For example, 5 weak classifiers may predict the values 1.0, 1.0, -1.0, 1.0, -1.0. From a majority vote, it looks like the model will predict a value of 1.0 or the first class. These same 5 weak classifiers may have the stage

values 0.2, 0.5, 0.8, 0.2 and 0.9 respectively. Calculating the weighted sum of these predictions results in an output of -0.8, which would be an ensemble prediction of -1.0 or the second class.

Data Preparation for AdaBoost

This section lists some heuristics for best preparing your data for AdaBoost.

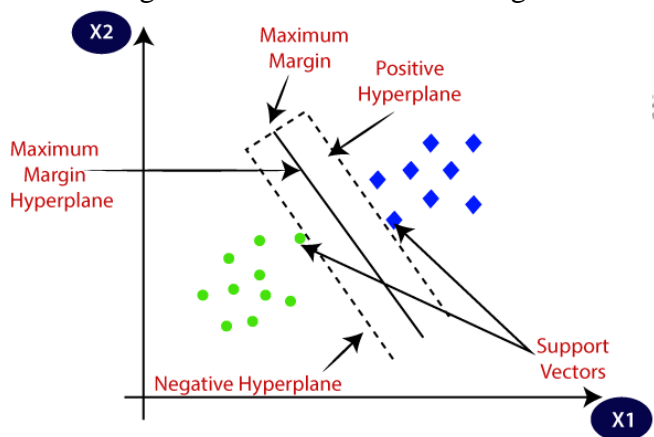
- **Quality Data:** Because the ensemble method continues to attempt to correct misclassifications in the training data, you need to be careful that the training data is of a high-quality.
- **Outliers:** Outliers will force the ensemble down the rabbit hole of working hard to correct for cases that are unrealistic. These could be removed from the training dataset.
- **Noisy Data:** Noisy data, specifically noise in the output variable can be problematic. If possible, attempt to isolate and clean these from your training dataset.

Support Vector Machine Algorithm

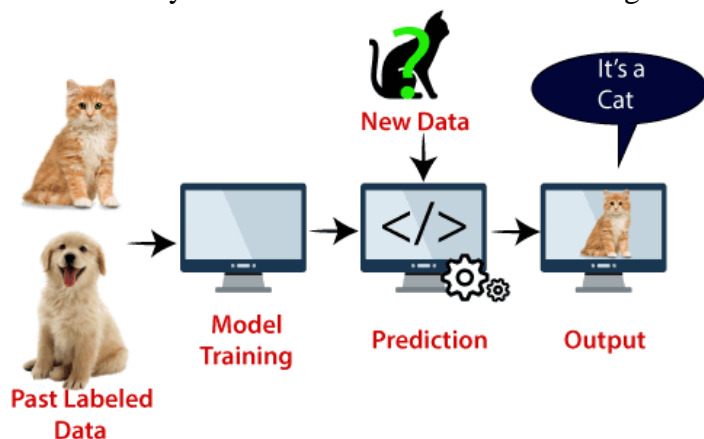
Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



Example: SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:



SVM algorithm can be used for **Face detection, image classification, text categorization**, etc.

Types of SVM

SVM can be of two types:

- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

Hyperplane and Support Vectors in the SVM algorithm:

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

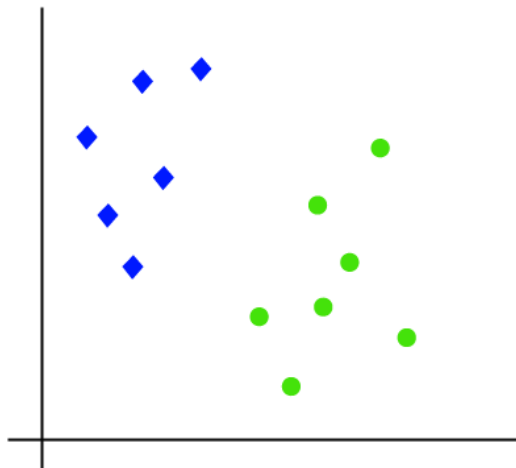
Support Vectors:

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

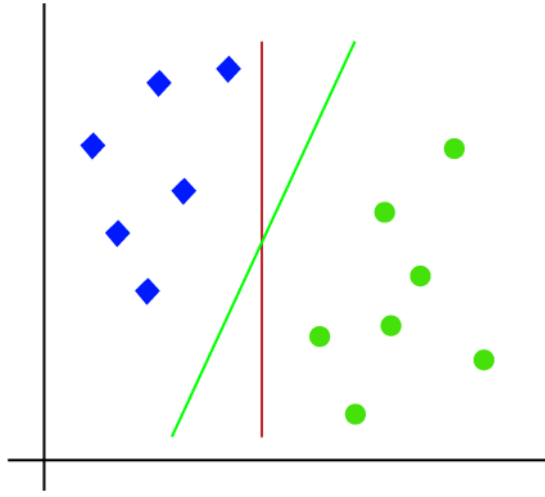
How does SVM works?

Linear SVM:

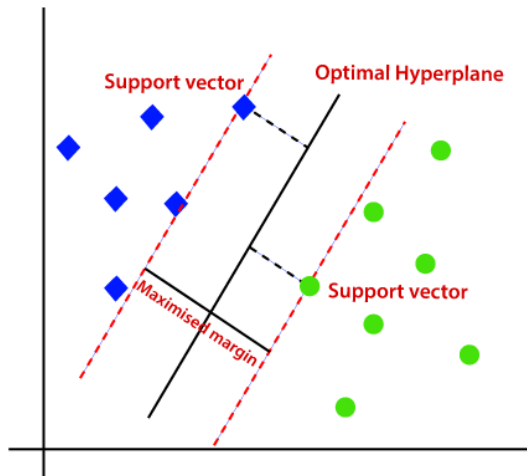
The working of the SVM algorithm can be understood by using an example. Suppose we have a dataset that has two tags (green and blue), and the dataset has two features x_1 and x_2 . We want a classifier that can classify the pair(x_1 , x_2) of coordinates in either green or blue. Consider the below image:



So as it is 2-d space so by just using a straight line, we can easily separate these two classes. But there can be multiple lines that can separate these classes. Consider the below image:

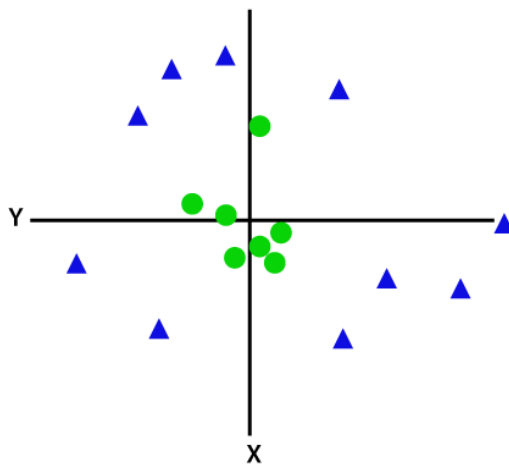


Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a **hyperplane**. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as **margin**. And the goal of SVM is to maximize this margin. The **hyperplane** with maximum margin is called the **optimal hyperplane**.



Non-Linear SVM:

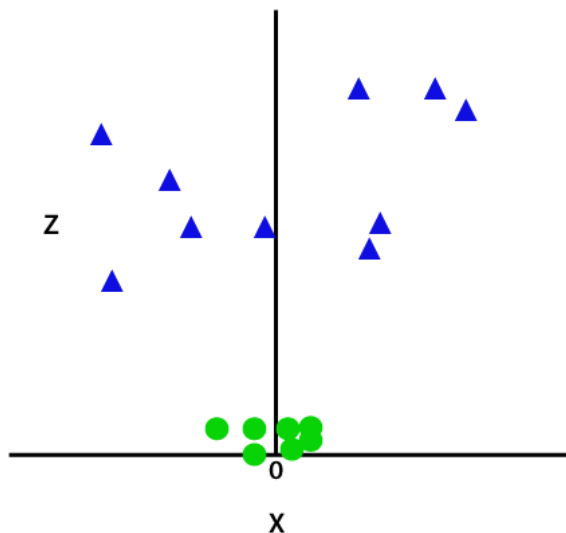
If data is linearly arranged, then we can separate it by using a straight line, but for non-linear data, we cannot draw a single straight line. Consider the below image:



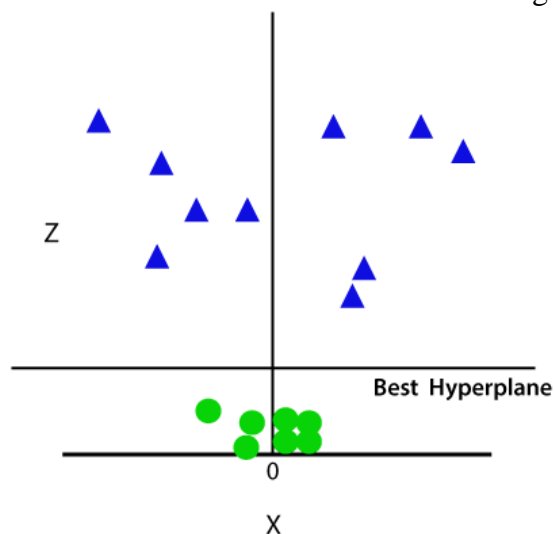
So to separate these data points, we need to add one more dimension. For linear data, we have used two dimensions x and y, so for non-linear data, we will add a third dimension z. It can be calculated as:

$$z=x^2 +y^2$$

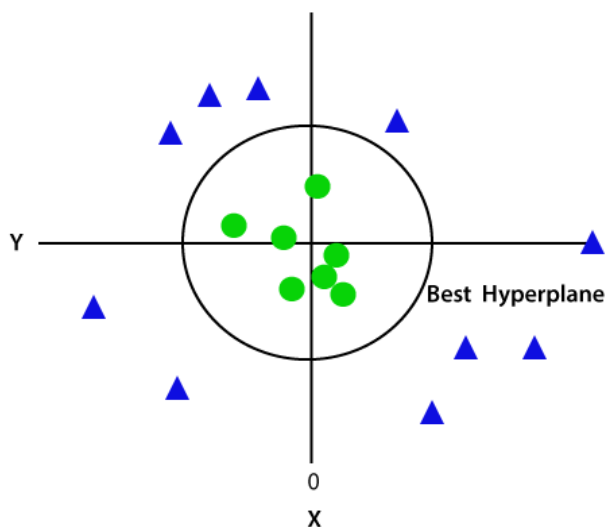
By adding the third dimension, the sample space will become as below image:



So now, SVM will divide the datasets into classes in the following way. Consider the below image:



Since we are in 3-d Space, hence it is looking like a plane parallel to the x-axis. If we convert it in 2d space with $z=1$, then it will become as:



Hence we get a circumference of radius 1 in case of non-linear data.

Python Implementation of Support Vector Machine

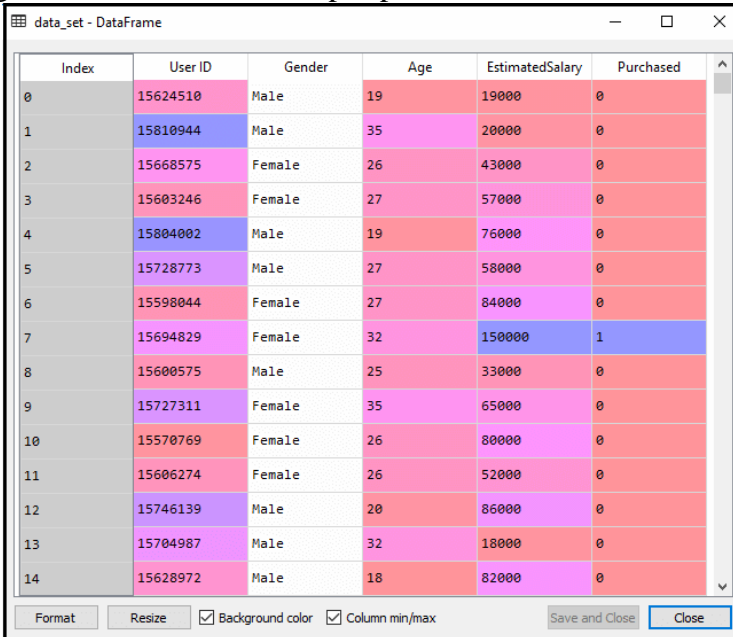
Now we will implement the SVM algorithm using Python. Here we will use the same dataset **user_data**, which we have used in Logistic regression and KNN classification.

Data Pre-processing step

Till the Data pre-processing step, the code will remain the same. Below is the code:

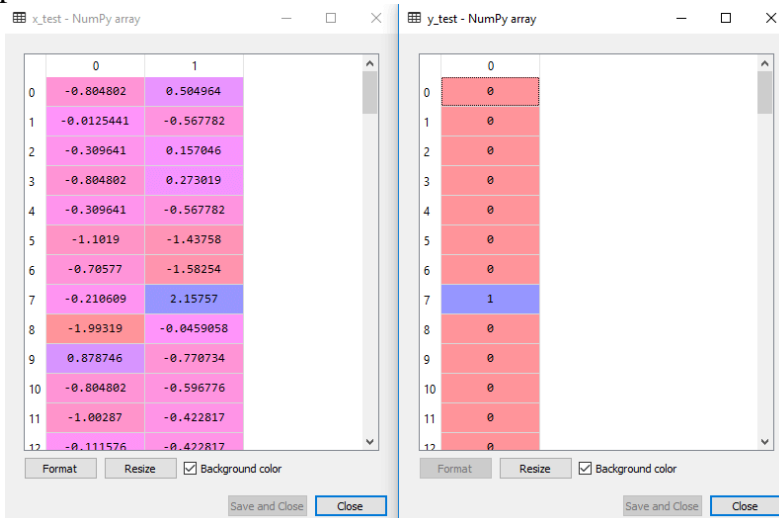
```
#Data Pre-processing Step
# importing libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd
#importing datasets
data_set=pd.read_csv('user_data.csv')
#Extracting Independent and dependent Variable
x= data_set.iloc[:, [2,3]].values
y= data_set.iloc[:, 4].values
# Splitting the dataset into training and test set.
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=0)
#feature Scaling
from sklearn.preprocessing import StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
```

After executing the above code, we will pre-process the data. The code will give the dataset as:



Index	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15660575	Male	25	33000	0
9	15727311	Female	35	65000	0
10	15570769	Female	26	80000	0
11	15606274	Female	26	52000	0
12	15746139	Male	20	86000	0
13	15704987	Male	32	18000	0
14	15628972	Male	18	82000	0

The scaled output for the test set will be:



	0	1
0	-0.804802	0.504954
1	-0.0125441	-0.567782
2	-0.309641	0.157046
3	-0.804802	0.273019
4	-0.309641	-0.567782
5	-1.1019	-1.43758
6	-0.70577	-1.58254
7	-0.210609	2.15757
8	-1.99319	-0.0459058
9	0.878746	-0.770734
10	-0.804802	-0.596776
11	-1.00287	-0.422817
12	-0.111576	-0.422817

	0
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	1
8	0
9	0
10	0
11	0
12	0

Fitting the SVM classifier to the training set:

Now the training set will be fitted to the SVM classifier. To create the SVM classifier, we will import **SVC** class from **Sklearn.svm** library. Below is the code for it:

```
from sklearn.svm import SVC # "Support vector classifier"  
classifier = SVC(kernel='linear', random_state=0)  
classifier.fit(x_train, y_train)
```

In the above code, we have used **kernel='linear'**, as here we are creating SVM for linearly separable data. However, we can change it for non-linear data. And then we fitted the classifier to the training dataset(x_train, y_train)

Output:

Out[8]:

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,  
    decision_function_shape='ovr', degree=3, gamma='auto_deprecated',  
    kernel='linear', max_iter=-1, probability=False, random_state=0,  
    shrinking=True, tol=0.001, verbose=False)
```

The model performance can be altered by changing the value of **C(Regularization factor)**, **gamma**, and **kernel**.

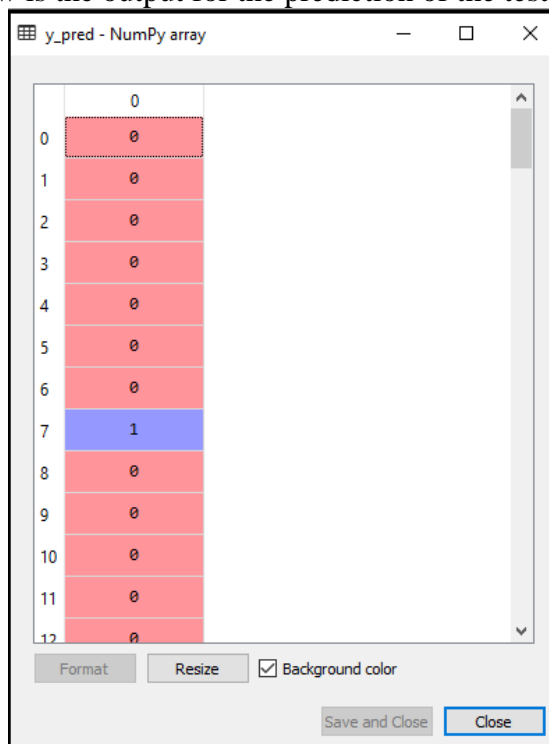
Predicting the test set result:

Now, we will predict the output for test set. For this, we will create a new vector y_pred. Below is the code for it:

```
#Predicting the test set result  
y_pred= classifier.predict(x_test)
```

After getting the y_pred vector, we can compare the result of **y_pred** and **y_test** to check the difference between the actual value and predicted value.

Output: Below is the output for the prediction of the test set:



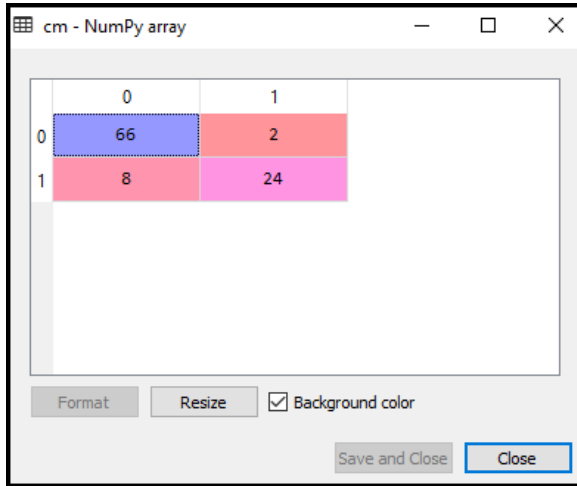
Creating the confusion matrix:

Now we will see the performance of the SVM classifier that how many incorrect predictions are there as

compared to the Logistic regression classifier. To create the confusion matrix, we need to import the **confusion_matrix** function of the sklearn library. After importing the function, we will call it using a new variable **cm**. The function takes two parameters, mainly **y_true**(the actual values) and **y_pred** (the targeted value return by the classifier). Below is the code for it:

1. #Creating the Confusion matrix
2. from sklearn.metrics **import** confusion_matrix
3. cm= confusion_matrix(y_test, y_pred)

Output:



As we can see in the above output image, there are $66+24= 90$ correct predictions and $8+2= 10$ correct predictions. Therefore we can say that our SVM model improved as compared to the Logistic regression model.

Visualizing the training set result:

Now we will visualize the training set result, below is the code for it:

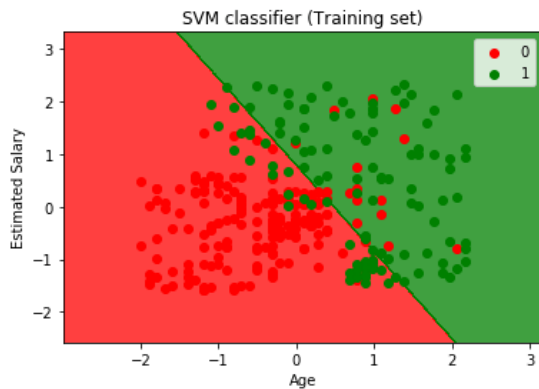
```

from matplotlib.colors import ListedColormap
x_set, y_set = x_train, y_train
x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),
nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))
mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
alpha = 0.75, cmap = ListedColormap(('red', 'green')))
mtp.xlim(x1.min(), x1.max())
mtp.ylim(x2.min(), x2.max())
for i, j in enumerate(nm.unique(y_set)):
    mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
        c = ListedColormap(('red', 'green'))(i), label = j)
mtp.title('SVM classifier (Training set)')
mtp.xlabel('Age')
mtp.ylabel('Estimated Salary')
mtp.legend()
mtp.show()

```

Output:

By executing the above code, we will get the output as:



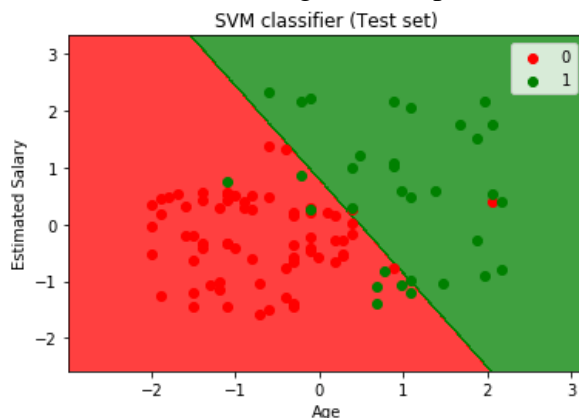
As we can see, the above output is appearing similar to the Logistic regression output. In the output, we got the straight line as hyperplane because we have **used a linear kernel in the classifier**. And we have also discussed above that for the 2d space, the hyperplane in SVM is a straight line.

Visualizing the test set result:

```
#Visulaizing the test set result
from matplotlib.colors import ListedColormap
x_set, y_set = x_test, y_test
x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),
nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))
mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
alpha = 0.75, cmap = ListedColormap(('red','green' )))
mtp.xlim(x1.min(), x1.max())
mtp.ylim(x2.min(), x2.max())
for i, j in enumerate(nm.unique(y_set)):
    mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
c = ListedColormap(('red', 'green'))(i), label = j)
mtp.title('SVM classifier (Test set)')
mtp.xlabel('Age')
mtp.ylabel('Estimated Salary')
mtp.legend()
mtp.show()
```

Output:

By executing the above code, we will get the output as:



As we can see in the above output image, the SVM classifier has divided the users into two regions (Purchased or Not purchased). Users who purchased the SUV are in the red region with the red scatter points. And users who did not purchase the SUV are in the green region with green scatter points. The hyperplane has divided the two classes into Purchased and not purchased variable.

Large Margin Intuition

SVM Decision Boundary

Consider a case where we set constant C to be a very large value, when minimizing the optimization objective, we are going to be highly motivated to choose a value, so that the first term is equal to 0. So what would it take to make this first term equal to 0.

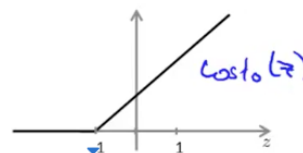
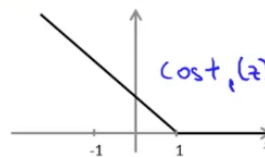
Whenever $y^{(i)} = 1$:

$$\theta^T x^{(i)} \geq 1$$

Whenever $y^{(i)} = 0$:

$$\theta^T x^{(i)} \leq -1$$

$C = \infty$

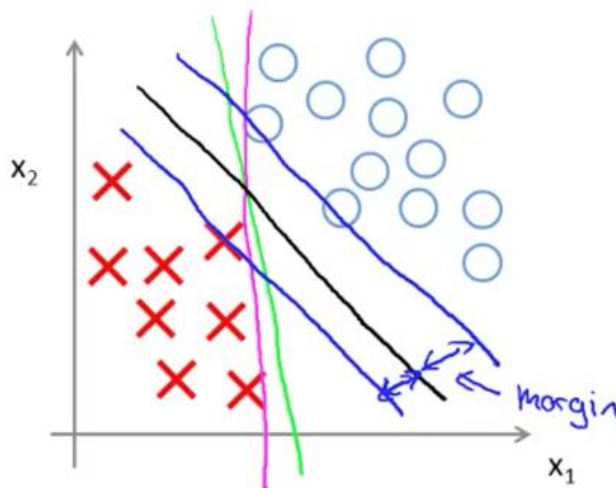


When the first term is equal to 0, we need to minimize (ignored θ_0).

Linear separable case

The obtained decision boundary when minimizing the optimization objective will have the margin as large as possible (hence the name **Large Margin Intuition**).

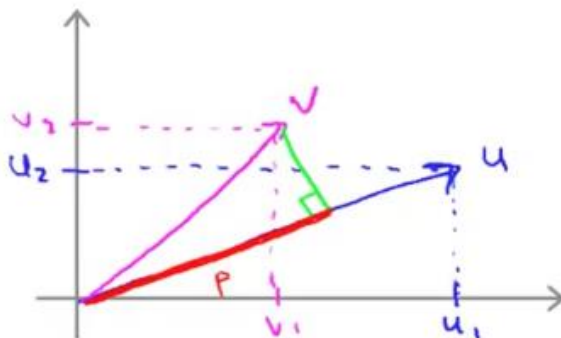
This means SVM will choose the black decision boundary instead of the pink and green one:



Mathematics Behind Large Margin Intuition

Vector Inner Product

p = length of projection of v onto u . p can be positive or negative.



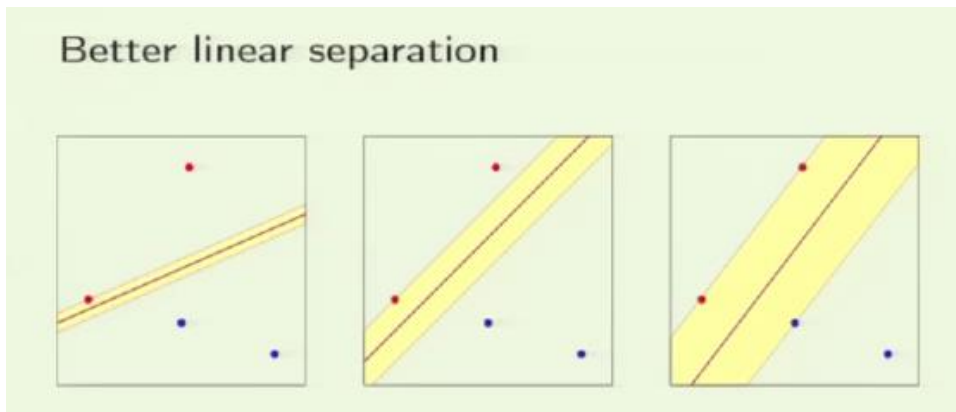
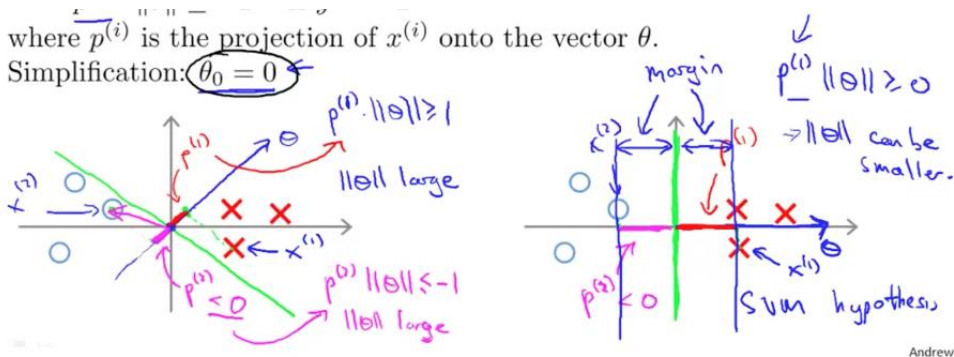
SVM Decision Boundary

We can rewrite the optimization objective of SVM as follow:

where $p^{(i)}$ is the projection of $x^{(i)}$ onto the vector θ .

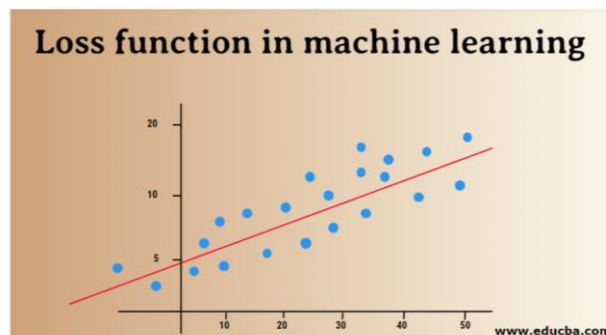
Simplification: $\theta_0 = 0$.

According to the illustration below, with the minimal value of the magnitude of θ , the absolute value of p will large as much as possible (hence the large margin).



In logistic regression, we take the output of the linear function and squash the value within the range of $[0,1]$ using the sigmoid function. If the squashed value is greater than a threshold value(0.5) we assign it a label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify it with another class. Since the threshold values are changed to 1 and -1 in SVM, we obtain this reinforcement range of values $[-1,1]$ which acts as margin.

Loss Function



In Machine learning, the loss function is determined as the difference between the actual output and the predicted output from the model for the single training example while the average of the loss function for all the training examples is termed as the cost function. This computed difference from the loss functions(such

as Regression Loss, Binary Classification, and Multiclass Classification loss function) is termed the error value; this error value is directly proportional to the actual and predicted value.

How does Loss Functions Work?

The word 'Loss' states the penalty for failing to achieve the expected output. If the deviation in the predicted value than the expected value by our model is large, then the loss function gives the higher number as output, and if the deviation is small & much closer to the expected value, it outputs a smaller number.

It is important to note that, amount of deviation doesn't matter; the thing which matters here is whether the value predicted by our model is right or wrong. Loss functions are different based on your problem statement to which machine learning is being applied. The cost function is another term used interchangeably for the loss function, but it holds a slightly different meaning. A loss function is for a single training example, while a cost function is an average loss over the complete train dataset.

Types of Loss Functions in Machine Learning

Below are the different types of the loss function in machine learning which are as follows:

1. Regression loss functions

Linear regression is a fundamental concept of this function. Regression loss functions establish a linear relationship between a dependent variable (Y) and an independent variable (X); hence we try to fit the best line in space on these variables.

$$Y = X_0 + X_1 + X_2 + X_3 + X_4 \dots + X_n$$

- X = Independent variables
- Y = Dependent variable

Mean Squared Error Loss

MSE(L2 error) measures the average squared difference between the actual and predicted values by the model. The output is a single number associated with a set of values. Our aim is to reduce MSE to improve the accuracy of the model.

Consider the linear equation, $y = mx + c$, we can derive MSE as:

$$MSE = \frac{1}{N} \sum_{i=1}^n (y(i) - (mx(i) + b))^2$$

Here, N is the total number of data points, $\frac{1}{N} \sum_{i=1}^n$ is the mean value, and y(i) is the actual value and $mx(i) + b$ its predicted value.

Mean Squared Logarithmic Error Loss (MSLE)

MSLE measures the ratio between actual and predicted value. It introduces an asymmetry in the error curve. MSLE only cares about the percentual difference between actual and predicted values. It can be a good choice as a loss function when we want to predict house sales prices, bakery sales prices, and the data is continuous. Here, the loss can be calculated as the mean of observed data of the squared differences between the log-transformed actual and predicted values, which can be given as:

$$L = \frac{1}{n} \sum_{i=1}^n (\log(y(i)+1) - \log(\hat{y}(i)+1))^2$$

Mean Absolute Error (MAE)

MAE calculates the sum of absolute differences between actual and predicted variables. That means it measures the average magnitude of errors in a set of predicted values. Using the mean square error is easier to solve, but using the absolute error is more robust to outliers. Outliers are those values, which deviate extremely from other observed data points.

MAE can be calculated as:

$$L = \frac{1}{n} \sum_{i=1}^n |y(i) - \hat{y}(i)|$$

2. Binary Classification Loss Functions

These loss functions are made to measure the performances of the classification model. In this, data points are assigned one of the labels, i.e. either 0 or 1. Further, they can be classified as:

Binary Cross-Entropy

It's a default loss function for binary classification problems. Cross-entropy loss calculates the performance of a classification model, which gives an output of a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability value deviate from the actual label.

Hinge loss

Hinge loss can be used as an alternative to cross-entropy, which was initially developed to use with a support vector machine algorithm. Hinge loss works best with the classification problem because target values are in the set of $\{-1,1\}$. It allows to assign more error when there is a difference in sign between actual and predicted values. Hence resulting in better performance than cross-entropy.

Squared Hinge loss

An extension of hinge loss, which simply calculates the square of the hinge loss score. It reduces the error function and makes it numerically easier to work with. It finds the classification boundary that specifies the maximum margin between the data points of various classes. Squared hinge loss fits perfect for YES OR NO kind of decision problems, where probability deviation is not the concern.

3. Multi-class Classification Loss Functions

Multi-class classification is the predictive models in which the data points are assigned to more than two classes. Each class is assigned a unique value from 0 to (Number_of_classes - 1). It is highly recommended for image or text classification problems, where a single paper can have multiple topics.

Multi-class Cross-Entropy

In this case, the target values are in the set of 0 to n i.e $\{0,1,2,3\dots n\}$. It calculates a score that takes an average difference between actual and predicted probability values, and the score is minimized to reach the best possible accuracy. Multi-class cross-entropy is the default loss function for text classification problems.

Sparse Multi-class Cross-Entropy

One hot encoding process makes multi-class cross-entropy difficult to handle a large number of data points. Sparse cross-entropy solves this problem by performing the calculation of error without using one-hot encoding.

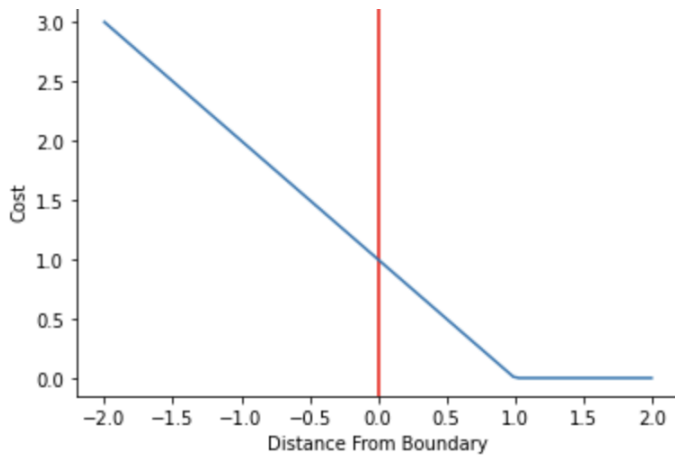
Kullback Leibler Divergence Loss

KL divergence loss calculates the divergence between probability distribution and baseline distribution and finds out how much information is lost in terms of bits. The output is a non-negative value that specifies how close two probability distributions are. To describe KL divergence in terms of probabilistic view, the likelihood ratio is used.

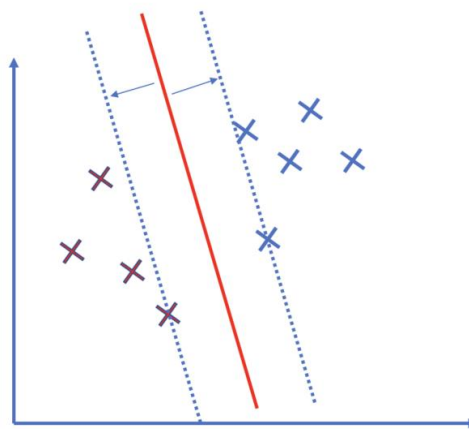
Hinge Loss

The hinge loss is a specific type of cost function that incorporates a margin or distance from the classification boundary into the cost calculation. Even if new observations are classified correctly, they can incur a penalty if the margin from the decision boundary is not large enough. The hinge loss increases linearly.

The hinge loss is mostly associated with soft-margin support vector machines.



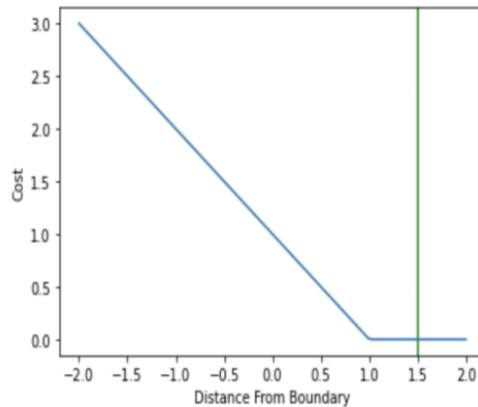
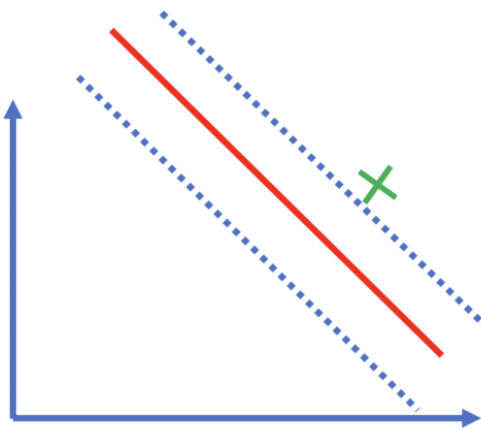
If you are familiar with the construction of hyperplanes and their margins in support vector machines, you probably know that margins are often defined as having a distance equal to 1 from the data-separating-hyperplane. Otherwise, check out my [post on support vector machines](#) (link opens in new tab), where I explain the details of maximum margins classifiers. We want data points to not only fall on the correct side of the hyperplane but also to be located beyond the margin.



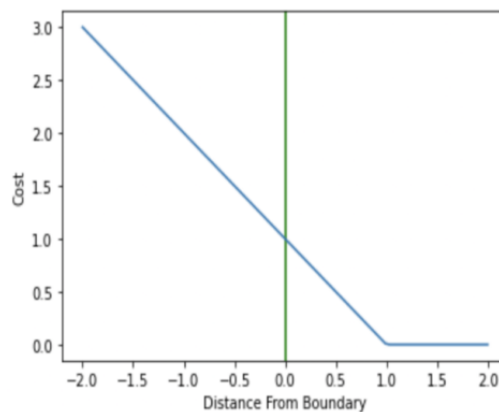
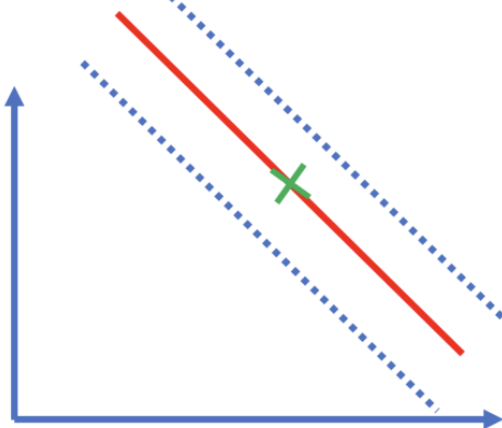
Support vector machines address a classification problem where observations either have an outcome of +1 or -1. The support vector machine produces a real-valued output that is negative or positive depending on which side of the decision boundary it falls. Only if an observation is classified correctly and the distance from the plane is larger than the margin will it incur no penalty. The distance from the hyperplane can be regarded as a measure of confidence. The further an observation lies from the plane, the more confident it is in the classification.

For example, if an observation was associated with an actual outcome of +1, and the SVM produced an output of 1.5, the loss would equal 0.

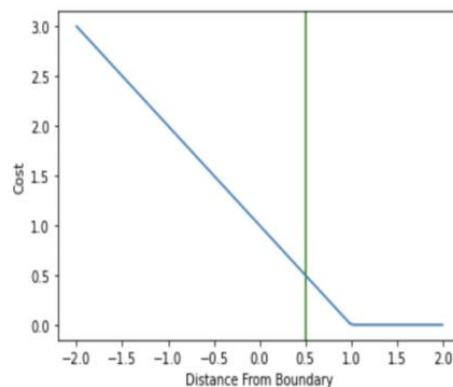
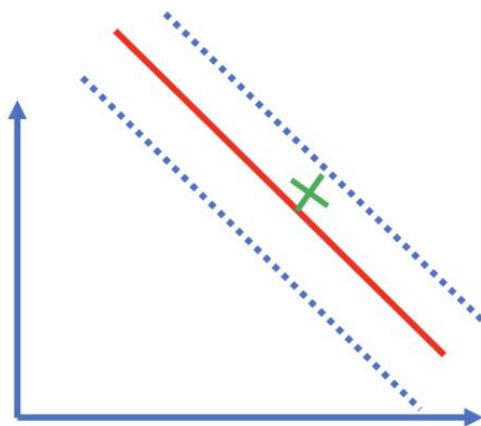
Contrary to methods like linear regression, where we try to find a line that minimizes the distance from the data points, an SVM tries to **maximize** the distance. If you are interested, check out my post on [constructing regression lines](#). Comparing the two approaches nicely illustrates the difference between the nature of regression and classification problems.



An observation that is located directly on the boundary would incur a loss of 1 regardless of whether the real outcome was +1 or -1.



Observations that fall on the correct side of the decision boundary (hyperplane) but are within the margin incur a cost between 0 and 1.



All observations that end up on the wrong side of the hyperplane will incur a loss that is greater than 1 and increases linearly. If the actual outcome was 1 and the classifier predicted 0.5, the corresponding loss would be 0.5 even though the classification is correct.

Now that we have a strong intuitive understanding of the hinge loss, understanding the math will be a breeze.

Hinge Loss Formula

The loss is defined according to the following formula, where t is the actual outcome (either 1 or -1), and y is the output of the classifier.

$$l(y) = \max(0, 1 - t \cdot y)$$

Let's plug in the values from our last example. The outcome was 1, and the prediction was 0.5.

$$l(y) = \max(0, 1 - 1 \cdot 0.5) = 0.5$$

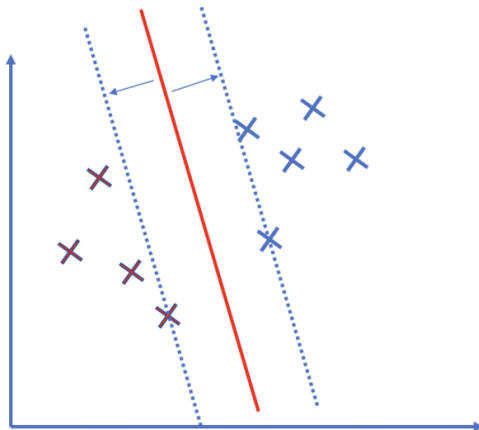
If, on the other hand, the outcome was -1, the loss would be higher since we've misclassified our example.

$$l(y) = \max(0, 1 - (-1) \cdot 0.5) = 1.5$$

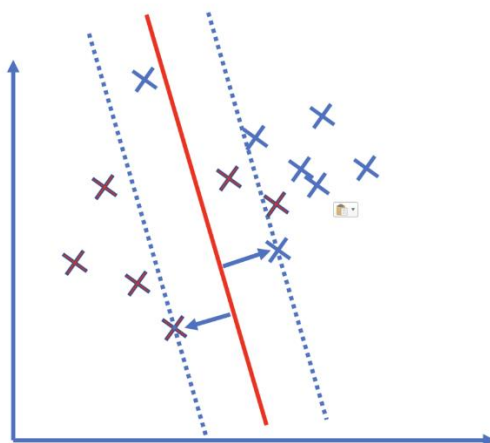
Instead of using a labelling convention of -1, and 1 we could also use 0 and 1 and use the formula for cross-entropy to set one of the terms equal to zero. But the math checks out more beautifully in the former case. With the hinge loss defined, we are now in a position to understand the loss function for the support vector machine. But before we do this, we'll briefly discuss why and when we actually need a cost function.

Hard Margin vs Soft Margin Support Vector Machine

In a hard margin SVM, we want to linearly separate the data without misclassification. This implies that the data actually has to be linearly separable.



In this case, the blue and red data points are linearly separable, allowing for a hard margin classifier. If the data is not linearly separable, hard margin classification is not applicable. Even though support vector machines are linear classifiers, they are still able to separate data points that are not linearly separable by applying the kernel trick.



The blue and the red data points are not linearly separable. Furthermore, if the margin of the SVM is very small, the model is more likely to overfit. In these cases, we can choose to cut the model some slack by allowing for misclassifications. We call this a soft margin support vector machine. But if the model produces too many misclassifications, its utility declines. Therefore, we need to penalize the misclassified samples by introducing a cost function. In summary, the soft margin support vector machine requires a cost function while the hard margin SVM does not.

SVM Cost

In the [post on support vectors](#), we've established that the optimization objective of the support vector classifier is to minimize the term w , which is a vector orthogonal to the data-separating hyperplane onto which we project our data points.

$$\min_w \frac{1}{2} \sum_{i=1}^n w_i^2 \quad \text{minimize } \frac{1}{2} \sum_{i=1}^n w_i^2$$

This minimization problem represents the primal form of the hard margin SVM, which doesn't account for classification errors.

For the soft-margin SVM, we combine the minimization objective with a loss function such as the hinge loss.

$$\min_w \frac{1}{2} \sum_{i=1}^n w_i^2 + \sum_{j=1}^m \max(0, 1 - t_j \cdot y_j)$$

The first term sums over the number of features (n), while the second term sums over the number of samples in the data (m).

The t variable is the output produced by the model as a product of the weight parameter w and the data input x.

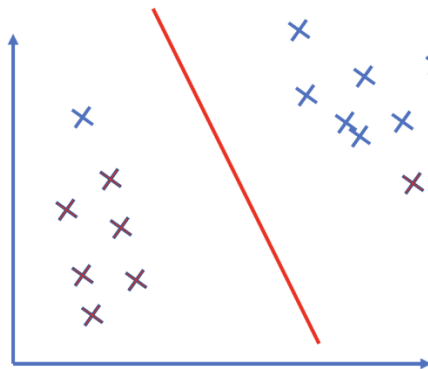
$$t_i = w^T x_j$$

To understand how the model generates this output, refer to the [post on support vectors](#) (link opens in new tab).

The loss term has a regularizing effect on the model. But how can we control the regularization? That is how can we control how aggressively the model should try to avoid misclassifications. To manually control the number of misclassifications during training, we introduce an additional parameter, C, which we multiply with the loss term.

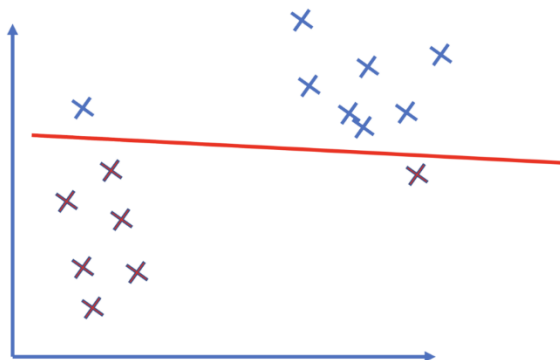
$$\min_w \frac{1}{2} \sum_{i=1}^n w_i^2 + C \sum_{j=1}^m \max(0, 1 - t_j \cdot y_j)$$

The smaller C is, the stronger the regularization. Accordingly, the model will attempt to maximize the margin and be more tolerant towards misclassifications.



Cost function with a small regularization parameter C

If we set C to a large number, then the SVM will pursue outliers more aggressively, which potentially comes at the cost of a smaller margin and may lead to overfitting on the training data. The classifier might be less robust on unseen data.



Cost function with a large regularization parameter C leading to less regularization.

SVM Kernels

Kernel Function is a method used to take data as input and transform it into the required form of processing data. “Kernel” is used due to a set of mathematical functions used in Support Vector Machine providing the window to manipulate the data. So, Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Basically, It returns the inner product between two points in a standard feature dimension.

Standard Kernel Function Equation :

$$K(\bar{x}) = 1, \text{ if } \|\bar{x}\| \leq 1$$
$$K(\bar{x}) = 0, \text{ Otherwise}$$

Major Kernel Functions :-

For Implementing Kernel Functions, first of all, we have to install the “scikit-learn” library using the command prompt terminal:

```
pip install scikit-learn
```

Gaussian Kernel: It is used to perform transformation when there is no prior knowledge about data.

$$K(x, y) = e^{-\left(\frac{\|x-y\|^2}{2\sigma^2}\right)}$$

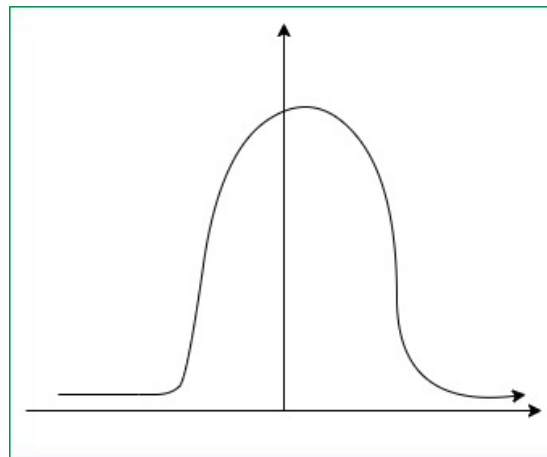
- **Gaussian Kernel Radial Basis Function (RBF):** Same as above kernel function, adding radial basis method to improve the transformation.

$$K(x, y) = e^{-\left(\gamma\|x - y\|^2\right)}$$

$$K(x, x1) + K(x, x2) \text{ (Simplified - Formula)}$$

$$K(x, x1) + K(x, x2) > 0 \text{ (Green)}$$

$$K(x, x1) + K(x, x2) = 0 \text{ (Red)}$$



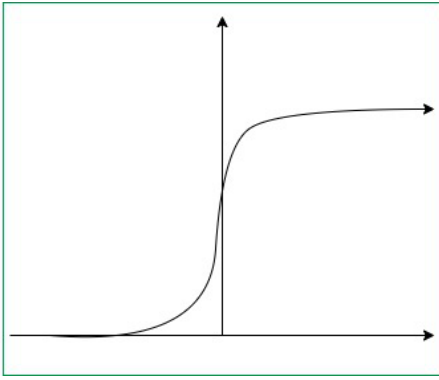
Gaussian Kernel Graph

Code:

```
from sklearn.svm import SVC
classifier = SVC(kernel='rbf', random_state = 0)
# training set in x, y axis
classifier.fit(x_train, y_train)
```

Sigmoid Kernel: this function is equivalent to a two-layer, perceptron model of the neural network, which is used as an activation function for artificial neurons.

$$K(x, y) = \tanh(\gamma \cdot x^T y + r)$$



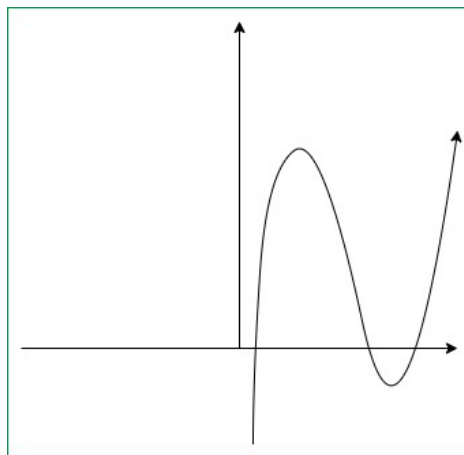
Sigmoid Kernel Graph

Code:

```
from sklearn.svm import SVC
classifier = SVC(kernel='sigmoid')
classifier.fit(x_train, y_train) # training set in x, y axis
```

Polynomial Kernel: It represents the similarity of vectors in the training set of data in a feature space over polynomials of the original variables used in the kernel.

$$K(x, y) = \tanh(\gamma \cdot x^T y + r)^d, \gamma > 0$$



Polynomial Kernel Graph

Code:

```
from sklearn.svm import SVC
classifier = SVC(kernel='poly', degree = 4)
classifier.fit(x_train, y_train) # training set in x, y axis
```


Unit IV – Parametric Machine Learning.

Logistic Regression: Classification and representation

Introduction to Logistic Regression:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts $P(Y=1)$ as a function of X . It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1.**

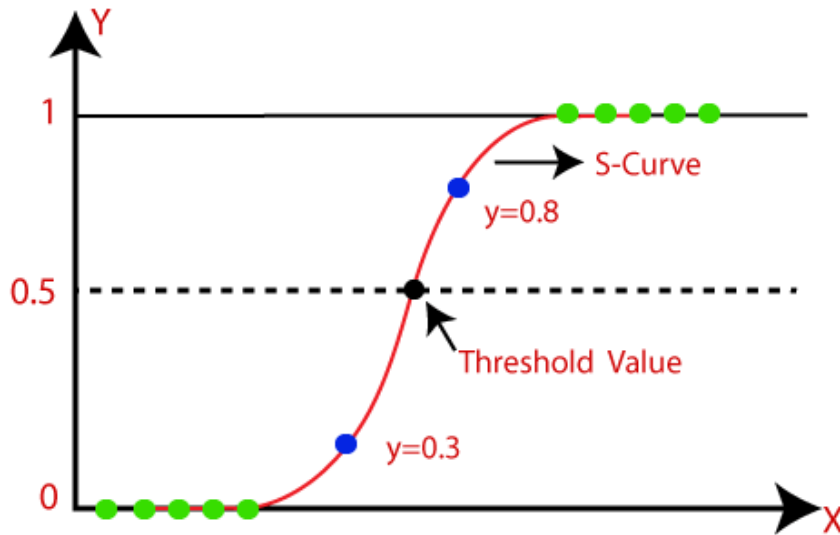
Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems.**

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Function (Sigmoid Function):

1. The sigmoid function is a mathematical function used to map the predicted values to probabilities.
2. It maps any real value into another value within a range of 0 and 1.
3. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the sigmoid function or the logistic function.
4. In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

1. The dependent variable must be categorical in nature.
2. The independent variable should not have multi-collinearity.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

1. We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

2. In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}; 0 \text{ for } y=0, \text{ and infinity for } y=1$$

3. But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$\log \left[\frac{y}{1-y} \right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

- **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
- **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Steps in Logistic Regression:

To implement the Logistic Regression using Python, we will use the same steps as we have done in previous topics of Regression. Below are the steps:

1. Data Pre-processing step
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

1. Data Pre-processing step: In this step, we will pre-process/prepare the data so that we can use it in our code efficiently. It will be the same as we have done in Data pre-processing topic. The code for this is given below:

```
#Data Pre-processing Step
# importing libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd

#importing datasets
data_set= pd.read_csv('user_data.csv')
#Extracting Independent and dependent Variable
x= data_set.iloc[:, [2,3]].values
y= data_set.iloc[:, 4].values
# Splitting the dataset into training and test set.
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.25, random_state=0)
#feature Scaling
from sklearn.preprocessing import StandardScaler
st_x= StandardScaler()
x_train= st_x.fit_transform(x_train)
x_test= st_x.transform(x_test)
```

2. Fitting Logistic Regression to the Training set:

We have well prepared our dataset, and now we will train the dataset using the training set. For providing training or fitting the model to the training set, we will import the LogisticRegression class of the sklearn library.

After importing the class, we will create a classifier object and use it to fit the model to the logistic regression. Below is the code for it:

```
#Fitting Logistic Regression to the training set
from sklearn.linear_model import LogisticRegression
classifier= LogisticRegression(random_state=0)
classifier.fit(x_train, y_train)
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, l1_ratio=None, max_iter=100, multi_class='warn', n_jobs=None, penalt
y='l2', random_state=0, solver='warn', tol=0.0001, verbose=0, warm_start=False)
```

3. Predicting the Test Result

Our model is well trained on the training set, so we will now predict the result by using test set data. Below is the code for it:

```
# Predicting the test set result
y_pred= classifier.predict(x_test)
```

4. Test Accuracy of the result:

Now we will create the confusion matrix here to check the accuracy of the classification. To create it, we need to import the confusion_matrix function of the sklearn library. After importing the function, we will call it using a new variable cm. The function takes two parameters, mainly y_true(the actual values) and y_pred (the targeted value return by the classifier). Below is the code for it:

```
#Creating the Confusion matrix
from sklearn.metrics import confusion_matrix
cm= confusion_matrix()
```

We can find the accuracy of the predicted result by interpreting the confusion matrix. By above output, we can interpret that $65+24= 89$ (Correct Output) and $8+3= 11$ (Incorrect Output).

5. Visualizing the training set result:

Finally, we will visualize the training set result. To visualize the result, we will use ListedColormap class of matplotlib library. Below is the code for it:

```
#Visualizing the training set result
```

```
from matplotlib.colors import ListedColormap
x_set, y_set = x_train, y_train
x1, x2 = nm.meshgrid(nm.arange(start = x_set[:, 0].min() - 1, stop = x_set[:, 0].max() + 1, step =0.01),
nm.arange(start = x_set[:, 1].min() - 1, stop = x_set[:, 1].max() + 1, step = 0.01))
mtp.contourf(x1, x2, classifier.predict(nm.array([x1.ravel(), x2.ravel()]).T).reshape(x1.shape),
alpha = 0.75, cmap = ListedColormap(('purple','green' )))
```



```

mtp.xlim(x1.min(), x1.max())
mtp.ylim(x2.min(), x2.max())
for i, j in enumerate(nm.unique(y_set)):
    mtp.scatter(x_set[y_set == j, 0], x_set[y_set == j, 1],
               c = ListedColormap(('purple', 'green'))(i), label = j)
mtp.title('Logistic Regression (Training set)')
mtp.xlabel('Age')
mtp.ylabel('Estimated Salary')
mtp.legend()
mtp.show()

```

In the above code, we have imported the ListedColormap class of Matplotlib library to create the colormap for visualizing the result. We have created two new variables x_set and y_set to replace x_train and y_train. After that, we have used the nm.meshgrid command to create a rectangular grid, which has a range of -1 (minimum) to 1 (maximum). The pixel points we have taken are of 0.01 resolution. To create a filled contour, we have used mtp.contourf command, it will create regions of provided colors (purple and green). In this function, we have passed the classifier.predict to show the predicted data points predicted by the classifier.

Output: By executing the above code, we will get the below output:



The graph can be explained in the below points:

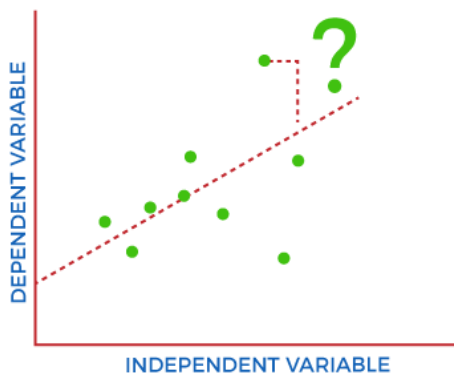
- In the above graph, we can see that there are some **Green points** within the green region and **Purple points** within the purple region.
- All these data points are the observation points from the training set, which shows the result for purchased variables.
- This graph is made by using two independent variables i.e., **Age on the x-axis** and **Estimated salary on the y-axis**.
- The **purple point observations** are for which purchased (dependent variable) is probably 0, i.e., users who did not purchase the SUV car.
- The **green point observations** are for which purchased (dependent variable) is probably 1 means user who purchased the SUV car.
- We can also estimate from the graph that the users who are younger with low salary, did not purchase the car, whereas older users with high estimated salary purchased the car.

- But there are some purple points in the green region (Buying the car) and some green points in the purple region (Not buying the car). So we can say that younger users with a high estimated salary purchased the car, whereas an older user with a low estimated salary did not purchase the car.

Cost Function in Machine Learning

A Machine Learning model should have a very high level of accuracy in order to perform well with real-world applications. But how to calculate the accuracy of the model, i.e., how good or poor our model will perform in the real world? In such a case, the Cost function comes into existence. It is an important machine learning parameter to correctly estimate the model.

COST FUNCTION IN MACHINE LEARNING



Cost function also plays a crucial role in understanding that how well your model estimates the relationship between the input and output parameters.

In this topic, we will explain the cost function in Machine Learning, Gradient descent, and types of cost functions.

What is Cost Function?

A cost function is an important parameter that determines how well a machine learning model performs for a given dataset. It calculates the difference between the expected value and predicted value and represents it as a single real number.

In machine learning, once we train our model, then we want to see how well our model is performing. Although there are various accuracy functions that tell you how your model is performing, but will not give insights to improve them. So, we need a function that can find when the model is most accurate by finding the spot between the undertrained and overtrained model.

In simple, "Cost function is a measure of how wrong the model is in estimating the relationship between X(input) and Y(output) Parameter." A cost function is sometimes also referred to as Loss function, and it can be estimated by iteratively running the model to compare estimated predictions against the known values of Y.

The main aim of each ML model is to determine parameters or weights that can minimize the cost function.

Types of Cost Function

Cost functions can be of various types depending on the problem. However, mainly it is of three types, which are as follows:

1. Regression Cost Function
2. Binary Classification cost Functions
3. Multi-class Classification Cost Function.

1. Regression Cost Function

Regression models are used to make a prediction for the continuous variables such as the price of houses, weather prediction, loan predictions, etc. When a cost function is used with Regression, it is known as the "Regression Cost Function." In this, the cost function is calculated as the error based on the distance, such as:

$$\text{Error} = \text{Actual Output} - \text{Predicted output}$$

There are three commonly used Regression cost functions, which are as follows:

a. Means Error

In this type of cost function, the error is calculated for each training data, and then the mean of all error values is taken.

It is one of the simplest ways possible.

The errors that occurred from the training data can be either negative or positive. While finding mean, they can cancel out each other and result in the zero-mean error for the model, so it is not recommended cost function for a model.

However, it provides a base for other cost functions of regression models.

b. Mean Squared Error (MSE)

Mean Square error is one of the most commonly used Cost function methods. It improves the drawbacks of the Mean error cost function, as it calculates the square of the difference between the actual value and predicted value. Because of the square of the difference, it avoids any possibility of negative error.

The formula for calculating MSE is given below:

$$\text{MAE} = \frac{\sum_{i=0}^n |y - y'|}{n}$$

Mean squared error is also known as L2 Loss.

In MSE, each error is squared, and it helps in reducing a small deviation in prediction as compared to MAE. But if the dataset has outliers that generate more prediction errors, then squaring of this error will further increase the error multiple times. Hence, we can say MSE is less robust to outliers.

c. Mean Absolute Error (MAE)

Mean Absolute error also overcome the issue of the Mean error cost function by taking the absolute difference between the actual value and predicted value.

The formula for calculating Mean Absolute Error is given below:

$$\text{MAE} = \frac{\sum_{i=0}^n |y - y'|}{n}$$

This means the Absolute error cost function is also known as L1 Loss. It is not affected by noise or outliers, hence giving better results if the dataset has noise or outlier.

2. Binary Classification Cost Functions

Classification models are used to make predictions of categorical variables, such as predictions for 0 or 1, Cat or dog, etc. The cost function used in the classification problem is known as the Classification cost function. However, the classification cost function is different from the Regression cost function.

One of the commonly used loss functions for classification is cross-entropy loss.

The binary Cost function is a special case of Categorical cross-entropy, where there is only one output class. For example, classification between red and blue.

To better understand it, let's suppose there is only a single output variable Y

$$\text{Cross-entropy}(D) = -y * \log(p) \text{ when } y = 1$$

$$\text{Cross-entropy}(D) = -(1-y) * \log(1-p) \text{ when } y = 0$$

The error in binary classification is calculated as the mean of cross-entropy for all N training data. Which means:

$$\text{Binary Cross-Entropy} = (\text{Sum of Cross-Entropy for N data})/N$$

3. Multi-class Classification Cost Function

A multi-class classification cost function is used in the classification problems for which instances are allocated to one of more than two classes. Here also, similar to binary class classification cost function, cross-entropy or categorical cross-entropy is commonly used cost function.

It is designed in a way that it can be used with multi-class classification with the target values ranging from 0 to 1, 3, ...,n classes.

In a multi-class classification problem, cross-entropy will generate a score that summarizes the mean difference between actual and anticipated probability distribution.

For a perfect cross-entropy, the value should be zero when the score is minimized.

Gradient Descent in Machine Learning

Gradient Descent is known as one of the most commonly used optimization algorithms to train machine learning models by means of minimizing errors between actual and expected results. Further, gradient descent is also used to train Neural Networks.

In mathematical terminology, Optimization algorithm refers to the task of minimizing/maximizing an objective function $f(x)$ parameterized by x . Similarly, in machine learning, optimization is the task of minimizing the cost function parameterized by the model's parameters. The main objective of gradient descent is to minimize the convex function using iteration of parameter updates. Once these machine learning models are optimized, these models can be used as powerful tools for Artificial Intelligence and various computer science applications.

In this tutorial on Gradient Descent in Machine Learning, we will learn in detail about gradient descent, the role of cost functions specifically as a barometer within Machine Learning, types of gradient descents, learning rates, etc.

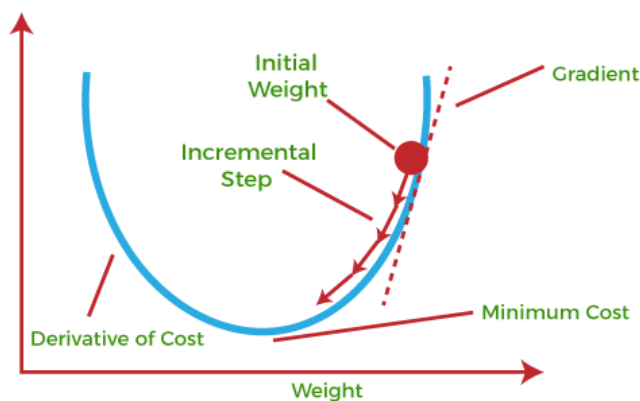
What is Gradient Descent or Steepest Descent?

Gradient descent was initially discovered by "Augustin-Louis Cauchy" in mid of 18th century. Gradient Descent is defined as one of the most commonly used iterative optimization algorithms of machine learning to train the machine learning and deep learning models. It helps in finding the local minimum of a function.

The best way to define the local minimum or local maximum of a function using gradient descent is as follows:

If we move towards a negative gradient or away from the gradient of the function at the current point, it will give the local minimum of that function.

Whenever we move towards a positive gradient or towards the gradient of the function at the current point, we will get the local maximum of that function.



This entire procedure is known as Gradient Ascent, which is also known as steepest descent. The main objective of using a gradient descent algorithm is to minimize the cost function using iteration. To achieve this goal, it performs two steps iteratively:

Calculates the first-order derivative of the function to compute the gradient or slope of that function.

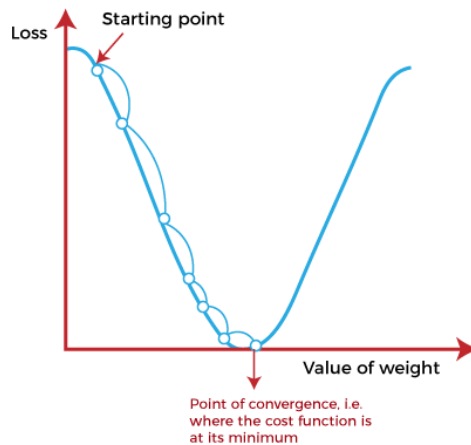
Move away from the direction of the gradient, which means slope increased from the current point by alpha times, where Alpha is defined as Learning Rate. It is a tuning parameter in the optimization process which helps to decide the length of the steps.

How does Gradient Descent work?

Before starting the working principle of gradient descent, we should know some basic concepts to find out the slope of a line from linear regression. The equation for simple linear regression is given as:

$$\text{Equation : } Y=mX+c$$

Where 'm' represents the slope of the line, and 'c' represents the intercepts on the y-axis.



The starting point (shown in above fig.) is used to evaluate the performance as it is considered just as an arbitrary point. At this starting point, we will derive the first derivative or slope and then use a tangent line to calculate the steepness of this slope. Further, this slope will inform the updates to the parameters (weights and bias).

The slope becomes steeper at the starting point or arbitrary point, but whenever new parameters are generated, then steepness gradually reduces, and at the lowest point, it approaches the lowest point, which is called a **point of convergence**.

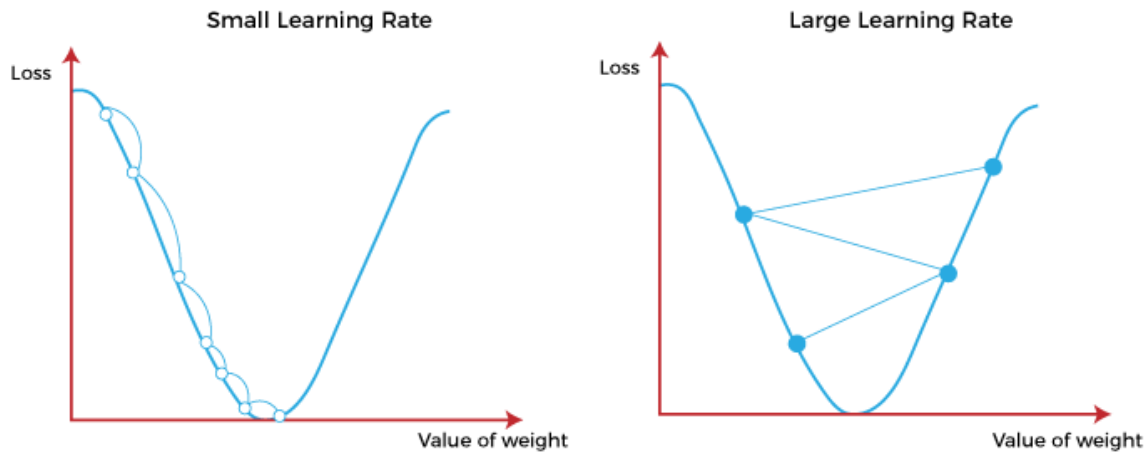
The main objective of gradient descent is to minimize the cost function or the error between expected and actual. To minimize the cost function, two data points are required:

1. Direction & Learning Rate

These two factors are used to determine the partial derivative calculation of future iteration and allow it to the point of convergence or local minimum or global minimum. Let's discuss learning rate factors in brief;

Learning Rate:

It is defined as the step size taken to reach the minimum or lowest point. This is typically a small value that is evaluated and updated based on the behavior of the cost function. If the learning rate is high, it results in larger steps but also leads to risks of overshooting the minimum. At the same time, a low learning rate shows the small step sizes, which compromises overall efficiency but gives the advantage of more precision.



Types of Gradient Descent

Based on the error in various training models, the Gradient Descent learning algorithm can be divided into **Batch gradient descent**, **stochastic gradient descent**, and **mini-batch gradient descent**. Let's understand these different types of gradient descent:

1. Batch Gradient Descent:

Batch gradient descent (BGD) is used to find the error for each point in the training set and update the model after evaluating all training examples. This procedure is known as the training epoch. In simple words, it is a greedy approach where we have to sum over all examples for each update.

Advantages of Batch gradient descent:

- It produces less noise in comparison to other gradient descent.
- It produces stable gradient descent convergence.
- It is Computationally efficient as all resources are used for all training samples.

2. Stochastic gradient descent:

Stochastic gradient descent (SGD) is a type of gradient descent that runs one training example per iteration. Or in other words, it processes a training epoch for each example within a dataset and updates each training example's parameters one at a time. As it requires only one training example at a time, hence it is easier to store in allocated memory. However, it shows some computational efficiency losses in comparison to batch gradient systems as it shows frequent updates that require more detail and speed. Further, due to frequent updates, it is also treated as a noisy gradient. However, sometimes it can be helpful in finding the global minimum and also escaping the local minimum.

Advantages of Stochastic gradient descent:

In Stochastic gradient descent (SGD), learning happens on every example, and it consists of a few advantages over other gradient descent.

- It is easier to allocate in desired memory.
- It is relatively fast to compute than batch gradient descent.
- It is more efficient for large datasets.

3. MiniBatch Gradient Descent:

Mini Batch gradient descent is the combination of both batch gradient descent and stochastic gradient descent. It divides the training datasets into small batch sizes then performs the updates on those batches separately. Splitting training datasets into smaller batches make a balance to maintain the computational efficiency of batch gradient descent and speed of stochastic gradient descent. Hence, we can achieve a special type of gradient descent with higher computational efficiency and less noisy gradient descent.

Advantages of Mini Batch gradient descent:

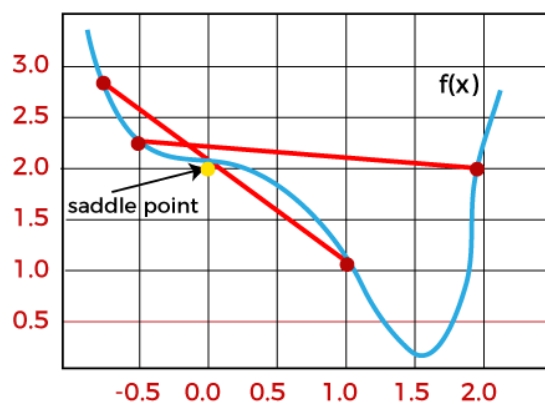
- It is easier to fit in allocated memory.
- It is computationally efficient.
- It produces stable gradient descent convergence.

Challenges with the Gradient Descent

Although we know Gradient Descent is one of the most popular methods for optimization problems, it still also has some challenges. There are a few challenges as follows:

1. Local Minima and Saddle Point:

For convex problems, gradient descent can find the global minimum easily, while for non-convex problems, it is sometimes difficult to find the global minimum, where the machine learning models achieve the best results.



Whenever the slope of the cost function is at zero or just close to zero, this model stops learning further. Apart from the global minimum, there occur some scenarios that can show this slop, which is saddle point and local minimum. Local minima generate the shape similar to the global minimum, where the slope of the cost function increases on both sides of the current points.

In contrast, with saddle points, the negative gradient only occurs on one side of the point, which reaches a local maximum on one side and a local minimum on the other side. The name of a saddle point is taken by that of a horse's saddle.

The name of local minima is because the value of the loss function is minimum at that point in a local region. In contrast, the name of the global minima is given so because the value of the loss function is minimum there, globally across the entire domain the loss function.

2. Vanishing and Exploding Gradient

In a deep neural network, if the model is trained with gradient descent and backpropagation, there can occur two more issues other than local minima and saddle point.

Vanishing Gradients:

Vanishing Gradient occurs when the gradient is smaller than expected. During backpropagation, this gradient becomes smaller that causing the decrease in the learning rate of earlier layers than the later layer of the network. Once this happens, the weight parameters update until they become insignificant.

Exploding Gradient:

Exploding gradient is just opposite to the vanishing gradient as it occurs when the Gradient is too large and creates a stable model. Further, in this scenario, model weight increases, and they will be represented as NaN. This problem can be solved using the dimensionality reduction technique, which helps to minimize complexity within the model.

Example program:

```

import numpy as np
def gradient_descent(x,y):
    m_curr = b_curr = 0
    iterations = 10000
    n = len(x)
    learning_rate = 0.08

    for i in range(iterations):
        y_predicted = m_curr * x + b_curr
        cost = (1/n) * sum([val**2 for val in (y-y_predicted)])
        md = -(2/n)*sum(x*(y-y_predicted))
        bd = -(2/n)*sum(y-y_predicted)
        m_curr = m_curr - learning_rate * md
        b_curr = b_curr - learning_rate * bd
        print ("m {}, b {}, cost {} iteration {}".format(m_curr,b_curr,cost, i))
x = np.array([1,2,3,4,5])
y = np.array([5,7,9,11,13])

```



```
gradient_descent(x,y)
```

Out[]:

```
m 4.96, b 1.44, cost 89.0 iteration 0
m 0.4991999999999983, b 0.2687999999999993, cost 71.10560000000002 iteration 1
m 4.4515840000000002, b 1.4261760000000001, cost 56.8297702400001 iteration 2
.
.
m 2.0000000000000002, b 2.999999999999995, cost 1.0255191767873153e-29 iteration 9997
m 2.0000000000000001, b 2.9999999999999947, cost 1.0255191767873153e-29 iteration 9998
m 2.0000000000000002, b 2.999999999999995, cost 1.0255191767873153e-29 iteration 9999
```

Optimization in a Machine Learning

Machine learning optimization is the process of adjusting hyperparameters in order to minimize the cost function by using one of the optimization techniques. It is important to minimize the cost function because it describes the discrepancy between the true value of the estimated parameter and what the model has predicted. Optimization plays an important part in a machine learning project in addition to fitting the learning algorithm on the training dataset.

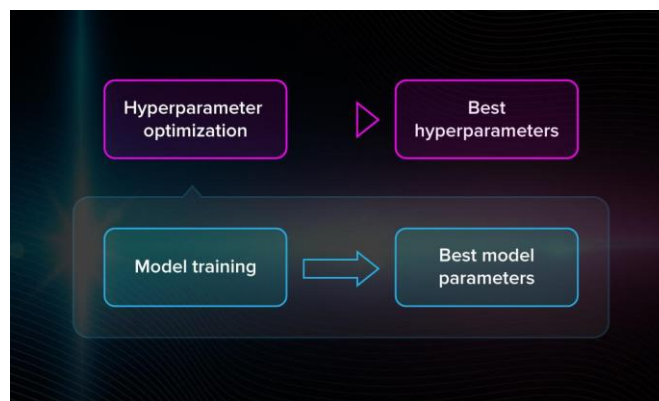
The step of preparing the data prior to fitting the model and the step of tuning a chosen model also can be framed as an optimization problem. In fact, an entire predictive modeling project can be thought of as one large optimization problem.

Parameters and hyperparameters of the model

Before we go any further, we need to understand the difference between parameters and hyperparameters of a model. These two notions are easy to confuse but we ought not to.

You need to set hyperparameters before starting to train the model. They include a number of clusters, learning rate, etc. Hyperparameters describe the structure of the model.

On the other hand, the parameters of the model are obtained during the training. There is no way to get them in advance. Examples are weights and biases for neural networks. This data is internal to the model and changes based on the inputs.



To tune the model, we need hyperparameter optimization. By finding the optimal combination of their values, we can decrease the error and build the most accurate model.

How hyperparameter tuning works:

As we said, the hyperparameters are set before training. But you can't know in advance, for instance, which learning rate (large or small) is best in this or that case. Therefore, to improve the model's performance, hyperparameters have to be optimized.

After each iteration, you compare the output with expected results, assess the accuracy, and adjust the hyperparameters if necessary. This is a repeated process. You can do that manually or use one of the many optimization techniques, which come in handy when you work with large amounts of data

Top optimization techniques in machine learning

Now let us talk about the techniques that you can use to optimize the hyperparameters of your model.

Exhaustive search

Exhaustive search, or brute-force search, is the process of looking for the most optimal hyperparameters by checking whether each candidate is a good match. You perform the same thing when you forget the code for your bike's lock and try out all the possible options. In machine learning, we do the same thing but the number of options is quite large, usually.

The exhaustive search method is simple. For example, if you are working with a k-means algorithm, you will manually search for the right number of clusters. However, if there are hundreds and thousands of options that you have to consider, it becomes unbearably heavy and slow. This makes brute-force search inefficient in the majority of real-life cases.

Gradient descent

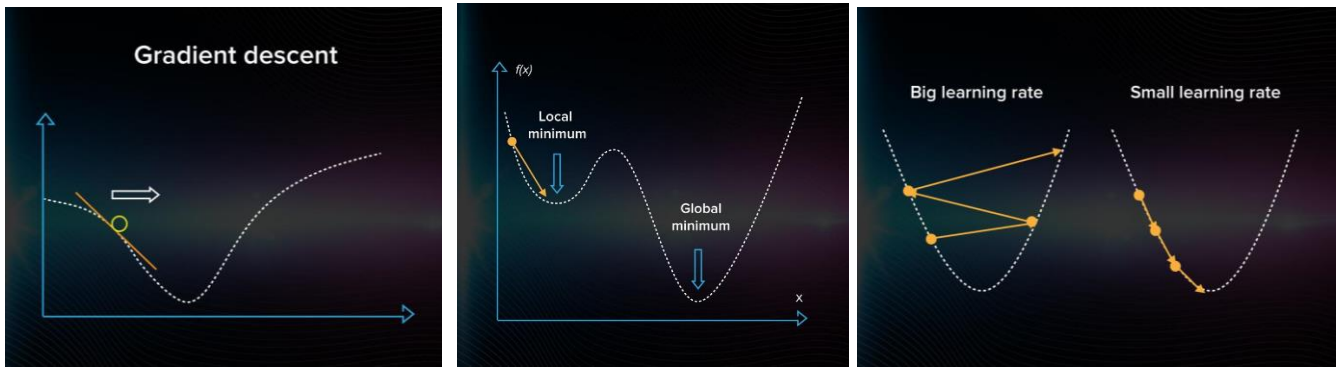
Gradient descent is the most common algorithm for model optimization for minimizing the error. In order to perform gradient descent, you have to iterate over the training dataset while re-adjusting the model.

Your goal is to minimize the cost function because it means you get the smallest possible error and improve the accuracy of the model.

On the graph, you can see a graphical representation of how the gradient descent algorithm travels in the variable space. To get started, you need to take a random point on the graph and arbitrarily choose a direction. If you see that the error is getting larger, that means you chose the wrong direction.

When you are not able to improve (decrease the error) anymore, the optimization is over and you have found a local minimum. In the following video, you will find a step-by-step explanation of how gradient descent works.

Looks fine so far. However, classical gradient descent will not work well when there are a couple of local minimums. Finding your first minimum, you will simply stop searching because the algorithm only finds a local one, it is not made to find the global one.



Note: In gradient descent, you proceed forward with steps of the same size. If you choose a learning rate that is too large, the algorithm will be jumping around without getting closer to the right answer. If it's too small, the computation will start mimicking exhaustive search take, which is, of course, inefficient.

So you have to choose the learning rate very carefully. If done right, gradient descent becomes a computation-efficient and rather quick method to optimize models.

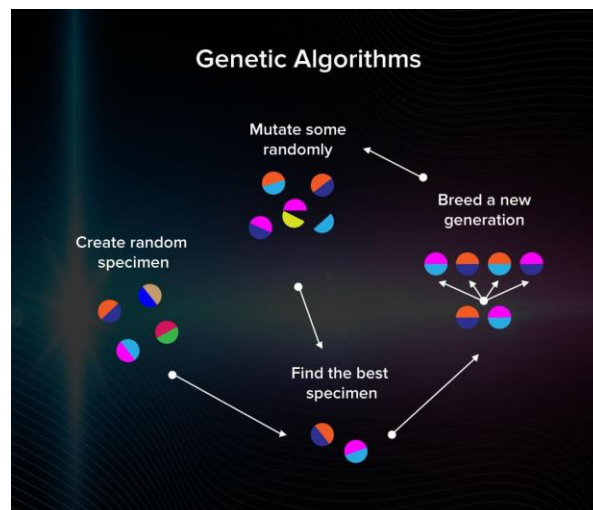
Genetic algorithms

Genetic algorithms represent another approach to ML optimization. The principle that lays behind the logic of these algorithms is an attempt to apply the theory of evolution to machine learning.

In the evolution theory, only those specimens get to survive and reproduce that have the best adaptation mechanisms. How do you know what specimens are and aren't the best in the case of machine learning models?

Imagine you have a bunch of random algorithms at hand. This will be your population. Among multiple models with some predefined hyperparameters, some are better adjusted than the others. Let's find them! First, you calculate the accuracy of each model. Then, you keep only those that worked out best. Now you can generate some descendants with similar hyperparameters to the best models to get a second generation of models.

We repeat this process many times and only the best models will survive at the end of the process. Genetic algorithms help to avoid being stuck at local minima/maxima. They are common in optimizing neural network models.



Regularization in Machine Learning

What is Regularization?

Regularization is one of the most important concepts of machine learning. It is a technique to prevent the model from overfitting by adding extra information to it.

Sometimes the machine learning model performs well with the training data but does not perform well with the test data. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the model is called overfitted. This problem can be deal with the help of a regularization technique.

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "*In regularization technique, we reduce the magnitude of the features by keeping the same number of features.*"

How does Regularization Work?

Regularization works by adding a penalty or complexity term to the complex model. Let's consider the simple linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n + b$$

In the above equation, Y represents the value to be predicted

X_1, X_2, \dots, X_n are the features for Y.

$\beta_0, \beta_1, \dots, \beta_n$ are the weights or magnitude attached to the features, respectively. Here represents the bias of the model, and b represents the intercept.

Linear regression models try to optimize the β_0 and b to minimize the cost function. The equation for the cost function for the linear model is given below:

$$\sum_{i=1}^M (y_i - y'_i)^2 = \sum_{i=1}^M (y_i - \sum_{j=0}^n \beta_j * X_{ij})^2$$

Now, we will add a loss function and optimize parameter to make the model that can predict the accurate value of Y. The loss function for the linear regression is called as **RSS or Residual sum of squares**.

Techniques of Regularization

There are mainly two types of regularization techniques, which are given below:

- **Ridge Regression**
- **Lasso Regression**

Ridge Regression

- Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions.

- Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as **L2 regularization**.
- In this technique, the cost function is altered by adding the penalty term to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
- The equation for the cost function in ridge regression will be:

$$\sum_{i=1}^M (y_i - y'_i)^2 = \sum_{i=1}^M \left(y_i - \sum_{j=0}^n \beta_j * x_{ij} \right)^2 + \lambda \sum_{j=0}^n \beta_j^2$$

- In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.
- As we can see from the above equation, if the values of λ **tend to zero, the equation becomes the cost function of the linear regression model**. Hence, for the minimum value of λ , the model will resemble the linear regression model.
- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
- It helps to solve the problems if we have more parameters than samples.

Lasso Regression:

- Lasso regression is another regularization technique to reduce the complexity of the model. It stands for **Least Absolute and Selection Operator**.
- It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.
- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as **L1 regularization**. The equation for the cost function of Lasso regression will be:

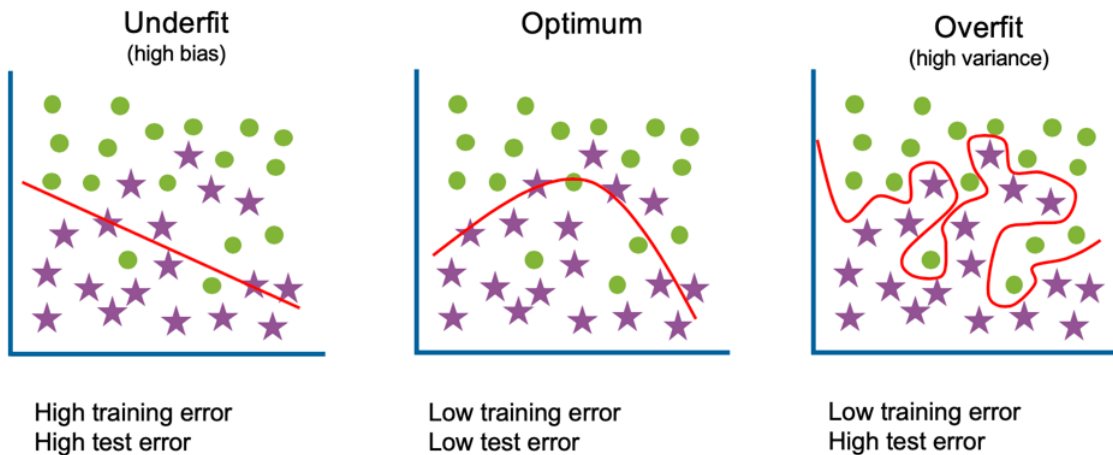
$$\sum_{i=1}^M (y_i - y'_i)^2 = \sum_{i=1}^M \left(y_i - \sum_{j=0}^n \beta_j * x_{ij} \right)^2 + \lambda \sum_{j=0}^n |\beta_j|$$

- Some of the features in this technique are completely neglected for model evaluation.
- Hence, the Lasso regression can help us to reduce the overfitting in the model as well as the feature selection.

Key Difference between Ridge Regression and Lasso Regression

- **Ridge regression** is mostly used to reduce the overfitting in the model, and it includes all the features present in the model. It reduces the complexity of the model by shrinking the coefficients.
- **Lasso regression** helps to reduce the overfitting in the model as well as feature selection.

What is Overfitting?



- Overfitting & underfitting are the two main errors/problems in the machine learning model, which cause poor performance in Machine Learning.
- Overfitting occurs when the model fits more data than required, and it tries to capture each and every datapoint fed to it. Hence it starts capturing noise and inaccurate data from the dataset, which degrades the performance of the model.
- An overfitted model doesn't perform accurately with the test/unseen dataset and can't generalize well.
- An overfitted model is said to have low bias and high variance.

Example to Understand Overfitting

We can understand overfitting with a general example. Suppose there are three students, X, Y, and Z, and all three are preparing for an exam. X has studied only three sections of the book and left all other sections. Y has a good memory, hence memorized the whole book. And the third student, Z, has studied and practiced all the questions. So, in the exam, X will only be able to solve the questions if the exam has questions related to section 3. Student Y will only be able to solve questions if they appear exactly the same as given in the book. Student Z will be able to solve all the exam questions in a proper way.

The same happens with machine learning; if the algorithm learns from a small part of the data, it is unable to capture the required data points and hence under fitted.

Suppose the model learns the training dataset, like the Y student. They perform very well on the seen dataset but perform badly on unseen data or unknown instances. In such cases, the model is said to be Overfitting.

And if the model performs well with the training dataset and also with the test/unseen dataset, similar to student Z, it is said to be a good fit.

How to detect Overfitting?

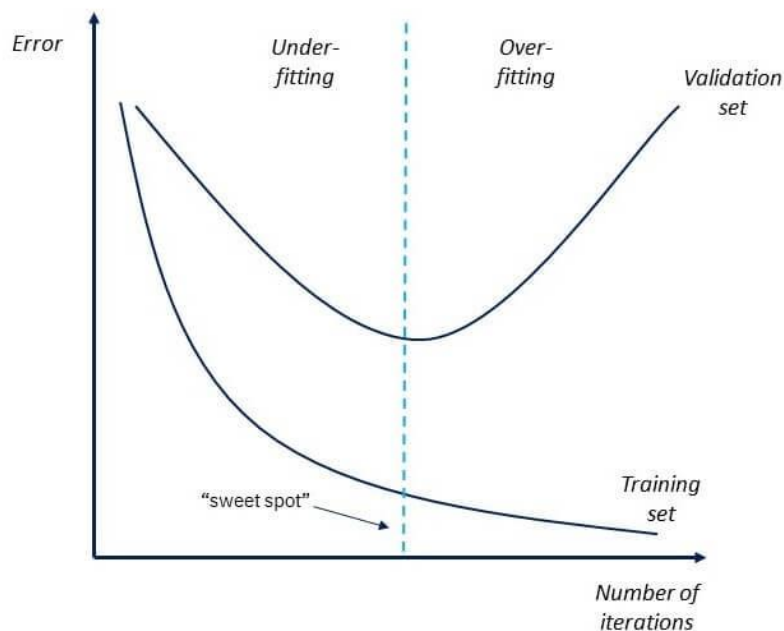
Overfitting in the model can only be detected once you test the data. To detect the issue, we can perform **Train/test split**.

In the train-test split of the dataset, we can divide our dataset into random test and training datasets. We train the model with a training dataset which is about 80% of the total dataset. After training the model, we test it with the test dataset, which is 20 % of the total dataset.



Now, if the model performs well with the training dataset but not with the test dataset, then it is likely to have an overfitting issue.

For example, if the model shows 85% accuracy with training data and 50% accuracy with the test dataset, it means the model is not performing well.



Ways to prevent the Overfitting

Although overfitting is an error in Machine learning which reduces the performance of the model, however, we can prevent it in several ways. With the use of the linear model, we can avoid overfitting; however, many real-world problems are non-linear ones. It is important to prevent overfitting from the models. Below are several ways that can be used to prevent overfitting:

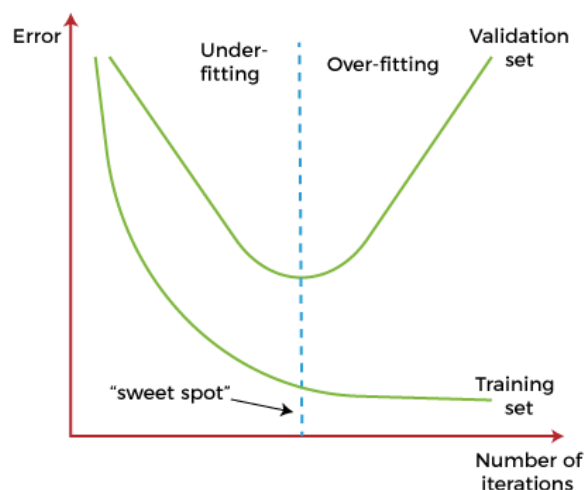
1. **Early Stopping**
2. **Train with more data**
3. **Feature Selection**
4. **Cross-Validation**
5. **Data Augmentation**
6. **Regularization**

1. Early Stopping

In this technique, the training is paused before the model starts learning the noise within the model. In this process, while training the model iteratively, measure the performance of the model after each iteration. Continue up to a certain number of iterations until a new iteration improves the performance of the model.

After that point, the model begins to overfit the training data; hence we need to stop the process before the learner passes that point.

Stopping the training process before the model starts capturing noise from the data is known as **early stopping**.



However, this technique may lead to the underfitting problem if training is paused too early. So, it is very important to find that "sweet spot" between underfitting and overfitting.

2. Train with More data

Increasing the training set by including more data can enhance the accuracy of the model, as it provides more chances to discover the relationship between input and output variables.

It may not always work to prevent overfitting, but this way helps the algorithm to detect the signal better to minimize the errors.

When a model is fed with more training data, it will be unable to overfit all the samples of data and forced to generalize well.

But in some cases, the additional data may add more noise to the model; hence we need to be sure that data is clean and free from in-consistencies before feeding it to the model.

3.Feature Selection

While building the ML model, we have a number of parameters or features that are used to predict the outcome. However, sometimes some of these features are redundant or less important for the prediction, and for this feature selection process is applied. In the feature selection process, we identify the most important features within training data, and other features are removed. Further, this process helps to simplify the model and reduces noise from the data. Some algorithms have the auto-feature selection, and if not, then we can manually perform this process.

4.Cross-Validation

Cross-validation is one of the powerful techniques to prevent overfitting.

In the general k-fold cross-validation technique, we divided the dataset into k-equal-sized subsets of data; these subsets are known as folds.

5.Data Augmentation

Data Augmentation is a data analysis technique, which is an alternative to adding more data to prevent overfitting. In this technique, instead of adding more training data, slightly modified copies of already existing data are added to the dataset.

The data augmentation technique makes it possible to appear data sample slightly different every time it is processed by the model. Hence each data set appears unique to the model and prevents overfitting.

6.Regularization

If overfitting occurs when a model is complex, we can reduce the number of features. However, overfitting may also occur with a simpler model, more specifically the Linear model, and for such cases, regularization techniques are much helpful.

Regularization is the most popular technique to prevent overfitting. It is a group of methods that forces the learning algorithms to make a model simpler. Applying the regularization technique may slightly increase the bias but slightly reduces the variance. In this technique, we modify the objective function by adding the penalizing term, which has a higher value with a more complex model.

The two commonly used regularization techniques are L1 Regularization and L2 Regularization.

Ensemble Methods

In ensemble methods, prediction from different machine learning models is combined to identify the most popular result.

The most commonly used ensemble methods are **Bagging and Boosting**.

In bagging, individual data points can be selected more than once. After the collection of several sample datasets, these models are trained independently, and depending on the type of task-i.e., regression or classification-the average of those predictions is used to predict a more accurate result. Moreover, bagging reduces the chances of overfitting in complex models.

Perceptron in Machine Learning

In Machine Learning and Artificial Intelligence, Perceptron is the most commonly used term for all folks. It is the primary step to learn Machine Learning and Deep Learning technologies, which consists of a set of weights, input values or scores, and a threshold. **Perceptron is a building block of an Artificial Neural Network.** Initially, in the mid of 19th century, **Mr. Frank Rosenblatt** invented the Perceptron for performing certain calculations to detect input data capabilities or business intelligence. Perceptron is a linear Machine Learning algorithm used for supervised learning for various binary classifiers. This algorithm enables neurons to learn elements and processes them one by one during preparation. In this tutorial, "Perceptron in Machine Learning," we will discuss in-depth knowledge of Perceptron and its basic functions in brief. Let's start with the basic introduction of Perceptron.

What is the Perceptron model in Machine Learning?

Perceptron is Machine Learning algorithm for supervised learning of various binary classification tasks. Further, **Perceptron is also understood as an Artificial Neuron or neural network unit that helps to detect certain input data computations in business intelligence.**

Perceptron model is also treated as one of the best and simplest types of Artificial Neural networks. However, it is a supervised learning algorithm of binary classifiers. Hence, we can consider it as a single-layer neural network with four main parameters, i.e., **input values, weights and Bias, net sum, and an activation function.**

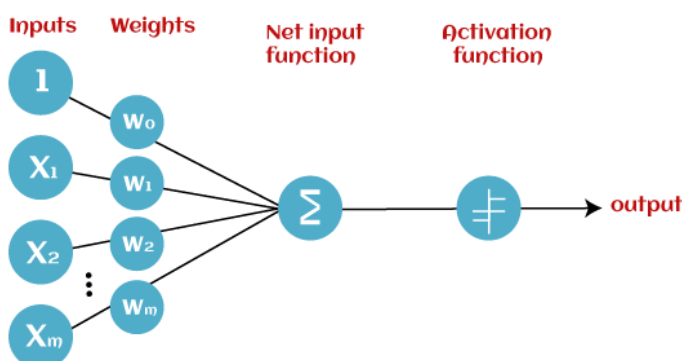
What is Binary classifier in Machine Learning?

In Machine Learning, binary classifiers are defined as the function that helps in deciding whether input data can be represented as vectors of numbers and belongs to some specific class.

Binary classifiers can be considered as linear classifiers. In simple words, we can understand it as a **classification algorithm that can predict linear predictor function in terms of weight and feature vectors.**

Basic Components of Perceptron

Mr. Frank Rosenblatt invented the perceptron model as a binary classifier which contains three main components. These are as follows:



- **Input Nodes or Input Layer:**

This is the primary component of Perceptron which accepts the initial data into the system for further processing. Each input node contains a real numerical value.

- **Wight and Bias:**

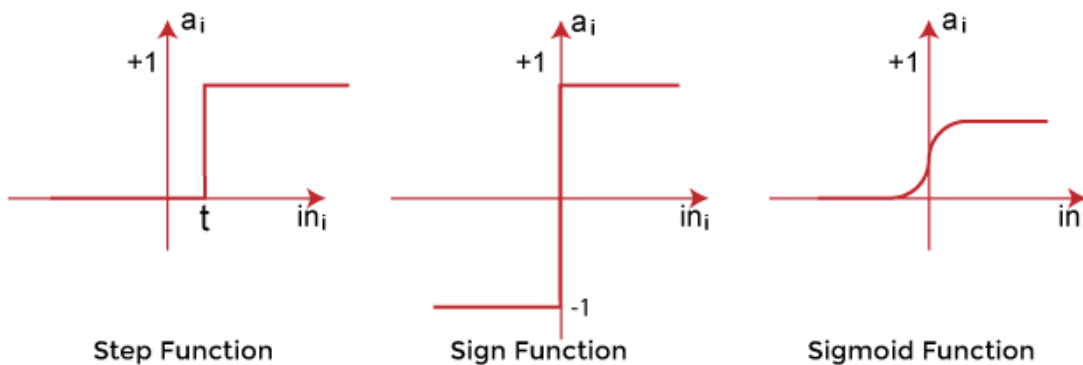
Weight parameter represents the strength of the connection between units. This is another most important parameter of Perceptron components. Weight is directly proportional to the strength of the associated input neuron in deciding the output. Further, Bias can be considered as the line of intercept in a linear equation.

o **Activation Function:**

These are the final and important components that help to determine whether the neuron will fire or not. Activation Function can be considered primarily as a step function.

Types of Activation functions:

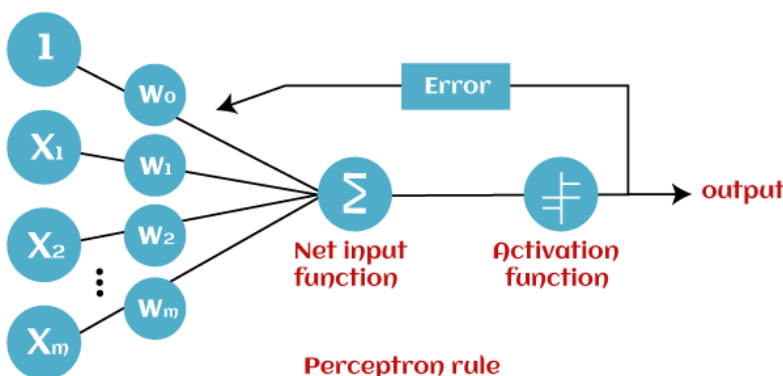
- o Sign function
- o Step function, and
- o Sigmoid function



The data scientist uses the activation function to take a subjective decision based on various problem statements and forms the desired outputs. Activation function may differ (e.g., Sign, Step, and Sigmoid) in perceptron models by checking whether the learning process is slow or has vanishing or exploding gradients.

How does Perceptron work?

In Machine Learning, Perceptron is considered as a single-layer neural network that consists of four main parameters named input values (Input nodes), weights and Bias, net sum, and an activation function. The perceptron model begins with the multiplication of all input values and their weights, then adds these values together to create the weighted sum. Then this weighted sum is applied to the activation function 'f' to obtain the desired output. This activation function is also known as the **step function** and is represented by 'f'.



This step function or Activation function plays a vital role in ensuring that output is mapped between required values (0,1) or (-1,1). It is important to note that the weight of input is indicative of the strength of a node. Similarly, an input's bias value gives the ability to shift the activation function curve up or down.

Perceptron model works in two important steps as follows:

Step-1

In the first step first, multiply all input values with corresponding weight values and then add them to determine the weighted sum. Mathematically, we can calculate the weighted sum as follows:

$$\sum w_i * x_i = x_1 * w_1 + x_2 * w_2 + \dots w_n * x_n$$

Add a special term called **bias 'b'** to this weighted sum to improve the model's performance.

$$\sum w_i * x_i + b$$

Step-2

In the second step, an activation function is applied with the above-mentioned weighted sum, which gives us output either in binary form or a continuous value as follows:

$$Y = f(\sum w_i * x_i + b)$$

Types of Perceptron Models

Based on the layers, Perceptron models are divided into two types. These are as follows:

1. Single-layer Perceptron Model
2. Multi-layer Perceptron model

Single Layer Perceptron Model:

This is one of the easiest Artificial neural networks (ANN) types. A single-layered perceptron model consists feed-forward network and also includes a threshold transfer function inside the model. The main objective of the single-layer perceptron model is to analyze the linearly separable objects with binary outcomes.

In a single layer perceptron model, its algorithms do not contain recorded data, so it begins with inconstantly allocated input for weight parameters. Further, it sums up all inputs (weight). After adding all inputs, if the total sum of all inputs is more than a pre-determined value, the model gets activated and shows the output value as +1.

If the outcome is same as pre-determined or threshold value, then the performance of this model is stated as satisfied, and weight demand does not change. However, this model consists of a few discrepancies triggered when multiple weight inputs values are fed into the model. Hence, to find desired output and minimize errors, some changes should be necessary for the weights input.

"Single-layer perceptron can learn only linearly separable patterns."

Multi-Layered Perceptron Model:

Like a single-layer perceptron model, a multi-layer perceptron model also has the same model structure but has a greater number of hidden layers.

The multi-layer perceptron model is also known as the Backpropagation algorithm, which executes in two stages as follows:

- **Forward Stage:** Activation functions start from the input layer in the forward stage and terminate on the output layer.
- **Backward Stage:** In the backward stage, weight and bias values are modified as per the model's requirement. In this stage, the error between actual output and demanded originated backward on the output layer and ended on the input layer.

Hence, a multi-layered perceptron model has considered as multiple artificial neural networks having various layers in which activation function does not remain linear, similar to a single layer perceptron model. Instead of linear, activation function can be executed as sigmoid, TanH, ReLU, etc., for deployment.

A multi-layer perceptron model has greater processing power and can process linear and non-linear patterns. Further, it can also implement logic gates such as AND, OR, XOR, NAND, NOT, XNOR, NOR.

Advantages of Multi-Layer Perceptron:

- A multi-layered perceptron model can be used to solve complex non-linear problems.
- It works well with both small and large input data.
- It helps us to obtain quick predictions after the training.
- It helps to obtain the same accuracy ratio with large as well as small data.

Disadvantages of Multi-Layer Perceptron:

- In Multi-layer perceptron, computations are difficult and time-consuming.
- In multi-layer Perceptron, it is difficult to predict how much the dependent variable affects each independent variable.
- The model functioning depends on the quality of the training.

What Is a Neural Network?

A neural network is a series of algorithms that endeavors to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates. In this sense, neural networks refer to systems of neurons, either organic or artificial in nature.

Neural networks can adapt to changing input; so the network generates the best possible result without needing to redesign the output criteria. The concept of neural networks, which has its roots in artificial intelligence, is swiftly gaining popularity in the development of trading systems.

Pros

- Can often work more efficiently and for longer than humans
- Can be programmed to learn from prior outcomes to strive to make smarter future calculations
- Often leverage online services that reduce (but do not eliminate) systematic risk
- Are continually being expanded in new fields with more difficult problems

Cons

- Still rely on hardware that may require labor and expertise to maintain
- May take long periods of time to develop the code and algorithms
- May be difficult to assess errors or adaptations to the assumptions if the system is self-learning but lacks transparency
- Usually report an estimated range or estimated amount that may not actualize

Multi-Class Classification

Multi-class classification is perhaps the most popular machine learning job, aside from regression.

The science behind it is the same whether it's spelled multiclass or multi-class. An ML classification problem with more than two outputs or classes is known as multi feature classification. Because each image may be classed as many distinct animal categories, using a machine learning model to identify animal species in photographs from an encyclopedia is an example of multi-class classification. Multi-class classification also necessitates the use of only one class in a sample (ie. an elephant is only an elephant; it is not also a lemur).

We are given a set of training samples separated into K distinct classes, and we create an ML model to forecast which of those classes some previously unknown data belongs to. The model learns patterns specific to each class from the training dataset and utilizes those patterns to forecast the classification of future data.

Approach –

1. Load dataset from the source.
2. Split the dataset into “training” and “test” data.
3. Train Decision tree, SVM, and KNN classifiers on the training data.
4. Use the above classifiers to predict labels for the test data.
5. Measure accuracy and visualize classification.

Decision tree classifier – A decision tree classifier is a systematic approach for multiclass classification. It poses a set of questions to the dataset (related to its attributes/features). The decision tree classification algorithm can be visualized on a binary tree. On the root and each of the internal nodes, a question is posed and the data on that node is further split into separate records that have different characteristics. The leaves of the tree refer to the classes in which the dataset is split. In the following code snippet, we train a decision tree classifier in scikit-learn.

Example:

```
# importing necessary libraries
from sklearn import datasets
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import train_test_split

# loading the iris dataset
iris = datasets.load_iris()

# X -> features, y -> label
X = iris.data
y = iris.target

# dividing X, y into train and test data
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state = 0)

# training a DescisionTreeClassifier
from sklearn.tree import DecisionTreeClassifier
dtree_model = DecisionTreeClassifier(max_depth = 2).fit(X_train, y_train)
dtree_predictions = dtree_model.predict(X_test)

# creating a confusion matrix
cm = confusion_matrix(y_test, dtree_predictions)
```

What is a backpropagation algorithm?

Backpropagation, or backward propagation of errors, is an algorithm that is designed to test for errors working back from output nodes to input nodes. It is an important mathematical tool for improving the accuracy of predictions in data mining and machine learning. Essentially, backpropagation is an algorithm used to calculate derivatives quickly.

There are two leading types of backpropagation networks:

1. **Static backpropagation.** Static backpropagation is a network developed to map static inputs for static outputs. Static backpropagation networks can solve static classification problems, such as optical character recognition (OCR).
2. **Recurrent backpropagation.** The recurrent backpropagation network is used for fixed-point learning. Recurrent backpropagation activation feeds forward until it reaches a fixed value.

What is a backpropagation algorithm in a neural network?

Artificial neural networks use backpropagation as a learning algorithm to compute a gradient descent with respect to weight values for the various inputs. By comparing desired outputs to achieved system outputs, the systems are tuned by adjusting connection weights to narrow the difference between the two as much as possible.

The algorithm gets its name because the weights are updated backward, from output to input.

The advantages of using a backpropagation algorithm are as follows:

- It does not have any parameters to tune except for the number of inputs.
- It is highly adaptable and efficient and does not require any prior knowledge about the network.
- It is a standard process that usually works well.
- It is user-friendly, fast and easy to program.
- Users do not need to learn any special functions.

The disadvantages of using a backpropagation algorithm are as follows:

- It prefers a matrix-based approach over a mini-batch approach.
- Data mining is sensitive to noise and irregularities.
- Performance is highly dependent on input data.
- Training is time- and resource-intensive.

What is a backpropagation algorithm in machine learning?

Backpropagation requires a known, desired output for each input value in order to calculate the loss function gradient -- how a prediction differs from actual results -- as a type of supervised machine learning. Along with classifiers such as Naïve Bayesian filters and decision trees, the backpropagation training algorithm has emerged as an important part of machine learning applications that involve predictive analytics.

What is the time complexity of a backpropagation algorithm?

The time complexity of each iteration -- how long it takes to execute each statement in an algorithm - depends on the network's structure. For multilayer perceptron, matrix multiplications dominate time.

Non-Linear Activation Functions

Examples of non-linear activation functions include:

1. Sigmoid function: The Sigmoid function exists between 0 and 1 or -1 and 1. The use of a sigmoid function is to convert a real value to a probability. In machine learning, the sigmoid function is generally used to refer to the logistic function, also called the logistic sigmoid function; it is also the most widely used sigmoid function (others are the hyperbolic tangent and the arctangent).

A sigmoid function is placed as the last layer of the model to convert the model's output into a probability score, which is easier to work with and interpret.

Another reason to use it mostly in the output layer is that it can otherwise cause a neural network to get stuck in training time.

2. TanH function: It is the hyperbolic tangent function whose range lies between -1 and 1, hence also called the zero-centred function. Because it is zero centred, it is much easier to model inputs with strongly negative, positive or neutral values. TanH function is used instead of sigmoid function if the output is other than 0 and 1. TanH functions usually find applications in RNN for natural language processing and speech recognition tasks.

On the downside, in the case of both Sigmoid and TanH, if the weighted sum input is very large or very small, the function's gradient becomes very small and closer to zero.

3. ReLU function: Rectified Linear Unit, also called ReLU, is a widely favoured activation function for deep learning applications. Compared to Sigmoid and TanH activation functions, ReLU offers an upper hand in terms of performance and generalisation. In terms of computation too, ReLU is faster as it does not compute exponentials and divisions. The disadvantage is that ReLU overfits more, as compared with Sigmoid.

4. Parametric ReLU (PReLU): ReLU has been one of the keys to the recent successes in deep learning. Its use has led to better solutions than that of sigmoid. This is partially due to the vanishing gradient problem in case of sigmoid activations. But, we can still improve upon ReLU. LeakyReLU was introduced, which doesn't zero out the negative inputs as ReLU does. Instead, it multiplies the negative input by a small value (like 0.02) and keeps the positive input as is. But this has shown a negligible increase in the accuracy of our models.

Dropout is a regularization

Dropout is a regularization technique for reducing overfitting in neural networks by preventing complex co-adaptations on training data. It is a very efficient way of performing model averaging with neural networks. The term "dropout" refers to dropping out units (both hidden and visible) in a neural network.

A simple and powerful regularization technique for neural networks and deep learning models is dropout. This notebook will uncover the dropout regularization technique and how to apply it to deep learning models in Python with Keras.

Dropout is a technique where randomly selected neurons are ignored during training. They are "dropped-out" randomly. This means that their contribution to the activation of downstream neurons is temporally removed on the forward pass and any weight updates are not applied to the neuron on the backward pass.

As a neural network learns, neuron weights settle into their context within the network. Weights of neurons are tuned for specific features providing some specialization. Neighboring neurons become to rely on this specialization, which if taken too far can result in a fragile model too specialized to the training data. This reliance on context for a neuron during training is referred to as complex co-adaptations.

COURSE OUTCOMES:

CO1:Understand about Data Preprocessing, Dimensionality reduction

CO2:Apply proper model for the given problem and use feature engineering techniques

CO3:Make use of Probability Technique to solve the given problem.

CO4:Analyze the working model and features of Decision tree

CO5:choose and apply appropriate algorithm to learn and classify the data

REFERENCES

1. Ethem Alpaydin, "Introduction to Machine Learning 3e (Adaptive Computation and Machine Learning Series)", Third Edition, MIT Press, 2014
2. Tom M. Mitchell, "Machine Learning", India Edition, 1st Edition, McGraw-Hill Education Private Limited, 2013
3. Saikat Dutt, Subramanian Chandramouli and Amit Kumar Das, "Machine Learning", 1st Edition, Pearson Education, 2019
4. Christopher M. Bishop, "Pattern Recognition and Machine Learning", Revised Edition, Springer, 2016.
5. Aurelien Geron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow", 2nd Edition, O'Reilly, 2019
6. Stephen Marsland, "Machine Learning – An Algorithmic Perspective", Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

MC4302

INTERNET OF THINGS

**L T P C
3 0 0 3**

COURSE OBJECTIVES:

- To understand the concepts of IoT and its working models
- To know the various IoT protocols
- To understand about various IoT Physical devices and Endpoints
- To know the security and privacy issues connected with IoT
- To apply the concept of Internet of Things in a real world scenario.

UNIT I FUNDAMENTALS OF IOT

9

Definition and Characteristics of IoT, Sensors, Actuators, Physical Design of IoT – IoT Protocols, IoT communication models, IoT Communication APIs, IoT enabled Technologies – Wireless Sensor Networks, Cloud Computing, Embedded Systems, IoT Levels and Templates, Domain Specific IoTs – Home, City, Environment, Energy, Agriculture and Industry.

UNIT II IOT PROTOCOLS

9

Protocol Standardization for IoT – Efforts – M2M and WSN Protocols – SCADA and RFID Protocols – Issues with IoT Standardization – Unified Data Standards – Protocols – IEEE802.15.4–BACNet Protocol– Modbus – KNX – Zigbee– Network layer – APS layer – Security

UNIT III IOT PHYSICAL DEVICES AND ENDPOINTS

9

Introduction to Arduino and Raspberry Pi- Installation, Interfaces (serial, SPI, I2C), Programming – Python program with Raspberry PI with focus on interfacing external gadgets, controlling output, and reading input from pins.

UNIT IV INTERNET OF THINGS PRIVACY, SECURITY AND GOVERNANCE 9

Introduction, Overview of Governance, Privacy and Security Issues, Contribution from FP7 Projects, Security, Privacy and Trust in IoT-Data-Platforms for Smart Cities, First Steps Towards a Secure Platform, Smartie Approach. Data Aggregation for the IoT in Smart Cities, Security

UNIT V APPLICATIONS 9

IOT APPLICATIONS - IoT applications for industry: Future Factory Concepts, Brownfield IoT, Smart Objects, Smart Applications. Study of existing IoT platforms /middleware, IoT- A, Hydra etc.

SUGGESTED ACTIVITIES:

- 1: Study of 5 different types of sensors and actuators available in Market
- 2: Study of commercial IoT available in any one domain
- 3: Study the recent developments in IoT Protocol
- 4: Implement simple Python programs for IoT
- 5: Study on the latest government policies on IoT security and Privacy
- 6: A study on how to use IoT to solve some problems in your neighborhood.

TOTAL: 45 PERIODS

COURSE OUTCOMES:

Able to

CO1: Define the infrastructure for supporting IoT deployments

CO2: Understand the usage of IoT protocols for communication between various IoT devices

CO3: Design portable IoT using Arduino/Raspberry Pi /equivalent boards.

CO4: Understand the basic concepts of security and governance as applied to IoT

CO5: Analyze and illustrate applications of IoT in real time scenarios

REFERENCES

1. Internet of Things - A Hands-on Approach, Arshdeep Bahga and Vijay Madiseti, Universities Press, 2015, ISBN: 9788173719547
2. Olivier Hersent, David Boswarthick, Omar Elloumi , "The Internet of Things – Key applications and Protocols", Wiley, 2012. .
3. David Hanes, Gonzalo Salgueiro, Patrick Grossetete, Rob Barton, Jerome Henry, "IoT Fundamentals, Networking Technologies, Protocols, and Use cases for the Internet of Things", Cisco Press, First Edition,2017.
4. Dieter Uckelmann, Mark Harrison, Michahelles, Florian (Eds), "Architecting the Internet ofThings", Springer, 2011
5. Raspberry Pi Cookbook, Software and Hardware Problems and solutions, Simon Monk, O'Reilly (SPD), 2016, ISBN 7989352133895
6. Peter Friess,'Internet of Things – From Research and Innovation to Market Deployment', River Publishers, 2014

MC4311

MACHINE LEARNING LABORATORY

**L T P C
0 0 4 2**

COURSE OBJECTIVES:

- To understand about data cleaning and data preprocessing
- To familiarize with the Supervised Learning algorithms and implement them in practical situations.

THE INTERNET OF THINGS

THE INTERNET OF THINGS

KEY APPLICATIONS AND PROTOCOLS

Olivier Hersent

Actility, France

David Boswarthick

ETSI, France

Omar Elloumi

Alcatel-Lucent, France



A John Wiley & Sons, Ltd., Publication

This edition first published 2012
© 2012 John Wiley & Sons Ltd

Registered office

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

ETSI logo reproduced by kind permission of © ETSI, All Rights Reserved.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Library of Congress Cataloging-in-Publication Data

Hersent, Olivier.

The internet of things : key applications and protocols / Olivier Hersent,
David Boswarthick, Omar Elloumi.

p. cm.

Includes bibliographical references and index.

ISBN 978-1-119-99435-0 (hardback)

1. Intelligent buildings. 2. Smart power grids. 3. Sensor networks. I. Boswarthick, David.
II. Elloumi, Omar. III. Title.

TH6012.H47 2012

681'.2--dc23

2011037210

A catalogue record for this book is available from the British Library.

ISBN: 9781119994350 (H/B)

Typeset in 10.5/13pt Times by Aptara Inc., New Delhi, India

Contents

List of Acronyms	xv
Introduction	xxiii
Part I M2M AREA NETWORK PHYSICAL LAYERS	
1 IEEE 802.15.4	3
1.1 The IEEE 802 Committee Family of Protocols	3
1.2 The Physical Layer	3
1.2.1 <i>Interferences with Other Technologies</i>	5
1.2.2 <i>Choice of a 802.15.4 Communication Channel, Energy Detection, Link Quality Information</i>	7
1.2.3 <i>Sending a Data Frame</i>	8
1.3 The Media-Access Control Layer	8
1.3.1 <i>802.15.4 Reduced Function and Full Function Devices, Coordinators, and the PAN Coordinator</i>	9
1.3.2 <i>Association</i>	12
1.3.3 <i>802.15.4 Addresses</i>	13
1.3.4 <i>802.15.4 Frame Format</i>	13
1.3.5 <i>Security</i>	14
1.4 Uses of 802.15.4	16
1.5 The Future of 802.15.4: 802.15.4e and 802.15.4g	17
1.5.1 <i>802.15.4e</i>	17
1.5.2 <i>802.15.4g</i>	21
2 Powerline Communication for M2M Applications	23
2.1 Overview of PLC Technologies	23
2.2 PLC Landscape	23
2.2.1 <i>The Historical Period (1950–2000)</i>	24
2.2.2 <i>After Year 2000: The Maturity of PLC</i>	24

2.3	Powerline Communication: A Constrained Media	27
2.3.1	<i>Powerline is a Difficult Channel</i>	27
2.3.2	<i>Regulation Limitations</i>	27
2.3.3	<i>Power Consumption</i>	32
2.3.4	<i>Lossy Network</i>	33
2.3.5	<i>Powerline is a Shared Media and Coexistence is not an Optional Feature</i>	35
2.4	The Ideal PLC System for M2M	37
2.4.1	<i>Openness and Availability</i>	38
2.4.2	<i>Range</i>	38
2.4.3	<i>Power Consumption</i>	38
2.4.4	<i>Data Rate</i>	39
2.4.5	<i>Robustness</i>	39
2.4.6	<i>EMC Regulatory Compliance</i>	40
2.4.7	<i>Coexistence</i>	40
2.4.8	<i>Security</i>	40
2.4.9	<i>Latency</i>	40
2.4.10	<i>Interoperability with M2M Wireless Services</i>	40
2.5	Conclusion	40
	References	41

Part II LEGACY M2M PROTOCOLS FOR SENSOR NETWORKS, BUILDING AUTOMATION AND HOME AUTOMATION

3	The BACnet™ Protocol	45
3.1	Standardization	45
3.1.1	<i>United States</i>	46
3.1.2	<i>Europe</i>	46
3.1.3	<i>Interworking</i>	46
3.2	Technology	46
3.2.1	<i>Physical Layer</i>	47
3.2.2	<i>Link Layer</i>	47
3.2.3	<i>Network Layer</i>	47
3.2.4	<i>Transport and Session Layers</i>	49
3.2.5	<i>Presentation and Application Layers</i>	49
3.3	BACnet Security	55
3.4	BACnet Over Web Services (Annex N, Annex H6)	55
3.4.1	<i>The Generic WS Model</i>	56
3.4.2	<i>BACnet/WS Services</i>	58
3.4.3	<i>The Web Services Profile for BACnet Objects</i>	59
3.4.4	<i>Future Improvements</i>	59

4	The LonWorks® Control Networking Platform	61
4.1	Standardization	61
4.1.1	<i>United States of America</i>	61
4.1.2	<i>Europe</i>	62
4.1.3	<i>China</i>	62
4.2	Technology	62
4.2.1	<i>Physical Layer</i>	63
4.2.2	<i>Link Layer</i>	64
4.2.3	<i>Network Layer</i>	65
4.2.4	<i>Transport Layer</i>	66
4.2.5	<i>Session Layer</i>	67
4.2.6	<i>Presentation Layer</i>	67
4.2.7	<i>Application Layer</i>	71
4.3	Web Services Interface for LonWorks Networks: Echelon SmartServer	72
4.4	A REST Interface for LonWorks	73
4.4.1	<i>LonBridge REST Transactions</i>	74
4.4.2	<i>Requests</i>	74
4.4.3	<i>Responses</i>	75
4.4.4	<i>LonBridge REST Resources</i>	75
5	ModBus	79
5.1	Introduction	79
5.2	ModBus Standardization	80
5.3	ModBus Message Framing and Transmission Modes	80
5.4	ModBus/TCP	81
6	KNX	83
6.1	The Konnex/KNX Association	83
6.2	Standardization	83
6.3	KNX Technology Overview	84
6.3.1	<i>Physical Layer</i>	84
6.3.2	<i>Data Link and Routing Layers, Addressing</i>	87
6.3.3	<i>Transport Layer</i>	89
6.3.4	<i>Application Layer</i>	89
6.3.5	<i>KNX Devices, Functional Blocks and Interworking</i>	89
6.4	Device Configuration	92
7	ZigBee	93
7.1	Development of the Standard	93
7.2	ZigBee Architecture	94

7.2.1	<i>ZigBee and 802.15.4</i>	94
7.2.2	<i>ZigBee Protocol Layers</i>	94
7.2.3	<i>ZigBee Node Types</i>	96
7.3	<i>Association</i>	96
7.3.1	<i>Forming a Network</i>	96
7.3.2	<i>Joining a Parent Node in a Network Using 802.15.4 Association</i>	97
7.3.3	<i>Using NWK Rejoin</i>	99
7.4	<i>The ZigBee Network Layer</i>	99
7.4.1	<i>Short-Address Allocation</i>	99
7.4.2	<i>Network Layer Frame Format</i>	100
7.4.3	<i>Packet Forwarding</i>	101
7.4.4	<i>Routing Support Primitives</i>	101
7.4.5	<i>Routing Algorithms</i>	102
7.5	<i>The ZigBee APS Layer</i>	105
7.5.1	<i>Endpoints, Descriptors</i>	106
7.5.2	<i>The APS Frame</i>	106
7.6	<i>The ZigBee Device Object (ZDO) and the ZigBee Device Profile (ZDP)</i>	109
7.6.1	<i>ZDP Device and Service Discovery Services (Mandatory)</i>	109
7.6.2	<i>ZDP Network Management Services (Mandatory)</i>	110
7.6.3	<i>ZDP Binding Management Services (Optional)</i>	111
7.6.4	<i>Group Management</i>	111
7.7	<i>ZigBee Security</i>	111
7.7.1	<i>ZigBee and 802.15.4 Security</i>	111
7.7.2	<i>Key Types</i>	113
7.7.3	<i>The Trust Center</i>	114
7.7.4	<i>The ZDO Permissions Table</i>	116
7.8	<i>The ZigBee Cluster Library (ZCL)</i>	116
7.8.1	<i>Cluster</i>	116
7.8.2	<i>Attributes</i>	117
7.8.3	<i>Commands</i>	117
7.8.4	<i>ZCL Frame</i>	117
7.9	<i>ZigBee Application Profiles</i>	119
7.9.1	<i>The Home Automation (HA) Application Profile</i>	119
7.9.2	<i>ZigBee Smart Energy 1.0 (ZSE or AMI)</i>	122
7.10	<i>The ZigBee Gateway Specification for Network Devices</i>	129
7.10.1	<i>The ZGD</i>	130
7.10.2	<i>GRIP Binding</i>	131
7.10.3	<i>SOAP Binding</i>	132
7.10.4	<i>REST Binding</i>	132
7.10.5	<i>Example IPHA–ZGD Interaction Using the REST Binding</i>	134

8	Z-Wave	139
8.1	History and Management of the Protocol	139
8.2	The Z-Wave Protocol	140
8.2.1	<i>Overview</i>	140
8.2.2	<i>Z-Wave Node Types</i>	140
8.2.3	<i>RF and MAC Layers</i>	142
8.2.4	<i>Transfer Layer</i>	143
8.2.5	<i>Routing Layer</i>	145
8.2.6	<i>Application Layer</i>	148

Part III LEGACY M2M PROTOCOLS FOR UTILITY METERING

9	M-Bus and Wireless M-Bus	155
9.1	Development of the Standard	155
9.2	M-Bus Architecture	156
9.2.1	<i>Physical Layer</i>	156
9.2.2	<i>Link Layer</i>	156
9.2.3	<i>Network Layer</i>	157
9.2.4	<i>Application Layer</i>	158
9.3	Wireless M-Bus	160
9.3.1	<i>Physical Layer</i>	160
9.3.2	<i>Data-Link Layer</i>	162
9.3.3	<i>Application Layer</i>	162
9.3.4	<i>Security</i>	163
10	The ANSI C12 Suite	165
10.1	Introduction	165
10.2	C12.19: The C12 Data Model	166
10.2.1	<i>The Read and Write Minimum Services</i>	167
10.2.2	<i>Some Remarkable C12.19 Tables</i>	167
10.3	C12.18: Basic Point-to-Point Communication Over an Optical Port	168
10.4	C12.21: An Extension of C12.18 for Modem Communication	169
10.4.1	<i>Interactions with the Data-Link Layer</i>	170
10.4.2	<i>Modifications and Additions to C12.19 Tables</i>	171
10.5	C12.22: C12.19 Tables Transport Over Any Networking Communication System	171
10.5.1	<i>Reference Topology and Network Elements</i>	171
10.5.2	<i>C12.22 Node to C12.22 Network Communications</i>	173
10.5.3	<i>C12.22 Device to C12.22 Communication Module Interface</i>	174
10.5.4	<i>C12.19 Updates</i>	176

10.6	Other Parts of ANSI C12 Protocol Suite	176
10.7	RFC 6142: C12.22 Transport Over an IP Network	176
10.8	REST-Based Interfaces to C12.19	177
11	DLMS/COSEM	179
11.1	DLMS Standardization	179
	11.1.1 <i>The DLMS UA</i>	179
	11.1.2 <i>DLMS/COSEM, the Colored Books</i>	179
	11.1.3 <i>DLMS Standardization in IEC</i>	180
11.2	The COSEM Data Model	181
11.3	The Object Identification System (OBIS)	182
11.4	The DLMS/COSEM Interface Classes	184
	11.4.1 <i>Data-Storage ICs</i>	185
	11.4.2 <i>Association ICs</i>	185
	11.4.3 <i>Time- and Event-Bound ICs</i>	186
	11.4.4 <i>Communication Setup Channel Objects</i>	186
11.5	Accessing COSEM Interface Objects	186
	11.5.1 <i>The Application Association Concept</i>	186
	11.5.2 <i>The DLMS/COSEM Communication Framework</i>	187
	11.5.3 <i>The Data Communication Services of COSEM Application Layer</i>	189
11.6	End-to-End Security in the DLMS/COSEM Approach	191
	11.6.1 <i>Access Control Security</i>	191
	11.6.2 <i>Data-Transport Security</i>	192

Part IV THE NEXT GENERATION: IP-BASED PROTOCOLS

12	6LoWPAN and RPL	195
12.1	Overview	195
12.2	What is 6LoWPAN? 6LoWPAN and RPL Standardization	195
12.3	Overview of the 6LoWPAN Adaptation Layer	196
	12.3.1 <i>Mesh Addressing Header</i>	197
	12.3.2 <i>Fragment Header</i>	198
	12.3.3 <i>IPv6 Compression Header</i>	198
12.4	Context-Based Compression: IPHC	200
12.5	RPL	202
	12.5.1 <i>RPL Control Messages</i>	204
	12.5.2 <i>Construction of the DODAG and Upward Routes</i>	204
12.6	Downward Routes, Multicast Membership	206
12.7	Packet Routing	207
	12.7.1 <i>RPL Security</i>	208

13	ZigBee Smart Energy 2.0	209
13.1	REST Overview	209
13.1.1	<i>Uniform Interfaces, REST Resources and Resource Identifiers</i>	209
13.1.2	<i>REST Verbs</i>	210
13.1.3	<i>Other REST Constraints, and What is REST After All?</i>	211
13.2	ZigBee SEP 2.0 Overview	212
13.2.1	<i>ZigBee IP</i>	213
13.2.2	<i>ZigBee SEP 2.0 Resources</i>	214
13.3	Function Sets and Device Types	217
13.3.1	<i>Base Function Set</i>	218
13.3.2	<i>Group Enrollment</i>	221
13.3.3	<i>Meter</i>	223
13.3.4	<i>Pricing</i>	223
13.3.5	<i>Demand Response and Load Control Function Set</i>	224
13.3.6	<i>Distributed Energy Resources</i>	227
13.3.7	<i>Plug-In Electric Vehicle</i>	227
13.3.8	<i>Messaging</i>	230
13.3.9	<i>Registration</i>	231
13.4	ZigBee SE 2.0 Security	232
13.4.1	<i>Certificates</i>	232
13.4.2	<i>IP Level Security</i>	232
13.4.3	<i>Application-Level Security</i>	235
14	The ETSI M2M Architecture	237
14.1	Introduction to ETSI TC M2M	237
14.2	System Architecture	238
14.2.1	<i>High-Level Architecture</i>	238
14.2.2	<i>Reference Points</i>	239
14.2.3	<i>Service Capabilities</i>	240
14.3	ETSI M2M SCL Resource Structure	242
14.3.1	<i>SCL Resources</i>	244
14.3.2	<i>Application Resources</i>	244
14.3.3	<i>Access Right Resources</i>	248
14.3.4	<i>Container Resources</i>	248
14.3.5	<i>Group Resources</i>	250
14.3.6	<i>Subscription and Notification Channel Resources</i>	251
14.4	ETSI M2M Interactions Overview	252
14.5	Security in the ETSI M2M Framework	252
14.5.1	<i>Key Management</i>	252
14.5.2	<i>Access Lists</i>	254

14.6	Interworking with Machine Area Networks	255
14.6.1	<i>Mapping M2M Networks to ETSI M2M Resources</i>	256
14.6.2	<i>Interworking with ZigBee 1.0</i>	257
14.6.3	<i>Interworking with C.12</i>	262
14.6.4	<i>Interworking with DLMS/COSEM</i>	264
14.7	Conclusion on ETSI M2M	266

Part V KEY APPLICATIONS OF THE INTERNET OF THINGS

15	The Smart Grid	271
15.1	Introduction	271
15.2	The Marginal Cost of Electricity: Base and Peak Production	272
15.3	Managing Demand: The Next Challenge of Electricity Operators . . . and Why M2M Will Become a Key Technology	273
15.4	Demand Response for Transmission System Operators (TSO)	274
15.4.1	<i>Grid-Balancing Authorities: The TSOs</i>	274
15.4.2	<i>Power Shedding: Who Pays What?</i>	276
15.4.3	<i>Automated Demand Response</i>	277
15.5	Case Study: RTE in France	277
15.5.1	<i>The Public-Network Stabilization and Balancing Mechanisms in France</i>	277
15.5.2	<i>The Bidding Mechanisms of the Tertiary Adjustment Reserve</i>	281
15.5.3	<i>Who Pays for the Network-Balancing Costs?</i>	283
15.6	The Opportunity of Smart Distributed Energy Management	285
15.6.1	<i>Assessing the Potential of Residential and Small-Business Power Shedding (Heating/Cooling Control)</i>	286
15.6.2	<i>Analysis of a Typical Home</i>	287
15.6.3	<i>The Business Case</i>	293
15.7	Demand Response: The Big Picture	300
15.7.1	<i>From Network Balancing to Peak-Demand Suppression</i>	300
15.7.2	<i>Demand Response Beyond Heating Systems</i>	304
15.8	Conclusion: The Business Case of Demand Response and Demand Shifting is a Key Driver for the Deployment of the Internet of Things	305
16	Electric Vehicle Charging	307
16.1	Charging Standards Overview	307
16.1.1	<i>IEC Standards Related to EV Charging</i>	310
16.1.2	<i>SAE Standards</i>	317
16.1.3	<i>J2293</i>	318
16.1.4	<i>CAN – Bus</i>	319

16.1.5	<i>J2847: The New “Recommended Practice” for High-Level Communication Leveraging the ZigBee Smart Energy Profile 2.0</i>	320
16.2	Use Cases	321
16.2.1	<i>Basic Use Cases</i>	321
16.2.2	<i>A More Complex Use Case: Thermal Preconditioning of the Car</i>	323
16.3	Conclusion	324
Appendix A	Normal Aggregate Power Demand of a Set of Identical Heating Systems with Hysteresis	327
Appendix B	Effect of a Decrease of T_{ref}. The Danger of Correlation	329
Appendix C	Changing T_{ref} without Introducing Correlation	331
C.1	Effect of an Increase of T_{ref}	331
Appendix D	Lower Consumption, A Side Benefit of Power Shedding	333
Index		337

List of Acronyms

6LoWPAN	6LoWPAN is the acronym of IPv6 over Low power Wireless Personal Area Networks and the name of a working group in IETF
ACL	Access Control List
ACSE	Association Control Service Element
AER	All Electric Range
AFE	Analog Front End
AIB	Application Layer Information Base
AIS	Application Interworking Specification
AMI	Automatic Metering Infrastructure
ANSI	American National Standards Institute
AODV	Advanced Ad-Hoc On-Demand Distance Vectoring
AP	Application Process
APDU	Application Protocol Data Unit
API	Application Programming Interface
aPoC	Application Point of Contact
APS	Application Support Sublayer
APSD-SE-SAP	Application Support Sublayer Data Entity Service Access Point
APSM-SE-SAP	Application Support Sublayer Management Entity Service Access Point
APSS-SE-SAP	Application Support Sublayer Security Entity Service Access Point
ARIB	Association of Radio Industries and Businesses is a standardization organization in Japan
ASDU	Aps Service Data Unit
ASK	Amplitude-Shift Keying
BbC	KNX Backbone Controller
BCI	Batibus Club International
BEV	Battery Electric Vehicle
BO	Beacon Order
BPSK	Binary Phase Shift Keying
BTT	Broadcast Transaction Table

CAN	Controller Area Network
CAP	Contention Access Period
CBC MAC	CBC Message Authentication Code
CC	Consistency Check
CCA	Clear Channel Assessment
CCM*	Extension of Counter with CBC-MAC Mode of Operation
CD range	Charge Depleting Range
CENELEC	European Committee for Electrotechnical Standardization
CER	Communication Error Rate
CFP	Contention Free Period
CI	Control Information
CNF	M-Bus CONFIRM Message
CRC	Cyclical Redundancy Check
CRL	X.509 Certificate Revocation List
CRUD	Create, Read, Update, Delete
CS mode	Charge Sustaining Mode
CSL	Coordinated Sampled Listening
CSMA	Carrier-Sense, Multiple Access
CSMA/CA	Carrier-Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier-Sense Multiple Access with Collision Detection
D device	ETSI M2M device without local M2M capabilities and interfaced to a gateway via the mId interface
D' device	ETSI M2M device implementing ETSI M2M capabilities and the mId interface to the network domain (does not interface via a gateway)
DA	Device Application
DAG	Direct Acyclic Graph
DAG root	A Node within the DAG that has no outgoing edge
DAO	Destination Advertisement Object
DER	Distinguished Encoding Rule
dIa	ETSI M2M Reference point between an application and ETSI M2M service capabilities
DIB	Data Information Block
DIO	DODAG Information Object
DIS	DODAG Information Solicitation
DLL	Data Link Layer the layer 2 specified in the seven-layer OSI model
DLMS	Device Language Message Specification is a specification for Data exchange for meter reading, tariff and load control
DODAG	Oriented Direct Acyclic Graph
DODAG Version	Specific iteration (“Version”) of a DODAG with a given DODAGID
DODAGID	The identifier of a DODAG Root
DR	Demand Response

DRH	Data Record Header
DSSS	Direct Spread Spectrum Destination
DTSN	Destination Advertisement Trigger Sequence Number
ED	Energy Detection
EFF	Extended Frame Format
EHS	European Home System
EIB	European Installation Bus
EIBA	The European Installation Bus Association
EMC	Electromagnetic Compatibility
EMS	Energy Management System
EN 50065-1	CENELEC standard for Powerline transmission on low-voltage electrical installations in the frequency range 3 to 148,5 kHz
EP	Enforcement Point
EPID	Extended PAN ID
ESI	Energy Services Interface
ESP	Energy Service Portal
eTag	Entity Tag
ETSI	European Telecommunications Standards Institute is an independent, nonprofit, standardization organization in the telecommunications industry
ETSI PLT	The ETSI Powerline working group
EUI	Extended Unique Identifier
EV	Electric Vehicle
EVCC	Electric Vehicle Communication Controller
EVSE	Electric Vehicle Charging Equipment
EXI	Efficient XML Interchange Encoding
FCC	Federal Communications Commission
FFD	Full Function Device
FHSS	Frequency Hopping Spread Spectrum
FLiRS	Frequently Listening Routing Slave
FSK	Frequency-shift keying is a frequency modulation scheme in which digital information is transmitted through discrete frequency changes of a carrier wave
GA	Gateway Application
GBA	Generic Bootstrapping Architecture
GCM	Galois/Counter Mode
GMO	Gateway Management Object
GO	Group Object
GRE	Gestionnaire de réseau de transport
GRIP	Gateway Remote Interface Protocol
HC	Header Compression
HEV	Hybrid Electric Vehicle

HLS	High-Level Security
HomePlug Alliance	The HomePlug Alliance is a group of electronics manufacturers, service providers, and retailers that establishes standards for power line communication
IANA	Internet Assigned Number Authority
I-Band	Industrial Band, see ISM
IC	Interface Class
IEC TC13	International Electrotechnical Commission, Technical Committee 13
IEEE	The Institute of Electrical and Electronics Engineers
IEEE 1901	IEEE 1901 is an IEEE working group developing a global standard for high speed Powerline communications
IEEE 802.15.4	IEEE 802.15.4-2006 is a standard that specifies the physical layer and media access control for low-rate wireless personal area networks
IEEE P1901.2	IEEE 1901.2 is an IEEE working group developing a Powerline communications standard for metering applications
IETF	Internet Engineering Task Force
IHD	In Home Display
IID	Interface Id
IO	Interface Object
IPHA	IP Host Application
IPHC	IP Header Compression
IPSO	Internet Protocol for Smart Objects is a industry alliance promoting Internet of Objects
ISM	Industrial Scientific and Medical
ISO	International Organization for Standardization
ISP	Intersystem Protocol
ITS	Intelligent Transport System
ITU	International Telecommunication Union is the specialized agency of the United Nations which is responsible for information and communication technologies
ITU G.9972	ITU G.9972 (also known as G.cx) is a recommendation developed by ITU-T that specifies a coexistence mechanism for networking transceivers
ITU G.hn	G.hn is the common name for ITU recommendation G.9960, a home network technology standard being developed under the International Telecommunication Union
ITU G.hnem	An ITU project addressing the home networking aspects of energy management
LC	Line Coupler
LDN	Logical Device Name

LLC	Logical Link Control layer
LLN	Low Bitrate and Lossy Network
LLS	Low-Level Security
LN	Logical Name
LonWorks	LonWorks is a networking platform created to control applications The platform is built on a protocol created by Echelon Corporation
LowPAN	Low-power Wireless Personal Area Networks
LQI	Link Quality Information
LRWBS	Low Rate Wide Band Services are emerging services on Powerline transmitting in the 2–4 MHz band
LV-MV	Low Voltage (less than 600 Volts) and Medium Voltage (in the order of magnitude of 20 000 Volts)
M2M	Machine-to-Machine
MAC	Media Access Control
MAS	M2M Authentication Server
MCPS	MAC Common Part Sublayer
MCPS-SAP	MAC Common Part Service Access Point
MDU	Multidwelling Unit
mIa	Reference Point between a M2M application and the M2M Service Capabilities in the Networks and Applications Domain
MIC	Message Integrity Protection Code
mId	Reference point between an M2M Device or M2M Gateway and the M2M Service Capabilities in the Network and Applications Domain
MLDE	MAC Layer Management Entity
MLME-SAP	MAC Layer Management Entity Service Access Point
MP2P	Multipoint To Point Traffic
MSBF	M2M Service Bootstrap Function
MSP	Manufacturer Specific Profile
MTU	Maximum Transmission Unit
NA	Network Application
NAN	Neighborhood Area Network
NAPT	Network Address and Port Translation
NIB	Network Information Base
NIF	Node Information Frame
NIP	Network Interworking Proxy
NIST	National Institute of Standards and Technology is a measurement standards laboratory in USA
NLDE-SAP	Network Layer Data Entity Service Access Point
NLME	Network Layer Management Entity
NLME-SAP	Network Layer Management Entity Service Access Point
NLSE-SAP	Network Layer Security Entity Service Access Point

NREL	National Renewable Energy Laboratory
NRZ	Nonreturn to Zero
NUD	Neighbor Unreachability Detection
OBIS	Object Identification System
OCP	Objective Code Point
OF	Objective Function
OFDM	Orthogonal Frequency-Division Multiplexing
OOK	On-off keying the simplest form of modulation that represents digital data as the presence or absence of a carrier wave
O-QPSK	Offset-Quadrature Phase-Shift Keying
OSI	Open Systems Interconnections
OTA	Over-the-Air
OUI	Organizationally Unique Identifier
P2MP	Point to Multipoint Traffic
PAA	PANA Authentication Agent
PaC	PANA Client
PAN	Personal Area Network
PAN ID	Personal Area Network Identifier
PANA	Protocol for Carrying Authentication for Network Access
PCT	Programmable Communicating Thermostat
PEV	Plug-in Electric Vehicle
PHEV	Plug-in Hybrid Electric Vehicle
PHR	Physical Header
PHY	Physical Layer
PIB	PAN Information Base
PIO	Prefix Information Option
PLC	Powerline Communication
PLT	Powerline Technology
PN	Parent Node
PoC	Point of Contact
PRE	PANA Relay Element
PRIME	Powerline Intelligent Metering Evolution
PSDU	Physical Service Data Unit
PSEM	Protocol Specification for Electric Metering
PSSS	Parallel Spread Spectrum modulation
PWM	Pulse Width Modulation
Rank	A node's individual position relative to other nodes with respect to a DODAG root
REQ	M-Bus REQUEST message
REST	Representational State Transfer
RFD	Reduced Function Device
RIT	Receiver-Initiated Transmission

ROLL	Routing over Low-power and Lossy network
RPF	Reverse Power Flow
RPL	RPL IPv6 Routing Protocol over Low-power and Lossy Networks
RPL Instance	A set of one or more DODAGs that share a RPLInstanceID
RPLInstanceID	A unique identifier within a RPL LLN. DODAGs with the same RPLInstanceID share the same Objective Function
RSP	M-Bus RESPOND Message
RTE	Réseau Transport Electricité
RTU	Remote Terminal Unit
RZtime	Rendezvous Time
SA	Secure Association
SAP	Service Access Point
S-Band	Scientific Band, <i>see</i> ISM
SCDE	Secured Connection Protocol
SCL	Service Capability Layer
SCME	SCoP Management Entity
SCoP	SCoP Data Entity
SCPT	Standard Configuration Property Type
SCSS	SCoP Security Service
SDP	SECC Discovery Protocol
SDU	Service Data Unit
SECC	Supply Equipment Communication Controller
SFD	Start Frame Delimiter
SHR	Synchronous Header
SKKE	Symmetric-Key Key Exchange
SLAAC	IPv6 Stateless Address Autoconfiguration
SN	Short Name
SND	M-Bus SEND Message
SNVT	Standard Network Variable Type
SoC	System on Chip
SUN	Smart Utility Network
TDMA	Time division multiple access is a channel access method for shared medium networks
TL	Transport Layer
TLS	Transport Layer Security
ToU	Time of Use
TP1	KNX Twisted Pair Physical Media
TSCH	Time-Synchronized Channel Hopping
TSO	Transmission System Operator
UC	Upgrade Client
UID	Unique Node Identifier
U-NII	Unlicensed National Information Infrastructure

UNVT	User Network Variable Type
US	Upgrade Server
V2GTP	Vehicle to Grid Transfer Protocol
VIB	Value Information Block
VIF	Value Information field, see M-Bus
WADL	Web Application Description Language
xAE	Application Enablement M2M Service Capability
xBC	Compensation Broker M2M Service Capability
XCAP	Extensible Markup Language (XML) Configuration Access Protocol (RFC 4825)
xCS	Communication Selection M2M Service Capability
xHDR	History and Data Retention M2M Service Capability
xIP	Interworking Proxy M2M Service Capability
xRAR	Reachability, Addressing and Repository M2M Service Capability
xREM	Remote Entity Management M2M Service Capability
xSEC	Security M2M Service Capability
xTM	Transaction Management M2M Service Capability
xTOE	Telco Operator Exposure M2M Service Capability
ZBD	ZigBee Bridge Device
ZC	ZigBee Coordinator
ZCL	ZigBee Cluster Library
ZCP	ZigBee Compliant Platform
ZDO	ZigBee Device Object
ZDP	ZigBee Device Profile
ZED	ZigBee End Device
Zero-crossing	In alternating current, the zero-crossing is the instantaneous point at which there is no voltage present
ZGD	ZigBee Gateway Device
ZigBee Alliance	ZigBee Alliance is a group of companies that maintain and publish the ZigBee standard
ZIPT	ZigBee IP Tunneling Protocol
ZR	ZigBee Router
ZSE	ZigBee Smart Energy

Introduction

Innovation rarely comes where it is expected. Many governments have been spending billions to increase the Internet bandwidth available to end users . . . only to discover that there are only a limited number of HD movies one can watch at a given time. In fact, there are also a limited number of human beings on Earth.

The Internet is about to bring us another ten years of surprises, as it morphs into the “Internet of Things” (IoT). Your mobile phone and your PC are already connected to the Internet, maybe even your car GPS too. In the coming years your car, office, house and all the appliances it contains, including your electricity, gas and water meters, street lights, sprinklers, bathroom scales, tensiometers and even walls¹ will be connected to the IoT. Tomorrow, several improvements will be made to these appliances such as not heating your house if hot weather is forecast, watering your garden automatically only if it doesn’t rain, getting assistance immediately on the road, and so on. These improvements will facilitate our lives and utilize natural resources more efficiently.

Why is this happening now? As always, there is a combination of small innovations that, together, have reached a critical mass:

- Fieldbus technologies, using proprietary protocols and standards (LON, KNX, DALI, CAN, ModBus, M-Bus, ZigBee, Zwave . . .), have explored many vertical domains. Gradually, these domains have started to overlap as use cases expanded to more complex situations, and protocols have emerged to facilitate interoperability (e.g., BACnet). But in many ways, current fieldbus deployments continue to use parallel networks that do not collaborate. The need for a common networking technology that would run over any physical layer, like IP, has become very clear.
- Despite the need for a layer 2 independent networking technology for fieldbuses, IP was not considered as a possible candidate for low-bitrate physical layers typically used in fieldbus networks, due to its large overheads. But the wait is now over: with 6LoWPAN not only has IP technology found its way onto low-bitrate networks but – surprise, surprise – it is IPv6 ! As an additional bonus, the technology comes with a state-of-the-art, standardized IP level mesh networking protocol, which makes multiply

¹ Sensors for structural monitoring.

mesh networking a reality: finally different layer 2 fieldbus technologies can collaborate and form larger networks.

- Today, local fieldbus networks optimize the HVAC² regulation in your office and perhaps your home, with sophisticated algorithms. The energy-efficiency regulation for new building construction has created a need for even more sophisticated algorithms, like predictive regulation that takes into account weather forecasts or load shifting that incorporates the CO₂ content of electricity. In many automation sectors, the current state-of-the-art tool requires the local fieldbus to collaborate with hosted centralized applications and data sources. The technology required to enable this progressed in steps: oBix introduced the concept of a uniform (REST) interface to sensor networks, ETSI M2M added the management of security and additional improvements required in large-scale public networks.

The industry was only missing a really, really compelling business case to trigger the enormous amount of R&D that will be required to integrate all these technologies and build a bulletproof Internet of Things.

This business case is coming from the energy sector:

- The accelerated introduction of renewable-energy sources in the overall electricity production park brings an increasing degree of randomness to the traditionally deterministic supply side.
- In parallel, the mass introduction of rechargeable electric and hybrid vehicles is making the demand side more complex: EVs are roaming objects that will need to authenticate to the network, and will require admission control protocols.

The current credo of electricity operators “demand is unpredictable, and our expertise is to adapt production to demand”, is about to be reversed into “production is unpredictable, and our expertise is to adapt demand to production”.

As the rules of the game change, the key assets of an energy operator will no longer be the means of production, but the next-generation communication network and information system, which they still need to build entirely, creating an enormous market for mission-critical M2M technology. This dramatic change of how electricity will be distributed prefigures the more general evolution of the Internet towards the Internet of Things, where telecom operators and network-based application developers will have an increasing impact on our everyday lives, including the things that we touch and use.

This book targets an audience of engineers who are involved or want to get involved in large-scale automation and smart-grid projects and need to get a feel for the “big picture”.

Many such projects will involve interfaces with existing systems. We included detailed overviews of many legacy fieldbus and automation technologies: BACnet, CAN, LON, M-Bus/wMBUS, ModBus, LON, KNX, ZigBee, Z-Wave, as well as C.12 and

² Heating, ventilation and air conditioning.

DLMS/COSEM metering standards. We also cover in detail two common fieldbus physical layers: 802.15.4 and PLC.

This book will not make you an expert on any of these technologies, but provides enough information to understand what each technology can or cannot do, and the fast-track descriptions should make it much easier to learn the details by yourself.

The future of fieldbus protocols is IP: we introduce 6LoWPAN and RPL, as well as the first automation protocol to have been explicitly designed for 6LoWPAN networks: ZigBee SE 2.0. We also provide an introduction to the emerging ETSI M2M standard, which is the much-awaited missing piece for service providers willing to provide a general-purpose public M2M infrastructure, shared by all applications.

I would like to thank Paul Bertrand, the inventor of the lowest-power PLC fieldbus technology to date (WPC) and designer of the first port of 6LoWPAN to PLC for accepting to write – guess what – the Powerline Communications chapter of this book. I am also grateful for the C.12 and DLMS chapters that were provided by Jean-Marc Ballot (Alcatel), and required a lot of documentation work.

Despite my efforts, there are probably quite a few errors remaining in the text, but there would have been many more without the help of the expert reviewers of this book: Cedric Chauvenet for 6LoWPAN/RPL, Mathieu Pouillot for ZigBee, Juan Perez (EPEX) for the smart-grid section, François Collet (Renault) for EV charging, Alexandre Ouimet-Storrs for his insights on energy trading, and the companies who provided internal documentation or reviews: Echelon for LON (with special thanks to Bob Dolin, Jeff Lund, Larry Colton and Mark Ossel), and Sigma Designs for Z-Wave. I am also grateful to Benoit Guennec and Baptiste Vial (Connected Object), who supplied me with the temperature and consumption profiles of their homes and shared their field experience with Z-Wave. Please let me know of remaining errors, so that we can improve the next edition of this book, at olivier.hersent@actility.com.

Gathering and reading the documentation for this book has been an amazing experience discovering new horizons and perspectives. I hope you will enjoy reading this book as much as I enjoyed writing it.

Olivier Hersent

Part One

M2M Area Network Physical Layers

1

IEEE 802.15.4

1.1 The IEEE 802 Committee Family of Protocols

The Institute of Electrical and Electronics Engineers (IEEE) committee 802 defines physical and data link technologies. The IEEE decomposes the OSI link layer into two sublayers:

- The media-access control (MAC) layer, sits immediately on top of the physical layer (PHY), and implements the methods used to access the network, typically the carrier-sense multiple access with collision detection (CSMA/CD) used by Ethernet and the carrier-sense multiple access with collision avoidance (CSMA/CA) used by IEEE wireless protocols.
- The logical link control layer (LLC), which formats the data frames sent over the communication channel through the MAC and PHY layers. IEEE 802.2 defines a frame format that is independent of the underlying MAC and PHY layers, and presents a uniform interface to the upper layers.

Since 1980, IEEE has defined many popular MAC and PHY standards (Figure 1.1 shows only the wireless standards), which all use 802.2 as the LLC layer.

802.15.4 was defined by IEEE 802.15 task group 4/4b (<http://ieee802.org/15/pub/TG4b.html>). The standard was first published in 2003, then revised in 2006. The 2006 version introduces improved data rates for the 868 and 900 MHz physical layers (250 kbps, up from 20 and 40 kbps, respectively), and can be downloaded at no charge from the IEEE at <http://standards.ieee.org/getieee802/download/802.15.4-2006.pdf>

1.2 The Physical Layer

The design of 802.15.4 takes into account the spectrum allocation rules of the United States (FCC CFR 47), Canada (GL 36), Europe (ETSI EN 300 328-1, 328-2, 220-1) and

MAC layer		BAND
802.11	WiFi	802.11, 802.11b, 802.11g, 802.11n : ISM 802.11a : U-NII
802.15.1	Bluetooth	ISM 2.4 GHz
802.15.4	ZigBee, SLOWPAN	ISM 2.4 GHz worldwide ISM 902–928 MHz USA 868.3 MHz European countries 802.15.4a: 3.1–10.6 GHz
802.16	Wireless Metropolitan Access Networks Broadband Wireless Access (BWA) WiMax	802.16 : 10–66 GHz 802.16a: 2–11 GHz 802.16e: 2–11 GHz for fixed/2–6 GHz for mobile

Figure 1.1 IEEE-defined MAC layers.

Japan (ARIB STD T66). In the United States, the management and allocation of frequency bands is the responsibility of the Federal Communications Commission (FCC). The FCC has allocated frequencies for industrial scientific and medical (ISM) applications, which do not require a license for all stations emitting less than 1 W. In addition, for low-power applications, the FCC has allocated the Unlicensed National Information Infrastructure (U-NII) band. Figure 1.2 lists the frequencies and maximum transmission power for each band.

IEEE 802.15.4 can use:

- The 2.4 GHz ISM band (S-band) worldwide, providing a data rate of 250 kbps (O-QPSK modulation) and 15 channels (numbered 11–26);
- The 902–928 MHz ISM band (I-band) in the US, providing a data rate of 40 kbps (BPSK modulation), 250 kbps (BPSK+O-QPSK or ASK modulation) or 250 kbps (ASK modulation) and ten channels (numbered 1–10)
- The 868–868.6 MHz frequency band in Europe, providing a data rate of 20 kbps (BPSK modulation), 100 kbps (BPSK+O-QPSK modulation) or 250 kbps (PSSS: BPSK+ASK

FCC band	Maximum transmit power	Frequencies
Industrial Band	<1W	902 MHz–928 MHz
Scientific Band	<1W	2.4 GHz–2.48 GHz
Medical Band	<1W	5.725 GHz–5.85 GHz
U-NII	<40 mW	5.15 GHz–5.25 GHz
	<200 mW	5.25 GHz–5.35 GHz
	<800 mW	5.725 GHz–5.82 GHz

Figure 1.2 FCC ISM and U-NII bands.

modulation), and a single channel (numbered 0 for BPSK or O-QPSK modulations, and 1 for ASK modulation).

In practice, most implementations today use the 2.4 GHz frequency band. This may change in the future as the IP500 alliance (www.ip500.de) is trying to promote applications on top of 6LoWPAN and 802.15.4 sub-GHz frequencies and 802.15.4g introduces more sub-GHz physical layer options. More recently, a new physical layer has been designed for ultrawide band (3.1 to 10.6 GHz).

Overview of O-QPSK Modulation at 2.4 GHz

The data to be transmitted is grouped in blocks of 4 bits. Each such block is mapped to one of 16 different *symbols*. The symbol is then converted to a 32-bit chip sequence (a pseudorandom sequence defined by 802.15.4 for each symbol). The even bits are transmitted by modulating the inphase (I) carrier, and the odd bits are transmitted by modulating the quadrature phase (Q) carrier (Figure 1.3). Each chip is modulated as a half-sine pulse. The transmitted chip rate is 2 Mchip/s, corresponding to a symbol rate 32 times slower, and a user data bitrate of 250 kbps. The sum of the I and Q signals is then transposed to the 2.4 GHz carrier frequency.

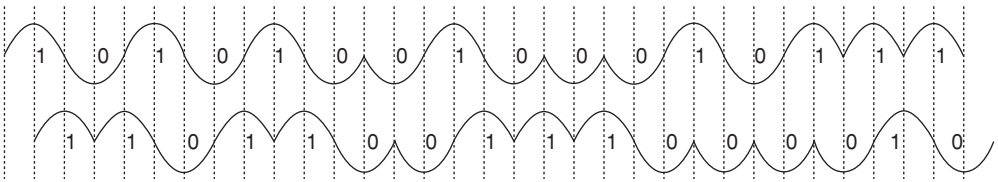


Figure 1.3 O-QPSK I and Q components.

802.15.4 uses a 32-bit encoding when it needs to refer to a specific frequency band, modulation, and channel. The first 5 bits encode a page number, and the remaining 27 bits are used as channel number flags within the page. The mapping of page and channel number to the frequency band, modulation and center frequency is shown in Figure 1.4.

1.2.1 Interferences with Other Technologies

Because the scientific band (2.4–2.48 GHz) is also unlicensed in most countries, this frequency band is used by many wireless networking standards, among which are WiFi (802.11, 802.11b, 802.11g, 802.11n), 802.15.4, and other devices such as cordless phones and microwave ovens.

Frequency band	Modulation	Page number	Channel number and center frequency
2.4 GHz	0-QPSK	0	11 : 2405 MHz
			12 : 2410 MHz
			13 : 2415 MHz
			14 : 2420 MHz
			15 : 2425 MHz
			16 : 2430 MHz
			17 : 2435 MHz
			18 : 2440 MHz
			19 : 2445 MHz
			20 : 2450 MHz
			21 : 2455 MHz
			22 : 2460 MHz
			23 : 2465 MHz
24 : 2470 MHz			
25 : 2475 MHz			
26 : 2480 MHz			
915 MHz	BPSK	0	1 : 906 MHz
	BPSK+ ASK	1	2 : 908 MHz
	BPSK+ 0-QPSK	2	3 : 910 MHz
			4 : 912 MHz
			5 : 914 MHz
			6 : 916 MHz
			7 : 918 MHz
			8 : 920 MHz
			9 : 922 MHz
			10 : 924 MHz
868 MHz	BPSK	0	0 : 868.3 MHz
	BPSK+ ASK	1	1 : 868.3 MHz
	BPSK+ 0-QPSK	2	0 : 868.3 MHz

Figure 1.4 802.15.4 frequency bands, modulations and channels.

1.2.1.1 FHSS Wireless Standards

The 802.11 physical layer uses frequency hopping spread spectrum (FHSS) and direct spread spectrum modulation. Bluetooth (802.15.1) uses FHSS in the ISM band.

The FHSS technology divides the ISM band into 79 channels of 1 MHz (Figure 1.5). The FCC requires that a transmitter should not use any channel more than 400 ms at a time (dwell time), and should try to use at least 75 channels (but this may not always be possible if some channels are too noisy).

FHSS Channel	Frequency (GHz)		
2	2.401–2.402		
3	2.402–2.403		
4	2.403–2.404		
...			
80	2.479–2.480		

Figure 1.5 FHSS channels defined by the FCC in the S-Band.

1.2.1.2 DSSS Wireless Standards

802.11b and 802.11g use only direct spread spectrum (DSSS). 11 DSSS channels have been defined, each of 16 MHz bandwidth, with center frequencies of adjacent channels separated by 5 MHz. Only 3 channels do not overlap (outlined in bold font in Figure 1.6): these channels should be used in order to minimize interference issues in adjacent deployments (3 channels are sufficient for a bidirectional deployment, however in tridimensional deployments, for example, in a building, more channels would be required).

1.2.2 *Choice of a 802.15.4 Communication Channel, Energy Detection, Link Quality Information*

In practice, only the 2.4 GHz frequency band is commonly used by the network and applications layers on top of 802.15.4, typically ZigBee and 6LoWPAN. The transmission power is adjustable from a minimum of 0.5 mW (specified in the 802.15.4 standard) to a maximum of 1 W (ISM band maximum). For obvious reasons, on links involving a battery-operated device, the transmission power should be minimized. A transmission power of 1 mW provides a theoretical outdoor range of about 300 m (100 m indoors).

DSSS channel	Frequency (GHz)
1	2.404–(2.412)–2.420
2	2.409–(2.417)–2.425
3	2.414–(2.422)–2.430
4	2.419–(2.427)–2.435
5	2.424–(2.432)–2.440
6	2.429–(2.437)–2.445
7	2.434–(2.442)–2.450
8	2.439–(2.447)–2.455
9	2.444–(2.452)–2.460
10	2.449–(2.457)–2.465
11	2.456–(2.462)–2.470

Figure 1.6 DSSS channels used by 802.11b.

Synchronous header (SHR)		Physical header (PHR)		Physical Service Data Unit
Preamble	SFD 111100101	Frame length (7 bits)	Ibit (reserved)	0 to 127 bytes

Figure 1.7 802.15.4 physical layer frame.

802.15.4 does not use frequency hopping (a technique that consumes much more energy), therefore the choice of the communication channel is important. Interference with FHSS technologies is only sporadic since the FHSS source never stays longer than 400 ms on a given frequency. In order to minimize interference with DSSS systems such as Wi-Fi (802.11b/g) set to operate on the three nonoverlapping channels 1, 6 and 11, it is usually recommended to operate 802.15.4 applications on channels 15, 20, 25 and 26 that fall between Wi-Fi channels 1, 6 and 11.

However, the 802.15.4 physical layer provides an energy detection (ED) feature that enables applications to request an assessment of each channel's energy level. Based on the results, a 802.15.4 network coordinator can make an optimal decision for the selection of a channel.

For each received packet, the 802.15.4 physical layer also provides link quality information (LQI) to the network and application layers (the calculation method for the LQI is proprietary and specific to each vendor). Based on this indication and the number of retransmissions and lost packets, transmitters may decide to use a higher transmission power, and some applications for example, ZigBee Pro provide mechanisms to dynamically change the 802.15.4 channel in case the selected one becomes too jammed, however, such a channel switch should remain exceptional.

1.2.3 Sending a Data Frame

802.15.4 uses carrier-sense multiple access with collision avoidance (CSMA/CA): prior to sending a data frame, higher layers are first required to ask the physical layer to perform a clear channel assessment (CCA). The exact meaning of "channel clear" is configurable: it can correspond to an energy threshold on the channel regardless of the modulation (mode 1), or detection of 802.15.4 modulation (mode 2) or a combination of both (energy above threshold *and* 802.15.4 modulation: mode 3).

After a random back-off period designed to avoid any synchronization of transmitters, the device checks that the channel is still free and transmits a data frame. Each frame is transmitted using a 30- to 40-bit preamble followed by a start frame delimiter (SFD), and a minimal physical layer header composed only of a 7 bits frame length (Figure 1.7).

1.3 The Media-Access Control Layer

802.15.4 distinguishes the part of the MAC layer responsible for data transfer (the MAC common part sublayer or MCPS), and the part responsible for management of the MAC layer itself (the Mac layer management entity or MLME).

The MLME contains the configuration and state parameters for the MAC layer, such as the 64-bit IEEE address and 16-bit short address for the node, how many times to retry accessing the network in case of a collision (typically 4 times, maximum 5 times), how long to wait for an acknowledgment (typically 54 symbol duration units, maximum 120), or how many times to resend a packet that has not been acknowledged (0–7).

1.3.1 802.15.4 Reduced Function and Full Function Devices, Coordinators, and the PAN Coordinator

802.15.4 networks are composed of several device types:

- 802.15.4 networks are setup by a *PAN coordinator* node, sometimes simply called the coordinator. There is a single PAN coordinator for each network identified by its PAN ID. The PAN coordinator is responsible for scanning the network and selecting the optimal RF channel, and for selecting the 16 bits PAN ID (personal area network identifier) for the network. Other 802.15.4 nodes must send an association request for this PAN ID to the PAN coordinator in order to become part of the 802.15.4 network.
- *Full Function Devices* (FFD), also called coordinators: these devices are capable of relaying messages to other FFDs, including the PAN coordinator. The first coordinator to send a beacon frame becomes the PAN coordinator, then devices join the PAN coordinator as their parent, and among those devices the FFDs also begin to transmit a periodic beacon (if the network uses the beacon-enabled access method, see below), or to respond to beacon requests. At this stage more devices may be able to join the network, using the PAN coordinator or any FFD as their parent.
- *Reduced Function Devices* (RFD) cannot route messages. Usually their receivers are switched off except during transmission. They can be attached to the network only as leaf nodes.

Two alternative topology models can be used within each network, each with its corresponding data-transfer method:

- The *star topology*: data transfers are possible only between the PAN coordinator and the devices.
- The *peer to peer topology*: data transfers can occur between any two devices. However, this is simple only in networks comprising only permanently listening devices. Peer to peer communication between devices that can enter sleep mode requires synchronization, which is not currently addressed by the 802.15.4 standard.

Each network, identified by its PAN ID, is called a *cluster*. A 802.15.4 network can be formed of multiple clusters (each having its own PAN ID) in a tree configuration: the root PAN coordinator instructs one of the FFD to become the coordinator of an adjacent PAN.

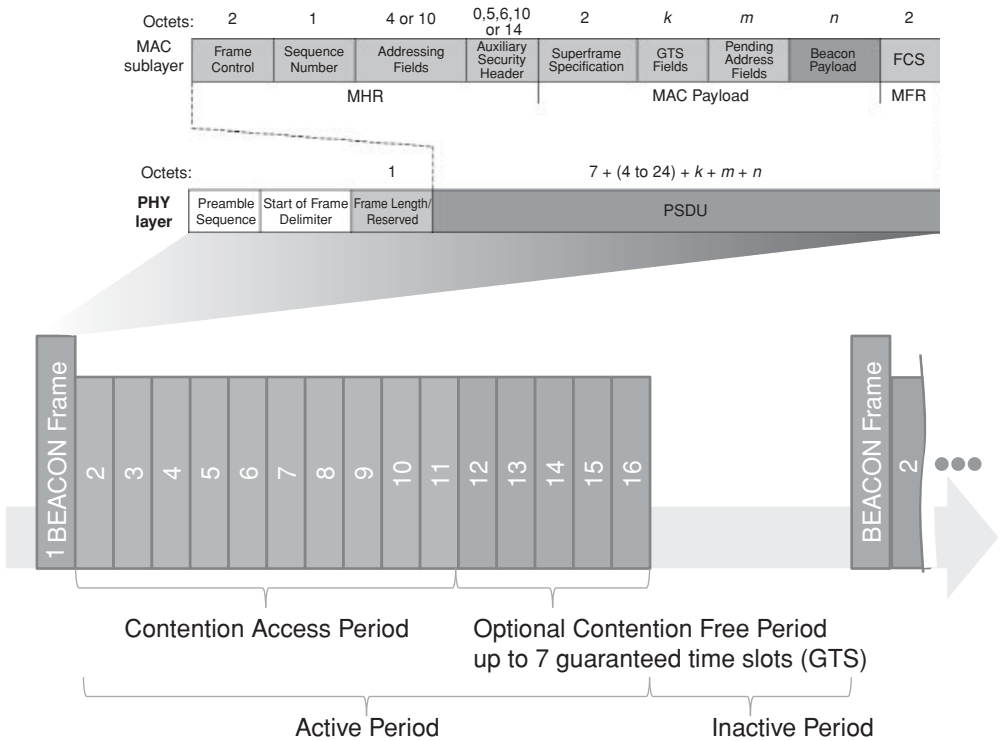


Figure 1.8 802.15.4 Superframe structure.

Each child PAN coordinator may also instruct a FFD to become a coordinator for another PAN, and so on.

The MAC layer specified by 802.15.4 defines two access control methods for the network:

- The *beacon-enabled access method* (or slotted CSMA/CA). When this mode is selected, the PAN coordinator periodically broadcasts a *superframe*, composed of a starting and ending beacon frame, 15 time slots, and an optional inactive period during which the coordinator may enter a low-power mode (Figure 1.8). The first time slots define the contention access period (CAP), during which the other nodes should attempt to transmit using CSMA/CA. The last N ($N \leq 7$) time slots form the optional contention free period (CFP), for use by nodes requiring deterministic network access or guaranteed bandwidth.

The beacon frame starts by the general MAC layer frame control field (see Figures 1.8 and 1.9), then includes the source PAN ID, a list of addresses for which the coordinator has pending data, and provides superframe settings parameters. Devices willing to send data to a coordinator first listen to the superframe beacon, then synchronize to the

	Bytes	
Frame Control Field	2	000-----: Beacon frame 001-----: Data Frame 010-----: Ack Frame 011-----: Command frame ---1-----: Security enabled at MAC layer ----1-----: Frame pending -----1-----: Ack request -----1-----: PAN ID compression (source PAN ID omitted, same as destination) -----XXX-----: reserved -----XX-----: Destination address mode 00 : PAN ID and destination not present (indirect addressing) 01 : reserved 10 : short 16-bit addresses 11 : extended 64-bit addresses -----XX--: Frame version (00 : 2003, 01 : 2006) -----XX: Source address mode
Sequence number	1	
Destination PAN ID	0 or 2	
Destination address	0 or 2 or 8	
Source PAN ID	0 or 2	
Source address	0 or 2 or 8	
Auxiliary security	variable	Contains security control, Frame counter, Key identifier fields
Payload	variable	
FCS	2	CRC 16 frame check sequence

Figure 1.9 802.15.4 MAC layer frame format.

superframe and transmit data either during the CAP using CSMA/CA, or during the CFP. Devices for which the coordinator has pending data should request it from the coordinator using a MAC data request command (see Figure 1.10).

When multiple coordinators transmit beacons, the active periods of the super frames should not overlap (a configuration parameter, *StartTime*, ensures that this is the case).

- The *nonbeacon-enabled access method* (unslotted CSMA/CA). This is the mode used by ZigBee and 6LoWPAN. All nodes access the network using CSMA/CA. The coordinator provides a beacon only when requested by a node, and sets the beaconorder (BO) parameter to 15 to indicate use of the nonbeacon-enabled access method. Nodes (including the coordinator) request a beacon during the *active scan* procedure, when

01	Association request
02	Association response
03	Disassociation notification
04	Data request
05	PAN ID conflict notification
06	Orphan notification
07	Beacon request
08	Coordinator realignment
09	GTS request

Figure 1.10 802.15.4 command identifiers.

trying to identify whether networks are located in the vicinity, and what is their PAN ID.

The devices have no means to know whether the coordinator has pending data for them, and the coordinator cannot simply send the data to devices that are not permanently listening and are not synchronized: therefore, devices should periodically (at an application defined rate), request data from the coordinator.

1.3.2 Association

A node joins the network by sending an association request to the coordinator's address. The association request specifies the PAN ID that the node wishes to join, and a set of capability flags encoded in one octet:

- *Alternate PAN*: 1 if the device has the capability to become a coordinator
- *Device type*: 1 for a full function device (FFD), that is, a device capable of becoming a full function device (e.g., it can perform active network scans).
- *Power source*: 1 if using mains power, 0 when using batteries.
- *Receiver on while transceiver is idle*: set to 1 if the device is always listening.
- *Security capability*: 1 if the device supports sending and receiving secure MAC frames.
- *Allocation address*: set to 1 if the device requests a short address from the coordinator.

In its response, the coordinator assigns a 16-bit short address to the device (or 0xFFFE as a special code meaning that the device can use its 64-bit IEEE MAC address), or specifies the reason for failure (access denied or lack of capacity).

Both the device and the coordinator can issue a disassociation request to end the association.

When a device loses its association with its parent (e.g., it has been moved out of range), it sends orphan notifications (a frame composed of a MAC header, followed by the orphan

command code). If it accepts the reassociation, the coordinator should send a realignment frame that contains the PAN ID, coordinator short address, and the device short address. This frame can also be used by the coordinator to indicate a change of PAN ID.

1.3.3 802.15.4 Addresses

1.3.3.1 EUI-64

Each 802.15.4 node is required to have a unique 64-bit address, called the *extended unique identifier* (EUI-64). In order to ensure global uniqueness, device manufacturers should acquire a 24-bit prefix, the *organizationally unique identifier* (OUI), and for each device, concatenate a unique 40-bit *extension identifier* to form the complete EUI-64.

In the OUI, one bit (M) is reserved to indicate the nature of the EUI-64 address (unicast or multicast), and another bit (L) is reserved to indicate whether the address was assigned locally, or is a universal address (using the OUI/extension scheme described above).

1.3.3.2 16-Bit Short Addresses

Since longer addresses increase the packet size, therefore require more transmission time and more energy, devices can also request a 16-bit short address from the PAN controller.

The special 16-bit address FFFF is used as the MAC broadcast address. The MAC layer of all devices will transmit packets addressed to FFFF to the upper layers.

1.3.4 802.15.4 Frame Format

The MAC layer has its own frame format, which is described in Figure 1.9.

The type of data contained in the payload field is determined from the first 3 bits of the frame control field:

- *Data frames* contain network layer data directly in the payload part of the MAC frame.
- The *Ack frame* format is specific: it contains only a sequence number and frame check sequence, and omits the address and data fields. At the physical layer, Ack frames are transmitted immediately, without waiting for the normal CSMA/CA clear channel assessment and random delays. This is possible because all other CSMA/CA transmissions begin after a minimal delay, leaving room for any potential Ack.
- The payload for *command frames* begins with a command identifier (Figure 1.10), followed by a command specific payload.

In its desire to reduce frame sizes to a minimum, 802.15.4 did not include an upper-layer protocol indicator field (such as Ethertype in Ethernet). This now causes problems, since both ZigBee and 6LoWPAN can be such upper layers.

1.3.5 Security

802.15.4 is designed to facilitate the use of symmetric key cryptography in order to provide data confidentiality, data authenticity and replay protection. It is possible to use a specific key for each pair of devices (link key), or a common key for a group of devices. However, the mechanisms used to synchronize and exchange keys are not defined in the standard, and left to the applications.

The degree of frame protection can be adjusted on a frame per frame basis. In addition, secure frames can be routed by devices that do not support security.

1.3.5.1 CCM* Transformations

802.15.4 uses a set of security transformations known as CCM* (extension of CCM defined in ANSI X9.63.2001), which takes as input a string “a” to be authenticated using a hash code and a string “m” to be encrypted, and delivers an output ciphertext comprising both the encrypted form of “m” and the CBC message authentication code (CBC MAC) of “a”. Figure 1.11 shows the transformations employed by CCM*, which uses the AES block cipher algorithm E.

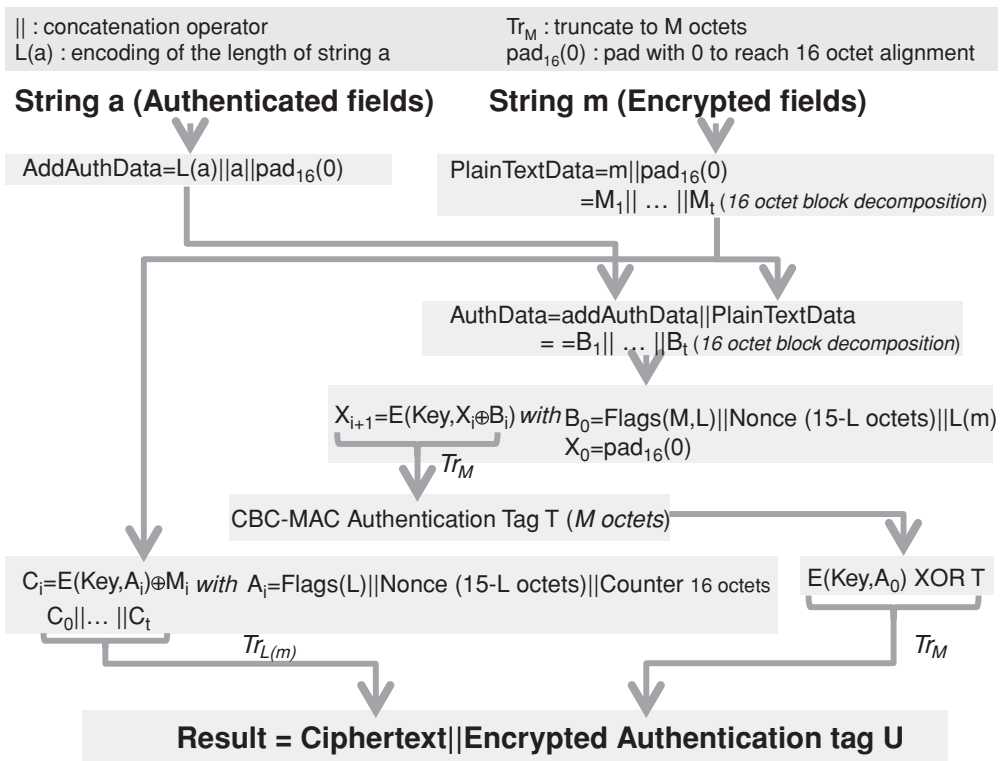


Figure 1.11 Overview of CCM* security transformations.

Security control field	Security attributes	Data confidentiality (data in “m” string)	Data authenticity (data in “a” string)
‘000’	None	OFF	No
‘001’	MIC-32	OFF	MHR, Auxiliary security header, Nonpayload fields, Unsecured payload fields
‘010’	MIC-64	OFF	
‘011’	MIC-128	OFF	
‘100’	Encrypted fields	Unsecured payload fields	No
‘101’	Encr. Fields+MIC-32		MHR, Auxiliary security header, Nonpayload fields
‘110’	Encr. Fields +MIC-64		
‘111’	Encr. Fields +MIC-128		

Figure 1.12 Security control field codes.

In the case of 802.15.4, $L = 2$ octets, and the nonce is a 13-octet field composed of the 8-octet address of the device originating the frame, the 4-octet frame counter, and the one-octet security-level code.

1.3.5.2 The Auxiliary Control Header

The required security parameters are contained in the *auxiliary control header*, which is composed of a security control field (1 octet), a frame counter (4 octets) ensuring protection against replay attacks, and a key identifier field (0/1/5 or 9 octets).

The first 3 bits of the security control field indicate the security mode for this data frame (Figure 1.12), the security mode determines the size of M in the CCM* algorithm (0, 4, 8 or 16 octets), and the data fields included in the “a” and “m” strings used for the computation of the final ciphertext (security attributes). The next 2 bits indicate the key identifier mode (Figure 1.13) and the remaining bits are reserved.

Key identifier mode	Description	Key Identifier field length
‘00’	Key determined implicitly from the originator and recipient of the frame	0
‘01’	Key is determined from the 1-octet Key-index subfield of the Key identifier field, using the MAC layer default Key source	1
‘10’	Key is determined explicitly from the 4-octet Key source subfield, and the 1-octet Key index subfield of the Key identifier field (part of the auxiliary security header)	5
‘11’	Key is determined explicitly from the 8-octet Key source subfield, and the 1-octet Key index subfield of the Key identifier field (part of the auxiliary security header)	9

Figure 1.13 Key identifier mode codes.

1.3.5.3 Key Selection

802.15.4 does not handle distribution of keys: the interface between the MAC layer and the key storage is a *key lookup* function, which provides a lookup string parameter that is used as an index to retrieve the appropriate key.

The lookup material provided depends on the context (see Figure 1.13):

- With implicit key identification (KeyIdMode = “00”), the lookup data is based on the 802.15.4 addresses. The design implies that, in general, the sender indexes its keys according to destinations, and the receiver indexes its keys according to sources.

Addressing mode	Sender lookup data (based on <i>destination</i> addressing mode)	Receiver lookup data (based on <i>source</i> addressing mode)
Implicit	Source PAN short or extended address	Destination PAN short or extended address
Short	Destination PAN and destination node address	Source PAN and destination node address
Long	Destination node 802.15.4 8 octet extended address	Source node 802.15.4 8 octet extended address

- With explicit key identification, the lookup data is composed of a key source identifier, and a key index. The design implies that the key storage is organized in several groups called key sources (one of which is the *macDefaultKeySource*). Each key source comprises several keys identified by an index.

The CCM standard specifies that a given key cannot be employed to encrypt more than 261 blocks, therefore the applications using 802.15.4 should not only assign keys, but also change them periodically.

1.4 Uses of 802.15.4

802.15.4 provides all the MAC and PHY level mechanisms required by higher-level protocols to exchange packets securely, and form a network. It is, however, a very constrained protocol

- It does not provide a fragmentation and reassembly mechanism. As the maximum packet size is 127 bytes (MAC layer frame, see Figure 1.7), and the MAC headers and FCS will take between 6 and 19 octets (Figure 1.9), applications will need to be careful when sending unsecured packets larger than 108 bytes. Most applications will require

security: the security headers add between 7 and 15 bytes of overhead, and the message authentication code between 0 and 16 octets. In the worst case, 77 bytes only are left to the application.

- Bandwidth is also very limited, and much less than the PHY level bitrate of 250 kbit/s. Packets cannot be sent continuously: the PHY layer needs to wait for Acks, and the CSMA/CA has many timers. After taking into account the PHY layer overheads (preamble, framing: about 5%) and MAC layer overheads (between 15 and 40%), applications have only access to a theoretical maximum of about 50 kbit/s, and only when no other devices compete for network access.

With these limitations in mind, 802.15.4 is clearly targeted at sensor and automation applications. Both ZigBee and 6LoWPAN introduce segmentation mechanisms that overcome the issue of small and hard to predict application payload sizes at the MAC layer. An application like ZigBee takes the approach of optimizing the entire protocol stack, up to the application layer for use over such a constrained network. 6LoWPAN optimizes only the IPv6 layer and the routing protocols, expecting developers to make a reasonable use of bandwidth.

1.5 The Future of 802.15.4: 802.15.4e and 802.15.4g

In the last few years, there has been an increased focus on the use of 802.15.4 for mission critical applications, such as smart utility networks (SUN). As a result, several new requirements emerged:

- The need for more modulation options, notably in the sub-GHz space, which is the preferred band for utilities who need long-range radios and good wireless building penetration.
- The need for additional MAC layer options enabling channel hopping, sampled listening and in general integrate recent technologies improving power consumption, resilience to interference, and reliability.

1.5.1 802.15.4e

Given typical sensor networks performance and memory buffers, it is generally considered that in a 1000-node network:

- Preamble sampling low-power receive technology allows one message per node every 100 s;
- Synchronized receive technology allows one message per node every 33 s;
- Scheduled receive technology allows one message per node every 10 s.

Working group 15.4e was formed in 2008 to define a MAC amendment to 802.15.4:2006, which only supported the last mode, and on a stable carrier frequency. The focus of 802.15.4e was initially on the introduction of time-synchronized channel hopping, but in time the scope expanded to incorporate several new technologies in the 802.15.4 MAC layer. 802.15.4e also corrects issues with the 802.15.4:2006 ACK frame (no addressing information, no security, no payload) and defines a new ACK frame similar to a normal data frame except that it has an “ACK” type. The currently defined data payload includes time-correction information for synchronization purposes¹ and optional received quality feedback.

Some of the major new features of 802.15.4e are described below.

1.5.1.1 Coordinated Sampled Listening (CSL)

Sampled listening creates an illusion of “always on” for battery-powered nodes while keeping the idle consumption very low. This technology is commonly used by other technologies, for example, KNX-rf. The idea is that the receiver is switched on periodically (every macCSLperiod, for about 5 ms) but with a very low duty cycle. On the transmission side, this requires senders to use preambles longer than the receiving periodicity of the target, in order to be certain that it will receive the preamble and keep the receiver on for the rest of the packet transmission. For a duty cycle of 0.05% and assuming a 5-ms receive period, the receive periodicity (macCSLperiod) will be 1 s, implying a receive latency of up to 1 s per hop. CSL is the mode of choice if the receive latency needs to be in the order of one second or less.

In 802.15.4e, CSL communication can be used between synchronized nodes (in which case the preamble is much shorter and simply compensates clock drifts), or between unsynchronized nodes in which case a long preamble is used (macCSLMaxPeriod). The latter case occurs mainly for the first communication between nodes and broadcast traffic: the 802.15.4e ACK contains information about the next scheduled receive time of the target node, so the sender can synchronize with the receiver and avoid the long preamble for the next data packet, as illustrated in Figure 1.14.

802.15.4e CSL uses a series of microframes (“chirp packets”, a new frame type introduced in 15.4e) as preamble. The microframes are composed of back-to-back 15.4 packets, and include a rendezvous time (RZtime) and optional channel for the actual data transmission: receivers need to decode only one chirp packet to decide whether the coming data frame is to their intention, and if so can decide to go back to sleep until RZtime and wake up again only to receive the data frame.

CSL supports streaming traffic: a frame-pending bit in the 15.4e header instructs the receiver to continue listening for additional packets.

¹ The value, in units of approximately approximately 0.954 μ s, reports the PDU reception time measured as an offset from the scheduled start time of the current timeslot in the acknowledger’s time base.

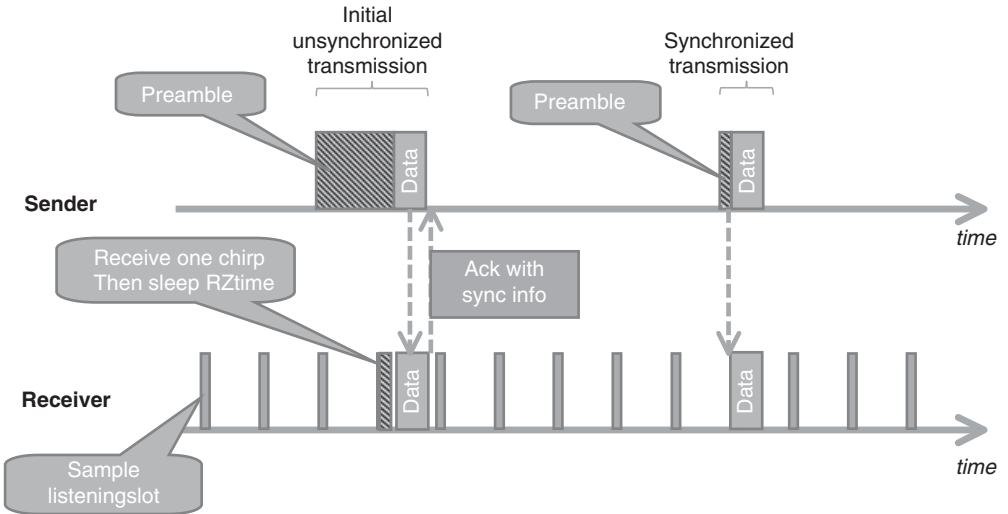


Figure 1.14 Overview of 802.15.4e CSL mode.

1.5.1.2 Receiver-Initiated Transmission (RIT)

The RIT strategy is a simple power-saving strategy that is employed by many existing wireless technologies: the application layer of the receiving node periodically polls a server in the network for pending data. When using the RIT mode, every `macRitPeriod`, the receiver broadcasts a `datarequest` frame and listens for a short amount of time (`macRitDataWaitPeriod`). The receiver can also be turned on for a brief period after sending data.

The downside of this approach is that the perceived receive latency is higher than in the CSL strategy, and multicast is not supported (must be emulated by multiunicast). The polling typically takes about 10 ms, so in order to achieve an idle duty cycle of 0.05% the `macRITPeriod` must be 20 s. RIT is adapted to sensor applications, which can tolerate long receive latency.

1.5.1.3 Time-Synchronized Channel Hopping (TSCH)

Channel hopping is a much-awaited feature of 802.15.4:

- It adds frequency diversity to other diversity methods (coding, modulation, retransmission, mesh routing), and will improve the resilience of 802.15.4 networks to transient spectrum pollution.
- In a multimode network, there are situations in which finding a common usable channel across all nodes is challenging. With channel hopping, each node to node link may use a specific set of frequencies.

Channel hopping is supported in the new ACK frame, which contains synchronization information. In an uncoordinated peer to peer network, the channel hopping penalty is only for the initial transmission, as the sender will need to continue to send “chirp packets” on a given send frequency until it becomes aligned with the receiver frequency. After the first ACK has been received, the sender and the receiver are synchronized and the sender will select the sending frequency according to the channel schedule of the receiver. If all joined nodes are in sync, then synchronizing to a single node is enough to be synchronized to the whole network.

The time-synchronized channel hopping (TSCH) mode defined by 802.15.4e defines the operation model of a 802.15.4e network where all nodes are synchronized. The MAC layer of 802.15.4e nodes can be configured with several “slotframes”, a collection of timeslots repeating in time characterized by the number of time slots in the cyclical pattern, the physical layer channel page supported, and a 27-bit channelMap indicating which frequency channels in the channel page are to be used for channel hopping. Each slotframe can be used to configure multiple “links”, each being characterized by the address list of neighboring devices connected to the link (or 0xffff indicating the link is broadcasting to everyone), a slotframeId, the timeslot within the slot frame that will be used by this link, the channel offset of the link,² the direction (receive, transmit or shared), and whether this link should be reported in advertisement frames. Each network device may participate in one or more slotframes simultaneously, and individual time slots are always aligned across all slotframes.

The FFD nodes in a TSCH mode 802.15.4 network will periodically send advertisement frames that provide the following information: the PAN ID, the channel page supported by the physical layer, the channel map, the frequency-hopping sequence ID (predefined in the standard), the timeslot template ID³ (predefined in the standard), slotframe and link information, and the absolute slot number⁴ of the timeslot being used for transmission of this advertisement frame. The advertisement frames are broadcast over all links configured to transmit this type of frame.

For PANs supporting beacons, synchronization is performed by receiving and decoding the beacon frames. For nonbeacon-enabled networks, the first nodes joining the network synchronize to the PAN coordinator using advertisement frame synchronization data, then additional nodes may synchronize to existing nodes in the network by processing advertisement frames. For networks using the time division multiple access mode, where precise synchronization of the whole network is essential, a new flag “clockSource” in the FFD state supports the selection of clock sources by 802.15.4e nodes without loops. A keep-alive mechanism is introduced to maintain synchronization.

² Logical channel selection in a link is made by taking $(\text{absolute slot number} + \text{channel offset}) \% \text{number of channels}$. The logical channel is then mapped to a physical channel using predefined conventions.

³ The timeslot template defines timing parameters within each timeslot, e.g. $TsTxOffset=2120 \mu s$, $TsMaxPacket=4256 \mu s$, $TsRxAckDelay=800 \mu s$, $TsAckWait=400 \mu s$, $TsMaxAck=2400 \mu s$.

⁴ The total number of timeslots that has elapsed since the start of the network.

1.5.2 802.15.4g

IEEE task group 802.15.4g focuses on the PHY requirements for smart utility networks (SUN).

802.15.4g defines 3 PHY modulation options:

- Multiregional frequency shift keying (MR-FSK): providing typically transmission capacity up to 50 kbps. “Multiregional” means that the standard maps a given channel page to a specific FSK modulation (2GFSK, 4GFSK . . .), frequency and bitrate. The current draft contains multiple variants for each region, implying that generic 802.15.4g radios will have to be extremely flexible.
- Multiregional orthogonal quadrature phase shift keying (O-QPSK): providing typically transmission capacity up to 200 kbps.
- Multiregional orthogonal frequency division multiplexing (OFDM): providing typically transmission capacity up to 500 kbps.

The number of frequency bands also increases to cover most regional markets:

- 2400–2483.5 MHz (Worldwide): all PHYs;
- 902–928 MHz (United States): all PHYs;
- 863–870 MHz (Europe): all PHYs;
- 950–956 MHz (Japan): all PHYs;
- 779–787 MHz (China): O-QPSK and OFDM;
- 1427–1518 MHz (United States, Canada): MR-FSK;
- 450–470 MHz, 896–901 MHz, 901–902 MHz, 928–960 MHz (United States): MR-FSK;
- 400–430 MHz (Japan);
- 470–510 MHz (China): all PHYs;
- 922 MHz (Korea): MR-OFDM.

802.15.4g is particularly interesting in Europe, where 802.15.4:2006 allowed a single channel (868.3 MHz). 802.15.4g now offers multiple channels:

- from 863.125 to 869.725 MHz in steps of 200 kHz (MR-FSK 200 kHz);
- from 863.225 to 869.625 in steps of 400 kHz (MR-FSK 400 kHz);
- from 868.3 to 869.225 MHz in steps of 400 kHz (O-QPSK);
- from 863.225 to 869.625 MHz in steps of 400 kHz (OFDM).

As the number of potential IEEE wireless standards and modulation options increases, the frequency scanning time would become prohibitively long if a coordinator was to scan all possible channels using all possible modulations. To solve this problem and improve coexistence across IEEE standards, 802.15.4g defines a new coex-beacon format, using a standard modulation method that must be supported by all coordinators (the common signaling mode or CSM defined in 802.15.4g).

2

Powerline Communication for M2M Applications

Paul Bertrand
Technology Consultant

2.1 Overview of PLC Technologies

For decades, powerline communication technologies (PLC) have made it possible to use power lines to send and receive data. This “no-new-wire” approach makes PLC one of the best communication technology candidates for the Smart Grid, compared to other wired technologies. On the other hand, as PLC technologies use a media that was not specified for communication, they have faced a number of technical challenges limiting diffusion to niche indoor markets or dedicated ultralow rate applications.

More recently, the booming of modern modulation techniques in integrated silicon made it possible to improve both communication reliability and data rate. Combined with the versatility of emerging protocols such as 6LoWPAN (see the 6LoWPAN chapter), a much larger market is opening for PLC.

Instead of offering here a detailed description of the modulation techniques in use by different vendors/alliances, this can be found for example in [1], this section is more focused on the evolution and comparison of emerging technologies, in the context of the specific requirements of M2M communication.

2.2 PLC Landscape

This section presents an overview of existing powerline technologies and standards. It is not exhaustive and focuses on the most widespread technologies.

2.2.1 *The Historical Period (1950–2000)*

This first period was driven mainly by utilities for outdoor applications at very low frequencies and with an extremely low rate.

The first experiment started in 1950 for remote street lighting. Basically it was one-way On/Off signaling of 10 kW switches at 10 Hz.

In the **mid-1980s** research began on the use of electrical distribution grids to support data transmission, in the [5–500 kHz] frequency band, always for one-way communication.

In **1989**, the ST7536 was the first monolithic half-duplex synchronous FSK modem suitable for applications according to EN 65 065-1 CENELEC and FCC specifications.

In **1997** the first tests for bidirectional data signal transmission over the electrical distribution network were conducted. A specific research effort was started by Ascom (Switzerland) and Norweb (UK).

2.2.2 *After Year 2000: The Maturity of PLC*

In the year 2000 the tremendous development of personal computer and home networking triggered more and more demand for high bitrate transmission technologies. As FSK modulation in the CENELEC band suffers strong data-rate limitations, the communication industry, inspired by the boom of ADSL, decided to study the implementation of OFDM (orthogonal frequency-division multiplexing) in the band above 2 MHz.

2.2.2.1 **High-Rate Modulations**

Homeplug and Homegrid are industrial alliances aiming to publish specifications or white papers on powerline technologies. Since 2000 Homeplug has issued different products standards allowing high data rate communication on existing home electric wires. Usually, these alliances are participating to standard organizations like IEEE and ITU.

Homeplug standards, as all high rate modulation technologies in powerline, use OFDM modulation in the frequency band above 2 MHz.

The typical performance of Homeplug implementations is outlined below:

- HP V1.0: peak PHY rate up to 14 Mb/s.
- HP AV: peak PHY rate up to 200 Mb/s.
- HP AV2 compliant with IEEE 1901: peak PHY rate up to 400–600 Mb/s.
- HP GreenPhy is a low-power profile of IEEE 1901 dedicated to smart grid applications and has a peak PHY rate of 10 Mb/s.

Such impressive performance levels, over such a harsh medium as residential powerline, are only possible through usage of complex signal processing, high power level of PLC

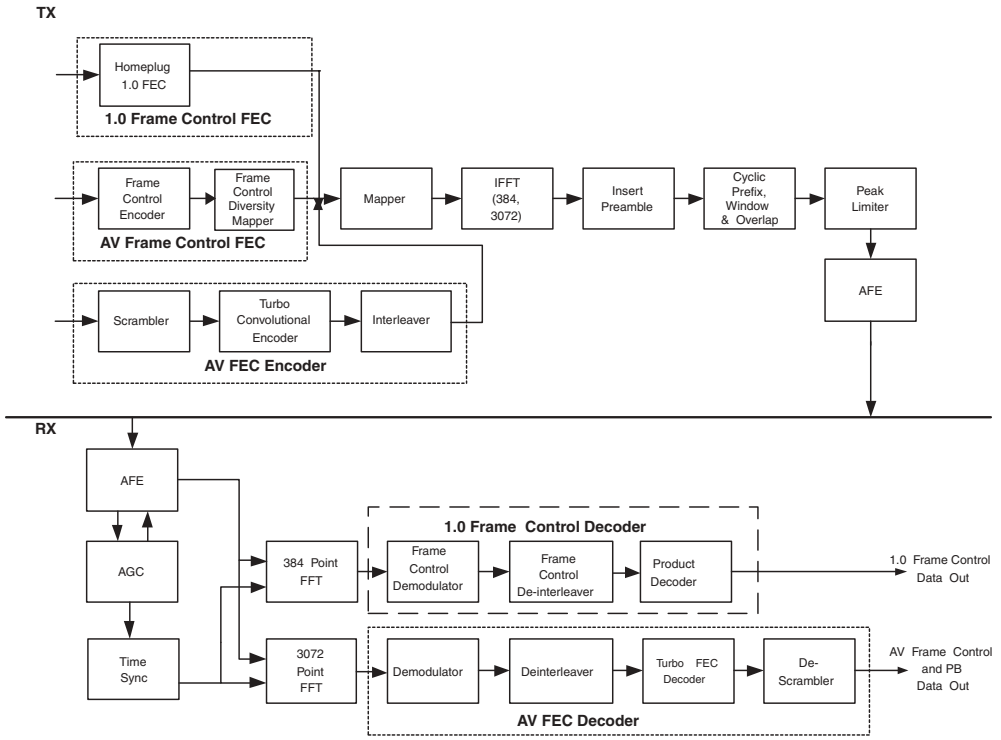


Figure 2.1 HPAV OFDM Transceiver Copyright © 2005, HomePlug® Powerline Alliance, Inc.

emissions and accordingly high power consumption. And still the effective real user data rate usually falls to less than $\frac{1}{4}$ of the PHY maximum data rate in real-life networks.

Figure 2.1, extracted from the HPAV-White-Paper from HomePlug alliance, shows the complexity involved in this type of modem.

2.2.2.2 Low-Rate Modulations

Compared to high-rate OFDM modulation, the complexity of low-rate modems is much lower and a standard DSP can, in most cases, implement the digital processing part of the modem. There are many existing technologies in this field, sometime with a very large installed base.

ISO/IEC 14 908-3 (LonWorks)

This international standard describes an FSK modulation in CENELEC A Band and is used for example in millions of meters in Italy. Refer to Section 4.2 for more details, including the application layer of LonWorks.

IEC 61 334 (S-FSK)

IEC 61 334 is an international standard for low-speed spread shift keying modulation on powerline. It is also known as EN 61 334-5-1:2001. Optional upper layer parts of 61 334 form the DLMS protocol used in metering applications (refer to Section 11.3 for more details on DLMS, including the application layer).

IEC 61 334 is the technology behind G1-PLC, deployed for example, in France as the last-mile communication technology for ERDF (France DSO) first-generation Linky meter.

G3-PLC and PRIME Alliance

G3-PLC and PRIME are two noninteroperable powerline technologies tested by utilities for metering communication. In Europe, both use the CENELEC bands to communicate but they intend also to operate in FCC band up to 500 kHz in the US.

G3-PLC is the powerline communications specification promoted by Maxim and ERDF (French DSO implementing G3 in the second-generation deployment of its “Linky” meter) for smart grid implementations worldwide.

The main specifications of G3 are as follows:

- 10 to 490 kHz operation complies with FCC, CENELEC, and ARIB;
- Coexists with IEC 61 334, IEEE 1901, and ITU G.hn systems;
- Transmission on low- and medium-voltage lines (LV–LV, MV–MV, MV–LV);
- OFDM modulation;
- IEEE 802.15.4-based MAC layer;
- 6LoWPAN adaptation layer supports IPv6 packets;
- AES-128 cryptographic engine;
- Adaptive tone mapping and channel estimation.

In the CENELEC A Band, used by utilities, the OFDM modulation is based on the division of the band into 70 tones. These tones can be modulated in either DBPSK (1 bit per tone), or DQPSK (2 bits per tone).

PRIME is an acronym (PRIME = powerline intelligent metering evolution) for an industry alliance focused on the development of a new open, public and nonproprietary powerline telecom solution for smart grid.

PRIME, created in May 2009, counts more than twenty members at the end of 2010. Principal members are ST Micro, Texas Instruments, ITRON, Landis & Gyr, Iberdrola, Current, ADD and ZIV.

PRIME is based on OFDM multiplexing in CENELEC-A band and is said to reach up to 100 kbps raw data rate. The OFDM signal itself uses 97 (96 data plus one pilot) equally spaced subcarriers with a short cyclic prefix.

Differential modulation schemes are used, together with three possible constellations: DBPSK, DQPSK or D8PSK. Thus, theoretical encoded speeds of around 47 kbps, 94 kbps and 141 kbps (if the cyclic prefix was not considered) could be obtained.

Both G3-PLC and PRIME specifications are the basis of current discussions within IEEE P1901.2 and ITU G.hnem.

Other Low-Rate Modulations

Many other modulations are used for low-rate applications. Most of them use simple modulations like FSK or OOK (X10 for example). Others use different models like spread spectrum or pulse modulation [3].

Table 2.1 shows a comparison between main low-rate technologies in term of standardization, frequency band and modulation.

2.3 Powerline Communication: A Constrained Media

Considering the advantages of powerline technology (no additional wires are required, the network already exists in every home), it is clear that this technology should already be deployed in every home for all types of home-automation applications. However, this is not the case.

We will see that reasons of this limited success are due to four key factors.

2.3.1 Powerline is a Difficult Channel

The channel mandated by CENELEC or FCC for communication is one of the noisiest existing. In homes most appliances now switch in the zero-crossing area of the voltage signal, creating strong spikes. Figures 2.2a and b show some examples of this noise.

Other categories of noise exist: a precise description can be found in [1].

In addition to strong noise levels, the CENELEC band is also known as a strong fading media.

Figure 2.3 shows an illustration of typical fading profiles in the CENELEC A, B, C, D bands.

The notches shown in Figure 2.2 are not stable and could change apparently randomly with time and location.

2.3.2 Regulation Limitations

By definition PLC injects high frequencies in the electric network wires. This injection may induce radio emissions in the HF spectrum and is likely to interfere with existing radio services. For this reason PLC emission and radiation have been regulated from the very beginning.

Table 2.1 Comparison of different powerline technologies

Organization Type	Organization(s)	Technology Name	Frequency Bands	Characterization	Modulation Methods	Signal Level Plan
International SDO	ISO/IEC, ANSI, LonMark	LonWorks, ISO/IEC 14908-3, ANSI 709.2	A (86 kHz and 75.453 kHz 125–140 kHz Fc = 131.579 kHz CENELEC A, B, C PL110 (95–125 kHz Fc = 110 kHz) PL132 (125–140 kHz Fc = 132.5 kHz)	Dual carrier A (86 kHz and 75.453 kHz) C (131.579 kHz and 86.232 kHz)	BPSK/NRZ	EN 50065-1, FCC Part 15
International SDO	ISO/IEC, BS	KNX, ISO/IEC 14543-3-5, EN 50090	PL110 (95–125 kHz Fc = 110 kHz) PL132 (125–140 kHz Fc = 132.5 kHz)	A(60, 66, 72, 76, 82.05 86 kHz) B 110 kHz, C 132.5 kHz	S-FSK/NRZ	EN 50065-1
International SDO	IEC TC57 WG09	IEC 61334-3-1, IEC 61334-5-1, IEC 61334-5-2, IEC 61334-4-32	FC = 132.5 kHz CENELEC A 20–95 kHz	10 kHz tone separations	S-FSK	EN 50065-1
Industry Specification PRIME Alliance	uSyscom, ADD, STM, TI	PRIME	CENELEC A (~42 – ~89 kHz), Capable up to 500 kHz	97 subcarriers, 488 Hz spacing	OFDM/DBPSK, DQPSK, D8PSK	EN 50065-1, FCC Part 15
Industry Specification Public	Maxim, ERDF, TI	LF NB OFDM G3 PLC	35.9–90.6 kHz/ CENELEC-A	36 carriers/ CENELEC-A	OFDM/DBPSK, DQPSK	EN 50065-1, FCC Part 15
Industry Specification	INSTEON Alliance	INSTEON	131.65 kHz	Single carrier	BPSK	FCC Part 15

Industry Specification	HomePlug Alliance	HomePlug C&C	FCC (120–400 kHz), CENELEC A,B	Spread Spectrum FCC 120–400 kHz CA 20–80 kHz, CB 95–125 kHz	Single carrier, spread spectrum DCSK6, DCSK4 PPM	EN 50065-1, FCC Part 15
Proprietary Specification	PCS	UPB	4–40 kHz			FCC Part 15
Proprietary Specification	<i>Pico Electronics</i>	X10	120 kHz	Single carrier	OOK	FCC Part 15
Proprietary Specification	ACT	A10	120 kHz	Single carrier	OOK	FCC Part 15
Proprietary Specification	Phillips	Phillips TDA5051A	Fc within 95–145 kHz (132.5 kHz typical)	Single carrier	ASK	FCC Part 15
Proprietary Specification	Ariane Controls	PLM-1	50–500 kHz, 262 kHz expected	Single carrier	FSK/NRZ	FCC Part 15
Proprietary Specification	ENEL	SITRED	CENELEC A 20 kHz–95 kHz	10 kHz tone separations	S-FSK	EN 50065-1
Proprietary Specification	Maxim	G3 Lite MAX2990	10–490 kHz CENELEC-FCC	Adjustable number of Subcarriers	OFDM	EN 50065-1
Proprietary Specification	Watteco	WPC	1.7–4 MHz	LRWBS for low-rate wide band services	Pulse modulation	LRWBS

(Continued)

Table 2.1 (Continued)

Organization Type	Organization(s)	Technology Name	Frequency Bands	Characterization	Modulation Methods	Signal Level Plan
Proprietary Specification	Yitran/Rensas	C&C Turbo	FCC (120–400 kHz) CENELEC A,B	Spread Spectrum FCC 120–400 kHz CA 20–80 kHz, CB 95–125 kHz	Single Carrier, spread spectrum, DCSK Turbo, DCSK6, DCSK4	EN 50065-1, FCC Part 15
Technologies under development (no detailed public information available)						
International SDO	IEEE P1901.2	LF NB OFDM (Draft)	CENELEC A/B/C/D		OFDM	EN 50065-1, FCC Part 15
International SDO	ITU-T SG-15/T4	G.hnem (Draft)	FCC 9 to 434 kHz CENELEC A/B/C/D		OFDM	EN 50065-1, FCC Part 15
International SDO	ISO 15118 CEI TC69 ETSI M2M	Electrical vehicle communication protocol vehicle and Grid	FCC 9 to 434 kHz CENELEC A/B/C/D			EN 50065-1
International SDO	ETSI PLT	Low-rate home automation	1.7–4 MHz	LRWBS	OFDM	prEN 50561-1

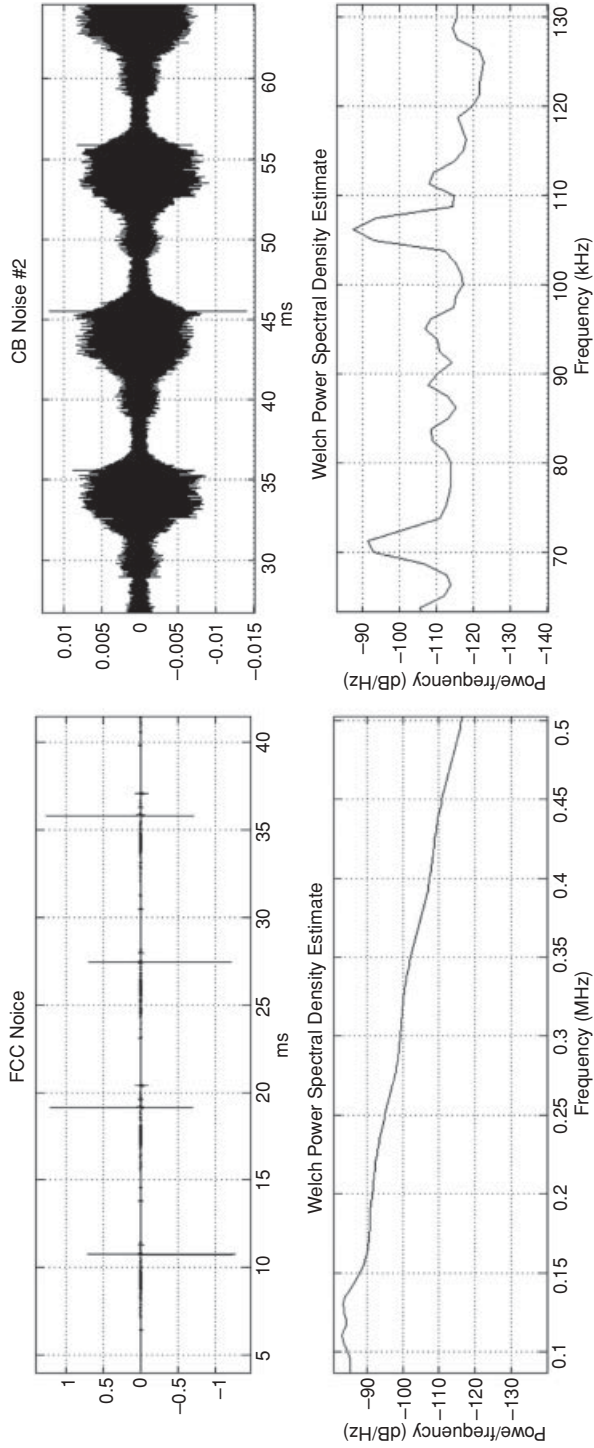


Figure 2.2 (a) and (b): Spikes and noise around the zero-crossing zone. Courtesy Yiran.

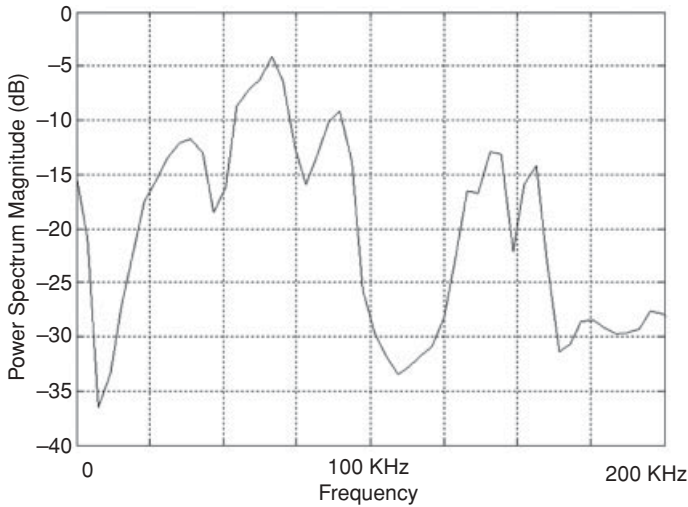


Figure 2.3 Example of absorption in the field. Courtesy Maxim.

Basically there is a distinction in term of regulation whether PLC is narrowband or broadband.

- Narrowband services are subject to CENELEC 50 065-1 or FCC Part15 regulations.
- Broadband services regulation is still in discussion in CENELEC (prEN 50 561-1).

For narrowband services, compared to FCC Part15 in the USA, PLC regulation in Europe appears to be more restrictive as only the frequency band less than 148.5 kHz is available for transmission while it is open up to 500 kHz in the US.

EN 50 065-1 allocates the 3–95 kHz frequency band to utilities for metering applications (it is known as the A band) and reserves the 95–125 kHz, 125–140 kHz and 140–148.5 kHz bands (known in early versions of the standard respectively as B, C, D bands) for analog and digital application within homes, commercial or industrial premises and for control of equipment connected to the low-voltage distribution network. Typical examples of B, C, D band applications are street lighting, electric vehicles or home automation.

As EN 50 065-1 conveys no rights to any user to communicate over any part of the electricity network owned by another party, services using narrowband PLC in home are limited to the 95–148.5 kHz band.

This 15-kHz-limited bandwidth severely limits services to extremely low rates applications and is one of the reasons for the limited diffusion of home automation PLC in homes in the CENELEC bands.

2.3.3 Power Consumption

Until recently, power consumption of PLC modems was not seen as a constraint due to the natural access to energy. But the increased awareness of the overall power consumption

of IT-related technologies, has now installed power consumption as a major constraint for new PLC technologies:

- Low power is now mandatory for many technologies including powerline systems for smart grid and smart metering deployments.
- In Europe, the “less than $1/2$ Watt” European Directive 2005/32/EC on standby power imposes new paradigms for powerline technologies.

If we look at the numbers: the average consumption of the best IEEE 1901 200 Mb/s powerline modem is around 6 W, while it is close to 4 W for a OFDM CENELEC modem. In a home environment, 4 or 6 W are insignificant compared to a 2 kW air-conditioning unit. But, in the context of M2M with a large number of connected devices, or compared to the consumption of a meter or replacement of a switch it is certainly prohibitive.

It is also prohibitive because the power supply size and heat dissipation are particularly challenging in the context of the form factor required by a meter or a switch.

If we look now at the main contributing factors to the overall PLC technology power consumption, it appears it is roughly balanced between analog subsystems and digital subsystems.

According to Moore’s law, digital parts have seen astounding progress in size and performance. But, at the end of the day, having to support increasing data rate, tough channels, sampling frequencies and signal processing are engaged in a never-ending race, thus limiting Moore’s Law benefits regarding power consumption.

On the other hand, except for some ultralow-power coupling technologies [3], power consumption of couplers and analog front end (AFE) are limited in their progress by the emission level required by the injection of high-frequency signals in a difficult media.

As a consequence of restrictive regulations and slow progress in power-consumption reduction, some companies are now working on efficient management of sleeping modes and standby states. Radio technologies, like ZigBee or 6LoWPAN, have implemented the same strategy in the past for battery-powered nodes. At first sight, it appears to be an efficient solution for AC powered nodes too . . . but it isn’t: sleeping modes unfortunately do not solve the power-supply size problem. It is a common [marketing] mistake to mix up overall peak power consumption and energy. A low-energy system, expressed in watt-hour, could need a relatively high peak power, in watt, when operating in “periodic wake-up” mode, and AC power supplies need to be dimensioned according to peak power requirements.

2.3.4 Lossy Network

As shown in Figure 2.4, the PLC link may be subject to as many disturbances as a wireless link, because every electrical device may inject noise and/or absorb the signal. Considering the number of electrical devices in the electrical network of a typical multidwelling

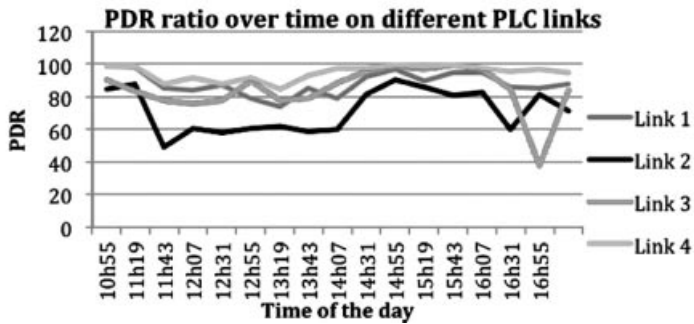


Figure 2.4 Packet delivery ratio (PDR) variation over time on several PLC links. Courtesy Wateco.

unit and their varying electrical behavior that randomly disturbs the communication channel, the routing mechanism over PLC networks has to cope with very lossy links, as well as dynamic loss characteristics. Furthermore, these noise/fading generators create asymmetric links that add routing complexity.

Implementation of special purpose routing protocols is now considered as the solution to reach full coverage with a lossy network and difficult channels. The Internet Engineering Task Force (IETF) recognized the need to form a new Working Group to standardize an IPv6-based routing solution for IP smart object networks, which led to the formation of a new Working Group called ROLL “Routing Over Low power and Lossy” networks in 2008.

Here is the charter of ROLL:

Low-power and lossy networks (LLNs) are made up of many embedded devices with limited power, memory, and processing resources. They are interconnected by a variety of links, such as (IEEE 802.15.4, Bluetooth, low-power WiFi, wired or other low power PLC (powerline communication) links. LLNs are transitioning to an end-to-end IP-based solution to avoid the problem of noninteroperable networks interconnected by protocol translation gateways and proxies.

The routing protocol RPL [2] developed in ROLL is the protocol chosen by ZigBee™ IP and 6LoWPAN. The advantage of RPL is that it is independent of the media and then it can be the basis of interoperability between PLC and radio sensors. Interoperability is simply achieved by routing messages from a PLC node to a wireless node. For more details on RPL, refer to the Section 12.4.

Figure 2.5 shows an illustration of this dual PHY sensor network.

The conclusion is that the powerline medium appears to be a very challenging medium, limited by regulation and power consumption and until recently no existing standard really offered a good alternative to wireless for M2M application in home. Fortunately,

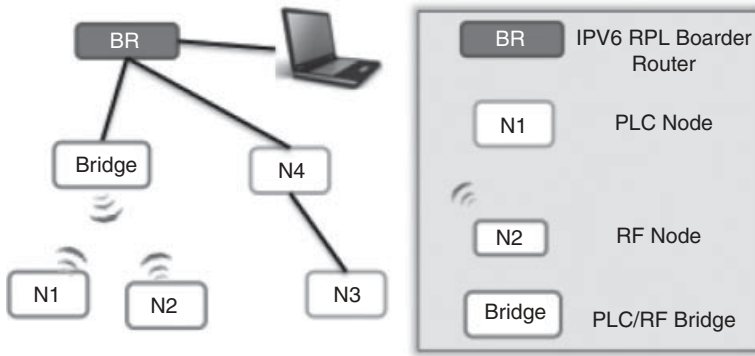


Figure 2.5 Wireless nodes and PLC nodes exchange messages through a RPL border router.

the recent development of ROLL and the requirement of using IPv6 open the door to new powerline standards interoperable with wireless services using the same routing protocol.

2.3.5 *Powerline is a Shared Media and Coexistence is not an Optional Feature*

2.3.5.1 Why is Coexistence So Important?

This question of coexistence in PLC has been discussed in many PLC standard organizations. Many articles can be found in the Smart Grid Interoperability Panel (<http://collaborate.nist.gov/twiki-sggrid/bin/view/SmartGrid/SGIP>).

Powerline cables are a shared medium. They cannot provide links dedicated to a single user and there is no practical way to insulate two neighbors. The signals transmitted within an electrical network can interfere with signals generated within an adjacent house or apartment. It is then likely that these interferences will produce data rate and quality of service drops. The same issues exist in other share media like radio.

For this reason, it is necessary to define mechanisms to limit the harmful interference caused by noninteroperable neighboring devices.

Different mechanisms exist or are proposed today:

- EN 50 065-1 provides a CSMA/CA algorithm and a 132.5-kHz “channel busy” signal for the C band.
- IEEE 1901 and ITU G9972 provide a standard called intersystem protocol (ISP) implementing a TDMA and frequency-division mechanism for a fair access to the shared resource.
- Current discussions in IEEE 1901.2 are also investigating the coexistence between OFDM CENELEC band transmission for metering applications and legacy narrowband services (based on FSK modulation for example).

- prEN50412-4 is a coexistence mechanism in the LRWBS band (2–4 MHz) proposed to CENELEC and based on CSMA/CA in subchannels)

2.3.5.2 Access to the Channel

Low-rate powerline systems have to implement access to channel mechanisms and especially when there is a large number of nodes or when noninteroperable technologies may transmit in the same electrical network. Access to a channel is classically ruled by energy detection in a channel coupled to a CSMA generic mechanism similar to the one described in IEEE 802.15.4.

Compared to the original one, and due to the difference between radio and powerline, some timing constants are different, in particular the backoff period. Basically, the backoff period is the base time a node has to wait before it can transmit after sensing the channel. Compared to 802.15.4 wireless systems, a low-rate PLC modem offers higher latency and lower data rates, however the backoff period has to be chosen short enough (1 ms for example) to allow a sufficient responsiveness for the latency requirements of home-automation applications like switching.

In order to evaluate the impact of data rate and the number of nodes versus average access time, simulations have been carried out with different situations:

- Number of sensors from 5 to 30. All sensors are independent and try randomly to access the network.
- The number of tries/minute is presented vertically.
- A maximum of fail/success rate of 3 is supposed.
- There are no propagation effects and the channel is perfect (best case).

The parameters are:

- backoff period = 1 ms;
- frame size varies randomly from 16 bytes to 128 bytes;
- a failed transmission occurs when the total backoff time > 100 ms (typical minimum latency for home automation HMI).

Figure 2.6 shows that the network supports 20 nodes communicating 400 ms every minute, provided the data rate is a least 100 kb/s.

With a lower data rate, on average every node will retry more than 3 times to access the channel.

Meters transmitting through a powerline modem in the CENELEC band might be impacted by the results of these simulations. For example in a multidwelling situation it is likely that more than 30 nodes might be within reach of the meters. If a high activity level in homes is required (real-time display of the consumption, thermal regulation . . .)

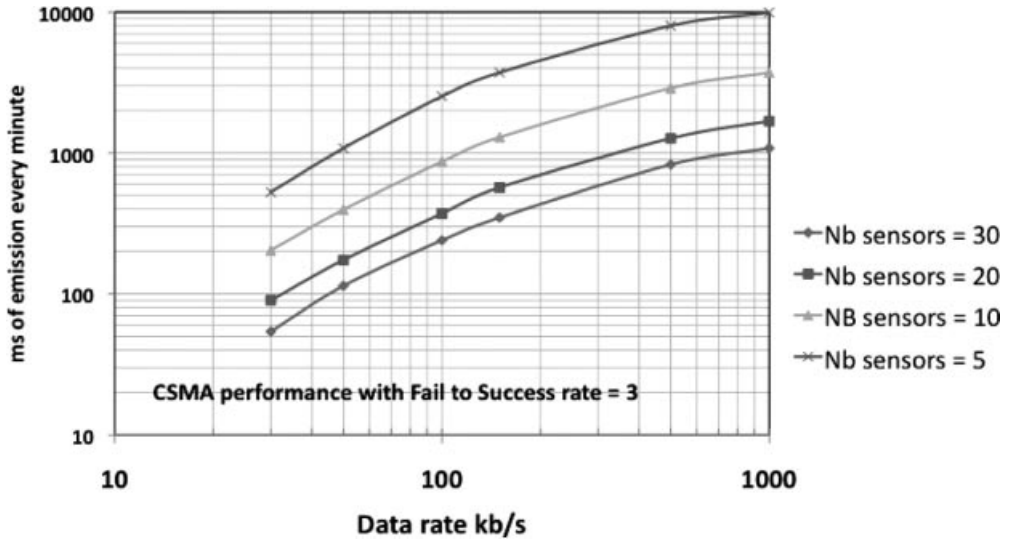


Figure 2.6 Simulations of the impact of number of nodes on access time. Courtesy Watteco.

the 30 nodes could saturate the channel and the latency of the communication between the meter and substations could increase.

Prioritizing meter communication and increasing data rate can minimize this effect but still not cancel it.

Another solution is to use different frequency bands in the home and for metering. For example, the CENELEC band for metering and the LRWBS band (2–4 MHz) for home automation.

2.4 The Ideal PLC System for M2M

PLC combined with recent IPv6 developments could lead to a new standard for wired communication in homes. Recently, in the context of European Mandate 441 for metering applications and home automation, ETSI PLT has defined the requirements of this new standard.

What are the main requirements of this new PLC M2M standard?

- Openness and availability;
- Range;
- Energy consumption;
- Data rate;
- Robustness;
- EMC regulatory compliance;

- Coexistence with other PLC technologies;
- Security;
- Latency;
- Interoperability with M2M wireless services.

2.4.1 *Openness and Availability*

Open standards, compared to proprietary developments, are accelerating factors for the dissemination and the success of any mass-market technology. Standard openness encourages interoperability, stimulates multisourcing and low-cost policies. The use of available standards, when adaptation to in-home PLC automation is possible, is a tremendous factor of stability and interoperability between existing mature technologies and new emerging ones.

- Home automation standards must use existing, available and internationally recognized PHY, MAC and DLL technologies. Examples are: IEEE 802.15.4, 6LoWPAN, IPv6, RPL . . .

2.4.2 *Range*

Range (or coverage) is one of the most important requirements in PLC. It is very challenging because of the extreme harshness of the medium. multidwelling units (MDU) where multiphases networks can exceed an internode distance of 100 m are probably the worst-case situation in terms of range performance.

- The standard should be able to cover all outlets of the home. Routing protocols like RPL can be used to ensure full coverage if needed.

2.4.3 *Power Consumption*

Power consumption is a very sensitive parameter for home automation applications.

- The standard must enable products to comply with local low power regulations like Directive 2005/32/EC on energy consumption in standby mode in Europe or Energy Star in the US.
- The standard must enable products with low power consumption in relation to the savings users may request for energy-efficiency products.
- The power consumption of a powerline node must be comparable to the power consumption of a wireless node.

2.4.4 Data Rate

Data rate is not a critical requirement per se for home automation applications. A data rate of 10 kb/s is, in most cases, enough to cover in-home lighting or switching applications. However, factors like routing protocols, security, access to media mechanisms in a home with dozen of nodes will probably increase communication stream and payload.

For that reason, it is critical not to limit the standard data rate to 10 kb/s. On the other hand, high-rate PLC systems (>1 Mb/s) are not a good compromise for evident reasons of power consumption. 100 kb/s appears to be a good compromise.

- The standard must provide a nominal rate of 100 kb/s in field installations.
- In the case of very noisy channels, the standard should support variable data rates to keep reliable links.

2.4.5 Robustness

The home environment can be very challenging for PLC nodes. Harsh channels in homes are due to various physical reasons:

- Low impedance appliances (from 1 ohm to 10s of ohms, pure capacitive loads, etc.);
- Disturbance from common electric devices (chargers, dimmers, switching power supplies, etc. . . .);
- Absorption from breakers and ground fault circuit interrupter;
- Electrical topology (multiphase wiring, neutral/ground connections, etc.).

In order to offer a good end-user experience:

- The standard must provide close to 100% connectivity in the home.
- The standard may use routing or/and data rate adaptation in order to keep connectivity in harsh environments.
- The standard must support multiphase topologies, optionally with phase couplers. Across the world, different home wiring topologies exist using single phase (France, Spain . . .), dual phase (US and Japan) or three phases (Germany and Northern countries). Usually, the PLC signal is injected in one phase and natural crosstalk between phases may not always be sufficient. In that case phase couplers ensure reinjection of signal from one phase to the other phases.

2.4.6 *EMC Regulatory Compliance*

- The standard must comply with local EMC regulations in force in the frequency band in use.

2.4.7 *Coexistence*

- The standard must implement existing coexistence standards when using frequency bands where other PLC systems are also transmitting.

Coexistence standards already in use by legacy PLC systems are:

- EN 50 065-1 C band: (channel busy signaling at 132.5 kHz);
- ISP mechanism in the 2–30 MHz band (as described in IEEE 1901 or ITU G.9972).

2.4.8 *Security*

- The standard must provide services to support encryption and secure data services. However, a compromise should be found between security and cost of implementation. Furthermore, plug and play installation may conflict with strong security requirements.
- The security suite must be open and available.

2.4.9 *Latency*

- The standard must support low-latency communication in conformance with end-user usual expectation in home automation. Usually, a latency of 100 ms is considered as a maximum for home automation applications.

2.4.10 *Interoperability with M2M Wireless Services*

- The standard must ensure interoperability with 6LoWPAN and other wireless compatible protocol through gateways embedding both nodes and running RPL.

2.5 **Conclusion**

PLC technologies, after a rather difficult and slow start, are now on track for mainstream deployments triggered by smart grid and home automation investment programs and mandates all over the world.

It is worth noting that new paradigms and business models will deeply influence the specifications of emerging standards. The capacity of the communication channel to support encrypted IPv6 frames, the large number of nodes, low power consumption and interoperability with radio transmission technologies are clearly the next challenges engineers will have to overcome.

References

- [1] PowerLine Communications (2010) *Theory and Applications for Narrowband and Broadband Communication Over Power Lines*, Wiley
- [2] RPL: IPv6 Routing Protocol for Low power and Lossy Networks - <http://tools.ietf.org/html/draft-ietf-roll-rpl>.
- [3] IPSO White Paper #6 – “A survey of several low power Link layers for IP Smart Objects” by JP Vasseur, Paul Bertrand, Bernard Aboussouan, Eric Gnoske, Kris Pister, Roland Acra and Allen Huotari.

Part Two

Legacy M2M Protocols for Sensor Networks, Building Automation and Home Automation

3

The BACnet™ Protocol

BACnet stands for Data Communication Protocol for **B**uilding Automation and Control **N**etworks. Unlike most other protocols that began as private implementations followed by standardization efforts, BACnet was built from the ground up as an independent, royalty-free, open standard control and automation protocol. The standard committee was chaired by university professors until 2004, its goal was to harmonize data types and formats, data exchange primitives, and common application services. Several open source BACnet stacks are available.(e.g., <http://bacsharp.sourceforge.net/>; <http://bacnet.sourceforge.net/>).

The scope of BACnet applications is very large, including HVAC (heating, ventilating, and air conditioning) applications, lighting control, fire control and alarm, security, and interfacing to utility companies.

Together with LonWorks, BACnet is one of the most popular industrial automation and control protocol, adopted in products of many leading vendors (Siemens Building Technologies, Johnson Controls, Inc., Teletrol Systems, @IC, TAC, KMC Controls, American Auto-Matrix, Contemporary Controls Ltd, Reliable Controls).

3.1 Standardization

The BACnet standardization effort began in 1987 during a Standard Project Committee meeting of ASHRAE (American Society of Heating, Refrigerating and Air-Conditioning Engineers). BACnet became an ISO standard in 2003 (ISO 16 484-5). In January 2006 the BACnet Manufacturers Association and the BACnet Interest Group of North America combined their operation in a new organization called BACnet International (<http://www.bacnetassociation.org/>), which provides conformance testing

services (BACnet Testing Laboratories) and promotes the adoption and development of the standard.

3.1.1 *United States*

BACnet became a standard in 1995 as ASHRAE/ANSI standard 135 and a conformance testing method was standardized in 2003 as BSR/ASHRAE Standard 135.1. The last revision of the standard was published in 2010.

3.1.2 *Europe*

BACnet was adopted in 2003 by CEN (Comité Européen de Normalization, <http://www.cen.eu>) Technical Committee 247, for the management level and automation level. For the Automation level, it coexists with EIBnet (Konnex), at the Field level, CEN adopted Konnex (merger of three European protocols EIB (European Installation bus), Batibus, EHS), and LonWorks/LONTalk.

Europe has a specific European user and interest group: <http://www.big-eu.org>.

3.1.3 *Interworking*

BACnet ability to interwork with other technologies has always been a key concern, and BACnet does provide enough flexibility to allow mapping of other common protocols to a BACnet model. However, there are often many ways of providing such a mapping, and there is a need to formally specify a standard mapping in order to ensure interoperability of interprotocol gateway implementations:

- BACnet interoperability with Konnex (KNX) control protocol has been specified in Annex H/5.
- BACnet interoperability with ZigBee has been specified in Annex X.

3.2 **Technology**

BACnet focuses on the network layer and above. At the presentation layer, it uses ASN.1 syntax¹ for the definition of all data structures and messages (application protocol data units or APDUs). The BACnet transport layer adds routing information to these APDUs,

¹ASN.1, or “Abstract Syntax Notation 1” is defined in ISO/IEC 8824. This syntax, widely used in the telecom world, is used to define precisely data structures, and also functional primitives. It includes a standard serialization mechanism : ASN.1 BER (simple but less efficient), and ASN.1 PER (complex but extremely efficient).

and the resulting messages may be carried on top of virtually any link layer, using the adaptation functions provided by the BACnet network layer.

3.2.1 *Physical Layer*

BACnet upper layers are independent from the underlying physical layer, facilitating the implementation of BACnet on most popular networks. BACnet physical layers have been defined for ARCNET, Ethernet, IP tunneling (defined for routers interconnecting BACnet segments over IP in Annex H), BACnet/IP (devices are IP aware and can communicate directly over IP networks), RS-232 (BACnet Point to Point), RS-485 (with BACnet specific Master-Slave/Token Passing LAN technology, up to 32 nodes on 1200 m, at 76 kbit/s on shielded twisted pairs), and LonWorks/LonTalk.

Since 2008, there is also a standard implementation of BACnet over ZigBee® (Annex X).

3.2.2 *Link Layer*

BACnet can be implemented directly on top of the LonTalk or IEEE802.2 (Ethernet and ArcNet) data link layers. It also defines a data link layer (Point to Point PTP) for RS232 serial connections, and a MS/TP data link layer for RS-485.

For IP or other network technologies that can be used as link layers, the standard defines a BACnet virtual link layer (BVLL) that formalizes all the services that a BACnet device might require from the link layer, such as broadcasts.

For instance, BACnet devices may implement the IP BVLL, which encapsulates the required control information not readily available from the native IP link layer (e.g., a flag indicating whether the message was received as a unicast or broadcast), in a BACnet virtual link control information (BVLCI) header (see Figure 3.1). Thanks to the IP BVLL, BACnet devices become full-fledged BACnet IP devices, able to communicate directly over IP without a need for an “Annex H” router. Similarly, a BACnet device could implement an ATM, frame relay or ISDN BVLL in order to become a native node in these networks.

On many link layers, broadcasts are difficult or have their own limitations. BACnet has a concept of a “BACnet broadcast management device” (BBMD), which implements the broadcast requirements of BACnet for the selected link layer, for example, it may convert a BACnet broadcast into IP-based multiunicast and/or broadcast messages. Devices can register with the BBMD to receive broadcast messages dynamically.

3.2.3 *Network Layer*

BACnet is primarily defined as a network layer protocol, which defines the network addresses required for the routing of messages. BACnet networks consist of one or more

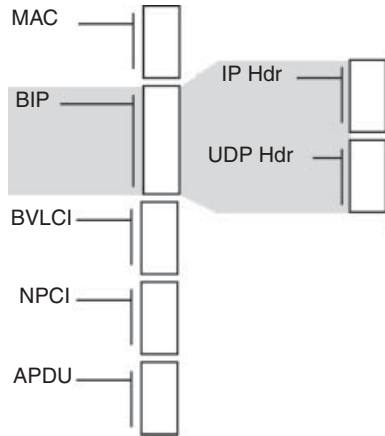


Figure 3.1 Transport of a BACnet application message (APDU) over IP/UDP.

segments consisting of single physical segments or multiple physical segments connected by repeaters. The BACnet segments are connected by bridges if they employ the same LAN technologies, or BACnet routers otherwise.

BACnet addresses are hierarchical: the formal separation of the network identifier and the address identifier simplifies routing. Addresses have a variable length, which makes it easy for BACnet to adapt to the native addresses of underlying link layers (Figure 3.2 shows the BACnet use of an IP address). The BACnet network header (NPCI) can include the following information elements:

- A 2-byte source network (SNet) and variable length source address (SAddr, SLen). For Ethernet, ArcNet and MS/TP, the native protocol address format is used, for LonTalk,

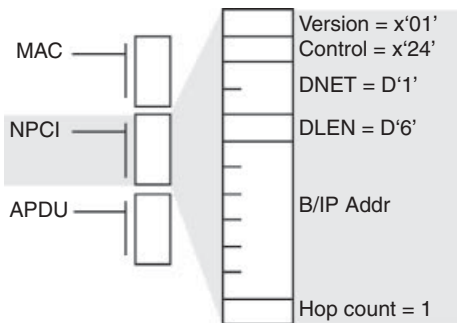


Figure 3.2 BACnet message from a BACnet non-IP device to a BACnet/IP device.

the concatenation of the subnetID and nodeID (2 bytes), or the concatenation of the subnetID and Neuron ID (7 bytes) is used.

- A 2-byte destination network (DNet) and variable-length destination address (Daddr,Dlen). For broadcast messages, DNet identifies the network on which a broadcast is required, and Dlen=0.
- A 4-level network priority indicator
- A 1-byte message type: 9 message types are used by the BACnet routing mechanisms (e.g., *Who-is-router-to-network* to discover a router to a specific networkID). Vendors can define specific extension message types.
- A 2-byte vendor ID.

Not all information elements have to be present, depending on the specific use-case: a *control* bitmask field specifies which information elements are present.

3.2.4 Transport and Session Layers

BACnet implements a collapsed OSI model in which the transport and session layers are not required. The application layer provides the required reliability mechanisms usually associated with the transport layer, as well as the segmentation and sequencing mechanisms usually associated with the session layer.

3.2.5 Presentation and Application Layers

BACnet does not attempt to formally separate the presentation layer and the application layer (a separation that is often a bit artificial for most protocols anyway). BACnet models the various features of devices as *objects*, exchanging *service* primitives. The service primitives are described using ASN.1 syntax and serialized using ASN.1 BER (basic encoding rules, ITU-T Recommendations X.209 and X.690, for a good introduction on ASN.1 and BACnet, see <http://bacnetbill.blogspot.com/2009/10/bacnet-tagging-rules.html>).

3.2.5.1 BACnet Objects

BACnet abstracts the device basic functions as *objects*: each *device* is decomposed into a collection of standardized objects, where physical inputs and outputs and other characteristics of the object (name, type, configuration parameters) are represented by *properties*. Each object is identified by a unique *Object_Identifier* within the device. See Table 3.1 for a list of standard BACnet objects.

BACnet currently lists 30 object types, for which it defines standard properties and the expected behavior:

```
BACnetObjectType ::= ENUMERATED {
    access-door           (30),
    accumulator          (23),
    analog-input         (0),
    analog-output        (1),
    analog-value         (2),
    averaging            (18),
    binary-input         (3),
    binary-ouput        (4),
    binary-value         (5),
    calendar             (6),
    command              (7),
    device               (8),
    event-enrollment    (9),
    event-log            (25),
    file                 (10),
    group                (11),
    life-safety-point    (21),
    life-safety-zone     (22),
    load-control         (28),
    loop                 (12),
    multi-state-input    (13),
    multi-state-output   (14),
    multi-state-value    (19),
    notification-class   (15),
    program              (16),
    pulse-converter      (24),
    schedule             (17),
    – see averaging     (18),
    – see multi-state-value (19),
    structured-view      (29),
    trend-log            (20),
    trend-log-multiple   (27),
    – see life-saftey-point (21),
    – see life-saftey-zone (22),
    – see accumulator    (23),
    – see pulse-converter (24),
    – see event-log      (25),
    – enumeration value 26 is reserved for a future addendum
    – see trend-log-multiple (27),
    – see load-control    (28),
    – see structured-view (29),
    – see access-door    (30),
```

Table 3.1 Standard BACnet objects

AnalogInput
AnalogOutput
AnalogValue
BinaryInput
BinaryOutput
BinaryValue
Calendar
Command
Device
EventEnrolment
File
Group
Loop
MultistateInput
MultistateOutput
NotificationClass
Program
Schedule
Averaging
MultistateValue
TrendLog
LifeSafetyPoint
LifeSafetyZone
Accumulator
PulseConverter

- **Device:** all devices are required to implement the device object. The *Object_Identifier* of the device object must be unique across the BACnet network and is the identifier of the physical device implementing that device object. All other objects are implemented only if needed. The device object lists all objects implemented by the device.
- **Binary input, Binary output, Binary value.**
- **Analog input, Analog output, Analog value, Averaging** (function that monitors a signal and records its min, max and average value).
- **Multistate input, Multistate output, Multistate value.**
- **Accumulator:** for devices implementing a feature that counts pulses. **Pulse converter:** counts pulses or takes an accumulator object as an input, but the count can be offset (adjusted) at any time, and scaled.
- **Loop:** properties modeling a feedback control loop.
- **LifeSafetyPoint** (such as smoke detectors, pull stations, sirens. . .), **LifeSafetyZone** (properties associated to a group of LifeSafetyPoints and LifeSafetyZones).

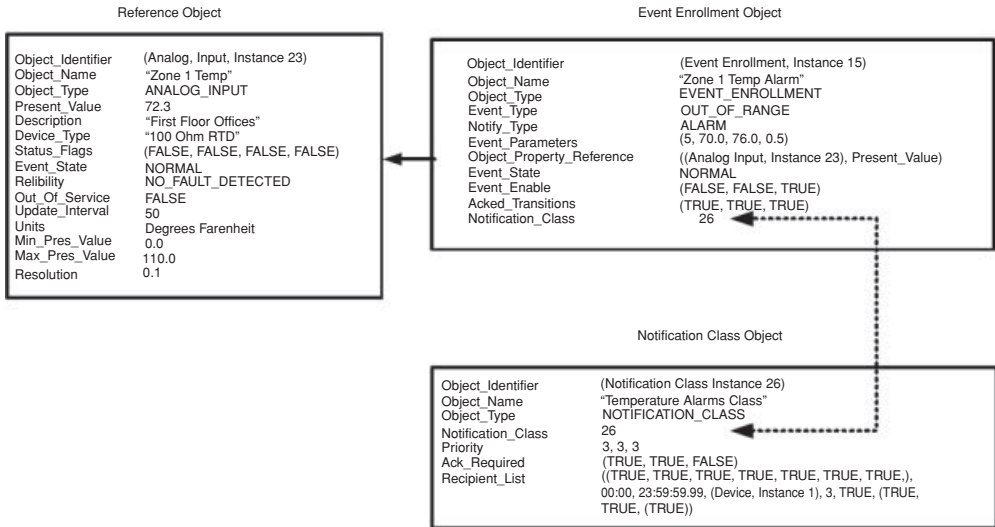


Figure 3.3 Example use of BACnet objects for event notification (from ANSI 135 standard document).

- **Access door:** object modeling a physical door.
- **Calendar** (list of dates). **Schedule:** the schedule object describes a periodic schedule (days and time of day, with exceptions), and for each period associates time-dependent values to the properties of other objects.
- **Event enrollment:** this object defines the required conditions for an event to occur (e.g., a change of state or value, a command failure, a value getting off range for a certain duration, a sudden change of a value, etc.), and which devices or objects must be notified if this happens. See Figure 3.3.
- **Notification Class:** this list of objects that are enrolled for an event notification is used as a property of the Event enrollment object. It also contains priority parameters. The list of notified elements may vary according to days of the week and time of day. See Figure 3.3.
- **Command** (a command object can be used to execute a set of actions changing properties of other objects).
- **File:** lists the properties of a file object that may be accessed using File services.
- **Program:** list the properties modeling an application program.
- **Trend Log:** this object models a function that logs, at periodic intervals, a given property value. Such a log can be triggered based on predefined conditions. **Trend Log Multiple Objects:** same as a Trend Log, but stores multiple property values in parallel. **Event Log:** "FIFO" log of timestamped event notifications.
- **Group:** a group of objects. **Structured View:** organizes other objects in an org-chart-type structure.

- **Load control:** this object provides the interfaces for power-load shedding, for example, processing of requests from a utility company. It support hierarchical load shedding, scheduled shedding, compliance reporting.

Typical object properties include the object name, a reliability indicator (no error detected, no sensor, etc.), the present value of a counter and its unit, a scaling factor, maximum and minimum values. They also include configuration parameters for object features, for example, how long a value must remain out of range before reporting an error condition, or the minimal increment of a real value triggering an update notification.

3.2.5.2 BACnet Services

BACnet considers that all objects are servers that provide *services*. It defines 5 classes of services, the description of each service can be found in ANSI/ASHRAE 135 clauses 13 to 17 (see Table 3.2).

Each service uses a set of messages supporting the related communication needs. The messages are defined using ASN.1 syntax (ANSI/ASHRAE 135 clause 21) and exchanged using standard remote operation primitives (request, indication, response, confirm):

- **Alarm and event services:** BACnet provides multiple event reporting options: objects may support “intrinsic reporting” (e.g., report an event periodically, report error conditions, status updates), or may be configured by means of *Event enrollment* object (Figure 3.3) to report specific conditions such as a change of value (*COV reporting*), or a value out of range. The latter mechanism, called *algorithmic reporting* implements a subscribe-notify model for events. The objects that requested to be notified are listed in *Notification Class* objects (Figure 3.3).

The following service primitives are defined for event management and reporting: **AcknowledgeAlarm** (self-explanatory), **ConfirmedCOVNotification** (*Change of value* event notification primitive in which receivers must report the success or failure of actions taken as a result of the event), **UnconfirmedCOVNotification**, **ConfirmedEventNotification**, **UnconfirmedEventNotification**, **GetAlarmSummary** (BACnet events can be flagged as alarms, in which case a list of active alarms is returned by this primitive), **GetEnrollmentSummary** (returns a list of event-notifying objects according to specified filters, such as objects with an active event enrollment from another object), **GetEventInformation** (returns a list of active event states within a device), **LifeSafetyOperation** (e.g., silence a siren), **subscribeCOV** (subscribe to *Change of value* notifications for an object), **SubscribeCOVProperty** (subscribe to *Change of value* notifications for a property).

- **File access services:** read and write primitives are atomic, that is, a single operation is executed at a time.

Table 3.2 Standard BACnet services

Who Is
I Am
Who Has
I Have
Read Property
Write Property
Device Communication Control
ReinitializeDevice
Atomic Read File
Atomic Write File
Time Synchronization
UTC Time Synchronization
Subscribe COV
Subscribe COV Property
Confirmed COV Notification
Unconfirmed COV Notification
Read Property Multiple
Read Property Conditional
Read Range
Write Property Multiple
Get Alarm Summary
Get Event Information
Get Enrollment Summary
Acknowledge Alarm
Confirmed Event Notification
Unconfirmed Event Notification
Unconfirmed Text Message
Confirmed Text Message
Add List Element
Remove List Element
Create Object
Delete Object
Unconfirmed Private Transfer
Confirmed Private Transfer
VT Open
VT Data
VT Close
Life Safety Operation
Get Event Information

- **Object access services:** a set of self-explanatory primitives: **ReadProperty**, **ReadPropertyConditional**, **ReadPropertyMultiple**, **WriteProperty**, **WritePropertyMultiple**, **CreateObject**, **DeleteObject**, **AddListElement**, **RemoveListElement**.
- **Remote device management services:** a set of primitives for maintenance purposes (start and stop BACnet message transmission, send vendor specific commands, reinitialize a device, time synchronization). Among those primitives, the *Who-Has/I-Have* services are used to discover which devices on the network have a given object name or object ID, the *Who-Is/I-Am* primitives are used to discover devices on a BACnet network.
- **Virtual terminal services:** primitives for bidirectional exchange of character-oriented data, “Telnet like”.

3.3 BACnet Security

BACnet device A supporting security can request a session key from a key server for a future communication with device B. The key server will generate a session key SK_{AB} and transmit it securely to A and B (encrypted with the private keys of A, respectively B). BACnet uses 56-bit DES encryption.

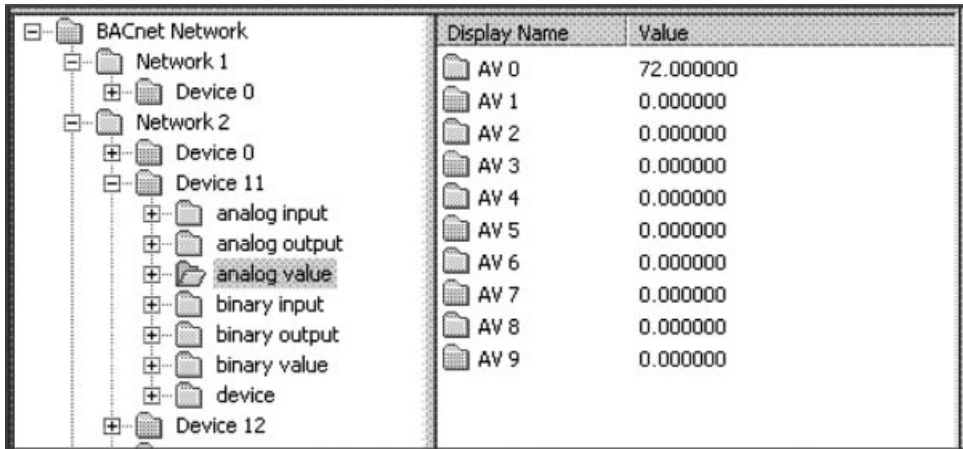
Device A may then authenticate a future transaction with B: A and B authenticate each other by exchanging challenges (based on random numbers encrypted with the session key), the challenge message includes the identifier (InvokeID) of the future transaction to be authenticated.

A may also ensure the confidentiality of the future transaction by encrypting the corresponding application message with the session key.

3.4 BACnet Over Web Services (Annex N, Annex H6)

The XML working group of ASHRAE SSPC 135 has introduced Addendum c to BACnet-2004 that specifies a Web Services interface to building automation and control systems. The addendum is in two parts:

- Annex N to BACnet defines the BACnet Web Services interface, BACnet/WS. BACnet/WS is a connectionless protocol using a Simple Object Access Protocol 1.1 interface (<http://www.w3.org/2002/ws/>) over HTTP (RFC 2616). The model and primitives exposed in Annex M are independent from the underlying protocol and could apply to any building automation protocol (LonWorks, KNX, ModBus . . .).
- Annex H6, Combining BACnet Networks with Non-BACnet Networks, that prescribes the gateway mapping specifically to and from BACnet messages. Annex H6 exposes a specific profile for Web Services access to underlying BACnet objects.



Display Name	Value
AV 0	72.000000
AV 1	0.000000
AV 2	0.000000
AV 3	0.000000
AV 4	0.000000
AV 5	0.000000
AV 6	0.000000
AV 7	0.000000
AV 8	0.000000
AV 9	0.000000

Figure 3.4 Snapshot of the SCADA engine BACnet web service.

3.4.1 The Generic WS Model

The BACnet/WS fundamental data structure is a tree of *nodes* under a single *root node*. In order to allow more flexibility in the tree structure, BACnet/WS has a notion of *reference node* that can serve as an alias to a *referent node*.

Each node and each property is identified by a path:

- /Floor2/Room3/Discharge Temp identifies a node;
- /Floor2/Room3/Discharge Temp:inAlarm identifies property inAlarm of node DischargeTemp;
- Another format commonly used by BACnet WS vendors (Figure 3.4) is the following /[Network]/[Device]/[ObjectType]/[Instance], for instance /2/11/2/0 is a path to **network 2, device 11, analog value (object type 2) instance 0**;
- Special value <Path>:Children indicates all the nodes one level below the current node. For instance /:Children returns all the children of the root path. /1/1/0:Children returns all paths to Analog Inputs on Device 1.

The network visible state of each node is exposed as the node value, and a collection of attributes (Figure 3.5). BACnet uses Primitive attributes (native XML types), Array attributes (arrays of Primitive values), Enumerated attributes (choice of XML strings specified by BACnet/WS). Only the Value attribute is writable with the services defined by the standard.

BACnet/WS defines a small set of 5 standard nodes required in any implementation, for example:10.5

.sysinfo/software-version, a string containing the software revision of the software running on the server

Attribute	XML type	Description
NodeType	String	Hint about a node content. One of Unknown /System /Network /Device /Functional /Organizational /Area /Equipment /Point /Collection /Property /other
NodeSubType	String	
Description	String	
DisplayName	String	
Aliases	String	
Reference		
Attributes	String	Array containing the list of all attributes present in this node.
Children	String	Array containing the collection of identifiers for the children of this node when accessed through this path. The path to the each child is obtained by concatenation of the specified child identifier with the path of the current object.
HasDynamic Children	Boolean	
Value	Depends on ValueType	e.g. if the ValueType is 'OctetString', the XML type of the Value property will be base64binary.
ValueType	String	One of None /String /OctetString /Real /Integer /Multistate /Boolean /Date /Time /DateTime /Duration
Units	String	If ValueType is 'Real' or 'Integer', Engineering unit for the Value attribute of the node, expressed as the enumeration identifier of the corresponding unit in the BACnetEngineeringUnits ASN.1 production (if the canonical service option is used, otherwise arbitrary units can be used)
ValueAge	Double	In seconds
HasHistory	Boolean	
Writable	Boolean	True if the value is writable through web services.
WritableValues	String	Array containing all possible values that may be written to the Value attribute (when ValueType is Multistate or Boolean)
InAlarm	Boolean	True to indicate an alarm condition
PossibleValues	String	Array containing all possible values of the Value attribute (when ValueType is Multistate or Boolean)
MinimumLength	nonNegativeInteger	
MaximumLength	nonNegativeInteger	
Resolution	Depends on ValueType	
Maximum	Depends on ValueType	
Minimum	Depends on ValueType	
Overridden	Boolean	

Figure 3.5 BACnet/WS mandatory and most common attributes.

3.4.2 BACnet/WS Services

BACnet/WS defines the services used to access and manipulate the data on the server:

- **getValue/getRelativeValues:** from a path parameter, the `getValue` service retrieves a single value for a single attribute on a single node (Figure 3.6). The value can be a primitive type or an array attribute (in which case the results are concatenated with a semicolon separation in the result string). **getValues:** accepts multiple paths parameters and returns multiple values. **getArray/getArrayRange:** these services accept a single path to an array parameter and return an array of strings (the entire array or a portion for `getArrayRange`). **getArraySize** returns the size of an array.

Request

```
<?xml version="1.0" encoding="UTF-8"?>
<SOAP-ENV:Envelope
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:ns="urn:bacnet_ws">
  <SOAP-ENV:Body SOAP-ENV:encodingstyle=
    "http://schemas.xmlsoap.org/soap/encoding/">
    <ns:getvalue>
      <ns:optionsx/ns:options>
      <ns:path>/.sysinfo/.vendor-name</ns:path>
    </ns:getvalue>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Response

```
<?xml version="1.0" encoding="UTF-8"?>
<SOAP-ENV:Envelope
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:ns="urn:bacnet_ws">
  <SOAP-ENV:Body SOAP-ENV:encodingstyle=
    "http://schemas.xmlsoap.org/soap/encoding/">
    <ns:getvalueResponse>
      <ns:result>SCADA Engine</ns:result>
    </ns:getvalueResponse>
  </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Figure 3.6 Example WS `getValue` request and response for the value of `.sysinfo/.vendor-name`.

- **setValue:** sets a new value for a single attribute on a single node identified by a path. **setValues** sets multiple values identified by multiple paths.
- **getHistoryPeriodic:** specifies a sampling interval and a start time for a property identified by its path, and returns a specified number of samples interpolated according to a specified method.
- **getDefaultLocale/getsupportedLocales:** retrieves the locale(s) that the server has configured for its default locale/supports.

Some services accept service options that modify their behavior or their return values: readback (reads back the value is the result string of setValue), errorString, errorPrefix, locale, writeSingleLocale, canonical, precision, noEmptyArrays, writePriority.

3.4.3 The Web Services Profile for BACnet Objects

This profile specifies the mapping of some BACnet object properties to BACnet/WS node attributes (Figure 3.7).

BACnet object property	WS node attribute
Object_Name	DisplayName
Present_Value	Value
Units	Units
Status_flags[IN_alarm]	InAlarm
Action_Text	PossibleValues/WritableValues
Active_Text, Inactive_Text	PossibleValues/WritableValues

Figure 3.7 Some recommended mappings of BACnet object properties to BACnet/WS node attributes.

3.4.4 Future Improvements

3.4.4.1 Updated BacNet/WS Annex N: An ATOM Interface

As this book was going to press, Addendum 2010 a.m. to BACnet was about to be released. It contains a major update of the BACnet/WS interface specification. The major features are:

- The adoption of the ATOM Publishing protocol (defined by IETF RFC 5023) for the REST version of the interface.
- The adoption of the PubSubHubbub subscription model for data push services. PubSubHubbub was defined as a simple subscribe/notify extension to ATOM and RSS. See <http://code.google.com/p/pubsubhubbub/> for more details.
- A comprehensive XML representation of BACnet structures (Annex Q), the control system modeling language (CSML).

With this new version, BACnet/WS ceases to be a simplified interface serving limited purposes: it really becomes a fully functional interface that provides access to the full functionality of BACnet.

3.4.4.2 Profile Names

Work is ongoing within the “Applications” (AP) working group, to investigate the development of “macro” object types suitable for various application areas. This group is studying a proposal for a “BACnet modeling language” that is expected to provide a machine-readable way of representing the capabilities of individual BACnet devices such as services and objects supported. A new standard object property called “Profile_Name” now allows the extension of standard (or proprietary) object types in such a way that devices with knowledge of the named profile are able to interpret the extended properties. This is expected to form the basis for convenient BACnet interfaces to other protocols that support object-oriented representation of their functionality, for example, LonWorks.

4

The LonWorks[®] Control Networking Platform

The LonWorks series of networking protocols were developed by Echelon[®] Corporation for the needs of control and automation applications and are now managed by the LonMark[®] International trade group. The LonWorks platform was initially developed in an effort to move away from the proprietary centralized control model, where a central controller receives all measurements from remote sensors and sends all commands to remote actuators. In an effort to eliminate the controller as a single failure point and increase the efficiency and power of control systems, the LonWorks platform introduced a concept of “connection” enabling devices to exchange data directly, using a subscribe/notify model.

At the physical layer, the LonWorks platform is media independent; including media types for copper pairs (wires) and power lines, radio, infrared light, and optical fiber.

The LonWorks platform is one of the most popular protocols for building and industrial automation, claiming over 90 million installed devices.

4.1 Standardization

In 2008, LonWorks also was approved as ISO standards: ISO/IEC 14 908-1, -2, -3, and -4 for the protocol, twisted-pair channel, power-line channel, and IP-tunneled channel, respectively.

4.1.1 *United States of America*

The communication protocol (a.k.a., the LonTalk[®] protocol; Echelon’s trade name) was submitted to ANSI in 1999 and accepted as a standard for control networking

(ANSI/CEA-709.1; originally EIA-709.1). Shortly after, the power line and twisted-pair physical layers were accepted as part of the ANSI standard series.

4.1.2 Europe

The European Committee for Standardization (CEN) standardized the protocol for “buildings” use in 2005. Then in 2007, the LonWorks platform became part of the Application Interworking Specification (AIS) recognized by the European Committee of Domestic Equipment Manufacturers (<http://www.ceced.eu/>) for Household Appliances Control and Monitoring.

4.1.3 China

In China the LonWorks platform is both a national standard in the category of “controls” (GB/Z 20 177.1-2006) and in the category of “buildings” (GB/T 20 299.4-2006).

4.2 Technology

Figure 4.1 shows the structure of a LonWorks packet, including data fields used by each protocol layer. The protocol layers are detailed in the following sections.

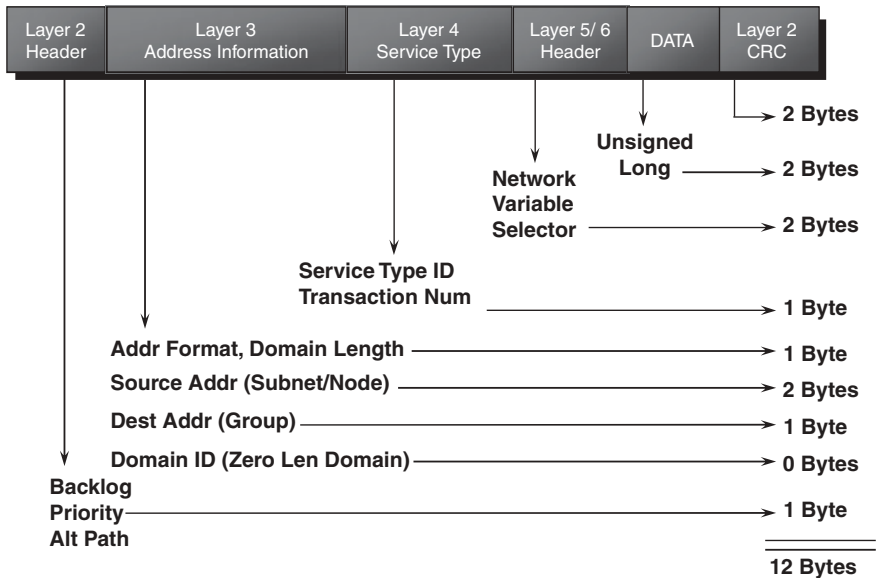


Figure 4.1 A typical ISO/IEC 14 908-1 Packet. Reproduced by permission of © Echelon Corporation.

4.2.1 Physical Layer

The LonWorks protocol is media independent, and assumes only a physical layer that can transmit binary signals, called a channel. Specific transceivers are required for each underlying physical layer. The series of standards defines transceivers for twisted pair, link power, power line, radio frequency, optical fiber, coaxial cable, and infrared media channels (Figure 4.2, the complete list can be found at <http://www.lonmark.org/spid>). Most of the transceiver channels use differential Manchester encoding, where each “1”

Channel Name	Media	Bit rate and capacity	Reference
FO-20L/ FO-20S	Fiber optic	1.25 Mbps	ANSI/CEA-709.4
IP-852	LonWorks over IP	Over 10000 packets per second	ANSI/CEA-852
PL-20 (L-N)/ PL-20 (L-E)	126 kHz or 140 kHz BPSK, line to neutral or line to earth	5 kbps	
PL-20A	CENELEC A-band power line (75 kHz or 86 kHz BPSK)	2613 bps (about 11 packets per second)	ANSI/CEA-709.2
PL-20C/ PL-20N	CENELEC C-band Power line (115 or 132 kHz BPSK) with/without access protocol	156.3 k/3987 bps (about 18 (20C) to 20 (20N) packets per second)	ANSI/CEA-709.2
TP/FT-10	Free topology twisted pair (can be branched at any point) Max wire length : about 500m (900 to 2700m with bus termination). Devices per link : 64 (when power provided by twisted pair), 128 (if devices have a separate device power source)	78.13 kbps [(about 180 packets per second, peaks up to 225 packets per second)	ANSI/CEA-709.3
TP/RS485-39	RS-485 Twisted Pair	39.06 kbps	EIA/TIA-232-E
TP/XF-1250	Transformer-Isolated Twisted Pair	1.25 Mbps(about 576 packets per second, peaks up to 720 packets per second)	LONMARK® Interoperability Guidelines

Figure 4.2 LonWorks media channels (most common in bold). Reproduced by permission of © Echelon Corporation.

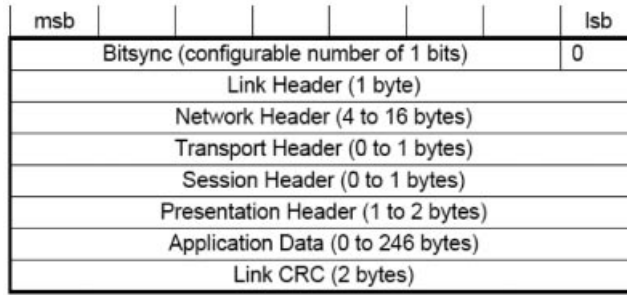


Figure 4.3 A typical LonWorks data frame, least significant bits are transmitted first. Reproduced by permission of © Echelon Corporation.

is transmitted as a polarity reversal for a full period, and each “0” is represented as two polarity reversals during a single, full period. This type of encoding ensures that there is no continuous component in the transmission (it averages to 0 regardless of the information transmitted), and that connections – particularly those using two wires – that not need to care about polarity.

Each physical communication link may be interconnected by means of a LonWorks router, or extended by means of a physical layer repeater. Channels connected by a repeater form a *segment*.

4.2.2 Link Layer

The protocol’s link layer provides cyclical redundancy check (CRC) error checking in order to detect most transmission errors; an access, collision avoidance and priority mechanism; and a data-frame format (Figure 4.3).

4.2.2.1 Access and Priority Mechanism

The media access control (MAC) algorithm employed is carrier-sense, multiple access (CSMA), in a variant called p-persistent CSMA: A LonWorks networking device is required to establish that the transmission medium is idle before it can start communicating (this is common to all CSMA protocols). In addition, in order to reduce the probability of collisions, it will begin to transmit, with probability p , in one of $1/p$ predefined time slots (called *beta-2* slots, during typically from 2 to 30 bit times). The number of time slots is dynamically adjusted based on the network load: with more time slots (smaller p), the network works better during high loads, but this adds to the transmission delay compared to fewer time slots. Each LonWorks networking packet includes the number of acknowledgments expected as a result of sending this packet, which allows receiving devices to estimate the upcoming network load and adjust the number of *beta-2* slots accordingly. Adjustments are made as multiples of 16 ($n \times 16$), where n is called the

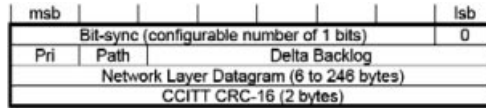


Figure 4.4 Details of the link layer header. Reproduced by permission of © Echelon Corporation.

current transmission-channel *backlog*. The required increment is indicated in the link header delta backlog field (Figure 4.4).

Some transceivers can send over two channels for redundancy purposes; the desired channel for a packet is indicated by the *Path* bit of the link header (Figure 4.4).

On each channel, a fixed number of the first *beta-2* time slots (up to 127 time slots) can be allocated to priority packets. Devices can send both priority packets and nonpriority packets. The *Pri* bit (Figure 4.4) of the link layer header indicates whether the packet is a priority packet.

4.2.3 Network Layer

The network layer provides the message-delivery mechanisms.

Each device (“node”) is identified by a unique 48-bit identifier, called the unique node identifier (UID) or the unique_node_ID, within device memory structures. It is also, colloquially and historically, known as the neuron ID, or *neuronID*. The UID does not change over the lifetime of the device. It is normally used only when the device is first inserted in the network, before it has been assigned a logical network address. This facilitates the replacement of a device by a new device of the same type, which will have a different UID but will be assigned the same logical network address as the replaced device. The UID is also utilized for applications requiring authenticated messaging service (for higher-security needs).

The protocol uses hierarchical addressing, and defines the *domain* (0, 1, 3, or 6 bytes), *subnet* (8 bits), and *node* (7 bits) subaddresses. Each device is assigned a unique *nodeID* in each subnet. Therefore there may be up to 32 385 devices (255 subnets × 127 nodes) per domain. The devices of a single domain or subnet may be on various channels; and devices from multiple domains may coexist on the same channel.

The source of each message is contained in the address field of the header, and specifies the sending node *subnetID* and *GroupID* (first two bytes). The target of a message may be, depending on the header *Addr* field value:

- A single node: the header comprises a 2-byte destination address specifying the *subnetID* and *nodeID* (header *Addr* format=2) or a 7-byte address specifying the destination *subnetID* and *neuronID* (header *Addr* format field=3).
- All devices in a subnet (3-byte address, header *Addr* format field=0).
- All devices in a domain (3-byte address, *subnetID*=0, header *Addr* format field=0).

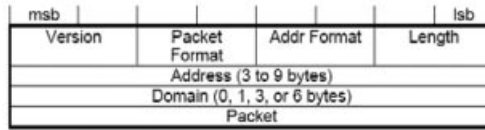


Figure 4.5 The network header format. Reproduced by permission of © Echelon Corporation.

- A group: the protocol defines group addresses (2-byte *domainID*, and 1-byte *groupID*, header Addr format header field=1), so there may be up to 256 groups per domain. Such addresses may be used to address groups of devices on different subnets. There is a maximum of 64 devices per group for acknowledged device-to-device messaging services and no limit for unacknowledged messaging services.

The packet format field specifies whether the packet is a transport packet (packet format field value = 0), a session packet (1), an authenticated packet (2) or a presentation packet (3).

4.2.4 Transport Layer

The protocol's transport layer provides the end-to-end reliability mechanisms. The protocol offers four basic types of messaging service, depending on the desired tradeoff between reliability and efficiency:

- *Acknowledged* (header transport packet format field = 0, see Figure 4.6): messages are sent in the context of a transaction identified by a *transactionID*. Each receiver sends an acknowledgment message (header transport packet format field = 2) with the *transactionID*. If not all acknowledgements have been received (until a configurable timeout), the message is retransmitted with the same transaction ID.
- *Request/response*: the request/response service is managed by the session layer.
- *Repeated* (header transport packet format field = 1): each message is repeated several times so that the probability of a device failing to receive one of the messages is reduced. However, the target devices do not acknowledge these messages, so the service is not fully reliable. Echelon Corporation estimates that 3 repeats results in a successful delivery probability greater than 99.999%. This service is useful for group addressing to large groups.

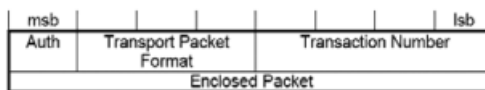


Figure 4.6 Transport layer header details. Reproduced by permission of © Echelon Corporation.

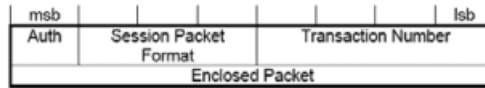


Figure 4.7 The session layer header format. Reproduced by permission of © Echelon Corporation.

- *Unacknowledged*: the message is sent as “best effort” only and the sender is not notified if the message is lost *en route*. This service is useful for periodic data reporting from sensors.

The device can select any of the mechanisms listed above to transport its presentation layer messages.

For authentication to work (Auth bit in the transport header set to 1, see Figure 4.6), a 48-bit key (one per domain) must be configured in each device sending or receiving authenticated messages. The authenticated message is sent as a normal message from A to B but before acting on this message, B will authenticate the message by challenging A. The challenge includes a random number and A is expected to reply with a hash (encrypted encoding) of the secret domain key and the challenge. B computes the same hash locally and compares the result with the challenge response of A. If they are identical, B successfully acknowledges the original message.

4.2.5 Session Layer

The session layer replaces the transport layer when the packet format field of the network layer is set to 1 (Figure 4.5). It offers authentication (see transport layer) and a request/response service.

Like the acknowledged messaging service, the request/response service is useful when a message is sent to a device or group of devices and individual responses are required from each receiver. The request message (session packet format=0, see Figure 4.7) may either be resent until all responses (session packet format=2) have been received (the transaction number provides the acknowledgment mechanism), or it may be duplicated several times to minimize the risk of packet loss. The responses from a request/response transaction, unlike the simple, low-level acknowledgment from an acknowledged messaging service, usually include application-level response data.

The enclosed packet data are formatted as a presentation-layer message.

4.2.6 Presentation Layer

4.2.6.1 Presentation-Layer Messages

The presentation layer defines the data-interpretation conventions of the protocol: it uses *messages* that are transported and retransmitted by the lower layers. Except for specific

Table 4.1 LonWorks presentation-layer message types

Message Type	1 byte Message code	Usage
User Application Message	00-2F	Message payload includes a 6-bit message code, followed by data. The applications exchanging application messages must agree on the interpretation of the message codes.
Standard Application Message	30-3E	Same as User application messages but using the standard message codes used for standard application-layer services (data log, file transfer, and self-installation functions).
Foreign-Frame Message	40-4E	Arbitrary data, which may encapsulate other protocols.
Network Diagnostic Message	50-5F	
Network-Management Message	60-7F	
Network Variable Message	80-FF	Identifier that identifies the data as a data value (or data structure) of 1 to 31 bytes that may be shared by multiple devices on a network.

needs, most applications typically exchange data using network variable messages, except for some specific needs (file transfer, self-installation, etc.) or when there are communications requirements beyond those specified by the network variable messages.

The LonWorks networking presentation-layer messages begin with a one-byte message code that defines the type of data contained within the message (Table 4.1), followed by 0 to 277 bytes of data.

4.2.6.2 Network Variables and the LonWorks Subscribe/Notify Model

Network variables are essential interfaces of most LonWorks networking devices. Network variables (Figure 4.8) have a direction (input to receive data, output to send data), type (scalar or aggregate of several fields), length, a self-documentation string. They are identified by a *network variable index* on the device, and are identified by a 14-bit *network variable selector* (0-3FFF) over the network (maintained in a configuration table on each device). A single network variable index can be associated to several network variable selectors on the same device; in which case, one is called the primary network variable

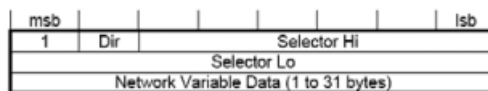


Figure 4.8 Format of a LonWorks network variable message (carried by the transport/session layer). Reproduced by permission of © Echelon Corporation.

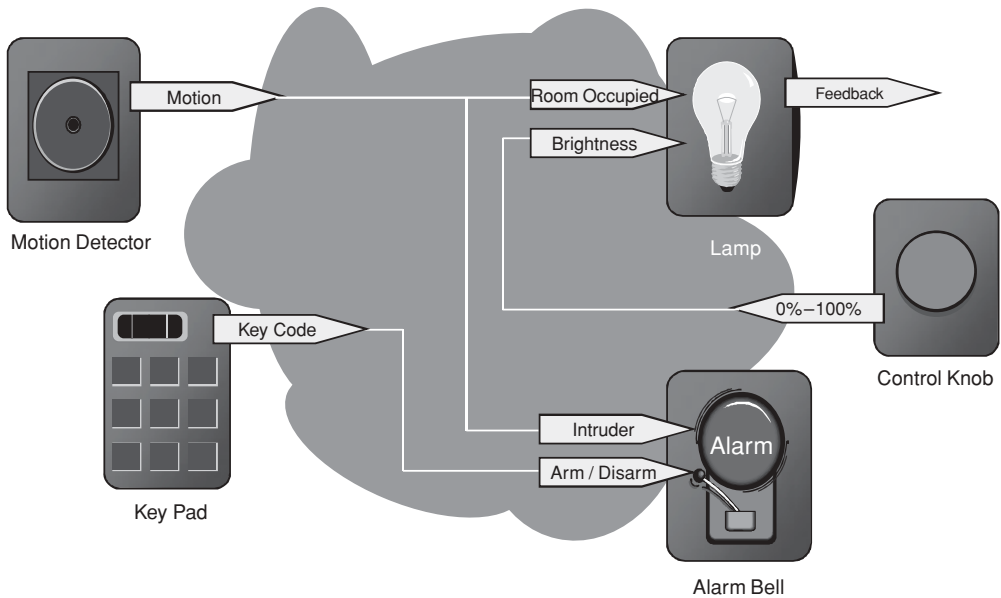


Figure 4.9 Example of devices exchanging information through connections. Reproduced by permission of © Echelon Corporation.

selector and the others are aliases. Network variables' values are exchanged over the network by network variable messages (Figure 4.10).

The LonWorks protocol provides two primitives to poll network variables over a network: one of which can poll multiple devices simultaneously using the network variable selector and one targeted to a single device using the network variable index.

The protocol also has a native subscribe/notify model – a key feature that supports the claim that the networking platform facilitates networks designed without central points of failure. Network variables belonging to different devices can be connected if they have the same type and length, for example, a state output (type: switch-type) of a switch device may be connected to the switch-type input of a lamp. There are several types of connections:

- Unicast: single output to single input;
- Multicast out: single output to multiple inputs on several devices;
- Multicast in: multiple outputs on several devices to a single input.

Connections are created by a process called *binding* and are performed by a network-management tool or by the self-installation process of the device. When binding network variables, the protocol implementation of the device is configured with:

- The list of addresses of the other devices (or groups of devices) in the network expecting that network variable's value;
- The target network variable selectors.

SNVT_temp_p	
Overview:	
Temperature (degrees Celsius).	
Details:	
Standard:	<i>yes</i>
Resource Set:	<i>Standard 00:00:00:00:00:00:00:00</i>
Index:	<i>105</i>
Obsolete:	<i>no</i>
Size:	<i>2</i>
Programmatic Name:	<i>SNVT_temp_p</i>
Neuron C Type:	<i>signed long</i>
	<i>Minimum: -27317</i>
	<i>Maximum: 32767</i>
	<i>Invalid: 32767</i>
	<i>Scaling (A,B,C): 1, -2, 0</i>
	<i>Scaled value: $1 * 10^{-2} * (Raw+0)$</i>
	<i>Resolution: 0.01</i>

Figure 4.10 Example specification of a LonMark standard network-variable type (SNVT_temp_p). Reproduced by permission of © Echelon Corporation.

The application on the device will only update the network variable, and the protocol implementation will ensure that it is sent to all configured target addresses (domain/subnet/node or group, using the appropriate network variable selector). See Figure 4.9 for an example network configuration showing several connections using network variables.

The network variable selector values 3000 to 3FFF hex are reserved for unbound network variables, with the selector value equal to 3FFF hex minus the network variable index. Selector values 0 to 2FFF hex are available for bound network variables. This provides a total of 12 288 network variable selectors for bound network variables. Each device can have up to 8192 network-variable aliases and 4096 of those bindable network aliases.

Standard network variable types (SNVTs, pronounced “SNIV-its”), specify standard data encodings (units, range, resolution, scaling, data structure, etc.) covering most

common usage cases. The list of SNVTs includes over 200 types and covers a wide range of applications. The complete list is available at types.lonmark.org.

If an application requires a network variable type that is not a SNVT, device manufacturers can define custom network-variable types. These are called *user network variable types (UNVTs)*.

4.2.7 Application Layer

The network configuration and network-diagnostic services are defined by the protocol standard. The following list summarizes the standard application-layer services. Additional standard application-layer services (Figure 4.11) are published at www.lonmark.org

SFPTsccVAV 

[Root](#) [Print](#) [Contact](#) [LonMark](#)

Space Comfort Controller (SCC) - Variable Air Volume. Type of HVAC unit controller that provides temperature control for a space within a building

Network Variables	Configuration Properties
nviSpaceTemp	nciSetpoints
nvoSpaceTemp	nciSndHrtBt
nvoUnitStatus	
nviAirflow	nciBypassTime
nviAirFlowSetpt	nciDuctArea
nviApplicMode	nciDuctAreaHeat
nviAuxHeatEnable	nciFanOperation
nviComprEnable	nciFlowGain
nviCoolPriSlave	nciFlowGainHeat
nviCoolSecSlave	nciFlowOffset
nviCoolSrcTemp	nciHvacType
nviEconEnable	nciLocation
nviEmergOverride	nciManualTime
nviEnergyHoldOff	nciMaxFlow
nviFanSpeedCmd	nciMaxFlowHeat
nviFlowOverride	nciMinFlow
nviHeatCool	nciMinFlowHeat
nviHeatPriSlave	nciMinFlowStdby
nviHeatSecSlave	nciMinFlowUnitHt
nviHeatSrcTemp	nciMinOutTm
nviMinAirFlow	nciMinFlwStdbyHt
nviMinAirFlowHt	nciNomFlow
nviOAMinPos	nciNomFlowHeat
nviOccManCmd	nciNumDampers
nviOccSchedule	nciNumValve
nviOccSensor	nciOAMinPos
nviOutdoorDewPt	nciRcvHrtBt
nviOutdoorRH	nciSatTime
nviOutdoorTemp	nciSpaceCO2Lim
	nciSpaceRHSetpt

Figure 4.11 Example specification of a LonMark functional profile: the space comfort controller – variable air-volume controller. Reproduced by permission of © Echelon Corporation.

- **Network configuration** – configuration of the network attributes of a device (network address and binding information for the device’s network variables).
- **Network diagnostics** – diagnostics commands.
- **File transfer** – the largest practical amount of data that can be transferred in a single packet is 228 bytes, but the LW-FTP file-transfer method transfers data using a stream of 32-byte packets.
- **Application configuration** – provides a standard interface to configure the behavior of a device. The interface is based on configurable data values called *configuration properties*. Standard configuration property types (SCPTs, pronounced “SKIP-its”), are the configuration equivalent to SNVTs. An up-to-date list of SCPTs can be found at types.lonmark.org.
- **Application specification** – documentation of a device’s functions as a set of function blocks (a distinct set of complementary network variables and configuration properties).
- **Application diagnostics** – standard testing primitives for function blocks and devices.
- **Application management** – standard primitives to enable, disable, and override function blocks on a device.
- **Alarming** – standard primitives for a device to report alarm conditions.

SCPTminSendTime	
Overview:	
Minimum send time. The minimum period of time between consecutive transmissions of the current value	
Details:	
Standard:	yes
Resource Set:	Standard 00:00:00:00:00:00:00:00-0
Index:	52
Obsolete:	no
Size:	2
Programmatic Name:	SCPTminSendTime
Default:	0.0
Neuron C Type:	SNVT time sec

Figure 4.12 Example definition of a standard configuration-property type (SCPTminSendTime). Reproduced by permission of © Echelon Corporation.

- **Scheduling** – standard primitives for scheduling events based on time of day, day of week, and date.
- **Time and date management** – standard primitives for synchronizing the time-of-day and date for devices within a network.

Some application-layer services are defined for device developers through a committee of those developers within LonMark International, the trade association in support of the protocol standards. The culmination of those application-specific functional definitions are known as *functional profiles*, or simply *profiles* (Figure 4.11 and 4.12), and are implemented in part or in their entirety by developers as *function blocks* within the device.

4.3 Web Services Interface for LonWorks Networks: Echelon SmartServer

The Echelon SmartServer is an Internet gateway and local computing platform (e.g., edge control node) that connects LonWorks, ModBus, M-Bus, local I/O to other devices and networks, in addition to providing a SOAP Web Service-based interface for data access and configuration. Besides using SOAP (a W3C recommendation), the SmartServer interface is not formally standardized.

The SmartServer provides access to LonWorks networking devices through *data points*, which contain a value, data type, and format properties. The gateway implements the following functions that can be used to directly access data points: **DataPointRead**, **DataPointWrite**. For each write, it is possible to specify whether the value must be immediately propagated to the LonWorks network or not. This makes it possible to set several items in a structure sequentially and to propagate all values to the network at once; in an atomic way.

The SmartServer gateway also provides a set of applications accessible through SOAP (Figure 4.13).

The *DataServer* application can be used to create, manage, delete, and access data points. The *Datalogger* can sample and store data-point values in logs and circular buffers. The *AlarmGenerator* generates alarms – for instance, when certain limits of data-point values are exceeded – while the *AlarmNotifier* logs alarm conditions, sends notifications via SMTP e-mail, or updates specific data points. The *EventScheduler* and *EventCalendar* applications can be used for periodic updates of data points. The *TypeTranslator* can be utilized to translate values of data points with a specific variable type into a different type.

4.4 A REST Interface for LonWorks

As we have seen in the previous section, a web services interface already exists for LonWorks. However, the new trend in M2M architecture is to use a REST model, where

DataServer List Get Set Delete Read Write ResetPriority	Datalogger: List Get Set Read Clear Delete	AlarmGenerator: List Get Set Delete	AlarmNotifier: List Get Set Delete Read Write Clear
AnalogFB: List Get Set Delete	EventScheduler: List Get Set Delete	EventCalendar: List Get Set Delete	TypeTranslator: List Get Set Delete RuleList RuleGet RuleSet RuleDelete

Figure 4.13 SmartServer applications and functions. Reproduced by permission of © Echelon Corporation.

the types of interactions are restricted to only the CRUD¹ verbs. This requires a specific design of the representation of the underlying M2M network or device.

Echelon Corporation published a first version of a REST interface for LonWorks in July 2010. In this version 1.0, the “LonBridge Proxy Server REST API” supports the basic CRUD REST interactions, but did not introduce yet a subscribe/notify model: this type of interaction is feasible in a REST architecture but requires both interface sides to act as client and server. Standardized subscribe/notify REST models have been introduced for instance by ETSI TC M2M (refer to Chapter 14 this book): it is expected that ETSI M2M interfaces to LonWorks will also exist in the near future, and introduce the additional interactions and features made possible by the ETSI TC M2M REST architecture.

4.4.1 LonBridge REST Transactions

4.4.2 Requests

The LonBridge API supports the following HTTP request methods:

- GET: retrieve resource data from the LonBridge server;
- POST: create a new resource;

¹ Create, read, update, delete.

- PUT: update an existing resource managed by the LonBridge proxy server;
- DELETE: delete a resource.

The specifications for each resource describe how the commands are applied.

4.4.3 Responses

Responses include a response body and a status code.

The response body may be formatted as JSON, XML, HTTP, or text. The default is JSON. The response format may be specified as a suffix to the URL – for example **GET server/api/devices.xml** returns a list of all devices in XML format. The response format may also be specified in the accept header.

The status code is a standard HTTP status code. Typical status codes include the following:

- 200 – OK (standard response for successful request);
- 201 – Created (standard response after successfully creating a resource);
- 400 – Bad Request (request has invalid syntax or cannot be fulfilled);
- 404 – Not Found (requested resource could not be found but may be available in the future);
- 500 – Internal Server Error (generic error message when other messages don't apply);
- 501 – Not Implemented (request not recognized).

4.4.4 LonBridge REST Resources

The various LonBridge REST resources made available by the API are regrouped in seven functional groups addressing the following domains: network, device, device type, connections, groups and measures.

4.4.4.1 Network

These resources allow to access or modify the main LonWorks network parameters.

Syntax	http://server[:port]/api/network/{resource} [?params]
Methods	PUT, GET
Resources	Network resources: name, domainId, domainLength, key

4.4.4.2 Devices

Device resources are used to retrieve and update resources on an individual device or a set of devices. The device ID is the LonBridge device ID, which is the letter “o” followed by an identifier, for example: **o0**, **o1**, or **o2**.

Syntax	http://server[:port]/api/devices[/]{id}[/{resource[=value]}][/?params]
Methods	PUT, GET, DELETE
Resources	<p>Device resources: name, type, location, scene, active, and data points defined per device type:</p> <ul style="list-style-type: none"> • blinds: angle, level, motion, scene • dimmer (Lamp Module) resources: brightness, state (on, off), power, energy, scene • switch (Appliance Module) resources: state (on, off), power, energy, scene • occupancy: occupied • thermostat: fan (auto, on), humidity, mode (auto, heat, cool, off), schedule, setback, setpoint, temperature, message, pricing <p>Device parameters: startDate (default current date; specified as <i>day[-month[-year]]</i> where <i>month</i> defaults to current month and <i>year</i> defaults to current year), startTime (default current time; 24-h time), interval (default 60 minutes), maxCount (default 100), and deviceType (default all). When a startDate or startTime parameter is specified, up to maxCount records may be returned. Each record includes a timestamp in “<i>year-month-day hour:minute:second</i>” format, for example: 2010-07-05 15:43:10.</p>

Examples

GET server.com/api/devices – returns list of all devices. The following is an example XML encoded response body:

```
<devices>
  <o0 type="switch" name="Appliance Module 1"
    brand="Echelon" active="true" state="on"
    power="27" energy="8.5913" />
  <o1 type="dimmer" name="Lamp Module 1" brand="Echelon"
    active="true" brightness="93" state="on" power="58"
    energy="0.5219" />
</devices>
```

The following is an example JSON encoded response body:

```
{
  "o0": {
    "type": "switch",
    "name": "Appliance Module 1",
    "brand": "Echelon",
    "active": "true",
    "state": "on",
    "power": 27,
    "energy": 8.5913
  },
  "o1": {
    "type": "dimmer",
    "name": "Lamp Module 1",
    "brand": "Echelon",
    "active": "true",
    "brightness": 93,
    "state": "on",
    "power": 58,
    "energy": 0.5219
  }
}
```

GET `server.com/api/devices?deviceType="switch"` – returns a list of all switch devices. This corresponds to the LonBridge `<get TBD />` command.

GET `server.com/api/devices/o2` – returns a list of all resources defined for device o2. This corresponds to the LonBridge `<o2.get/>` command.

GET `server.com/api/devices/o2/power` – returns the last power-consumption reading for device o2. This corresponds to the LonBridge `<o2.get select="state"/>` command.

PUT `server.com/api/devices/o2/state` – turns on device o2 on or off (the state is sent in the request body). This corresponds to the LonBridge `<o2.set state="on"/>` command.

DELETE `server.com/api/devices/o2` – deletes device o2. This corresponds to the LonBridge `<o2.delete/>` command.

4.4.4.3 Device Types

Lists devices by device type.

Syntax	<code>http://server[:port]/api/{deviceType}/{id}/{resource}[?params]</code>
Methods	PUT, GET, DELETE
Resources	Device resources: same as for devices . Device parameters: same as device parameters.

4.4.4.4 Connections

Syntax	<code>http://server[:port]/api/connections/{id}/{resource}[?params]</code>
Methods	PUT, GET, POST, DELETE
Resources	Connection resources: state and setting .

4.4.4.5 Scenes

Syntax	http://server[:port]/api/scenes/{id}/{resource}[?params]
Methods	GET, POST
Resources	Scene resources: state and setting .

4.4.4.6 Groups

Syntax	http://server[:port]/api/devices/{id}/groups/{id}/{resource} [?params] http://server[:port]/api/groups/{id}/{resource}[?params]
Methods	PUT, GET, DELETE
Resources	Group resources for devices: membership (true or false) . Group resources: state .

4.4.4.7 Measures

Syntax	http://server[:port]/api/devices/{id}/groups/{id}/{resource} [?params] http://server[:port]/api/groups/{id}/{resource}[?params]
Methods	GET, POST (define new measure), DELETE (delete measure)
Resources	Measure resources are used to retrieve aggregate calculations for current and historical data.

Examples

GET server.com/api/measures – returns list of all measures.

GET server.com/api/measures/energy – returns aggregate energy usage.

GET server.com/api/measures/energy?startDate=1 – returns up to 100 aggregate energy usage historical values since the first of the month.

GET server.com/api/measures/energy?category=lighting – returns aggregate energy usage for lighting devices.

GET server.com / api / measures / energy?category = lighting&location = “Living Room” – returns aggregate energy usage for lighting devices in the living room location.

GET server.com / api / measures / energy?category = lighting&location = “Living Room”&startDate=1 – returns up to 100 aggregate energy usage historical values for lighting devices in the living room location.

5

ModBus

5.1 Introduction

Many protocols have been designed for the needs of industrial automation and metering. These protocols generally use simple query/response models and allow for extremely simple implementations. Many protocols derived from the frame formats defined by IEC 870-5 such as:

- T101 (IEC 870-5-101) that was generated by the IEC TC57 for electric utility communication between master stations and remote terminal units, it is also based on the IEC-870-5-x link layer, using frame format FT 1.2.
- DNP 3.0, a protocol originally designed by Westronic, Inc. that was released into the public domain in 1993, based on the IEC-870-5-x link layer with a few modifications (e.g., use of FT3 frames for asynchronous, rather than synchronous, communication, inclusion of both source and destination addresses).
- M-Bus (see Section 9.3)
- Profibus, a fieldbus initially designed by Siemens and later standardized as IEC 61 158 (“Digital Data Communication for Measurement and control, Fieldbus for use in industrial control systems” for versions DP-V0, DP-V1 and DP-V2) and IEC 61 784 (Communication Profile Family DPF3). The protocols user’s association website is <http://www.profibus.com/>.

Other protocols developed independently into *de-facto* standards, such as ModBus, a very common protocol that is used in many industrial and HVAC installations.

5.2 ModBus Standardization

ModBus is a trademark of Modicon inc (Schneider Electric group), which also maintains the standard. The ModBus standard specification over a serial line can be found at http://www.modbus.org/docs/Modbus_over_serial_line_V1_02.pdf.

ModBus is an application layer messaging protocol that provides client/server communication between devices connected on different types of buses or networks. Because of its simplicity, ModBus has become one of the *de-facto* standards for industrial serial-message-based communications since 1979.

ModBus typically runs on top of RS 232, RS 442 point to point or RS 485 point to multipoint links. The ModBus/TCP specification, published in 1999 defines an IP-based link layer for ModBus frames.

ModBus devices communicate using a master-slave model: one device, the master, can initiate transactions (called *queries*), which can address individual slaves or be broadcast to all slaves. The slaves take action as specified by the query, or return the requested data to the master.

5.3 ModBus Message Framing and Transmission Modes

The transmission mode defines the framing and bit encoding of the messages to be transmitted on the ModBus network. In a given ModBus network, all nodes must use the same mode and serial parameters:

- In the *ASCII Transmission Mode*, each byte is encoded on the serial link as 2 ASCII characters. Each ASCII character is sent separately as 1 start bit, 7 data bits, zero or one parity bit, one or two stop bits. The message is framed by a starting “:” ASCII byte, and ends with a “CR-LF” byte sequence (see Figure 5.1).
- In the *RTU (remote terminal unit) transmission mode*, the message is transmitted in a continuous stream. Each 8-bit byte is framed by 1 start bit, 8 data bits, zero or one parity bit, one or two stop bits. The message itself starts after a silent period of at least 3.5 character times.

ModBus Addresses: ModBus messages begin by the target 8-bit address that can take any decimal value between 1 and 247. 0 is used for broadcasts. The address field of the message frame contains two characters in ASCII mode, or 8 bits in RTU Mode. Each query contains the address of a specific slave. When it responds, the slave includes its own address in the message.

ModBus Functions: The function code field contains two characters in ASCII mode, and 8 bits in RTU mode, which can take any decimal value between 1 and 255 and are selected based on the device application profile. Some example functions are listed:

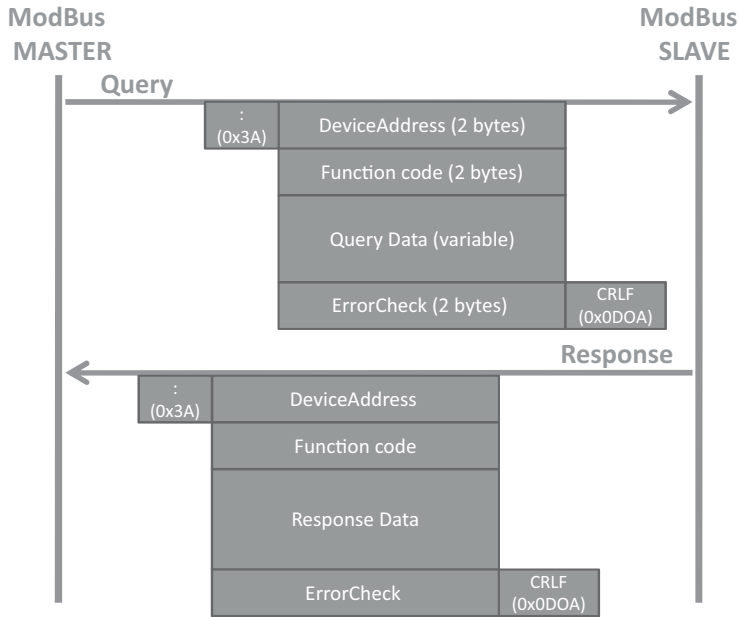


Figure 5.1 ModBus message framing (ASCII mode).

- 0x02: Read Input Status. Parameters: starting register address, and number of consecutive addresses to read. Response data: 1 bit per input read.
- 0x11: Report Slave ID. Parameters: none. Response data: slave ID, run indicator, device specific data.

ModBus Data Field: The data field provides the application level information, as required by the ModBus function. When a given ModBus function requires variable size data, the data field begins with the “byte count” of the data.

5.4 ModBus/TCP

The ModBus/TCP specification can be found at http://www.eecs.umich.edu/~modbus/documents/Open_ModbusTCP_Standard.doc

ModBus/TCP provides TCP/IP access to the ModBus functionality. Each ModBus Request/response is sent over a TCP connection established between the master and the slave, using well-known port 502. The TCP connection may be reused for several query/response exchanges.

The byte content of the ModBus request and response frames (i.e. without framing start-stop-parity bits specific to the serial physical layer) is simply transported over the TCP

connection, in big indian order. The only addition of ModBusTCP is to add a seven-byte message prefix:

```
ref ref 00 00 00 len unit
```

The ref bytes are simply copied by the slave from the request, and may be used as a handle by the master. The length information in the message prefix allows proper reassembly of the ModBus message when it has been segmented in several IP packets. The slave address has been renamed “unit identifier” and is contained in unit. The rest of the message conforms to the regular ModBus structure, but the error check fields may be omitted for obvious reasons.

6

KNX

6.1 The Konnex/KNX Association

The Konnex (or KNX) Association was set up in 1999 on the merger between three former European associations promoting intelligent homes and buildings:

- Batibus Club International (BCI France) promoting the Batibus system;
- The European Installation Bus Association (EIBA) promoting the EIB system;
- European Home Systems Association (Holland) promoting the EHS system.

The goal of the KNX Association was to define and offer certification services for the KNX open standard, while offering legacy support and certification services for Batibus, EIB¹ and EHS. Membership is limited to manufacturers, there are over 200 members from 22 countries as of 2010, including ABB, Agilent, Bosch, Electrolux, Hager, Legrand, Merten, Moeller, Schneider, Siemens and many more leading vendors of home and building automation equipment.

KNX technology is royalty free for KNX members.

6.2 Standardization

In order to standardize the specifications, the KNX association cooperates with CENELEC TC 205. The KNX protocol has become an international standard in Europe as EN 50 090 (media and management procedures), EN 13 321-1 (media) and EN 13 321-2 (CEN, KNXnet/IP).

¹ EIB is backward compatible to KNX, most devices can be labeled both with the KNX as well as the EIB logo.

At an international level, KNX is standardized by ISO and IEC (ISO/IEC 14 543-3). KNX is in prestandard stage in China as GB/Z 20 965.

There are several versions of KNX, which are all backwards compatible. The current version is 2.0 (since August 2009).

The overall specification counts over 6000 pages, divided into 10 volumes.² Individual specification documents can be purchased from the KNX association. The KonCert group manages KNX certification and testing.

Gateway specifications exist between BACnet (ISO 484 Annex 5 H.5 mapping KNX and BACnet, see also Chapter 3), DALI (lighting control) and KNX.

6.3 KNX Technology Overview

The overall KNX architecture is documented in Vol 3, part 3/1. The KNX architecture is decentralized: nodes can interact with other nodes without the need for a central controller.

The protocol stack uses the OSI model with a null session and presentation layer. It is based on the original work of EIB, which is therefore backward compatible to KNX.

KNX standardizes the protocol, but also the data model (EN 50 090-3-3, KNX volume 3/7) for basic types (integer and float values, percentage) and common device functions such as switching, dimming, blinds control, HVAC and so on . . .

6.3.1 Physical Layer

The physical layer of KNX is specified in Vol 3, Chapter 3/3/1 of the specifications. KNX can use a variety of physical layers

- **TP1³: Twisted pair** (Chapter 3/2/2). TP was the first physical layer that was defined as part of EIB, and is still the dominant physical layer used in KNX deployments. The TP bus provides both power and communication, using inductive coupling (Figure 6.2). A twisted-pair installation is made of lines, each line is composed of up to 4 line segments interconnected by repeaters, and each segment interconnects up to 64 devices. Lines are interconnected by line couplers (LC). The line couplers interconnect to the KNX backbone via a backbone controller (BbC), and the devices that can be accessed via a given BbC are part of the same KNX area (or zone). Line couplers and backbone controllers act as routers, that is, filter the messages that they relay based on

² Vol 1 Primer (deprecated), Vol 2 Cookbook (deprecated), Vol 3 System specifications, Vol 4 Hardware requirements (link to relevant standards, e.g., safety and environmental requirement), Vol 5 Certification manual, Vol 6 profiles, Vol 7 application descriptions (actual functional profiles), Vol 8 conformance testing, Vol 9 Basic and system Components (physical couplers, bus interface modules and couplers), Vol 10 Specific standards (Extended Tag format).

³ There was a TP0 defined, which was deprecated by KNX.

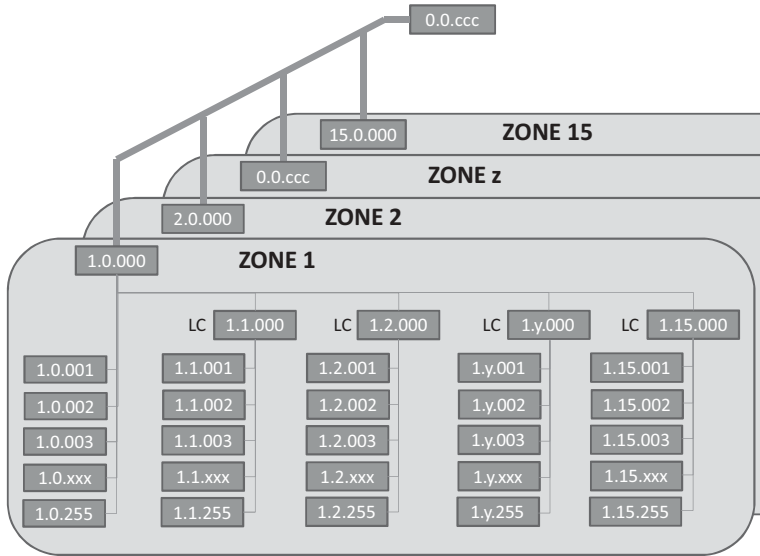


Figure 6.1 KNX network topology.

the destination address and the domain id (when present). The address space allows up to 15 areas (Figure 6.1), each with 15 lines, and a KNX TP installation can manage a maximum of 61 249 devices.

The physical layer uses a CSMA/CA medium access control using inductive coupling (Figure 6.2), and performs error detection (horizontal or vertical parity checks with acknowledgements (Ack, Nack, busy). The bus provides a 21 to 29 V power supply, and low-power KNX nodes may draw power from the TP line (typically up to 150 mW).

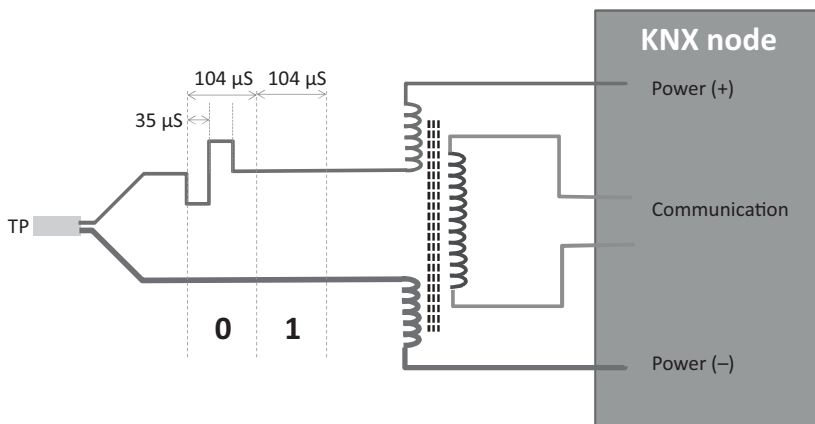


Figure 6.2 KNX TP node, modulation and inductive coupling.

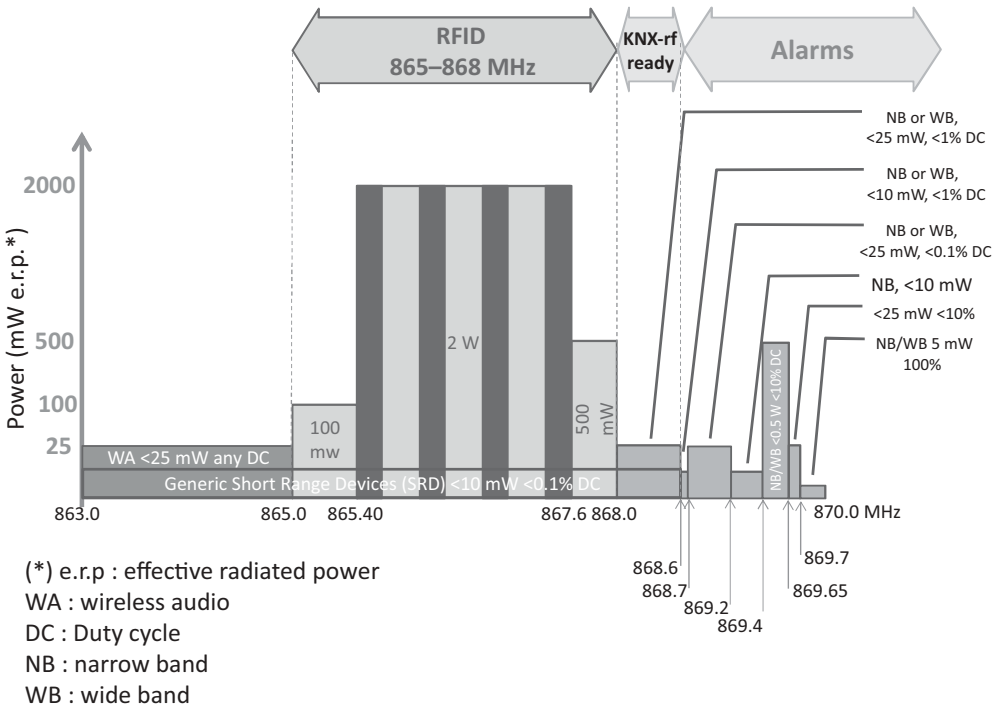


Figure 6.3 Usage of the 863–870 MHz band in Europe.

The transmission begins with a start bit (0), followed by a application octet, a parity bit, a stop bit (1) and a mandatory pause (11). The theoretical throughput is 9600 bps.

- **PL110 Over PLC⁴** (Chapter 3/2/3). PL110 uses a FSK modulation scheme and was also part of the original EIB specification. Each PLC line can have up to 64 devices. Since PLC is inherently a broadcast open media, the separation of domains (the portion of the KNX network logical topology over which the data signals of one physical layer type propagate) is ensured by a 48-bit domain address, in addition of the zone/line/node Id address (see TP1 for a description of these addresses).
- **Over RF** (Chapter 3/2/5 defined in 2001). This physical layer uses the 868-870 MHz band (Figure 6.3).

The KNX-rf 1.1 specification was updated in 2010, introducing a “push button” and easy controller mode setup specification, and using a 1% duty cycle on the center frequency 868.3 MHz: this version is called “KNX-rf ready”. It allows bidirectional communication with low duty cycle devices by sending a 4.8-ms preamble for transmissions. Devices are preconfigured with group addresses for multicast communication, and unicast communication uses an “extended group address” composed of the group address of the sender and its serial number.

⁴There was a PLC 132, which was removed from the specification.

Time slot	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Fast1	█				█				█					█				█				█					█	
Fast2			█													█												
Fast3							█													█								
Slow1											█	█																
Slow2																								█	█			

Figure 6.4 KNX-rf multireceiver scanning.

A further KNX-rf update is expected in 2011 called “KNX-rf multi” that introduces 3 RF “fast channels” and 2 RF “slow channels” (for battery-powered devices and energy-harvesting products) as well as fast acknowledge services. Fast channel 1 would use the existing KNX 1.1 center frequency (868.3 MHz) and a duty cycle of 1%, it is the configuration channel and the default call channel. Fast channel 2 would use a center frequency of 868.95 MHz and a duty cycle of 0-1%, Fast channel 3 (optional, coexisting with slow channel 1) would use a center frequency of 869.85 MHz and a duty cycle up to 100%. Slow channels use a preamble length of 500 ms and duty cycle of 10%. Slow channel 2 will use a center frequency of 869.525 MHz. KNX-rf “multi” products will not be compatible with KNX-rf 1.1. products, but will be compatible with KNX ready products. Because of the preamble, a KNX-rf multireceiver can scan both fast and slow channels (Figure 6.4).

- **Over IP** (Chapter 3/2/6). KNX may use IP as a native communication medium. KNXnet/IP uses binary or XML encoded PDUs to emulate bus communication. KNX/IP defines a tunneling mechanism over IP, using datalink layer binary PDUs.

6.3.2 Data Link and Routing Layers, Addressing

The data link layer is specified in Vol 6, Chapter 3/3/2, and the routing layer is specified in Chapter 3/3/3.

KNX nodes exchange telegrams, formatted as in Figure 6.5. Telegrams are transmitted octet by octet on the physical layer. Each telegram is acknowledged by the recipient node after a mandatory pause (equivalent to 11 bits), and retransmitted, if needed, up to 3 times. On TP, a 9 octet command will be transmitted and acknowledged (when unicast) in about 15 ms.

Each KNX node has a unique 2 byte address, used mainly for configuration purposes and as the source of telegrams. The source address is encoded on 16 bits as an area identifier (4 bits), a line identifier (4 bits) and a device number (8 bits). The source address

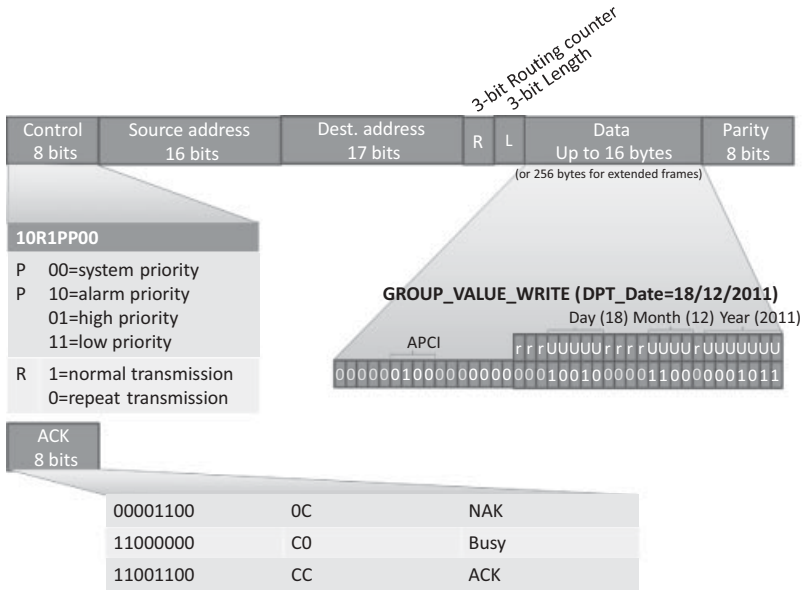


Figure 6.5 KNX telegram format.

appearing in a telegram is always a physical address. This address is configured during the installation process, typically by pressing a key on the device that causes it to enter a signaled state that allows the KNX ETS installation tool to assign an address to the device. Device number 0 is reserved for line couplers and backbone couplers.

The destination address may be a logical group address (bit 17 = 1) or a physical address (bit 17 = 0). The group address may be formatted on 2 levels (4-bit main group ID, 11-bit secondary group ID) or 3 levels (4-bit main group ID, 3-bit auxiliary group ID, 8-bit secondary group ID). The group destination address is the abstraction for a command wire: a KNX switch “connects” to a KNX lamp controller by sending a message to the group that is configured as the lamp controller input, the message acts as a data container for values that the sending nodes wants to share with the receiving node(s). Such group bindings are defined during the configuration of KNX nodes. Special group address 0000 h addresses all KNX nodes.

The KNX installation has a 64 k addressing space. The individual address and all group addresses of a device are stored in the device “address table”.

KNX supports 4 priority levels encoded as specific values of the control bits (see Figure 6.5). In the case of collisions among same priority nodes, the node with the lowest physical address transmits first. On the TP1 physical layer, this priority mechanism derives from the fact that “0” is dominant when a simultaneous 0/1 transmission collision occurs.

The 3-bit routing counter implements a “hop count”: a line coupler decides if a telegram can be transmitted to the other side based on the remaining hop count. The number of hops is normally limited to 6.

An 8-bit parity field secures the KNX telegram, it is based on a vertical parity scheme: bit i of the security field is set to 1 if the octet-wise sum of bits i of the previous telegram data is an even number.

6.3.3 *Transport Layer*

The KNX transport layer is specified in Volume 3, Chapter 3/4.

The transport layer (TL) provides a connection-oriented peer to peer communication service, providing a connect and disconnect primitive, a TL-acknowledgment, sequence counter and time-out management (typically 6 s for the configuration mode).

The transport layer also removes the source of the message before calling the application layer, and therefore the behavior of actuators never depends on the source of messages.

6.3.4 *Application Layer*

The KNX application layer is defined in Vol 3, Section 3/3/7. The application layer defines the group objects, and the exchange of group-object values via service requests, for instance “group value write” illustrated in Figure 6.5.

The application layer also defines the “property value write” service, which is used to set values and configuration parameters to KNX device interface objects.

6.3.5 *KNX Devices, Functional Blocks and Interworking*

Volume 3, part 7 is dedicated to interworking. The overall interworking model is outlined in 3/7/1. Chapter 3/4 defines the application environment, and Chapter 3/4/1 defines the application interface layer, including data representation models for group objects and interface objects. Such as data structures or flags (e.g., transmission allowed, write allowed, data has been written by the bus . . .).

KNX devices are defined by the functional blocks that they support. A functional block is a logical grouping of inputs, outputs and parameters that are useful to perform a certain function. For instance the “sunblind actuator basic” is one of the functional blocks defined by KNX, which specifies:

- A list of mandatory and optional inputs (Move UpDown, StopStep UpDown, Set Absolute Position blinds Percentage, WindAlarm . . .).
- A list of mandatory and optional outputs (Info Move UpDown, Current Absolute Position Blinds Length, . . .).
- A list of mandatory and optional parameters (Reversion Pause time, Move Up/Down time, Preset Slat angle, . . .).

Inputs, outputs and parameters are specified by their datapoint types (see Section 6.3.5.3, and a specific functional interpretation of the Datapoint values in the specific context of

the functional block. Inputs, outputs and parameters may be published as properties or group objects.

KNX volume 6 provides an extensive library of functional blocks (over 146 as of 2011) such as dimming controllers, room temperature controllers, schedulers, system clock, alarm, and so on.

6.3.5.1 Group Objects

The process information of KNX functional blocks (input, output or parameter) may be published as a group object (GO). A group object may be read or written over the bus via dedicated multicast service primitives. The KNX specification of each functional block defines which of the inputs, outputs and parameters may, or must, be published as a group object.

The type of the group object is described by a datapoint (see Section 6.3.6.). A group object that sends its value may be configured with one and only one destination group address, but a group object may listen to several group addresses. The target and monitored group addresses are configured in the KNX device *Address Table*. All group objects linked to the same group address must be of the same datapoint.

For example, if the on/off group object of a presence sensor (an output) is assigned group address 1/1, and the on/off group object of a light controller (an input) is also assigned group address 1/1, then the presence sensor will control the light controller.

6.3.5.2 Interface Objects

Interface objects (IO) store certain properties of the device, mostly parameters (Figure 6.6). A node can have up to 256 interface objects. The type of each interface object is identified

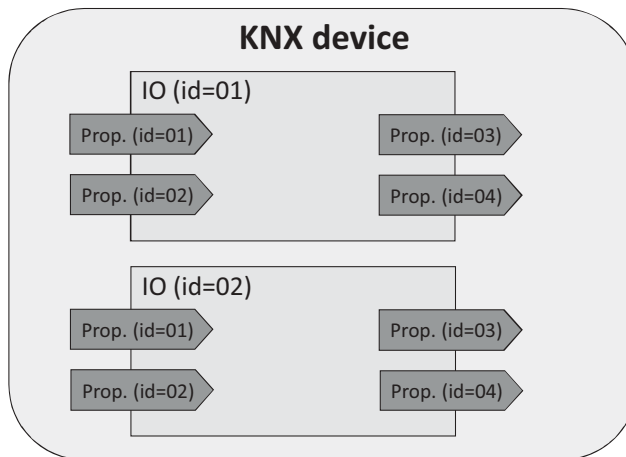


Figure 6.6 KNX interface objects and properties.

by a 16-bit identifier, the type of property is given by an 8-bit identifier. Chapter 3/7/3 contains the standard identifier tables.

The KNX application layer provides unicast primitives to read or set property values, using messages addressed to the physical address of the device.

The value of a group object can be reflected in the value of a property, in order to make it possible to read or set a property via group communication.

6.3.5.3 KNX Datapoints

Chapter 3/7/2 specifies the datapoint types. KNX provides an extensive library of datapoints (over 350 as of 2011), which are used to express the properties of KNX devices, and parameters of commands sent over the network (see Figure 6.5). Datapoints are defined by:

- Their data type (format and encoding);
- Their dimension (range and unit).

Each datapoint type is identified by a 16-bit main number.16-bit subnumber identifier. The main number identifies the format and encoding, the subnumber identifies the range and unit. Subnumbers are allocated by the KNX association based on the application domain: 0 to 99 for common use range and units, 100 to 499 for HVAC applications, 500 to 599 for load management, 600 to 999 for lighting, 1000 to 1999 for system applications. Subnumbers greater than 60 000 are used by manufacturer-specific extensions.

KNX has defined its own syntax to define Datapoints types, with a letter representing the data type of each field (e.g., unsigned value, see Figure 6.7), and a subscript number indicating the number of bits used to encode the data type. For instance, U₈ is an unsigned

A	Character
A[n]	String of <i>n</i> characters
B	Boolean/Bit set
C	Control
E	Exponent
F	Floating point value
M	Mantissa
N	eNumeration
R	Reserved bit or field
S	Sign
U	Unsigned value
V	2's Complement signed value
Z ₈	Standardized Status/Command B8. Encoding as in DPT_StatusGen

Figure 6.7 KNX Datapoint field definition symbols.

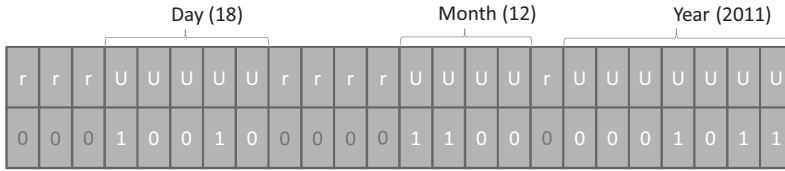


Figure 6.8 DPT_Date KNX datapoint.

number field encoded over 8 bits. DPT_Date is encoded as $r_3N_5r_4N_4r_1U_7$, and illustrated on Figure 6.8.

For metering applications, KNX has defined a number of datapoints aimed at tunneling M-Bus addresses and metering values, in order to facilitate interworking. See Section 9.3 for more details.

The library of datapoint types may be downloaded from the KNX Association’s website for free.

Type information is used mainly at configuration time: it is not transmitted for better performance and to avoid imposing unnecessary restrictions on the combinations of devices.

6.4 Device Configuration

Device configuration uses mainly point to point (unicast) telegrams. There are three options to configure a node:

- In the system mode or “S-mode”, the management configuration tool runs on a PC (using KNX ETS™ software, which stores configuration data as XML schema).
- In easy mode or “E-mode”, several strategies are employed to avoid the use of a PC to configure the network. An embedded “master controller” may be activated to search partner devices, and connects to further devices one by one (identify/discover device). Group bindings may then be configured via a controller menu, or in “push button” mode. In the “push button” mode, links are configured one by one by first activating the actuator (“push-button”), then the sensor to be enrolled.
- Devices may also be preconfigured using a logical tag (extended possibility specified in Volume 10). The devices are preconfigured and use a specific framing format (EFF extended frame format), in addition of the standard format. This mode introduces semantical and geographical zoning tags. Configuration tuning is performed with ETS.

7

ZigBee

7.1 Development of the Standard

The 802.15.4 standard provides a physical and link layer technology optimized for low bitrate, low duty cycle applications. However, in practice sensor and control applications also need a mesh networking layer, and a standard syntax for application layer messages. In 2002, several companies decided to form the ZigBee alliance to build the missing standard layers that would be required to enable a multivendor mesh network on top of 802.15.4 radio links.

In 2008, the ZigBee alliance counted more than 200 members:

- *Promoter* members get early access to, contribute to and vote on the specifications of the alliance. They can veto decisions made by other participants in the alliance and get special marketing exposure in ZigBee events. New candidates for the promoter status must get co-opted by existing promoter members.
- *Participant* members have the same contribution and voting rights as promoters, but without veto rights.
- *Adopters* also get early access at the specifications, but can contribute only to the application profile working groups, and do not have voting rights.

The ZigBee alliance regularly organizes interop events, called ZigFests, and organize a developers conference twice a year. In order to ensure interoperability across vendors, the use of the ZigBee Compliant Platform (ZCP) certification and logo is reserved for products passing the ZigBee test suite, which includes interoperability tests with the “Golden units” (stacks from four reference implementations: Freescale, Texas Instruments, Ember, and Integration).

The deployment of many telecom standards either failed or was slowed down by multiple patent claims, many of which were not disclosed during the design phase of the

standard. While the ZigBee alliance can do nothing against potential patent claims coming from nonmembers, it did verify that no technology included in the standard was subject of a known patent. In addition, every new member of the ZigBee alliance must sign a disclosure statement regarding patents that could potentially apply to ZigBee technology.

There are several versions of ZigBee. The current versions of ZigBee are ZigBee 2006/2007 (stack profile 0x01, ZigBee 2007 adds optional frequency agility and fragmentation), and ZigBee Pro (stack profile 0x02) that adds support for more nodes and more hops through source routing (it does not support tree routing), multicasting, symmetric links and a high security level. There was a ZigBee 2004 version, which is now deprecated.

7.2 ZigBee Architecture

7.2.1 *ZigBee and 802.15.4*

ZigBee sits on top of 802.15.4 physical (PHY) and medium-access control (MAC) layers, which provide the functionality of the OSI physical and link layers.

So far ZigBee uses only the 2003 version of 802.15.4. All existing ZigBee commercial devices use the 2.4 GHz S-Band as the 2003 version of 802.15.4 does not allow sufficient bandwidth on other frequencies. The 2006 version adds improved data-transfer rates for 868 MHz and 900 MHz but is not yet part of the ZigBee specification.

802.15.4 offers 16 channels on the 2.4 GHz, numbered 11 to 26. ZigBee uses only the nonbeacon-enabled mode of 802.15.4, therefore all nodes use CSMA/CA to access the network, and there is no option to reserve bandwidth or to access the network deterministically. ZigBee restricts PAN IDs to the 0x0000 – 0x3FFF range, a subset of the 802.15.4 PAN ID range (0x0000-0xFFFFE).

All unicast ZigBee commands request a hop by hop acknowledge (optional in 802.15.4), except for broadcast messages.

7.2.2 *ZigBee Protocol Layers*

The ZigBee network layer provides the functionality of the OSI network layer, adding the missing mesh routing protocol to 802.15.4. It also encapsulates the network formation primitives of the 802.15.4 MAC layer (network forming and joining).

The rest of the ZigBee protocol layers (Figure 7.1) do not follow the OSI model:

- The **Application Support Sublayer (APS)** layer has several functions:
 - Multiplexing/demultiplexing: it forwards the network layer messages to the appropriate application objects, according to their endpoint ID (each application is allocated an endpoint ID).
 - Binding: the APS layer maintains the local binding table, that is, records remote nodes and endpoints which have registered to receive messages from a local endpoint.
 - 64-bit IEEE to 16-bit ZigBee network node address mapping.

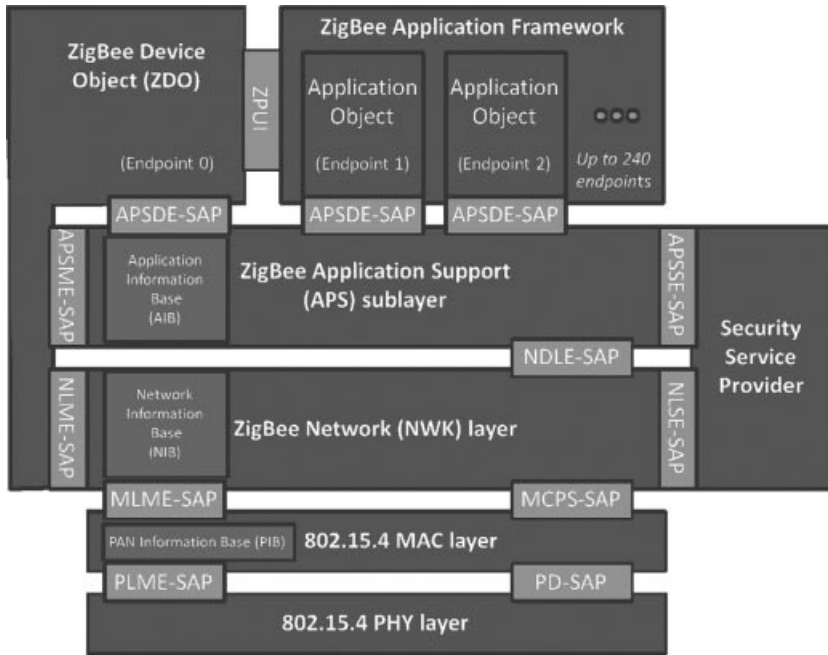


Figure 7.1 ZigBee architecture overview.

- Management of end to end acknowledgements. The application layer supports acknowledgements independently of the link layer acknowledgements of 802.14.4. The APS manages retries and duplicate filtering as required, simplifying application programming.
 - Fragmentation.
- Also, as part of the application support sublayer management entity, or APSME:
- Group addressing: the APSME allows to configure the group membership tables of each endpoint ID, and forwards messages addressed to a group ID to the application objects with relevant endpoint IDs.
 - Security: management of keys.
- The **ZigBee Device Object (ZDO)** layer is a specific application running on endpoint 0, designed to manage the state of the ZigBee node. The ZDO application implements the interfaces defined by the ZigBee device profile (ZDP, application profile ID 0x0000). These primitives encapsulate the 802.15.4 network formation primitives of the ZigBee network layer (node discovery, network joining), as well as additional primitives supporting the concept of binding (see Section 7.5.2.2).
 - The **ZigBee Cluster Library (ZCL)** was a late addition to ZigBee, specified in a separate document. It consists in a library of interface specifications (cluster commands and attributes) that can be used in public and private application profiles. It is now

considered as one of the key assets of ZigBee: while the ongoing evolution of ZigBee towards a 6LoWPAN-based networking layer is likely to replace the original networking layers of ZigBee, the ZCL is likely to remain the “lingua franca” of application developers. One important addition of the ZCL is the group cluster, which provides the network interface for group formation and management.

- The **Application Framework** layer provides the API environment of ZigBee application developers, and is specific of each ZigBee stack. Each application is assigned an Endpoint ID.

The interfaces of ZigBee layers are called “service access points” (SAP), as in 802.15.4. One interface, the layer management entity ([layer name]-ME) is responsible for configuring internal data of the layer. Another interface, the data entity ([layer name]-DE), provides the data send/receive and other nonmanagement primitives.

7.2.3 ZigBee Node Types

The ZigBee node types listed below are not mutually exclusive. A given device could implement some application locally (e.g., a ZigBee power plug) acting as a ZigBee End Device, and also be a ZigBee router and even a ZigBee coordinator.

- **ZigBee End-Device (ZED)**: this node type corresponds to the 802.15.4 reduced function device. It is a node with a low duty cycle (i.e. usually in a sleep state and not permanently listening), designed for battery operation. ZEDs must join a network through a router node, which is their parent.
- **ZigBee router (ZR)**: this node type corresponds to the 802.15.4 full function device (FFD). ZigBee routers are permanently listening devices that act as packet routers, once they have joined an existing ZigBee network.
- **ZigBee Coordinator (ZC)**: this node type corresponds to a 802.15.4 full function device (FFD) having a capability to form a network and become a 802.15.4 PAN coordinator. ZigBee coordinators can form a network, or join an existing network (in which case they become simple ZigBee routers). In nonbeacon-enabled 802.15.4 networks, coordinators are permanently listening devices that act as routers, and send beacons only when requested by a broadcast beacon request command.

The ZigBee coordinator also contains the trust center, which is responsible for admission of new nodes on the network and management of security keys (see Section 7.7).

7.3 Association

7.3.1 Forming a Network

When forming a network, a ZigBee coordinator first performs an active scan (it sends beacon requests) on all channels defined in its configuration files. It then selects the

channels with the fewer networks, and if there is a tie performs a passive scan to determine the quietest channel. It finally broadcasts a 802.15.4 beacon for the selected PAN ID on the selected radio channel, then remains silent (or repeats the beacon periodically, depending on the implementation). Depending on the configuration of the stack, the scan duration on each channel can range from 31 ms to several minutes, so the network-forming process can take significant time. If there are any ZigBee routers associated to the network, they will typically repeat the beacon with an offset in time relative to their parent's beacon (an extension of 802.15.4:2003).

The ZigBee specification allocates range 0x0000 to 0x3FFF for PAN IDs (a subset of the range defined by 802.15.4: 0x0000 to 0xFFFE). The PAN-ID should be unique for a given channel for networks not capable of dynamic channel change (ZigBee 2006), and unique on all channels if channel agility is enabled (ZigBee 2007, ZigBee PRO). A ZigBee coordinator beacon may also include an extended PAN ID (64 bit EPID), in addition of the 16-bit 802.15.4 PAN identifier, in order to facilitate vendor specific network selection for joining nodes. This EPID identifier is only used in the beacon frames and has no other uses, while the 16-bit 802.15.4 PAN identifier is always used for joining and addressing purposes.

7.3.2 *Joining a Parent Node in a Network Using 802.15.4 Association*

ZigBee devices that are not yet associated either capture by chance the beacon, or try to locate a network by broadcasting a 802.15.4 beacon request on each of the 16 radio channels (active scan, see Figure 7.2), unless the radio channel has been preconfigured or determined in the application profile. If a coordinator has formed a network on one of those channels, it responds to the beacon request by broadcasting a 802.15.4 beacon, which specifies the 16-bit PAN ID of the network, the address of the coordinator in short 16-bit format or extended 64-bit format, and optionally an extended PAN ID (EPID). Any ZigBee router that has already joined the network will also respond with a beacon if they hear the beacon request.

The ZigBee payload of the 802.15.4 beacon also contains the ZigBee stack profile supported by the network, a flag indicating whether the responding node has remaining capacity for routers or end devices joining as new children, and the device depth of the sending device, that is, its level in the parent/child tree rooted at the coordinator.

Once it has discovered the PAN ID of the network, its radio channel, and the address of a router or coordinator within radio reach, the new ZigBee node sends a standard 802.15.4 *association request* command to the address of the specific parent node it wants to join as a child node (0x01 profile nodes must join the node with the smallest device depth). The association request message uses the extended 64-bit address of the joining node as the source address. Devices may wish to join a specific PAN ID, or may use a special PAN ID value 0xFFFF to signal that they are willing to join any PAN ID.

The parent node acknowledges the command, and then if it accepts the association responds with a 802.15.4 association response command sent to the extended address of the device. The association response specifies the 16 bit short address that the device

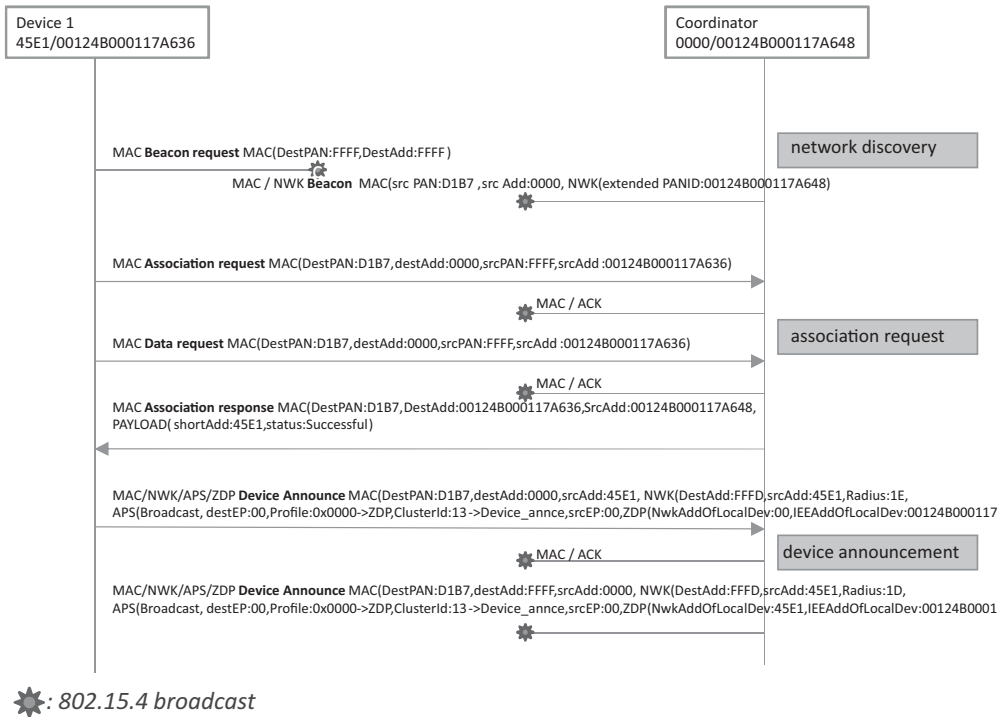


Figure 7.2 End device 1 joins the ZigBee network.

should use in the future (in order to save transmission time and therefore energy). The association response is acknowledged by the device. When the joining device is a sleeping node (`RxOnIdle=false`), the association response is not sent immediately, but stored until the sleeping nodes polls it using a “data request”, as in the example of Figure 7.2.

Once associated, ZigBee devices usually send a data request command (now using the short 16-bit address assigned by the parent as source address) to its parent in order to receive any pending configuration data. After waiting for a response, battery-powered ZigBee end devices go to sleep until the next scheduled wake-up time or interrupt.

In the example of Figure 7.2, the joining device is within radio range of the coordinator. In a more general case, broadcast and unicast messages will be relayed by one or more ZigBee routers, and a new joining device may join the network from any location accessible through mesh networking.

ZigBee joining, at the lowest level of security, only uses the “permit joining” flag of the beacon: nodes can join a network only when this flag is set in the beacon response. At the application level, most implementations allow administrators to “permit joining” for a limited amount of time, after which the network will not accept further joins. If the device is allowed to join and the `nwkSecurityLevel` parameter is set to `0x00`, then the node becomes a new child of the parent node with relationship type `0x01` (child), otherwise set

as type 0x05 (unauthenticated child). When security is enabled, interactions with the trust center (see Section 11.2.1) follow the unauthenticated joining process for key distribution.

In order to facilitate commissioning, nodes may implement the *commissioning ZDP cluster* (see Section 7.6) to preconfigure security material and other parameters, and reset the node. Some nodes may be set up to join any network with permit-join enabled, or may be preconfigured to join the well-known commissioning network with extended PAN ID 0x00f0c27710000000.

In theory, up to 31 100 nodes (9330 routers) can join a given network in stack profile 0x01, and over 64 000 nodes in stack profile 0x02.

7.3.3 Using NWK Rejoin

A device that loses connection to the network can attempt to rejoin using the ZigBee NWK layer rejoin command, which also triggers a beacon request. Since the NWK layer rejoin command use NWK layer security, the difference from a join based on 802.15.4 association is that no additional authentication step needs to be performed when security is enabled, and that nodes may rejoin any parent as long as it has available capacity, regardless of the status of the accept joining flag of the beacon. If it rejoins a different parent (e.g., because the original parent no longer responds), the node will be allocated a different short address, and must broadcast a device announce to the network in order to update bindings that may be configured in other nodes (see Figure 7.2).

After power cycles, most implementations do not immediately attempt an explicit rejoin in order to avoid network overloads, if they still have the address of their parent node and their own short address in nonvolatile memory. It is assumed that all nodes will restart in the same state as before the power cycle. An explicit rejoin is triggered only if the node fails to communicate with its parent. Such a procedure is often referred to as “silent rejoin”. It is also the default procedure, in ZigBee Pro/2007, when the coordinator triggers a channel change (annex A).

7.4 The ZigBee Network Layer

The network layer is required for multihop routing of data packets in the mesh network, and is one of the key missing elements of 802.15.4. ZigBee uses the AODV public-domain mesh algorithm. The ZigBee network layer uses a specific data frame format, documented in Table 7.1, which is inserted at the beginning of the 802.15.4 payload.

7.4.1 Short-Address Allocation

ZigBee uses the 0x0000 – 0xFFFF7 range for network node short addresses. The ZigBee coordinator uses short address 0x0000. The allocation of other network

addresses, under control of the ZigBee Coordinator, depends on the routing technology in use:

ZigBee supports two address allocation modes:

- In stack profile 0x01, the network address depends on the position of the node in the tree. The distributed address assignment mechanism uses CSkip, a tree-based network address partition scheme designed to provide every potential parent with a subblock of network addresses. In addition to the default meshed routing, a tree-based routing can be used as a back-up (routers use the address allocation to decide whether to forward the packet to a parent or to a child).
- In stack profile 0x02 (ZigBee 2007, ZigBee Pro), a stochastic address assignment mechanism is used and ZigBee provides address-conflict detection and resolution mechanisms.

7.4.2 Network Layer Frame Format

The network layer PDU format is illustrated in Table 7.1, and is transported as 802.15.4 payload (see Chapter 1).

Table 7.1 The ZigBee network layer frame format

Field name	Size (octets)	Field details
Frame Control	2	-----XX : Frame type (00 : network data) -----0010-- : Protocol version (always 0x02 for ZigBee 2006/2007/Pro) -----XX----- : Route discovery (0x01:enable) -----X----- : Multicast (0 : unicast) -----X----- : Security (0 : disabled) ----X----- : Source route (0 : not present) ---X----- : Destination IEEE address (0 : not specified) --X----- : Source IEEE address (0 : not specified) 000----- : Reserved
Dest. Address	2 or 8	0xffff broadcast to all nodes including sleeping devices 0xffffd broadcast to all awake devices (RxOnIdle = True) 0xffffc broadcast only to routers, not to sleeping devices
Source address	2 or 8	
Radius	1	Maximum number of hops allowed for this packet
Sequence number	1	Rolling counter
Payload	Variable	APS data, or network layer commands

7.4.3 Packet Forwarding

At the network layer, ZigBee packets can be:

- *Unicast*: the message is sent to the 16-bit address of the destination node
- *Broadcast*: if broadcast address 0xFFFF is used, the message is sent to all network nodes. If broadcast address 0xFFFD is used, the message is sent to all nonsleeping nodes. If broadcast address 0xFFFC is used, the message is broadcast to routers only (including the ZigBee coordinator). A radius parameter adjusts the number of hops that each broadcast message may travel. The number of simultaneous broadcasts in a ZigBee network is limited by the size of the broadcast transaction table (BTT), which requires an entry for each broadcast in progress. The minimal size of the BTT is specified in ZigBee application profiles, for example, 9 for HA.
- *Multicast* that is, sent to a 16-bit group ID.

ZigBee unicast packets are always acknowledged hop by hop (this is optional in 802.15.4). Broadcast packets are not acknowledged and are usually retransmitted several times by the ZigBee stacks of the originating node and by ZigBee routers on the path. Note that broadcast messages and messages sent to group IDs are not always broadcast at the link layer level: since sleeping nodes do not receive such messages, after wake up they send a data request to their parent node, and the queued messages are sent as unicast messages specifically to the sleeping node (in which case they are acknowledged at the link-layer level). In most implementations, the messages are queued only 7 to 10 seconds in the parent node (ZigBee specifies 7 seconds) so sleeping nodes should wake up at least once every 6 seconds. However in practice several vendors manufacture sleeping nodes with wake up periods of up to 5 minutes . . . in this case applications need to be prepared to resend commands every 6 seconds until the target node wakes up.

Typically, a ZigBee node forwards a packet in about 10 ms. The propagation of data packets through the meshed network is limited by the initial value of Radius, a hop counter decremented at each routing node.

Delivery of packets to sleeping nodes uses the IEEE “Data request” packet. Parent nodes buffer received packets for their sleeping children. When it wakes up, the sleeping child sends a IEEE “Data request” packet to its parent. If it has data pending, the parent sets a specific “more data” flag in the ACK response, then the pending data.

7.4.4 Routing Support Primitives

The network layer provides a number of command frames listed in Table 7.2.

The route request command enables a node to discover a route to the desired destination, and causes routers to update their routing tables. At the MAC level, the route request command is sent to the broadcast address (0xffff) and the current destination PAN ID.

The response, if any, is a route reply command that causes routers on the path to update their routing tables.

Table 7.2 ZigBee routing layer primitives

Command Frame Identifier	Command Name Reference
0x01	Route request
0x02	Route reply
0x03	Network Status
0x04	Leave
0x05	Route record
0x06	Rejoin request
0x07	Rejoin response
0x08	Link status
0x09	Network report
0x0a	Network update

7.4.5 Routing Algorithms

7.4.5.1 Broadcast, Groupcast, Multicast

At the 802.15.4 level, messages sent to multiple destinations are always broadcast. At the network layer, however, ZigBee offers more possibilities, depending on the destination address:

- 0xffff broadcast to all nodes including sleeping devices;
- 0xfffd broadcast to all awake devices (RxOnIdle = True);
- 0xfffc broadcast only to routers, not to end devices.

In order to avoid 802.15.4 collisions, broadcast packets are relayed after a random delay of about 100 ms and therefore propagate ten times more slowly than unicast messages. The radius parameter is decremented at each hop, so the broadcast propagation can be controlled with the initial radius value (see Table 7.1).

ZigBee also uses a broadcast transaction table (BTT) in order to avoid any looping of broadcast messages: each broadcast packet is uniquely identified by its source address and network sequence number. When relaying a broadcast message, routers keep a copy of this unique identifier for 9 s (broadcast timeout), and will drop any looped packet. If the BTT is full, all broadcast messages are dropped. Routers that do not hear all neighbor routers retransmit a broadcast message may retransmit the broadcast message, implementing a form of implicit acknowledge mechanism.

Groupcast and multicast are implemented by the APS layer:

- APS messages sent to group addresses are filtered by the APS layer of the receiving node, so that only endpoints (and all of them) of member nodes will receive the message. However, all nodes receive the message (destination set to 0xffff at the network layer)

- In ZigBee Pro, the radius is not decremented when the message is forwarded by a group member. This makes it possible to restrict the 802.15.4 broadcast propagation to group members only, allowing some slack (`apsNonmemberRadius`) in order to cope with disconnected groups. ZigBee Pro calls this “multicast”.

7.4.5.2 Neighbor Routing

This mode is not formally documented in ZigBee, but most vendors use it. If a router R already knows that the destination of a packet is a neighbor router or a child device of R, it can send the packet directly to this node. ZigBee end devices, however, must always route outgoing packets to their parent.

7.4.5.3 Meshed Routing

This is the default routing model of ZigBee. It implements the advanced *ad-hoc* on-demand distance vectoring (AODV) algorithm.

The principle of AODV is illustrated on Figure 7.3.

Node A needs to set up a route to node D. It broadcasts a route request (see Table 7.3) to network address 0xffff (routers only), which propagates through the network. Each ZigBee router that receives that message forwards it to its neighbors, adding their local estimation of quality of the link over which they received the route request to the path cost parameter of the route request. Note that the route we are discovering is in the A to D direction, while the path costs actually used are in the D to A direction. This is because the sender of the route request, which is broadcasting the message, cannot transmit different values of the link cost, therefore the receiving node needs to update the path cost. *ZigBee mesh routing assumes symmetrical link quality.*

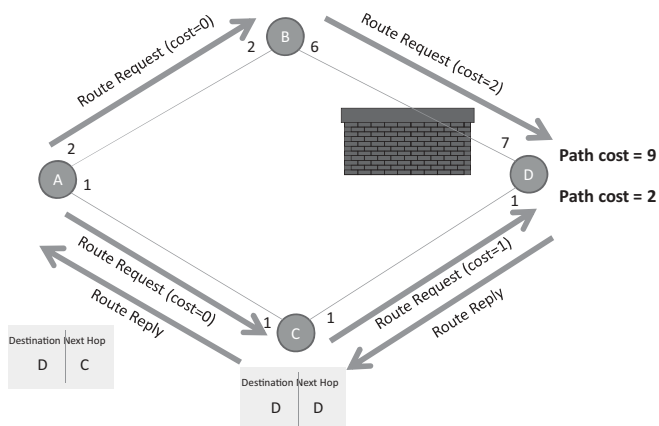


Figure 7.3 ZigBee mesh route discovery.

Table 7.3 Route request parameters

Command options	Route request identifier	Destination address	Path cost	Destination IEEE address
(1 octet)	(1 octet)	(2 octets)	(1 octet)	(0 or 8 octets)
xxx00xxx : not a many-to-one route request	Sequence number	Intended destination	Accumulator for path length as the command is propagated	Only if bit 5 of the command option is set to 1
xxx01xxx : many-to-one route request and the sender supports a route record table				
xxx10xxx : many-to-one route request and the sender does not support a route record table				
xxxxx1xx : the command frame indicates the destination IEEE address, otherwise set to 0				
xxxxxx1x : route request for a multicast group, and the destination address field contains the group ID, otherwise set to 0				

Node D will wait for a while until it believes it has received all broadcast route requests, then computes the lowest path cost and responds by a unicast route reply along the best route that was just discovered (the parameters of the response are listed in Table 7.4). Router C creates a routing table entry for D, recording the next hop to reach D (in this simple case, D itself), and node A also creates a routing table entry for D, recording C as the next hop router.

Table 7.4 Route response parameters

Command options	Route request identifier	Originator address	Responder address	Path cost	Originator IEEE address	Destination IEEE address
(1 octet)	(1 octet)	(2 octets)	(2 octets)	(1 octet)	(0 or 8 octets) Address of the originator of the route request.	(0 or 8 octets) Address of the responder.

Node A will continue to use this route as long as it works, or until the applications requests calculation of a new route.

In ZigBee 2006 and 2007, routes are unidirectional: D will need to discover a route if it needs to send a packet to A. In ZigBee Pro, D would also store the route to A, assuming symmetry. ZigBee Pro routers “ping” each other every 15 s to make sure links are indeed bidirectional, and eliminate one-way links for both directions.

7.4.5.4 Tree-Based Routing

Tree-based routing is a back-up routing mechanism (when no route exists or can be discovered) that can be used only in ZigBee 2006/2007 networks which use CSkip-based address allocation. Tree-based routing simply uses the fact that routers know the blocks of addresses allocated to their router children, and to their children ZigBee end devices: if the destination address is one of those, the router forwards the packet to the appropriate child, otherwise it propagates the packet to its parent.

Because of the limited address space of 802.15.4 (64 K addresses), which limits the size of CSkip address blocks, tree-based routing is limited to 5 parent/child relationships, each node having a maximum of 20 children and 6 routers. This limits the number of nodes to 31 101 in ZigBee 2006/2007.

7.4.5.5 Source Routing

ZigBee mesh networking is limited by the size of the routing table of routers. It works and scales well in a peer to peer environment, but in environments where one node is a preferred communication source that frequently communicates to all other nodes, then this node and adjacent routers would need to store a routing table with as many entries as nodes. This situation happens with data concentrators, used for metering applications.

ZigBee Pro solves this problem by introducing source routing:

- The concentrator broadcasts a many to one route request (up to five hops), which enables routers along the path to record the shortest path to the concentrator.
- When a node first sends a packet to the concentrator, it first sends a route record message towards the concentrator, so that the concentrator will have a chance to learn the optimal path back towards the node (assuming symmetric links).
- The concentrator can now reach any node using source-routed messages (up to 5 intermediary routers can be specified). It still needs a lot of memory, but all routers in the ZigBee network can now work with very small routing tables.

7.5 The ZigBee APS Layer

The APS layer is responsible for management and support of local applications. It defines all the concepts that make it possible to develop and interconnect ZigBee applications: endpoints, groups, bindings, and so on.

7.5.1 Endpoints, Descriptors

A given ZigBee device may implement multiple applications at the same 802.15.4 address. Clearly some multiplexing mechanism is required to identify the source and destination application of a message. This multiplexing identifier is called *endpoint* in the ZigBee specification. Think of it as the equivalent of a port number in a TCP/IP network.

Each endpoint is further characterized by a simple descriptor. The simple descriptor contains the endpoint number, application profile ID, application device ID (a 16-bit number referring to a device definition, for example, a HA thermostat, used only for informative purposes since it contains no technical data), an application version ID, the list of input clusters and the list of output clusters.

The descriptor of an endpoint can be retrieved from any other node by using the ZDP simple descriptor request command (*Simple_Desc_req*, see also Section 7.8.1).

In addition to the simple descriptor, specific to each endpoint, ZigBee devices have additional descriptors applying to the whole node:

- A node type descriptor: capabilities of the node;
- A node power descriptor: node power characteristics;
- A complex descriptor (optional): Further information about the device descriptions, described as pairs of compressed XML tag and related field data;
- A user descriptor: User-definable descriptor.

Profiles can be discovered by using the ZigBee device profile request primitive, addressed to endpoint 0 (ZDO) of the device.

7.5.2 The APS Frame

APS data frames can be sent unicast (with or without application level end to end acknowledgment, in addition to the MAC level hop per hop acknowledgment), groupcast, multicast (ZigBee 2007 and ZigBee Pro), or broadcast. Groupcasts and broadcasts are both supported by network-level broadcasts, and are not acknowledged.

At the application level, ZigBee allows application developers to use 64-bit or 16-bit addresses, group addresses or indirect addressing in order to identify the destination node, for instance in the *APSDE-DATA.request*. In all cases, the ZigBee stack resolves that address to a 16-bit node address or to a group address before transmission.

This resolution mechanism uses the APS address map. This cache stores the mapping of 64-bit IEEE addresses to 16-bit ZigBee short network addresses. It is used, for instance, to resolve binding requests (which specify only a 64-bit IEEE address) to a 16-bit address. The maintenance of this table is performed by listening to broadcast device announce commands (e.g., when a device changes location and its 16-bit address changes, see Figure 7.2 for an example).

If a node does not have a cached route to the destination, it performs a route discovery using ZDP commands `IEEE address request` and `NWK address requests`. When a frame that required an end to end acknowledgment has not been acked after 3 retries (typically a retry every 1.5 s, this delay is adjustable in most stacks), another route discovery may be performed.

7.5.2.1 Groups

A group identifier (in the range 0x0000 to 0xFFFF) is an address that can be used at APS layer level to send a message to multiple ZigBee applications residing on other nodes (see Section 7.4.5.1). Any ZigBee node can belong to up to 16 groups. An application residing on a ZigBee node on endpoint E adds itself to a group G by calling the local `AdGroupRequest` APS primitive: this function adds endpoint E to the list of local members of group G.

Messages addressed to a group are broadcast at the ZigBee network layer (the ZigBee network frame destination address is 0xffff)¹: ZigBee routers will forward a copy of the packet to any neighbor. Group messages are therefore received and processed by all nodes in the PAN network at the MAC layer, but the APS layer forwards the message only to the endpoints (individual applications residing on the node) that have registered to be members of the group ID.

The ZigBee HA profile recommends group addressing each time a message needs to be sent to more than 4 nodes.

7.5.2.2 Indirect Addressing, Binding

Bindings are one of the publish/subscribe models implemented in the ZigBee specification (together with attribute reporting, see Section 7.8).

Cluster C (see Section 7.8) on source endpoint E1 is bound to destination endpoint E2 (typically hosted by a different node) if it sends events related to its output cluster ID(s) to the corresponding input cluster ID(s) of E2. E1 can be bound to multiple target endpoints. Each binding is unidirectional and independent, if E1 is bound to E2, E2 may or may not be bound to E1 (it is a totally different binding).

The binding table can be managed locally through an API (`APSME-BIND.request`) or remotely via ZDP commands: The ZDP `end-Device-Bind` request is sent to endpoint 0 of the target node and specifies:

- The target endpoint of the binding;
- The source 64-bit IEEE address (the 16-bit ZigBee network address is resolved by the APS network address map);

¹ At the 802.15.4 level the ZigBee group messages might be unicast (if the sending node only has one neighbor) or broadcast.

- The source endpoint;
- The list of input clusters and output clusters of the source endpoint.

The local binding table lists, for each binding:

- The local source endpoint;
- The application layer destination address that can be a 802.15.4 address (64-bit format), or a group ID (16-bit);
- The destination endpoint if the destination is not a group address;
- A cluster ID.

A typical use case is that a device looks for another node in the network with capabilities corresponding to a match descriptor (supported application profile, cluster ID and direction, e.g., a lamp supporting the on/off cluster as an input). It then binds to that node using the End-Device-Bind command. In order to facilitate this configuration operation, bindings may be specified for groups. ZigBee devices that can initiate or process events have a button that places them in “identify mode” for about 10 s. Command AddGroupIdentifying can be broadcast and will automatically place the nodes in “Identify” mode in the group.

At the application level, the binding table can be used through the indirect addressing mode, for example, in the *APSDE-DATA.request* primitive. When indirect addressing is used, the destination address (node address or group address) is resolved using the local binding table, based on the endpoint ID of the sending application. Indirect addressing is very flexible as it allows external nodes to configure the routing of messages across ZigBee applications residing on different nodes (e.g., instruct a switch to send its on/off events to a ZigBee-controlled relay).

7.5.2.3 APS Frame Format

The APS frame format is outlined in Table 7.5.

The format of the application level payload depends on the value of the application profile identifier and the cluster identifier:

- Application Profile ID 0x0000: the payload format is defined by the ZigBee device profile (ZDP).
- Application Profile IDs 0x0000 to 0x7FFF are reserved for public application profiles, the payload format is defined by the ZigBee cluster library (ZCL)
- Application Profile IDs 0xBF00 to 0xFFFF are reserved for manufacturer specific profiles (MSP). The payload format is defined by the manufacturer but may also use the ZCL.

Table 7.5 The APS frame format

Field name	Bytes	Field details
Frame Control	1	-----XX : Frame type (00 : APS data) ----XX-- : Delivery mode (00 : unicast; 11 : group addressing) ---X---- : Indirect address mode (0 : ignored) --X----- : Security (0 : none) -X----- : Ack (0 : not required) 0----- : Reserved
Dest. Endpoint	1	16-bit destination address or group ID
Cluster identifier	2	0x0006: On/Off
Application Profile identifier	2	0x0104: HA
Source endpoint	1	
Counter	1	APS level counter
AF payload	Variable up to 80 bytes	APS service data unit (ASDU): a ZCL frame, ZDP frame or application-specific payload.

The ZDP and ZCL are described in the following Sections 7.6 and 7.8.

7.6 The ZigBee Device Object (ZDO) and the ZigBee Device Profile (ZDP)

The **ZigBee Device Object (ZDO)** layer is a specific application running on endpoint 0, designed to manage the state of the ZigBee node. The ZDO application implements the interfaces defined by the ZigBee device profile (ZDP, application profile ID 0x0000).

The clusters defined within the ZDP are similar to those defined in application-specific profiles, but unlike them, the clusters within the ZigBee device profile define capabilities supported in all ZigBee devices. All ZDP client-side transmission of cluster primitives is optional.

The ZDO implements a number of configuration attributes (e.g., the various node descriptors), as well as a number of local APIs and network primitives.

7.6.1 ZDP Device and Service Discovery Services (Mandatory)

These primitives support the discovery of nodes based on some of their characteristics. Since sleeping devices are not capable of receiving such requests, a cache mechanism is provided. The discovery cache is a database of nodes that registered to this cache after a find node cache request, and stores cached descriptor data (stored using Node_Desc_store_req).

ZDP primitives with mandatory server-side processing:

- NWK_addr_req (NWK_addr_rsp): finding a network address from a given IEEE address.
- IEEE_addr_req (IEEE_addr_rsp): finding an IEEE address from a given NWK address.
- Node_Desc_req, Power_Desc_req, Simple_Desc_req (and corresponding responses): finding the node/power/simple descriptor of a device from a given NWK address.
- Active_EP_req (Active_EP_rsp): acquire a list of endpoints on a remote device with simple descriptors.
- Match_Desc_req (Match_Desc_rsp): finding a list of devices with matching profile ids and cluster IDs.
- Device_annce.

ZDP primitives with optional server-side processing:

- Complex_Desc_req, User_Desc_req, Discovery_Cache_req, User_Desc_set, System_Server_Discover_req, Discovery_store_req, Node_Desc_store_req, Power_Desc_store_req, Active_EP_store_req, Simple_Desc_store_req, Remove_node_cache_req, Find_node_cache_req, Extended_Simple_Desc_req, Extended_Active_EP_req ... and corresponding responses.

7.6.2 ZDP Network Management Services (Mandatory)

These services implement the network features corresponding to the node type (coordinator, router or end device), for example, managing network scan procedures, interference detection, and so on. It provides the related interfaces for local applications. Remote nodes can also send a remote management command to permit or disallow joining on particular routers or to generally allow or disallow joining via the trust center.

Only the Mgmt_Permit_Joining_req/rsp server-side processing is mandatory, all other primitives are optional:

- Mgmt_NWK_Disc_req / Mgmt_NWK_Disc_rsp (control network scanning).
- Mgmt_Lqi_req / Mgmt_Lqi_rsp (getting the neighbor list from a remote device).
- Mgmt_Rtg_req / Mgmt_Rtg_rsp (getting the routing table from a neighbor device).
- Mgmt_Bind_req / Mgmt_Bind_rsp (getting the binding table from a neighbor device).
- Mgmt_Leave_req / Mgmt_Leave_rsp (request a remote device to leave the network).
- Mgmt_Direct_Join_req / Mgmt_Direct_Join_rsp (requesting that a remote device permit a device designated by DeviceAddress to join the network directly).
- Mgmt_Cache_req / Mgmt_Cache_rsp (allows to retrieve a list of ZigBee end devices registered with a primary discovery cache device).
- Mgmt_NWK_Update_req / Mgmt_NWK_Update_rsp (allows communication of updates to the network configuration parameters).

7.6.3 ZDP Binding Management Services (Optional)

These primitives enable binding management and maintenance of the bindings table (e.g., processes device replacement notifications). The concept of binding and indirect addressing is discussed in Section 7.5.2.2.

- End_Device_Bind_req / End_Device_Bind_res;
- Bind_req / Bind_res;
- Unbind_req / Unbind_res;
- Bind_Register_req / Bind_Register_res;
- Replace_Device_req / Replace_Device_res;
- Store_Bkup_Bind_Entry_req / Store_Bkup_Bind_Entry_res;
- Remove_Bkup_Bind_Entry_req / Remove_Bkup_Bind_Entry_res;
- Backup_Bind_Table_req / Backup_Bind_Table_res;
- Recover_Bind_Table_req / Recover_Bind_Table_res;
- Backup_Source_Bind_req / Backup_Source_Bind_res;
- Recover_Source_Bind_req / Recover_Source_Bind_res.

7.6.4 Group Management

Although one would expect group management over the network to be part of the core ZDP primitives, these were added later as part of the ZCL groups cluster (cluster ID 0x0004). See Section 7.8.

7.7 ZigBee Security

7.7.1 ZigBee and 802.15.4 Security

ZigBee networks can choose whether to enable security or not. Devices compliant with a public application profile must conform to their profile security settings.

ZigBee offers security services at two levels:

- Network (NWK) level security;
- Application (APS) level security.

None of these security services uses the MAC-level security defined by 802.15.4 (the 802.15.4 frame control field security bit is set to 0), which would encrypt the ZigBee NWK header that is required by ZigBee routing. However, ZigBee simply transposes the exact same mechanisms to the ZigBee network and application layer, and uses the

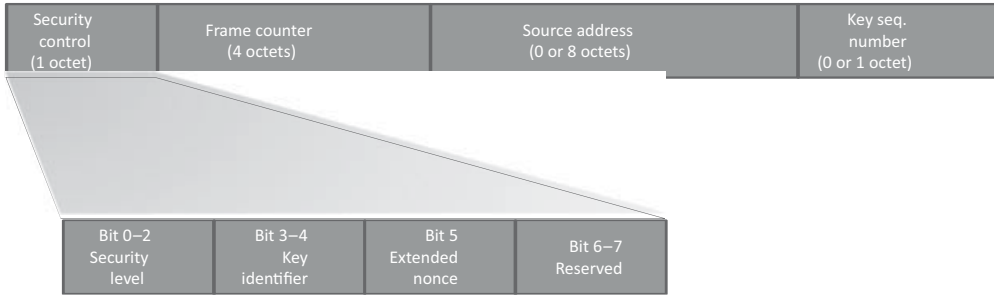


Figure 7.4 Auxiliary Security header format.

same encryption and hash algorithm (AES-CCM*), so encryption acceleration modules of 802.15.4 chips can still be used.

The ZigBee device object (ZDO) manages the security policies and configuration of a device.

7.7.1.1 NWK Level Security

ZigBee provides optional integrity protection and encryption, as illustrated in Figure 7.6. When network-level integrity protection or encryption is used, the security bit in the NWK control field is set to 1, indicating the presence of a security auxiliary frame header, and a message-integrity protection code (MIC).

The NWK security auxiliary header is composed of four subfields, illustrated in Figure 7.4.

In the security control subfield, the security level is set to the value of the MIB nwk security-level parameter, by default 0x05. This value specifies whether the frame is only integrity protected but not encrypted, or that it should also be encrypted, in which case the entire network payload is encrypted. Network security is applied as configured in the device network information base (NIB, nwkSecureAllFrames=TRUE to secure all frames) by default, but the application may override this setting frame by frame by specifying the SecurityEnable parameter of the NLDE-data.Request primitive.

The services provided by each security level are listed in Figure 7.5.

The MIC, as well as encryption, are computed using AES and CCS* (see Section 1.1), using the main or alternate network key.

7.7.1.2 Application Layer Security

The APS layer provides a number of security primitives that can be used by application developers:

- APSME-ESTABLISH-KEY to establish a link key with another ZigBee device using the SKKE protocol.
- APSME-TRANSPORT-KEY to transport security material from one device to another.
- APSME-UPDATE-DEVICE to notify the trust center when a device joins or leaves the network.
- APSME-REMOVE-DEVICE to instruct a router to remove a child from the network (used by the trust center).
- APSME-REQUEST-KEY to ask the trust center an application master key or the active network key.
- APSME-SWITCH-KEY used by the trust center to tell a device to switch to a new network key.
- APSME-AUTHENTICATE used by two devices to authenticate each other.

Security control	0x00	0x01	0x02	0x03	0x04	0x05	0x06	0x07
encryption	No				Yes			
MIC bit size	No MIC	32	64	128	No MIC	32	64	128

Figure 7.5 AUX header security control field values. The key identifier is set of 0x01 for the active network key (0x00 for link keys used by the APS layer).

In addition, the application-layer security provides its own optional integrity and encryption services, based on the network key or on a link-specific key (associated to the destination of the packet), under control of the application (TxOptions parameter).

Just like NWK security, the APS layer security will add an auxiliary security header (AUX header), and an integrity code, as illustrated on Figure 7.6. The difference is that the integrity protection scheme protects the APS header, auxiliary header and APS payload, and when encryption is used, only the APS payload is encrypted.

7.7.2 Key Types

– Master Keys

- Application master keys are distributed by the trust center (via unsecured key transport) and used to set up link keys between two devices, and for mutual authentication of devices.
- Trust center master keys are used to derive a link key for communication with the trust center.

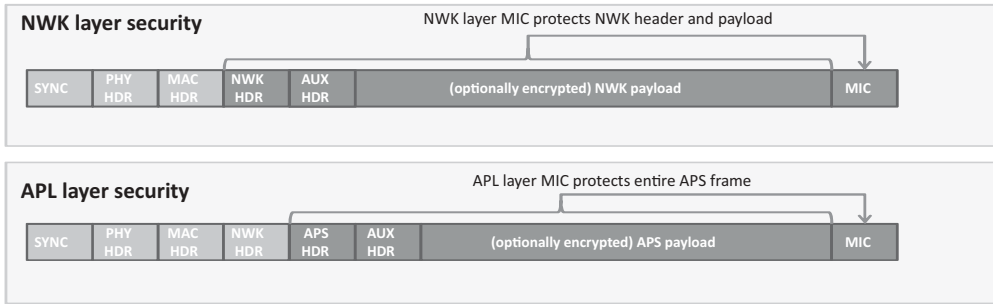


Figure 7.6 ZigBee NWK and APS security services.

- **Network keys** are used for network management. They are preconfigured or configured by the trust center, using unsecured key transport when standard security is used or using the trust center link key in high-security mode.
 - Standard network keys are used in the standard security mode, and can be used to secure general application layer commands.
 - High-security network keys are used in the high-security mode, they are not used for communication between secure devices, which use link-specific keys instead.
- **Link keys** can either be negotiated using SKKE, or configured by the trust center under request of a device. Service specific keys are derived by hashing of the link key:
 - The **key-load** key is used to protect transported master and link keys.
 - The **key-transport** key is used to protect transported network keys.

A ZigBee device stores a (master key/link key) key pair for each device with which they may use link-key-based communication. The device is identified by its 64-bit IEEE address.

7.7.3 The Trust Center

Key distribution is not addressed by 802.15.4. For security purposes, ZigBee defines the role of “trust center”, which is responsible for key distribution and joining policy.

In *high-security mode*, the trust center maintains a list of devices, master keys, link keys and network keys. A device can be preloaded with the trust center address and initial master key, or the master key can be sent via an unsecured key transport primitive. The trust center, by default, is the ZigBee coordinator, but the coordinator can designate another device, or the trust center can be preconfigured in devices.

In *low-security mode*, a device communicates with its trust center using the current network key, which can be preconfigured (as in the commercial building automation profile) or configured via an unsecured key transport primitive during the joining process (as in the home automation profile).

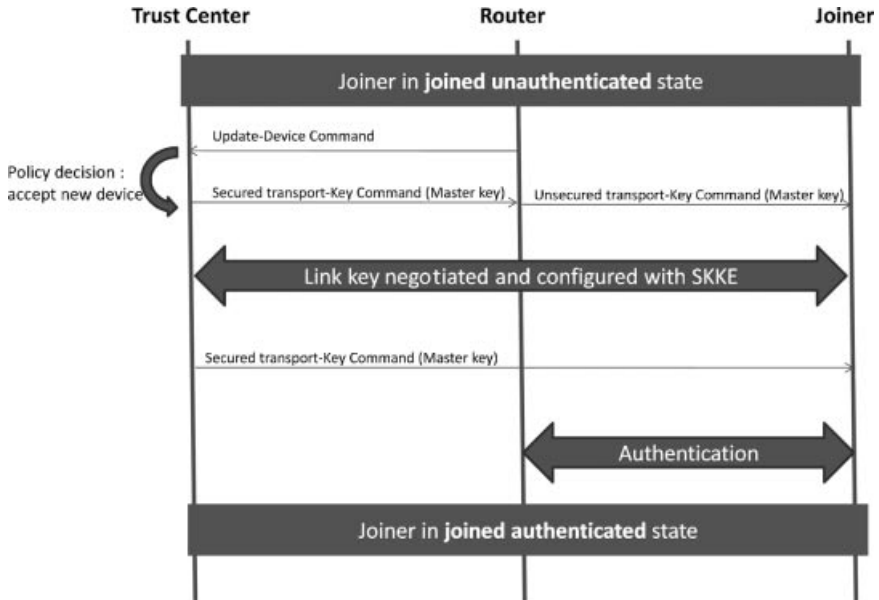


Figure 7.7 Security bootstrapping procedure in high-security mode.

The trust center is notified by ZigBee routers of joining devices by means of an Update-Device command (see Figure 7.7). They can reject or accept the new joiner. In the latter case the trust center communicates the joiner master key to the router, which relays it to the joiner. This enables the trust center to establish a link key to the joiner, and to finally securely communicate the NWK key to the joiner using a secure Transport-Key command (key-transport key, derived from the link key).

Devices can also request the trust center to compute a link key pair for communication with another device: the trust center will communicate the link key to each device using Transport-Key commands.

7.7.3.1 SKKE

Secure ZigBee devices are configured with a link specific master key for each device they may need to communicate securely with, or need to authenticate. A link key, distinct from the master key, can be negotiated between ZigBee endpoints by using a symmetric key establishment (SKKE) scheme using procedures defined in ANSI X9.63-2001.

7.7.3.2 Entity Authentication

ZigBee provides a APSME-AUTHENTICATE.request primitive is used for initiating entity authentication or responding to an entity authentication initiated by another device. The request includes a random challenge and the response is a hash based on the shared master key for the device pair and frame sequence numbers.

7.7.4 The ZDO Permissions Table

The ZDO optionally comprises a permissions configuration table. This table lists a number of tasks categories (e.g., *ApplicationSettings* for authorization to configure bindings, groups, and other application configuration commands) that can be requested from an external entity to the ZDO. For each task category, the permissions configuration table lists the addresses of devices authorized to perform these tasks (or specifies that any device is authorized), and specifies whether the related commands need to be secured with a link-specific key.

7.8 The ZigBee Cluster Library (ZCL)

The cluster library provides support for communication between applications, including over the air configuration of group tables, reading and setting attribute values, subscribing to certain attribute-value changes, and so on.

7.8.1 Cluster

A cluster is a set of related commands and attributes. Each cluster is defined by a ZigBee application profile and identified by a cluster ID (0x0000 to 0xFFFF). Commands are identified by a command ID (0x00 to 0xFF), and attributes by an attribute ID (0x0000 to 0xFFFF).

Clusters are composed of a server side and a client side. Each application residing on a ZigBee device lists the clusters that it supports as a server in its simple descriptor input cluster list, and the clusters that it supports as a client in the simple descriptor output cluster list.

An application that implements the server side typically exposes a set of attributes, and must be able to receive and process the primitives defined by the cluster to read or manipulate these attributes. It may also support attribute-reporting commands, which are sent to requesting devices that implement the client side of the cluster.

Vice versa an application that implements the client side of the cluster may use any of the commands supported by the server side, and must support receiving attribute reporting commands, if it requests such notifications.

Each cluster is defined only within a given application profile. However, in order to avoid any confusion and to redefine common clusters multiple times, the clusters defined by the ZigBee cluster library (ZCL) use the same cluster IDs for each ZigBee public application profile. Table 7.6 lists some of the clusters defined by the ZCL, which therefore have the same meaning for all public application profiles defined by ZigBee.

Clusters are directional: for each endpoint, the endpoint simple descriptor contains the list of input clusters (commands that it accepts and properties that can be written to the endpoint), and the list of output clusters (commands that it may send and properties that may be read).

Table 7.6 Some clusters defined by the ZCL

Cluster ID	Cluster Name
0x0000	Basic cluster
0x0001	Power configuration cluster
0x0002	Temperature configuration cluster
0x0003	Identify cluster
0x0004	Groups cluster
0x0005	Scenes cluster
0x0006	OnOff cluster
0x0007	OnOff configuration cluster
0x0008	Level control cluster
0x000a	Time cluster
0x000b	Location cluster

Each ZigBee public application profile defines a list of devices, each implementing a standard set of features defined by a list of input and output clusters. Within a given public-application profile, manufacturers can extend that list by using the “manufacturer specific extension” field of the ZCL frame (see Section 7.8.4). An endpoint declares the clusters it supports in its simple descriptor, which contains the endpoint number, application profile ID, application device ID, and application version ID (see Section 7.9), the list of input and output clusters.

7.8.2 Attributes

Attributes are identified by a 16-bit number. The ZCL library provides commands to:

- read or set attributes;
- require reporting on an attribute, either periodic or based on a change of value.

7.8.3 Commands

Commands are identified by an 8-bit number. The first 4 bits of cluster-specific commands (defined only within the scope of a given cluster) are set to zero. For instance, command 0x00 of the OnOff cluster (0x0006) means “Off”.

Other identifiers are reserved for cross-cluster commands.

7.8.4 ZCL Frame

The ZigBee cluster library frame (Table 7.7) carries application-specific commands.

The general commands (frame type = 00) are listed in Table 7.8.

Table 7.7 ZCL frame format

Field	Bytes	Field details
Frame Control	1	<p>-----XX : Frame type (00 : cluster for entire app. Profile, 01 : cluster specific command, other values reserved)</p> <p>----0-- : Manufacturer specific (0 : defined by ZCL, 1 manufacturer specific)</p> <p>---X--- : Direction (0 : client to server)</p> <p>000X---- : Default response (0 enabled, 1 disabled)</p> <p>000----- : Reserved</p>
Manufacturer code	0/2	Present only if the frame is manufacturer specific (Frame control bit 6)
Transaction sequence number	1	Matching of request command frames and response command frames
Command identifier	1	e.g., 0x42: toggle

Table 7.8 ZCL general commands

Command identifier	Command	Usage
0x00	Read attributes	Read a list of attributes, whose 16-bit identifiers are passed as payload
0x01	Read attributes response	Includes a list of attribute records as payload. Each attribute record is as follows: Attribute identifier (16 bit) Status(Success or error: 8 bit) Attribute data type (0/8 bit) Attribute value (variable). Array types are encoded as element type number of elements list of elements.
0x02	Write attributes	Write a list of attributes, as described by a list of attribute records in payload.
0x03	Write attributes undivided	Write attributes, all or nothing mode.
0x04	Write attributes response	Response with a list of status codes, for each parameter.
0x05	Write attributes no response	Write attributes, best effort mode.
0x06	Configure reporting	Only specific attributes support reporting in a cluster. The list of attribute reporting configuration records passed as parameter includes minimal and maximal reporting intervals, the minimum change of the attribute that should trigger reporting. This command may be sent by a client to a server to configure reporting, or by a server to a client to describe its future reporting parameters.

(Continued)

Table 7.8 (Continued)

Command identifier	Command	Usage
0x07	Configure reporting response	
0x08	Read reporting configuration	Request to get the reporting configuration parameters for one or more attributes
0x09	Read reporting configuration response	
0x0a	Report attributes	Reporting for one or more attributes of a cluster, according to a previously configured reporting relationship (binding)
0x0b	Default response	Basic success/error response
0x0c	Discover attributes	Specifies a starting 16-bit attribute identifier and maximum number of identifiers to report
0x0d	Discover attributes response	
0x0e	Read attributes structured	Used to read only specific index positions (up to 15) for array-type attributes
0x0f	Write attributes structured	Used to write only specific index positions (up to 15) for array-type attributes
0x10	Write attributes structured response	

7.9 ZigBee Application Profiles

An application profile defines a set of messages and attributes standardized for use in a particular context. Application profiles are identified by an application profile ID (0x0000 to 0xFFFF). Each manufacturer can define proprietary messages within its own private application profile, but the ZigBee alliance defines a set of public application profiles, which enable cross-vendor interoperability for the target applications. Profile IDs 0xBF00 to 0xFFFF are reserved for manufacturer specific profiles (MSP), and must be requested from the ZigBee alliance. Profile IDs 0x0000 to 0x7FFF are reserved for public application profiles, and are listed in Table 7.9.

Each ZigBee public profile contains a list of devices identified by a device ID. Each device implements a standard set of features defined by a list of input and output clusters, and a list of attributes (some optional, some mandatory). Within a given public application profile, manufacturers can extend that list by using the “manufacturer-specific extension” field of the ZCL (see Section 7.8.4).

7.9.1 The Home Automation (HA) Application Profile

The HA profile is by far the most widely implemented by manufacturers. While ZigBee does not specify use of nonvolatile memory, the HA profile does: nodes should retain

Table 7.9 ZigBee public application profiles

Public Application Profile		Profile ID	Usage
Home Automation	HA	0x0104	Security, HVAC, LIGHTING CONTROL, ACCESS CONTROL, IRRIGATION. . .
Commercial Building Automation	CBA	0x0105	Security, HVAC, AMR, lighting control, access control
Industrial Plant Monitoring	IPM	0x0101	Asset management, process control, environmental control, energy management
Telecommunications Applications	TA	0x0107	Information delivery in hot zones, public information enquiry, location-based services, remote control (TV, DVD), cell phone
Automatic (Advanced) Metering Initiative or Smart Energy 1	AMI ZSE 1	0x0109	
Personal Home and Hospital (Health) Care	PHHC	0x0108	Patient monitoring, Fitness monitoring.

their configuration (PAN ID, address, group IDs, etc. . . .) even after a power down. This facilitates battery replacement and network restarts after power outages (silent rejoin, see Section 7.3.3).

The HA profile determines the following settings:

- It uses stack profile 0x01 or 0x02.
- Channels 11, 14, 15, 19, 20, 24, 25 are preferred.
- The trust center link key is hardcoded in the profile, the trust center distributes a network key.
- The minimum number of entries of entries in the broadcast transaction table (BTT, see Section 7.4.5.1) is nine.

The standard devices defined by the home automation public application profile are characterized by the mandatory clusters that they must support. Mandatory and optional “common clusters” are defined for all HA devices:

- Server side: mandatory support of basic and identify clusters. Optional support for power configuration, device temperature configuration, alarms, meter, and manufacturer-specific clusters.
- Client side: optional support for meter, and manufacturer-specific clusters.

The HA profile also lists mandatory and optional clusters specific to each device type. Table 7.10 lists some examples.

Table 7.10 Some devices defined by the HA public application profile

	Device name	Device ID	Supported clusters Mc/Ms: mandatory on client or server side, Oc/Os: optional on client or server side
Generic	On/Off switch	0x0103	Ms: OnOff (0x0006) Os: OnOffSwitch Config (0x0007) Oc: Scenes (0x0005) Oc: Groups (0x0004) Oc: Identify (0x0003)
	Range Extender	0x0008	Only common clusters
	Mains Power Outlet	0x0009	Ms: OnOff (0x0006) Ms: Scenes (0x0005) Ms: Groups (0x0004)
Lighting	On/Off Light	0x0100	Ms: OnOff (0x0006) Ms: Scenes (0x0005) Ms: Groups (0x0004)
	DimmableLight	0x0101	Ms: OnOff (0x0006) Ms LevelControl (0x0008) Ms: Scenes (0x0005) Ms: Groups (0x0004) Oc: Occupancy sensing (0x0406)
	Light Sensor	0x0106	Ms: Illuminance measurement (0x0400) Oc: Groups (0x0004)
	DimmerSwitch	0x0104	Mc: OnOff (0x0006) Mc LevelControl (0x0008) Oc: On Off switch configuration (0x0007) Os: Scenes (0x0005) Os: Groups (0x0004)
	Shade	0x0200	Ms: OnOff (0x0006) Ms LevelControl (0x0008) Ms: On Off switch configuration (0x0007) Ms: Scenes (0x0005) Ms: Groups (0x0004)
Closures	Shade Controller	0x0201	Mc:OnOff (0x0006) Mc LevelControl (0x0008) Oc: Shade configuration (0x0100) Oc: Scenes (0x0005) Oc: Groups (0x0004) Oc: Identify (0x0003)

Table 7.10 (Continued)

		Supported clusters Mc/Ms: mandatory on client or server side, Oc/Os: optional on client or server side	
	Device name	Device ID	
HVAC	Heating / Cooling unit	0x0300	Ms:OnOff (0x0006) Mc: Thermostat (0x0201) Os: Fan control (0x0202) Os: Level control (0x0008) Os: Groups (0x0004)
	Thermostat	0x0301	Ms: Thermostat (0x0201) Os: Scenes (0x0005) Os: groups (0x0004) Os: Thermostat user interface configuration (0x0204) Os/ Oc: Fan control (0x0202) Os / Oc: Temperature measurement (0x0402) Os / Oc:: Occupancy sensing (0x0406) Os/Oc: Relative humidity measurement (0x0405)
	Temperature sensor	0x0302	Ms: Temperature measurement (0x0402)

7.9.2 ZigBee Smart Energy 1.0 (ZSE or AMI)

ZigBee Smart Energy 1.0 is a public application profile (profile 0x0109), documented in “ZigBee SMART ENERGY PROFILE SPECIFICATION (rev 15, 1/12/2008)”. It defines the smart energy devices and clusters required to build an energy-management system (EMS).

The ZigBee Smart Energy 1 (ZSE) public application was defined to enable usage of ZigBee for automatic metering, demand response and prepayment applications required by utilities. The ZSE was defined just before ZigBee decided that its next version would rely on IP, and ZSE was the first application profile to be entirely redesigned in the context of IP, the new specification is called ZigBee Smart Energy 2.0 (see Section 13.4).

ZSE dedicates a secure HAN (use of security is mandatory) to the utility, and defines the communication primitives used between the energy service portal controlled by the utility, and devices located in the end user primitives, such as home displays or load control devices. As the displays and other ZigBee devices may be deployed by end users on their home network using HA, ZSE also defines an extension (the “stub APS”) to the ZigBee core specification enabling limited single hop communications between two PANs (the utility PAN and the user home area network).

ZSE defines several new clusters listed in Table 7.11.

7.9.2.1 Security

SE makes a mandatory use of link keys (preconfigured or commissioned), which are otherwise optional in ZigBee. However, a master key is not used or preconfigured in ZigBee SE devices, which do not operate in “high-security” mode.

ZigBee smart energy devices use a network key allocated by the trust center, using the key establishment cluster with a preconfigured trust center link key. The link key is also replaced by the trust center as part of this security bootstrapping process. ZigBee smart energy networks will not generally send keys in the clear.

ZigBee SE envisions that two separate home area networks will be used:

- One network for the exclusive use of the utility company, interconnecting the energy services portal, in-home display(s) and load control devices.
- One separate network for the home owner use, including home automation devices, in-home displays (ihd), a home energy management console, and smart appliances.

The links between the automatic metering infrastructure (AMI) servers and the home area networks may use a combination of non-ZigBee and ZigBee networks. The energy services portal may control a collection of sub-ESPs, in a cascading fashion, so that ZigBee can optionally also be used as a neighborhood area network (NAN).

Since all ZigBee SE devices are configured with multiple keys (link keys and the network key), ZigBee SE defines which cluster uses which key. All the ZigBee SE clusters use application link keys, but most general clusters (e.g., basic cluster, alarm cluster, identify cluster) use the network key.

7.9.2.2 Smart Energy Extensions of ZigBee

ZigBee SE defines a simplified APS layer designed for basic “inter-PAN” communication, that is, communication between a PAN and a device that has not joined. The specification mentions the “refrigerator magnet” with an LCD screen as a target.

Such messages can be sent unicast, broadcast or sent to a group, but without any security.

7.9.2.3 Smart Energy Devices

ZigBee SE defines the following “smart energy” devices:

Energy Service Portal (ESP, Device Id 0x0500)

The energy service portal is a server controlled by the utility company that connects to the metering and energy management devices within the home. The ESP acts as the coordinator and trust center of the network.

The ESP must support the server side of the price, message, demand response/load control and time clusters, and optionally of the complex metering, simple metering and prepayment. It may support the client side of the simple metering, complex metering, price and prepayment clusters.

Metering Device (Device Id 0x0501)

A metering device must support the client side of the metering cluster, optionally of the complex metering cluster. It may support the client side of the metering prepayment, price and message clusters.

In-Premises Display Service (Device Id 0x0502)

The device is designed to be used as a simple user interface, displaying graphs or messages, and able to signal button press events. It must implement the client side of at least one of the price, simple metering and messaging clusters.

Programmable Communicating Thermostat (PCT, Device Id 0x0503)

This device is designed to control heating and cooling systems. It must implement the client side of the Demand response/load control and time clusters. It may implement the client side of the prepayment, price, simple metering and message clusters.

Load Control Device (Device Id 0x0504)

This device is designed to manage the electric consumption of devices in a generic way. It must implement the client side of the demand response/load control and time clusters. It may implement the client side of the price message cluster.

Range Extender Device (Device Id 0x0008)

A device announcing this device Id must be a pure ZigBee router, not supporting any other function.

Smart Appliance Device (Device Id 0x0505)

Smart appliances are able to participate in energy-management policies. They must implement the client side of the price and time clusters, and optionally of the demand response/load control and message clusters.

Prepayment Terminal Device (Device Id 0x0506)

This device is not fully specified yet. It is designed to support advance payment of services, and various display functions.

7.9.2.4 Smart Energy Clusters

Demand Response and Load Control Cluster (0x0701)

Server-Side Commands

Table 7.11 New clusters defined by ZigBee smart energy

Price	0x0700
Demand response and load control	0x0701
Simple metering	0x0702
Message	0x0703
Registration	0x0704
AMI tunneling (complex metering)	0x0705
Prepayment	0x0706

Load control event (0x00). This event is sent to the devices asked to implement a load control action, and specifies the actions required. It includes the following parameters:

- *Issuer Event ID*: unique identifier that will be used to identify future event reports related to this demand response and load control event.
- *Device Class*: bit-encoded field indicating the device classes (end devices actually performing the energy-demand response, as opposed to their ZigBee controllers) needing to participate in an event. The classes defined include HVAC compressors, Strip heaters, water heaters, electric vehicles, and so on.
- *Utility Enrolment Group*: both the utility enrolment group field and the device class must match the target device configured values, otherwise it will ignore the command.
- *Start Time*: UTC Timestamp or 0x00000000 for “now.”
- *Criticality Level*: levels 1 to 9 are currently defined. Participation in levels 1 to 6 is voluntary, participation in level 9 is mandatory. Level 1 signals an abnormal percentage of nongreen sources in the delivered energy.
- *Cooling and heating temperature offsets*: offsets, in units of 0.1 °C, to the local temperature setpoint of the thermostat (added for cooling systems, subtracted for heating systems), noncumulative across sequential demand response events. 0xFF when not used.
- *Cooling and heating temperature setpoint*: request to replace the current setpoint by the indicated value between -273 °C and 327 °C in units of 0.01 °C, set to 0x8000 when not used.
- *Average Load Adjustment Percentage*: defines a load offset of -100 to +100 points relative to 100%, with a resolution of 1 point. Interpretation is specific to the client implementation. Set to 0x80 when not used.
- *Duty cycle*: a percentage of time between 0 and 100%, 0xFF when not used.
- *Event Control*: flags to indicate whether randomization of the start and end times is required.

Cancel load control event (0x01): cancellation order specifying the issuer event Id, utility enrollment group and device classes concerned. Flag to optionally override the end

randomization settings of the original load control event, and desired cancellation UTC time (0x00000000 for “now”).

Cancel all load control events (0x02): same as the cancel load control event command, without the filtering parameters.

Client Side

The client side of the load control/demand response cluster must be able to store at least 3 scheduled events. Events exceeding the storage capacity should be retrieved as soon as possible using command “get scheduled events”.

The client side maintains several attributes:

- The utility enrolment group;
- The start randomization period, and the stop randomization period, in minutes;
- The device class bitmask.

It implements two commands:

- **Get Scheduled Events (0x01)**, which asks the server side to resend up to a certain number of load control commands scheduled to start at or after the specified UTC time stamp.
- **Report Event Status (0x00)**, which reports the time at which the load-control event (identified by its event ID) was effectively executed, the criticality level, the cooling or heating set points, load adjustment percentage, duty cycle that were applied. The event status parameter contains the current state of the load control event: started, completed, user opted in or opted out before or during the event, canceled, superseded, and error conditions. An electronic signature ensures nonrepudiation.

Simple Metering Cluster (0x0702)

Server-Side Attributes

Attribute set identifier	Attributes	
0x00 Reading information set	CurrentSummationDelivered (0x00), CurrentSummationReceived (0x01),	Most recent summed value of energy/gas/water delivered to, or provided by the premises.
	CurrentMaxDemandDelivered (0x02), CurrentMaxDemandReceived (0x03), CurrentMaxDemandDeliveredTime (0x08), CurrentMaxDemandReceivedTime (0x09)	Instant value of maximum rates, and when they were measured.
	DFTSummation (0x04), DailyFreezeTime (0x05),	Snapshot of CurrentSummationDelivered taken at instant DailyFreezeTime

	PowerFactor (0x06), ReadingSnapShotTime (0x07),	Average power factor
0x01 TOU information Set	CurrentTierNSummationReceived CurrentTierNSummationDelivered	Where $N = 1$ to 6, partial summation counters per price tier (defined per period in the time of use schedule or per price tier).
0x02 Meter status	Status (0x00)	Status flags: tamper detected, leaks, etc.
0x03 Formatting	UnitOfMeasure (0x00), Multiplier (0x01), Divisor (0x02), SummationFormatting (0x03), DemandFormatting (0x04), HistoricalConsumptionFormatting (0x05),	Set of attributes used to transform counters to displayable values: units, scaling factors.
0x04 ESP historical consumption	InstantaneousDemand (0x00), CurrentDayConsumptionDelivered (0x01), CurrentDayConsumptionReceived (0x02), PreviousDayConsumptionDelivered (0x03), PreviousDayConsumptionReceived (0x04), CurrentPartialProfileIntervalStartTimeDelivered (0x05), CurrentPartialProfileIntervalStartTimeReceived (0x06), CurrentPartialProfileIntervalValueDelivered (0x07), CurrentPartialProfileIntervalValueReceived (0x08)	Accumulators since midnight for the current day, since the start time of the current profile interval, for the previous day
0x05 Load profile configuration	MaxNumberOfPeriodsDelivered (0x00)	maximum number of intervals the device is capable of returning in one get profile response command.
0x06 Supply Limit	CurrentDemandDelivered (0x00), DemandLimit (0x01), DemandIntegrationPeriod (0x02), NumberOfDemandSubintervals (0x03)	Demand is integrated during the demand integration period, and the integration result is written to currentDemandDelivered at the end of each subinterval.

Server-Side Commands

GetProfileResponse (0x00) returns a number of period accumulators for periods ending before a specified EndTime.

RequestMirror (0x01): request the ESP to mirror metering device data, using RequestMirrorResponse command.

command

RemoveMirror (0x02): request the ESP to remove its mirror of metering device data.

Client-Side Commands

GetProfile (0x00): requests a number of period accumulators for periods ending before a specified EndTime, for received or for delivered quantities.

RequestMirrorResponse (0x01): the ESP informs a sleepy metering device it has the ability to store and mirror its data, which should be sent to the indicated endpointID.

RemoveMirror (0x02): the ESP no longer has the ability to mirror data.

Price Cluster (0x0700)*Server-Side Attributes*

6 price tiers labels are defined, by attributes “TierXpriceLabel” (0x0000 to 0x0006). The price tiers are defined by command publish price.

Client-Side Commands

- **getCurrentPrice (0x00)**: initiates a publish price command for the current time.
- **getScheduledPrices (0x01)**: initiates a publish price command for all currently scheduled times after the provided time stamp (up to a maximum number of scheduled times also specified in the command).

Server-Side Commands

Publish price (0x00): this command defines a new price tier and is sent in response to a getCurrentPrice or getScheduledPrices command. It contains the following subfields:

- *ProviderID*: unique Id of the commodity provider;
- *Rate Label*: 12 character UTF8 string related to current rate;
- *Issuer Event ID*: unique identifier for this pricing information, must increase when newer prices are published for the same period;
- *Current Time*: UTC time of the sending node;
- *Unit of measure*: 8-bit field defining the commodity and unit of measure;
- *Currency*
- *Price Trailing Digit and Price tier*: bit field indicating the price tier for this rate, and the position of the decimal point in the published price;

- *Number of price tiers and Register tier*: number of price tiers in use (0 to 6), for the current tier, indication of which CurrentTierXSummationDelivered accumulator relates to the current price tier;
- *StartTime*: UTC start time of the current rate signal, 0x00000000 means “now”;
- *Duration in minutes*: validity of this price signal. 0xffffffff means “until changed”;
- *Price*: price per base unit;
- *PriceRatio* (optional): ratio to the “normal” tariff;
- *Generation Price* (optional): price for commodity received from the premises;
- *AlternateCostDelivered*, Alternate cost unit and alternate cost trailing digit: cost using an alternate measure, for example, grams of CO₂ per kWh.

Messaging Cluster (0x0703)

Server-Side Commands

DisplayMessage (0x00): specifies the level of importance of the message, whether a confirmation is required, and the number of minutes the message must be displayed.

CancelMessage (0x01)

Client-Side Commands

GetLastMessage (0x00): request to send a DisplayMessage command

MessageConfirmation (0x01): used to acknowledge a message, provides a timestamp.

Smart Energy Tunneling (Complex Metering) Cluster (0x0704)

Not defined yet, this cluster is a placeholder for future tunneling mechanisms for more sophisticated metering protocols, for example, C.12 or DLMS/COSEM.

Prepayment Cluster (0x0705)

Not defined yet.

7.10 The ZigBee Gateway Specification for Network Devices

The ZigBee Gateway specification for network devices, Version 1.0, was released on July 27th 2011. It had been a work in progress since 2007. This specification defines several possible communication protocols between ZigBee gateway devices (ZGD) that implement a bidirectional interface between ZigBee 1.0 PANs and IP networks, and IP host applications (IPHA).

The ZGD provides access to:

- ZCL operations: read and write attributes, configure notifications and report events;
- ZDO operations;
- Macro operations simplifying network and service discovery;
- Endpoint management;

- Its own ZigBee information bases (AIB, NIB, and PIB attributes);
- Network startup and join functions on behalf of the IPHA;
- Security material configuration and operations.

The communication between ZGDs and IPHAs is bidirectional, both act as client and server. The IPHA calls procedures implemented by the ZGD, and the ZGD calls event handlers of the IPHA. These RPC functions have been specified in an abstract, protocol-independent request-response format that needs to be complemented by a protocol binding specification.

Some ZGD procedures operate only in blocking mode, while other procedures offer a choice of blocking or nonblocking mode (such nonblocking procedures provide a *CallbackDestination*, which is an IPHA callback URL or an empty string if the IPHA uses polling). All IPHA event handlers are nonblocking.

ZigBee PAN messages received by the ZGD are processed by one or more ZGD callback handlers, which decode and feed them to requesting IPHAs. The configuration of these callback handlers is persistent across ZGD power cycles.

Version 1.0 of the ZigBee gateway specification proposes three possible network bindings implementing the ZGD to IPHA remote procedure calls: a SOAP binding, a REST binding, and a GRIP (gateway remote interface protocol) binding.

7.10.1 The ZGD

The ZGD itself is a ZigBee device. As such it should support the ZigBee commissioning cluster, as well as mandatory ZDO client and server clusters.

The ZGD functions belong to several categories: gateway management object, ZigBee device profile, ZigBee cluster library, application support sublayer, inter-PAN, and network layer.

An IPHA can access:

- APS commands to read and modify the ZGD configuration by using a set of APS functions. It can read or set descriptors (*ConfigureNodeDescriptor*, *GetNodeDescriptor*, *GetNodePowerDescriptor*, *ConfigureUserDescriptor*, *GetUserDescriptor*), manage local ZGD endpoints (*ConfigureEndpoint*, *ClearEndpoint*), send and receive APS frames (*SendAPSMesssage* / *NotifyAPSMesssageEvent*), manipulate local groups (*AddGroup*, *RemoveGroup*, *RemoveAllGroups*, *GetGroupList*), and read local bindings (*GetBindingList*).
- Network layer commands to get access to the Network layer management entity (NLME) of the ZGD (*FormNetworkProcedure*, *FormNetworkEvent*, *StartRouter*, *StartRouterEvent*, *Join*, *JoinEvent*, *Leave*, *LeaveEvent*, *Reset*, *ResetEvent*, *DiscoverNetworks*, *DiscoverNetworksEvent*, *PerformEnergyScan*, *PerformEnergyScanEvent*,

NetworkStatusEvent, PerformRouteDiscovery, PerformRouteDiscoveryEvent, Send-NWKMessage, NotifyNWKMessageEvent).

An IPHA can send arbitrary ZDP frames by using the SendZDPCommand ZGD function and receive arbitrary ZDP frames by implementing a NotifyZDPEvent event handler.

The ZGD also provides access to the ZCL. An IPHA can send and receive arbitrary ZCL commands by using the SendZCLCommand ZGD function and implementing the NotifyZCLEvent event handler.

In addition the ZGD supports the specific gateway management object (GMO). The GMO provides functions:

- Enabling the IPHA to retrieve the ZGD version and feature set (GetVersion).
- Enabling read/write of ZGD information base (Get/Set).
- Event CallBack management (CreateCallBack, DeleteCallBack, ListCallBacks) and polling by IPHA (PollCallBack, PollResults, UpdateTimeOut).
- Facilitating network and service discovery (StartNodeDiscovery, NodeDiscoveryEvent, NodeLeaveEvent, ReadNodeCache, StartServiceDiscovery, ServiceDiscoveryEvent, GetServiceDescriptor, ServiceDiscoveryEvent, ReadServiceCache).
- Controlling the Gateway insertion into the PAN (StartGatewayDevice, StartGateway-DeviceEvent, ConfigureStartupAttributeSet, ReadStartupAttributeSet).
- Managing address aliases (CreateAliasAddress, DeleteAliasAddress, ListAddresses):
In order to facilitate the mapping of addresses between the PAN network (short 16-bit addresses may change in the PAN) and the IPHA, the ZGD maintains an address alias table where IPHAs can associate their own alias to any 64-bit address. The IPHA or ZGD may use special code 0x10 in the Destination/source address mode² of ZDP messages to indicate use of an alias address: the ZGD will translate to 64-bit or 16-bit addresses as required.

7.10.2 GRIP Binding

The gateway remote interface protocol (GRIP) is a lightweight request/response protocol built over SCoP.

The secured connection protocol (SCoP) was designed in the context of the ZigBee bridge device specification as an IP tunneling protocol between ZigBee PAN gateways. It provides support for datagram and stream-oriented communications as well as fragmentation by means of a socket-like

² **APSDE-DATA.request DstAddrMode parameter values** : 0x00 = DstAddress and DstEndpoint not Present; 0x01 = 16-bit group address for DstAddress; DstEndpoint not present; 0x02 = 16-bit address for DstAddress and DstEndpoint present; 0x03 = 64-bit extended address for DstAddress and DstEndpoint present; 0x04 – 0xff = reserved.

interface provided by the SCoP data entity (SCDE), on top of UDP, TCP or TLS. SCoP security leverages CCM* (SCoP security service SCSS). Management services are provided by the SCoP management entity (SCME).

SCoP messages target a specific service (identified by a service identifier code), on the target gateway. Currently the supported services are SCoP service (hello, goodbye, keepalive commands), bridge service and GRIP service.

The GRIP API provides a GRIDE-DATA request that must be provided with target information (DestIPVersion, DestIPAddress, DestPort), desired transport parameters (TransportMode, SecurityLevel) and application-level parameters (target FunctionDomain, FunctionCategory, ManufacturerCode, FunctionIdentifier, FunctionParameters).

Requests are provided to the target GRIP application by means of the GRIDE-DATA.indication API. Responses are returned to the querying GRIP application by means of the GRIDE-DATA.response API, and specifies a Status code as well as a function result (an octet string).

The GRIP binding specifies specific GRIP function codes and parameter formats to transport all the generic ZGD procedures defined in Section 7.10.1. The following function categories are defined: Manufacturer specific, GMO, ZDP, ZCL, APS, INTERPAN, NWK. Parameter encoding uses ASN.1 distinguished encoding rules (DER).³

7.10.3 SOAP Binding

The SOAP binding specifies a standard web services interface by means of a WSDL document. All the generic ZGD procedures defined in Section 7.10.1 are covered.

7.10.4 REST Binding

Due to the resource oriented design pattern of REST, the REST binding for the ZGD functional protocol is not as straightforward as the GRIP or SOAP bindings.

A set of resources are defined to represent the ZGD (Figure 7.8), each ZigBee network and each ZigBee node (Figures 7.9 and 7.10), each resource is addressable with an

³ISO 8825, 1998: Information Technology. ASN.1 Encoding rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER). International Standard ITU-T X690 (1997) | ISO/IEC 8825-1:1998.

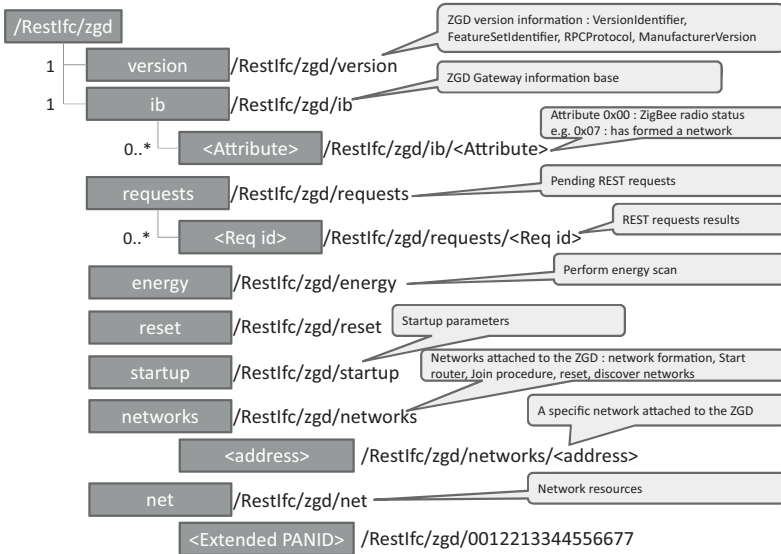


Figure 7.8 ZGD REST resources overview.

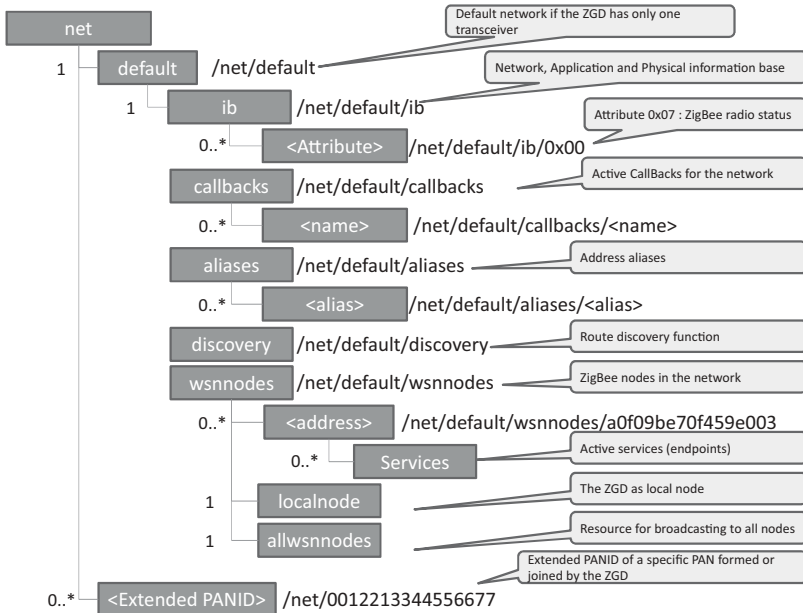


Figure 7.9 Overview of network and node resources.

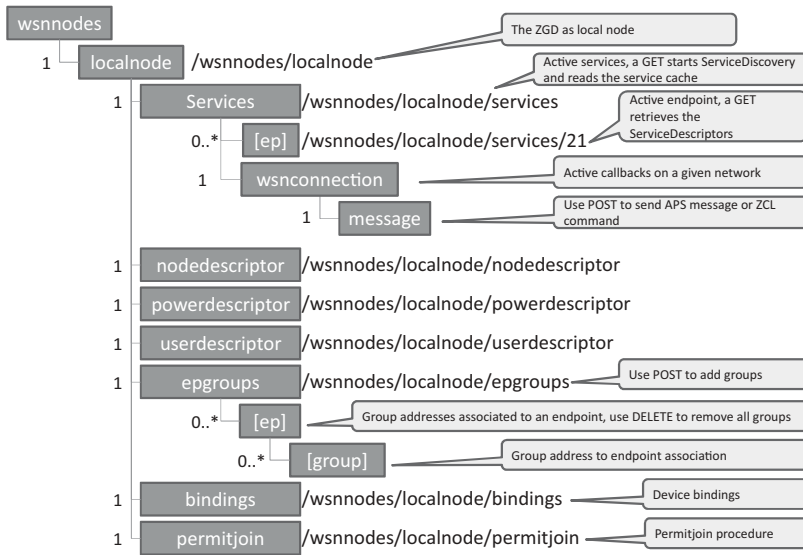


Figure 7.10 Overview of ZigBee node resources.

URL. Each ZGD function is implemented as a state transfer between the IPHA and the appropriate resource: For instance, the GetVersion function, used by the IPHA to discover the features of a ZGD, is implemented as a read operation on the “version” resource of the ZGD.

Asynchronous notifications (responses and events) use a special resource hosted by the IPHA. The IPHA declares the URI of this resource to the ZGD using the CreateCallback REST operation, and can also specify a specific callback URI in each REST request supporting the CallbackDestination parameter.

The resources are represented by XML documents, specified as part of the REST binding.

7.10.5 Example IPHA–ZGD Interaction Using the REST Binding

The IPHA may read any resource exposed by the ZGD, for instance

```
HTTP GET:
http://''ZGD_IP_Addr:ZGD_PORT''/RestIfc/zgd/version
```

And the ZGD responds with the requested resource representation


```
<tns:Info xmlns:tns='http://www.zigbee.org/GWGSchema' >
  <tns:Status>
    <tns:code>0x00</tns:code>
  </tns:Status>
  <tns:Detail>
    <tns:Version>
      <tns:VersionIdentifier>0x01</tns:VersionIdentifier>
      <tns:FeatureSetIdentifier>0x00</tns:
        FeatureSetIdentifier>
      <tns:RPCProtocol>0x0004</tns:RPCProtocol>
      <tns:ManufacturerVersion>1.1</tns:ManufacturerVersion>
    </tns:Version>
    <tns:Detail>
  </tns:Info>
```

In this example we suppose that the IPHA needs to allocate a dedicated endpoint on the ZGD. It therefore begins by registering a new endpoint on the ZGD. In the following example we use a synchronous transaction. The IPHA could have use an asynchronous method by specifying a callback URI in parameter *CallbackDestination* of request URI sent to the ZGD.

HTTP POST: HTTP POST: http://''ZGD_IP_Addr:ZGD_PORT''/
localnode/services

```
<?xml version='1.0' encoding='UTF-8'?>
<tns:SimpleDescriptor
xmlns:tns='http://www.zigbee.org/GWGSchema' >
<tns:ApplicationProfileIdentifier>0x0104</tns:ApplicationProfile
Identifier>
<tns:ApplicationDeviceIdentifier>0x0002</tns:ApplicationDevice
Identifier>
<tns:ApplicationDeviceVersion>0x00</tns:ApplicationDevice
Version>
<tns:ApplicationInputCluster>0x0000</tns:ApplicationInput
Cluster>
<tns:ApplicationInputCluster>0x0003</tns:ApplicationInput
Cluster>
<tns:ApplicationInputCluster>0x0004</tns:ApplicationInput
Cluster>
<tns:ApplicationInputCluster>0x0005</tns:ApplicationInput
Cluster>
<tns:ApplicationInputCluster>0x0006</tns:ApplicationInput
Cluster>
<tns:ApplicationOutputCluster>0x0003</tns:ApplicationOutput
Cluster>
</tns:SimpleDescriptor>
```

The IPHA did not request a specific endpoint, therefore it will be allocated by the ZGD. If the request succeeds, the IPHA will receive a response similar to the following:

```
<tns:Info xmlns:tns='http://www.zigbee.org/GWGSchema' >
  <tns:Status>
    <tns:code>0x00</tns:code>
  </tns:Status>
  <tns:Detail>
    <tns:endpoint>0x23</tns:endpoint>
  </tns:Detail>
</tns:Info>
```

This is an example of a synchronous response. When using an asynchronous mode, the response would include a request identifier attribute allocated by the ZGD.

The IPHA has the option of registering a callback function bound to this endpoint or to all endpoints, by invoking an HTTP POST request to `[NwkRootURI]/callbacks` that specifies filters which will be used by the ZGD to identify messages that will be notified to the IPHA.⁴

The IPHA may then instruct the ZGD to send any command. The endpoint identifier used in the URI is 35 (0x23). For instance an APSMessage:

```
POST [NwkRootURI]/localnode/services/35/wsnconnection/message
<?xml version='1.0' encoding='UTF-8'?>
<tns:APSMessage xmlns:tns='http://www.zigbee.org/GWGSchema' >
  <tns:DestinationAddress>
    <tns:NetworkAddress>0x0001</tns:NetworkAddress>
  </tns:DestinationAddress>
  <tns:DestinationEndpoint>0x02</tns:DestinationEndpoint>
  <tns:SourceEndpoint>0x23</tns:SourceEndpoint>
  <tns:ProfileID>0x0104</tns:ProfileID>
  <tns:ClusterID>0x0000</tns:ClusterID>
  <tns:Data>0102030405060708090a0b0c0d0e0f</tns:Data>
  <tns:TxOptions>
    <tns:SecurityEnabled>true</tns:SecurityEnabled>
    <tns:UseNetworkKey>true</tns:UseNetworkKey>
    <tns:Acknowledged>true</tns:Acknowledged>
    <tns:PermitFragmentation>true</tns:PermitFragmentation>
  </tns:TxOptions>
  <tns:Radius>3</tns:Radius>
</tns:APSMessage>
```

⁴Callbacks may also be registered using resources `[NwkRootURI]/localnode/services/[ep]/wsnconnection` or `[NwkRootURI]/localnode/allservices/wsnconnection`, in which case filters are preset for all APS messages matching the endpoint id.

If the APS command is successful, the IPHA will receive the following response from the ZGD:

```
<tns:Info xmlns:tns='http://www.zigbee.org/GWGSchema' >
  <tns:Status>
    <tns:code>0x00</tns:code>
  </tns:Status>
  <tns:Detail>
    <tns:APSMessageresult xmlns:tns='http://www.zigbee.org/GWGSchema' >
      <tns:ConfirmStatus>0x00</tns:ConfirmStatus>
      <tns:TxTime>0x01234567</ts:TxTime>
    </tns:APSMessageresult>
  </tns:Detail>
</tns:Info>
```

8

Z-Wave

8.1 History and Management of the Protocol

The Z-wave protocol was designed by private company Zensys (www.zen-sys.com), based in the US and Denmark. Zensys is now a subsidiary of SIGMA Designs, a provider of system on chip (SoC) products for the multimedia and entertainment industry. Zensys started by introducing to the market, in 2001, a light-control system for consumers, and evolved its product to a full-fledged home area network meshed protocol implemented in a proprietary SoC.

Z-wave quickly became a very popular home automation protocol, with hundreds of products sold on web sites like www.zwaveproducts.com or www.zwaveworld.com. Zensys customers appreciate that Z-wave offers approximately the same features as its standard competitor 802.15.4 (ZigBee), but with fewer interoperability issues (due to the multi-vendor nature of the ZigBee ecosystem), and an unlicensed frequency band (868 MHz) that is often perceived to be less problematic than the crowded 2.4 GHz band.

Zensys OEMs using the protocol are grouped in the Z-wave alliance (www.z-wavealliance.org), which promotes awareness of the product, organizes developer's forums and interoperability testing events ("Unplugfest"). Z-wave products are certified by Zensys or agreed labs, using tools provided by Zensys, which test various network management functionalities and assess the communication error rate (CER) of the device under test at various distances.

The evolution of the system is managed by Zensys technical service (TS, <http://support.zen-sys.com>). This section describes the third generation of the chip and protocol stack, as available in March 2010 (ZW0302 SoC). There is another version in preparation, but not yet released.

8.2 The Z-Wave Protocol

8.2.1 Overview

In the design of Z-wave, Zensys targeted a precise market segment: low-cost and reliable home area networks. In other words “better than X10”¹ (that means reliable), but still low cost: Zensys targets the 3 USD per SoC price segment, about 30% of the cost of RF systems capable of multimedia communications such as WiFi.

The chip has been optimized for battery-powered nodes: it remains in “power down” mode (consuming about 2.5 μA for the ZW0302 SoC) most of the time, and “Wakes-up” only for brief intervals from time to time to perform its function, receive and transmit. Wake-up occurs at programmable intervals or if an interrupt is raised, and at these moments the SoC power consumption is of course higher (about 25 mA when receiving or transmitting).

The Z-wave offering, as provided by Zensys developer’s kits to its customers, covers the physical layer (RF link), the link and routing layers (called by Zensys the MAC, transfer and routing layers). The transport layer is a simple resend scheme, and there is no concept of a session. APIs are provided at the routing-layer level for application developers.

RF home area networks typically use very low transmission power (many RF nodes are battery powered), in a complex environment with multiple reflections of the signal, diffraction, and so on. In this context, the strategy selected by Zensys for Z-wave is to build a mesh network, so that nodes with no power-consumption constraints can relay the radio signals from room to room. In a mesh network, there is no need for a node to be able to communicate directly with any other node: the mesh protocol builds routing tables ensuring that any node has a valid path, via selected nodes, to any other node in the network.

8.2.2 Z-Wave Node Types

From a *radio perspective*, there are two types of nodes:

- Devices that can enter sleep mode. All battery-powered devices are in this situation. They cannot be used by the Z-wave network as relays to other nodes.
- *Always listening* nodes, identified by a specific flag, can be used as repeaters in the Z-wave networks because their radio is always listening. Usually, most devices powered from AC mains are “always listening”.

From a *radio-routing perspective*, Z-wave defines two main types of nodes:

- *Controllers*, which have and maintain the full topology of the network, and can assign routes to slaves;

¹ X10 is a popular CPL-based low-cost home automation protocol, also known for its unreliability (commands are not acknowledged and easily affected by parasitic noise on power lines).

- *Slaves*, which have a limited knowledge of the network topology and no functionality related to the maintenance of the network topology (e.g., including or excluding nodes).

8.2.2.1 Controllers

Controllers maintain a complete network topology map, and therefore can calculate routes to reach any node in the network. Zensys defined several subtypes of controllers:

- *Primary Controller*. A controller that has the capability to include or exclude nodes in the network, maintains the list of nodes in the network and calculates the route table.
- *Secondary Controller*. A controller that receives a copy of the list of nodes and of the route table. In the case of a failure of the primary controller, a secondary controller can take the role of a primary controller.

Some other variants have appeared in successive versions to fulfill more evolved needs:

- *Portable Controller*. A controller that is typically mobile and battery powered, therefore not always listening. In the first generations of Z-wave networks, the node inclusion was performed at a reduced radio power, therefore there was a need for a portable controller. In more recent networks, other strategies can be used (inclusion controller, inclusion at full power or explorer frame, see below) and therefore a portable controller is not required.
- *Static Controller*. Installed in a fixed location and mains powered, a static controller is always listening. A static controller can be primary, but should preferably be a secondary controller if the network uses low power inclusion, due to range limitations. A static controller can optionally be configured as:
 - a *Static Update Controller (SUC)*: a SUC is a network database that gets updates from the primary controller about all network changes, and can send network topology information to secondary controllers and routing slaves that request updates. The SUC function is assigned to a static controller by the primary controller. Secondary controllers with SUC functionality can take over the role of the primary controller if it fails.
 - *SUC ID Server (SIS)*. This optional functionality of a SUC was introduced with Zensys Developer's Kit 3.40. The SIS enables other controllers to include/exclude nodes in the network on its behalf: when a SIS is present all controllers will become *inclusion controllers* instead of primary and secondary controllers. They can include/exclude nodes to a network and provide correct home ID and node ID as long as these inclusion controllers can reach the controller with SIS functionality.
- *Bridge Controller*: designed to control up to 128 devices located on other automation networks (e.g., LonWorks, BACnet, KNX or X.10) by emulating as many virtual slave Z-wave nodes itself.

8.2.2.2 Slaves

A Z-wave node that is mains powered (always listening) and is able to execute or route commands from other Z-wave devices in the network, with no other network

functionalities, is a *Slave*. Typically, a Z-wave power plug is a slave node. A slave device must be in a fixed position and must be listening at all times. Slave devices have no topology map and cannot calculate routes. Since they are always listening, slaves can receive commands via routes (from controllers or routing slaves) and send replies back using the same route (the *response route*). Slaves must act as repeaters in the network, forwarding messages to the next hop that must be provided in the route that is part of the message.

A node that can send unsolicited messages to other nodes in the network is a *routing slave*, for example a battery-powered temperature sensor. Routing slaves can receive commands and send replies, but they have a partial topology map and cannot calculate routes. Therefore, they can only address a subset of the network to send unsolicited frames (up to 5 nodes), and must store static routes to the targets of the unsolicited messages. They can also store several routes to the static controller with SUC/SIS functionality and ask for route updates. Routing slaves can optionally act as a repeater (*slave*) in the network if they are mains powered.

Z-wave defines 3 subtypes of routing slaves:

- *FLiRS Routing Slave*. Frequently listening routing slave (FLiRS). This is a normal routing slave configured to listen for a wakeup beam in every wake-up interval. This enables other nodes to wake up the FLiRS and sends a message to it. Only ZDK 5.0x supports FLiRS nodes. One usage for a FLiRS could be as the chime node in a wireless doorbell system.
- *Enhanced Slave*: is a routing slave, with an external EEPROM to store application data.
- *ZensorNet™ Routing Slave*: these nodes add low latency point to point communication support for battery-powered nodes in a Z-Wave® network, as required by simple smoke-alarm applications. The ZensorNet routing slave node is basically a routing slave node configured as FLiRS with the additional functionalities of Zensor net binding and Zensor net flooding. See Section 8.2.4.1 for details.

8.2.3 RF and MAC Layers

Z-wave chips use a B-FSK “spectrum shaping” modulation over a 868.42 MHz carrier frequency in the EU, 908.42 MHz in the USA, 921.42 MHz in Australia, and 919.82 MHz in Hong Kong.

The chip output power is programmable from -20 to 0 dBm, and the chip is able to decode signals as low as -102 dBm for a data rate of 9.6 kbps (about 30 m indoors, and 150 m in line of sight conditions). The maximum bitrate attainable with the current version is 40 kbps in favorable link conditions. It is lower than the maximum bitrate of 802.15.4 (ZigBee) systems, which operate around 100 kbps. However, Zensys plans to add support for other frequencies (950 MHz and 2.4 GHz) in Z-wave 6.00, which also targets a higher maximum bitrate over 100 kbit/s.

7	6	5	4	3	2	1	0
Home ID (0)							
Home ID (1)							
Home ID (2)							
Home ID (3)							
Source Node ID							
Routed	Ack	Low Power	Reserved	Header Type			
Reserved							
Length							
Destination Node ID							
Payload byte 0-x							
...							
Checksum							

Figure 8.1 Z-wave singlecast frame format.

Z-wave frames begin with a preamble, and encapsulate the transmitted data between a start of frame marker and an end of frame marker. The data itself is Manchester encoded at 9.6 kbps, and NRZ encoded at 40 kbps.

The collision-avoidance mechanism is based on delaying transmissions randomly after the last occurrence of RF activity on the channel.

8.2.4 Transfer Layer

The Z-wave transfer layer controls the transmitting and receiving of frames.

The general format of a Z-wave datagram transfer layer frame is illustrated in Figure 8.1.

Singlecast frames are sent to one specific node ID, and can optionally include a route (list of slave nodes to the destination). A flag indicates if an acknowledgment is desired or not (Acks are singlecast frames with no data, and generated hop by hop, with some optimization mechanism). Singlecasts get retransmitted if no Ack is received.

A **Multicast frame** can target a range of selected nodes (1 to 232). The format of this frame is identical to that of a unicast frame, except that the destination node ID is replaced by an address offset and one or more of mask bytes. The receiving nodes do not acknowledge a multicast frame. If reliable communication is required, a multicast frame must be followed by a singlecast frame.

A **Broadcast frame** targets all nodes in the Z-wave network with the specified Home ID and is not acknowledged. It is used for instance to transmit the *node information frame* when a node action button is pressed. The format of this frame is identical to that of a unicast frame, except that the destination node ID is set to FFh (255).

7	6	5	4	3	2	1	0
Home ID (0)							
Home ID (1)							
Home ID (2)							
Home ID (3)							
Source Node ID							
Routed	Ack	Low Power	Reserved	Header Type			
Reserved							
Length							
Destination Node ID							
Ver			Cmd				
Reserved						Direction	Source routed
SessionTxRandomInterval							
Session TTL				Repeater Count			
Repeater 0							
Repeater 1							
Repeater 2							
Repeater 3							
Checksum							

Figure 8.2 A Z-wave explorer frame.

An *Explorer frame* (Figure 8.2) is a special class of the broadcast frame. All nodes in the direct range of the originator receive an explorer frame. The handling of an explorer frame then depends on the destination address (a single node or all nodes), and a number of configuration flags, which determine:

- whether the frame should be forwarded by the receiving node specifically, or all nodes that receive it;
- whether frames with identical source node ID and sequence number should be discarded.

A *SearchRequest* addresses a particular node but must be forwarded by all nodes (network flooding). A *SearchResult* addresses a particular node and must be forwarded only by all nodes in a route. A *SearchStop* addresses a particular node and must be forwarded only by all nodes in a route but all nodes must discard frames resembling the frame (identified by srcNodeId + seqNo), in order to stop the network flooding initiated by a *SearchRequest*.

An explorer frame can carry a command to a destination NodeID through a list of repeaters, and is used for autoinclusion (network wide inclusion) and route resolution. The SessionTTL parameter, initially set to 4, is decremented by each forwarding node, ensuring that the flooding process terminates.

8.2.4.1 Battery-Powered Nodes

Battery-powered nodes pose a specific problem, because they are awake only periodically, and the device timers are not accurate enough to maintain any form of synchronization that would allow the sender of a frame to “guess” when to send it.

In a Z-wave network, sending configuration information to a routing slave that can enter sleep mode involves polling: the routing slave can use a “dial up” service to for example, signal to a static controller that they are ready to receive new configuration information. For this purpose the *wake-up command class* is used. Such nodes must know the route to the static controller (these routes are called *return routes*), and periodically issue a *wake up notification* command to any always listening device (it may be broadcast if the target node ID has not been configured by the *wake up interval set* command), requesting for pending commands. The target node will send any pending commands, then issue a *wake up no more information* command so that the battery-powered node can go back to sleep.

This communication method implies a relatively high latency, which is not acceptable to all applications. An example of such an application is a network of smoke detectors, which are all required to sound an alarm signal as soon as one detector has been triggered. For this type of application Zensys introduced the notion of ZensorNet™ routing slaves.

ZensorNet™ routing slaves use a specific a specific link layer mechanism. They use a wake-up beam, ZensorNet binding and ZensorNet flooding (no routing). When a ZensorNet™ device needs to send a frame to a network of ZensorNet™ routing slaves, it uses a wake-up beam longer than the configured wake up interval of devices in the network. This ensures that all other ZensorNet™ devices will receive the wake-up beam. After receiving such a beam, each ZensorNet™ devices starts beaming its own wake-up beam, flooding the network. Each device beams a given frame only once, preventing flooding loops.

Up to 16 ZensorNet™ routing slave nodes can exist in a ZensorNet™ network.

8.2.5 Routing Layer

The routing layer controls the routing of frames in the network. Z-wave networks use a source-routing mechanism: the initiator of a frame generates a complete route to the end destination through a number of repeaters.

The route consists in a sequence of node IDs that is placed in the frame, as illustrated in Figure 8.3. The frame becomes a *routed singlecast* (Figure 8.3). An always listening node receiving a *routed singlecast* with its node ID at the top of the repeater’s list will repeat this frame, just removing its node ID from the list of repeaters. This ensures that routing loops cannot occur.

A routed acknowledge frame has the same structure, but the Ack flag is set and it carries no data in it.

If a device has tried all configured routes to a target without success, it issues a special “SearchRequest” explorer frame for the target node ID, embedding the command it was

7	6	5	4	3	2	1	0
Home ID (0)							
Home ID (1)							
Home ID (2)							
Home ID (3)							
Source Node ID							
Routed	Ack	Low Power	Reserved	Header Type			
Reserved							
Length							
Destination Node ID							
Reserved				Routed Err	Routed Ack	Direction	
Repeaters				Hops			
Repeater 0							
Repeater 1							
...							
Repeater n							
Reserved							
Reserved							
Payload byte 0-x							
..							
Checksum							

Figure 8.3 Z-wave routed singlecast frame.

trying to send. All slave nodes that receive this frame forward a copy, and hopefully the frame finally reaches the target device. The command is executed only once even if multiple copies reach the destination, because they all have the same sequence number. The target device then sends a “SearchReply” back along the route that was just discovered, and the initiating device learns this new route. A SearchStop explorer frame is sent immediately, ensuring that all pending search frames are killed, stopping network flooding as soon as possible.

8.2.5.1 Overview of Z-Wave Addressing: Home ID, Node ID, Primary and Secondary Controller, Inclusion

Z-wave uses a 32-bit *Home ID* identifier to separate different Z-wave networks (e.g., two adjacent houses). Each device is allocated an 8-bit *node ID*, unique for a given network identified by its *home ID*. A Z-wave network can have up to 232 nodes for any given *home ID*.

The *home ID* is configured² in each Z-wave *primary controller*. During the *inclusion* process the Z-wave *primary controller* allocates a unique *node ID* to each Z-wave device, and also configures the *home ID* allocated to each device.

²Generated randomly and renewed after each controller exclusion, for Zensys SDK 4.5 and above.

When a *controller* is included in an existing Z-wave network (controlled by a *primary controller*), it becomes a *secondary controller*, and is assigned a unique *Node ID* and the same *home ID* as the *primary controller*. In addition, for battery-powered nodes, the primary controller assigns a return route to the device (see Section 8.2.4.1)

8.2.5.2 Inclusion

With most current Z-wave networks, the inclusion process is manual. An action must be taken on the controller to enter a waiting mode, then a button must be pressed on the Z-wave peripheral to be included (one peripheral at a time). The peripheral sends a *node information frame* (NIF) to the controller, which must be reachable directly by the device during the association process. The NIF specifies:

- The *basic device class* that defines the protocol library used (portable controller, static controller, slave, and routing slave).
- The *device class*, for example, generic controller, static controller, binary switch, multilevel switch, binary sensor, multilevel sensor The device class describes the type and functionality of a device, and implies mandatory command classes that must be supported. The generic device class defines the broad functionality of a device (e.g., multilevel switch), and the specific device class defines more precisely the function of the device (e.g., multilevel power switch for a light dimmer, motor control class for automated window shades, both of generic device class multilevel power switch).
- All *command classes* supported by the device (except secure commands, see security).

The controller replies by assigning a Node ID to the device, and formally ending the transfer.

On recent Z-wave networks supporting the “*explorer frame*” (see Figure 8.2), the peripheral being associated does not need to be reachable directly by the controller during the association process: the association request and subsequent controller configuration commands can be routed through other nodes already part of the same network. Zensys calls this the “*autoinclusion*”, because devices supporting this feature enter “autoassociation mode” automatically on power-up. A periodic association request is sent (after increasing random intervals in order to minimize collisions) via explorer frames. The explorer frame floods the network, via adjacent nodes, with a TTL that prevents loops, trying to reach a static controller. When the static controller finally receives the request, it returns a routed response.

Immediately after inclusion, a node or secondary controller is requested by the controller to look for its neighbor nodes on the network (by sending NOP frames and listening for ACKs), and reports the node IDs of the devices within range back to the controller.

For any given node N (and secondary controllers), the primary controller knows which nodes are reachable by N (the node neighbor list). From this information it builds its own

routing table, and sends all relevant information to the secondary controller to build its local routing table:

- The node table (node IDs as well as device/command classes);
- The node neighbor list;
- The SUC ID server (see controllers), if any.

Optionally, the controller can exchange information about groups (names and nodes) and scenes (names, nodes and levels in nodes), see application layer for details.

8.2.5.3 Network Management

When an “always listening” node is moved or becomes unreachable with the current routing configuration due a changing RF environment, Primary controllers or controllers with SUC/SIS functionality can initiate a “rediscovery” procedure to get new routes to the moved nodes, healing the network. A similar rediscovery procedure exists for “not listening” nodes.

As the node ID is allocated by the network controller, in case a node fails, it is possible to replace that node while preserving the node ID.

8.2.6 Application Layer

8.2.6.1 Commands

The Z-wave protocol supports the applications by defining a standard command structure, ensuring interoperability. Each command encapsulated in a Z-wave frame is composed of:

- An 8-bit command class identifier. Command classes are defined for the Z-wave protocol and for Z-wave applications, for example, `COMMAND_CLASS_BATTERY`.
- An 8-bit command identifier (an extension mechanism ensures that up to 4000 commands can be defined).
- A list of command parameters. For any command, the list of parameters can be extended as new versions are released. Devices supporting older version truncate the list of parameters to just those that they can understand. Devices can be queried for the command class version that they support.

Devices declare the supported command classes during the inclusion process, as part of the *node information frame* that is sent as a broadcast any time the node action button is pressed or as a response to a controller “*get node information*” command.

Most devices support the *basic command class*. Through primitives such as `BASIC_SET` or `BASIC_GET`, it is possible to set or read values for simple devices (like

dimmers, power plugs or thermometers) which do not require a more complex command syntax.

Most devices will support additional command classes besides the basic command class, for example, `COMMAND_CLASS_BATTERY` for battery-powered devices, `COMMAND_CLASS_SENSOR_MULTILEVEL` for multilevel sensors (e.g., temperature, luminosity, hygrometry sensors)

For some uses, the list of commands already defined might be insufficient: an extension mechanism is defined for proprietary commands. Zensys assigns manufacturer IDs to ensure that proprietary commands from distinct manufacturers do not interfere.

8.2.6.2 Multi-Instance Devices

Some devices implement several functions, for example, multiple relays, multiple switches, or sensors for multiple physical parameters (e.g., temperature, humidity, pressure). As a consequence, it may not be enough to send a command to a node ID to identify the exact action desired. Z-wave solves that issue with the concept of multi-instance devices (MI). Multi-instance devices advertise support for the *Multi-Instance* command class in the *network information frame*.

The *multi-instance command encapsulation* command is used to encapsulate commands sent to a Z-Wave node, and simply adds an instance parameter to the encapsulated command.

8.2.6.3 Associations

Most home area networking and fieldbus protocols define a form of “publish/subscribe” model. This intent of this feature is to allow devices to interact directly without requiring services from a central node: this improves the reactivity of the system (e.g., the time it takes to switch on a lamp after operating the switch), and the resistance to failures of the central controller.

Z-wave implements the “publish/subscribe” model through the concept of *associations*. Z-wave calls an “association” the operation by which a Z-wave device is configured to control another device or set of devices. For instance, a switch or movement sensor can control a lamp or group of lamps.

For each type of event that a controlling device can generate (e.g., “switch on event”), it should allocate a locally unique 8-bit *grouping ID*. Up to 255 types of events can be managed per device.

Associations are handled at the application level of the device code, not by the Zensys libraries. Associations are supported by two command classes:

- `COMMAND_CLASS_ASSOCIATION`: The *association set* command is used to add nodes, identified by their node IDs, to a given grouping identifier. The *multi-instance*

association set command can be used for multi-instance devices (both a node ID and an Instance ID are provided). Routing slaves must also be configured by a controller with return routes in order to be able to reach the target nodes. Additional commands allow the controller to retrieve the association status of the nodes, remove associations, request the number of groupings, and so on.

- **COMMAND_CLASS_ASSOCIATION_COMMAND_CONFIGURATION** defines the commands necessary to configure which commands should be sent to a node part of a grouping ID when the corresponding event occurs. A *command record* consists of the grouping identifier, the Node ID and the complete command. The *command configuration set* command is used to specify which command should be sent to a specific node ID within a given grouping identifier. Additional commands allow retrieval of the maximum number of command records, the remaining number of free command records, and so on.

Both primary and secondary controllers can configure groupings and the required return routes. Ordinary nodes can setup command records in other nodes if they support commands from the **COMMAND_CLASS_ASSOCIATION_COMMAND_CONFIGURATION** class.

8.2.6.4 Scenes

Z-wave *scenes* provide the toolbox that facilitates the simultaneous configuration of multiple devices, with minimal network activity and avoiding the flicker effects that might appear if individual configuration commands were sent to each individual device. This can be used, for instance, to preset lamp levels for common situations, for example, “home theater”, “dinner”, “chat around the fireplace” configurations.

Scene settings can be configured in devices using commands of defined by the scene actuator configuration command class (management of settings associated to a given scene ID), and activated by means of the commands defined in the scene activation command class.

Several scenes can be configured in a controller. Each scene defines a list of node IDs part of this scene, and an 8-bit level parameter (a value that will have the same effect on devices as if it was passed through a basic set command).

Once a scene has been configured, it can be activated using the **SCENE_ACTIVATION_SET** command (which can be multicast), specifying a scene ID and a dimming duration.

8.2.6.5 Security

Z-wave network security is optional, and implemented by commands defined in the *security command class*. When security is provided, for example, for door locks, it is

based on a network-wide secret key (created by the primary controller at startup) which is used to encrypt *security-encapsulated secure commands*.

A device willing to send a *security-encapsulated secure command* to another starts by requesting a nonce from the target Node ID (using a nonce get, nonce report message exchange). The value of this nonce, as well as a source generated initialization vector, will be included in the security-encapsulated secure command, ensuring replay prevention. A message authentication code is also added to the message to prevent tampering. The destination node ID is sent in clear as part of a standard singlecast, therefore secure messages can be routed through nonsecure nodes.

AES 128 is used for authentication (message authentication code based on a Davies–Meyer hash) and encryption. The current SoC generation requires a software implementation. Hardware support for security has been announced for the fourth generation of the chip, ZW0401, which will implement AES-128 in hardware.

The network secret key is distributed during initial installation: right after the inclusion of a secure node, the primary controller sends *key set* security encapsulated commands to the secure node, using a temporary key (all zeroes).³

Developers decide, at the application level, which commands should be supported securely, and declare these commands to the controller separately (security commands supported get/report). All commands declared through the standard *node information frame* (NIF) are nonsecure.

³ Future implementations may use a PIN-code-based encryption during the initial network key exchange.

Part Three

Legacy M2M Protocols for Utility Metering

9

M-Bus and Wireless M-Bus

9.1 Development of the Standard

The M-Bus standard was the result of collaboration between Dr. Horst Ziegler of the University of Paderborn, a chip maker (Texas Instruments) and a company focused on metering data management (Techem). In 1990 there was no established communication standard for the reading of utility metering devices: such a standard had to be low cost and adapted to battery-powered meters.

The original design uses a simple two-wire serial communication bus, and is documented at <http://www.m-bus.com/>. One major advantage of the new bus was that all meters and the reading device could be connected to the same wire (which is why it is called a bus).

The link layer used by M-Bus was initially standardized in 1990 as IEC 870-5-1 (Telecontrol Equipment and Systems/Transmission Protocols/Transmission Frame Formats) and IEC 870-5-2 (Link Transmission Procedures, 1992). The first standard to be published related to the M-Bus application layer was EN 1434 in 1997, where parts -2 and -3 define an application layer for a wire communication protocol dedicated to heat meters.

The standardization work is now managed by Cenelec Technical committee TC 294, which generalized the use of M-Bus for any type of meter readout in the EN 13 757 series:

- EN13757-1:2002 Data exchange (DLMS);
- EN13757-2:2004 Physical and link layer (M-Bus);
- EN13757-3:2004 Dedicated application layer (M-Bus);
- EN13757-4:2005 Wireless meter readout (wM-Bus);
- EN13757-5:2007 Relaying (network aspects);
- EN13757-6:2007 Data exchange (local bus).

The Internet of Things: Key Applications and Protocols, First Edition.

Olivier Hersent, David Boswarthick and Omar Elloumi.

© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

Table 9.1 M-Bus wired data transmission, physical layer bit encoding

	Master transmission	Slave transmission
Bit 0	+36 V	<1.5 mA
Bit 1	+24 V	11–20 mA

EN 13 757-2 defines the physical and link layer for the wired bus, while EN 13 757-3 defines the application layer (actual messages). The wireless version of M-Bus is defined in EN 13 757-4.

9.2 M-Bus Architecture

M-Bus was not originally designed for complex multihop networking, and it uses a simplified 4-layer model.

9.2.1 Physical Layer

The two-wire bus interconnects one master and several slaves, using asynchronous half-duplex serial data transmission. Slaves can be powered from the bus.

The nominal voltage of the bus is 36 V. Serial data transmission from the master to slaves uses bus voltage level shifts, while data transmission from the slaves to the master uses a modulation of the slave current consumption (Table 9.1).

The quiescent state is always a 1. The start bit of the serial transmission is therefore a “0”, followed by 8 data bits, followed by a parity bit and a stop bit (always 1).

The communication speed can vary between 300 and 9600 bits/s, and up to 250 slaves can be connected over a single twisted pair, which can span up to 1000 m, depending on the number of slaves.

9.2.2 Link Layer

The data link layer is based on IEC 870-5, which defines several frame formats depending on the level of integrity protection required. M-Bus uses frame format class FT 1.2, and defines 4 types of frames (Figure 9.1).

The **single-character frame** is used to acknowledge transmissions.

The **short frame** has a fixed size.

The **long frame** has a length field (user data bytes + 3) that is transmitted twice, followed again by the start character.

The **control frame** is a long frame without user data.

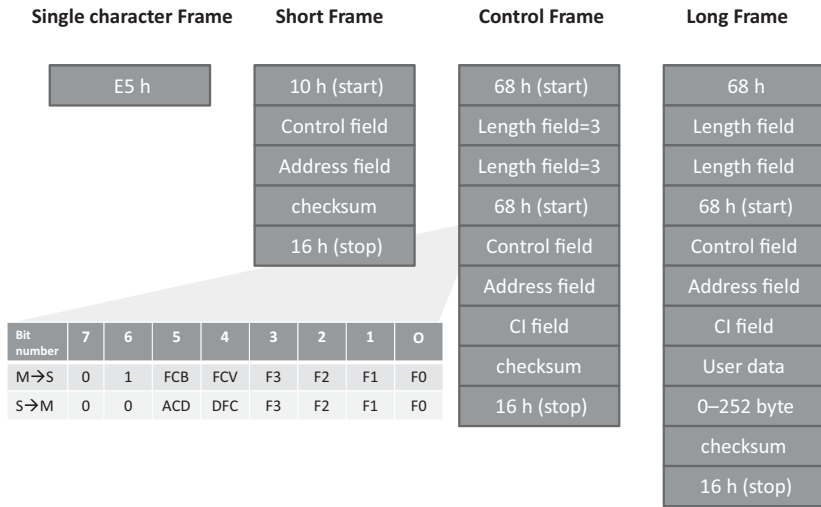


Figure 9.1 M-Bus frames.

Primary addresses 1 to 250 can be allocated to slaves. Address 0 is reserved for unconfigured slaves. Address 253 is reserved to indicate that the destination address has been set at the network layer. Addresses 254/255 are reserved for managing the local physical layer.

The bits of the control field (C) have a different use depending on the direction of the communication. The FCB bit is used as a one-bit frame counter (used if FCV bit is set). The DFC bit is used by slaves to signal that it can accept no further data. The ACD bit is set when a slave wants to transmit higher-priority data.

The CI field (control information) is part of the application layer.

The link layer defines two transmission services:

- Send/Confirm (SND/CNF): the response is a single-character frame. SND_NKE is used to start or restart a communication (the next master frame will have NCB = 1). SND_UD (optional) is used by the master to send user data to the slave.
- Request/Response (REQ/RSP): REQ_UD2 is used by the master to request data from the slave, and the slave can transfer its data to the slave using RSP_UD.

9.2.3 Network Layer

The M-Bus network is structured into zones, which consists of one or more segments interconnected by repeaters.

The standard provides a way to switch a given slave in a “selected” state, without using or requiring a primary address. In order to achieve this, the master sends a special

Table 9.2 M-Bus, common CI field values

00h-4Fh	Reserved for DLMS applications
54h-58h	
50h	Application reset
51h	Data send (master to slave)
5Ch	Synchronize action
70h	Slave to master: report of application errors
71h	Slave to master: report of alarms
72h	Slave to master: 12-byte header followed by variable format data
78h	Slave to master: variable data format response without header

SND_UD message with CI=52h or 56h to primary address 253 (FDh), which specifies the secondary address of the slave (identification number, manufacturer, version and medium, specified or wildcarded). All slaves receiving this message compare these information elements with their own, the slave that matches, if any, enters “selected” state and sends a single-character response.

Subsequent requests to the selected slave(s) may be sent by SND_UD messages addressed to primary address FDh. Slaves remain in selected state until they receive a new selection command with a nonmatching secondary address.

The selection procedure, because it allows wildcarded parameters for individual address digits composing the secondary address, can also be used by masters for network discovery.

9.2.4 Application Layer

Control information (CI) field values indicate the formatting of the rest of the application data payload. Table 9.2 lists common values of the CI field.

Example user data formatting for CI=72h:

- A 12-byte header shown in Table 9.3.
- Variable data blocks: composed of one or more data records. Each data record is composed of a data record header (DRH), followed by the data itself.

The DRH describes the data:

- The data information block (DIB) of the DRH describes the length, type and coding of the data (e.g., 6-digit BCD, variable length 8-bit text string) as well as the storage number of the data concerned (register identifier of the value being read), and whether the value being read is an instant value, minimum value or invalid.
- The value information block (VIB) contains the value of the unit and the multiplier. For instance a value information field of “E000 0nnn” corresponds to an energy measurement in Wh from 0,001 Wh to 10 000 Wh with a multiplier of $10^{(nnn-3)}$. See Table 9.4 for common unit codes.

Table 9.3 Application data header example

Ident. Nr	Manufacturer	Version	Device Access		Status	Signature
			type	No		
Serial number allocated during manufacture. 8 BCD digits	Derived from 3-letter EN 61107 ASCII code of manufacturer ID	Determined by manufacturer	Response counter	Error flags, e.g., Power Low	Used for encryption	
4 byte	2 byte	1 byte	1 byte	1 byte	1 byte	2 byte

Table 9.4 M-Bus value information field unit codes

Coding	Description	Range Coding	Range
E000 0nnn	Energy	10(nnn-3) Wh	0.001 Wh to 10 000 Wh
E000 1nnn	Energy	10(nnn) J	0.001 kJ to 10 000 kJ
E001 0nnn	Volume	10(nnn-6) m ³	0.001 l to 10 000 l
E001 1nnn	Mass	10(nnn-3) kg	0.001 kg to 10 000 kg
E010 00nn	On Time	nn = 00 s nn = 01 min nn = 10 h nn = 11 day	
E010 01nn	Operating Time	<i>as above</i>	
E010 1nnn	Power	10(nnn-3) W	0.001 W to 10 000 W
E011 0nnn	Power	10(nnn)J/h	0.001 kJ/h to 10 000 kJ/h
E011 1nnn	Volume Flow	10(nnn-6) m ³ /h	0.001 l/h to 10 000 l/h
E100 0nnn	Volume Flow ext.	10(nnn-7) m ³ /min	0.0001 l/min to 1000 l/min
E100 1nnn	Volume Flow ext.	10(nnn-9) m ³ /s	0.001 ml/s to 10 000 ml/s
E101 0nnn	Mass flow	10(nnn-3) kg/h	0.001 kg/h to 10 000 kg/h
E101 10nn	Flow Temperature	10(nn-3) °C	0.001 °C to 1 °C
E101 11nn	Return Temperature	10(nn-3) °C	0.001 °C to 1 °C
E110 00nn	Temperature Difference	10(nn-3) K	1 mK to 1000 mK
E110 01nn	External Temperature	10(nn-3) °C	0.001 °C to 1 °C
E110 10nn	Pressure	10(nn-3) bar	1 mbar to 1000 mbar
E110 110n	Time Point	n = 0 date n = 1 time and date	IEC 870-5-4 CP16 8-bit encoded date IEC 870-5-4 CP32 32-bit encoded date
E110 1110			dimensionless

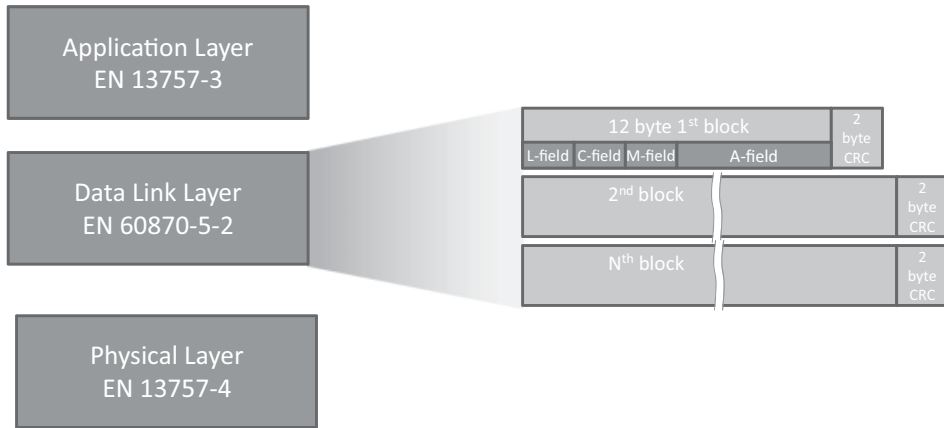


Figure 9.2 The wireless M-Bus protocol layers, details of the link layer.

9.3 Wireless M-Bus

Wireless M-Bus is specified EN13757-4:2005 (wireless meter readout) and uses the 868 MHz frequency, a lower frequency (169 MHz) enabling better penetration within buildings is being standardized as well (prEN13757-4:2011 Mode N). The application messages are specified in EN 13 757-3 as for the wired M-Bus.

The wireless M-Bus specification defines communications between a “meter” device and a “other” device, which usually corresponds to a data concentrator.

Wireless M-Bus uses the 3-layer IEC model, outlined in Figure 9.2.

9.3.1 Physical Layer

The physical layer is described by EN 13 757-4 “Communication system for meters and remote reading of meters, Part 4: Wireless meter readout, Radio meter reading for operation in the 868 to 870 MHz SRD band”.

Several data transmission modes have been defined on the 868 MHz license free ISM band defined in Europe (Table 9.5): the stationary mode (S-mode), the frequent-transmit mode (T-mode) and the frequent-receive mode (R-mode). The physical and data link layers of the S-mode have been designed jointly with the KNX association, and are identical to KNX-RF (see Chapter 6). More recently, the N-mode, at 169MHz, has been defined.

All modes have been designed to optimize the battery lifetime, the meter remains in sleep mode except during short transmission periods. Some modes enable bidirectional communication: the meter switches to receive mode for a short period after transmission. Even in bidirectional mode, the “other” side can never initiate a transmission, but needs to wait until the meter wakes up, transmits at least a message, and then switches to receive mode.

Table 9.5 wM-Bus transmission modes

Mode	Data rate from meter (chip rate)	Direction	
R2	4.8 kbps	Meter → Other	
S1/S1m	32 768 kbps	Meter → Other	868.3 MHz (600 kHz channel), 2FSK modulation.
S2		Meter ↔ Other	
T1	100 kbps	Meter → Other	
T2		Meter ↔ Other	868.95 MHz for other to meter, 500 kHz channel.

Mode T provides the shortest transmission time and the longest battery life for a wireless meter. The battery-life requirements are typically 10 to 20 years.

Wireless M-Bus bit encoding use a higher frequency for encoding chip “1” and a lower frequency for encoding chip “0”. Transmissions begin by a preamble composed of an integer number of “01”, followed by a synchronization word.

For data, the wireless M-Bus specification uses two channel-encoding methods:

- Manchester encoding: data bit “1” is encoded by chip pattern “01”, while bit “0” is encoded by chip pattern “10”. This is a classic encoding that was used to facilitate clock recovery, although with most modern radio chips, it would no longer be required. The resulting data rate is half of the chip rate.
- Three out of six encoding: each byte of data is split in nibbles of 4 bits, each 4-bit nibble is then encoded as 6 chips. As a result, each data byte is encoded in one and a half chip bytes.

The encoded data is then followed by a trailer of two to eight chips.

The various physical layer parameters used for each transmission mode are summarized in Table 9.6.

Table 9.6 Wireless M-Bus physical layer parameters

Mode	Preamble (chips)	Sync word (chips)	Encoding
R2	78	18	Manchester
S1	558 (about 17 ms !)	18	Manchester
S2 (two-way)	30	18	Manchester
T 1(one-way)	38	10	3 out of 6 encoding
T2 other to meter			Manchester

9.3.2 Data-Link Layer

The data-link layer is defined in IEC 60 870-2-1:1992, Telecontrol equipment and systems, Part 5: Transmission protocols, Section 1: Link transmission procedure, and IEC 60 870-1-1:1990, Telecontrol equipment and systems, Part 5: Transmission protocols, Section 1: Transmission frame formats.

The data-link layer manages framing and CRC generation and detection, provides physical addressing and manages acknowledges (bidirectional modes only).

The wireless M-Bus frame format used in EN 13 757-4:2005 is derived from the frame type 3 (FT3) defined in EC60870-5-2. The frame consists of one or more blocks of data, as illustrated in Figure 9.2. Each block includes a 16-bit CRC field. The first block is a fixed length block of 12 bytes that includes the L-field, C-field, M-field, and A-Field.

- L-Field: in variable packet size mode, the transmission frame format begins by a length indication (L-field), which indicates the number of link-layer payload bytes not including the L, M and A fields, CRC bytes or encoding, from 0 to 255. The L-field itself is encoded using Manchester or 3 out of 6 encoding. Due to the various overheads, the actual maximum value of the L field may be limited by the underlying radio chip packet handler.
- C-Field: the C-field identifies the frame type: SEND, CONFIRM, REQUEST, or RESPOND. For SEND and REQUEST frames, the C-field also indicates whether a CONFIRM or RESPOND is expected.
- M-Field: the manufacturer's 3-letter code (encoded as 3×5 bits by taking the ASCII code and subtracting $0x40$). Assigned codes are listed at <http://www.dlms.com/flag/INDEX.HTM>
- A-Field: each device is assigned a unique 6-byte address, assigned by the manufacturer. As in the case of the wired M-Bus, only meters are assigned addresses: the address included in the frame is always the originator address for send and request frames, and the target address for confirm and response data frames. The "other" side remains anonymous.
- CI-Field: the application header specified the application payload type. The possible values are specified in EN13757-4:2005.
- CRC: a 2-byte CRC is added for each 16-byte block, using polynomial $x^{16} + x^{13} + x^{12} + x^{11} + x^{10} + x^8 + x^6 + x^5 + x^2 + 1$.

9.3.3 Application Layer

The application layer is defined in EN 13 757-3, Communication system for meters and remote reading of meters, Part 3: Dedicated application layer. It is identical to the application layer of the wired M-Bus, which facilitates bridged M-Bus applications connecting a wired side with a wireless side.

9.3.4 *Security*

M-Bus data can optionally be encrypted. The original specification used the DES algorithm, while more recent specifications now use the AES 128 algorithm.

When using encryption, the meter is configured with a key, which is shared by the master. Some headers are always sent unencrypted (e.g., the header of Table 9.3), and a portion of the header indicates the encryption method (high byte of the signature field in the example of Table 9.3, for instance DES cipher block chaining). The encrypted data follows.

10

The ANSI C12 Suite

Jean-Marc Ballot

Alcatel-Lucent

10.1 Introduction

Before the publication of C12, meters from different vendors in the USA did not use the same data formats or communication protocols. C12 is a standard suite specified by the American National Standards Institute (ANSI), which provides an interoperable solution for data formats, data structures, and communication protocols used in automatic metering infrastructure (AMI) projects. Although this standard is focused on the American national market, C.12 is used in smart metering systems of many countries (in particular its variant OSGP, deployed by Echelon, Inc).

The standard data structure is specified in the ANSI C12.19 document. It is defined as a set of tables. When these tables share a common purpose or they are relative to a common feature of the meter, then the tables are included in a specific chapter called a “decade”.

The first communication protocol that made use of the C12.19 tables was specified in the ANSI C12.18 document. The first release was published in 1996 and describes the communications between a C12.18 meter and a C12.18 client by means of an optical port. A C12.18 client could be a hand-held reader, a laptop, or any device with an optical port.

In 1999, ANSI C12.21 was specified for communications between a C12 device and C12 client via a modem. This offered a first solution for AMI projects.

In 2007, a new specification took into account the development of data networks not using modems. ANSI C12.22 is a “Protocol Specification for Interfacing to Data Communication Networks”. The goal of this new item of the C12 suite is to allow interactions with C12.19 table data over any networking communications system.

Beyond the ANSI specifications, RFC 6142 “ANSI C12.22, IEEE 1703 and MC12.22 Transport Over IP”, published in March 2011, proposes a framework for transporting ANSI C12.22 Application Layer messages on an IP network.

10.2 C12.19: The C12 Data Model

ANSI C12.19 is the “American National Standard for Utility Industry End Device Data Tables”.

C12.19 defines a data structure used for representing metering data and metering functions exposed by a metering equipment to a client machine. C12.19 does not contain any protocol for the transport of the data, only the data structure is specified. As briefly mentioned in the introduction, the data structure is defined as a set of standard tables. When these standard tables share a common purpose or they are relative to a common feature of the meter, then the tables are included in a specific chapter called a “decade”. Each decade covers a specific area of functionality. The version of ANSI C12.19 published in 2007 contains 17 decades (the original version published in the 1990s contained fewer decades than the current version). Table 10.1 provides the list of C12.19 decades:

Beyond these standard tables, ANSI C12.19 also provides a standard way to add proprietary tables. These tables are called manufacturer tables. If they follow the general

Table 10.1 C12.19 table decades

Decade number	Name of the Decade	Number of Tables in the Decade
0	Configuration Tables	9
1	Data Source Tables	9
2	Register Tables	9
3	Local Display Tables	5
4	Security Tables	7
5	Time-of-Use Tables	7
6	Load Profile Tables	8
7	History & Events Logs	10
8	User-Defined Tables	10
9	Telephone Control Tables	9
10	Extended Source Tables	4
11	Load Control & Pricing Tables	9
12	Network Control Tables	Temporarily defined in ANSI C12.22
13	Relay Control Tables	Temporarily defined in ANSI C12.22
14	Extended User Defined Tables	4
15	Quality of Service Tables	9
16	One-Way Tables	5

rules for the table format, it is possible for a manufacturer to introduce some value-added functions in its products.

ANSI C12.19 carefully defines all the data types that are used in the definition of the data structure. The communication protocol is not described, only the data format is specified.

The transport of table structures is not specified by C12.19 but it is mentioned that this transport “*is dependent only on the presence of basic Read and Write services (e.g., those as defined in ANSI C12.18, ANSI C12.21 and ANSI C12.22)*” (extract from C12.19 section 8). The design of C12.19 is natively “RESTful” (see Chapter 13)! The structure of the tables ensures that any operation required to manipulate the C12.19 tables can be performed only with the basic read and write services. ANSI C12.19 provides some basic requirements for the read and write services, which must be implemented by all C12 devices, but manufacturers may additionally implement more primitives to interact with C12.19 tables.

10.2.1 *The Read and Write Minimum Services*

- The read service request allows the transfer of table data from a sending party to a receiving party. The read service can be used for full table read or a partial table read. In the case of full transfer, only the Table_Identifier is provided in the read request. In the case of partial table read, some additional optional parameters have to be provided in order to choose the records and record fields that are requested. Two addressing methods are specified for a partial read:
 - a first method by providing up to 5 indexes relative to the table and optionally an element count starting at the indexed position;
 - a second one by providing an offset (in octets) relative to the beginning of the table and optionally an octet count starting at the indicated position.

If the end device that receives the read request does not support the method, the entire table is retrieved.

- The write service allows nonsolicited data to be sent to a receiving party. As for the read service, the write service request allows a complete or a partial table write to be performed, and the partial table write service can use indices or an offset.

Besides the read and write service, C12.19 supports 27 standard commands, and even manufacturer-defined commands, but the implementation of commands only uses the basic read and write services, using special-purpose tables (Tables 07 and 08). C12.19 does not compromise with the REST design!

10.2.2 *Some Remarkable C12.19 Tables*

The first table of the first decade, Table 00, plays a special role. It is called GEN_CONFIG_TBL and contains all the information relative to the configuration of the end device. For example, it contains the full list of supported tables and procedures.

Tables 07 and 08 in Decade 0 also play a specific role. They are called “Table 07: Procedure Initiate Table” and “Table 08: Procedure Response Table” and are used for enabling the execution of commands. When an initiator wants to request the execution of a command in a meter, it has to write in Table 07 some parameters that explicitly provide information about the procedure to be executed. The command response, that is, the result of the procedure execution, is placed in Table 08 in order to be read by the initiator of the command.

It has to be noted that these two special-purpose Tables 07 and 08 are not able to buffer commands. They enable execution of only one command at a time: the specification explicitly mentions that “*If a procedure initiate request is followed by another procedure initiate request, the procedure response for the first procedure initiate request may be lost*”.

The list of procedures that may be executed by using Tables 07 and 08 contains 28 standard procedures. Among the 28 procedures we can mention:

- cold start;
- warm start;
- save configuration;
- remote reset;
- set date and/or time;
- execute diagnostics.

10.3 C12.18: Basic Point-to-Point Communication Over an Optical Port

ANSI C12.18 was the first standardized protocol that was specified to interact with ANSI C12.19 Data Tables. The first release was published in 1996 then revised in 2006. It describes the communications between an electric metering equipment and another device used as a client via an optical port. The client device is typically a hand-held reader or a laptop used for reading or writing the meter internal data.

ANSI C12.18 focuses on the physical, data link and application layers. Layers 3 to 6 of the OSI model are out of scope. The three main functional areas covered by C12.18 are the following:

- modification of the communication channel;
- transport of information to and from the metering device;
- closure of the communication channel when communications are complete.

The application layer defines the PSEM (protocol specifications for electric metering) language that provides basic services that are used for channel configuration and

information retrieval. Each service uses a request–response scheme. Nine services are defined:

- Identification service: this is the first service that shall be invoked after the establishment of the physical connection. The version and revision numbers of the protocol are returned by the service.
- Read service: this is used for triggering the transfer of table data from the requested device to the requesting one. As mandated by ANSI C12.19, both complete and partial transfer are possible. The complete transfer is mandatory. The partial transfer may use one of the two possible options: index based or offset based.
- Write service: this is used to transfer a table data to a target device. As mandated by ANSI C12.19, both complete and partial transfer are possible. The complete transfer is mandatory. The partial transfer may use one of the two possible options: index based or offset based.
- Logon service: this is used to setup a session without establishing the access permissions yet. These permissions will be established later through the security service.
- Security service: this is used for establishing access permissions. It is based on the use of a password as a mean for selecting access permissions. This service cannot be invoked before the logon service because a session has to be established as a prerequisite. The received password is compared with the one stored in the password table of the security decade.
- Logoff service: this is used for terminating the session previously established via the logon service.
- Negotiate service: this is an optional service used to reconfigure the communication channel in the case of the desired communication parameters differ from the default values. Baud rate and packet size are among the negotiable parameters.
- Wait service: this is used for maintaining an already established communication channel beyond the time-out value that ensures automatic termination. The value of the time-out will be reset to the previous value once a valid packet is received.
- Terminate service: this is used for immediately interrupting the communication channel. Generally, this service is used in the case of excessive errors or security issues.

Besides these high-level application layer services, ANSI C12.18 also provides settings for Layer-2 and Layer-1 establishment. Baud rate, number of packets, packet size channel traffic time-out, data type, data format and data polarity are among the handled parameters.

10.4 C12.21: An Extension of C12.18 for Modem Communication

ANSI C12.21 “Protocol Specification for Telephone Modem Communication” is an extension to the C12.18 standard. C12.18 was the first standardized protocol allowing interaction with ANSI C12.19 tables, but it was still necessary to be in the immediate

vicinity of the C12 Device when handling the tables. C12.21 allows remote interactions over a telephone network.

The three main area of functionalities already provided by C12.18 are not modified (i.e. modification of the communication channel, transport of information, and closure of the communication channel).

The PSEM (protocol specifications for electric metering) now contains 12 services instead of 9 in the C12.18 specification.

- 7 services are identical to those in C12.18: read, write, logon, security, logoff, negotiate, wait. Actually, logoff service is very slightly modified in the terminology of its description but not in its functionality.
- 2 services (identification and terminate service) are modified compared with their C12.18 versions.
 - Identification service: the modification implements basic negotiation of the authentication algorithm. If authentication is supported, then the authentication itself will be performed by calling a new service of C12.21: the authenticate service (see below).
 - Terminate service: it is used for immediately returning the communication channel to its “base state” that is, the state in which the channel is still established but with the default parameters. In this state there is no established session.
- 3 new services are provided: timing setup, disconnect, and authenticate.
 - Timing_Setup service: this is an optional service that allows configuring some timers or number of retry attempts used in the communication channel establishment, when these values differ from the default values.
 - Disconnect service: this is used for immediately interrupting the communication channel. The disconnect service in the C12.21 is a redesign of the terminate service in the C12.18, main use case for this service is to interrupt the communication when too many errors or security issues are observed.
 - Authenticate service: this new service was required as a result of the modification of the identification service. The C12.21 identification service negotiates the authentication algorithm supported by the end device. After establishing a session with the logon service, the authenticate service will be used in order to perform mutual authentication at session level.

10.4.1 Interactions with the Data-Link Layer

They are quite limited. The communication channel of the modem is established with a set of default parameters. The service layer only has the possibility, after calling the identification service and before calling the logon service, to call either the negotiate service or the Timing_Setup service (or both) in order to modify packet size, packet number for reassembly, timers, or retry attempts number.

10.4.2 *Modifications and Additions to C12.19 Tables*

Beyond the modification of existing services or addition of new services, C12.21 also specifies some changes in the C12.19 tables. Some existing tables were modified and some new tables were added.

Some of the most significant changes are listed below:

- C12.19 Table 07 (procedure initiate table) was modified in order to add a new standard procedure that did not exist in the original version of the C12.19. This new procedure triggers an immediate call establishment with a phone number specified as a procedure parameter.
- A new decade (no. 9) was added to the original C12.19. This decade contains 7 new tables associated with the use of a telephone modem.

10.5 C12.22: C12.19 Tables Transport Over Any Networking Communication System

ANSI C12.22 “Protocol Specification for Interfacing to Data Communication Networks” was made necessary by the development of new networking technologies. The approach of C12.18 that defines a communication protocol for a given network was no longer practical. C12.19 introduces new concepts enabling transport C12.19 data from meters to a back-end central system over any kind of communication network.

ANSI C12.22 defines several types of network elements that are used in a reference topology. Interfaces between different types of network element are described in the standard. Some new data tables are also added to the ANSI C12.19-1997 standard as required by the new C12.22 interfaces. Some existing tables are also modified in order to ensure compatibility with the new C12.22 standard.

10.5.1 *Reference Topology and Network Elements*

The reference topology defined by C12.22 is outlined in Figure 10.1.

This reference topology makes use of different types of network elements:

- C12.22 Host: this is a termination point in a C12.22 network. It may be an authentication host or/and notification host. An authentication host performs the authentication tasks for a registering node. A Notification Host is the applicative part that is able to interpret the C12.19 data structures and that needs to be notified when new nodes are registered.
- C12.22 Device: this is a network element that contains a C12.22 application (the C12.19 data structures, the associated protocol, and the control of associations). In order to enable communications between a C12.22 device and the C12.22 network, a

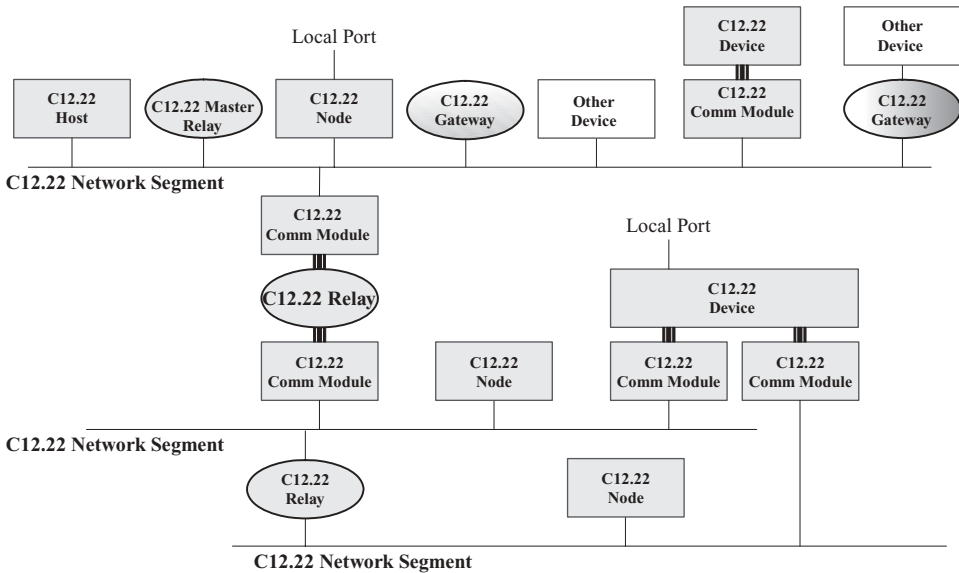


Figure 10.1 C12.22 reference topology (ANSI C12.22, Chapter 5, Figure 5.1).

C12.22 device has to implement one (at least) interface to a C12.22 communication module.

- **C12.22 Communication Module:** this is a hardware device that implements the Layers 6 to 1 of the OSI model allowing communications between a C12.22 Device and a C12.22 network through an interface fully defined in the C12.22 specification.
- **C12.22 Node:** it is a combined C12.22 communication-module/device network element.
- **C12.22 Master Relay:** this is a network element in charge of receiving the registration requests coming from the C12.22 network elements that belong to its domain. It is also in charge of propagating the registration request to the appropriate C12.22 authentication host. A C12.22 master relay is on top of a hierarchical topology of C12.22 relays. A C12.22 master relay contains all the functionalities of a C12.22 Relay.
- **C12.22 Relay:** this is a network element that ensures the address translation between the Layer 7 address of a C12.22 node and its network address. This layer 7 address is called ApTitle (application process title). A C12.22 relay also uses the ApTitle in order to provide a C12.22 message-forwarding service when the lower layers do not provide such forwarding capability. In this case, when a message is sent to a C12.22 Relay with a called ApTitle different from the relay's ApTitle, then the C12.22 relay is in charge to forward the message to the final destination or to the first C12.22 relay in the path. A C12.22 relay maintains internal routing tables in order to provide this routing service.

- C12.22 Gateway: this is a protocol converter from the C12.22 protocol to any other protocol. It is used for enabling communications between a C12.22 node and a non-C12.22 node.

10.5.2 C12.22 Node to C12.22 Network Communications

The protocol stack used for communicating between a C12.22 node and a C12.22 network is only defined at layer 7 that is, at the application layer. The other layers (6 to 1) are “open to any network protocol”. This application layer obviously contains the C12.19 data tables and also provides an evolution of the C12.21 PSEM (protocol specification for electric metering). This new version of the PSEM protocol contains 13 services:

- Three services are unchanged: the read, write and security services.
- Six services are modified (compared to C12.21): identification, logon, logoff, terminate, disconnect, and wait services.
- Four new services are provided: registration, deregistration, resolve, and trace services.
 - Registration service: this is used by a C12.22 network element in order to declare itself to the hierarchical structure of C12.22 relays. A C12.22 element has to send a registration request to a C12.22 master relay. During the initial registration a new routing table entry is added to all the C12.22 relays on the path to the master relay. Routing table entries are a soft state, and subsequent periodic registration requests and necessary to keep the routing table entry valid. After receiving a registration request, a C12.22 master relay has to forward it to the C12.22 authentication host.
 - Deregistration service: the effect of this service when it is called by a C12.22 network element is a deletion of the corresponding routing table entry from all the C12.22 master relay and relays. The removal of this network element is taken into account by all other C12.22 elements including authentication and notification hosts.
 - Resolve service: this enables communication between two C12.22 nodes that belong to the same local area network. When a requesting C12.22 node (X) needs to directly communicate with another local C12.22 node Y, it sends a resolve request with the ApTitle of node Y to its C12.22 relay in order to retrieve the native network address of node Y.
 - Trace service: when invoked by a C12.22 node, this service returns the path of C12.22 relays between the requesting node and a target C12.22 Node. The target node is not involved in the processing of the trace service that is only performed at the C12.22 relay level.

As in C12.18 and C12.21 standards, partial table access is possible in C12.22 specification by using one of the two defined methods: index or offset based.

An extended mode of the PSEM, called EPSEM (extended PSEM) is specified. EPSEM allows sending multiple requests and receiving multiple responses simultaneously.

In order to convey the APDUs (application protocol data units) that contain EPSEM services and their associated payload, the C12.22 standard uses the ACSE (association control service element) encoding method specified in ISO 8650-1. ACSE is an envelope for EPSEM primitives that also allows to transport association parameters and some security parameters when a secure transaction is required (C12.22 security mechanism supports both authentication and encryption).

10.5.3 C12.22 Device to C12.22 Communication Module Interface

In order to model the communication ports of C12.22 meters, C12.22 introduces the concept of C12.22 communication modules. A given meter may support multiple types of “plug-in” communication modules, using a standard connector and serial protocol.

As represented in Figure 10.2, a communication module is connected to the C12.22 Device through an interface that is fully defined in the C12.22 standard. It is also connected to any LAN (e.g., ZigBee, . . .), WAN (DSL, GPRS, . . .), or MAN (Ethernet, . . .). When a short-range connection is used, the communication module may communicate with a C12.22 relay that implements the same network technology.

The C12.22 device/C12.22 communication module protocol stack is fully defined in C12.22, and is outlined in Figure 10.3.

The application layer in a C12.22 device (and optionally in a C12.22 communication module) is identical to the one implemented in a C12.22 node and described in the previous section.

As mentioned in Figure 10.3, the transport layer specifies a set of 6 services that have the following functionalities:

- Negotiate service: this is used when a C12.22 communication module detects an attached C12.22 device. The service negotiates communication parameters settings with the C12.22 device. Typical examples of communication parameters that may be negotiated

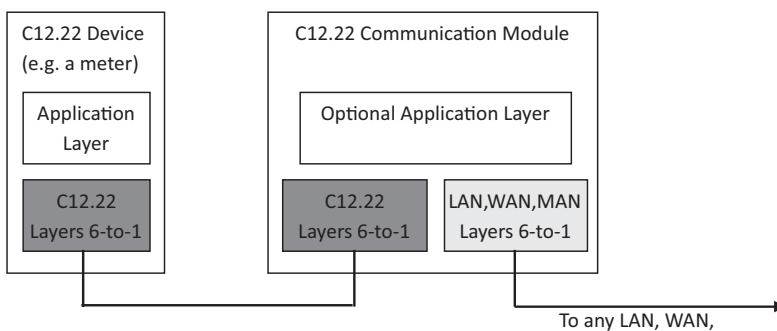


Figure 10.2 C12.22 Communication module.

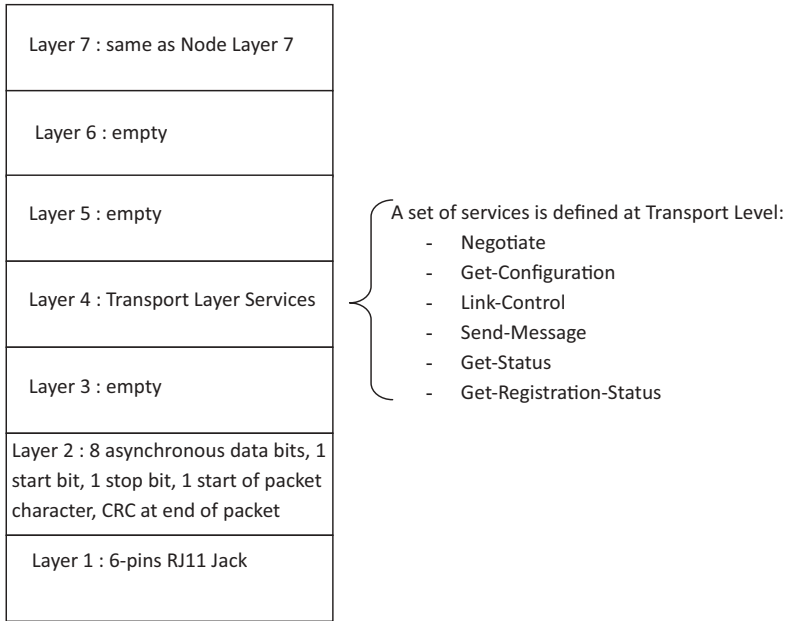


Figure 10.3 C12.22 Communication module to node protocol stack.

by invoking this service are baud-rate, maximum-packet-size, maximum-number-of-packet for reassembly function, . . .

- **Get-Configuration service:** this service is invoked by the communication module in order to request its configuration from the C12.22 Device. ApTitles, native address, or device class are among the exchanged parameters.
- **Link-Control service:** by using this service the C12.22 device can control communications from the communication module to the C12.22 network. It can enable or disable the communication interface, the registration process, direct communication with target nodes on the same network segment, and other functions.
- **Send-Message service:** this is the service enabling exchange of ACSE-PDUs. It is used by the C12.22 device when sending a message to a C12.22 communication module or by a communication module when transmitting a message received from the C12.22 network.
- **Get-Status service:** the C12.22 device may use this service in order to get information about the communication interface of the C12.22 communication module. The device may also request some network statistics about the communication status.
- **Get-Registration-Status service:** after a registration is performed, by invoking this service the C12.22 device can retrieve registration information from the communication module. The returned information includes relay address and ApTitle, registration period, and the amount of time left before the end of the current registration.

10.5.4 C12.19 Updates

C12.22 updates some existing tables and adds new decades to the existing C12.19 data tables:

- Decade 12 “Node Network Control Tables” is added and contains 8 tables modeling the C12.22 node access to a C12.22 network
- Decade 13 “Relay Control Tables” is added and contains 7 tables in relation with the management of a C12.22 relay.
- The content of the Table 07 (procedure initiate table) is augmented with 4 new procedures related to the newly added Decade 12: Registration, Deregistration, Network Interface Control, and Exception Report.

10.6 Other Parts of ANSI C12 Protocol Suite

The previously described C12 standards relate to networking and communication aspects. C12 contains additional specifications, for example:

- ANSI C12.01: “Code for Electricity Metering” defines some conditions for a set of tests performed on electricity metering equipment. C12.01 defines these tests and the associated acceptable performance criteria in terms of overload support, the effect of the variation of numerous parameters, the effect of heat, the effect of a magnetic field, and so on.
- C12.10: “Physical Aspects of Watthour Meters – Safety Standard”. As mentioned in the NEMA abstract “This standard covers the physical aspects of both detachable and bottom-connected watthour meters and associated registers. These include ratings, internal wiring arrangements, pertinent dimensions, markings and other general specifications”.
- C12.20: “Electricity Meters – 0.2 and 0.5 Accuracy Classes”. This standard defines requirements for electricity meters in terms of voltage or frequency ranges, form designation, displays, and so on. Based on a list of 38 standardized tests, it also defines the acceptable performance for an electricity meter.

10.7 RFC 6142: C12.22 Transport Over an IP Network

RFC 6142 “ANSI C12.22, IEEE 1703, and MC12.22 Transport Over IP” provides a framework for transporting ANSI C12.22 data over an IP network. IEEE 1703 and MC12.22 are similar specifications in IEEE and Measurement Canada environments.

When ANSI C12.22 defines a layer 7 protocol and proposes to transport it over any underlying protocol, RFC 6142 proposes to restrict the scope to the transport of C12.22 messages by using TCP and UDP transports over an IP network. RFC 6142 more precisely

focuses on the adaptation of Chapter 5 of the ANSI C12.22, which is related to the “C12.22 Node to C12.22 Network Segment Details”.

All the C12.22 network elements considered in RFC 6142 are natively IP aware. In case of a C12.22 IP relay, RFC 6142 only deals with the IP interface and not with the other possibly non-IP aware interfaces that may be used for message forwarding to C12.22 non-IP nodes.

RFC 6142 describes a C12.22 IP network segment in a general manner, without any intention to provide any guidelines on its size. A small LAN or a full C12.22 IP network are equally possible.

In order to convey the RFC 6142 C12.22 messages in a standardized way, port number 1153 was assigned by IANA (Internet Assigned Number Authority) for both TCP and UDP.

RFC 6142 specifies an encoding for the native IP address in order to standardize the use of IP addresses in the appropriate fields of ANSI C12.19 Tables. IPv4 and IPv6 are two possible options.

The support of IP multicast is required in all C12.22 hosts, relays and master relays and recommended in the C12.22 nodes in order to facilitate the reading of numerous C12.22 meters. In this case the meters have a common C12.22 multicast group `ApTitle` and can be reached by sending a single EPSEM read request. Two specific IPv4 and IPv6 multicast addresses have been assigned by IANA to a newly created “All C1222 Nodes” multicast group (224.0.2.4 for IPv4 address and `FF0X::24` for IPv6). The use of a TTL (time to live) attribute in an IP packet header allows the propagation of C12.22 IP multicast messages to be limited.

C12.22 allows the use of two connection modes: a connection-oriented mode and a connection-less mode. RFC 6142 maps these modes to the use of TCP or UDP. For each type of C12.22 network elements, depending on their ability to support TCP or UDP and depending on their ability to be able to accept unsolicited new datagrams or connection requests, RFC 6142 defines a set of basic rules for correctly handling the application associations and the exchanges of UDP or TCP messages.

ANSI C12.22 contains its own security mechanism and does not mandate any transport layer security. RFC 6142 allows the use of a transport layer security mechanism as an enhancement to the C12.22 security feature.

10.8 REST-Based Interfaces to C12.19

Although this was not an explicit design principle, the design of C12 happens to be fully REST compliant. This is a lucky circumstance as modern smart-grid designs recommend the use of a REST style architecture. At present, no formal IP based REST interface for C12 has been proposed, however, a tentative interface for the ETSI TC M2M is presented in Chapter 14.

11

DLMS/COSEM

Jean-Marc Ballot,* and Olivier Hersent
* *Alcatel-Lucent*

11.1 DLMS Standardization

11.1.1 *The DLMS UA*

The Device Language Message Specification¹ user association was formed in 1997 by utilities and manufacturers to develop open standards for multiutility (all energy types) meter data exchange, for all application segments. As of 2010 it counts over 180 members, as well as multiple associate member organizations: ESMIG, M-Bus, Euridis.org, Selma, DVGW, PPCEM and the ZigBee Alliance. Over 140 meter types, from over 40 manufacturers, have been certified.

The DLMS UA maintains the specification, is the registration authority for IEC 62 056 (OBIS codes), performs technical support and training, and operates the conformance specification scheme. The DLMS UA is organized in two working groups:

- The Maintenance and Development WG, handling the development of the standard.
- The Final End Users and Developers WG, focused on use cases and gathering feedback from deployment and interop testing, led by French utility EDF.

11.1.2 *DLMS/COSEM, the Colored Books*

DLMS/COSEM separates the aspects of data modeling, data identification, messaging and transport:

¹ The initial name, from French Utility EDF, was “Distribution Line Message Specification”.

- COSEM, the companion specification for energy metering, specifies the data model, that is, the standard object interfaces, with their attributes and methods. It maintains a registry of object interfaces (OBIS data identification codes). xDLMS messaging is used to access COSEM objects attributes and methods.
- DLMS itself is an application layer protocol that defines abstract object-related services and protocols. Out of the original 22 services defined by DLMS, DLMS/COSEM uses only a subset of 4 messaging services, as well as a few extensions. This profile is named xDLMS.
- DLMS supports multiple transport layers: Twisted pair, power line, IP, and so on.

The DLMS specification is documented in 4 “books”, which can be purchased on DLMS.com:

- The Blue book “COSEM – Identification System and Interface Classes” – DLMS UA 1000-1:2009, Ed. 9.0, 2009-02-09, specifies the data model (COSEM interface classes and OBIS codes for various energy types). The Blue book has been internationally standardized by IEC and CEN.
- The Green book “DLMS/COSEM – Architecture and Protocols” – DLMS UA 1000-2:2009, Ed. 7.0, 2009-12-22, specifies the protocols with DLMS on top, for the various media specific communication profiles. The Green book has been internationally standardized by IEC and CEN.
- The Yellow book specifies conformance test plans for COSEM object model.
- The White book is a glossary of terms.

Together, these books represent over 600 pages of specifications, and the specification is still rapidly expanding. The first implementation of DLMS/COSEM was deployed in 1999. In 2002 the specification was published as IEC and CEN standards. More recently, the standard was adopted by SM-CCG and OPENmeter consortium as the core standard for smart metering. DLMS/COSEM is also published as a standard in China and in India.

11.1.3 DLMS Standardization in IEC

The DLMS-UA work was co-opted by IEC TC13 (International Electrotechnical Commission, Technical Committee 13). The IEC TC13 is in charge of “electrical energy measurement, tariff and load control”. In Europe, the CENELEC TC13 mirrors the IEC TC13. IEC TC 13 endorses the DLMS colored books in their IEC 62 056 series.

The contents of the “Green book” are reflected in the following IEC standard documents:

- **IEC 62 056-42:** Physical layer services and procedures for connection-oriented asynchronous data exchange;
- **IEC 62 056-46:** Data-link layer using HDLC protocol;

- **IEC 62 056-47:** COSEM transport layers for IPv4 networks;
- **IEC 62 056-53:** COSEM application layer.

The contents of the “Blue book” are reflected in the following standard documents:

- **IEC 62 056-61:** Object Identification System (OBIS).
- **IEC 62 056-62:** Interface classes.

The colored books are also reflected in CENELEC standards, for example, EN 13 757 part 1 for the Blue book COSEM interface object model.

11.2 The COSEM Data Model

COSEM is an object model for metering applications that is utility-type and communication-media independent. COSEM uses the client–server paradigm. In the COSEM model, the meter is the server. A COSEM server only models the elements of the meter that are visible externally. COSEM data structures are specified in ASN.1 syntax.

A COSEM server (a physical metering device) is modeled as a set of “logical devices”, hosted in a single physical device. Each logical device models a subset of the functionalities of the physical meter. A logical device is implemented as an application process (AP).

A logical device is composed of a set of COSEM interface objects. Interface objects model various functions of the meter and they are accessible from the client side through the communication interfaces of the meter. Each interface object is a collection of attributes and methods. The structure of objects that have common characteristics is described once for all in an interface class. The interface classes are specified in the DLMS Blue book.

These high-level data model principles are represented in Figure 11.1.

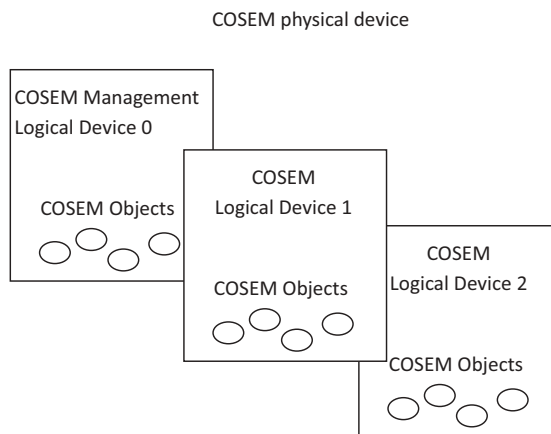


Figure 11.1 DMLS data model overview.

Each physical DLMS device shall provide one “management logical device” which contains the list of logical devices in the physical device (“service access point assignment”) and may contain one or several other logical devices. Each logical device is identified by a unique COSEM logical device name object (LDN), an octet string of up to 16 octets starting with a 3-octet manufacturer identifier. Each logical device also contains an application association object, which contains in its object_list attribute the list of visible COSEM objects in the context of the application association session with a given client (the “association view”).

Note: The DLMS and ZigBee high-level data models are very similar: DLMS Logical Devices are equivalent to ZigBee applications, DLMS interface objects are equivalent to ZigBee clusters. The DLMS Blue book is the equivalent of the ZigBee cluster library (ZCL). The fundamental difference is that ZigBee clusters behave as servers or clients, enabling peer to peer communication, while DLMS interface objects are only servers. Another difference is that the Blue book focuses on metering and basic meter related I/O control, while the ZCL scope extends beyond metering.

The object-oriented concept behind COSEM is that any real-world “thing” can be described by some attributes and methods. For the various applications (e.g., energy, billing, load profiles, instant power measurement, I/O control, access control), COSEM maps the “thing” to a COSEM object composed of:

- A set of attributes. Attributes have a meaning, a data type, a value range, and access rights.
- A set of methods (e.g., reset, start).

Similar objects make up an interface class (IC), specified using ASN.1 syntax.

For instance, a meter can offer two registers:

- A register measuring the active energy $T1 = 1234 \text{ kWh}$;
- A second register measuring the reactive energy $T2 = 0123 \text{ kWh}$.

Both registers are very similar, and can be represented by a name, a value, a unit, and methods for example, reset. They can be modeled by the same interface class, the REGISTER class.

11.3 The Object Identification System (OBIS)

OBIS defines identification codes for commonly used data items in energy metering. The DLMS-UA Blue book and IEC 62 056-61 specify the overall structure of the identification system and the mapping of all data items to their identification codes.

Value group A	Value group B	Value group C	Value group D	Value group E	Value group F
---------------	---------------	---------------	---------------	---------------	---------------

Figure 11.2 Structure of an OBIS code.

All data items exposed by a meter are uniquely identified by an OBIS code. This is true for measurement data, but the scope of OBIS is wider than measurement data: OBIS codes also cover data used for configuration of metering equipment; or related to the meter status. OBIS codes are defined for all types of utility metering applications: electricity, water, gas and heating.

The concept of an OBIS code is based on a hierarchical structure composed of 6 different “value groups”, from A to F (Figure 11.2 and Figure 11.3).

- Value group A is for the identification of the energy type. For example, electricity, hot or cold water, gas, heater, cooling are possible energy types.
- Value group B is used to distinguish between several possible inputs in a metering equipment. A data concentrator is a typical example of such equipment.
- Value group C is used for identifying the type of physical quantity, for example, voltage, volume, temperature, power. Value group C items clearly depend on the content of value group A.
- Value groups D and E are used for identifying additional data that can for example be the result of an internal processing by using a specific algorithm applied on data already identified from value groups A to C. It is also used for consortia- or country-specific applications.
- Value group F was originally planned for identifying historical data (billing periods) when needed. If there is no such data then value group E may be used for improving classification.

Obis code element	Range		Manufacturer extension codes
A	0–15	Abstract objects : 0	
		Energy type (1: electricity, 4 Heat, 5 Cooling, 6 Heat, 7 Gas, 8 Cold water, 9 Hot water)	
B	0–255	Channel	128–199
C	0–255	Type of physical quantity measured	128–199, 240
D	0–255	Processing, consortia or country specific	128–254
E	0–255	Classification	128–254
F	0–255	Historical	128–254

Figure 11.3 OBIS code value groups.

A part of the range of values in value groups B to F are reserved for manufacturer-specific data.

OBIS codes can be represented by 6 integers in dotted A.B.C.D.E.F format, for example, 1.0.1.8.0.255.

Thousands of OBIS codes have been defined, the complete list is available on the DLMS user association web site (DLMS.com). Interface classes are versioned, and can be extended over time.

In the case of very simple devices, the logical name (LN) referencing method using the OBIS system is not used. A simpler system, using a 13-bit integer for referencing any attribute of a COSEM interface object is used. This simpler system is called short name (SN) referencing. This is useful for ensuring compatibility with the older versions of DLMS.

11.4 The DLMS/COSEM Interface Classes

COSEM defines standard objects, defined by their interface classes (IC), for data storage, access control and management, time and event bound control. Interfaces classes are specified in the DLMS-UA Blue book and in the IEC 62 056-62.

Each interface object is a collection of attributes and methods:

- Attributes represent the characteristics of the object. The first attribute, mandatory, is the “logical name” and its value is an OBIS code or a short name that identifies the measurement category applying to the object instance. For instance, a “register” object with logical name [1 1 1 8 0 255] measures electric total positive active energy, while a “register” object with logical name [1 1 3 8 0 255] measures electric total positive reactive energy. Each interface class definition also allocates an index to each attribute. Each attribute is uniquely identified:
 - by the class ID and “logical name” of the object instance to which it belongs, and its index within this instance (LN referencing)
 - or
 - by a short 13 bit-integer (SN referencing), for simple devices. Some SN values are reserved for special objects, for example, 0xFA00 for the Association SN.
- **Methods:** in the object-oriented model of DLMS, external entities can act on the object only through defined methods, for example, for accessing attribute values. For instance, the “reset” method, on a register interface class, sets the current consumption value to the default value. More complex methods are defined, for instance methods that trigger authentication procedures. Within an object instance, methods are identified by their index (LN referencing) or by a short integer (SN referencing).

The set of interface classes represents a tool box that a manufacturer can use when building a meter product, and facilitating interoperability. The model can be extended, new objects only need to be added to the OBIS registry and defined using the appropriate ASN.1

description. Some OBIS code ranges are reserved, for example, for national extensions (specific attributes, interface classes) by using the E164 country code in field C of the interface OBIS code.

Manufacturers may decide not to implement standard interface classes for all objects and use the DLMS manufacturer extension mechanisms. However, when a standard interface class is used, it must be implemented in conformance with the DLMS Blue book.

11.4.1 Data-Storage ICs

- **Register (class ID 1):** this object contains a value, and an enumerated pointer to a unit.
- **Extended register (class ID 3):** this object extends the register by providing a time stamp.
- **Demand register (class ID 5):** extends the register object by storing the current value, as well as maximum and minimum values.
- **Register activation (class ID 6):** this object specifies at which periods of the day which register is activated.
- **Profile generic (class ID 7):** this is a generic “spreadsheet-like” object.
- **Utility table (class ID 26):** this IC encapsulates ANSI C12.19 table data. Each “table” is represented by an instance of this IC, identified by its logical name. The IC attributes are the ANSI Table-Id, the length of the table, and a buffer containing the table data.
- **Register table (class ID 61):** a simpler version of the profile generic object, which can be used to store multiple similar values.
- **Status mapping (class ID 63):** while status codes can hardly be standardized, this table maps custom-status codes to utility-specified values.

11.4.2 Association ICs

These objects are specified as gatekeepers to other objects:

- **Association SN (class ID 12):** list of SN references to objects of a given logical device that are accessible in a given association context with a COSEM client. This object may be present multiple times if a logical device supports multiple application associations.
- **Association LN (class ID 15):** same as above using LN referencing.
- **SAP Assignment (class ID 17):** the service access point assignment object contains the list of logical devices within a physical device and their respective service access points.
- **Image transfer (class ID 18):** this object is used to manage the upload of software images.
- **Security setup (class ID 64):** contains information on security policies within a particular application association, and methods to set up security keys.

11.4.3 Time- and Event-Bound ICs

- **Clock class (class ID 8):** the clock object, including timezone and daylight saving data.
- **Script table (class ID 9):** scripts that can be used for the activation of tariffs, upload of a new firmware, and so on . . . Scripts are a sequence of method invokes or attribute modifications.
- **Schedule object (class ID 10):** the “to do list” object, specifying time- or date-driven activities.
- **Special days table (class ID 11):** list of special days for use with the schedule object or the activity calendar.
- **Activity calendar (class ID 20):** defines a calendar-based schedule of actions.
- **Register monitor (class ID 21):** can be used to configure the monitoring of values of several registers and, if certain triggers are met, to execute action scripts.
- **Single action schedule (class ID 22):** for example, execute firmware.
- **Disconnect control (class ID 70):** manages a disconnect unit of the meter, for example, a contactor.
- **Limiters (class ID 71):** triggers an action script when the value attribute of a monitored object crosses a threshold for a certain amount of time.

11.4.4 Communication Setup Channel Objects

Multiple objects have been defined to manage the physical layer parameters and communication setup over these physical layers, for instance:

- IEC local port for IEC 62 056-21 ports;
- IEC HDLC setup;
- TCP-UDP setup;
- IPv4 setup;
- IPv6 setup;
- M-Bus slave.
- M-Bus client (meter acts as master), enable mapping of M-Bus data identifiers (data information block, variable information block) to M-Bus value objects of “extended register” interface class objects.
- M-Bus master port setup, to set EN 13 757-2 interfaces.

11.5 Accessing COSEM Interface Objects

11.5.1 The Application Association Concept

In order to allow the client party to access COSEM interface objects in the server, the DLMS-UA defined the concept of “application association”. This application association

is an application-level connection. It is established between a Client AP (application process) and a server AP (one of the logical devices that are modeled in the metering equipment). There is only one Association per logical device. The client AP always initiates the establishment of the association. For very simple devices, one-way communicating devices, and for multicasting and broadcasting pre-established associations are also allowed.

During the association establishment, some contextual data is exchanged and the authentication mechanisms are selected.

After the association establishment, the client AP and the server AP can exchange application data: some of the COSEM interface objects in the server (i.e. one of the logical devices of the metering equipment) become accessible for the client AP. Several data communications services are specified in order to exchange data. Once data exchanges are finished, the association has to be released.

The association establishment is performed by using some basic services of the COSEM application layer that is presented in the next section.

11.5.2 The DLMS/COSEM Communication Framework

The DLMS/COSEM protocol stack contains a metering application, the COSEM application layer and COSEM transport layers. The COSEM application layer is unique for any type of transport layer. Data are exchanged between a server AP and a client AP by using communication profiles (one in the server, one in the client). DLMS-UA defined several communication profiles (implemented in the COSEM transport layer) in the DLMS Green book (IPv4, HDLC, PLC, M-Bus, . . .).

Note: The DLMS version used in DLMS/COSEM is an extension of the original DLMS specified in IEC 61 334-4-41. This extended version is referred to as xDLMS. However, in the text, we continue to use DLMS.

For a better readability, only the IPv4 transport layer is represented in Figure 11.4.

The COSEM application layer provides a set of services in order to access to the application interface objects and methods. COSEM application layer services are split into 3 categories:

- application association establishment and release;
- data transfer;
- layer management (for local management, then out of scope of DLMS specifications).

Due to the existence of two different referencing methods (LN and SN) for accessing the meter objects, the COSEM application layer in the client side contains two different

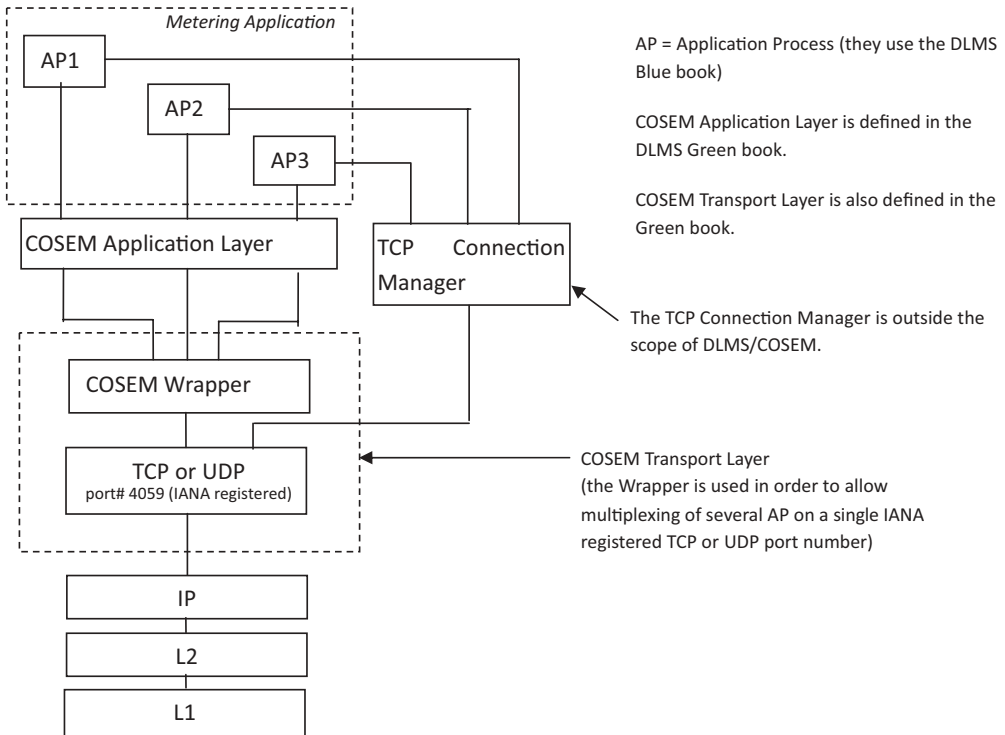


Figure 11.4 The DLMS/COSEM protocol stack.

sets of services: one for the logical referencing method, the other one for the short name referencing method.

The DLMS application protocol is connection oriented: in the previous section we explained that an application association has to be established between a server AP and a client AP before any communication with COSEM objects can occur. The set of services in charge of application association handling is composed of three services:

- COSEM-OPEN.request;
- COSEM-RELEASE.request;
- COSEM-ABORT.request.

The principle is the following:

- COSEM-OPEN.request sets up an application association. During the association establishment, a specific COSEM interface object is created: the “association” object. Among several attributes, this association object contains the list of all visible COSEM interface objects in the context of this association: after Association establishment, the client application process can read the list of visible interface objects, and perform some operations on these objects.

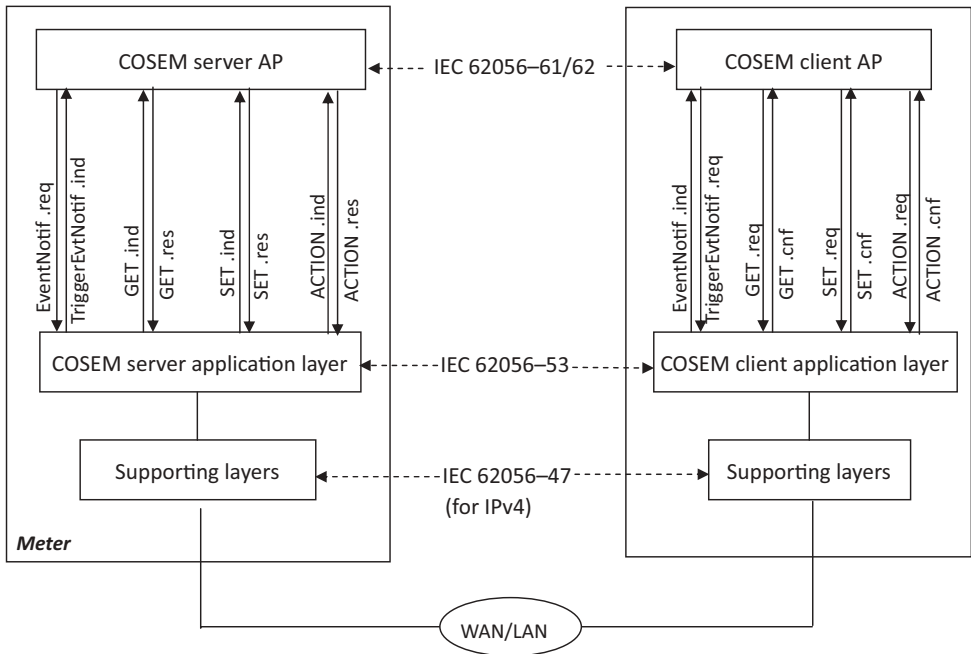


Figure 11.5 DLMS data communication services (LN referencing case).

- Application data exchange takes place using DLMS data communication services (refer to the next section for details).
- COSEM-RELEASE.request releases the application association.

for pre-established AAs, OPEN/RELEASE/ABORT requests are not used.

11.5.3 The Data Communication Services of COSEM Application Layer

The data communications services applicable to LN referencing are summarized on the Figure 11.5.

The set of services is: GET, SET, ACTION and EventNotif.

- The GET service is invoked by the client AP to request the value(s) of one or all attributes of one or more COSEM interface objects from the server AP. For example, the GET service is used for reading the value of an electricity counter. In this case, the class_id is 3 (for register class), and the attribute targeted by the GET service is the “value” attribute.
- The SET service is invoked by the client AP to request the remote server AP to set the value of one or more attributes of a COSEM interface object. For example, the SET service may be used for changing the electricity tariff for a specific period of time.

- The ACTION service is used by the client AP to remotely invoke one or more methods of one or more COSEM interface objects in the remote server AP. The server AP executes the requested ACTION. The reset of a register is a typical example of an invocable method.
- The EventNotification Service: this service is used in order to enable the server AP to send an unsolicited notification of the occurrence of an event to the remote client AP. This notification contains the value of a COSEM interface object attribute. It is an exception to the client–server paradigm. The client AP may explicitly solicit sending an EventNotification message by calling the Trigger_EventNotification_sending service primitive.

In the case of SN referencing, the list of data communication services is different.

- The read service is used to read the value of one or more attributes or to invoke one or more methods of COSEM interface objects.
- The write service is used to write the value of one or more attributes or to invoke one or more methods of COSEM interface objects.
- The UnconfirmedWrite service is used to write the value of one or more attributes or to invoke one or more methods of COSEM interface objects. It is an unconfirmed service.
- The InformationReport service: upon the occurrence of a specific event, the server can inform the client party of the value of one or more COSEM interface object attributes. It is an exception to the client–server paradigm.

The parameters for read/write must include:

- the physical layer MAC address of the meter (this is used by lower layers to establish communication with the meter or concentrator);
- an InvokeID also encoding the priority of the message;
- the interface class OBIS code, for example, REGISTER class;
- the interface class instance (e.g., multiple REGISTER classes may exist on a meter);
- the identifier of the attribute (for get and set), or the identifier of the method (for action).

Referencing may also use short-name mapping of logical name using the interface class mapping table.

The responses include:

- the destination physical layer MAC address;
- an InvokeID also encoding the priority of the message;
- the response data.

Data formats are described in the OBIS profile for the relevant object class. For instance, for the active energy register, attribute 2 is used to store the register using long unsigned encoding, while attribute 3 is an enumerated value mapped to the physical unit.

The lower layers encode the DLMS messaging primitives to PDUs, using A-XDR encoding, a specific version of ASN.1 BER optimized for COSEM data types specified in IEC 61 334-6.

11.6 End-to-End Security in the DLMS/COSEM Approach

DLMS/COSEM provides security features in two different domains:

- Access control security: controls the server data that a given client may access using role-based access rules.
- Security for data transport: provides security during the transport of data from a DLMS/COSEM end-point to another DLMS/COSEM endpoint.

11.6.1 Access Control Security

Access control security is provided as part of the application association establishment procedure.

In order to be able to access server side data, the client has to be authenticated. This is performed during the association establishment. Depending on the capabilities of the meter, the level of the security for the data access is negotiated. DLMS/COSEM provides three different levels of data access security:

- Lowest-level security: in this case there is no security at all. Peer authentication is not needed. This level allows direct access to the data contained in the server.
- Low-level security (LLS).
- High-level security (HLS).

In the LLS security model, the security is ensured via a username/password scheme. The goal is not the authentication of the server. Only the client is authenticated by providing a secret (generally a password) during the application association establishment procedure. The server checks whether the password is correct then the association is considered as established.

The association interface class provides a way to access the password in the server by using the “change_secret” method.

In the HLS security model, a mutual authentication is a prerequisite for application association establishment. Different HLS_Authentication_Mechanisms may be negotiated during the application association establishment (e.g., with different methods for generating a digest, based on MD5, SHA-1, ...).

Once the client is authenticated, the list of objects that may be accessed is determined by the server and presented in the AA object_list attribute. This doorkeeper function controls access to associations, registers, profiles, clocks, and so on, using access tables is according to the requester role determined by its identity.

11.6.2 Data-Transport Security

This part of the security scheme provides cryptographic data protection. Ciphering and deciphering is performed by the COSEM application layer on a per-message basis. In order to decide whether ciphering protection is needed, the COSEM AL uses information contained in the security context that was negotiated during the application association establishment. The security context is contained in a security setup object associated to the application association and specifies:

- The level of security to be applied to messages:
 - no security;
 - all messages have to be authenticated;
 - all messages have to be encrypted;
 - all messages have to be both authenticated and encrypted.
- The security algorithm to be used: currently the DLMS specifications contain only one security suite, the Galois/counter mode (GCM) with AES-128 symmetric encryption algorithm. Some additional security suites may be added in the future.
- The different security materials and credentials: among them, the master key, the ciphering keys, the authentication keys, the initialization vectors, . . .
- Security setup object linked to association object, specifying which security services to apply (e.g., encryption).

All meters must have a master key that is pre-established (and communicated via database transfer to the meter controller).

Part Four

The Next Generation: IP-Based Protocols

12

6LoWPAN and RPL

12.1 Overview

Traditionally, battery-powered networks or low-bitrate networks, such as most fieldbus networks or 802.15.4 (see Chapter 1 for details) were considered incapable of running IP. In the home and industrial automation networks world, the situation compares to the situation of corporate LANs in the 1980s: “should I run Token-Ring, ATM or IPX/SPX?” translates to “should I run ZigBee, LON or KNX?”

IP, with its concept of layer 3 routing and internetwork technology, has made those debates about incompatible networks obsolete: the vast majority of LANs and WANs today run IP, and many people can hardly remember which layer 2 technology their IP networks are running on. Almost any layer 2 technology can be used and will simply extend the IP internetwork.

The same transition to IP is now happening in the home and industrial automation worlds. 6LoWPAN and RPL have made this possible.

12.2 What is 6LoWPAN? 6LoWPAN and RPL Standardization

The Internet Engineering Task Force (IETF) 6LoWPAN Working Group was formed in 2004 to design an adaptation layer for IPv6 when running over 802.15.4 low-power and lossy networks (LowPAN or LLN). The work included a detailed review of requirements, which were released in 2007 (RFC 4919).

In practice, however, the 6LoWPAN is not restricted to radio links, and the technology can be extended to run over other media, for instance it has been extended to run over low-power CPL (www.watteco.com) or G3 OFDM CPL. IPv6 is also being adapted to other physical layers, independently of 6LoWPAN, for example, for Home-Plug CPL. Many fieldbus vendors are now considering an IPv6 adaptation layer for their products.

802.15.4 and most low-power transmission technologies must rely on mesh networking to create large networks. Two techniques may be used:

- “Mesh under”: the link layer (layer 2) supporting the IP network takes care of mesh networking and packet forwarding, and the IP layer sees a large subnet. An example of such a mesh under protocol is GeoNET, currently under development to support car to car transmission as part of the ETSI intelligent transport system (ITS) technical committee (<http://www.geonet-project.eu/>, <http://www.etsi.org/website/Technologies/IntelligentTransportSystems.aspx>). Mesh under is also used in the large 6LoWPAN backbone of the smart metering project of France DSO ERDF.

Obviously, mesh under works only within the context of a single link-layer technology.

- “Route over”: IP level (layer 3) mesh routing. If multiple underlying networking technologies need to be used simultaneously (e.g., wireless 802.15.4 and CPL), or when the underlying networking technology supports only point to point or local broadcast link layer communication capabilities, then IP level mesh routing becomes necessary to form the internetwork.

The IETF Routing Over Low-power and Lossy networks (ROLL) Working Group was formed in 2008 to create such an IP level routing protocol adapted to the requirements of mesh networking for the Internet of Things: the first version of RPL was finalized in April 2011 (at the time of writing the RFC was not yet allocated). See Section 12.4 for more details.

At present the reference documents for 6LoWPAN are:

- IETF RFC4919: “IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs): Overview, Assumptions, Problem Statement, and Goals”.
- IETF RFC4944: “Transmission of IPv6 Packets over IEEE 802.15.4 Networks”.
- Internet draft ID-6LoWPAN-HC defines HC (header compression), which will replace the header compression mechanism defined in RFC4944 (now “NOT RECOMMENDED”). This document is in “RFC queue” status at the time of writing.
- draft-ietf-6LoWPAN-nd-16, which updates the original IPv6 neighbor discovery mechanism for use in LowPANs.

Some existing networks use alternative pre-RPL routing methods, such as IETF draft-daniel-6LoWPAN-load-adhoc-routing-03: “6LoWPAN *ad hoc* on-demand distance vector routing (LOAD)”, used in the French AMI initiative of ERDF.

12.3 Overview of the 6LoWPAN Adaptation Layer

6LoWPAN is designed to work on top of 802.15.4 networks. The optional hop by hop acknowledgment feature of 802.15.4 is used, but the `macMaxFrameRetries` should be set

to a relatively low value (e.g., the default of 3) in order to make sure the 802.15.4 layer will not continue to retry when IP and application-level retransmission mechanisms trigger.

6LoWPAN needs to solve 4 issues:

- Header compression: on battery-powered networks, long packet headers is synonymous with energy waste. Native IPv6, with its 40-byte header, was probably one of the worst possible candidates for such networks: without compression, the payload of a single IPv6 UDP packet transmitted over a 802.15.4 link layer would not be able to exceed 53 bytes! In the most favorable case, the LowPAN and UDP compressed headers require just 6 bytes.
- Packet fragmentation and reassembly: low-power networks usually provide small MTUs, because transmission uses energy, and transmission time is proportional to the packet size. Also, small packets are less subject to packet loss that may occur over lossy networks such as 802.15.4. For instance, on 802.15.4 networks, the frame size is only 127 bytes, and the MAC level overheads (addressing fields, FCS, security headers, see Chapter 1 for more details) may leave as little as 81 bytes for IP. IPV6 normally requires a MTU of 1280 bytes!
- Adaptation of IPv6 neighbor discovery defined in RFC4861 and 4862.
- Support for “mesh under” layer 2 forwarding.

One of the issues of 802.15.4 is that it forgot to define a field to identify the “next higher protocol” (e.g., the equivalent of the Ethernet “Ethertype” field). Therefore, there is no reliable mechanism to share a given 802.15.4 PAN among multiple L3 protocols, like 6LoWPAN and ZigBee 1.0.

6LoWPAN currently defines several headers, which appear in the following order when present:

- The mesh addressing header;
- Hop by hop processing header, which encode hop-by-hop options such as BC0 broadcast sequence number;
- Destination processing: for example, the fragment header;
- Payload transport: for example, the IPv6 and UDP compression headers.

The first byte of each header, called the dispatch byte, identifies the nature of the header (Figure 12.1). A large subset of the dispatch byte space is currently reserved, leaving some room for future 6LoWPAN extensions or future coexistence with other protocols that would use the same dispatch byte.

12.3.1 Mesh Addressing Header

Currently, no “mesh-under” protocol is defined for 802.15.4, so this header is only a facility provided to make it possible in the future. When 802.15.4 mesh-under routing is

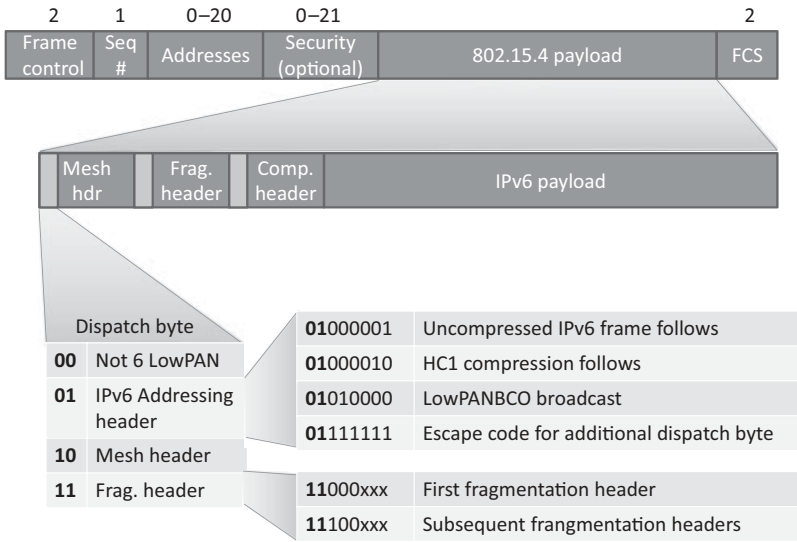


Figure 12.1 6LoWPAN header stacking, the dispatch byte.

enabled, the 802.15.4 MAC frame contains the source and destination addresses for each hop, therefore a container is needed for the original and final 802.15.4 addresses. The mesh addressing header provides such a container, and also contains a “HopLeft” counter that should be decremented by each layer2 hop.

12.3.2 Fragment Header

The fragment header for the first fragment specifies the full (reassembled) packet size, and uses a datagram tag common for all fragments of this IP packet, which will be used by the receiver, together with the sender and destination MAC addresses, to identify fragments belonging to the same packets. Subsequent fragments also specify the offset of the fragment in the full IP packet, in multiples of 8 bytes (see Figure 12.2).

12.3.3 IPv6 Compression Header

12.3.3.1 Forming an IPv6 Unicast Address from the 802.15.4 EUI64 or 16-bit Short Address

The 802.15.4 EUI64 is composed of a 24-bit OUI (organizationally unique identifier) and a 40-bit extension identifier chosen by the manufacturer). The OUI has two reserved bits in its first octet: the least significant bit is reserved to define a space for multicast addresses, and the second least significant bit (L) is used to distinguish locally assigned addresses from universal addresses formed as OUI+extension.

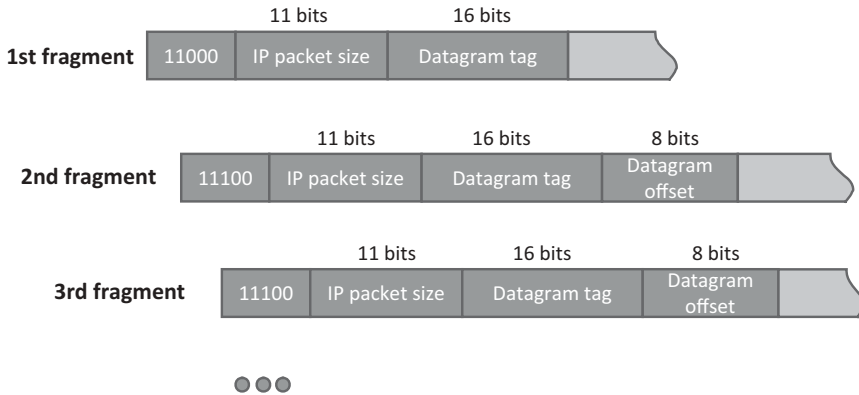


Figure 12.2 6LoWPAN fragment header.

In order to form the interface Id (IID), RFC4291 chooses to invert the L bit of the EUI64 and calls it the U bit (universal address bit), so that locally assigned addresses will have all-zero prefixes and be more compact and easier to remember.

6LoWPAN mandates that IPv6 local addresses will be derived from the EUI64 addresses using the following convention: 64 bit prefix + Ubit formatted EUI64. Prefix FE80:: (111111010 followed by 54 zeroes) is used for link local addresses, prefixes beginning with 001 are global prefixes for unicast addresses.

The IPv6 address can also be derived from the 16-bit 802.15.4 short address, by concatenating it with the PAN identifier, or with 16 zeroes if the PAN ID is unknown (RFC 4944 Section 6). 6LoWPAN requires bit 6 (Ubit) of the PAN identifier to be zero as it is not a universal address. Short addresses beginning with a 0 bit are reserved for unicast addresses, and short addresses starting with 100 are reserved for multicast addresses. Other values are reserved.

12.3.3.2 HC1-HC2 Compression (Now “Not Recommended” and Replaced by HC)

The original 6LoWPAN standard (RFC4944) defined a simple stateless compression mechanism compatible with the capabilities of resource constrained nodes, which exploits the redundancies between the MAC layer and the IPv6 layer, and encodes most likely values of variable fields in a more compact format. The “v6” version field is elided.

Figure 12.3 shows the conventions used by HC1 and HC2. The “C” flag indicates that the flow label and traffic class are all zeroes and elided.

When the payload is UDP (NH=01), then HC2 compression can be used to reduce the size of UDP ports (only 4 bits are sent inline if the ports are in the 61 616–61 631 range), and to omit the redundant UDP size field.

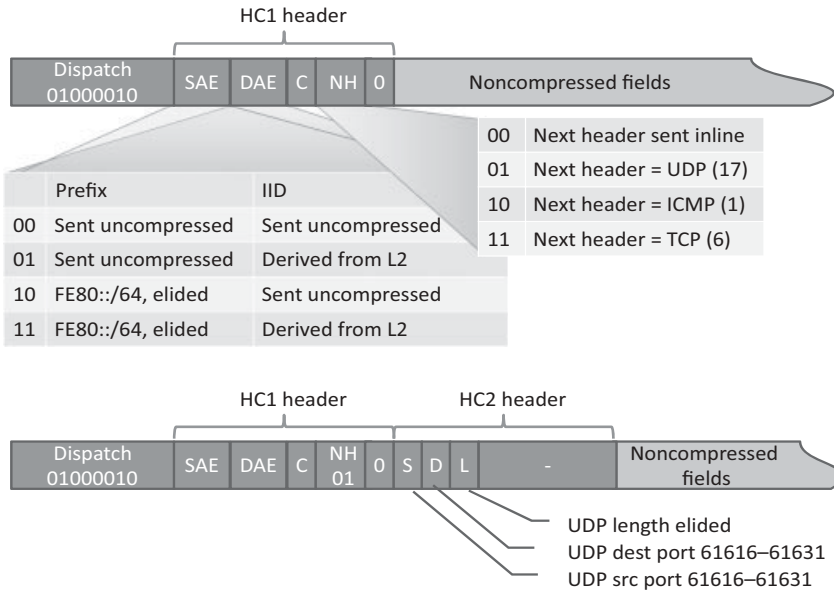


Figure 12.3 HC1, HC2 compression.

The noncompressed IP fields are sent starting with the hop counter, then in the order of the HC1 header elements. Uncompressed UDP fields are sent in the order of standard UDP fields.

12.4 Context-Based Compression: IPHC

HC1 compression works well when using link-local addresses, but any communication with IPv6 nodes located outside the local network will require globally routable IPv6 addresses, which are not compressed with HC1. Internet draft ID-6lowpan-hc defines a new header for IPHC compression with simple support for shared context information between sender and receiver.

IPHC “steals” 5 bits out of the reserved dispatch value field for its own 13-bit base header, outlined in Figure 12.4.

The new IPHC header format adds an ability to selectively compress IPv6 flow labels (RFC2460 and RFC3697), 6-bit differentiated services code points (DSCP, RFC2474 and RFC3260) and explicit congestion notification (ECN, RFC 3168). The address compression takes into account prefixes indexed by the optional context IDs (SCI and DCI).

If the N bit is set, the IPHC header and uncompressed IPv6 fields are followed by a LOWPAN_NHC header, otherwise the IPv6 next headers are transmitted inline.

The format of the LOWPAN_NHC header is illustrated in Figure 12.5.

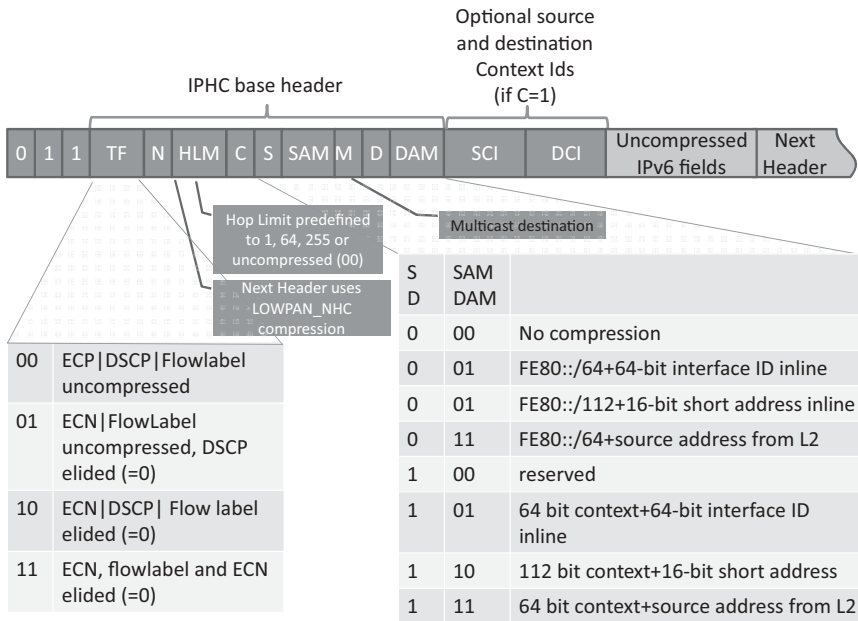


Figure 12.4 The IPHC 6LoWPAN header structure.

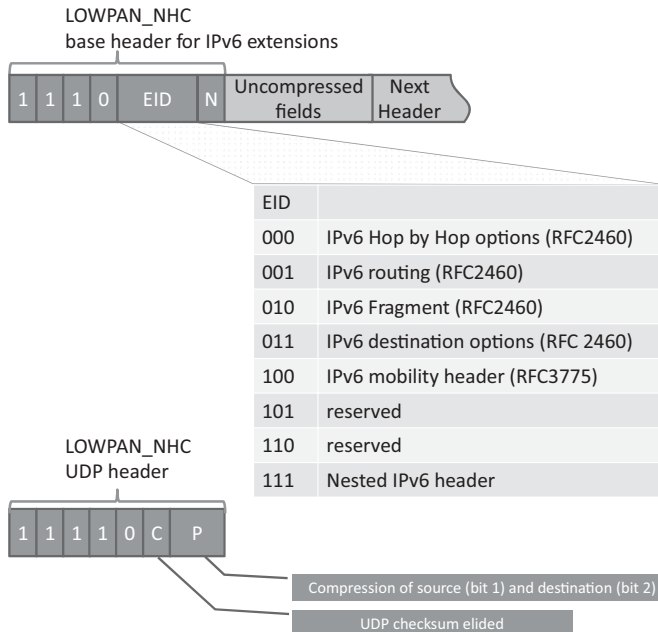


Figure 12.5 LOWPAN_NHC format for UDP and IPv6 options.

The LOWPAN_UDP header optionally compresses the UDP ports using the same approach as HC2 (compressed port format F0xx sending just 4 bits inline).

12.5 RPL

The IETF Routing Over Low-power and Lossy networks (ROLL) Working Group was formed in 2008 to create an IP level routing protocol adapted to the requirements of mesh networking for the Internet of Things: the first version of RPL (Routing Protocol for Low-power and lossy networks) was finalized in April 2011 (at the time of writing the RFC was not yet allocated).

The reference documents for ROLL are (at the time of writing, April 2011):

- <https://datatracker.ietf.org/doc/draft-ietf-roll-rpl/>, which defines RPL, the IPv6 routing protocol for low power and lossy networks;
- RFC 6206 that defines the RPL objective function 0;
- draft-ietf-roll-terminology, which defines the terminology used by ROLL;
- draft-ietf-roll-security-framework, which defines a security framework for ROLL;
- draft-ietf-roll-trickle, “the trickle algorithm” defines a dynamically adjustable transmission window scheme to optimize RPL traffic;
- draft-ietf-roll-routing-metrics for the computation of metrics;
- draft-ietf-roll-minrank-hysteresis-of that defines a hysteresis-based mechanism to prevent topology oscillations;
- draft-ietf-roll-p2p-rpl that defines an optimization mechanism for point to point communication (e.g., in automation scenarios for sensor/actuator messages flows).

RPL specifies a routing protocol specially adapted for the needs of IPv6 communication over “low-power and lossy networks” or LLNs, supporting peer to peer traffic (point to point), communication from a central server to multiple nodes on the LLN(point to multipoint P2MP) and *vice versa* (multipoint to point MP2P). The base RPL specification is optimized only for MP2P traffic (upward routing or convergecast used, e.g., in metering networks) or P2MP, and P2P is optimized only through use of additional mechanisms such as draft-ietf-roll-p2p-rpl.

Such LLNs are a constrained environment, which imply specific requirements explored by the IETF ROLL working group in RFC5867, RFC5826, RFC5673, and RFC5548. RPL has been designed according to these LLN specific requirements (typically on networks supporting 6LoWPAN), but is not limited to operation over LLNs.

Multiple concurrent instances of RPL may operate in a given network, each RPL instance is characterized by a unique RPLinstanceID. The following sections describe the behavior of an individual RPL instance.

The RPL routing protocol builds one or more destination oriented direct acyclic graphs (DODAG). Each DODAG is a directed graph with no cycles and with a single root node (see Figure 12.6). The graph is built according to optimization objectives specified by an

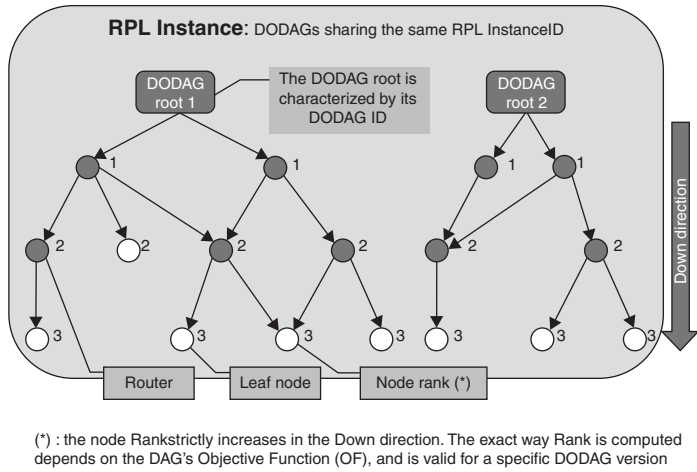


Figure 12.6 RPL builds a destination-oriented direct acyclic graph (DODAG).

objective function (OF, defined by the OCP field of a DIO DODAG configuration option). The objective function is not specified by RPL itself, but in other companion documents according to domain-specific requirements: for the available network metrics, the OF computes the “rank” measuring the “distance” between the node and the DODAG root and also defines the parent node selection policy, for instance an objective function could seek to minimize the expected packet delay, while another might want to avoid routing through any battery-operated node (see [I-D.ietf-roll-routing-metrics]).

RPL requires bidirectional links. Bidirectional connectivity must be verified before accepting a router as a parent, for example, by using IPv6 neighbor unreachability detection

ICMPv6 Type=155	Code	Checksum
Security (secure RPL msgs only)	0x00	DODAG Information Solicitation
	0x01	DODAG Information Object
	0x02	Destination Advertisement Object
Base	0x03	Destination Advertisement Object Ack
	0x80	Secure DODAG Information Solicitation
	0x81	Secure DODAG Information Object
options	0x82	Secure Destination Advertisement Object
	0x83	Secure Destination Advertisement Object Ack
	0x8A	ConsistencyCheck

Figure 12.7 Structure of ICMPv6 RPL control message.

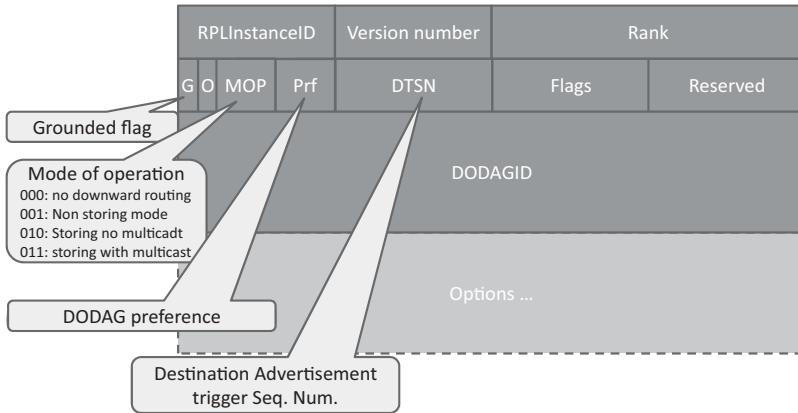


Figure 12.8 RPL DIO base object (followed by options).

(NUD), bidirectional forwarding detection (RFC5881) and hints from lower layers via layer 2 triggers like RFC5184.

12.5.1 RPL Control Messages

RPL routers need to exchange information in order to build the DODAG and populate routing tables. RPL defines a new ICMPv6 (RFC 4443) message, type 155, for this purpose.

RPL defines the following base objects:

- The DODAG information solicitation (DIS) message;
- The DODAG information object (DIO), see Figure 12.8;
- The destination advertisement object (DAO);
- The DAO Ack object;
- The consistency check (CC) object, which is used to check secure message counters and to carry RPL challenges and responses, and is always carried in a secure RPL message.

12.5.2 Construction of the DODAG and Upward Routes

The DODAG information object (DIO) is used to build the DODAG: it carries general DODAG configuration parameters and information that allows listening RPL routers to select a set of DODAG parents. Several type-length-value encoded options in the same RPL control message may specify:

- The address of the sending RPL router, and prefixes that may be used for IPv6 stateless autoconfiguration (0x08 prefix information option, or PIO). The PIO contains the same

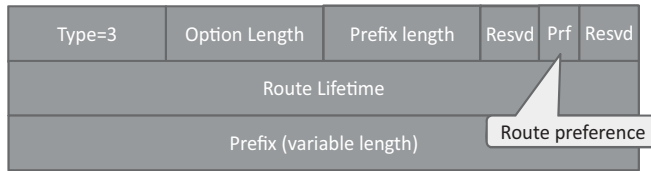


Figure 12.9 RPL Route Information option.

fields as the IPv6 neighbor discovery prefix information option defined in RFC4861, RFC4862 and RFC3775. A 1-bit “L flag” indicates that addresses derived from the prefix can be considered “on-link”, a 1-bit “A flag” indicates that the prefix can be used for stateless address autoconfiguration.

- Metrics allowing estimation of the cost to reach destinations starting with each prefix (0×02 metric container option, formatted as specified in ID.IETF-roll-routing-metrics),
- One or more prefixes that are reachable by the advertising node (0×03 routing information option, illustrated in Figure 12.9 and containing the same fields as the IPv6 neighbor discovery route information option defined in RFC4191).
- Additional DODAG configuration information (0×04 DODAG information option) such as the values of MaxRankIncrease and MinHopRankIncrease used to constrain the rank a node can advertise when reattaching to a DODAG, or the default lifetime of all RPL routes.

RPL nodes send DIOs periodically via link-local multicasts, and joining nodes may request DIOs from their neighbors by multicasting ICMPv6 control messages containing a DODAG information solicitation Object (DIS). DIO parameters are explained in Figure 12.8, the DTSN is an 8-bit unsigned integer number set by the issuer of the message. In the storing mode of operation, incrementing the DTSN is a way to request updated DAO messages from child nodes.

Each DODAG, identified by a unique RPLInstanceID and DODAGID, is built incrementally from the root to leaf nodes:

- RPL nodes, starting by the DODAG root, advertise their presence, affiliation with a DODAG, routing cost, and related metrics by sending link-local multicast DIO messages to the all-RPL-nodes address. The DODAG root advertises predefined rank ROOT_RANK (=MinHopRankIncrease), and also specifies if it is “grounded”, that is, if it can reach the set of destinations specified by the local DODAG policy (the “goal”). A DODAG is said to be floating if it cannot satisfy the goal.
- Nodes use the received DIO information to join a new DODAG and select their parents in the DODAG, or to maintain their affiliation to an existing DODAG. Nodes select parents according to the policy specified by the objective function and the rank of their neighbors as advertised by DIO messages. For the determination of parent

relationships, the ranks of potential parent nodes are compared with a granularity of `MinHopRankIncrease` (specified in the DIO messages), so that `parent1` and `parent2` will be considered of equal rank if $\text{floor}(\text{rank}(\text{parent1}) / \text{MinHopRankIncrease}) = \text{floor}(\text{rank}(\text{parent2}) / \text{MinHopRankIncrease})$.

A first set of nodes will attach to the DODAG root and start to advertise DIO messages with the corresponding RPL instance and DODAG ID, expanding the reach of the DODAG. As new nodes will start to hear the RPL instance DIO messages and attach to it, the DODAG reach expands further until it reaches all nodes willing to attach to this RPL instance.

Nodes provision upward routing table entries according to their local policies (e.g., least cost), for the destinations specified by the DIO message, setting one or more DODAG parents as the next hop.

Each DIO announcement is attached to a specific DODAG version, therefore if the root decides to change the DODAG version it triggers a complete recalculation of the DODAG topology: this is a global DODAG repair.

A node may poison previously announced routes by advertising a special rank value of `INFINITE_RANK (=0xFFFF)`. Note that if the destination cannot be reached temporarily, the node should rely on the local repair procedure and not poison the routes. The node may also decide to create a floating DAG. Poisoning a route implies that all sub-DAGs will also have infinite rank and therefore breaks the DAG topology.

12.6 Downward Routes, Multicast Membership

RPL uses destination advertisement object (DAO, Figure 12.10) messages to establish downward routes and to indicate multicast group membership. DAO messages are not mandatory and are required only by RPL instances that provide support for point to multipoint or peer to peer traffic. RPL control messages carrying a DAO object may also transport a list prefixes announced as reachable RPL targets or multicast groups (`0x05` RPL target option), opaque RPL target descriptors (`0x09` RPL target descriptor option),

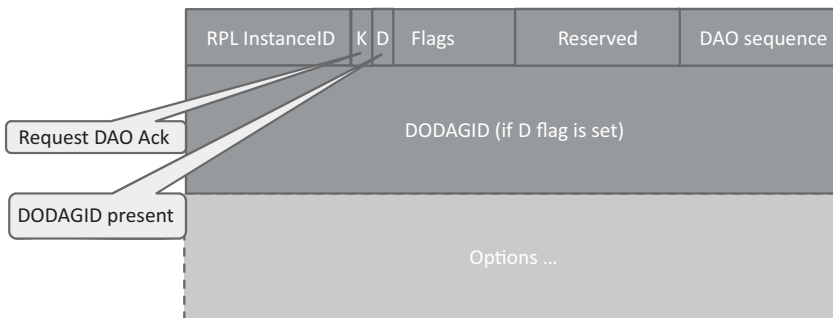


Figure 12.10 DAO object format.

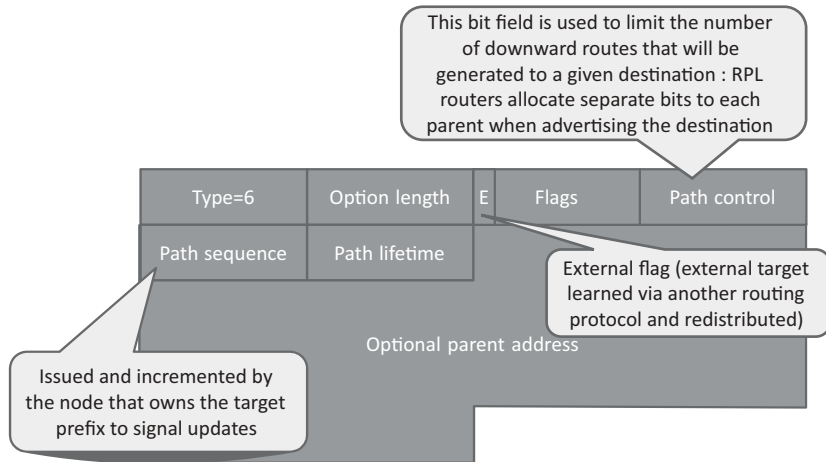


Figure 12.11 RPL transit information option.

and transit information (0x06 transit information option, Figure 12.11) which is used to indicate attributes for a path to one or more destinations, for instance its lifetime (a lifetime of 0x00000000 indicates loss of reachability to a target).

RPL supports two models for downward routing, each RPL instance supporting downward traffic selects one of the two models:

- In the storing model, RPL routing nodes are stateful. DAO messages, including the prefixes and addresses reachable by the sending node, are sent to the parents. Parents store the preferred downward routes and propagate aggregated DAOs upward.
- In the nonstoring model, all downward traffic includes a source routing header specifying each hop along the path, and intermediary routers do not store any routing information. Nodes send unicast DAO messages to the DODAG root, which include prefixes and addresses directly reachable by the node, and the node parents (in a transit information option as illustrated in Figure 12.11). The set of DAO messages enables the DODAG root to calculate an optimal hop by hop source routing path for each advertised destination. This mode has important implications: messages will be much longer (include source routing information), and P2P traffic is always routed to the DAG root.

12.7 Packet Routing

IP packets injected in an RPL network must have a RPL header that specifies the RPL instance, except when strict source routing is used.

If some leaf nodes send IP packets without such an RPL header, the first RPL router is required to add it, and select a default RPL instance. The RPL specification in itself does

not specify the header format, but points to the ID-ietf-6man-rpl-option that places the RPL information into an IPv6 hop-by-hop option header.

The RPL information includes:

- “O”: the down 1-bit flag indicating the intended direction of the packet.
- “R”: the Rank-Error 1-bit flag signaling that a mismatch has occurred during forwarding between the rank relationship of the sender and receiver, and the effective direction of the packet. Such inconsistency, which can happen during the construction of a new DODAG version of a given RPL instance, is allowed to happen only once. RPL routers will absorb packets that are already “R” flagged in case such rank inconsistency is detected again.
- “F”: the forwarding error 1-bit flag which indicates the node cannot forward the packet further towards the destination. In storing mode, routers that receive packets that they cannot route to their intended destination from a parent will loop back the packet to this parent with the F flag set: this allows the parent to remove the erroneous DAO routing entry from its routing table (“DAO inconsistency detection and recovery”).
- The 8-bit RPLInstanceID.
- The 16-bit SenderRank, which must be set to zero by the packet source and then is set to the rank value of the forwarding RPL router.

P2P packets travel up toward a DODAG root, then are routed down to the final destination by the first RPL router capable of reaching the destination. In the case of “nonstoring” RPL instances packets will travel all the way to a DODAG root, which will add the source routing header (RH4 header as specified in ietf-6man-rpl-routing-header) and reinject the packet in the down direction.

12.7.1 RPL Security

RPL defines three security models:

- The “unsecured” model does not implement specific security features at RPL level, however the layer 2 network may implement some level of security (e.g., a 802.15.4 network key).
- The “preinstalled” model requires all RPL nodes to be configured with preprovisioned keys, which they use to code and decode secure RPL messages. Secure RPL messages have the high order bit of the code field set (see Figure 12.7).
- The “authenticated” mode that also uses a preinstalled key, but only to join the network as a leaf node. The node will need to obtain a key or a certificate from an authentication authority to join an authenticated RPLInstance as a router. This last mechanism is not fully defined yet and will require future companion specifications.

13

ZigBee Smart Energy 2.0

13.1 REST Overview

“Representational state transfer”, or REST is a distributed software architecture style that was described by Roy Fielding in a thesis presented in 2000. The thesis discusses client-server based architectures, analyses the reasons for the success of HTTP and hypertext, and presents a number of constraints that define a RESTful architecture, that is, an architecture that will use the same design principles, and share the same desirable properties as HTTP (scalability, simplicity, reliability . . .).

13.1.1 Uniform Interfaces, REST Resources and Resource Identifiers

The first design constraint is that interfaces should be “uniform”, based on the concept of exchanging *resources*. Roy Fielding introduces the concept of a resource, an abstraction for server-side information (and associated native data representation): “*Any information that can be named can be a resource*”. Resources are associated to resource identifiers. A REST interface will transmit only a *representation* of such resource, which is a specific way of presenting the resource to a client that can evolve over time, or depend on the client type, while the native data representation evolves independently. Technically, a representation is “*a sequence of bytes*”, plus *representation metadata* (to describe the structure and semantics of this byte sequence), and optional *resource metadata* (to represent information about the resource independent of its representation).

The REST messages exchanged between a client and a server include the identifier of the resource, an optional resource representation data, and optional control data that may indicate the purpose of a message (e.g., action being requested) or be used to parameterize the requests (e.g., select a specific representation of a resource).

It is perhaps easier to understand the specific approach of REST by comparing with other traditional programming architecture styles. Most IT systems interfaces are based

on two concepts: a verb describing the action to be taken, and some form of data. This is typical of object-oriented systems, where all data is encapsulated in the server, and manipulated only through methods (verbs) and associated method parameters. Usually, the goal of such interfaces is to isolate clients from the internal data model of the server. With such verb/parameter interface models a developer needs to be familiar with two dimensions of the interface definition: the various verbs available for each interface, and the associated parameters. Reading the interface documentation is a prerequisite to coding.

In contrast, as REST interfaces are representation centric, a small set of verbs, uniform across all use cases, can be used. Usually, this set of verbs is referred to as CRUD for create, read, update and delete. Developers need to focus only on the resource representation format. This approach is very developer friendly: in many cases, knowing the resource identifier is enough to start coding. Reading the resource will provide a representation, and often representations are sufficiently self-descriptive for a developer to have a good intuition of what to do next to manipulate the resource. Many web portals provide a REST interface with very little documentation for the general public, giving access to a comprehensive feature set, and still integrating functionality from those portals “feels” easy.

An additional advantage of verb standardization is that the REST transport protocol can be specified independently of any use case, which makes it possible to define standard transport protocols (HTTP, CoAP), as well as generic application-level components (Roy calls them “*connectors*”) such as proxies, load balancers, firewalls, and so on.

Of course, while this “visibility” is one of the design goals of REST, the REST constraints still provide some flexibility that can be used to defeat the original intention: for instance, it is possible to map an object-oriented interface by implementing one resource per verb, or by using one control message per verb.

At present, while REST is probably more human friendly, it is not quite as machine friendly as other interface models, such as SOAP. So far, it fails to provide a comprehensive interface description language. The Web application description language (WADL) serves that purpose, but only for a subset of potential REST interfaces. In practice, most recent RESTful architecture standards, such as oBIX, ZigBee SEP 2.0 or ETSI M2M use text specification for the definition of REST interfaces.

13.1.2 REST Verbs

The paper of Roy Fielding never stated which verbs a REST architecture had to provide, as the central idea was that those verbs should aim at manipulating resources. However, the design principles of HTTP, and its evolution to HTTP 1.1, are discussed at length, and in practice many recent standards start by specifying their HTTP binding, before considering other potential bindings for example, to CoAP or other REST-capable protocols.

The exact use of HTTP verbs in a REST context is sometimes ambiguous. Recent standards (ZigBee SEP 2.0, ETSI M2M) have converged on the following usage guidelines:

- GET: request verb for reading a representation of a resource in a *safe* fashion, that is, the request does not change the state of the resource on the server. If the resource URI represents a collection, a list of URIs of collection members will be returned. Generally, resources should be exposed according to the principle of “*gradual reveal*”, that is, structured in a way that complex structure subelements will be represented by reference in the representation of the parent resource. Another way to see this is that representations should include references to related representations, or “*hypermedia as the engine of application state*”.
- PUT: request verb for creating or replacing a resource, in an *idempotent* fashion, that is, multiple identical requests should have the same effect as a single request. If the URI represents a collection, the entire collection is replaced.
- POST: request verb for appending to a resource or creating a subordinate resource (not safe nor idempotent). For instance, a POST /item would typically result in the creation of /item/<subresource instance number assigned by the server> for example, /item/1. The URI of the created resource is returned in the location header as part of the 201 “created” response.
- DELETE: interface for deleting a resource (*idempotent*). If the URI represents a collection, the entire collection is replaced.
- HEAD (optional): interface to request metadata regarding a resource, in a safe fashion.
- OPTIONS: interface to request the methods available at the server for a resource, for the authorization level of the client.

13.1.3 Other REST Constraints, and What is REST After All?

The RESTful architectural style is defined by additional constraints:

- Communications should be stateless, “each request from client to server must contain all of the information necessary to understand the request, and cannot take advantage of any stored context on the server. Session state is therefore kept entirely on the client. This constraint induces the properties of visibility, reliability, and scalability. Visibility is improved because a monitoring system does not have to look beyond a single request datum in order to determine the full nature of the request. Reliability is improved because it eases the task of recovering from partial failures. Scalability is improved because not having to store state between requests allows the server component to quickly free resources [...]”.
- Server responses should classify responses as cacheable or not. The use of caching improves scalability and the user experience (reduced latency).
- The interfaces should facilitate layered architectures (e.g., use intermediate load-balancing servers).

However, the core constraints are those outlined in Sections 13.1.1 and 13.1.2.

There were several recurrent discussions on the Internet and within standard bodies debating whether this or that architecture was RESTful or not. Here are some common topics:

- Are “servers” and “clients” defined at function level or interface by interface? Both ZigBee/Homeplug and ETSI concluded that many “real-world” applications cannot strictly separate functions as server or client. Many functions are both server and client for different interfaces. Therefore, the REST concepts are usually understood and applied on a per interface basis (client and server interfaces).
- What about subscribe/notify? This key functionality appears to be missing in the original REST paper, and is also missing in HTTP, leading to workarounds such as polling for AJAX interfaces. Recent standards reintroduce a subscribe/notify model by defining a dedicated resource to store subscriptions to resource R, and define a resource that need to be implemented by hosts interested in notification related to R, where notifications will be posted. This is a typical example of a case where “clients” of a resource R hosted on a server, will need to implement a server function, while the “server” will act as a client to post notifications.
- Concurrent access control to a resource. When the HTTP binding is used, the etag value is used to prevent simultaneous resource updates race conditions. All resource representations for a given resource are required to have the same etag, and the etag should be changed each time the resource is updated. Clients that want to perform a resource modification based on the assumption that the resource has not been updated since the last time they read it should include a condition based on the etag value (If-Match HTTP header).

With these clarifications, it seems that there is now a fairly good agreement among standard bodies on the practical implementation of REST style architectures. Among the work topics for further alignment:

- URI naming conventions, for example, how to represent collections;
- agreement on partial resource access/update methods, for instance using XCAP (RFC 4825) or xPath (defined by the W3C), at least for XML-encoded resources.

13.2 ZigBee SEP 2.0 Overview

The ZigBee Smart Energy Profile 2.0 protocol is the result of a joint work of the ZigBee alliance and of the HomePlug powerline Alliance, who is behind the HomePlug/AV (draft IEEE P1901) CPL standard and working on the “Green Phy” CPL standard (see Section 2.1 for more information on CPL technology).

The idea was to redesign an equivalent of ZigBee SE 1.0, but in a physical-layer-independent way, based on an IP networking layer and using a RESTful design. The

clusters of ZigBee SE 1.0 are redesigned as “function sets” in SEP 2.0. In addition, the working group took into account a comprehensive requirements list coming from utility companies as well as the Society of Automotive Engineers (SAE) for aspects related to electric vehicles.

This chapter summarizes the 0.7 draft version of Smart Energy 2.0 which was published in April 2010 and is available on the web site of the ZigBee alliance: <http://www.zigbee.org/Standards/ZigBeeSmartEnergy/Version20Documents.aspx>.

SEP 2.0 assumes an all IPv6 network. The protocol is designed according to the REST paradigm, and the data model is intended to map directly to IEC 61968 (the “common information model”). Resources are modeled using XML, and resource representations are compressed using EXI.

As this book was going to press the draft specification was updated. The overall functionality and design remains similar, but unfortunately the data model included too many changes for us to be able to fully update this chapter. We have inserted notes in the text to outline the main changes.

13.2.1 ZigBee IP

Since SEP 2.0 relies on an IP stack, and the typical implementation targets low-powered radio device, the first step was to define an IP transport layer over for 802.15.4 networks. ZigBee decided to adopt the work of IETF 6LoWPAN and ROLL working groups and mainly focused on selecting among the various options proposed:

- It uses **IEEE 802.15.4-2006 physical and mac layers**. The non-IP version of ZigBee uses the 2003 version of 802.15.4: this update means that ZigBee IP, unlike its predecessor, will be able to run on all the frequencies supported by 802.15.4, including 900 MHz and 868 MHz, since the 2006 version introduced higher-bitrate options for these frequencies (250 kbps, up from 20 and 40 kbps, respectively, see Chapter 1).
- **6LoWPAN is used with the hc adaptation layer** (RFC 6282 “Compression Format for IPv6 Datagrams over IEEE 802.15.4-Based Networks”), and uses 6LoWPAN neighbor discovery (RFC 4861, updated by <https://datatracker.ietf.org/doc/draft-ietf-6lowpan-nd/>)
- The **IETF ROLL RPL routing protocol** (see Chapter 12 for more details) **is used in nonstoring mode** : source routing must be used by the 6LoWPAN DODAG root, using the new **RH4 routing header** (draft-ietf-6man-rpl-routing-header, at the time of writing).
- As mandated by RPL, IPv6 packets are injected into the RPL router network with the new **RPL option for the hop-by-hop** header (see draft-ietf-6man-rpl-option-01).

- Standard TCP and UDP transport layers are used. In practice, for constrained devices UDP is more likely to be used, in conjunction with CoAP.
- Regarding security, ZigBee/IP uses **PANA** (Protocol for Carrying Authentication for Network Access, RFC 5191 and 5193) on top of UDP to transport the EAP authentication during the client–network authentication phase for the **EAP** (RFC 3748)/**EAP-TLS** (RFC 5216)/**TLS** (RFC 5246) security stack that use the ECC and RSA public key mechanisms and the PSK cipher suite. See Section 12.4 on security for more details.

13.2.2 ZigBee SEP 2.0 Resources

13.2.2.1 ZigBee SEP 2.0 and REST

According to the REST model, communicating entities are classified as clients or servers. SEP2.0 recognizes that in practice, most nodes are clients and servers, and that notion has a meaning only in relation to a given transaction, where the server hosts the representation of the resource being read, updated or deleted.

SEP 2.0 interacts with resources by means of the standard REST verbs for the HTTP binding: GET, PUT, POST, DELETE. A subscription mechanism has been added that can be used by nonsleepy devices (see Section 12.2.2.2).

When HTTP is used, the transport protocol is TCP on port 80 or TLS on port 443. CoAP on constrained devices will use UDP instead, however the use of CoAP was still debated by the ZigBee alliance at the time of writing. The media type used by SE 2.0 is `application/exi` (see Figure 13.5), which refers to the efficient XML interchange (EXI) encoding format defined by the W3C. EXI uses prior knowledge of the XML grammar of the documents exchanged between a client and a server to encode the document in a compressed format. EXI has reached “proposed recommendation” status on January 2011.

13.2.2.2 SE 2.0 Lightweight Subscription/Notification Mechanism

A device (A) that supports the subscription mechanism must expose the `/sub` collection resource. It will send notifications to all devices that have subscribed for event notifications by inserting a subscription subresource in the `/subcollection`.

A device (B) that supports receiving event notifications must expose the `/ntfy` resource. In order to subscribe to notifications regarding `//{host A}/resource`, device B must send a POST to `//{host A}/sub/{IPv6 address of B}` containing the URI of the monitored resource (`//{host A}/resource`) and the URI of its own notification resource (`//{host B}/ntfy`), as in the example of Figure 13.1.

It is possible to list all active subscribers of (A) by reading `/sub`. Figure 13.2 shows an example of the `/sub` resource representation (as for other examples of this chapter, the resource representation is shown decoded, it would actually be EXI encoded).

```

PUT /sub/{IPv6 Address of A} HTTP/1.1
Host: {IPv6 Address of B}
Content-Type: application/exi

<?xml version='1.0' encoding='UTF-8'?>
<SubscriptionList xmlns='http{s}://www.zigbee.org/doc/se-2-0-0'>
  <Subscription>
    <Resource>http{s}://{IPv6 Address of A}/resource
    </Resource>
    <NotificationURI>http{s}://{IPv6 Address of B}/ntfy
    </NotificationURI>
  </Subscription>
  <Subscription>
    <Resource>http{s}://{IPv6 Address of A}/resource2
    </Resource>
    <NotificationURI>http{s}://{IPv6 Address of B}/ntfy
    </NotificationURI>
  </Subscription>
</SubscriptionList>

```

Figure 13.1 SE 2.0 notification subscription example.

If (A) detects a change in the monitored resource, it will send a POST to `://{host B}/ntfy` containing the URI of the modified resource, and the URI of (B)'s subscription (to be used as a handle to the trigger subscription, or simply a reminder of the URI that can be used to change the subscription). Figure 13.3 shows an example notification.

Note: As this book was going to press, an update of the draft specification introduced a EndDevice list resource where a server stores data related to each client. A subresource (e.g. `edev/1`) is created for each client. The client subscriptions are now stored in `edev/{#}/sub`.

```

<?xml version='1.0' encoding='UTF-8'?>
<SubscriberList xmlns='http{s}://www.zigbee.org/doc/se-2-0-0'>
  <Subscriber href="http{s}://{IPv6 Address}/sub/{IPv6
    Address 1}" name="{IPv6 Address 1}" />
  <Subscriber href="http{s}://{IPv6 Address}/sub/{IPv6
    Address 2}" name="{IPv6 Address 2}" />
</SubscriberList>

```

Figure 13.2 SE 2.0, example list of subscribers (/sub).

```

POST /ntfy HTTP/1.1
Host: {IPv6 Address of A}
Content-Type: application/exi

<?xml version='1.0' encoding='UTF-8'?>
<Notification xmlns='http{s}://www.zigbee.org/doc/se-2-0-0'>
  <Resource>http{s}://{IPv6 Address of A}/resource4</Resource>
  <SubscriptionURI<http{s}://{IPv6 Address of B}/SUBSCRIBE/
{IPv6 Address of A}
  </SubscriptionURI>
</Notification>

```

Figure 13.3 Example SE 2.0 notification posted to /ntfy.

13.2.2.3 SE 2.0 Collection and Event Resources

Figure 13.4 lists the conventions used by SE 2.0 to represent collection resources and resource instances that are part of a collection. Events are stored as a collection, with a specific alias to the active event.

URI http(s): //{address}/...	Description
/path to a collection of events}/act	Alias of the currently active event resource
/mrr	List of links to mirrored feature sets, e.g. /mrr/0
/egg	Collection of 'egg' resources, GET /egg typically returns : <pre> <EggList xmlns='http{s}://www.zigbee.org/doc/ se-2-0-0'> <Egg href="http{s}://{IPv6 Address}/egg/0" name="0" /> <Egg href="http{s}://{IPv6 Address}/egg/1" name="1" /> </EggList> </pre> POST /egg creates a new egg instance e.g. /egg/1. PUT is not allowed on collections. DELETE /egg deletes all sub-resources of /egg.
/egg/1	Egg resource instance # 1, as part of a collection. PUT /egg/1 replaces this instance or creates it if it did not exist.

Figure 13.4 SE 2.0 collection and event resources.

13.2.2.4 Resource Discovery, /rsc

For the resolution of host names, device and resource discovery, SE 2.0 uses mDNS (IETF draft-cheshire-dnsext-multicastdns) and DNS-SD (draft-cheshire-dnsext-dns-sd). DNS service discovery uses DNS PTR records mapping a <Service>.<Domain> (e.g. Test\032Server._smartenergy._tcp.local.) name to a list of <Instance>.<Service>.<Domain> names. Each <Instance>.<Service>.<Domain> can be located by using DNS SRV records, and additional service information is stored in TXT records in key/value pair format. The URI of an Smart Energy 2.0 DeviceCapabilities resource is stored in the path key (e.g. path=/dcap).

The list of resources on a given host (identified by its IP address) is represented at URI /rsc, and enumerates all logical device types supported, and the URI of the related function set (see Section 12.3). In ZigBee 1.0 terms, this is equivalent to the list of “endpoints”, each endpoint corresponding to a function set.

The list may also contain mirrored resources for other devices (e.g., sleeping or mobile devices). The list of mirrored resources may also be found at /mrr.

13.3 Function Sets and Device Types

SE 2.0 defines a “function set” as a group of related functionalities (the equivalent of a cluster in ZigBee 1.0). Each function set defines a list of REST resources and associated transactions, and is identified by a resource name. The resource name is used for the resource discovery mechanism.

SE 2.0 defines the following function sets:

- demand response/load control;
- messaging;
- confirmation;
- pricing;
- prepayment;
- metering;
- plug-in electric vehicles;
- distributed energy resource;
- billing;
- registration;
- base;
- device management/configuration;
- firmware download server;
- firmware download client;
- diagnostics and monitoring.

```

HTTP/1.1 200 OK
Content-Type: application/exi

<?xml version='1.0' encoding='UTF-8'?>
<DeviceList xmlns='http{s}://www.zigbee.org/doc/se-2-0-0'>
  <Common>
    <Profile>Common</Profile>
    <Basic href="http{s}://{IPv6 Address}/Basic/0" name="0"/>
  </Common>
  <MeteringDevice>
    <Profile>Smart Energy</Profile>
    <Type>Electric</Type>
    <Meter href="http{s}://{IPv6 Address}/Meter/0" name="0"/>
  </MeteringDevice>

  <ESI>
    <Profile>Smart Energy</Profile>
    <Time href="http{s}://{IPv6 Address}/Tnme/0" name="0"/>
  </ESI>

  <ThermostatDevice>
    <Profile>Home Automation</Profile>
    <Thermostat href="http{s}://{IPv6 Address}/TSTAT/0"
      name="0"/>
  </ThermostatDevice>
</DeviceList>

```

Figure 13.5 Example response to GET /rsc (data represented in bold is actually compressed using EXI).

Figure 13.6 lists the device types that are defined in SE 2.0. Each device type is characterized by the function sets it must implement on the client or the server side. This is a transposition of the model of ZigBee 1.0, where each device type was characterized by the supported clusters on the server and client sides.

13.3.1 Base Function Set

The base function set groups general support resources that are useful to most end devices. Some of them can also be implemented by the ESI (e.g., the time resource). The resources currently defined as part of the base function set are briefly described in Figure 13.7 and Figure 13.8. The usage of most of them is self-explanatory.

The usage for the randomize resource and the power configuration resources is described in more detail in the following sections.

Device type	Mandatory function set client or server
In Premises Display	Client of the Metering, Price, or Message function sets.
Load Control	Client of the Demand Response Load Control function set.
Smart Thermostats	Client of the Demand Response Load Control or Price function set.
Meters	Server of the Metering function set
Smart Appliances	Client of the Demand Response Load Control or Price function set.
Premises Energy Management Systems	Both a client and a server of either the Demand Response Load Control or Price function sets
Energy Services Interface	Server of the Message, Price, and Demand Response Load Control Function Sets.
Prepayment Terminals	Client of the Price and Billing function sets
Inverters	Server of the DER function set.
Electric Vehicle Supply Equipment (EVSE)	Client of the Plug-In Electric Vehicle List, Price, and Demand Response Load Control function sets, and a server of the Plug-In Electric Vehicle.
End Use Measurement Device (EUMD)	Server of the Metering function set

Figure 13.6 ZigBee SEP 2.0 device types and mandatory function sets.

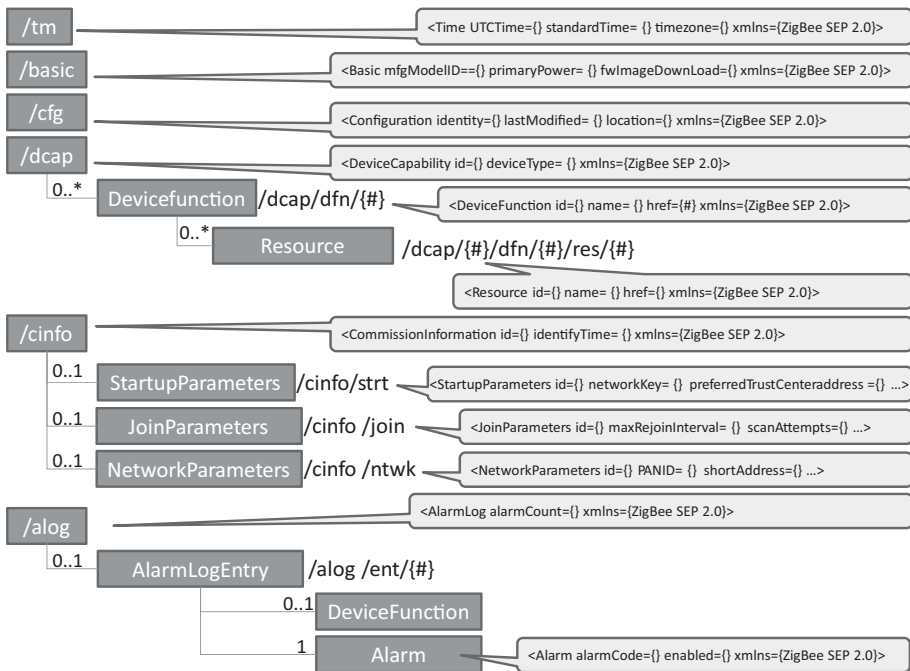


Figure 13.7 ZigBee SEP 2.0 base resources (part 1).

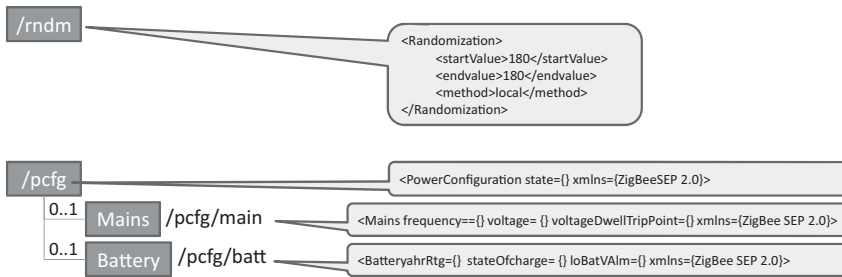


Figure 13.8 ZigBee SEP 2.0 base resources (part 2).

13.3.1.1 Randomization

Telecom service providers know the potential catastrophic effects of massive synchronization of client devices (see Appendix B). In order to provide a generic tool to avoid these situations, ZigBee SE 2.0 provides a randomization mechanism that can be used by any function set potentially affected by this problem: at present, the demand response load control, and price function sets.

Randomization is supported through the /rndm resource (see Figure 13.8 and Figure 13.9). Depending on the service-provider policy, the implementation may rely on the effective calculation of a random value, within the specified bounds, by the device, or on a fixed random value preconfigured on each device.

```
<Randomization>
  <startValue>180</startValue>
  <endvalue>180</endvalue>
  <method>local</method>
</Randomization>
```

Figure 13.9 ZigBee SE 2.0 randomization resource example.

13.3.1.2 Firmware Download

The firmware upload function is implemented by an upgrade client (UC), which polls, or optionally subscribes to the resources of an upgrade server (US). ZigBee SEP 2.0 defines a digitally signed firmware file format similar to that already used for ZigBee 1.0 over-the-air (OTA) upgrading cluster, with minor changes. Several types of files are defined (security credential, log, configuration) and identified by a two-octet value, with 0x0000-0xffbf reserved for manufacturer use (one identifier should be used per device type). This makes it possible to update or subscribe to changes of only certain file types (0xffff is defined as the wildcard file type).

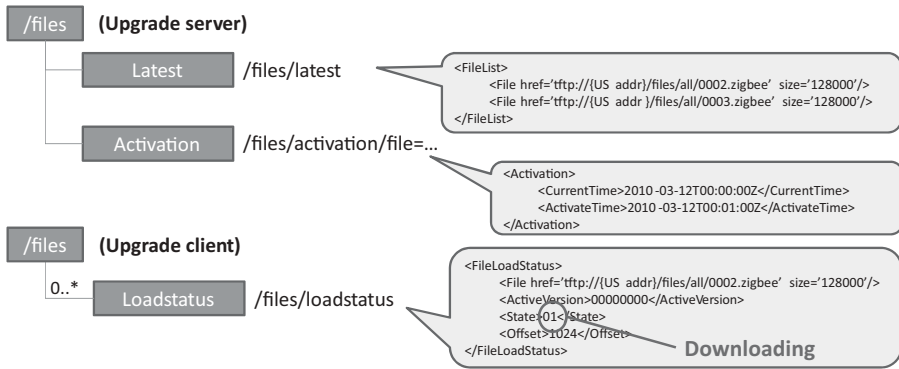


Figure 13.10 ZigBee SEP 2.0 firmware upload resources.

On the upgrade server, the URI `/files/latest` returns the list of new files targeted for the client. The request optionally specifies the manufacturer code, file type and current file version. Deferred activation is supported by means of the Activation resource. The upgrade server resources are outlined in Figure 13.10.

On the client side, the status of configuration files can be verified by reading the `/files/loadstatus` resource.

13.3.2 Group Enrollment

Energy service providers usually want to direct their commands (e.g., demand-response commands) to only a subset of all potential controlled devices. This might be to avoid synchronization effects, to control the aggregate energy volume affected by the command, or to scope the command geographically. SE 2.0 supports these requirements by providing the notion of a group enrollment resource.

The SE 2.0 group enrollment URI is `/enrl`.

```
<?xml version='1.0' encoding='UTF-8'?>
<EnrollmentList xmlns='http{s}://www.zigbee.org/doc/se-2-0-0'>
  <EnrollGroup>
    <Group>0</Group>
    <Group>1</Group>
  </EnrollGroup>
</EnrollmentList>
```

As this book was going to press, an update of the draft specification was published. Device group enrollments are now configured as a subresource of the EndDevice resource instance on the ESI. Each group contains links to the specific function set instances applicable to the EndDeviceGroup. `/enrl` resource is no longer used.

1	TimeAttribute	12 = instantaneous
2	DataQualifier	0=N/A
3	AccumulationBehaviour	6= indicating
4	FlowDirection	1=Forward
5	UomCategorySubclass	0=N/A
6	UomCategoryIndex	8=demand
7	MeasurementCategory	0=N/A
8	Enumeration	
9	Phase	0=n/a to all phases
10	Multiplier	3=kilo
11	UnitOfMeasure	38=W

Figure 13.11 IEC TC57 61968 ReadingTypeID structure.

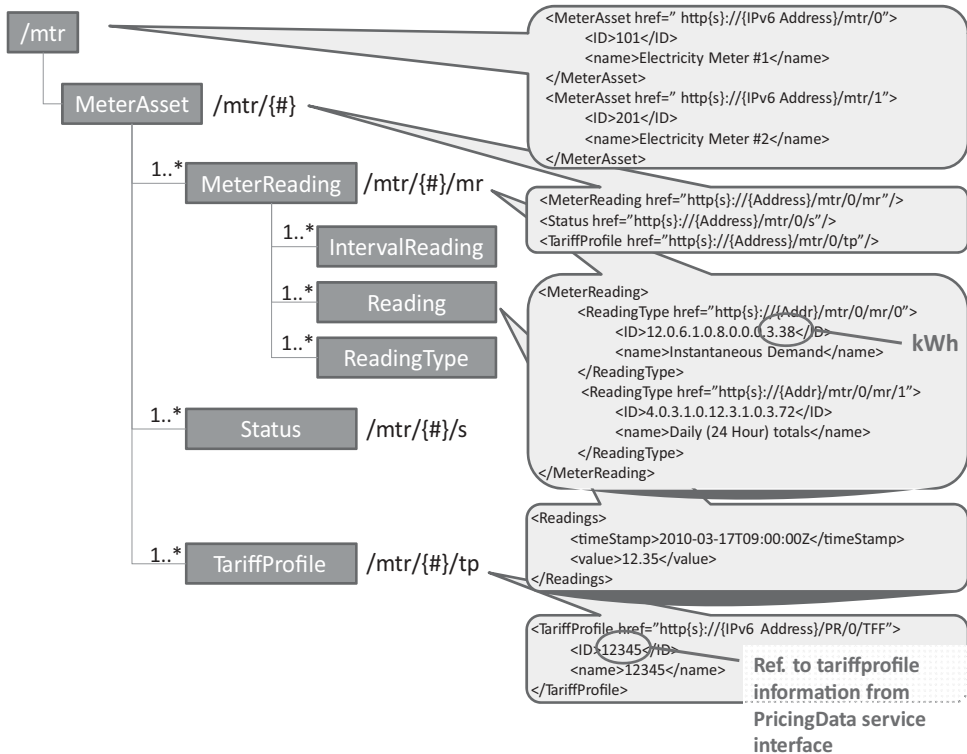


Figure 13.12 ZigBee SE 2.0 meter resources.

13.3.3 Meter

The meter structures of ZigBee SEP 2.0 have been designed in close coordination with IEC TC57 61968 and captures a baseline metering functionality. SEP 2.0 does not implement, however, more advanced functions such as programmable autoreads. The structure of the metering resources is outlined in Figure 13.12.

The ReadingTypeID format is imported from IEC TC57 61968, it is a concatenation of 9 attributes each represented by one or two integers for a total of 11 integers (Figure 13.11). For instance the present maximum indicating forward water (m³/h) is represented by 15.8.6.1.0.63.0.0.0.0.121.

13.3.4 Pricing

The price resources allow utilities to publish a description of their tariff structures on an ESI. Figure 13.13 illustrates the relationships between price resources and shows the “well-known” URIs defined by SE 2.0.

The ZigBee SE 2.0 pricing resources support both time of use (ToU) pricing and consumption interval pricing, in any combination. The example of Figure 13.14, adapted from draft 0.7, shows how time-based pricing and volume-based pricing can coexist.

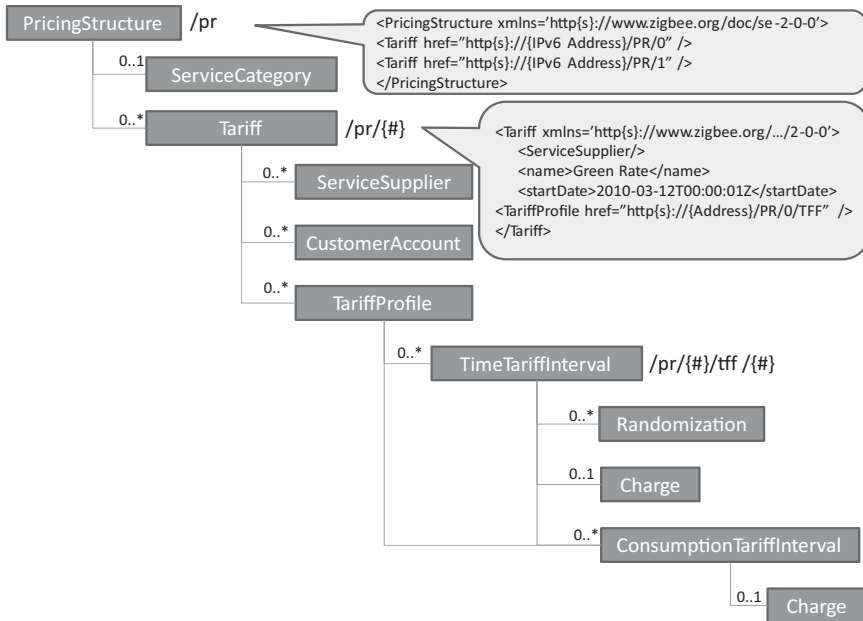


Figure 13.13 ZigBee SE 2.0 pricing resources.

```

<!-- Midnight to 8:00am -->
<?xml version='1.0' encoding='UTF-8'?>
<TimeTariffInterval xmlns='http(s)://www.zigbee.org/doc/se-2-0-0'>
  <startDateTime>2010-03-12T00:00:01Z</startDateTime>
  <!-- 0 to 100 kWh -->
  <ConsumptionTariffInterval>
    <startValue>0</startValue>
    <Charge>
      <fixedPortion>
        <value>0.05</value>
        <monetaryUnit>USD</monetaryUnit>
      </fixedPortion>
    </Charge>
  </ConsumptionTariffInterval>
  <!-- > 100 kWh -->
  <ConsumptionTariffInterval>
    <startValue>100</startValue>
    <Charge>
      <fixedPortion>
        <value>0.07</value>
        <monetaryUnit>USD</monetaryUnit>
      </fixedPortion>
    </Charge>
  </ConsumptionTariffInterval>
  <Randomization>
    <flag>TRUE</flag>
    <type>Start</type>
  </Randomization>
  <Randomization>
    <flag>TRUE</flag>
    <type>End</type>
  </Randomization>
</TimeTariffInterval>

```

This tariff interval implicitly ends at the startDateTime of the next timeTariffInterval

This consumption interval implicitly ends at the startValue of the next consumption interval

Figure 13.14 Example TimeTariffInterval resource.

13.3.5 Demand Response and Load Control Function Set

The clients of the DR/LC function set are typically smart thermostats or any device that supports load control. The server side is typically the ESI and implements the following resources (see Figure 13.15):

- /dr : a collection of DemandResponseProgram collections;
- /dr/{#} : a specific DemandResponseProgram collection resource;
- /dr/{#}/nm : a specific DemandResponseProgram name attribute resource;
- /dr/{#}/edc : a collection of EndDevicesControls;
- /dr/{#}/edc /{#} : a specific EndDevicesControl resource;
- /dr/{#}/edc /act : the active EndDevicesControl resource.

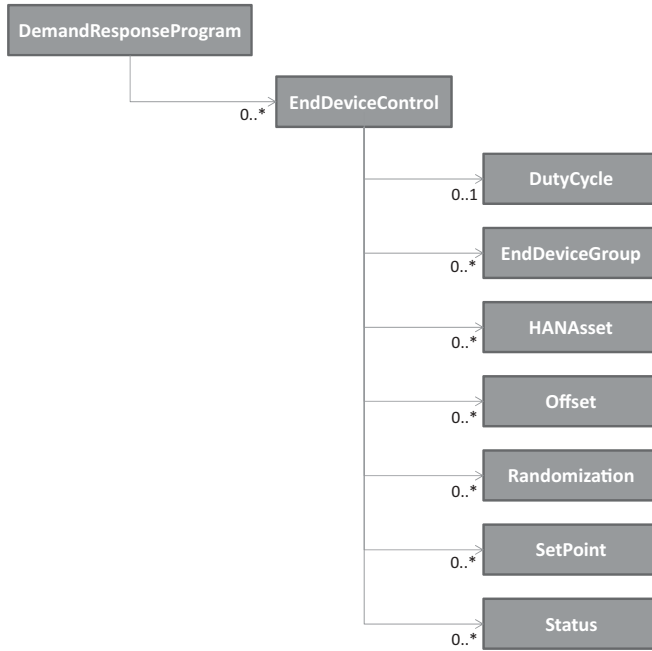


Figure 13.15 ZigBee SE 2.0 DR/LC server resources.

The EndDeviceControl resource contains or points to all the information required by a client to implement a particular dr/lc event for a period of time.

In addition to the attributes listed in Figure 13.16, each EndDeviceControl resource may also contain the subresources listed in Figure 13.17.

DR/LC attributes	Description
ProgramLevel (integer)	Level of a demand response program request, where 0=emergency.
drProgramMandatory (boolean)	Whether a demand response program request is mandatory
duration (Minutes)	Event duration (end - start)
href (anyURI)	Hypertext reference pointing to a URI
ID (string)	Object identifier
name (string)	Name of the EndDeviceControl resource
scheduledInterval (DateTimeInterval)	(if control has scheduled duration) Date and time interval the control has been scheduled to execute within.
type (string)	Type

Figure 13.16 EndDeviceControl attributes.

Sub-Resource	Sub-Resource attributes
DutyCycle	<ul style="list-style-type: none"> – name (string): Duty cycle name – normalValue (PerCent): Duty cycle value such as 80% – state (string): State such as on or off
EndDeviceGroup	<ul style="list-style-type: none"> – groupAddress (integer) : Address of this end device group. – href (anyURI) : Hypertext reference pointing to a URI – ID (object identifier)
HANAsset	<ul style="list-style-type: none"> – category (string): Utility-specific categorization of this document. – href (anyURI) : Hypertext reference pointing to a URI – ID (string) : HAN asset identifier
Offset : Offset such as cooling or heating offset	<ul style="list-style-type: none"> – name (string) – normalValue (PerCent) – Offset as per cent – type (string) – Offset type – value (string) : offset value
Randomization : Randomization for start or end of an event	<ul style="list-style-type: none"> – endValue (unsignedInt) : End randomization value in SEP UTCTime format such 300 for 5 minutes – flag (boolean) : Randomization or not – href (anyURI) : Hypertext reference pointing to a URI – ID (string): Identifier – method (string) – Local (maximum) or static randomization – name (string) : Randomization name – startValue (unsignedInt) : Start randomization value in SEP UTCTime format such 300 for 5 min – type (string) : Randomization type (start or end randomization)
SetPoint : A SetPoint is an analog control used for supervisory control.	<ul style="list-style-type: none"> – maxValue (float) : Normal value range maximum for any of the Control.value. Used for scaling, e.g. in bar graphs. – minValue (float) : Normal value range minimum for any of the Control.value. Used for scaling, e.g. in bar graphs. – name attribute (string) : Name of an attribute. – normalValue (float) : Normal value for Control.value e.g. used for percentage scaling – value (float) : Value in type of float
Status: Current status information relevant to an entity.	<ul style="list-style-type: none"> – dateTime (unsignedInt) : Date and time for which status ‘value’ applies. – href (anyURI) : Hypertext reference pointing to a URI – reason: Reason code or explanation for why an object went to the current status ‘Value’. – value (string) : Value in string

Figure 13.17 DR/LC EndDeviceControl subresources.

The DR client would typically send the following request to the ESI:

```
GET /esi HTTP/1.1
Host: {IPv6 Address}
```

The server would respond:

```
HTTP/1.1 200 OK
Content-Type: application/exi
<?xml version='1.0' encoding='UTF-8'?>
<EndDeviceControl href='http{s}://{IPv6 Address}/dr/0/edc/0'>
<ID>101</ID>
<Randomization>
<flag>TRUE</flag>
<type>Start</type>
</Randomization>
<Randomization>
<flag>TRUE</flag>
<type>End</type>
</Randomization>
</EndDeviceControl>
```

13.3.6 Distributed Energy Resources

The ZigBee S2.0 SEP resources that model distributed energy resources are outlined in Figure 13.18.

The power-generation curve is perhaps the most important piece of information from a utility perspective. This resource stores the stepwise linear approximation of the power-generation curve, an example is given in Figure 13.19.

The DERstatus object contains a description of the operational state of the generating unit, such as “starting up”, the number of times the generator has been started since the last counter reset, and the total operation time of the generator since the last counter reset.

The DERcontrol object allows a utility to offset the production level to a certain percentage, and includes options to randomize such commands.

13.3.7 Plug-In Electric Vehicle

One of the priority action plans (PAP 11¹) of the US National Institute of Standards and Technology (NIST) Smart Grid Interoperability Panel (SGIP) is to design common object models for electric transportation. PAP11 formulates the issue as follows:

¹ <http://collaborate.nist.gov/twiki-sggrid/bin/view/SmartGrid/PAP11PEV>.

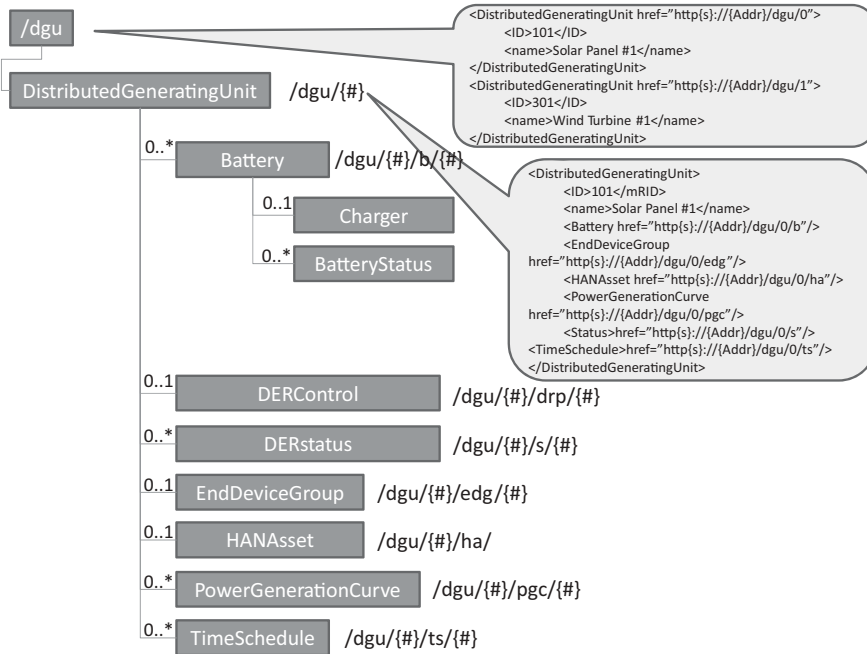


Figure 13.18 ZigBee 2.0 SEP distributed energy resources data model.

```

<PowerGenerationCurve href="http{s}://{IPv6 Address}/dgu/0/pgc/0">
  <curveType></curveType>
  <description></description>
  <ID></ID>
  <xUnit>s</xUnit>
  <y1Unit>W</y1Unit>
  <CurveData>
    <xvalue>1.0</xvalue>
    <y1value>12.34</y1value>
  </CurveData>
  <CurveData>
    <xvalue>2.0</xvalue>
    <y1value>42.34</y1value>
  </CurveData>
  <CurveData>
    <xvalue>3.0</xvalue>
    <y1value>72.34</y1value>
  </CurveData>
</PowerGenerationCurve>
    
```

Figure 13.19 ZigBee 2.0 SEP PowerGenerationCurve example.

The introduction of mobile plug-in electric vehicles (PEVs) to the grid creates some interoperability challenges around exchanging price, demand response (DR), and settlement information. The impact of PEVs on the grid is expected to be significant, and the ability to control the charging profiles through price or direct control, the need for cyber security (including appropriate privacy), the issues of safety, the possibility of allowing customers to sell PEV electricity back into the grid, and complexity of providing fair settlement to everyone in the value chain when vehicles charge away from their home base, requires common object models to manage all these aspects.

As of March 2011, no firm decision had been made by the SGIP governing board to adopt the ZigBee SE 2.0 information model (repackaged as SAE J2847/1) as a standard, however, it was the most likely candidate to be adopted during the ongoing standardization process to be held within the SGIP Vehicle to Grid (V2G) working group,² and within IEC TC 57 (ZigBee SE 2.0 PEV and DER information model are designed as an extension of the common information model, or CIM, defined in IEC 61968 and IEC 61850). For more background information on EV charging, refer to Chapter 6.

The structure of the ZigBee 2.0 SEP resources that compose the PEV function set is outlined in Figure 13.20.

The ElectricVehicle object is characterized by the objects listed in Figure 13.20, and following attributes:

- **ID**: a string;
- **odometerReadDateTime**: an unsigned integer indicating the date of the odometer³ reading;
- **odometerReading**: a string indicating the actual reading;
- **status**: an indicator of the status of the battery, which includes indicators for over and under charge conditions, and a per cent indicator of the state of charge.

The battery object indicates the battery technology (batTyp attribute), Ah capacity rating of the battery (ahrRtg), and battery nominal voltage.

A specific charger resource is designed to retrieve the battery charge information: `http{s}://{IPv6 Address}/pev/{#}/btrr/{#}/ct` where “ct” stands for “charge transaction”. The charger resource indicates, among other things, the active power charge rate (reChArTc attribute) of the battery, and the charging schedule (start and end time).

The battery status subresource: `http{s}://{IPv6 Address}/pev/{#}/btrr/{#}/s` includes indicators for over- and undercharge conditions, and a per cent indicator of the state of charge for the specific battery.

The DemandResponseProgram collection contains a list of pointers to DemandResponsePrograms, identified by their ID (the actual DemandResponseProgram resource resides on the ESI), and applying to this PEV.

² <http://collaborate.nist.gov/twiki-sgrid/bin/view/SmartGrid/V2G>.

³ “An odometer indicates distance traveled by a car or other vehicle” (Wikipedia).

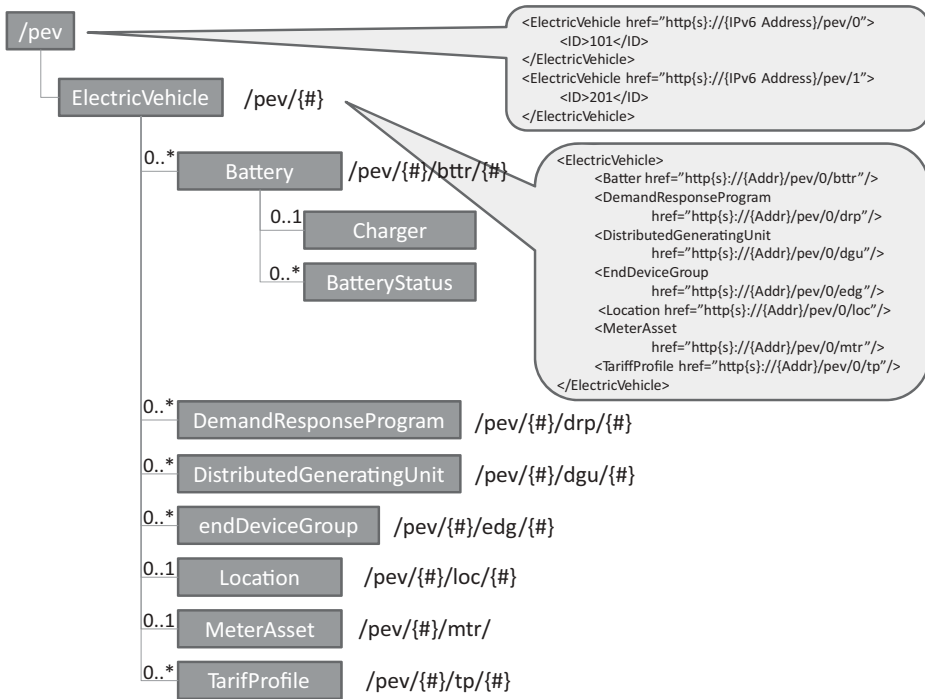


Figure 13.20 ZigBee 2.0 SEP PEV resources.

13.3.8 Messaging

The messaging function set uses two types of resources:

- TextMessages (see Figure 13.21) is typically implemented in the ESI and must be polled or subscribed to by the clients. The resource URI `http{s}://{server IPv6 Address}/msg/` provides access to a collection of message collections. Individual messages are accessed by their individual URIs `http{s}://{server IPv6 Address}/msg/{#}/txt/{#}`. The structure of each TextMessage, as illustrated on Figure 13.21, enables the ESI to display each message to only certain groups of devices, or to certain device types, at during certain periods of time. Each message may request a confirmation.
- Confirmations. That resource is implemented (or mirrored to a parent device) by in premises displays that are capable of message confirmation. The resource URI `http{s}://{IPv6 Address}/rsp/` provides access to a collection of responses. The structure of the confirmation resource includes the confirmed message ID and a time stamp. It can also be used to confirm prices or billing.

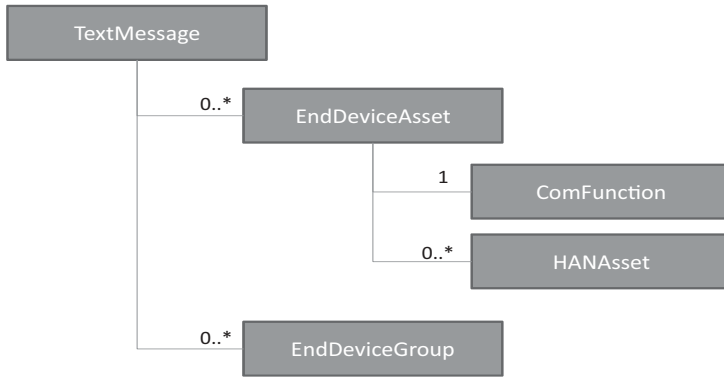


Figure 13.21 ZigBee SE 2.0 TextMessage resource.

13.3.9 Registration

The Registration function set and associated resources (Figure 13.22) are used to define which devices are authorized to access which function sets.

The information related to each end device is stored in the HANAsset resource and subresources. Keeping the harmonization with the IEC Common Information Model, the HANAsset class is a type of CIM “EndDeviceAsset”.

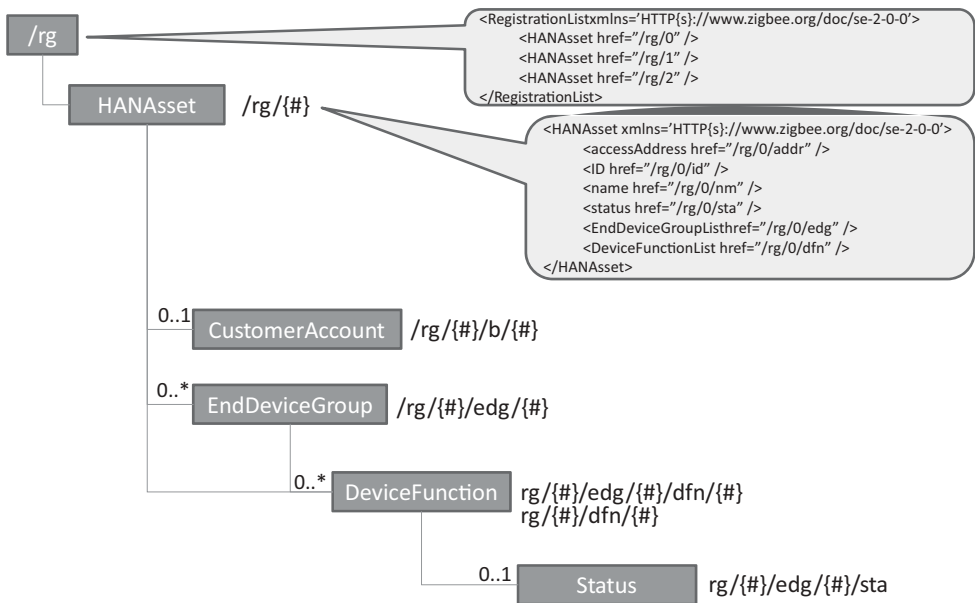


Figure 13.22 ZigBee SEP 2.0 registration resources.

The DeviceFunctions list the FunctionSets that the device has been granted access to prior to registration, or has requested access to through the registration process. The “status” of the DeviceFunction registration reflects whether access has been granted or if the request is still pending. An end device may declare its membership in one of several EndDeviceGroups. The DeviceFunctions that are declared as subresources of an EndDeviceGroup are function sets that can be accessed by all members of the group.

A number of shortcut REST URIs have been defined to access HANAsset attributes, which are listed in Figure 13.23.

Note: As this book was going to press, an update of the draft specification was published. The registration of each device is now materialized by an instance of EndDevice resource on the server to which the device registered, e.g. /edev/1 instead of rg/1.

13.4 ZigBee SE 2.0 Security

13.4.1 Certificates

A device certificate is installed during manufacturing on each SE device, and is not expected to change during the lifetime of the device (as long as the network ID, e.g., EUI-64, of the device does not change).

An operational certificate is used to secure a given session or relationship. The registration server is used to perform registration and grants an operational certificate.

13.4.2 IP Level Security

The first thing an SE 2.0 device needs to do is to access the network.

Each layer 2 technology defines specific means for the authentication of new nodes, which is problematic for technologies that want to remain layer 2 agnostic, like SE 2.0. PANA (RFC 5191) solves this issue by defining a protocol that allows clients to authenticate themselves to the access network using IP protocols. The PANA client (PaC) is configured with an IPv6 address, and exchanges IP UDP PANA authentication messages with the PANA authentication agent (PAA). The network enforcement point (EP) is supposed to let unauthenticated nodes to communicate with the IP address of the PAA : it may be collocated with the PAA, or it acts as a PANA relay element (PRE). Communication between the PaC and the PRE uses UDP over port 716. Typically, on a 802.15.4 radio network, the PRE would be the parent node (PN) and the joining node would be the PaC, both would be using their link local IPv6 addresses (see Chapter 1 for more information on 802.15.4).

Upon successful authentication, the PaC is granted broader network access possibly by a new IP address assignment (typically, add a IPv6 global address using the prefix obtained during earlier IPv6 router discovery), by enforcement points changing filtering rules for

/rg/{#}/addr	Access address for a specific HANAsset object. E.g. : <HANAsset accessAddress={Addr} xmlns=' HTTP{s}://www.zigbee.org/doc/se-2-0-0' />
/rg/{#}/id	unique resource id for a specific HANAsset object.
/rg/{#}/nm	name of a specific HANAsset object E.g: <HANAsset name="Device0" xmlns=' HTTP{s}://www.zigbee.org/doc/se-2-0-0' />
/rg/{#}/sta	status value for a specific HANAsset object.
/rg/{#}/edg	is a collection of EndDeviceGroup objects. E.g.: <EndDeviceGroupList xmlns=' HTTP{s}://www.zigbee.org/doc/se-2-0-0'> <EndDeviceGroup href="/rg/0/edg/0" /> <EndDeviceGroup href="/rg/0/edg/1" /> </EndDeviceGroupList>
/rg/{#}/edg/{#}	specific EndDeviceGroup object.
/rg/{#}/edg/{#}/addr	Group address of a specific EndDeviceGroup object. <EndDeviceGroup groupAddress={IPv6 Address} xmlns=' HTTP{s}://{ZigBee SEP}' />
/rg/{#}/edg/{#}/id	unique resource id for a specific EndDeviceGroup object.
/rg/{#}/edg/{#}/dfn	Collection of DeviceFunction objects associated with a specific EndDeviceGroup object. <DeviceFunctionList xmlns=' HTTP{s}://www.zigbee.org/doc/se-2-0-0'> <DeviceFunction href="/rg/0/edg/0/dfn/0" /> <DeviceFunction href="/rg/0/edg/0/dfn/1" /> </DeviceFunctionList>
/rg/{#}/edg/{#}/dfn/{#}	specific DeviceFunction object associated with an EndDeviceGroup
/rg/{#}/edg/{#}/dfn/{#}/id	unique resource id for a specific DeviceFunction object associated with an EndDeviceGroup.
/rg/{#}/edg/{#}/dfn/{#}/id	unique resource id for a specific DeviceFunction object associated with an EndDeviceGroup.
/rg/{#}/edg/{#}/dfn/{#}/nm	name of a specific DeviceFunction object associated with an EndDeviceGroup.
/rg/{#}/edg/{#}/dfn/{#}/sta	Registration status for a specific HANAsset, DeviceFunction, and EndDeviceGroup.

Figure 13.23 ZigBee SEP 2.0 registration resource URIs.

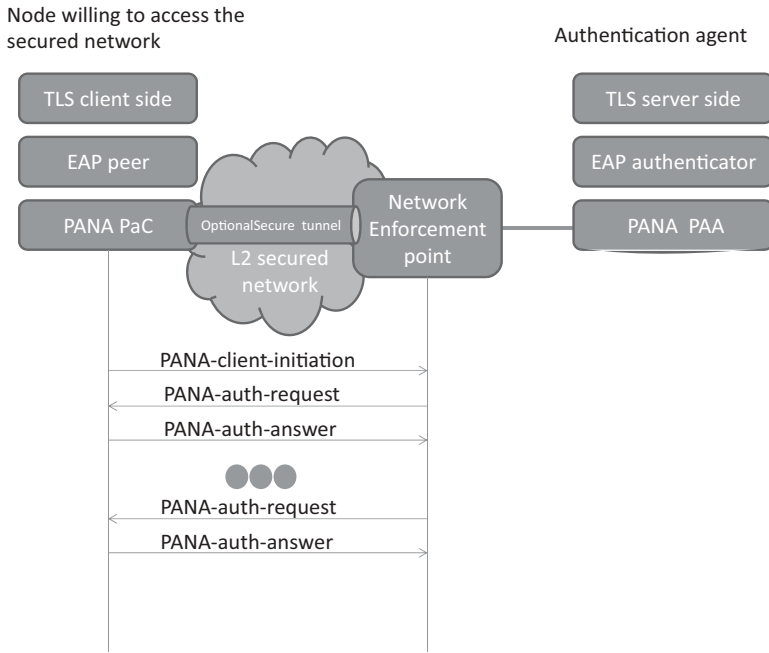


Figure 13.24 Overview of PANA.

the same IP address, or in the specific case of 802.15.4 by obtaining the network key. Extensions to PANA are currently being introduced to support such group key distribution (draft-ohba-pana-keywrap-02 as of March 2011).

Other than that, PANA simply encapsulates EAP payload, which enables use of any authentication method supported by EAP. Figure 13.24 illustrates the security stack used to access the network, and the use of PANA.

Integrity protection of messages between the PaC and PAA is possible as soon as EAP exchanges have generated a secure association (SA) shared key. In addition PANA can establish an optional IPsec tunnel to the PAA if the layer two is not secure and encryption needs to be provided.

ZigBee SE 2.0 uses EAP-TLS (RFC 5216), an EAP transport method for transport layer security (TLS). It uses TLS 1.2 (RFC 5246), and mandates support for the cipher suites TLS_PSK_WITH_AES_128_CCM_8 and TLS_ECDHE_ECDSA_WITH_AES_128_CCM_8. The AES-CCM cipher is implemented in many 802.15.4 chipsets in hardware. The standard TLS handshake is used to provide mutual authentication and to derive a shared security association.

The ZigBee *authentication server* manages incoming EAP-TLS transactions from joining nodes, performs network authentication and controls network access. It is typically

hosted by the ESI. Its functional role is roughly equivalent to that of the application trust center in ZigBee 1.0.

13.4.3 Application-Level Security

13.4.3.1 Registration

Once the device is able to communicate to other nodes on the IP network, it still needs to register to the utility or service provider registration server (e.g., the ESI) in order to be able to access the service provider SE 2.0 resources.

The device authenticates with the registration server using TLS handshake based on device certificates, and the registration server grants one or more operational certificates to the device (using the TLS record protocol): one operational certificate is associated to the device, and operational certificates are associated to individual resources on the device. At the application level, ZigBee SE 2.0 uses TLS cipher suites TLS_DHE_RSA_WITH_AES_128_GCM_SHA256 and TLS_ECDHE_ECDSA_WITH_1854_AES_128_GCM_SHA256, and X.509v3 certificates. A device may register with multiple registration servers (e.g., multiple utilities or service providers).

13.4.3.2 Authorization Server, ACLs

Any device that contains protected resources must also implement the authorization server function: it authenticates the client requesting access to the resource using TLS and the operational certificates, and then uses TLS negotiation to establish a secure tunnel. Therefore, in addition to native local network security, ZigBee SE 2.0 provides application-level secure tunnels. These secure tunnels are specific to each set of operational certificates, therefore to each utility registration: multiple utilities may securely share the same SE 2.0 network. Once the identity of the client has been asserted, the authorization server then uses application-defined ACLs to restrict access to the protected resources according to its security policy.

Privilege	Description
GET	allowed to perform the GET method on the resource
PUT	allowed to perform the PUT method on the resource
POST	allowed to perform the POST method on the resource
DELETE	allowed to perform the DELETE method on the resource
GET_ACL	allowed to perform the GET method on the acl subresource
PUT_ACL	allowed to perform the PUT method on the acl subresource
POST_ACL	allowed to perform the POST method on the acl subresource
DELETE_ACL	allowed to perform the DELETE method on the acl subresource

Figure 13.25 ACL privileges defined in ZigBee SE 2.0.

```

<AccessControlList xmlns='http{s}://www.zigbee.org/doc/se-2-0-0'>
  <Grant>
    <ID>{IPv6 Address}</ID>
    <Privilege>GET_ACL</Privilege>
  </Grant>
  <Grant>
    <ID>{IPv6 Address2}</ID>
    <Privilege>GET</Privilege>
    <Privilege>PUT</Privilege>
    <Privilege>POST</Privilege>
    <Privilege>DELETE</Privilege>
    <Privilege>GET_ACL</Privilege>
    <Privilege>PUT_ACL</Privilege>
    <Privilege>POST_ACL</Privilege>
    <Privilege>DELETE_ACL</Privilege>
  </Grant>
  <!-- Grant GET Access to ULAs -->
  <Grant>
    <ID>fc00: :/7</ID>
    <Privilege>GET</Privilege>
  </Grant>
</AccessControlList>

```

Figure 13.26 Example ZigBee SE 2.0 ACL.

SE 2.0 ACLs are designed as white lists: access to a resource via a REST method is not granted unless an ACL explicitly allows it. The ACL of a resource is defined as subresource `//{host address}/{resource}/acl`, and lists which privileges are granted to each client host (see Figure 13.25).

An example ACL resource is provided in Figure 13.26.

14

The ETSI M2M Architecture

14.1 Introduction to ETSI TC M2M

At present, there are about 50 to 70 billion “machines” in the world, about 1% of which are connected to a communication network. There is obviously an enormous growth potential for M2M, but the transition from current midscale M2M applications (about 500 000 devices) to the next level (applications managing tens of millions of devices) will require new standards.

While current M2M standards address the transport level, and client to server communication protocols, the future “Internet of Things” will require a system-level architecture:

- Enabling application developers to focus on functionality, not lower-level tasks like network access control, authentication or routing;
- Enabling any application to read or control any sensor, under control of a horizontal security framework;
- Providing network-based services, such as data publication and subscription.

In order to achieve these goals, common functions and network elements need to be identified and standardized at part of the M2M infrastructure: the ETSI M2M technical committee was created in January 2009 at the request of many telecom operators to create a standard system-level architecture for mass-scale M2M. ETSI TC M2M does not address one domain in particular; on the contrary, its ambition is to become the common backbone of all mass-scale M2M applications. The following domains are explicitly covered:

- Security/serenity: surveillance applications, alarms, object/people tracking;
- Transportation: fleet management, road safety;
- Health care: personal security, e-health;
- Smart energy: measurement, provisioning and billing of utilities;
- Supply and provisioning: freight supply and distribution monitoring, vending machines;
- City automation: public lighting management, waste management;
- Manufacturing: production chain monitoring and automation.

The Internet of Things: Key Applications and Protocols, First Edition.

Olivier Hersent, David Boswarthick and Omar Elloumi.

© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

As for all recent automation protocols, the ETSI M2M architecture is resource centric and adopts the RESTful style (refer to Section 13.1 for more details on REST). As usual, the four basic verbs of REST (create, read, update, delete) are complemented at the functional level by execute, subscribe and notify primitives, which are implemented, at a lower level, by helper resources manipulated by the CRUD verbs. See Section 14.3.6 for example.

We expect ETSI M2M to become the system-level architecture of the Internet of Things, much in the same way as GSM or UMTS (also from ETSI) have become the dominant system-level architectures for mobile communications.

ETSI M2M does not aim at replacing existing standard or proprietary automation protocols, such as those described in the other chapters of this book. It aims at integrating all of these protocols into a common architecture, facilitating access to any of these vertical protocols and networks from any hosted service, in an operator-controlled way. The companion book “M2M Communications: A Systems Approach” (David Boswarthick, Omar Elloumi, Olivier Hersent) describes ETSI M2M general architecture in more details. In this chapter we will just give an overview of ETSI M2M architecture, and explain in more detail how ETSI M2M can integrate existing automation and metering protocols, using the example of ZigBee 1.0, DLMS and C.12.

14.2 System Architecture

14.2.1 High-Level Architecture

The ETSI M2M functional architecture is presented in ETSI TS 102 690. The ETSI M2M system architecture separates the M2M device domain and the network and applications domain (Figure 14.1):

- The **device domain** is composed of **M2M devices** and **M2M gateways**. ETSI M2M devices can connect to the M2M network domain directly (D’ devices) or via M2M gateways acting as a network proxy (D devices). M2M gateways can be cascaded, or operate in parallel mode (e.g. for redundancy purposes).
- The **network and applications domain** comprises:
 - The access and transport network (e.g., an xDSL access network and an IP transport network).
 - The **M2M core**, which itself is composed of:
 - a **Core network** (which provides IP connectivity, service and network control functions, network to network interconnect and roaming support); and
 - **M2M service capabilities**, the functional modules implementing the M2M functions shared by multiple applications through open interfaces.
 - The **M2M applications** that run the M2M service logic and use the M2M service capabilities. The ETSI M2M architecture supports multiagent applications, which can have components running in the end devices (device application or DA), in the gateways (gateway application or GA) and in the network (network application or NA).

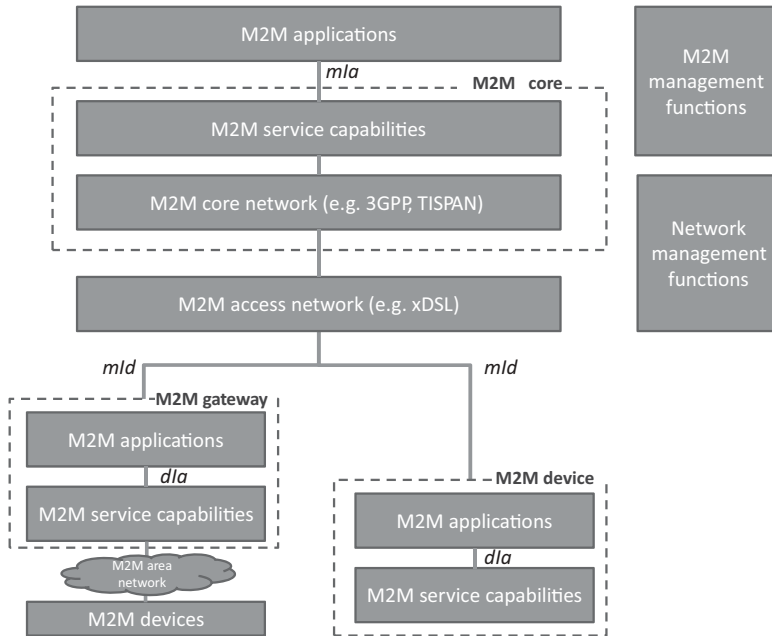


Figure 14.1 ETSI M2M high-level architecture.

- The network management functions address the access, transport and core network provisioning and supervision.
- The M2M management functions provide the services required for M2M service bootstrapping (M2M service bootstrap function or MSBF), and M2M security (M2M authentication server or MAS).

14.2.2 Reference Points

TS 102-690 defines three reference points:

- **m1a** between a M2M Network Application (NA) and the M2M service capabilities in the networks and applications domain. It provides registration and authorization primitives for the NA, service session management (event reporting or streaming sessions), and read/write/execute/subscribe/notify primitives for objects or groups of objects residing in M2M devices or gateways, as well as group objects managed by the network-domain capabilities.
- **d1a** between:
 - (a) a device application (DA) and M2M service capabilities in the same M2M device or in a M2M gateway;
 - (b) a gateway application (GA) and M2M service capabilities in the same M2M gateway.

GAE	This function registers gateway applications and is the single contact point for gateway applications via interface dIa, it hides the service capabilities topology and performs routing towards capabilities.
GGC	Manages the secure transport session establishment and policing according to the messages service class, using GSEC to retrieve session keys. Performs routing between Service Capabilities and the NGC Domain over interface mId.
GRAR	Provides a network storage capability for state associated to named M2M devices and handles subscriptions to state changes. Key information items include routable addresses or reachability status. The GRAR also acts as a group manager for M2M devices.
GCS	The GCS is the network selection function when the core network is reachable via several alternative access networks or when the gateway owns several routable addresses, according to the service class of the messages to be routed or other policies. It also provides alternative network or communication service selection in case of failures.
GREM	Acts as a management proxy for the NREM.
GSEC	Implements key management, service layer registration, session key management.
GTM	Optional transaction management.

Figure 14.2 Gateway capabilities.

dIa provides registration and authorization primitives for DAs and GAs to the device/gateway, service session management (event reporting or streaming sessions), and read/write/execute/subscribe/notify primitives for objects or groups of objects residing in M2M devices or gateways, as well as group objects managed by the device/gateway capabilities.

- **mId** between an M2M device or M2M gateway and the M2M service capabilities in the network and applications domain. mId provides registration and authorization primitives for DAs and GAs to the M2M core, service session management (event reporting or streaming sessions), and read/write/execute/subscribe/notify primitives for objects or groups of objects residing in M2M devices or gateways, as well as group objects managed by the devices, gateways or in the network core capabilities.

TS 102 921 “machine-to-machine communications (M2M); mIa, dIa and mId interfaces” specifies the actual implementation of these interfaces over several protocol bindings, currently HTTP and CoAP.

14.2.3 Service Capabilities

The service capabilities can reside in the end device (in the acronyms below x=D for device), in the gateway (x=G, Figure 14.2), or in the network (x=N, Figure 14.3). At present, the following M2M service capabilities have been defined:

- application enablement (xAE);
- generic communication (xGC);
- reachability, addressing and repository (xRAR);
- communication selection (xCS);
- remote entity management (xREM);
- security (xSEC);
- history and data retention (xHDR);
- transaction management (xTM);
- compensation broker (xCB);
- telco operator exposure (xTOE);
- interworking proxy (xIP).

Figure 14.4 illustrates the role of service capabilities in the overall ETSI M2M architecture.

NAE	This function registers network applications and is the single contact point for network applications via interface mla, it hides the service capabilities topology and performs routing towards capabilities.
NGC	Manages the secure transport session establishment and policing according to the messages service class, using NSEC to retrieve session keys. Performs routing between M2M Devices, M2M Gateways, Service Capabilities and M2M Application residing in the Network and Applications Domain. It supports unicast, multicast and anycast.
NRAR	Provides a network storage capability for state associated to named M2M devices, M2M gateways or groups of M2M devices or gateways, and handles subscriptions to state changes. Key information items include routable addresses or reachability status. The NRAR also acts as a group manager for M2M devices and gateways.
NCS	The NCS is the network selection function for devices reachable via several routable addresses, according to the service class of the messages to be routed or other policies. It also provides alternative network or communication service selection in case of failures.
NREM	Provides firmware update support, configuration management functions, performance and fault management. Typically proxies requests to a Broadband forum TR-069 auto-configuration server.
NSEC	Implements key management, service layer registration, session key management. Interfaces with a M2M Authentication server (MAS) (e.g. via Diameter) to obtain authentication data.
NDHR	Optional transaction data archival for legal requirements.
NTM	Optional transaction management.
NCB	Optional financial compensation service.
NTOE	Exposes services such as SMS, MMS, USSD, localization, etc.

Figure 14.3 Network-domain capabilities.

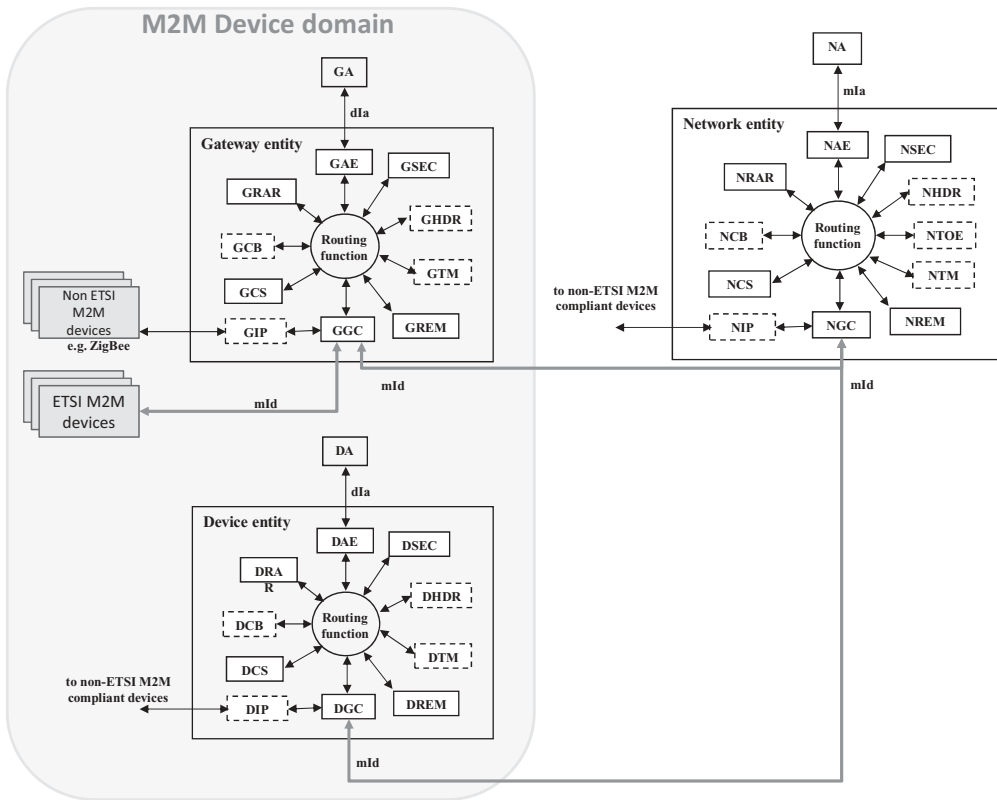


Figure 14.4 M2M system architecture overview.

14.3 ETSI M2M SCL Resource Structure

As expected in a REST architecture style, the functionality of the ETSI M2M service capability layer (SCL) is exposed as a set of addressable resources. Currently, TS 102 921 specifies two serialization options for resources: XML or JSON. For XML, support for FastInfoset (ITU-T X.891 | ISO/IEC 24 824-1) and efficient XML interchange (EXI, <http://www.w3.org/TR/exi/>) optimizations are recommended.

A given M2M entity exposes its resources as subresources of sclBase. The sclBase resource is represented by an absolute URI. All other resources hosted or registered in the SCL are identified by a URI that is hierarchically derived from the URI of the sclBase. Each subresource “x” of multiplicity 1 is modeled as an attribute named xReference and containing the subresource URI, and each subresource “y” of multiplicity “0..unbounded” is represented as one collection attribute named yCollection that contains a list of subresource URIs.

The ETSI M2M resource hierarchy is outlined in Figure 14.5. For clarity, not all resources currently defined by ETSI M2M are represented. Some subresources are “virtual”,

ScIbase	Attributes including references to sub-resources	Sub-resource attributes
	accessRightID creationTime lastModifiedTime pocs searchStrings sclsReference/	ID of the access rights resource applicable to this SCL Creation time attribute Last modification time attribute Points of Contact attribute Tokens used as keys for searching resources and contents Reference to SCLs collection subresource
		accessRightId Access right id of the parent resource creationTime Creation time lastModifiedTime Last modification time mgmtObjReference Reference to Management object resources sclCollection Reference to registered SCL resources subscriptionsReference Reference to Subscription resources, see Figure 12
	applicationsReference	Reference to Applications collection subresource
		accessRightId Access right id of the parent resource creationTime Creation time lastModifiedTime Last modification time applicationCollection Reference to application resources, see Figure 7 mgmtObjsReference Reference to management objects collection resources subscriptionsReference Reference to Subscriptions collection resource
	containersReference	Reference to Containers collection subresource
		accessRightId Access right id of the parent resource creationTime Creation time lastModifiedTime Last modification time containerCollection References to Container resources, see Figure 10 locationCollection References to Location resources subscriptionsReference Reference to Subscription collection resource
	groupsReference	Reference to Groups collection subresource
		accessRightId Access right id of the parent resource creationTime Creation time lastModifiedTime Last modification time groupCollection References to group resources, see Figure 11 subscriptionsReference Reference to Subscription collection resource
	accessRightsReference	Reference to AccessRights collection subresource
		accessRightId Access right id of the parent resource creationTime Creation time lastModifiedTime Last modification time accessRightCollection Reference to access right resources, see

Figure 14.5 ETSI M2M resource hierarchy overview.

	subscriptionsReference	Figure 9 Reference to Subscription collection resource
subscriptionsReference		Reference to Subscriptions collection subresource
	subscriptionCollection	Reference to access right resources, see Figure 12
discoveryReference		Reference to Discovery resource (virtual)
	discoveredURIs	List of resource URIs matching searchPrefix and filterCriteria
	matchSize	number of resources that matched the filterCriteria
	statusCode	
	Truncated	If URI list had to be truncated

Figure 14.5 (Continued)

that is, they serve to implement a specific function and typically can only be read and do not have an e-tag.

More detailed views can be found in Figure 14.6 for SCL resources, in Figure 14.7 for application resources, in Figure 14.10 for container resources, in Figure 14.9 for access right resources, in Figure 14.11 for group resources, in Figure 14.12 for subscription resources, in Figure 14.13 for notification channels resources.

Applications register to a SCL simply by creating a new application resource under the applications collection subresource. Remote SCLs register to a local SCL by creating a new SCL resource under the SCLs collection subresource.

Basic REST primitives define how to create, read, update or delete a resource in its entirety, but there are many cases where an application wants to interact with only a part of the resource. TS102921 provides a number of conventions enabling partial resource read, write or subscription. For instance, single attributes or elements can be addressed. For resources formatted as XML content, ETSI M2M supports the XCAP (XML configuration access protocol, RFC 4825) approach, which allows arbitrary xpath expressions to manipulate parts of the XML representation of a resource.

14.3.1 SCL Resources

The `sclCollection` attribute of the SCLs collection subresource contains references to one or more SCL resources. The structure of an SCL resource representation is outlined in Figure 14.6. `sclName` represents the specific SCL resource name.

14.3.2 Application Resources

Application resources represent applications (DA, GA or NA). Applications register under one or more SCLs. They interact with the rest of the system via `mIa` (NA) or `dIa` (DA, GA)

<scl>	Attributes including references to sub-resources	Sub-resource attributes
	<p>accessRightID</p> <p>creationTime</p> <p>expirationTime</p> <p>id</p> <p>lastModifiedTime</p> <p>onlineStatus</p> <p>pocs</p> <p>searchStrings</p> <p>Schedule</p> <p>serverCapability</p> <p>discoveryReference</p>	<p>ID of the access rights resource applicable to this SCL</p> <p>Creation time attribute</p> <p>Expiration time attribute</p> <p>Id attribute</p> <p>last modification time attribute</p> <p>ONLINE OFFLINE NOT_REACHABLE</p> <p>Points of Contact (network URI) of this SCL</p> <p>Tokens used as keys for searching resources and contents</p> <p>Schedule that tells when an SCL is available</p> <p>If TRUE the registered SCL can handle incoming requests.</p> <p>Reference to Discovery resource (virtual resource)</p>
	<p>discoveredURIs</p> <p>matchSize</p> <p>statusCode</p> <p>Truncated</p>	<p>List of resource URLs matching searchPrefix and filterCriteria</p> <p>number of resources that matched the filterCriteria</p> <p>If URI list had to be truncated</p>
	<p>notificationChannelReference</p>	<p>Reference to Notification channel resources</p>
	<p>creationTime</p> <p>lastModifiedTime</p>	<p>Creation time of the resource</p> <p>Last modification time</p>
	<p>notificationChannelCollection</p>	<p>List of notification channel resources, see Figure 13</p>
	<p>mngtObjsReference</p>	<p>Reference to Management object collection subresource</p>
	<p>...</p> <p>applicationsReference</p>	<p>(TBD at the time of writing)</p> <p>Reference to Application collection subresource</p>
	<p>(see Figure 5)</p> <p>containersReference</p>	<p>Reference to Containers collection subresource</p>
	<p>(see Figure 5)</p> <p>groupsReference</p>	<p>Reference to Groups collection subresource</p>
	<p>(see Figure 5)</p> <p>accessRightsReference</p>	<p>Reference to AccessRights collection subresource</p>
	<p>(see Figure 5)</p> <p>subscriptionsReference</p>	<p>Reference to Subscriptions collection subresource</p>
	<p>(see Figure 5)</p>	

Figure 14.6 SCL resource representation.

<app>	
accessRightId	Access right id of the parent resource
announceTo	list of resources to which the resource is announced at the moment.
aPoC	Local application point of contact
aPoCPaths	Retargeting application paths and optional accessRight associated with the path. The paths are relative to the path of the application.
applicationStatus	ONLINE OFFLINE NOT_REACHABLE
creationTime	Creation time of the resource
expirationTime	Expiration time. The expiration mechanism is optional. The expiration mechanism is mainly useful when the application is connected over a network access (e.g. CoAP devices).
id	Application id
lastModifiedTime	Last modification time
locRequester	
searchString	Generic search string
<i>Extension(s)</i>	<i>1 or several additional attributes. These extensions are optional and depend on the context.</i>
accessRightsReference	Reference to Access rights collection subresource
	(see Figure 5)
containersReference	Reference to Containers collection subresource
	(see Figure 5)
groupsReference	Reference to Groups collection subresource
	(see Figure 5)
subscriptionsReference	Reference to Subscriptions collection subresource
	(see Figure 5)
notificationChannelsReference	Reference to Notification Channels collection subresource
	(see Figure 5)

Figure 14.7 Detail of a DA/GA/NA representation.

```

GET /m2m/applications/firstApp HTTP/1.1
Host: m2m.example.com

HTTP/1.1 200 OK
Content-Type: application/xml

<?xml version="1.0" encoding="UTF-8"?>
<application xmlns="http://uri.etsi.org/m2m">
  <creationTime>2001-12-31T12:00:00.000</creationTime>
  <searchStrings>
    <searchString>tag1</searchString>
    <searchString>tag2</searchString>
  </searchStrings>
  <containersReference>
    http://m2m.operator.org/m2m/applications/firstApp/
containers
  </containersReference>
</application>

```

Figure 14.8 Example application resource application/xml serialization.

interfaces. Basic interactions may involve reading or writing any resource accessible by the SCL (typically reading or writing container subresources), provided that the application has the relevant access rights to the resource (see Figure 14.9). This notion of generic access right is one of the key features of ETSI M2M, enabling secure sharing of the M2M network, gateways and devices by multiple applications.

Figure 14.8 shows an example serialization of an application resource, using application/xml.

<Access right>	
announceTo	list of resources to which the resource is announced at the moment
creationTime	Creation time of the resource
expirationTime	
lastModifiedTime	Last modification time
searchStrings	Tokens used as keys for searching resources and contents
permissions	Permissions
selfPermissions	Permissions on this resource
subscriptionsReference	Reference to Subscriptions collection subresource
(see Figure 5)	

Figure 14.9 Access right resource representation.

This type of interaction based on reading or writing resources is adapted to many situations, but not to all. For instance, an application may have a transient variable that it wishes to expose to other applications (for instance a timer), but it would clearly be inefficient to report the value of that transient variable to a network buffer. Other applications may offer access to a very large dataset, and the probability of external applications requesting access to some items in the dataset is very low, therefore publishing the whole dataset would be a very inefficient model.

The notion of application point of contact (aPoC) provides a more dynamic interaction model with the application: retargeting. This feature requires that the application register an application point of contact (see “aPoC” attribute in the resource tree). When an application does not provide an application point of contact, the SCL retargeting behavior is not available. In addition, the application may publish a set of application point of contact paths (“aPoCPaths” attribute), in order to restrict the SCL retargeting behavior to only resources addressed by specific subpaths. In the absence of this parameter, the application accepts that any resource request may be retargeted if it has published an application point of contact. Each application point of contact path is a tuple containing:

- A path relative to the URI of the application path resource;
- An optional access right identifier, which applies to all subresources subordinate URIs of under this path. In the absence of such access right identifier, the access rights of the application registration resource apply.

The REST requests of an external application trying to interact with a subresource identified by a URI matching any of the application point of contact paths prefixes is retargeted by the SCL to the aPoC of the application. This enables the application to respond to the request directly without a need for an intermediary buffer.

14.3.3 Access Right Resources

The access right resource (Figure 14.9) is used to configure the permitted actions for a given resource according to the action issuer.

The permissions attribute contains a list of “permission” elements. Each permission element is a triplet that associates:

- a permission id;
- permission flags (READ, WRITE, DELETE, CREATE and DISCOVER);
- permission holders (all, or specific permission holder URIs and specific domains identified by a path prefix).

14.3.4 Container Resources

Container resources provide a generic buffering capability which facilitates the exchange of data between applications, particularly for applications residing on sleeping devices.

<container>	
accessRightId	Access right id of the parent resource
announceTo	list of resources to which the resource is announced at the moment.
containerType	
creationTime	Creation time of the resource
expirationTime	
id	Id of the Container resource
lastModifiedTime	Last modification time
searchStrings	Tokens used as keys for searching resources and contents
maxByteSize	Max byte size of instances
maxNrOfInstances	Max number of instances
maxInstanceAge	Max instance age. Define the max age of each content instance. The age is initialized when the content instance is created. The age is re-initialized each time that the content instance is updated.
<i>Extension(s)</i>	<i>1 or several additional attributes. These extensions are optional and depend of the context.</i>
contentInstancesReference	Reference to Content instances collection subresource
creationTime	Creation time of the resource
currentNrOfInstances	Current number of instances in the resource
currentByteSize	Current size in bytes of data stored in a container resource
lastModifiedtime	Last modification time
latest	Reference to the latest instance
oldest	Reference to the oldest instance
contentInstanceCollection	references to contentInstance subresources
characterEncoding	
Content	Content (opaque to ETSI M2M layer) represented as MIME part
contentSize	Size in bytes of the content instance
contentType	MIME Content type of the content instance
creationTime	Creation time of the resource
delayTolerance	The time before the addition/change of the containing <instance> resource shall be notified to any subscribers
href	
id	
lastModifiedTime	Last modification time
<i>Extension(s)</i>	<i>1 or several additional attributes. These extensions are optional and depend of the context.</i>
subscriptionsReference	Reference to Subscriptions collection subresource
(see Figure 5)	
subcriptionsReference	Reference to Subscriptions collection subresource
(see Figure 5)	

Figure 14.10 Detail of a container representation.

ETSI M2M has structured this generic container into multiple time-indexed contentInstances. The contentInstance metadata identifies the MIME type of the contentInstance content, which is encoded as MIME part. The preferred MIME encoding is MTOM/XOP (application/xop+xml) as specified in http://www.w3.org/TR/xop10/ in order to avoid the overhead of base64 encoding.

This choice of a single index (time) for content instances is somewhat limitative compared to advanced metering protocols that handle multi-index storage structures (and provide read queries based on multiple indexes). We expect this limitation to be removed in future versions of ETSI M2M, in the interim searchstring tags may be used for multi-category indexing of contentInstances.

14.3.5 Group Resources

Group resources (Figure 14.11) make it possible to address a group of resources in a single operation, by addressing the group membersContentReference virtual resource. Currently, the types of resources that can be grouped are application, container, access right, and SCL.

<group>	
accessRightId	
announceTo	list of resources to which the resource is announced at the moment.
creationTime	Creation time of the resource
currentNrOfMembers	Current number of members in a group
expirationTime	
lastModifiedTime	
maxNrOfMembers	
Members	URI list of group members
memberType	APPLICATION CONTAINER ACCESS_RIGHT SERVICE_CAPABILITY_LAYER
searchStrings	Tokens used as keys for searching resources and contents
membersContentReference	Virtual resource that corresponds to requests that would be sent to resource of the corresponding memberType
membersContentResponses	Collection of MemberContentResponse
subscriptionsReference	
	(see Figure 5)

Figure 14.11 Group resource representation.

<subscription>	
contact	URI where the subscriber wants to receive its notifications
creationTime	Creation time
delayTolerance	The slack time allowed to notify a resource change
expirationTime	Expiration time for this subscription
filterCriteria	Optional reference e-tag and lastModifiedTime value of the subscribed resource. A notification should be immediately triggered if the resource is newer.
lastModifiedTime	
minimalTimeBetweenNotifications	Minimal time between notifications in milliseconds.
subscriptionType	ASYNCHRONOUS SYNCHRONOUS

Figure 14.12 Detail of subscription representation.

14.3.6 Subscription and Notification Channel Resources

Subscription resources provide RESTful support for the subscribe/notify primitives that are made available to applications at the functional level.

Clients willing to subscribe to changes of a given resource must add a subscription subresource (Figure 14.12) under the subscriptions collection. The notification can happen asynchronously (`subscriptionType=ASYNCHRONOUS`) in which case the client must specify a contact URI as part of the subscription) or may use the “long-polling model”. The long-polling model is used by applications that are not server capable. Such applications must first request the SCL to create a notificationChannel resource (see Figure 14.13). The SCL will populate the newly created notificationChannel with two parameters: the `contactURI` to be used in subscriptions, and the associated long polling URI in the `channelData` attribute. The nonserver-capable application will then read the long-polling URI, and the SCL will reply only when notifications happen related to the associated `contactURI`.

<notificationChannel>	
channelData	Long-polling URI for LONG_POLLING channelType
channelType	Only 'LONG_POLLING' for now.
contactURI	Contact URI to be used in Subscription resource.
creationTime	Creation time
id	
lastModifiedTime	

Figure 14.13 Detail of notification channel representation.

The notification content includes the subscription URI, as well as the latest representation of the subscribed-to resource.

14.4 ETSI M2M Interactions Overview

An application may interact with resources on the local SCL provided that the accessRights of that resource allow the type of action for the issuer application. An application may also interact with resources located on other SCLs, in which case the local SCL forwards the request to the target SCL, and access rights are checked by the target SCL.

ETSI M2M provides several options for the handling of request responses:

- **Synchronous response:** the expected response content is delivered in the response of the original request, which is blocked until a response becomes available or a timeout error occurs.
- **Semiasynchronous response:** requests include correlation data. If the expected response content cannot be delivered immediately, the target SCL sends an acknowledge indication and an indication of the minimum time to wait before sending a new request. The issuer SCL is expected to reissue the request with the same correlation data, until it obtains the expected response content in the target SCL response.
- **Asynchronous:** the local SCL needs to be server capable, and provides a contact_server URI for responses, as well as correlation data in each request, so that responses received over the server URI can be associated to the corresponding requests.

14.5 Security in the ETSI M2M Framework

14.5.1 Key Management

ETSI M2M uses a key hierarchy:

- The root key K_R is the long-term secure key and is stored in the network and the device. It is used to derive the service keys. Depending on the underlying network technology K_R can be stored at the network level in a secure element (e.g., SIM card) or at the service capability layer. K_R is typically preinstalled and shared between the network and the gateway.
- The service key K_S is derived from the root key during the authentication and authorization of the M2M service layer. K_S is derived during the gateway-registration procedure to the network, and associated to the session (it may expire and be renewed during the session). The service key is not available to applications.

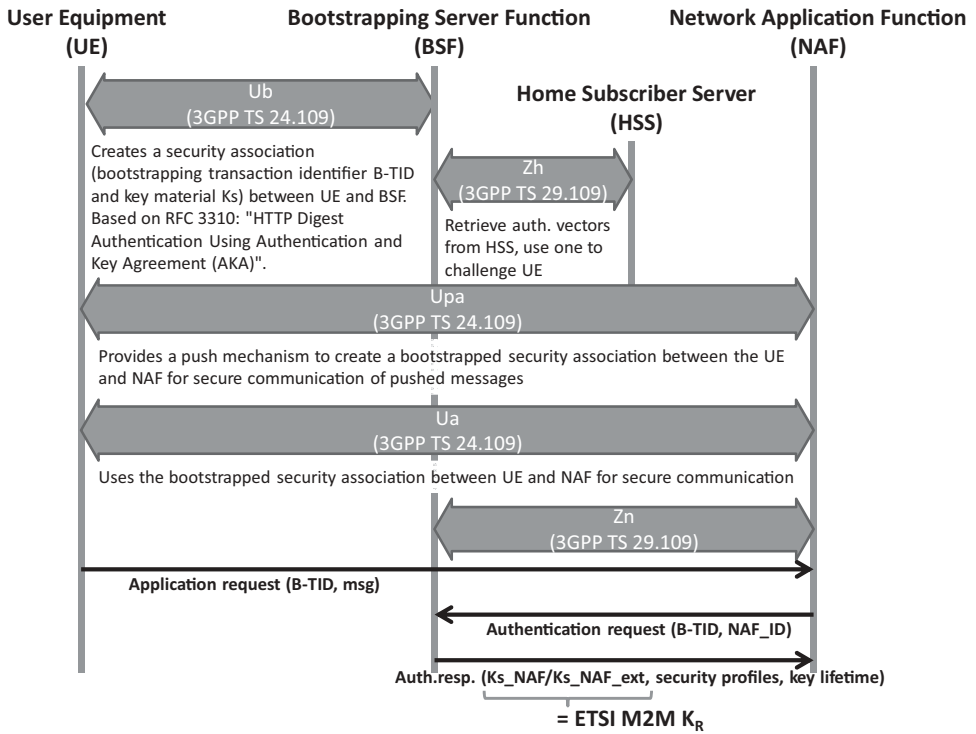


Figure 14.14 Generic authentication architecture (GAA, 3GPP TS 33.220).

- The optional application key K_A . There is one application key per application that wishes to manage its own security context. K_A will be used for authentication and authorization of M2M applications at the M2M device or gateway. At the time of writing, the exact use of K_A was not specified.

TR102921 describes the several security bootstrapping mechanisms. Some options leverage on existing security bootstrapping frameworks at the access-network level, for example, the 3GPP generic bootstrapping architecture (GBA over the 3GPP Ub interface defined in 3GPP TS 24.109, see Figure 14.14) that uses SIM or USIM card security material to derive a shared key between the M2M gateway (user equipment) and the MSBF (acting as 3GPP NAF, via the NSCL).

Other options are completely independent of the underlying transport network, for example, using EAP over PANA transport, with the M2M gateway acting as EAP peer and PANA client, the NSCL as EAP authenticator and PAN authentication agent, and the MSBF acting as EAP authentication server (Figure 14.15).

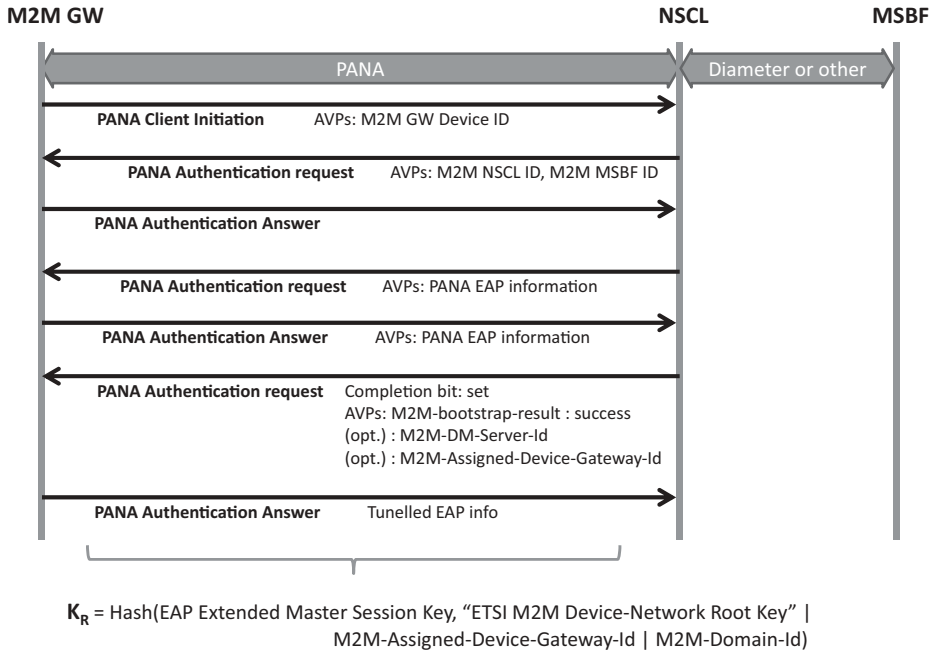


Figure 14.15 PANA/EAP ETSI M2M security bootstrapping.

14.5.2 Access Lists

The ETSI M2M framework is designed to enable usage of a shared M2M infrastructure (network, gateways, sensors and devices) by multiple applications.

The security aspects are very important in this context:

- A given end-user (e.g., a residential subscriber) owning several sensors under control of his local M2M gateway will of course not want to publish sensor data, or provide access to actuators, to anyone. ETSI M2M provides an access-control list (ACL) for most resources published by the M2M gateway SCL. In the example of Figure 14.7, the user, by configuring properly the ETSI M2M gateway SCL ACLs can decide which network application can perform which REST action (e.g., read or write) on which device application, or even on which container of which device application.

Of course, the average end user will never deal with the details of ETSI M2M. Instead, a user-friendly interface will be provided by the M2M service provider. In the example of Figure 14.16, the M2M service provider presents each “network application” (or device application, or group of related NA and DA applications) as an icon, using the now-familiar application store design. In Figure 14.17, the user is configuring a sensor, and is asked which applications can access this sensor information: behind the scenes,

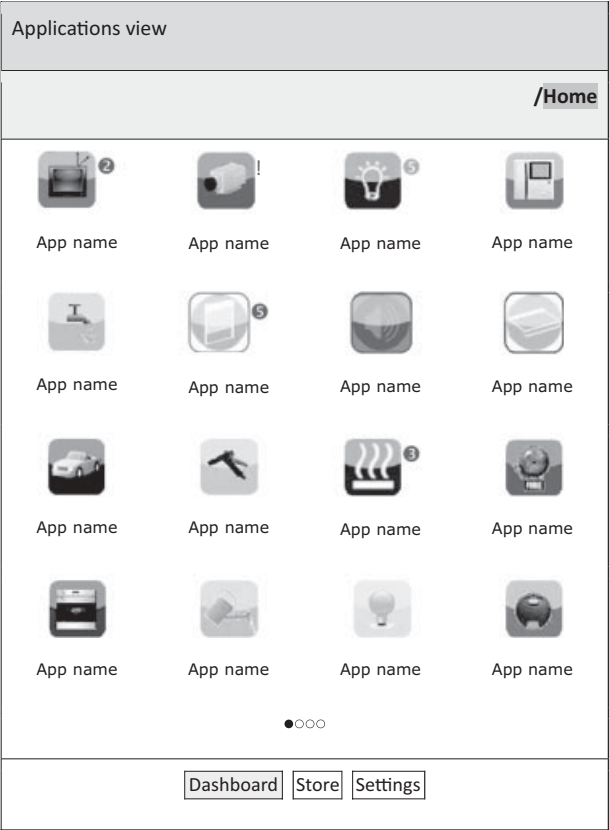


Figure 14.16 A possible user interface for ETSI M2M service configuration.

the middleware will configure the proper ACLs in the representation of that sensor on the gateway SCL.

14.6 Interworking with Machine Area Networks

ETSI M2M provides a very interesting model to interact with multiple machine area networks in a homogeneous way, enabling network or gateway applications to interact with these networks with minimal customization requirements. For most applications, only the specific sensor/actuator URIs and parameter names need to be configured or discovered: the actual underlying protocol details of the machine area networks are abstracted by the ETSI M2M framework.

ETSI TR 102 966 “Interworking with M2M Area Networks” explores the specific mappings recommended for interworking with other M2M technologies, such as ZigBee.

Device List Device Edit

Cancel Ok

Vendor name

Product name

Personal name [_____]

Place Bedroom [Configure](#)

Optional tech element 1 [_____]

Optional tech element 2 [_____]

Optional tech element 3 [_____]

Used by applications

Application name 1 Full access Read Only No link

Application name 2 Full access Read Only No link

Application name 3 Full access Read Only No link

Delete device

Figure 14.17 Security configuration screen for a sensor.

At the time of writing this section, the document was still in early draft stage, however, the author was an active participant in the elaboration of this technical report, and believes that the approach outlined below will be very close to the final recommendation.

14.6.1 Mapping M2M Networks to ETSI M2M Resources

There are many potential ways of mapping M2M networks to ETSI M2M standard resources. One possibility is to attempt to recreate the hierarchy of native M2M networks within the hierarchy of ETSI M2M resources (e.g., SCLs for top-level networks, applications for nodes, etc. . . .). However, obviously, it is impossible to replicate the exact

resource hierarchy of any network, which can get quite deep. In the case of ZigBee for instance (refer to Chapter 7), each controller may access a collection of networks, each with a number of devices hosting a number of applications. Each application publishes an interface composed of a number of clusters, and each cluster is composed of multiple attributes and primitives!

The preferred option is to represent external machine area networks as a collection of applications (GA or DA). Any resource hierarchy present in the native M2M network is recreated by means of pointer attributes: the ETSI M2M application representing the parent object will expose special attributes pointing to child objects.

Recognizing that most automation protocols use a similar datamodel structure, TR 102 966 introduces specific ‘tags’ (special searchString values) to identify the type of ‘object’ modeled by the application in a protocol independent way:

- ETSI.ObjectType tags (for instance ETSI.ObjectType/ETSI.AN_NWK) are used to discriminate between applications representing Interworking Proxy objects (ETSI.ObjectType/ETSI.IP), Network objects (ETSI.ObjectType/ETSI.AN_NWK), Device objects (ETSI.ObjectType/ETSI.AN_DEV), Application objects (ETSI.ObjectType/ETSI.AN_APP), and Point objects (ETSI.ObjectType/ETSI.AN_POINT). The various object types will be illustrated in the example of ZigBee (see Section 14.6.2).
- ETSI.ObjectSemantic tag (for instance ETSI.ObjectSemantic /OASIS.OBIX_1_1) is used to discover the semantic conventions supported by the object. The syntax of an object representation is usually indicated by its Content-Type, for instance application/xml. However, multiple semantic conventions may leverage the same syntactic rules. In the use case of interworking with control and sensor networks, examples of such semantic conventions leveraging application/xml syntax are OASIS oBix (ETSI.ObjectSemantic/OASIS.OBIX_1_1), ZigBee Gateway Device REST binding, or ASHRAE BACnet annex am (ETSI.ObjectSemantic/ASHRAE.CSML_1_0).

ETSI M2M implements a REST design model that allows multiple representations of the objects manipulated through the ETSI M2M SCL.

- ETSI.ApplicationProfile tags are reserved for future use. The intent is to be able to facilitate search of specific devices for example, “lamps”. Nomenclatures have been created by ZigBee, KNX and LonWorks, and one is being worked on by BACnet. Future work would lead to a harmonized nomenclature that would use this tag category

14.6.2 Interworking with ZigBee 1.0

Interworking with ZigBee 1.0 networks leverages the ZigBee Alliance “Gateway specification for network devices”, version 1.0 (refer to Chapter 7 for details). This specification exposes how a ZigBee gateway devices (ZGD) exposes the ZigBee network resources to IP host applications (IPHA), over several network bindings.

An ETSI M2M application (local gateway application or device application in the case of a separate ZigBee interworking device) models the ZGD and acts as an

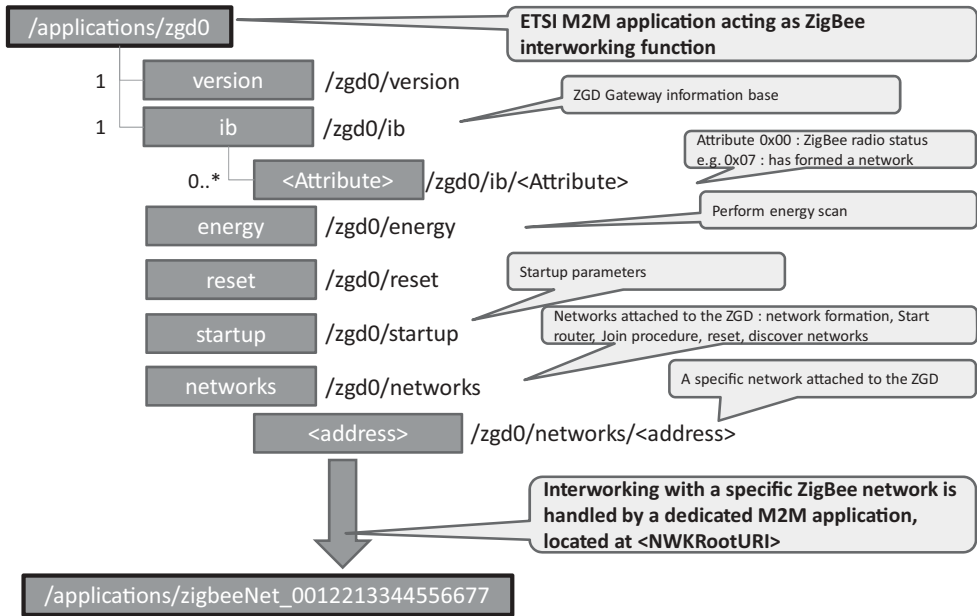


Figure 14.18 ZGD REST representation as a standard ETSI M2M application.

interworking proxy. It has a searchString attribute containing technology-independent tag “ETSI.ObjectType/ETSI.IP” so that applications may discover this application, which has interworking proxy capability. This application contains one <container> subresource that has a searchString containing tag “ETSI.ObjectType/ETSI.IP”, and a tag of category ETSI.ObjectSemantic (oBix for our examples below). The latest contentInstance of this container lists the networks supported by this interworking proxy (in a protocol-independent way) as well as proxy capability information. Each network element points to the ETSI M2M application that represents it. This container instance may also include links to the ZGD representation defined by the ZigBee alliance (Figure 14.18). An example interworking proxy representation is shown in Figure 14.19.

Each ZigBee network is represented by a specific ETSI M2M application located at the URI specified in the ‘networks’ list. Applications representing a network have a searchString containing tag: ETSI.ObjectType/ETSI.AN_NWK, and one container with the same tag. The latest contentInstance of this container contains the representation of the network. Figure 14.20 illustrates how a client application can read this resource.

The representation of the network contains the network identifier and the list of nodes (in a protocol independent way), as well as technology-specific information related to the network. Each ZigBee node is represented by a specific ETSI M2M application located at the URI specified in the ‘nodes’ list. Applications representing a node have a searchString containing tag: ETSI.ObjectType/ETSI.AN_DEV, and one container with the same tag.

```

>>> HTTP GET
/gsc/applications/zgd0/containers/descriptor/contentInstances/
last/content

<<< 200 OK
<obj>
  <str name="interworkingProxyID" val="Text for correla-
tion purpose"/>
  <list name="supportedTechnologies">
    <obj>
      <enum name="anStandard" val="ZigBee_1_0"/>
      <enum name="anProfile" val="ZigBee_HA"/>
      <enum name="anPhysical"
val="IEEE_802_15_4_2003_2_4GHz"/>
    </obj>
  </list>

  <list name="networks"/>
    <ref href="/gsc/applications/zbnw0/">
  </list>
</obj>

```

Figure 14.19 Retrieving the M2M application resource modeling the ZGD via HTTP GET.

```

>>> HTTP GET
/gsc/applications/zbnw0/containers/descriptor/contentInstances/
last/content

<<< 200 OK
<obj>
  <str name="networkID" val="Text for correlation
purpose"/>
  <str name="extendedPanID" val="0x685B3C34"/>

  <list name="nodes">
    <ref href="/gsc/applications/zbnode0/">
  </list>
</obj>

```

Figure 14.20 Retrieving the M2M application resource modeling the ZigBee network via HTTP GET.

```

>>> HTTP GET
/gsc/applications/zbnode0/containers/descriptor/contentInstances/
last/content

<<< 200 OK
<obj>
  <str name="nodeID" val="Text for correlation purpose"/>
  <str name="ieeeAddress" val="0x685B3C88"/>
  <enum name="type" val="endDevice"/>

  <list name="applications">
    <ref href="/gsc/applications/zbapp0"/>
  </list>
</obj>

```

Figure 14.21 Retrieving the M2M resource for a node via HTTP GET.

The latest contentInstance of this container contains the representation of the node. Figure 14.21 illustrates how a client application can read this resource.

Our example network has a single node. Each node may host multiple applications, for instance multiple switches or lamp controllers. ZigBee identifies each application as an ‘endpoint’. Each ZigBee endpoint is represented by a specific ETSI M2M application located at the URI specified in the node ‘applications’ list. ETSI M2M Applications representing a node application¹ have a searchString containing tag: ETSI.ObjectType/ETSI.AN_APP, and one container with the same tag. The latest contentInstance of this container contains the representation of the node application, in this case a ZigBee endpoint. Figure 14.22 gives an example representation of a node application resource.

Each node application may implement multiple interfaces. ZigBee identifies each interface as a ‘cluster’. Interfaces may be represented as separate applications (the element is a reference to the M2M application), but in the example of Figure 14.22 the contentInstance data includes the actual interface content. Each interface contains protocol dependent primitives (e.g. ‘toggle’) as well as a number of attributes (e.g. ‘OnOff’). TR 102 966 offers an abstraction for attributes representing measurements: the Point that may include protocol-independent specification of units, range, etc. (It is not used in the example).

When the ZigBee interworking application starts, it audits all networks and replicates a view of the ZigBee network topology in the GSC (Figure 14.23). This view includes all ZigBee nodes and all ZigBee applications (endpoints). When a new ZigBee node is

¹Node applications are applications running in sensors, not to be confused with the M2M application resources that model them.


```

...
<int name="endpoint" val="1"/>
<int name="applicationProfileID" val="0x0104"/>
<int name="applicationDeviceID" val="0x0100"/>

<list name="Interfaces">
  <obj>
    <str name="clusterID" val="0x0006"/>
    <enum name="clusterType" val="input"/>

    <list name="attributes">
      <ref name="0x0000"

      href="/<sclBase>/applications/<networkX_nodeY_application
Z>/containers/0x0006_OnOff"/>
    </list>

    <list name="operations">
      <op name="0x00"
href="/<sclBase>/applications/<interworking_proxy_unit>/0x0006_
off"/>
      <op name="0x01"
href="/<sclBase>/applications/<interworking_proxy_unit>/0x0006_
on"/>
      <op name="0x02"
href="/<sclBase>/applications/<interworking_proxy_unit>/0x0006_
toggle"/>
    </list>
  </obj>
</list>
...

```

Figure 14.22 Extract of the ZigBee endpoint representation.

added in the network, the ZigBee HAN representation is immediately updated to reflect this change in the GSC.

At this stage, other gateway or network applications can discover the elements of the ZigBee networks by retrieving the resources maintained in the GSC by the driver. They can also subscribe to the representation changes, if the driver supports it (and creates the appropriate subscriptions collections in the ZigBee network representation).

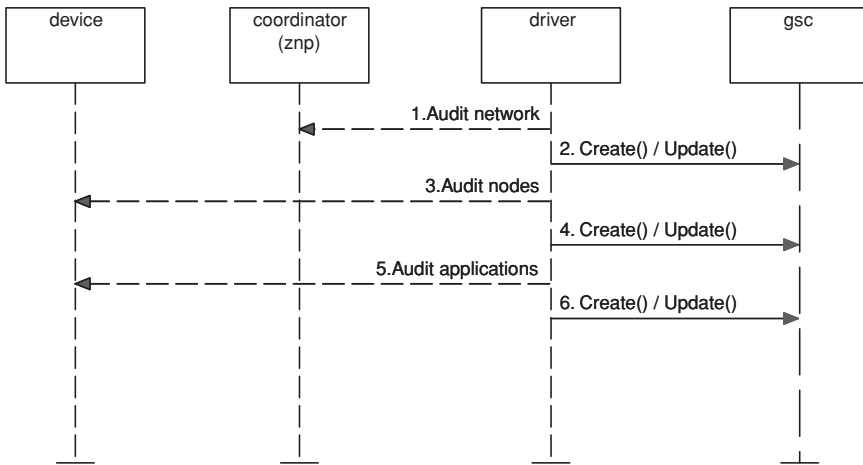


Figure 14.23 ZigBee driver explores the ZigBee network and creates the SCL representations.

14.6.3 Interworking with C.12*

In order to prove the versatility of ETSI M2M standards and their ability to provide a common framework for interfaces to existing standards, we explore in this section the interface to C.12 metering standards. As mentioned in Chapter 10, the communication protocol for the transport of C12.19 tables requires only basic read and write services. Such standardization of communication patterns using few verbs is typical of REST design principles.

As ETSI M2M decided to use a RESTful approach for specifying its interface between M2M applications and the service capability layer, we attempt to use this interface to interact with C12.19 tables accessible over a C12 network. At first sight, a REST-style interface can be achieved by mapping each C12.19 table to a corresponding REST resource. The basic read and write services can also easily be mapped on the classical CRUD verbs (create, retrieve, update, delete).

In the DLMS/COSEM case (see Chapter 11), the data communication services accessible to the metering application are a little more complicated than read/write due to the existence of the “action” service. An intermediate resource has to be used in order to communicate the name of the “action” and the associated parameters in a REST style. In the C12.19 data model it is easier since the possible “actions” are already included in two specific tables (Tables #07 and #08). The “action” is called “procedure” and a request for executing a specific procedure corresponds to the invoke of the basic write service for Table #07.

Our purpose here is not to propose a specification of what would interface to native RESTful C12 meters, but instead we attempt to define an interworking function between

*This section is contributed by Jean-Marc Ballot.

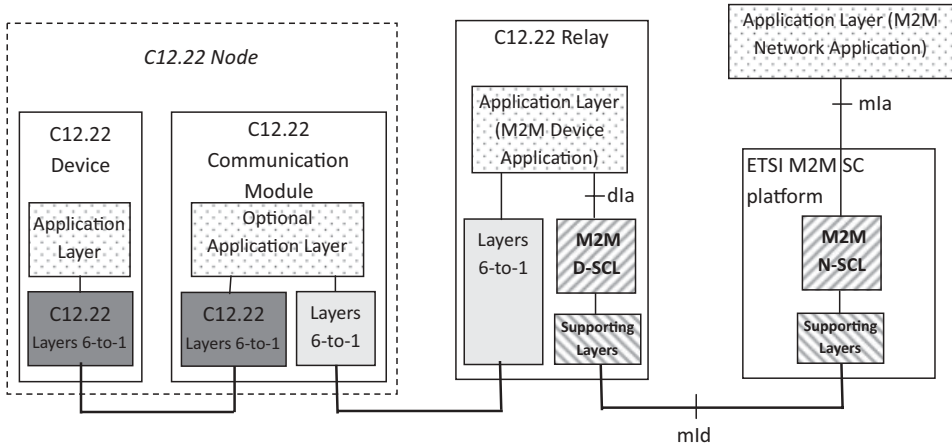


Figure 14.24 Interworking function implemented as C12.22 relay.

a standard C12 network and ETSI TC M2M, that is, exposing an ETSI M2M application interface. As an example, implementation for such interworking, we consider a C12.22 relay, as represented in Figure 14.24.

In Figure 14.24, mla, dla and mld are standardized ETSI M2M reference points. N-SCL and D-SCL are the ETSI M2M “network service capability layer” and “device service capability layer”.

On the left side of the C12.22 relay, the communications are fully compliant with the ANSI C12 suite of standards. On the right side of the C12.22 relay, the dla, mld and mla reference points support the RESTful communication patterns.

From the network application perspective, the C12.22 node can be represented as part of the C12.22 relay “device application” representation, or each node could be represented as a separate device application. During initialization steps, the device application in the C12.22 relay can read the C12.19 table #00 (called GEN_CONFIG_TBL) of the C12.22 node in order to get the full list of the supported tables and procedures. Then, the C12.22 relay is able to use the RESTful dla interface in order to request the creation of a set of REST resources that represent the set of C12.19 tables handled by the C12.22 node. A one-to-one mapping is performed between C12.19 tables and REST resources. The resources are physically created in the N-SCL (they could also be created locally in the C12.22 relay if the relay has REST server capability).

The interactions with these resources can be performed by using the CRUD verbs enhanced with notify and subscribe as specified by ETSI TC M2M. When a network application wants to read (or write) a C12.19 table contained in the C12.22 node, it uses a retrieve (or an update) query to the corresponding REST resource. The C12.22 interworking relay then contacts the relevant C12.22 node in order to read/write the actual content of the relevant C12.19 table. The C12.22 interworking relay may represent static

C12 node parameters as container data. Read/writes to dynamic parameters will preferably use the ETSI M2M retargeting mechanism already described above:

- The REST resources that have to be directly requested from the real legacy C12.22 node are not physically stored in the SCL, instead the SCL mirrored information will only contain a pointer to these resources, and specify that retargeting is supported. The retargeted queries can be sent to the application point of contact, which must be published by the device application (the interworking C12.22 relay) during the initialization phase.
- When a request is sent by a network application that targets a resource not stored in the SCL, the SCL retargets the request to the device application by forwarding it to the application point of contact. The device application can then read (or write) the real content by communicating with the C12.22 node.

This redirection mechanism is not only useful in the case of interworking but also make it possible to implement natively ETSI M2M compliant C12 meters. Such a C12 meter would implement a RESTful dIa interface, and would be able to mirror static C12.19 REST resources in the SCL. For C12.19 tables that are volatile or too frequently changing (e.g., time elapsed since a specific event, time remaining before a specific event, consumption metering counters), the redirection mechanism is used instead and the mirrored representation would store only pointers.

14.6.4 Interworking with DLMS/COSEM

ETSI TC M2M decided to use a RESTful approach for the interface between M2M applications and the service capability layer. The interworking with DLMS/COSEM meters is not trivial because DLMS/COSEM uses an object oriented approach that is not RESTful compliant.

DLMS/COSEM meters are “legacy devices” according to the ETSI M2M reference architecture document and interworking will be performed by an interworking proxy (NIP or GIP), located in a M2M gateway (GIP) or in the core network (NIP). At the time of writing, ETSI TC M2M had not yet fully specified the interworking proxy function. In the following text, we are proposing a possible approach in line with the current specification, based on a network-based interworking proxy (NIP), and illustrated on Figure 14.25.

We are assuming that the network application interacting with the DLMS meter is ETSI M2M compliant. The NIP could be implemented as a new interface on DLMS data concentrators. This would be an option to standardize the interactions between network applications and data concentrators (currently out of scope of DLMS/COSEM).

The network interworking proxy (NIP) is considered as a part of the ETSI M2M platform. It contains a COSEM client AP and a COSEM client AL. The communications between the client parts in the NIP and the server parts in the metering equipment use the DSML/COSEM specifications. Once the client AP has established an association with the server AP in the meter, a set of the COSEM interface objects becomes accessible.

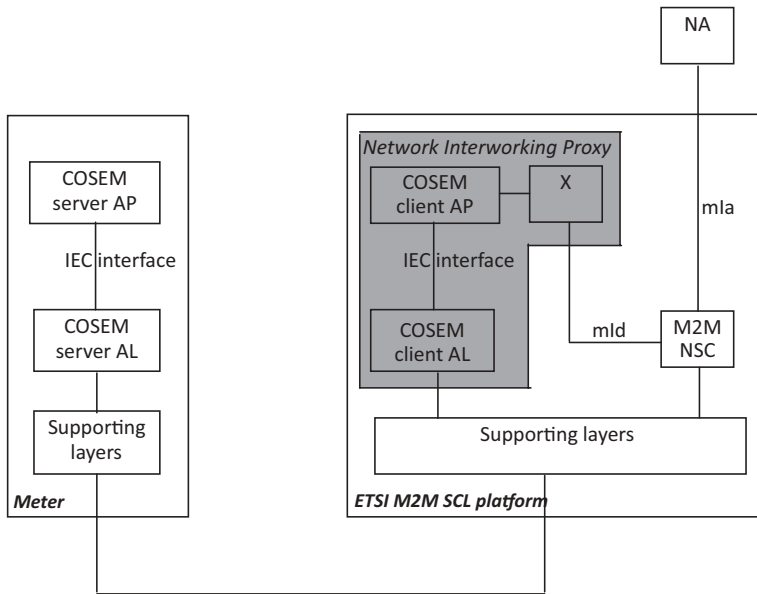


Figure 14.25 ETSI M2M – DLMS/COSEM interworking using a network interworking proxy (NIP).

The “X” part of the NIP in Figure 14.25 represents the new software added for ETSI M2M interworking.

From a RESTful point of view, each COSEM interface object can be represented as a resource: the NIP first reads each COSEM interface object by using the standardized COSEM GET service. Then, after reading the object, the corresponding resource is created in the SCL layer. Only static parameters should be stored in the SCL. Volatile parameters such as counter values should be listed in SCL resources only as indirections, that is, read requests from network applications will be retargeted by the SCL to the interworking proxy.

The DLMS specifications also define an ACTION service enabling a client AP to remotely invoke one or more methods of COSEM interface objects. This interaction pattern can also leverage the ETSI M2M retargeting mechanisms: SCL representations of COSEM interface objects will list a subresource for each supported action. The NA may invoke an action by sending a POST request targeted to the corresponding subresource (and containing the action parameters). The write request will be retargeted to the interworking proxy by the SCL, and the interworking proxy will convert the request to a native DLMS/COSEM action service invocation. Synchronous responses will be provided in the 200 OK response.²

² In the case of asynchronous responses, the NIP response may just be 201 created, and the asynchronous response may be sent later to a resource supplied a part of the invoke parameters.

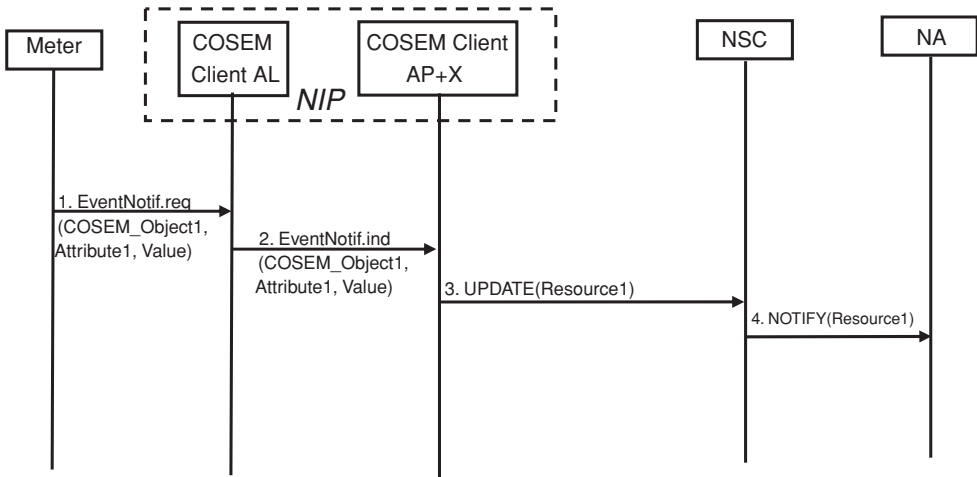


Figure 14.26 DLMS event-notification interworking with ETSI M2M.

The meter may also send an EventNotification, this case represents the exception of the DLMS client–server model. This notification contains the value of a COSEM interface object attribute. This interworking case is illustrated in Figure 14.26.

In step 1, the meter needs to inform the client AP of the value of one attribute (Attribute1) of one COSEM object (Object1). The meter uses the COSEM EventNotification service in order to provide the client AP with the attribute value (step 2).

The NIP has to update Resource1 that corresponds to Object1 in the REST environment. As a single attribute of Object1 is modified, a partial update is performed.³ The network application will be notified of this update if it subscribed to Resource1.

14.7 Conclusion on ETSI M2M

Just like networked information systems played a fundamental role in the transformation of almost every business, connected objects will fundamentally change the design of most industrial and automation processes (see for instance draft-ietf-6LoWPAN-usecases). The Internet emerged as the information backbone interconnecting all information systems, and the Internet of Things is now emerging as the backbone interconnecting all objects.

A common view is to consider that the IoT will be just one giant peer to peer IPv6 network (there are about 6×10^{27} IPv6 addresses per square meter on earth). But there are several problems with this view:

³ However, if partial updates are not allowed, due to some internal policy in the REST server, the NIP will have to retrieve Resource1, locally update its content with the new value of the Attribute1, then perform a full update of Resource1.

- For most actual sensor networks using low-bitrate and lossy networks (LLNs), standard IPv6 is not usable: 6LoWPAN will be used instead. The usable addressing space, notably for destination addresses, is very limited and not optimized for any to any worldwide communication. 6LoWPAN is optimized for a local or regional sensor network.
- Basic security considerations force the introduction of application-level gateways between private domains and the public side of the internetwork. Today, the much-touted exhaustion of IPv4 addresses does not have the screaming halt effect one might expect on the development of the Internet: this is because all corporations use private addresses internally, and their firewalls map on demand these private addresses to very few external public IP addresses using NAT (network address and port translation) or application-level proxies, e.g., for HTTP). The same will probably go for the Internet of Things, private 6LoWPAN clouds will be interconnected by application-level gateways.

ETSI M2M is currently the only candidate standard for such “inter-networks of things” application-level gateways. It provides a standard data model for machine area networks, a standard access control list format, and a global application-level addressing mechanism for connected objects. This design ensures that any application on earth can access any sensor, but not using the sensor IPv6 address directly (it may be a private address), using instead the application level URI for the sensor that can be used to route the application requests to the proper object network access gateway, under control of its ACLs. Such gateways are also the ideal place for fieldbus protocol interworking, based on the normalized REST representation of each fieldbus network.

We believe the ETSI M2M framework will become the necessary complement of 6LoWPAN for service providers envisioning very large scale, multipurpose M2M infrastructure deployments.

Part Five

Key Applications of the Internet of Things

15

The Smart Grid

15.1 Introduction

The smart grid will be one of the most important applications of the Internet of Things.

A major paradigm change is happening in electricity markets, driven by the convergence of several factors:

- The challenges posed by the accelerated introduction of renewable-energy sources in the overall electricity production, which brings an increasing degree of randomness to the traditionally deterministic supply side.
- The ubiquitous penetration of the Internet in homes and businesses, and the increased confidence in next-generation smart distributed networks for mission-critical applications (after years of experience of the successful migration of telephony networks to VoIP).
- The gradual opening of electricity markets, with new regulation opening production facilities and distribution networks to all actors, greater fluidity in electricity trade markets, and fast maturing of the regulatory framework for active utility operators, such as those implementing demand response.
- The increasing volatility of electricity prices, resulting from the underlying volatility of oil and natural gas, but also increasingly from the propagation of external shocks, such as exceptional climatic events, through energy exchanges.

The current credo of electricity operators “demand is unpredictable, and our expertise is to adapt production to demand”, is about to be reversed into “production is unpredictable, and our expertise is to adapt demand to production”.

As the rules of the game change, the key assets of an energy operator will no longer be the means of production, but the next-generation communication network and information system, which they still need to build entirely. M2M communications and emerging standards such as 6LoWPAN, RPL and ETSI M2M will be key enablers for this evolution.

15.2 The Marginal Cost of Electricity: Base and Peak Production

As for any other production type, the cost of electricity generated by a power plant is the sum of the cost of the primary energy supply and the amortization of the plant itself. For this reason, electricity produced by plants running continuously (amortized over the full year) is always cheaper than electricity produced by plants running only sporadically. Nuclear power plants, which produce (relatively) cheap electricity, typically operate continuously except during planned maintenance periods and their production can be adjusted only marginally, over long periods of time (48 h or more).

As a result, demand prediction is key for all electricity operators, as it allows them to plan their production investments. Demand can be decomposed into:

- “Base demand”: this is the component of demand that varies most slowly and can be produced from plants running continuously close to their maximum capacity. Electricity production for the base demand has the cheapest marginal production cost.
- Variable demand, which can be supplied from plants operating at a lower utilization rate, or purchased on the market. It has a higher marginal production cost. The “daily peak demand” represents the levels of demand reached only a few hours per day, typically at 7 p.m. in the evening. The yearly peak demand represents the levels reached only a few hours per year, and consequently at the highest marginal cost.

For all electricity operators the marginal production cost gets higher and higher as the current production level increases beyond the “base demand” (Figure 15.1).

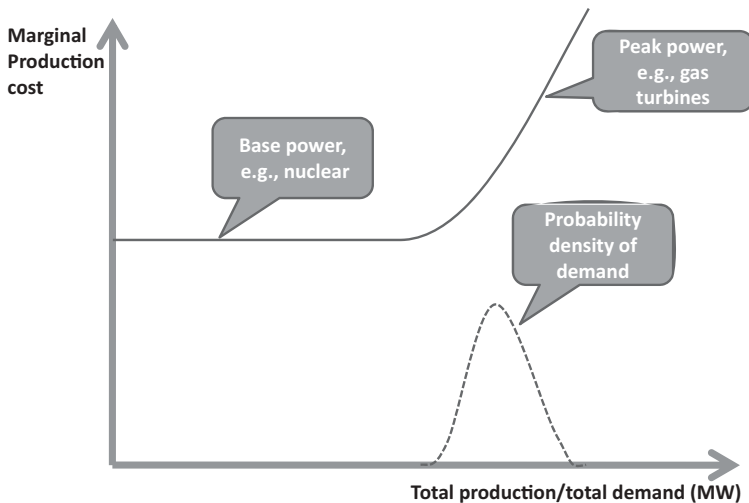


Figure 15.1 The marginal cost of power as a function of current demand/production level.

At some point, the cost of generating each additional MWh gets so high that it may exceed the final, usually fixed or slowly varying, selling price of the MWh to the end consumer. At this point, demand response becomes a no-brainer: instead of producing more power and losing money, the operator should create an incentive for its customers to use less energy.

15.3 Managing Demand: The Next Challenge of Electricity Operators . . . and Why M2M Will Become a Key Technology

Today, most operators model their residential and small-business customers statistically, using a generic demand profile that depends only on the average yearly consumption of the customer and his location. Operators only measure the total consumption of each customer, not its distribution over time. As a result, an operator's ability to reward their customers that have a lower "peak demand" component in their consumption is limited to simple time-based tariffs (typically a day and a night tariff for counters able to measure day and night consumption separately).

With more advanced counters able to report consumption profiles over time (such as those deployed in first-generation automatic metering projects), operators can introduce time-dependent tariffs that give customers an incentive to minimize their peak consumption. If such tariffs are published and are reasonably stable, users can schedule their consumption (e.g., use delayed start function of washing machines or dryers) and therefore "flatten" their consumption pattern and reduce their average electricity bill.

But unfortunately, published time-dependent tariffs are not sufficient: the increasing proportion of production coming from renewable sources creates random, unpredictable changes in total production power. The randomness of production power requires real-time adaptation of demand: tariffs would need to become dynamic, that is, change in real time, or if tariffs remain published, separate real-time incentives to adapt demand must be created.

Obviously, most residential customers or small businesses do not have the time or expertise required to control their heating, air conditioning, hot water tank, fridge, or ventilation system in real time. This is why commercial electricity suppliers will also need to actively manage the regulation of energy use in each home, using bidirectional M2M technology. "Intelligent meters" will be useless if not complemented by active energy management systems. Such bidirectional M2M energy-management systems will flatten the demand profile, and will also be able to react in real-time to network or production incidents, minimizing the costs associated to last-minute energy purchases.

In the future, the ability to provide such efficient demand response programs will become a key differentiator for electricity operators: the operators who will manage to best "flatten" the consumption of their customers, while preserving or enhancing comfort, will reduce their production costs and become more competitive.

15.4 Demand Response for Transmission System Operators (TSO)

15.4.1 Grid-Balancing Authorities: The TSOs

While multiple commercial electricity suppliers may exist in a country, they all share the same distribution grid. Because the network is interconnected, any commercial supplier can buy additional capacity from independent producers or other utilities and easily distribute this power to its own customers.

But what happens if a supplier does not generate/buy enough power for its customers? As the grid is fully interconnected, these customers will draw power from all other suppliers, causing instability for all customers. Clearly, some independent authority must be able to monitor production and demand for all operators, and ensure the proper balancing of production and demand.

Most countries have set up such an independent authority, the transmission system operator (TSO, see Figure 15.3 for a list of some European TSOs), which monitors in real time the production level of all operators connected to the grid, and the aggregate demand of all customers (see Figure 15.2, IESO forecasting demand in Ontario every five minutes).

As each operator is supposed to make sure its power production balances the demand of its customers through production planning and demand response, in an ideal world the actual aggregate production level should always match the actual aggregate demand. In reality, power consumption depends a lot on the weather (temperature, sun exposure), which is not completely predictable. In addition, some operators will sometimes have an issue that makes it very hard for them to meet the required production level: it might be an outage, a maintenance operation affecting some of their production facilities, or any unusual variation of actual demand compared to the projected demand. For all these reasons, actual demand will always differ slightly from the actual level of production.

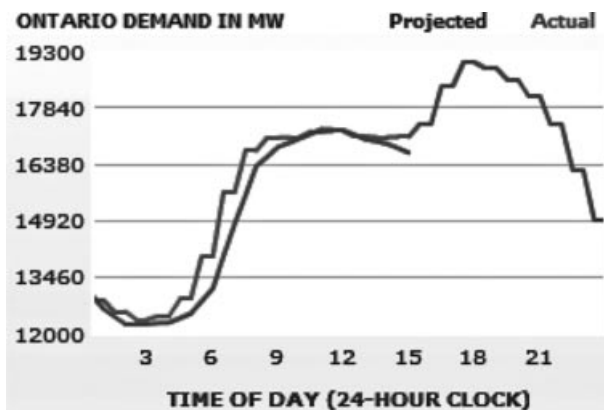


Figure 15.2 Demand projection by Ontario's IESO authority (independent electricity system operator), and actual demand.

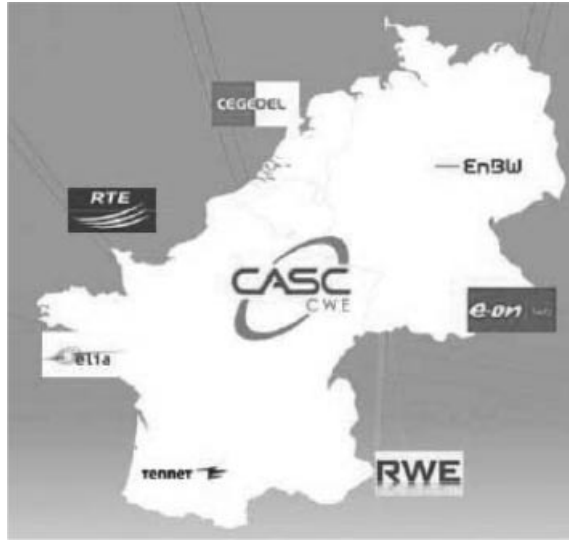


Figure 15.3 Some European TSOs and CWE, a cross-border capacity allocation agency.

The TSO in charge of balancing the network monitors in real time the production level of all power plants (and energy imports), as well as power consumption (using measurement systems along the main distribution lines). It also monitors the network frequency, which is the best instantaneous indicator of the production/demand imbalance: if the frequency is lower than the reference frequency (50 Hz in Europe), consumption exceeds supply, *vice versa* when the frequency is too high, supply exceeds demand.

The TSO has several tools to balance the network:

- Impose automatic production compensation algorithms to all major production plants: if the measured grid frequency is below target, the power plant will increase its power injection, and *vice versa*.
- If demand exceeds supply, the TSO can ask producers to increase their production (either as part of their contract, or by purchasing this energy), import energy from neighboring countries, or reduce demand (activation of power-shedding bids).
- If supply exceeds demand, the TSO can export energy, or sell excess energy to the best bidders. It can also ask producers to decrease their production.

The TSO expenses associated to this continuous balancing activity are funded by the electricity operators, usually *pro-rata* of the actual unbalance between their power production (own or acquired), and the power consumption of their customers.

The production capacity mobilized at the last minute by TSOs is often very expensive, and can even get prohibitively expensive in the case of major unplanned outages. In addition, this peak power production is also extremely inefficient in terms of CO₂ generation.

Again, asking users to shed power demand is an attractive alternative, both financially and from an environmental perspective.

More and more TSOs therefore complement their auctions for additional power by symmetrical auctions for power shedding: they receive bids to reduce power demand, and select all competitive offers instead of activating or sourcing additional power production.

15.4.2 *Power Shedding: Who Pays What?*

The allocation of TSO expenses (resulting from their grid-balancing activities) to each commercial operator is not a simple accounting formality. The TSOs have a lot of margin to implement their own policy. For instance, the price at which TSOs buy excess energy from operators in a positive imbalance situation (in their perimeter) during a grid power-deficit period is a pure matter of policy: TSOs willing to discourage any form of random imbalance should buy at very low prices, whereas TSOs considering that the operator contributed to the network stability (albeit by coincidence) may want to buy at a higher price.

The financial aspects of power shedding are even more complex:

- Companies specialized in demand management may be distinct from commercial suppliers, and are really a new type of player.
- The best policy from a financial fairness point of view may not be the optimal proposal from an environmental point of view
- The best power-reduction proposals may come from customers of other operators than the operators that caused the network unbalance. For each MW not consumed by these customers participating in the demand response scheme, their electricity supplier will not be charging its retail price R , and at the same time is still producing the electricity it had (correctly) planned its users would demand if the demand-response mechanism had not unexpectedly changed this demand profile. Therefore, the operator of customers participating in the demand-response program is still spending the then current marginal cost of power production M , but no longer charges R .

Obviously, the commercial suppliers of customers participating in the demand-response program will want to be indemnified for the additional production that contributed to the grid balance but was not charged to the customers.

TSOs and regulators need to decide whether there should be an indemnification, the level of indemnification, and who pays what. Obviously the arguments of each player will reflect their own interest:

- We will show below that the overall turnover of commercial suppliers is largely unaffected by demand response (the lower consumption is temporary and usually compensated by higher consumption later on). However, we will also see that these operators do produce more energy than what would be otherwise required, and their

costs increase. Commercial suppliers will claim that their business is dependent on the accuracy of generic demand profiles, and since the additional production costs are caused by demand-response operators, these operators should compensate the additional costs.

- Some demand-response companies will claim that they have no relationship with the commercial suppliers; they are only dealing with the final customers who delegated their energy management to them, and therefore have no reason to compensate the additional costs of the operator.

In fact, these choices are more political choices than network management matters. The rapidity of the migration of energy networks towards “smart grids”, and the resulting environmental impact will largely depend on the choices made.

We will study in the next section the principles and business case of demand response on a real network (the French grid managed by TSO RTE), in order to highlight the critical role of standards such as ETSI M2M in the future of such smart-grid applications.

15.4.3 Automated Demand Response

Demand response can apply to any process that can dynamically modify its consumption (or production) level. Many industries, for example, fresh water distribution (Figure 15.4), are ideal candidates for demand response because they can adjust their power consumption very quickly, and have important storage capacity.

Such industrial facilities typically use M2M technologies to retrieve real-time state information (in the example of Figure 15.4: reservoir level and pump power), which is used as an input for an optimization algorithm that manages demand-response participation. M2M is also used to transmit commands (e.g., starting or stopping a pump) during execution of the demand-response program.

In the following sections, we will study one particular way of participating in demand-response programs, by controlling the setpoint of heating/cooling systems in individual homes.

15.5 Case Study: RTE in France

15.5.1 The Public-Network Stabilization and Balancing Mechanisms in France

The French national consumption represents about 490 TWh (2008). The French power system has an installed capacity of over 118 GW, and typically provides 80 GW to over 90 GW at peak loads. Demand load depends a lot on the weather: in autumn, winter or spring, a temperature drop of 1 °C results in additional loads up to 2100 MW on the French network (data from RTE, 2009), while in summer, when the temperature exceeds 25 °C, a rise of 1 °C may bring about an extra load of up to 600 MW (data from RTE, 2004).

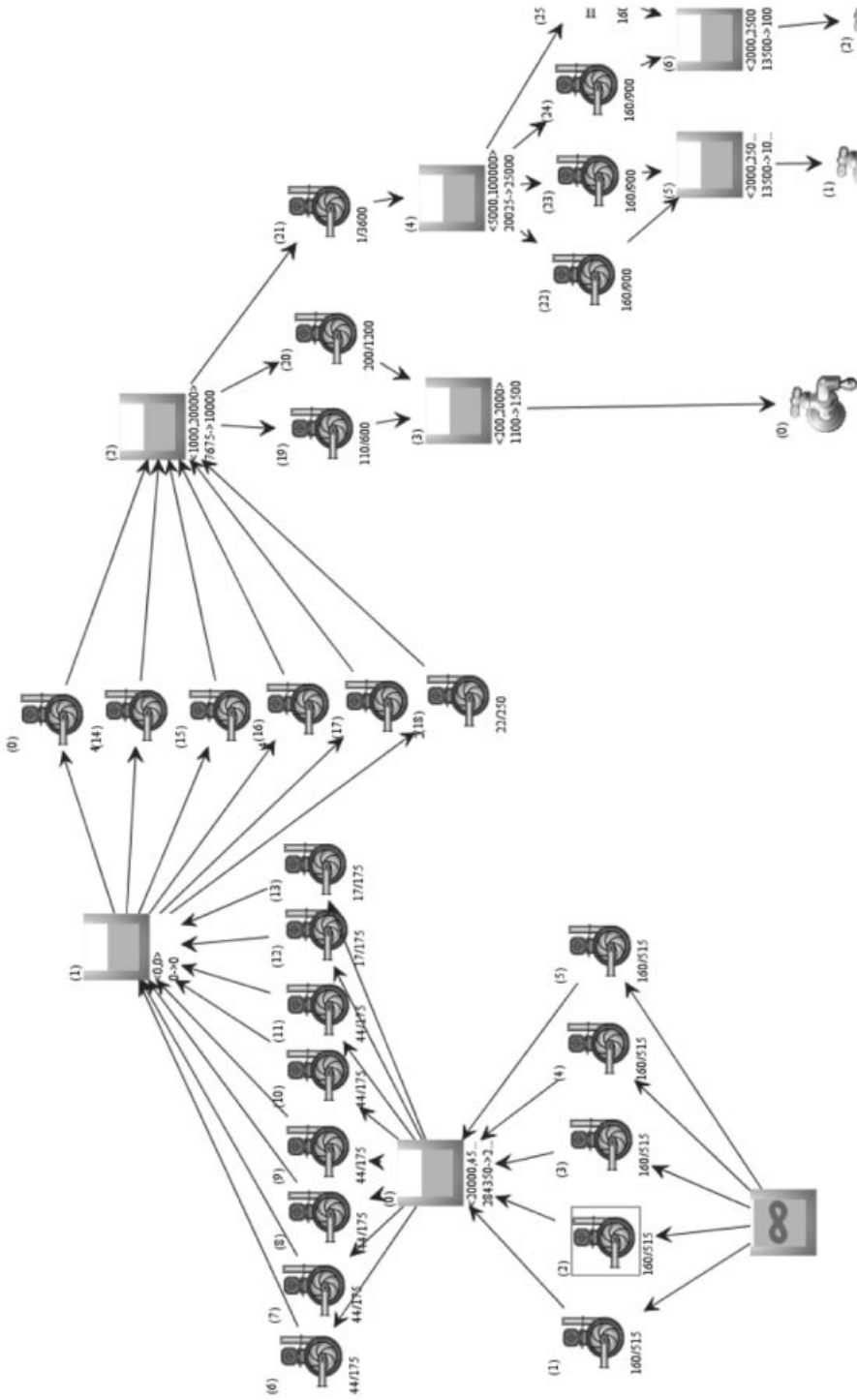


Figure 15.4 Model of a water-distribution system used for automated demand-response programs. Courtesy Activity, Smart-DR screenshot.

In France, the grid supply and demand adjustment became the responsibility of an independent authority, RTE (Réseau Transport Electricité), in 2002. Each French or foreign commercial electricity operator connected to the public grid must report their production, their projected and actual demand to RTE. They are also called “balance responsible entities”, and are responsible for properly balancing supply and demand within their own perimeter. Each balance-responsible entity may use its own production facilities, or buy in advance power production from others. RTE monitors in real time the balance of the grid and the load of main transmission lines, and takes the appropriate short-term actions to make sure the aggregate supply and demand are balanced, and that the load on transmission lines remains within acceptable limits.

During exceptional network imbalance events, RTE may put in place special network-protection measures, such as a 5% drop of tension or in extreme cases disconnection of high-voltage transformers, network split-up, and so on. Of course, such last-resort options must be avoided and remain exceptional. In order to adjust supply to the level of demand, RTE uses three levels of routine regulation mechanisms, which are permanently in place and used during normal network operations: the primary, secondary and tertiary adjustment mechanisms (Figure 15.5).

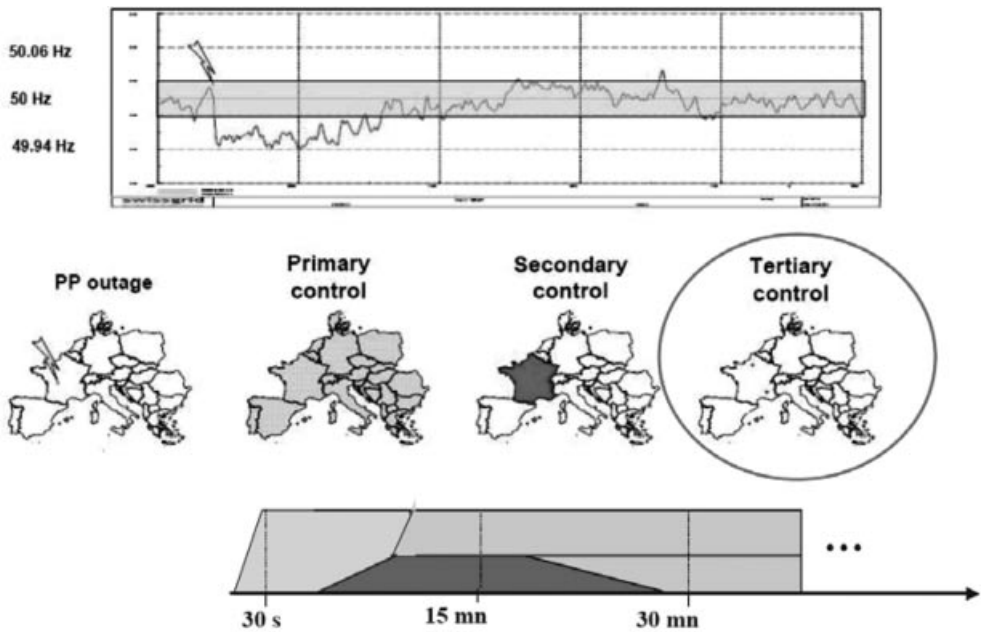


Illustration courtesy of RTE

Figure 15.5 Restoring balance: from immediate (primary) to midterm (tertiary) compensation mechanisms.

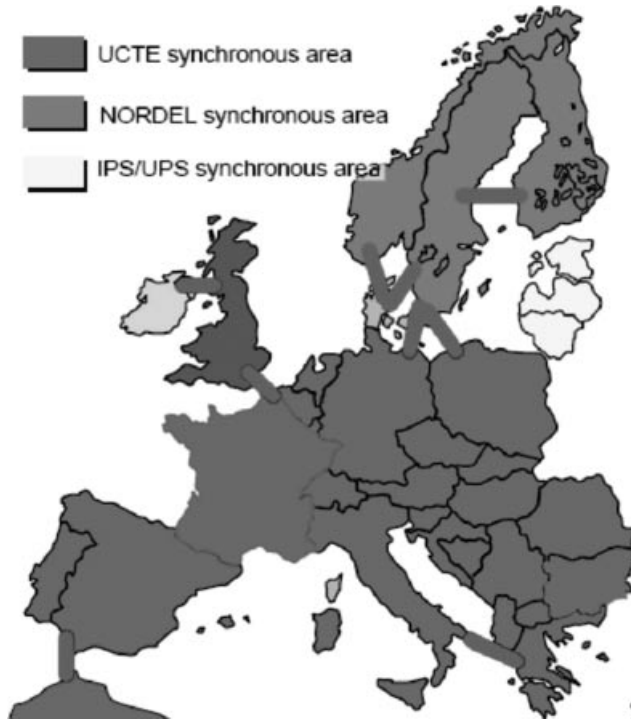


Figure 15.6 Interconnected synchronous areas in Europe.

The **primary adjustment mechanism** is based on static regulation loops in each production unit. Total primary capacity reserve is about 700 MW during winter in France. If, as a result of excess demand, the network frequency starts to drop, the primary adjustment mechanism reacts in each power plant, and adds an aggregate of 4400 MW of power or more per missing Hertz in less than 30 s. The primary power reserve ensures fast local regulation close to the power injection point. As this mechanism is in place throughout Europe and networks are interconnected (Figure 15.6), the primary reserves of all networks contribute to the stability of each other (about 20 000 MW/Hz of primary reserve power in the UCTE zone).

- The **secondary adjustment mechanism** is based on dynamic regulation loops in each production unit (with parameters adjustable by RTE in real time). Its capacity represents about 500 to 1000 MW, which can be mobilized in 2 to 15 min, depending on the extent of the adjustment required. Hydro power plants are a major contributor of the secondary reserves. Secondary-level adjustments provide coordinated regulation over wider areas, and aim at maintaining frequency stability and network synchronicity, while restoring the primary reserves.

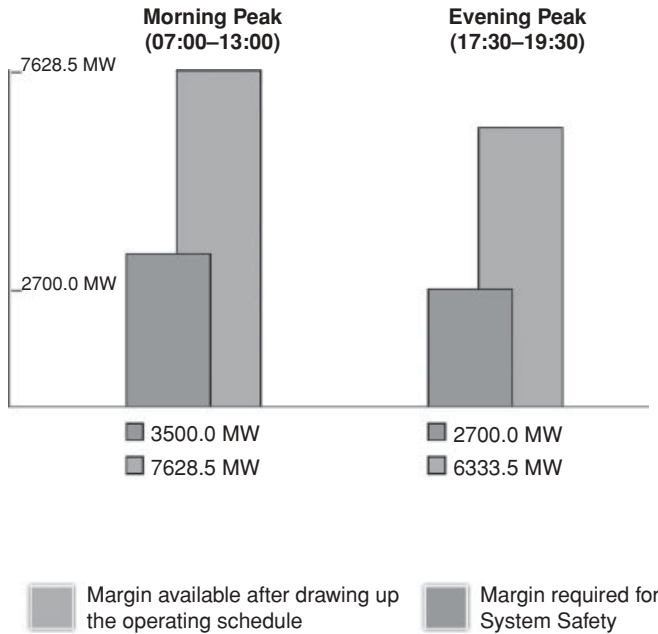


Figure 15.7 RTE uses the tertiary adjustment bids to maintain acceptable supply margins during demand peaks.

- The **tertiary adjustment mechanism**, based on semiautomatic and manual procedures, has a capacity of about 1000 MW within 15 min and an additional 500 MW within 30 min. The French demand-response market supplies the tertiary adjustment mechanism, and is used to readjust the daily power-generation schedule in order to reconstitute the primary and secondary adjustment capacities (Figure 15.7).

15.5.2 The Bidding Mechanisms of the Tertiary Adjustment Reserve

The management of the tertiary adjustment reserve is based on a permanent bidding mechanism in order to maintain an up to date inventory of potential short-term balancing offers: additional production or power shedding offers in case demand exceeds supply (see Figure 15.8), supply reduction or demand increases in case supply exceeds demand.

Bids for day D are open at D–1 4 p.m. and 9 p.m., and then can be added, modified or removed on D day every hour.

Each bid is characterized by:

- Its nature. Additional supply bids and power shedding bids allow RTE to compensate for an excess of demand: RTE will pay for the additional power injected in the public

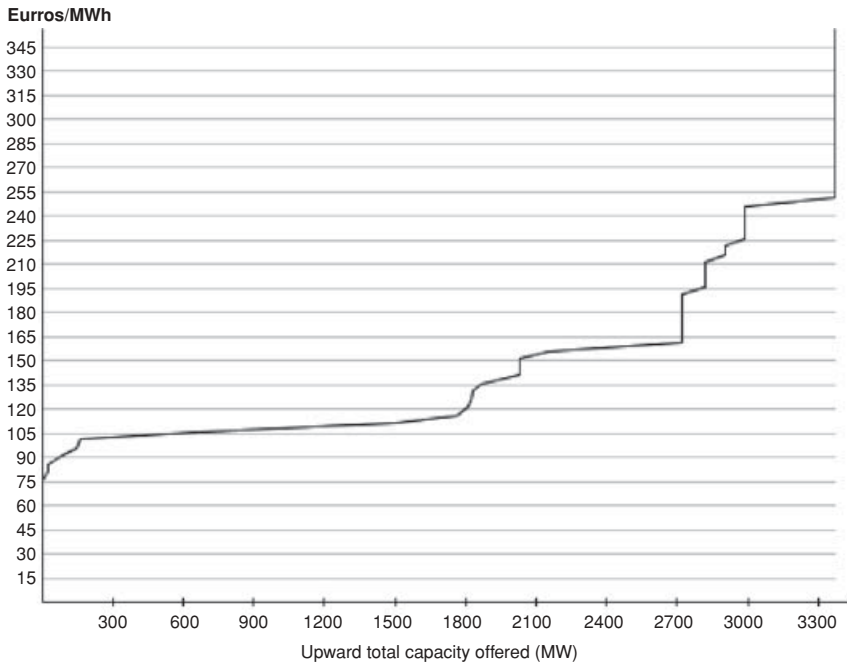


Figure 15.8 RTE typical cumulative upward bid capacity and marginal prices during the 17:30–19:30 peak time.

grid, and for the demand reduction of bidders. Supply-reduction bids and demand-increase bids allow RTE to compensate an excess of supply: the bidders will pay for the electricity sourced from the public grid, which they use either to replace their own supply reduction, or to power the extra demand.

- A validity period.
- A selling price (which might be a time-dependent price).
- An activation delay: the time that the bidder needs to make its offer effective and compliant with RTE activation level after it receives an activation notification.
- A minimal activation time.
- The amount of corrective power available (capacity available in MWh and peak power in MW), with a minimum of 10 MW for a half-hour. The corrective power may consist of additional supply, power shedding, supply reduction or demand increase depending on the nature of the bid.

In addition, for production plants participating in the adjustment power bidding, RTE requires to receive on D–1 the normal production/demand plan for day D, in order to check the effectiveness of additional supply bids.

The public network balancing state is assessed in real time by RTE. The network is self-stabilized by the primary adjustment and secondary mechanisms described above

(Figure 15.5), but RTE needs to restore the margins of these immediate reserves by using the corrective capacity made available by the bidders:

- **If demand exceeds supply:** RTE selects additional power-supply and power-shedding bids (what RTE calls the “fast reserve” of power) in increasing price order, and activates them totally or partially. The level of activation can be adjusted in real time by RTE within the limits of the proposed bid.
- **If supply exceeds demand:** RTE selects supply-reduction and additional demand-purchasing bids in decreasing price order, and activates them totally or partially.

15.5.3 Who Pays for the Network-Balancing Costs?

As illustrated in Figure 15.9, in a typical day RTE will activate more power to compensate for excess demand (paying prices above the spot price) than to compensate for excess supply (charging the sold excess power below spot market prices). As a result, this stabilization mechanism represents a net cost.

RTE decided to allocate this cost among equilibrium operators who contributed to the network imbalance, as illustrated in Figure 15.10.

RTE administratively classifies every half-hour in the day in the “upward adjustment” (excess demand) or “downward adjustment” (excess supply) category, based on the dominant trend in the half-hour.

If demand exceeded supply during the past half-hour (Figure 15.10, top):

- Some “balance responsible entities” may be in a situation where supply in their perimeter exceeded demand for this half-hour, effectively reducing the aggregate network

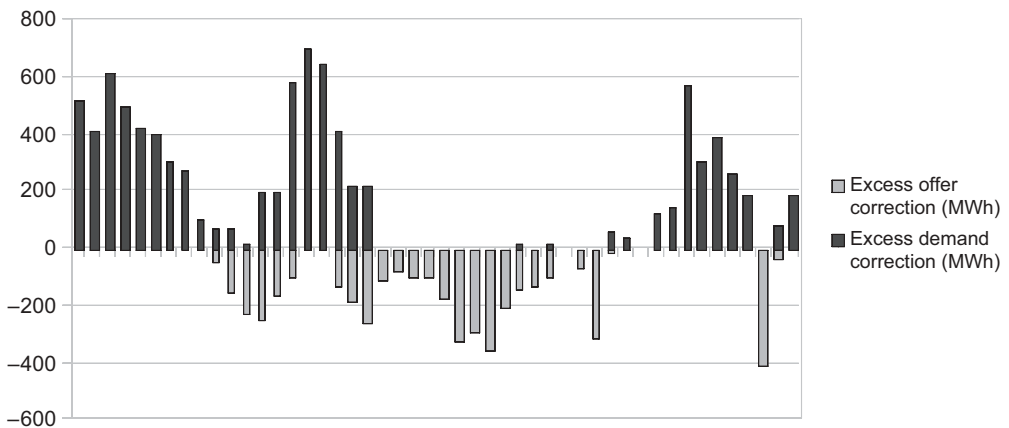


Figure 15.9 Typical RTE daily activation profile of adjustment bids (in half-hour steps).

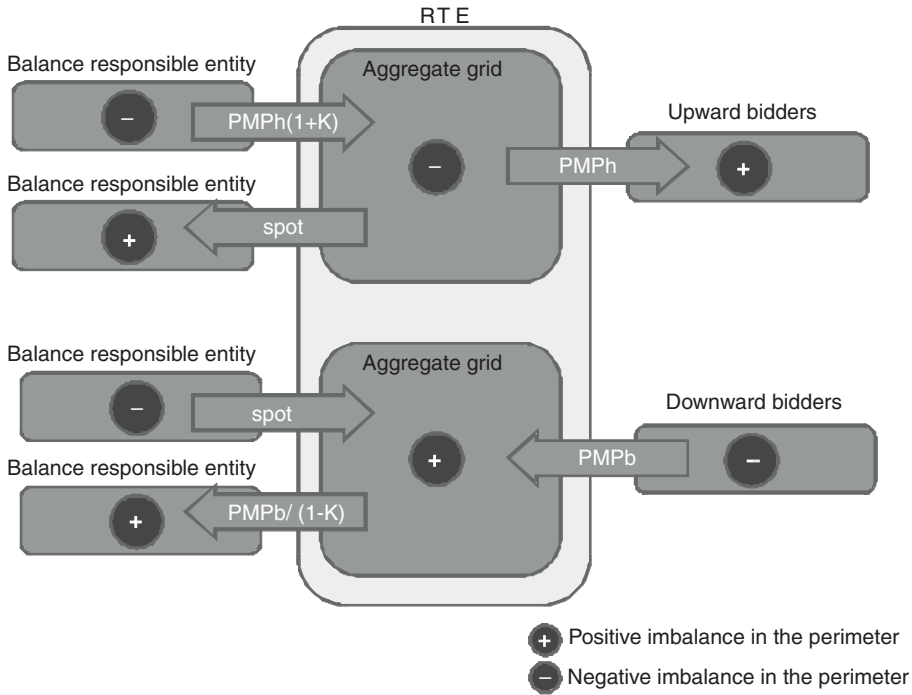


Figure 15.10 Financial flows between RTE, balance responsible entities and bidders, depending on the balancing state of their respective perimeters.

imbalance: in this case RTE buys the excess supply at the Pownext EPEX spot price determined at D-1 for delivery in this half-hour.

- The total balancing cost for the half-hour $PMPb^1$ (activated bids and purchases at the spot price), adjusted by a deterrent coefficient K (about 1.05), is shared among the “balance responsible entities” who were net users of the public network power during the half-hour, *pro-rata* of their actual use of the public grid resources to balance their perimeters. Since using these resources is costly, this of course encourages each “balance responsible entity” to properly predict demand within its perimeter in order to buy additional production in advance, rather than draw power from the last-minute “fast reserves” of the public grid.

If supply exceeded demand during the past half-hour (Figure 15.10, bottom):

- Some “balance responsible entities” may be in a situation where demand in their perimeter exceeded supply, effectively reducing the aggregate network imbalance: in this case

¹ For “Prix Moyen Pondéré à la Hausse”: average weighted price of increase bid offers (additional power, demand reduction) that were activated.

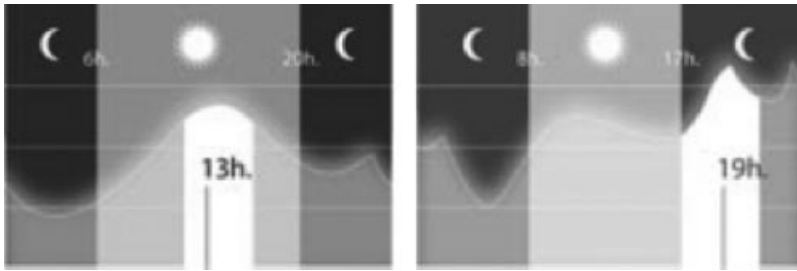


Figure 15.11 Typical demand profile during summer and winter: peak hours.

RTE provides the required additional supply, charging the Powernext EPEX spot price determined at D-1 for delivery in this half-hour.

- The total balancing revenues for the half-hour PMPb² (activated bids and power sales at the spot price), adjusted by a deterrent coefficient $1/K$, is shared among the “balance responsible entities” who were net suppliers of the public network during the half-hour, *pro-rata* of their actual supply. Since the purchasing price is below the spot price, this encourages each “balance responsible entity” to properly predict demand within its perimeter in order to sell excess production in advance, rather than below the spot price at the last minute.

15.6 The Opportunity of Smart Distributed Energy Management

In the past, only industrial sites were bidding as part of the demand-shedding mechanism. In addition, the residential and SME demand factor have been modulated by flexible tariffs³ in order to reduce demand during predicted peak hours, and in the most sophisticated cases, during predicted peak days (Figure 15.11). But these tariffs gradually disappeared after the liberalization of the electricity markets, due to the fact that not all operators could send the necessary tariff change triggers over the electricity network. In France for instance, it has been estimated that about 150 MWh of demand-shedding power disappears every year.

But the fact that most homes are now connected to the Internet, and the relatively low cost of the systems required to regulate energy usage in the home, are changing the situation. Since 2008, RTE experiments distributed power-shedding mechanisms using the Internet as the control network. Such a system could potentially be extended to virtually all homes and businesses using electrical heating, as broadband Internet penetration is very high in France.

² For “Prix Moyen Pondéré à la Baisse”: average weighted price of decrease bid offers (power reduction, demand increase) that were activated.

³ Tarif EJP (“Effacement des jours de pointe” or peak day demand-shedding tariff. It is no longer sold to new customers.

15.6.1 Assessing the Potential of Residential and Small-Business Power Shedding (Heating/Cooling Control)

For simplification, we consider a perfect heating system that exactly maintains its reference temperature (setpoint) at any moment. In practice, most real residential heating systems are either switched on or switched off, with a certain temperature hysteresis. These systems are studied in Appendices A, B and C.

Each house can be characterized by its thermal capacity C , and its thermal transmission factor K . If the current outside temperature is T_{out} and the inside temperature maintained by the heating system is T_{ref} , then the average heating power consumed by the house is:

$$P = K \times (T_{\text{ref}} - T_{\text{ext}})$$

The demand shedding mechanism will reduce T_{ref} by a small amount Δ , the new reference temperature will become $T_{\text{ref}} - \Delta$. The heating system will switch off to let the house cool to the new desired temperature. This will last $(C \times \Delta) / [K(T_{\text{ref}} - T_{\text{ext}})]$ (first-order approximation if $\Delta \ll (T_{\text{ref}} - T_{\text{ext}})$, the exact formula for the temperature evolution is an exponential).

After this cool-off period, the heating system is turned back on, and the power consumption becomes:

$$P' = K(T_{\text{ref}} - \Delta - T_{\text{ext}})$$

For a typical house, let us evaluate the order of magnitude of these parameters during winter ($T_{\text{ref}} - T_{\text{ext}} = 10^\circ\text{C}$), for a power-shedding period of 3 h allowing a decrease of T_{ref} by 1°C ($\Delta = 1^\circ\text{C}$):

- Average heating power $P = K \times (T_{\text{ref}} - T_{\text{ext}}) = 3 \text{ kW}$
- Time to cool by 1°C when heating is switched off :

$$C \times \Delta / K(T_{\text{ref}} - T_{\text{ext}}) = 100 \text{ min}$$

- Energy saved during the cooling period = $3 \text{ kW} \times 100 \text{ min} = 5 \text{ kWh}$. The same amount of additional energy will need to be consumed later in order to reheat the house back to T_{ref} .
- Energy saved during the remaining 80 min (1.33 h) at a lower temperature of $T_{\text{ref}} - \Delta$:

$$K \times \Delta \times 1.33 = P \times \Delta / (T_{\text{ref}} - T_{\text{ext}}) \times 1.33 = 0.4 \text{ kWh}$$

This saving, which reflects the lower level of comfort accepted by the home owner during the power-shedding period, will never be compensated by higher consumption later on if the heating setpoint is reset to its original value T_{ref} .

For 5000 such houses, the energy saved during the power shedding period is about 27 MWh. In order to reheat the house back to T_{ref} , about 25 MWh of higher consumption will have to be planned later on.

This simplified analysis does not take into consideration the hysteresis of the heating systems. In practice, the temperature-evolution curve in most homes is a zig-zag function of time (Appendix A, Figure A.1): when the heating system is on, the temperature increases by X °C/h, when it is off, it decreases by Y °C/h. A traditional thermostat regulates the temperature with a hysteresis of H °C: the heating system is switched on when the temperature reaches T_{ref} , and switched off when it reaches $T_{ref} + H$. We take this hysteresis into account in Appendix A and show in Appendix B that, as a result of this hysteresis, it is possible for a demand-management system to introduce synchronization across the heating systems, that is, they will all switch off and on at the same time, instead of randomly. This can create unacceptable consumption peaks in the grid. Therefore, the exact thermostat control algorithm is important: Appendix C exposes a control law which does not introduce synchronization, but instead introduces gradual power decreases and increases at the beginning and end of the power-shedding period. These issues are the main motivation for the introduction of randomization factors in ZigBee SE 2.0 (see chapter 13).

15.6.2 Analysis of a Typical Home

Figure 15.12 shows the values for X , Y and H for a typical home, and Figure 15.13 shows the daily power consumption as a function of external temperature. When active, the

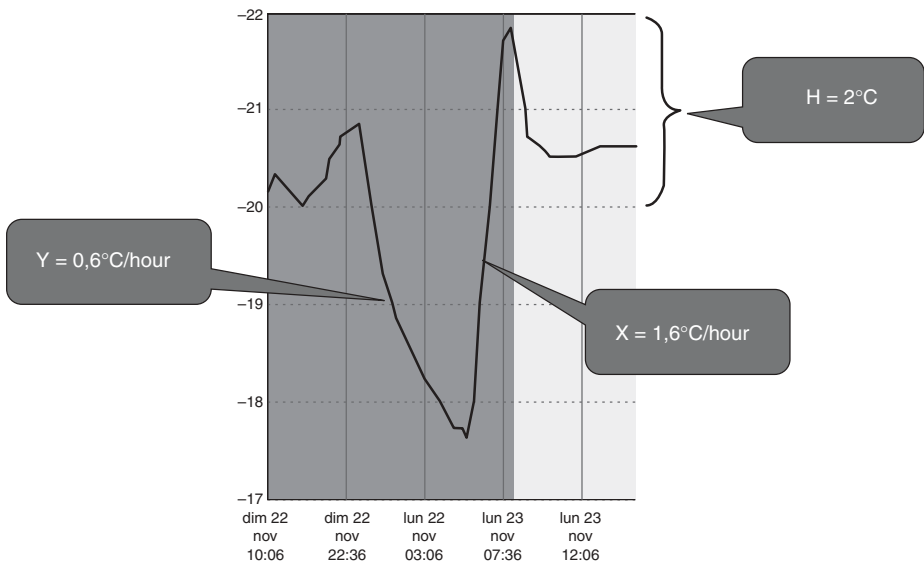


Figure 15.12 Typical parameters for a residential home.

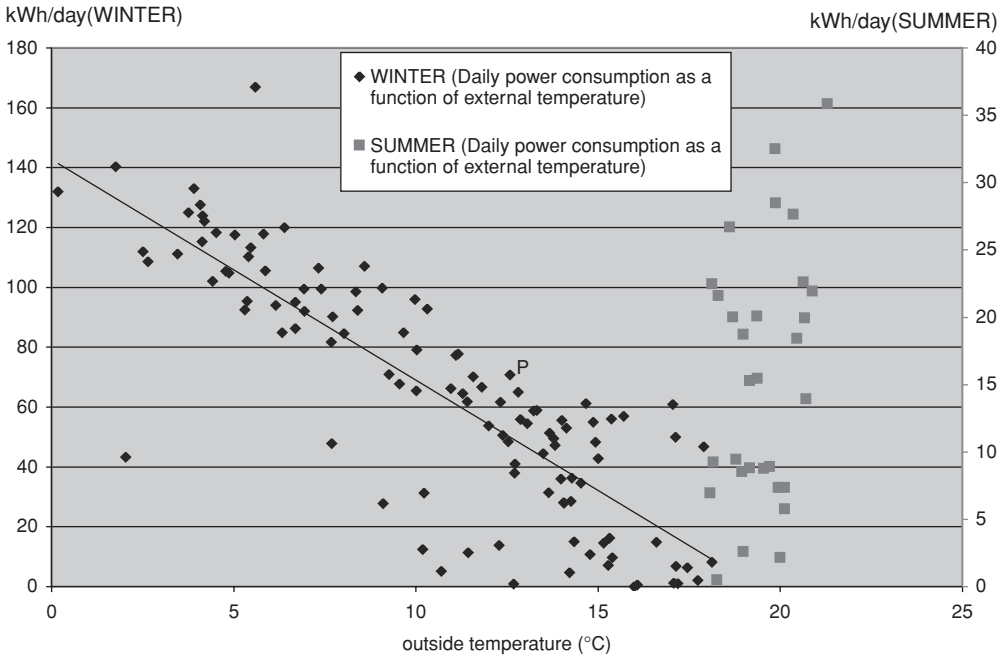


Figure 15.13 Daily power consumption as a function of external temperature.

electric heating system is clearly dominant over other uses. The yearly power usage of this 120 m² house is about 20 MWh.

The various physical characteristics of the home are related as follows:

$$Y = K(T_{ref} - T_{ext})/C \approx 0,6^{\circ}\text{C/h}$$

$X = P_{max}/C \approx 1,6^{\circ}\text{C/h}$, where P_{max} is the power of the heating system when turned on (this sample house has a basic on/off thermostat).

15.6.2.1 Potential Using Only the Daily Excess Demand / Excess Supply Volumes on the Public Grid

Let’s evaluate the potential demand response contribution of 5000 homes (with the same characteristics as the sample home we studied above) during a mild winter day: the outside temperature is 10 °C, the daily consumption is 90 kWh in order to maintain about 20 °C in the home. The average power consumption of these 5000 houses is:

$$P = 5000 \times 90/24 = 18.7\text{ MW}$$

If we decrease T_{ref} by 1 °C, 5000 similar homes will decrease their power demand from 18.7 MW to zero in 1 h 15 min (H/X , see Figure 15.14 and Appendix C where we

Δ	Tdown (100% to 0% P in h)	Tzero (0% P in h)	Tup (0% to 100% P in h)	Tzero+(Tdown+Tup)/2 Equivalent time at 0% P
-1°C	1:15	00:25	1:15	1:40
-2°C	1:15	2:05	1:15	3:20
-3°C	1:15	3:45	1:15	5:00

Figure 15.14 Demand response cycle parameters with $H = 2$, $X = 1.6$ and $Y = 0.6$.

use a demand-response policy that avoids synchronization effects), then power demand will remain null for an additional 25 min ($\Delta/Y-H/X$), and increase back to 18.7 MW in another 1 h 15 min (in reality the power consumption will be slightly less, due to the lower thermal dissipation when the house is 1 °C colder, see Appendix D for details). The full cycle saves 31 MWh compared to the normal consumption pattern and takes about 3 h to complete, leaving all homes 1 °C colder.

5000 120 m² homes, winter day (10 °C outside temperature)
 -1 °C → -31 MWh over 3 h

In order to maintain the average home temperature settings, such a -1 °C negative demand response cycle would need to be followed by a reheat cycle of +1 °C: this cycle would raise power demand from 18.7 to 28 MW ($P \times (H + \Delta) / H$, see Figure 15.15 and annex C) in 37 min (Δ/X), then remain at 28 MW for 2 hours 42 min ($H/Y - \Delta/X$), then decrease to 18.7 MW in another 37 min. The full reheat cycle drains an additional 31 MWh compared to the normal consumption pattern, and takes about 4 h to complete.

5000 120 m² homes, winter day (10 °C outside temperature)
 +1 °C → +31 MWh over 4 h

We can see that, **if the energy tariff is flat, the impact of a full power-shedding / reheat cycle on the home owner energy bill is nearly neutral:** if we neglect the slightly lower thermal dissipation of the house during the power-shedding period the energy saved during the power-shedding period is the same as the additional power consumed during the reheat period. However, the electricity tariff is rarely flat for homes using electric heating,

Δ	Tup (0% P to $P^*(H + \Delta)H$ in h)	Thigh (Demand= $P^*(H + \Delta)H$ in h)	Tdown ($P^*(H + \Delta)H$ to 0% P in h)
+1°C	0:37	02:42 (150%P)	0:37
+2°C	1:15	2:05 (200%P)	1:15
+3°C	1:52	1:27 (250%P)	1:52

Figure 15.15 Reheat cycle parameters with $H = 2$, $X = 1.6$ and $Y = 0.6$.

therefore electricity adjustment frameworks should seek to decorrelate the metering data used for billing purposes by the commercial supplier and the metering data used for the demand response participation. This is the case in France for industrial demand-response (>250 kW) where RTE automatically compensates metering data before sending it to commercial suppliers. The exact mechanism is still under discussion for lower power demand response use cases.

In order to evaluate the maximum potential of demand response in a favorable case, we are taking the following maximum tolerance assumptions of home owner for changes of the home minimal temperature (i.e. Δ), *in addition to that already planned* (Figure 15.12 shows that a $-3\text{ }^{\circ}\text{C}$ temperature drop is already planned at night for our sample house), depending on the time of day:

- between 7 a.m. and 9 a.m.: $-1\text{ }^{\circ}\text{C}$ (family wakes up, breakfast time);
- between 9 a.m. and 11:30 a.m.: $-2\text{ }^{\circ}\text{C}$ (children at school);
- between 11:30 a.m. and 2 p.m.: $-1\text{ }^{\circ}\text{C}$ (lunch time);
- between 2 p.m. and 6:30 p.m.: $-2\text{ }^{\circ}\text{C}$ (children at school);
- between 6:30 p.m. and 10 p.m.: $-1\text{ }^{\circ}\text{C}$ (dinner, going to bed);
- between 10 p.m. and 7 a.m.: $-2\text{ }^{\circ}\text{C}$ (everybody sleeping).

The home-owner allowance is at least $-1\text{ }^{\circ}\text{C}$ for the entire day: therefore over 24 h there could be up to 3 such negative demand-response/reheat cycles (7 hours for each cycle), totalling 93 MWh of negative demand response for 5000 similar homes (and 93 MWh of additional demand outside of the negative demand-response slots). It is rather difficult to use the additional allowance of $-2\text{ }^{\circ}\text{C}$, because at a minimum 7 h are necessary to decrease the temperature by $1\text{ }^{\circ}\text{C}$ and get back to normal. This seems possible only at night between 10 p.m. and 3 a.m. (the 3 a.m. to 7 a.m. period is required for the reheat cycle). This would add another 31 MWh of negative demand response between 3 a.m. and 7 a.m. (and 31 MWh of additional demand outside of the negative demand-response slots), bringing the total demand response potential to about 124 MWh for 5000 homes with the same characteristics as our sample home, assuming a “very” tolerant home owner.

Of course, ideally such demand-response/reheat mechanisms should always contribute to a better network equilibrium. This is possible only if excess demand and excess supply periods alternate during the day: fortunately, as illustrated in Figure 15.9, this is the case. However, as illustrated in Figure 15.16, the volume of upwards demand adjustments is often less than the volume of downwards adjustments.⁴ In the current RTE mechanism, the realistic maximum volume of demand response would need to be lower than about 1000 MWh per day in order to be compensated within 24 h during the daily excess supply periods. With 2 demand-response cycles of $-1\text{ }^{\circ}\text{C}$ and one cycle of $-2\text{ }^{\circ}\text{C}$ per day, this represents only about 40 300 homes!

⁴ See http://clients.rte-france.com/lang/an/clients_traders_fournisseurs/vie/mecanisme/histo/volume_type_offre.jsp.

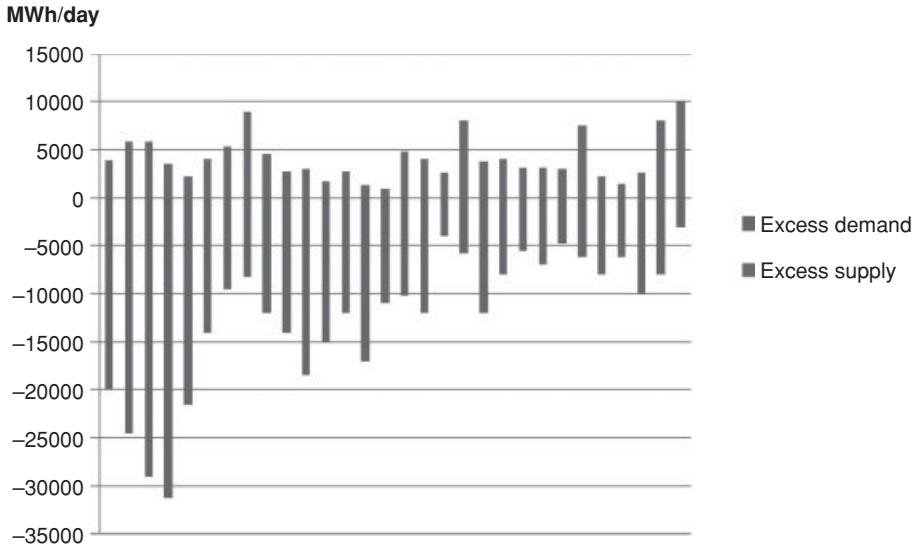


Figure 15.16 Daily volumes of excess demand and excess supply (September 2009).

If we had the possibility to get back to the normal temperature over a month, not within a day, then we would be theoretically limited only by the average monthly excess supply (about 4000 MWh/Day), not by the minimum daily excess supply. This makes the addressable demand response volume 4 times higher, but requires about 4×30 more homes (We need 4 times more energy, and 30 times more homes because it is possible to use only a 1/30 fraction of the daily temperature drift allowance per home every day in order to readjust over 30 days).

5 000 000 120 m² homes, winter day (10 °C outside temperature)
 -1 °C demand response cycles, FULLY compensated during public grid excess supply periods ↔ 4000 MWh of demand response per day (during winter).

15.6.2.2 Potential of Demand-Response Schemes that Modify the Average Demand

The demand-response potential that we evaluated above was limited by the volume of excess supply as part of the daily or monthly network-balancing bids, because we did not want to change the supplier production profile. However, if we have enough homes participating in the system and aim to fully address the monthly demand-response volume, we need to also modify the average power demand and therefore also the supplier-production profile.

We can address fully the average monthly volume of demand response (about 12 GWh per day) only if we are able to supply the peak demand-response volume every day, that is about 30 GWh. But now, out of the average 12 GWh required to reheat these homes back to the normal temperature every day, only a fraction (from about 1 GWh to 4 GWh) can be recovered from the public grid excess supply volume.

We therefore need to buy additional power from the energy supplier, for a total of αP . Since we do not want the supplier to deal with the randomness of demand response, this increased power αP will be provided constantly, even during power demand cycles: as a consequence our energy offer during power-demand cycles also increases, and the minimal number of homes required to participate in the program decreases.

Over an average day, we want the home temperature to remain constant, therefore we must provide an amount of energy equivalent to P over 24 h. We actually provide αP , except during the demand response cycles where this energy is provided to the public grid, not to the homes. We also recover a fraction s of the energy provided during the demand response cycles from the public grid excess supply. Therefore:

$$\alpha P (24 + (s - 1) [\text{total duration of demand response cycles}]) = 24 \times P.$$

$$\alpha = 24 / (24 + (s - 1) [\text{total duration of demand response cycles}]).$$

Given N_1 the number of -1°C power-shedding cycles per day, and N_2 the number -2°C power-shedding cycles per day, Figure 15.17 evaluates the necessary increase factor

Scenario	Number of homes required to fully balance the network (12000 MWh/day with peaks to 30000 MWh/day)			Required additional base power ($\alpha-1$) with the minimal number of homes required		
	s=0%	s=10%	s=33%	s=0%	s=10%	s=33%
N1=1, N2=0	4 470 000	4 500 000	4 580 000	+7,5% P	+6,7% P	+4,9% P
N1=2, N2=0	2 070 000	2 100 000	2 180 000	+16% P	+14% P	+10% P
N1=3, N2=0	1 270 000	1 300 000	1 380 000	+26% P	+23% P	+16% P
N1=2, N2=1	8 70 000	9 00 000	9 80 000	+38% P	+33% P	+22% P

Figure 15.17 Minimal number of participant homes and additional base power required as a function of the demand-response scenario.

of the base power production, as well as the number of homes required to participate in the demand response program in order to fully balance the network.

15.6.3 *The Business Case*

15.6.3.1 **The Residential and Tertiary Sector in Numbers**

Overall, the energy consumptions of the residential and tertiary sectors represent about 47% of the total energy consumption in France,⁵ compared with 28% for the industry and agriculture sector and 25% for the transports sector. The residential and tertiary sectors account for about 123 million tons of CO₂ emissions each year, or about a quarter of the total French CO₂ emissions. The residential sector alone accounts for about 2/3 of the total, while the tertiary sector accounts for only 1/3.

The business sector represents 850 million square meters (about 10 million equivalent homes, half public and half private), which are heated or cooled. The average energy efficiency varies widely depending on the sector, from about 131 kWh/m² in education to 322 kWh/m² in the transport sector.

The residential sector in France represents about 26 million homes totaling 3.5 billion square meters:

- about 13 million individual houses (including 2 million second homes and vacant homes);
- about 6 million condominiums (including 2 million second homes and vacant homes);
- about 4 million homes, apartments or houses, owned by institutional investors.

Most homes are still far from being energy efficient (see Figure 15.18): although there has been much progress, the average primary⁶ energy usage which was about 372 kWh/m² year in 1973 is still about 240 kWh/m² year for heating (hot water included). This represents an average CO₂ emission of 35 kg/m²/year. Despite the improvements in energy efficiency, the final energy demand in homes increased by 24% between 1973 and 2004. The 2008 statistics of the mandatory energy performance diagnostic (“Diagnostic de Performance Énergétique”⁷) show the following distribution:

- 31% of homes (most homes built between 1975 and 2000) consume between 150 and 230 kWh/m² year of primary energy (energy efficiency class D).

⁵ See <http://www.ifen.fr/acces-thematique/activites-et-environnement/construction-et-batiments/construction-et-batiments/la-consommation-energetique-des-batiments-et-de-la-construction.html>.

⁶ For electricity, statistics consider that the primary energy represents a multiple of the final consumption in kWh. France applies a multiple of 2.58 (2.58 kWh of primary energy is required to deliver 1 kWh to the home). Other countries of energy-efficiency labels use other values, 2 for Minergie (Switzerland), 2.85 for Passivhaus (Germany). See also: <http://www.fiabitat.com/labels-basse-energie.php>.

⁷ For a DPE simulator, see <http://www.outilssolaires.com/Archi/prin-perf.htm>.

kW/m ² /year	% of total	Average energy bill per room (Ile de France, 2008)
<= 50 (Class A)	–	<35€
51–90 (Class B)	–	35–63€
91–150 (Class C)	8%	63–106€
151–230 (Class D)	25%	106–162€
231–330 (Class E)	34%	162–232€
331–450 (Class F)	20%	232–316€
>450 (Class G)	12%	>316€

Figure 15.18 Energy efficiency of homes with electrical heating (air conditioning + sanitary hot water).

- 22% of homes consume between 230 and 330 kWh/m² year (class E).
- 18% of homes (most homes built after 2000) consume between 90 and 150 kWh/m² year (class C).

The sample house we studied above and used as a basis for our evaluations has a yearly consumption of 20 000 kWh, and 20 kWh/day for specific electricity uses, including about 10 kWh for sanitary hot water. It therefore uses about 16 000 kWh of electricity for heating and sanitary hot water, which represents about 350 kWh/m² of primary energy (Class F).

The typical uses of energy at home are listed in Figure 14.19. The relative weight of heating is expected to decrease significantly over the next 40 years, with the ever-increasing number of powered appliances in our homes, and the slow penetration of better energy-efficiency standards. The power consumption linked to IT and telecoms alone increases by about 10% per year, representing about 15% of the total consumer bills.⁸ A single Internet access device consumes over 50 kWh per year, and sometimes much more. Although the unitary consumption of these appliances improves generation after generation, there are more and more such devices in every home.

Like our sample home, **30% of principal residences in France (about 8 million) use electrical heating**. In France, during the winter electric heating is competitive on average: it releases about 180 g of CO₂ per kWh of heat due to the use of nuclear energy, compared to 300 g for fuel heating and 234 g for gas heating (source Ademe). But during the peak periods, traditional power plants are used and electric heating generates 500 to 600 g of

⁸ OCDE estimate (Christian REIMSBACH-KOUNATZE) is 15%, government estimate (rapport DETIC <http://www.telecom.gouv.fr/actualites/11-mars-2009-rapport-sur-les-tic-developpement-durable-2045.html>) is about 13.5% (55–60 TWh). The same study estimates that IT systems help save about 4 times the amount of CO₂ they generate.

Home energy usage	Today (240 kWh/m ² .year home)	2050 (60 kWh/m ² .year)
Heating	87%	30%
Hot water About 855 kWh/person.year, or about 25 kWh/m ² .	6%	30%
Cooking	3%	10%
Other	4%	30%

Figure 15.19 Energy uses in the home, now and in 2050.

CO₂ per kWh of heat. It is therefore very important to try to reduce the peak times as much as possible!

15.6.3.2 Demand Response: The Network-Balancing Business Case

5000 Homes, Single Power-Shedding Cycle per day (-1 °C), Reheat During Excess Supply

In the previous sections, we showed that by properly regulating the thermostats of about 5000 homes, with a single decrease of T_{ref} per day followed by a reheat cycle, we could bid for about 31 MWh/day as part of the network-balancing power-shedding bids, and reheat the house during the excess power supply periods (this ensures that the energy producer does not need to deviate from normal energy-production planning).

We will evaluate the revenues derived from the power-shedding bids using an average price per MWh for the activated power shedding bids of 80 €/MWh (a bit lower than the actual 1st of November to March 31st average in 2008).

The cost side is more subtle. Let us consider the normal situation (Figure 15.20), and the situation during a demand-response cycle (Figure 15.21), from the point of view of

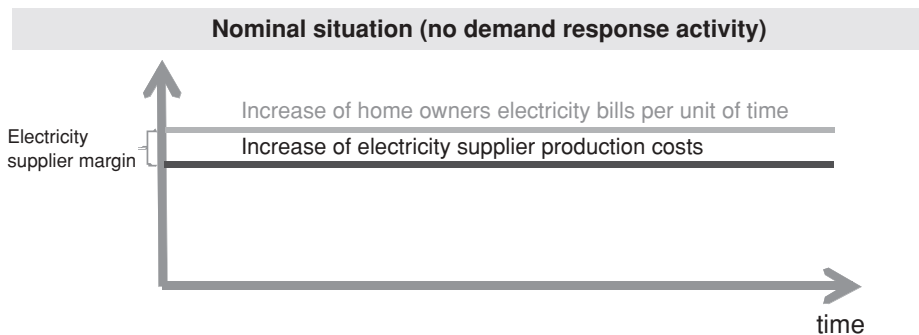
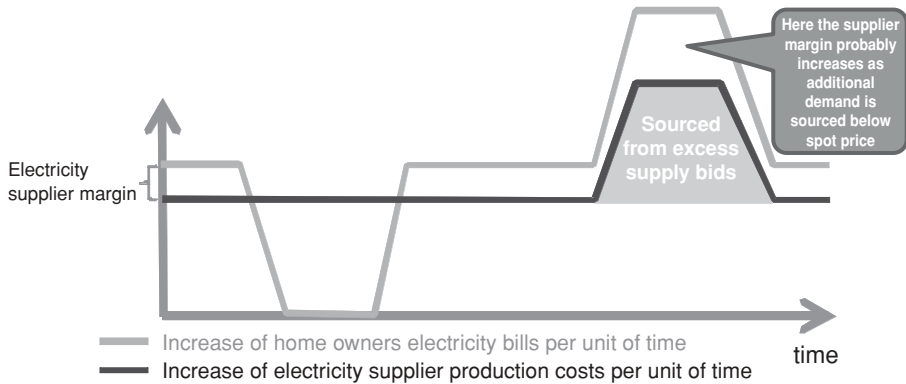


Figure 15.20 Home owners bills and supplier costs without demand response.

Demand response cycle, reheat during excess supply periods



The total sales of the electricity supplier have not changed over a full cycle. However, its cost base has increased of the amount of power sourced from the excess supply bids (below spot price).

Figure 15.21 Home owners bills and supplier costs during a demand-response cycle.

financial flows. In our simplified model the production cost of electricity is constant over time.

The total revenues of the energy supplier remain identical to the revenues it would have had without demand response (we neglect the effects of the variation of the home temperature on its thermal dissipation, these effects are evaluated in Appendix D). However, while the entity who organized the demand-response bids earns the bids revenues, the electricity supplier sees its costs increase, because it needed to source the reheat energy (or even worse . . . reheating may have caused a negative imbalance in its perimeter). Here, we further suppose that the company managing the power-shedding scheme is exchanging information with the energy supplier so that this energy can be purchased through excess supply bids.

The electricity supplier will ask to be indemnified for the additional costs. The negotiation of the amount of compensation will settle between two extremes:

- Exact compensation of the excess costs, that is, the volume of electricity supplied during the demand-response cycle, at the mean price of excess supply bids if the electricity is supplied by this mechanism, or the average production cost of energy at the moment of the demand response event. This would be the point of view of a regulator willing to maximize and accelerate the use of such demand-response strategies as opposed to additional production, and considering that the electricity supplier has no value added in this process (it only sources electricity, which is its regular business).
- Compensation of excess costs, plus a margin for the electricity supplier. This would be the default situation without regulatory pressure. In an extreme case the electricity

supplier could be willing to charge its regular residential tariffs, possibly higher than most demand response bids, severely impacting the business case for residential demand response.

Who should pay? From a purely *economic point of view*, it seems logical that the company who caused the additional cost (the power-shedding company) should pay for it. From a *legal point of view*, this is not totally clear as power-shedding companies may claim that they are only mandated by their customers to better regulate energy use in their homes, and customers never signed a contract forcing them to indemnify the operators for nonstandard demand profiles. From an *environmental perspective*, it might be optimal to consider that this additional energy supply is actually part of the grid-balancing system “costs”, due to a preference for demand shedding over additional production caused by the environmental constraints: in this case, it would ultimately be reflected by an increase of the K coefficient (see Section 15.5.3), and paid by the operators.

In the business cases exposed below, we consider that the power-shedding company will be responsible for the indemnification, which is obviously a worst case from the point of view of the power-shedding companies. For now (September 2011), the compensation level has not been settled in France, which is why we evaluate several scenarios in the business cases.

If we consider a cold season of 5 months, the yearly bid revenue for these 5000 homes is about 377 k€ ($31 \times 80 \times 365 \times 5/12$). The net revenue per home is estimated in Figure 15.22.

In addition, if the commercial price of electricity is about 100 €/MWh, and considering the average heating power of our sample house (3.74 kW), the home owner also sees his

	Gross revenue per home (€/year)	Compensation of electricity supplier (€/year)	Revenue sharing with user (50%) (€/year)	Net revenue per home per year (€/year)
Compensation at mean price of excess supply bids (~40€/MWh)	75	-37	-19	19
Compensation at mean cost of base production (~50€/MWh)	75	-47	-14	14
Compensation at mean cost of base production, plus 20% margin (~60€/MWh)	75	-56	-9	9

Figure 15.22 Business case estimation, 5000 homes, single power-shedding cycle per day.

energy bill decrease by about 15 € (See Appendix D, we use the approximation $H = 0$ to evaluate the period during which the heating system is off to 1 h 40 min (Δ/Y), and we arbitrarily consider that the average period during which the heating system maintains $T_{\text{ref}} - \Delta$ is 2 h as we do not know exactly when the next reheat cycle will be).

40 300 Homes, 4 Power-Shedding Cycles (-1°C) per day, Reheat During Excess Supply

This corresponds to the maximum potential of demand response with very tolerant home owners. We showed that by properly regulating the thermostats of about 40 300 homes, about 2 power shedding cycles of -1°C and one power shedding cycle of -2°C could be arranged, representing the maximum power-shedding capability of the home. With this strategy we can bid for about 1000 MWh/day as part of the network-balancing-power shedding bids, and still reheat the house during the excess power-supply periods.

The cold-season bid revenue for these 40 300 homes is about 12 M € (Figure 15.23).

In addition, if the commercial price of electricity is about 100 €/MWh, and considering the average heating power of our sample house (3.74 kW), the home owner also sees his energy bill decrease by about 60 €.

	Gross revenue per home (€/year)	Compensation of electricity supplier	Revenue sharing with user (50%)	Net revenue per home per year
Compensation at mean price of excess supply bids ($\sim 40\text{€/MWh}$)	301	-151	-75	75
Compensation at mean cost of base production ($\sim 50\text{€/MWh}$)	301	-189	-57	57
Compensation at mean cost of base production, plus 20% margin ($\sim 60\text{€/MWh}$)	301	-226	-38	38

Figure 15.23 Business case estimation, 40 300 homes, 4 power-shedding cycles per day.

Addressing the Full Power-Shedding Bidding Capacity

There are about 8 million homes in France using electric heating. We suppose that about a quarter of these homes can engage in a demand-response program: with these 2 million homes, and using about 2 demand-response cycles per day, we can fully balance the network and provide the required 12 000 MWh of demand response per day.

Scenario	Number of homes required to fully balance the network (12000 MWh/day with peaks to 30000 MWh/day)			Required additional base power ($\alpha-1$) with the minimal number of homes required		
	s= 0%	s=10%	s=33%	s=0%	s=10%	s=33%
N1=1, N2=0	4 470 000	4 500 000	4 580 000	+7.5% Pn	+6.7% Pn	+4.9% Pn
N1=2, N2=0	2 070 000	2 100 000	2 180 000	+16% Pn	+14% Pn	+10% Pn
N1=3, N2=0	1 270 000	1 300 000	1 380 000	+26% Pn	+23% Pn	+16% Pn
N1=2, N2=1	870 000	900 000	980 000	+38% Pn	+33% Pn	+22% Pn

Figure 15.24 Selected demand response scenario for our business case.

However, we also need to increase the normal demand of these homes by about 14%. Also, in order to accommodate the daily peaks of demand response (30 000 MWh), we will use, on average, only 40% of the available demand-response capacity of each home (Figure 15.24).

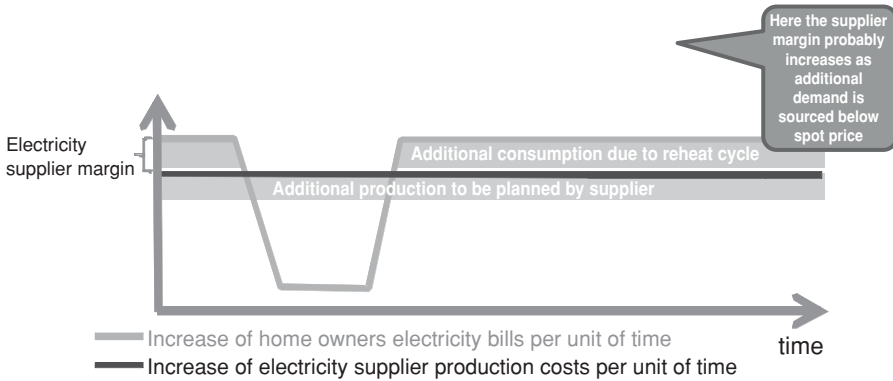
Now, the financial flows are also different (Figure 15.25).

We have spread the reheating energy demand evenly so that the electricity producer could cover it with its base production (the production that can be planned well in advance, and therefore sourced at the best prices). Clearly, this level of production cannot be stopped during the random demand-response periods so, for the energy supplier production planning, everything looks as if our homes had increased their energy demand by about 14%.

However, this additional cost is not covered by the home owners: their total bill over a demand response/reheat cycle is still identical to what it would have been without demand response.

Again, the energy supplier must be compensated for the additional cost. This time, the minimal compensation cost will be the base production cost of the supplier. We can source a fraction of this electricity from the public grid during the excess supply periods, but overall this will not represent more than 10 to 20% of the total, so the impact on the average electricity price is negligible for our evaluation. Figure 15.26 shows the resulting business case, depending on the purchasing price of electricity supplied during demand response bids.

Demand response cycle, reheat outside of demand response periods



The total sales of the electricity supplier have not changed over a full cycle. However, its cost base has increased of the amount of power provided during the demand response period, at the base production cost of the producer.

Figure 15.25 Home owners electricity bill and electricity supplier costs over a demand-response cycle (constant reheat outside of the demand-response cycle).

	Gross revenue per home (€/year)	Compensation of electricity supplier	Revenue sharing with user (50%)	Net revenue per home per year
Compensation at mean cost of base production (~50€/MWh)	70	-43	-13	13
Compensation at mean cost of base production, plus 20% margin (~60€/MWh)	70	-52	-9	9

Figure 15.26 Business case estimation, 2 100 000 homes, 2 power-shedding cycles per day (12 000 MWh/day on average).

15.7 Demand Response: The Big Picture

15.7.1 From Network Balancing to Peak-Demand Suppression

15.7.1.1 Feasibility of Peak-Demand Suppression

On a typical winter day, demand will fluctuate by about 50% from peak to trough (Figure 15.27). The peak usually occurs at 7 p.m. and can cause serious problems, especially

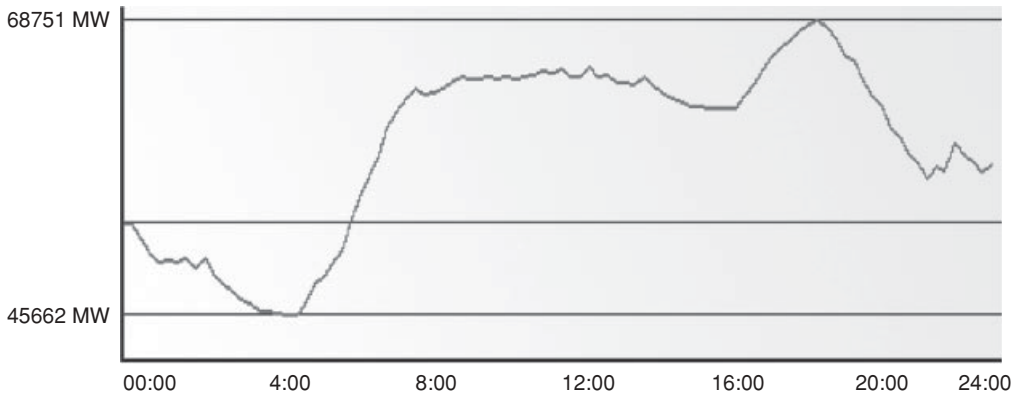


Figure 15.27 Typical daily demand profile (from http://clients.rte-france.com/lang/fr/clients_consommateurs/vie/courbes.jsp).

during the coldest days: $-1\text{ }^{\circ}\text{C}$ adds about 2100 MW to the national consumption, or twice the equivalent of the consumption of a large city for example, Marseilles (this sensitivity increases over time it was 1800 MW/ $^{\circ}\text{C}$ in 2006 and 1500 MW/ $^{\circ}\text{C}$ in 2000). Peak power consumption is in constant increase, and the trend is projected to continue : Figure 15.28 shows the forecast of RTE.

EDF, the incumbent national operator, decided to add, from 2006 to 2012, about 4000 MW of ultrafast thermal production capability (e.g., fuel turbines) that can start and provide power within 20 min. By comparison, an EPR nuclear reactor has a capacity of about 1600 MW. These turbines will be used typically less than a hundred hours per year.

The previous evaluations show that about 25% of homes using electric heating in France would need to participate in a demand-response program to completely balance the network during the winter.

It is interesting to evaluate also what could be achieved if participation in a demand response program became mandatory by law for all users of electric heating (this policy would have some logic, considering the need to reduce CO₂ emissions and the relative inefficiency of the primary energy to electrical heating conversion).

Within the framework of a mandatory program, 6 million additional homes would contribute to the peak-demand-suppression system, as well as a number of the 10 million equivalent homes of the business sector.

If a single $-1\text{ }^{\circ}\text{C}$ demand-response cycle is allowed per day, each home (arbitrarily based on our sample home, and increasing the base power production for the reheat cycles

2001	2002	2003	2004	2005	2006	2007	2008	2009	2015
79.6	79.7	83.5	81.4	86	86.3	59	84.4	91.5 (10/2009)	104 (forecast RTE 2009)

Figure 15.28 Peak consumption history in France and predictions (source RTE).

as explained above) can provide about 6 kWh over less than 3 h, within an hour or less, with a peak of about 3.75 kW. Together, the 6 million homes provide a fast capacity of over 30 GWh at a maximum power of 22 GW.

Most probably, the characteristics of our sample home do not reflect the average characteristics of the entire French residential market. Still, our evaluation gives us an order of magnitude that seems to clearly indicate that it is possible to completely replace the additional fast production capabilities, characterized by an extremely poor CO₂ performance for heating, by a voluntary expansion of demand response. The demand response is also more flexible than additional power plants, as it can provide additional power where it is required (e.g., in the often problematic western region of Brittany), without the constraints of the underlying transport network (limited incremental capacity during demand peaks, power losses).

Beyond the suppression of extreme peaks, there seems to be ample potential to also significantly smoothen the daily demand curve during winter: the portion of the demand beyond the daily average represents about 80 GWh, out of which about 20 GWh for the daily peak (Figure 15.27). From our evaluations above, it seems that the cumulated capacity of the 6 million homes would allow to shift about 30 GWh from peak to off-peak: enough to significantly flatten the daily demand curve, eliminating production and grid-capacity issues while decreasing the CO₂ emissions level and the average cost of electricity!

15.7.1.2 The Full Business Case of Demand Management

The business case associated to grid-balancing activities through demand shedding has been evaluated above. The yearly revenues associated to this activity are very dependent on the regulator decisions regarding the indemnification of additional costs of the electricity operator: while a standalone network balancing activity appears very profitable under the assumption that additional energy costs are considered a mutualized grid management cost, under any other assumption the yearly revenue per home is in the 10 to 20 €/per year range, unlikely to justify the investments necessary for such an activity.

However, the potential of demand management goes beyond grid balancing. The cost of electricity is highly variable, depending on the time of the day (Figure 15.29). These varying prices simply reflect the market estimate of the instantaneous marginal production price of electricity.

As electricity operators consider that the demand side cannot be influenced, today's predominantly fixed electricity residential prices simply reflect the average of the price curves weighted by the consumption profile of the customer.

This situation creates a significant opportunity for demand management. For each MWh of consumption "moved" from the peak hours to the lowest-cost hours the gain ranges from 40 up to 90 €! In the case of our sample home, a single demand shedding cycle of 1 °C shifts about 6.2 kWh. If this shift is optimized properly to the least-cost hours: the

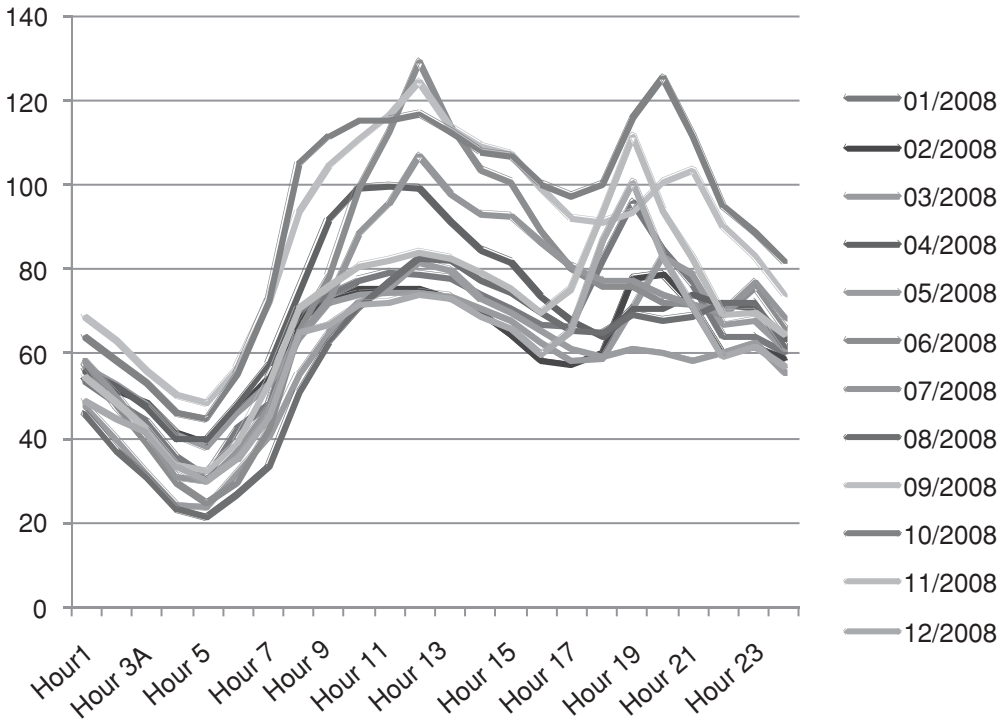


Figure 15.29 EPEX spot prices depending on delivery hour, monthly averages (2008).

resulting gain for the operator is in the 0.25 to 0.50 € per day. This represents 5 to 10% of the final consumer bill, as well as a significant marketing advantage : the CO₂ emission mix of the electricity sold is also improved. Of course, the same demand-response systems also introduce a better regulation of energy in the home, resulting in additional savings for the consumer.

The first operators to fully exploit these possibilities will be able to lower the electricity prices and will become more competitive.

Demand-response flexibility goes both ways: it is also possible to increase demand. This will make it easier for operators to source wind energy and other renewable-energy sources, as they will be able to temporarily “store” this randomly variable energy in the homes of their customers, in the form of hot water, slightly higher temperature, and so on.

We believe that the market dynamics will then lead to the general adoption of demand management, much in the same way as VoIP spread out from labs to every home in less than 10 years. In the same time frame, with the increasing penetration of electric cars, the potential of demand response – which will of course control the timing of battery recharging in homes and office parkings – becomes immense.

15.7.2 Demand Response Beyond Heating Systems

So far, we have only evaluated the potential of heating control during winter. Clearly the scope of demand response applies to other uses of electricity as well. Here is a list of the most promising domains:

- Sanitary hot water. This is one of the most common uses of electricity, and also one of the most flexible ones as no one really cares when sanitary hot water is heated, as long as one can take a hot shower in the morning.
- Air-conditioning systems. With the ever-rising comfort standards, the use of air conditioning during summer is getting more common. Already, in some countries for example, in Italy, the yearly peak demand is during summer.
- Controlling public lighting. In France only, there are about 8.5 million lamps representing 47% of the electricity bill of municipalities. Public lighting is notoriously inefficient, most lamps not only using obsolete technologies, but also wasting 30% of the energy to light up the sky, and adding to the light pollution. Replacing these lamps will take time and be very costly, however, a lot can be done to optimize the control of public lighting, such as using multilevel lighting (lower levels, e.g., one lamp out of two when there is no traffic and/or during peak power demand).
- Domestic and industrial refrigeration.
- Electric and hybrid car battery-recharging control.

15.7.2.1 Sanitary Hot Water

We consider a hot water tank of 200 l, with a daily hot water usage of 150 l. This corresponds to a daily water consumption of 500 l/day, with 50% warm water (warm water is cold water mixed with 60% of hot water from the hot-water tank).

In order to heat these 150 l from about 10 °C to about 60 °C, about 8.7 kWh are necessary each day. There are about 15% losses associated to thermal dissipation (the hot water temperature of sanitary hot-water tanks decreases by about 7 °C/day), therefore the total energy consumption will be about **10 kWh each day, representing about 3500 kWh/year**. It is generally considered that the energy consumption for sanitary hot water represents about 855 kWh per person per year, or about 25 kWh/m². This already represents about 10% of the total heating bill, but this share will increase in the future (see Figure 15.19).

The majority of sanitary hot-water tanks starts heating around 22:30, at the beginning of the reduced night tariff period. The heating period lasts about 4 h ($4 \times 2500 \text{ W} = 10 \text{ kWh}$).

15.7.2.2 Ventilation

In France, active ventilation systems are mandatory in new buildings since 1983. Most systems are single-flow ventilation systems which simply pump air from the house without trying to recover thermal energy.

The fan engine has a power of about 50 W. An average ventilation system renews about 30% of the building atmosphere per hour (this is extremely variable in a range of 10% to 100%). For a 100 m² house, with a volume of 250 m³, such an average ventilation system will pump out 75 m³ per hour. If $T_{\text{ref}} - T_{\text{ext}} = 10^\circ\text{C}$, this represents a power loss of about 325 W (plus the engine power).

The yearly energy loss caused by these simple ventilation systems represent about 2400 kWh per year (1950 kWh of thermal losses, plus 450 kWh for the fan consumption).

During the public grid peak consumption periods, the outside temperature is below 0 °C: the energy gain by stopping the ventilation system is about 700 watts (650 thermal loss watts, plus 50 watts for the fan consumption).

15.8 Conclusion: The Business Case of Demand Response and Demand Shifting is a Key Driver for the Deployment of the Internet of Things

For years, electricity operators have considered that demand was a random process governed only by statistical laws. As a result, the electricity industry has developed extremely sophisticated strategies to adapt production to demand in real time. The costs associated to this lack of control on demand are very high:

- dimensioning of public grids for peak transmissions;
- building of “peak power” plants used only a few hours during the year;
- extremely inefficient CO₂ emissions during peak hours.

However, there was no easy alternative as electricity operators, despite early attempts with powerline communications, never really developed a communication network capable of monitoring and controlling demand at the consumer level in real time.

With the ubiquitous presence of the Internet, it seems clear to us that times have changed, and that the conditions are present for a fast deployment of demand-response technologies, which seem to be capable of bringing considerable flexibility to today’s “statistical only” approach.

One of the interesting consequences of the development of demand response will be an increased level of competition among electricity operators.

Today, all operators work from the same statistical models of demand, and use the same power-plant technologies, facing identical costs. Competition is limited to limited marketing innovations and marginal production efficiency differences. If electricity operators were airlines, they would all be using the same planes, and there would be only one class of ticket: customers would choose according to seat color and meals.

After having deployed efficient demand-management tools, they will resemble more closely today’s airlines, using sophisticated yield management, several classes of tickets, low cost or business class only strategies. For airlines, this has resulted in a better usage of airplane, air traffic route and airport capacity, resulting in dramatically lower tariffs.

For electricity operators too, the end result will be a maximization of the infrastructure potential, and as a consequence lower costs and lower emission levels.

However, this communication network and home controllers, do have a cost. Deploying such networks and controllers in a silo mode (the initial temptation of all utilities and TSOs) is clearly not the optimal model. Instead, these smart-grid applications should be considered as applications of the Internet of Things: the communication network and the home controllers should support general-purpose M2M communications and applications, decreasing the marginal cost (energy, amortization) of the smart-grid use cases. In order to enable this infrastructure sharing, standards such as IP and ETSI M2M are critical: IP ensures that the communication network is application and physical layer agnostic, ETSI M2M forms the middleware layer controlling the information flows: which application can access which sensor/actuator, for which usage.

16

Electric Vehicle Charging

16.1 Charging Standards Overview

Until 2010, the standardization of EV charging addressed only very basic aspects, such as how the vehicle should behave when it detects charging power or go back to sleeping mode when the charging power disappeared (IEC 61 681:2001). The plugs themselves conformed to national standards (SAE J1772 shown in Figure 16.3, VDE-AR-E 2623-2-2/Mennekes shown in Figure 16.1, JARI/TEPCO), or to IEC industrial connector specifications: IEC 62 196-1, single phase, for the US and Japan, and IEC 62 196-2 or 62 196-3 (or 60309, see Figure 16.2) for Europe and China.

However, the standardization efforts for EV charging have intensified since 2008, and address mainly the following domains:

- **The definition of a physical connector.** On the EV side, the US and Japan have converged to the J1772 2010 plug (EV plug), which is now used by automakers Chrysler, Ford, GM, Honda, Mitsubishi, Renault-Nissan (Figure 16.5), Tesla, Toyota. Virtually all vehicles in production or planned in 2011 will use this plug. The J1772 plug is suitable only for single-phase charging.

In Europe in the beginning of March 2011 there was still some debate between a EV plug format proposed by German manufacturer Mennekes, and a format proposed by Italian manufacturer Scame. The additional complexity comes from the fact that three-phase power is much more common in Europe and actually usable in many homes for faster EV charging, and also from the various earth/ground connection regulations in each European country. The final consensus seems to be that there will be two standard EV side connectors:

- Type 1 connectors, which are simply J1772 single-phase connectors with 5 pins, a charging voltage up to 250 V and charging current up to 32 A, that is, AC charging power up to 7 kW (full charge in 4 to 6 h, or 50 to 70 km per hour of charge).



Figure 16.1 Mennekes EV plug (IEC 62196 v2 candidate).¹

- Type 2 connectors, originally proposed by Mennekes, which enable single- or three-phase charging with a charging voltage up to 500 V, and a charging current up to 63 A (70 A for single-phase charging). Three-phase 400 V charging at 32 A represents a charging power of 22 kW (full charge in a couple hours, or 100 to 150 km per hour of charge).

In Europe, there is also a lot of activity regarding the definition of an EV charging equipment (EVSE) side standard plug format, which would introduce some decoupling with the EV side plug, as each car could use its own cable. As of May 2011, the proposal of the EVplug alliance seemed to have gained a wide consensus (Figure 16.4). This EVSE plug is also called the “type 3” plug.

At the international level, the current standardization converged so far only on the requirements (IEC 62 196-1:2003 Plugs, socket-outlets, vehicle couplers and vehicle inlets – Conductive charging of electric vehicles – Part 1: Charging of electric vehicles up to 250 A AC and 400 A DC).



Figure 16.2 IEC 60 309-2 3P+N+E, 9h plug.

¹ Courtesy of loremo <http://www.flickr.com/photos/loremono/3499948469/>.



Figure 16.3 SAE J1772 plug.

- Enhanced security. Early EV chargers relied only on traditional 30 mA differential circuit breakers to detect current leakage. Modern EV chargers must have a dedicated pilot wire, which conforms to IEC 61 851:2010 (*Electric vehicle conductive charging system – Part 1: General requirements*), which implements proactive mass defect detection. This enhanced security mechanism is detailed in section “The IEC 61 851 pilot wire”.
- Basic charging power control by the charging infrastructure. IEC 61 851:2010 also specifies a simple pulse width modulation (PWM) protocol on the pilot wire enabling



Figure 16.4 Type 3 plug on one of the first EVplug alliance charging station, deployed in France.

the EV charging equipment (EVSE) to communicate to the EV the maximum charging current allowed. The initial idea was to make sure that the EV would never draw excessive currents capable of causing fires in the charging infrastructure, but the mechanism can also be used for admission control and demand response.

- Bidirectional charging control (DC charging). This functional area covers also the communication by the EV to the EVSE of EV requirements related to the charging of the EV battery by an external CC charger: like the constant current, constant power or constant tension required (depending on the battery technology). There are no agreed international standards yet at this level, IEC 61 851-23 was not published as of March 2011. The *de-facto* standard for existing vehicles is CHAdeMO, which uses dedicated pins on the J1772 connector.
- High-level communication. This functional area encompasses the requirements for charging control, but also covers security, charging and potentially many other services. At the international level, there are two standardization tracks: ZigBee SEP 2.0 (see Chapter 13), and IEC 15 118 (*Road vehicles - Vehicle to grid communication interface*). At the physical level, the proposals are to use IPv6 on top of Homeplug GreenPHY, G3 translated to the 150–450 kHz frequency band (300 to 400 kbit/s), or CAN (CHAdeMO proposal).

16.1.1 IEC Standards Related to EV Charging

The main IEC standard related to EV charging is IEC 61 851, managed by IEC TC 69. IEC 61 851 is split into several documents:

- **Part 1:** General requirements;
- **Part 21:** Electric vehicle requirements;
- **Part 22:** AC charging station requirements;
- **Part 23:** DC charging station requirements;
- **Part 24:** Communication protocol.

The IEC documents 3 cases for the physical connection:

- **Case “A”:** the cable and plug are attached to the EV.
- **Case “B”:** the cable is detachable. In case B1 the cable connects to a standard domestic plug, in case B2 the cable connects to a specific charging station. In practice, EVs sold in 2011 usually include a “B1” cable by default.
- **Case “C”:** the cable is attached to the EVSE.

The IEC also defines several charging modes. This nomenclature is now used by all EVSE vendors:

- **Mode 1** uses a standard 16 A socket outlet, single phase or 3-phase. The protection against electric shock relies only on differential breakers in the charging infrastructure or in the cable. This charging mode is prohibited in the US, but outside of the US most EVs currently circulating use mode 1 charging, and mode 1 charging is still proposed by default in many new EVs in 2011.
- **Mode 2** still uses a standard 16 A or 32 A socket outlet on the charging infrastructure side, but uses a specific plug on the EV side. The cable must include a protection device that implements proactive current-leakage detection through the pilot wire. The additional PWM communication over the pilot wire remains optional. This mode is the default for newer EVs in the US and Japan (the J1772 plug is used on the EV side), and is available as an optional cable for most recent EVs in other countries.
- **Mode 3** requires a dedicated EVSE. The pilot wire that controls the charging session is managed by the EVSE. No charging power is available over the charging cable until the EV has been detected by the pilot wire and the absence of current leaks or ground defects has been verified. The pilot wire also optionally indicates the maximum charge level by using PWM modulation: 16 A (normal), 32 A (semifast), > 32 A (fast). Mode 3 is safer than mode 1 or 2, because the plug is energized only if all of the following conditions are met:
 - the vehicle power plug is totally inserted (the pilot pin is last to connect);
 - ground continuity has been checked (pilot current present);
 - the vehicle has transmitted a signal confirming that everything is secured and charging is ready to begin.
- **Mode 4** applies to external chargers (e.g., DC charging) and also covers the inverse mode where the EV provides power to the charging infrastructure. This mode, covered by IEC 61 851-23, was not fully specified as of March 2011. It requires a serial bidirectional communication over the pilot wire.

In practice, the new EVs sold in 2011 usually had two plugs, a standard domestic male plug for mode 1, and a J1772 or equivalent plug for modes 2 and 3 charging (Figure 16.5).

16.1.1.1 The IEC 61 851 Pilot Wire

IEC 61 851:2001 requires for mode 2, 3 and 4 charging a dedicated “pilot wire” in the EV plug. This pilot wire is connected by a 1000 Ohm resistor to the mass (ground) of the vehicle and a weak current circulates from the pilot wire to the ground of the connector in order to test the correct grounding of the EV.

The new IEC 61 851-1 (edition 2 of November 2010) introduces an additional control function, and documents in annexes A and C a (non-normative) PWM protocol for the pilot wire (Figure 16.6).

In Figure 16.6, the EV indicates that it is ready to begin a charge session by closing switch S2, which changes the voltage measured at point Va. On the EVSE side, the duty

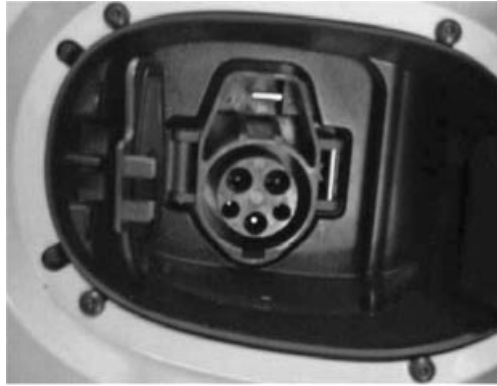


Figure 16.5 J1772 (type 1) connector on a Renault EV.

cycle of signal V_g indicates the maximum current that the EV is allowed to draw from the EVSE at any moment: no charging for a duty cycle lower than 3%, and then any value from 6 to 80 A coded by a duty cycle above 8%. A special value of the duty cycle of 5% indicates that a digital communication (“high-level communication”) is used instead.

The standard indicates that the dedicated pilot function could be replaced by a current modulation on the ground conductor (61851-1ed2, Appendix C), without defining it yet.

In this new version, the mandatory EVSE/charging cable functions for charging modes 2, 3 and 4 are:

- the verification that the EV is connected;
- continuous control of the correct grounding of the EV;
- energization and de-energization of the EV.

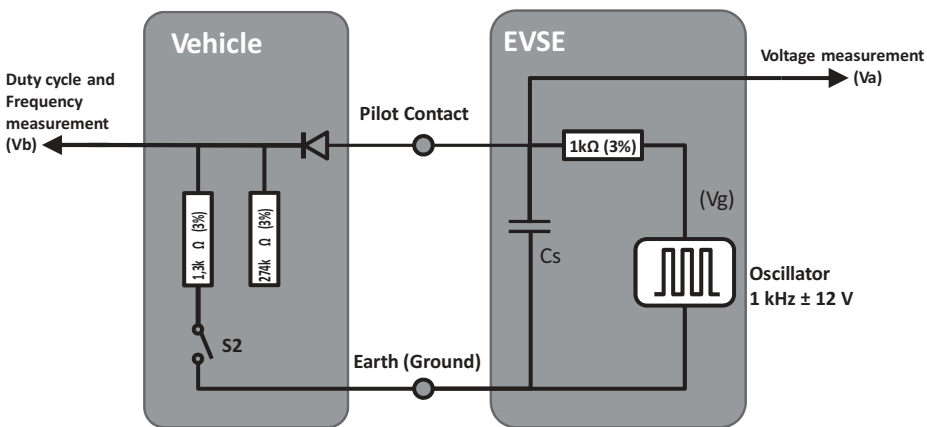


Figure 16.6 Pilot wire circuit, according to IEC 61851 and J1772.

The new optional functions are:

- selection of the charging power;
- determination of the ventilation requirements of the charging area;
- bidirectional power control (not yet specified in edition 2 of IEC 61851).

16.1.1.2 High-Level Communication: IEC 15 118

IEC 15 118 *Road vehicles - Vehicle to grid communication interface* was still a work in progress at the time of writing. This standard is managed by ISO/TC 22 (Road vehicles) jointly with IEC TC69.

- IEC 15 118-1 “General information and use-case definition” defines the vocabulary used in other parts of the standard, and focuses on the use cases for high-level communications between the electric vehicle communication controller (EVCC) and the supply equipment communication controller (SECC). Both online (Figure 16.8, SECC connected to the E-Mobility operator by a communication network as the service is provided) and semi-online (Figure 16.7, SECC-E-mobility provider communication is

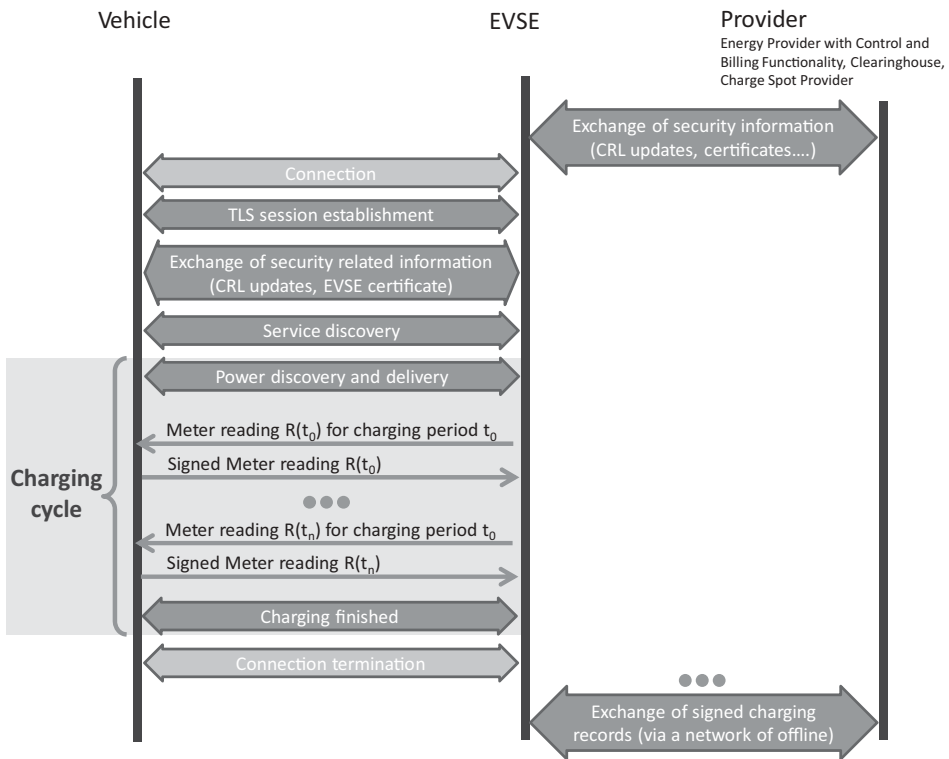


Figure 16.7 Overview of IEC 15 118-2 charging session semionline protocol flows.

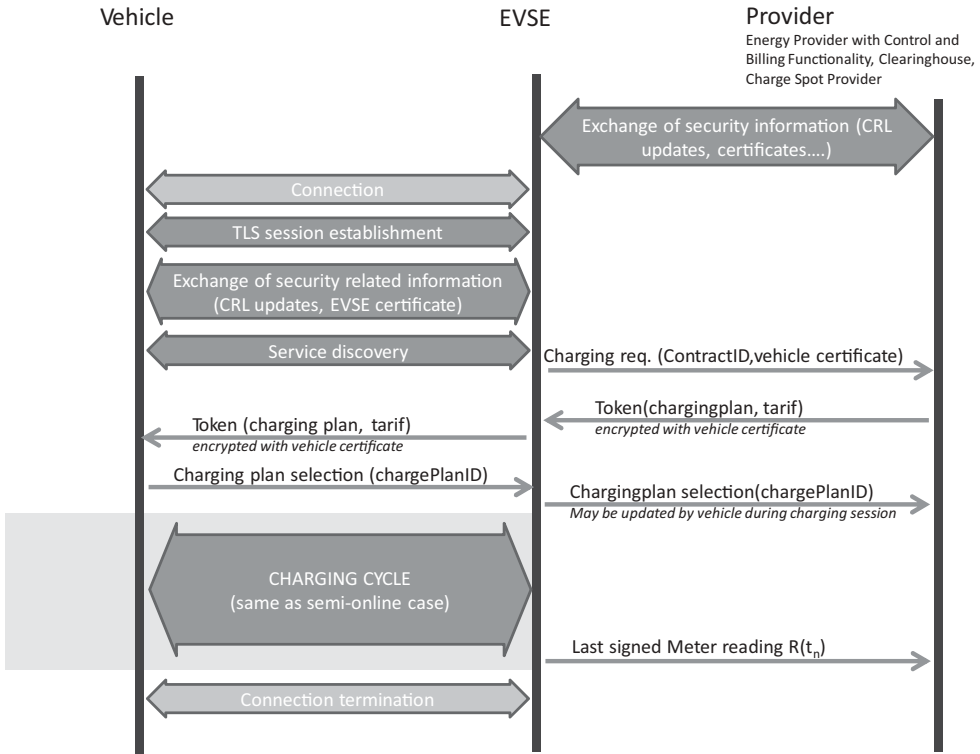


Figure 16.8 Overview of IEC 15 118-2 charging session online protocol flows.

not synchronous with the service provided) use cases are defined. IEC 15 118-1 defines four identifiers:

- **SECC Spot Operator ID:** the unique identification of the spot operator that provides the energy to the vehicle.
- **SECC Power Outlet ID:** the unique identification of the power outlet to the vehicle.
- **EVCC Provider ID:** the unique identification of the contract between the vehicle user or the vehicle itself and the energy provider given by the E-mobility operator. The E-mobility operator is defined as “the legal entity that the customer has a contract with for all services related to the EV operation”.
- **EVCC Contract ID:** it identifies the contract that will be used by the SECC to enable charging and related services (incl. billing). It is associated with the electricity consumer (who can be the driver, the owner of the vehicle or a E-mobility operator).
- **Session ID:** the identifier of the charging or value added service session, obtained after the authentication procedure.
- IEC 15 118-2 “Technical protocol description and open systems interconnections (OSI) requirements”, defines an application layer protocol on top of IPv6 (optionally IPv4) between the SECC and the EVCC. The communication uses a TCP connection secured

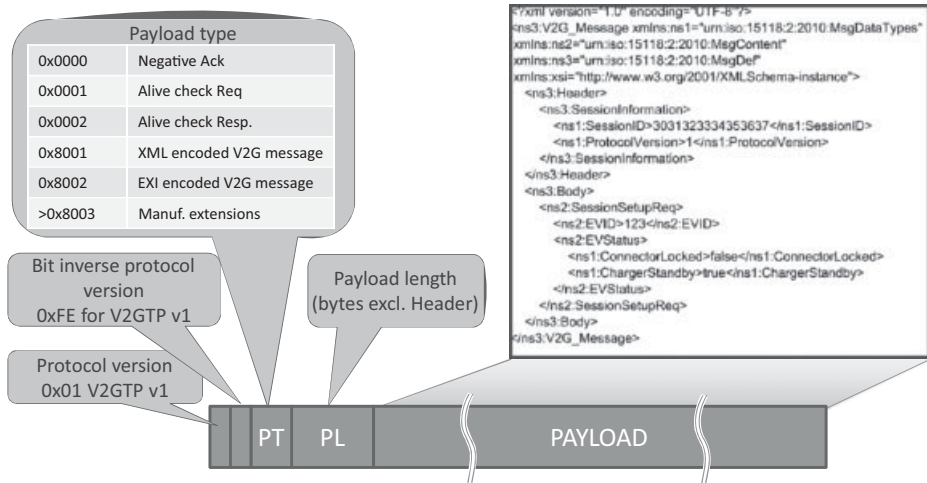


Figure 16.9 V2GTP header structure and example XML payload.

by TLS, the SECC acting as the TLS server, and uses the vehicle to grid transfer protocol (V2GTP) defined in IEC 15 118-2 (see below for more details on V2GTP). the SECC discovery protocol (SDP)

- IEC 15 118-3 “Physical and data link layer requirements”, will define the physical layer supporting the IP connectivity for the V2G protocol. It was not stable at the time of writing this document, although it seemed that PLC HomePlug GreenPhy technology (refer to the PLC chapter) had the best chance of being selected.

The V2GTP session is bootstrapped using the SECC discovery protocol (SDP). The EV begins by multicasting a UDP_SECC_LISTEN message to port 15 118 (!) as a means of discovering the SECC. The EVSe generates a random sessionID or provides the last valid sessionID if this is an attempt to resume a session. The SECC, if any, will respond to the source UDP port with a SECC discover response message, which specifies the server-side TCP port for the V2GTP session.

As V2GTP uses a TCP connection (which is stream oriented) to transport messages, it needs a framing mechanism. The V2GTP header (Figure 16.9) serves this purpose.

V2GTP defines an XML message set, which can be encrypted and signed using a certificate bound to a given contractID (current status at the time of writing). All messages have a common header format that carries the session ID.

The V2G protocol session is organized as follows:

Initialization of communication session: after session bootstrapping with the SECC discovery protocol and TLS session setup, the following messages are exchanged:

- **SupportedAppProtocolRequest/Response:** EVSE and EVCC exchange-supported protocol major/minor versions, associated namespaces and preference order.

- **Session Setup Request/Response:** the plug-in EV (PEV) specifies the PEV MAC address (PEVID) and PEV status code. The EVSE replies with a response code, an EVSEID and status code.

Service Discovery: Discover the services offered by the EVSE. Messages for this activity:

- **Service Discovery Request/Response:** the EV Service Discovery Request specifies the service scope (one or more URIs, each corresponding to a service provider), and optionally service type (e.g., charging, Internet access). The EVSE response specifies a list of supported services.
- **Service Selection Payment Request/Response:** the EV selects services and corresponding payment options in the service selection payment request, and the EVSE validates each service/payment option selected.
- **Payment Details Request/Response:** So far, only the details corresponding to payment method “contract” are specified. The EV sends a contractID, and the signature certificate chain of the EV. The EVSE responds with a random challenge that has to be signed by the EV.
- **Contract Authentication Request:** the EV sends a copy of the challenge and its signature, and the EVSE answers with a response code.

Charge vehicle: charging the EV is one possible service offered by the EVSE. It is divided into three phases:

(1) **Set up charging process:**

- **Charge Parameter Discovery Request/Response:** the PEV provides status information, charging mode (DC or AC), desired point in time for the end of charge, estimation of required recharge energy, maximum charge power, maximum charge voltage, minimum charge voltage, maximum number of charge phases, maximum charge current and minimum charge current. The EVSE responds with its status information and provides its own charging-limit parameters, as well as the identity of the relevant mobility/energy provider and a tariff table (Figure 16.10).
- **Line Lock Request/Response:** EV requests the EVSE to lock the connector on the EVSE side and specifies its own lock status.
- **Power Delivery Request/Response:** the PEV requests the EVSE to start providing charging power, specifies the selected tariff and optionally transmits the estimated charging profile (Figure 16.11) it will follow during the charging process. The EVSE replies with a response code.

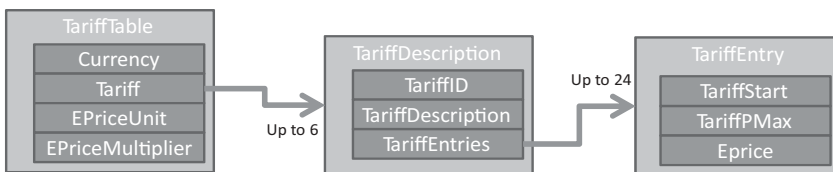


Figure 16.10 TariffTable structure IEC15118 draft (may 2011).

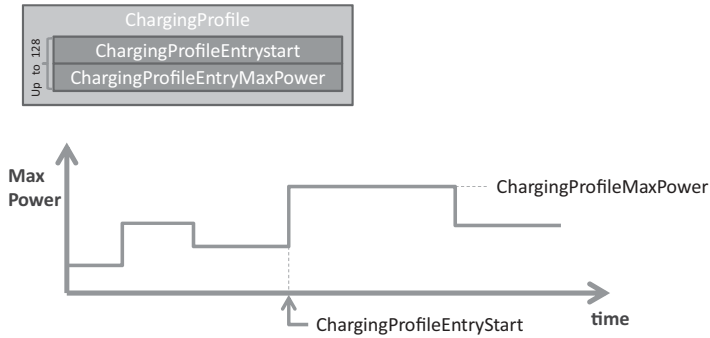


Figure 16.11 Charging profile structure in current IEC 15 118 draft (May 2011).

- **Metering Status Request/Response:** the EV request the EVSE to send its current metering status. The response of the EVSE contains a timestamp, an instantaneous estimate of the power the EVSE could deliver, optionally the current power it is delivering, a meterID and current metering information.
 - **Metering Receipt Request/Response:** the EV Metering Receipt Request contains the electronic signature of the following information elements: sessionID, MeterInfo, PEVstatus, time stamp, tariffID, metering information. This information ensures nonrepudiation for the billing process of the E-mobility provider. The EVSE acknowledges with a response code.
- (2) **Charging process:** periodic metering status request/response and metering receipt request/response (the draft standard mentions a periodicity of about 10 s), and if needed charge parameter discovery request/response and power delivery request/response.
 - (3) **Finalize charging process:** stop charging process (final power delivery request/response, metering status request/response, metering receipt request/response), unlock charge cord with line lock request/response.

At the time of writing (May 2011), there were some discussion ongoing as to whether to continue investigating a separate XML message set, or to use the ZigBee SEP 2.0. The openV2G project (<http://openv2g.sourceforge.net/>) is implementing both solutions. One of the issues is that the current design of V2GTP follows the style that was used in the 1990s for telecom protocols (e.g., H.323 for VoIP), and is not aligned with the resource-based (REST) programming style used by most recent protocols (ZigBee SE2.0, ETSI M2M . . .). This means that, as outlined in the current draft, V2GTP will not be able to leverage modern protocols designed for REST applications, like CoAP or HTTP.

16.1.2 SAE Standards

16.1.2.1 J1772 “Communications Between Plug-In Vehicles and the Utility Grid”

The original J1772 plug format was designed in 1996 by Japanese manufacturer Yazaki, and adopted by the SAE hybrid standard committee in 2001. This document defines the

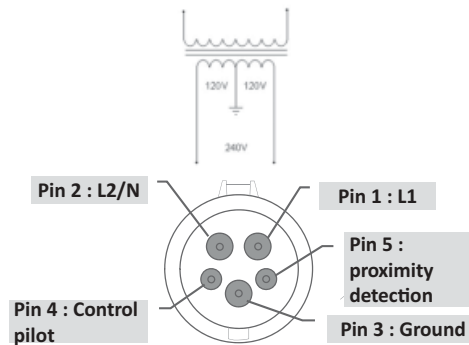


Figure 16.12 J1772 plug, and typical US residential wiring.

physical format of the EV plug for AC charging and addresses basic security and power control requirements by means of a proximity detector pin and a control pilot pin.

The original J1772:2001 specification used 9 pins, plus a proximity detector. 2 pins were used for DC charging up to 400 A, 2 pins for AC charging up to 40 A, 1 ground pin, 1 control pilot pin, and 3 pins for SAE J1850 data transmission. A magnetic proximity detector switch was provided on the EV side to prevent motion during charging.

J1772 was revised on January 2010. The J1772:2010 connector, illustrated on Figure 16.12, only uses 5 pins: two pins for single phase AC charging up to 80 A, one ground pin, one pilot wire pin, and one proximity detector pin. The proximity detector pin is connected to the ground via a resistor network in the plug, the resistor value changes when the latch release switch of the plug is pressed, signaling that the plug is about to get disconnected. The control pilot protocol is identical to that of IEC. The proximity detector and pilot wire pins are shorter, so that they connect last and disconnect first. The 3 dedicated pins provided by J1772:2001 were removed as SAE now plans to use PLC for high-level communications.

J1772 defines 3 charge levels:

- **“Level 1”**: charge via a charge cable including the pilot wire control equipment (“mobile EVSE”), connected to a US domestic plug (single-phase 120 V/15–20 A).
- **“Level 2”**: charging via a fixed EVSE station connected to a 240 V single phase, up to 80 A.
- **“DC charging”**, using an off-board charger.

16.1.3 J2293

Task force J2293 of the SAE is focused on the standardization of EVs and the communication between the EV and the EVSE. It initially addressed both conductive and inductive charging, but inductive charging (J1773) is no longer considered. The J1772 standard,

which focuses on conductive charging, implements the same pilot wire mechanism as the IEC.

The 2008 version of J2293² described a high-level charging protocol based on the J1850 serial communication protocol, which has never really been used and was formally made obsolete in 2010.

The SAE worked on a comprehensive list of use cases for EV charging, published as J2836/1 (a 244 page document!) in April 2010.

The SAE then worked on a revised architecture and communication protocol supporting all these use cases, and considered two options for the high-level communication protocol:

- The CAN bus that is adopted is all modern vehicles, but would need a specific message set to support EV charging.
- IPv6 and the ZigBee SEP 2.0 REST data model. Initially both 6LoWPAN over 802.15.4 and IPv6 over CPL were considered, at present it seems the SAE is focusing on IPv6 over CPL, most likely HomePlug GreenPhy (see Chapter 2).

This revised architecture was published in June 2010 as J2847-1 “*Communication between the Plug-in vehicle and the utility grid*” and J2847-2 “*Communication between Plug-in Vehicles and the Supply*”(which adds a message set to support DC charging and replaces J2293). Both documents are based on ZigBee SEP 2.0 (see Chapter 13).

16.1.4 CAN – Bus

The CAN bus is a proprietary protocol designed by Bosch GmbH in 1983, which was then published at SAE in 1986. CAN 2.0 was published in 1991, and uses an expanded 29-bit identifier format.

CAN is a half-duplex multimaster broadcast serial bus using NRZ encoding. Originally, only the interface to the physical layer was specified, but most recent CAN implementations rely on the physical layer specified by ISO 11 898-2:2003 (“*Road vehicles -Controller area network (CAN) - Part 2: High-speed medium access unit*”) or ISO 11 898-5:2007 (“*Road vehicles - Controller area network (CAN) - Part 5: High-speed medium access unit with low-power mode*”). The achievable bitrate varies between 1 Mbit/s (distances up to 40 m with ISO 11 898-2 or -5) and 125 kbit/s (distance up to 500 m with ISO 11 898-3).

The CAN data link layer uses ISO 11 898-1:2003 (“*Road vehicles – Controller area network (CAN) – Part 1: Data link layer and physical signaling*”), which specifies the medium access control sublayer and the logical link control sublayer. The MAC sublayer implements a CSMA collision avoidance mechanism: all messages begin with an ID,

² J2293-1 “Energy Transfer system for Electric Vehicles – part 1: Functional Requirements and system Architectures”, and J2293-2 “Energy Transfer system for Electric Vehicles – part 2: Communication Requirements and Network Architecture”.

which also encodes the priority level: in case a “1” and a “0” are simultaneously transmitted on the bus, only value “0” will be sensed by all nodes in the network. The node that had just transmitted a “1” then knows a collision has occurred and stops transmitting. The original CAN and CAN 2.0A specification uses an 11-bit identifier field. On recent CAN 2.0 implementations using a 29-bit identifier structure: the first 3 bits are reserved for the priority level, the next 18 bits encode the frame format (broadcast or unicast parameter group, and 8-bit destination address/group), the last 8 bits encode the sender address.

The ID field is followed by up to 8 payload bytes. Any higher-layer protocol may be used on top of CAN, for instance:

- The ISO transport protocol (ISO TP, 15 765-2) is used by existing diagnostics protocols on CAN (e.g., ISO 14230), and provides segmentation (transmission of up to 4095 bytes), flow control, broadcast and unicast addressing. Error recovery is left to the application layer.
- TP 2.0 that adds automatic error recovery (mainly used by Volkswagen).
- CAN Open, developed by Bosch GmbH and handed over to CAN in Automation user group (<http://www.can-cia.org/>) in 1995. It was standardized as EN 50 235-4 in 2002, which defines a object dictionary enabling cross-manufacturer compatibility.
- SAE J1939, which defines a segmentation and flow control mechanism, and an application layer for buses, trucks, and agricultural equipment.

At present, there is no proposal to port 6LoWPAN over CAN, which means the CAN option for EV to EVSE communication would not be IP based and would serve only for low-level control purposes. This is considered problematic by some car manufacturers who believe the EV to EVSE connection may be leveraged also for other uses, for example, downloading movies while charging, and in general would require the versatility of IP.

CAN, which was designed for closed systems, also lacks a set of specifications for strong authentication and security, which are required for communications between the EV and EVSE.

16.1.5 J2847: The New “Recommended Practice” for High-Level Communication Leveraging the ZigBee Smart Energy Profile 2.0

The J2847 series is the result of joint work since 2009 between SAE and the ZigBee and HomePlug alliances, with the following objectives:

- defining a message set that would support the requirements of J2836/1;
- integrate with home area networks and utility networks;
- follow the REST model recommended by the NIST, and be based on IPv6.

Because the high-level communication is based on IPv6, any underlying transport layer supporting IPv6 might be used. The REST interactions might be transported by HTTP

over IPv6 for high-speed physical layers such as HomePlug GreenPhy or G3, or by CoAP over 6LoWPAN for lower-speed physical layers such as 802.15.4 or low-energy CPL. Refer to Chapter 2 for more details on CPL technologies.

The original ZigBee SEP 2.0 specification does not address the EV charging use case only, therefore the SAE reworked the specification and split it in several documents focused on the specific requirements for EV charging:

- **J2847/Part 1** “Communication between Plug-in Vehicles and the Utility grid” lists the messages used to support the J2836/1 baseline requirements:
 - vehicle, customer, evse identification and authentication;
 - energy request management (energy and power requests, energy and power management and scheduling);
 - timing information (start and stop time, anticipated duration, charging profile, actual start time);
 - pricing, including time of use pricing, critical peak pricing;
 - load control (demand response, load shifting);
 - vehicle information and status.

The other J2847 documents were not published at the time of writing:

- **J2847/Part 2** “*Communication between Plug-in Vehicles and the Supply*” will replace the high-level messages specified in J2293 for the specific use case of DC charging.
- **J2847/Part 3** will focus on the *Reverse Power Flow* (RPF) use case, that is, when the EV provides energy to the utility network.
- **J2847/Part 4** will focus on charging system and EV diagnostics.
- **J2847/Part 5** will focus on vendor-specific extensions and options.

16.2 Use Cases

16.2.1 Basic Use Cases

The basic use case is, of course, the charging of the EV battery. The battery of a plug-in hybrid electric vehicle (PHEV) has a capacity of typically 5 to 6 times the capacity of a hybrid electric vehicle (HEV). The capacity of the battery of pure battery electric vehicles (BEV) is 3 times or more that of PHEVs (see Figure 16.13).

In order to allow for reasonable charging times, the charging power and current is typically 16 A or more. For an isolated home, this is enough to require load management and limiting in order to avoid simultaneous use of other high-power appliances, such as electric heating and sanitary water, electric ovens, and so on. In a residential complex, the simultaneous charging of several plug-in vehicles will also require careful synchronization of management in order to ensure fair charging of all vehicles, without exceeding the power limits of the connection between the building and the utility grid.

In both cases, this means that charging stations at home or in the work place will either require a separate utility connection (as in the case of Italy where most existing residential

		PHEV15	PHEV40	BEV
Battery capacity (kWh)		5	16	24
Charging time (20% to 100%)	1.4 kW (16 AWG)	2 h 50 mn	9 h 10 mn	13 h 45 mn
	3.3 kW (16 AWG)	1 h 15 mn	3 h 50 mn	5 h 50 mn
	7 kW (8 AWG)	35 mn	1 h 50 mn	2 h 45 mn
	19.2 kW (8 AWG)	15 mn	40 mn	1 h
	60 kW (4 AWG)	5 mn	15 mn	20 mn
Charge-depleting range or All Electric mode range		23 km–15 miles	64 km–40 miles	160 km–100 miles

Figure 16.13 Typical battery capacity and charging time of EVs.

electric connections cannot deliver more than 3 kW), or will need to be integrated in the local energy-management system. The latter is expected in all countries where the cost of upgrading the capacity of the existing home connection is lower than the cost of a new connection to the utility grid (e.g., in France where a new connection is charged about 1000€ in the best cases, and much more depending on the distance between the home and the closest distribution operator street cabinet).

Considering only the short term, EVs are beneficial for the stability of the electric grid. All transmission system operators (TSOs) face short-term mismatch problems between power production and demand (see Section 14.4 for more details). The charging power of EVs can be adapted in real time by using the pilot wire, and provide an ideal flexible load for demand-response policies.

Unfortunately, the EV is also problematic for utilities: EVs are ideal vehicles for commuters who run short distances between their work place and home. This means that the charging load will exhibit a pendular behavior as well: large and sudden charging requirement peaks in the suburbs after work hours and in business districts in the morning. However, since the midterm (hourly) pattern of such demand is predictable, utilities with smart-grid capabilities may implement load-shifting and load-capping policies to smooth the total demand in a given district: for instance the HVAC installations of business centers may be programmed to reach their setpoints just before the rush hour, possibly to slightly exceed the heating setpoint during winter and cooling setpoint during winter so that additional energy can be delivered to the EVs during a couple hours.

The basic capability provided by the pilot wire PWM charging rate control is sufficient to implement load limiting, load shifting and demand response, if all vehicles are treated the same, and of course in the specific case of a single home with a single car.

However, not all vehicles have the same needs:

- The battery of some vehicles may be charged at 90% of capacity already, while other cars may be at a critical battery level.
- Some vehicles may need to depart within the next hour, while other vehicles will not move for 8 h or more.
- Some users may be willing to pay more to be served in priority.

These examples show that more sophisticated arbitration policies that seek to optimize the aggregate satisfaction of users need more information on each EV state and requirements: the high-level communication protocol, such as ZigBee SEP 2.0/J2847, will become a mandatory feature as soon as EVs reach a penetration that make it very likely to have 5 EVs or more (including rechargeable hybrids) simultaneously connected in an office or residential complex parking, this is probably as soon as 2015!

16.2.2 A More Complex Use Case: Thermal Preconditioning of the Car

Traditional thermal engines have a nice side effect: heating is provided “for free”. However during hot summer days, we have all noticed how air conditioning affects the car performance and consumption.

This air-conditioning tax will be a burden for EVs, both during summer and winter. As much as 30% of the total energy of the battery may be spent by air conditioning, and much of it during the first 10 min for the initial conditioning: typically 6 kW for 10 min and 2 kW afterwards for heating, 3 to 4 kW initially then 2 kW for cooling.

In order to minimize this problem, a possible strategy is to precondition the car. A study published in November, 2010 by the US National Renewable Laboratory (NREL)³ quantified the improvement of preconditioning for several types of cars. The results provided in Figure 16.14 assume a mixed drive cycle of 55% city driving, and 45% highway driving. The impact of preheating would be even more favorable for short commuting cycles.

Of course, in addition to the increased range and reduced fuel consumption, the level of comfort is also improved by preconditioning.

This more complex use case shows that the communication protocols for future charging systems may need additional parameters, such as:

- The estimate of the preconditioning energy and power schedule, including the initial high-power conditioning and the ongoing conditioning. This should be calculated by the car based on external temperature measurement and air-conditioning setpoints.
- The preconditioning schedule, for example, when preconditioning is scheduled to begin.
- Whether or not the user charging profile allows preconditioning.

³“Analysis of Off-Board Powered Thermal Preconditioning in Electric Drive Vehicles” available at <http://www.nrel.gov/vehiclesandfuels/vsa/pdfs/49252.pdf>.

Car category	Charge-depleting (CD) or All-electric range (AER) reduction (additional fuel consumption)		Range increase by 20 mn preconditioning (compared to no pre conditioning)	
	Heating (-6°C ambient temperature)	Cooling (+35°C ambient temperature)	Heating (-6°C ambient temperature)	Cooling (+35°C ambient temperature)
100 mile BEV (e.g., Nissan Leaf)	-34%	-32%	+4%	+2%
40 miles PHEV (e.g., GM Volt)	-35% (+60% fuel)	-34% (+57% fuel)	+6% (-3% fuel)	+4% (-1%)
15 mile PHEV (e.g., Toyota Prius)	-19% (+3% fuel)	-32% (+49%)	+19% (-1% fuel)	+5% (-1% fuel)

Figure 16.14 Effect of air conditioning on range (55% city, 45% highway driving), from NREL study.

16.3 Conclusion

The need for communication systems in the car goes much beyond EV charging. The potential for M2M service providers is enormous: there are about 50 million cars produced in the world each year, for a total of over 600 million in circulation.

The European eCall directive, which mandates installation of automatic emergency calling in new cars as of 2014, has made it mandatory to install a cellular communication platform in all new European cars. Building on this, most automakers are now planning to use this cellular communication platform to implement a M2M GPRS connection for maintenance and monitoring of the vehicle. Such systems are already in place in all new EVs and will soon generalize to all new cars.

What is the next step? Automakers will soon provide on-board IP connectivity for the car appliances (GPS systems for instance). It would make no sense to build separate vertical systems for EV charging, car monitoring, and on-board M2M connectivity. Clearly, our vehicles will need to provide a generic M2M infrastructure able to leverage both the cellular network and IP connectivity provided through charging stations (or, in the future, by parking lots). With this in mind, we believe that the protocols as defined currently will need to evolve:

- IEC15118 current draft is “too vertical” but explores in details all the semantics for EV charging in the context of a multiactor deployment.
- ZigBee SE 2.0 is REST-based and a little more generic, but still focused on energy management. Multiactor interactions are not as clearly defined as in IEC 15 118.

ETSI M2M (see the ETSI M2M chapter) provides a generic REST framework for M2M, but so far no detailed work has been done to explore how the semantics of IEC15118 and SE2.0 could be integrated in this framework. The authors believe, however, that this is a very promising direction. The future communication architecture of our cars probably will need the use case versatility of frameworks such as ETSI M2M, complemented by specific interworking profiles for charging specific “REST resources” like those defined by IEC15118 and SE2.0. If that happens, EV charging “protocol” specifications should evolve into “REST resources” specifications, leaving the plumbing to other, more generic, protocols.

Appendix A

Normal Aggregate Power Demand of a Set of Identical Heating Systems with Hysteresis

For simplification, we suppose that weather conditions remain constant and that the temperature evolution curve in all homes is a zig-zag function of time (Figure A.1): when the heating system is on, the temperature increases by X °C/hour, when it is off, it decreases by Y °C per hour. All homes are equipped with a traditional thermostat that regulates the temperature with a hysteresis of H °C: the heating system is switched on when the temperature reaches T_{ref} , and switched off when it reaches $T_{\text{ref}} + H$. In reality, these temperature evolution curves are exponential functions of time, but this linear approximation is valid if H is small compared to the $T_{\text{ref}} - T_{\text{ext}}$, the temperature gradient between the inside and the outside of the house walls. The heating system therefore remains switched on for H/X hours (the part of the zig-zag during which temperature increases), and then remains switched off for H/Y hours (the part of the zig-zag during which temperature decreases).

We further suppose that initially thermostats are not correlated: at any moment, if we sort homes according to the last time the heating system was switched on, an identical number $k \cdot \delta t$ of homes had their heating system last switched on between T and $T + \delta t$. Note that the last time the heating system was switched on at any time T must be within interval $[T - H/X - H/Y, T]$, therefore we can write the total number of homes as $N = k(H/X + H/Y)$.

Let us calculate the number of heating systems that are switched on at any point in time T : the moment they have been last switched on must be in interval $[T - H/X, T]$, therefore we have $k \cdot H/X$ heating systems switched on.

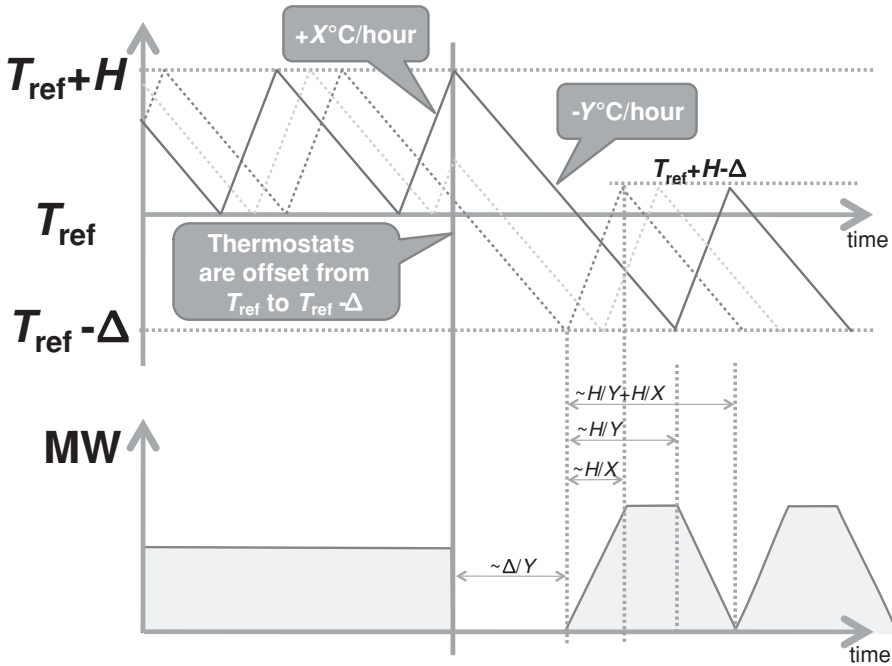


Figure A.1 Evolution of power demand after a synchronized thermostat offset.

If we call P_{\max} the cumulative power demand of all homes if they all had their heating systems switched on, the average aggregate power demand of a set of homes with decorrelated thermostats is $P = P_{\max}(H/X)/(H/X + H/Y) = P_{\max}Y/(X + Y)$.

Normal aggregate power demand of a set of homes with decorrelated thermostats:

$$P = P_{\max}Y/(X + Y).$$

Appendix B

Effect of a Decrease of T_{ref} . The Danger of Correlation

If we decrease T_{ref} by Δ °C at time 0, the effect on power demand depends on our thermostat strategy:

- We may switch off all heating systems (as by definition, all homes have a current temperature greater than T_{ref} , therefore also greater than $T_{\text{ref}} - \Delta$): the aggregate power demand of our homes instantaneously falls to zero.
- As the new hysteresis settings of our thermostat is now $[T_{\text{ref}} - \Delta, T_{\text{ref}} + H - \Delta]$, if $H > \Delta$ some homes have a current temperature in interval $[T_{\text{ref}}, T_{\text{ref}} + H - \Delta]$. We can choose to heat some or all of these homes up to $T_{\text{ref}} + H - \Delta$, maintaining residual power demand.

Let's choose the first strategy and evaluate its effects:

- Power demand will fall to 0 for Δ/Y hours: this duration represents the time it takes for the homes that were coldest (at T_{ref}) when the thermostat setting was changed to reach $T_{\text{ref}} - \Delta$.
- After Δ/H hours, heating systems are switched back on in an increasing number of homes. At time $\Delta/H + t$, the heating systems of all homes that reached temperature $T_{\text{ref}} - \Delta$ between time Δ/H and time $\Delta/H + t$ are switched on, that is, homes that had a temperature between T_{ref} and $T_{\text{ref}} + Yt$ ($t < H/Y$) when we changed the thermostat settings. These homes fall into two categories:
 - Homes that were in “cooling down” mode, the heating systems of which were last turned on in the interval $[-H/X - H/Y, -H/X - H/Y + t]$: there are kt such homes.

- Homes where the heating systems were on, and were turned off at time 0. The heating systems of these homes were last turned on in the interval $[-Yt/X, 0]$: there are kYt/X such homes.

Overall, at time $\Delta/H + t$, $kt(1 + Y/X)$ homes have had their heating system turned back on, and at time $\Delta/H + H/Y$, all homes have had their heating systems turned back on. But in the meantime, at time $\Delta/H + H/X$, the first homes that had their heating system turned back on reach temperature $T_{\text{ref}} + H - \Delta$ and get turned off: the number of active heating systems stabilizes to $kH/X \times (1 + Y/X)$, and power demand stabilizes to $P_{\text{max}} \times Y(1 + Y/X)/(X + Y) = P(1 + Y/X)$.

At time $\Delta/H + H/Y$, the heating system of the last homes to reach $T_{\text{ref}} - \Delta$ is turned back on. After this time, this additional demand therefore stops, while $k(1 + Y/X)$ heating systems per unit of time are still being switched off, having reached $T_{\text{ref}} + H - \Delta$: power demand decreases again, and zeroes at $t = \Delta/H + H/Y + H/X$.

At this point, the first homes that reached $T_{\text{ref}} + H - \Delta$ have had time to cool off, and reach $T_{\text{ref}} - \Delta$. $k(1 + Y/X)$ heating systems per unit of time are being switched on, and energy demand increases again. This scenario is illustrated in Figure A.1.

After an initial period where energy demand decreased to zero, energy demand becomes periodical, because we created a synchronizing event for all previously decorrelated heating systems.

Appendix C

Changing T_{ref} without Introducing Correlation

It is possible to decrease T_{ref} without creating a synchronizing event: for instance at $t = 0$ we can start to change T_{ref} to $T_{\text{ref}} - \Delta$ in each home, *but only when the home temperature reaches T_{ref}* (Figure C.1).

With this new strategy, the effect on power demand is progressive: for each unit of time the heating system in k homes reaching temperature T_{ref} remains off: aggregate energy demand decreases from $k.H/X$ to zero in H/X hours. After Δ/Y h, the first homes reach $T_{\text{ref}} - \Delta$, the thermostat turns the heating system back on, and aggregate energy demand increases again, before stabilizing to the same level as before we decreased the thermostat temperature.

C.1 Effect of an Increase of T_{ref}

Homes participating in a demand-response program accept only temporary changes to their home's temperature. At some point therefore the thermostat temperature must be increased again from $T_{\text{ref}} - \Delta$ to T_{ref} .

In order to avoid creating any synchronizing event, a possible strategy is to change the thermostat temperature for all homes that reach $T_{\text{ref}} + H - \Delta$ (Figure C.2).

If $\Delta/X < H/Y$ (case of (Figure C.1)), the number of active heating systems increases from $k.H/X$ to $k.(H + \Delta)/X$ in Δ/X hours: energy demand increases to $P(H + \Delta)/H$, then remains constant, above average, for $H/Y - \Delta/X$ hours, then returns to its average value in Δ/X hours.

If $\Delta/X > H/Y$, the number of active heating systems increases from $k.H/X$ to $k.(H/X + H/Y)$ in H/Y hours: energy demand increases to P_{max} , then remains constant at the maximum possible value P_{max} , for $\Delta/X - H/Y$ hours, then returns to its average value in H/Y hours.

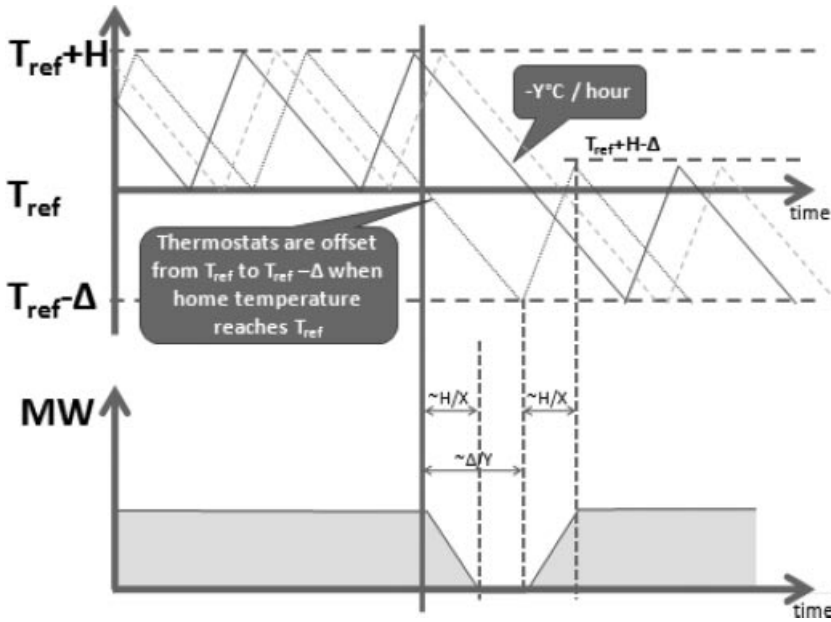


Figure C.1 Decreasing T_{ref} without creating a synchronizing event.

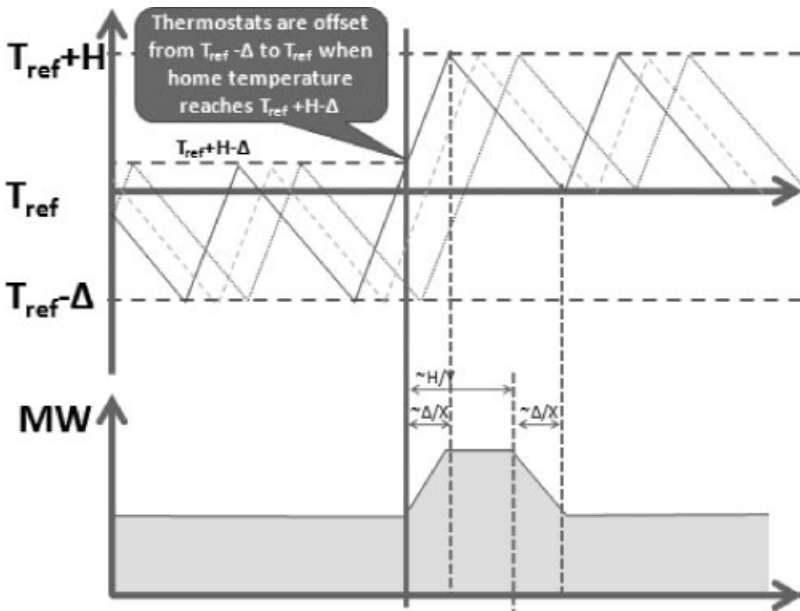


Figure C.2 Adjusting the thermostat back to its normal settings.

Appendix D

Lower Consumption, A Side Benefit of Power Shedding

When the user accepts to live in a home slightly colder than T_{ref} , the thermal dissipation of his home also decreases. The thermal dissipation of the home was $K \times (T_{\text{ref}} - T_{\text{ext}})$ and becomes $K \times (T_{\text{ref}} - T_{\text{ext}} - \Delta)$.

The effect of this lower thermal dissipation is to slightly increase X as the temperature decreases, and to slightly decrease Y (Figure D.1). These effects have been ignored in the evaluations of Appendices B and C.

However, the overall power savings resulting from the lower thermal dissipation during power shedding are easy to evaluate, in a simplified model where we ignore the heating hysteresis H :

- When the temperature is stabilized to $T_{\text{ref}} - \Delta$, the energy used to reheat the houses can be expressed as $P' = P_n \times (T_{\text{ref}} - T_{\text{ext}} - \Delta) / (T_{\text{ref}} - T_{\text{ext}})$: if $T_{\text{ref}} - T_{\text{ext}} = 10^\circ\text{C}$ and $\Delta = 1^\circ\text{C}$, P is 10% lower than P' .
- During the temperature decrease period, the dissipation power decreases quasilinearly (we approximate the exponential evolution of the home temperature by a linear function) from P_n to P' . If C is the thermal capacity of the house, k the thermal dissipation, t the duration of the house cooling period (heating off) if we neglect the variation of T_{ref} , t' the duration of the house cooling period (heating off) if we take the variation of T_{ref} in consideration we can write:

$$C \times \Delta - k(T_{\text{ref}} - T_{\text{ext}})t = 0$$

$$C \times \Delta - k \int (T_{\text{home}} - T_{\text{ext}})dt' = 0 \leftrightarrow C \times \Delta - k(T_{\text{ref}} - T_{\text{ext}} - \Delta/2)t' \\ = 0 \leftrightarrow t' = t \times (T_{\text{ref}} - T_{\text{ext}}) / (T_{\text{ref}} - T_{\text{ext}} - \Delta/2)$$

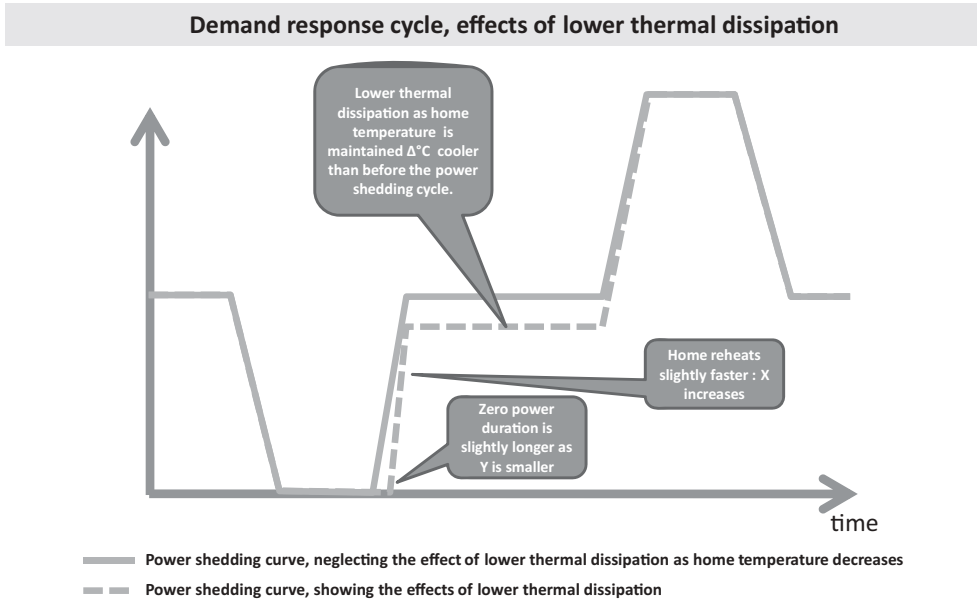


Figure D.1 Effects of the lower thermal dissipation on the power shedding cycle.

Therefore, with the same values of $(T_{ref} - T_{ext})$ and Δ as above, we can estimate the real house cooling period with the heating off to be about 5% longer than in our estimates of Appendices B and C.

- During the temperature increase period, the dissipation power increases linearly from P' to P_n . If P_{max} is the power of the heating system we can write:
 - $P_{max}t - k(T_{ref} - T_{ext})t = C \times \Delta$ (energy conservation as the house reheats).
 - Mean heating power = $P = P_{max}Y/(X + Y) = k(T_{ref} - T_{out})$.
 - $-C \times \Delta + P_{max}t' - k \int (T_{home} - T_{ext})dt' = 0 \Leftrightarrow -C^* \Delta + P_{max}t' - k(T_{ref} - T_{ext} - \Delta/2)t' = 0$.
 - By combining the above we obtain $t' = t[(T_{ref} - T_{ext})X/Y]/[(T_{ref} - T_{ext})X/Y + \Delta/2]$.
 - With the same values of $T_{ref} - T_{ext}$ and Δ as above, $X = 1,6 \text{ }^\circ/\text{h}$ and $Y = 0,6 \text{ }^\circ/\text{h}$ we see that the real heating time is about 2% shorter than what we obtain in Appendices B and C, when taking into account the variation of the house thermal dissipation during the reheat cycle. We obviously have the same ratio regarding the reheat energy, we use $t'P_{max}$ instead of tP_{max} .

In summary, the impact of the variation of the house thermal dissipation during the reheat cycle can be summarized as follows:

- The evaluations of Appendices B and C for the energy made available to the grid during the power shedding cycle are undervalued by about 5%. The undervaluation becomes more important as $T_{\text{ref}} - T_{\text{ext}}$ decreases (mild temperatures).
- The impact on the reheat cycle is negligible.
- The whole cooling/reheat cycle is not completely neutral for the home owner. He gains about 5% of the energy made available to the network as a result of the imbalance between the energy savings of the cooling period and the extra energy used during the reheat cycle. In addition, his bill is reduced by about 10% as long as the temperature is maintained 1 °C colder.

Index

- 2005/32/EC, 32–3
- 6LoWPAN, 195–208
- 802.15.4, 3–4, 93
 - Ack frame, 13
 - Active scan, 11–12
 - Association request, 9, 12
 - Auxiliary control header, 15
 - Beacon-enabled access method, 10–11
 - CAP, 10–11
 - CBC MAC, 14
 - CCM*, 14–15
 - CFP, 10–11
 - Channels, 4, 6
 - Clear Channel Assessment (CCA), 8
 - Cluster, 9–10
 - Command frames, 13
 - Contention Access Period, 10–11
 - Contention Free Period, 10–11
 - Coordinator, 9–12
 - CSMA/CA, 3, 8, 10–12
 - Data frames, 13
 - Dwell time, 6
 - Energy detection (ED), 8
 - EUI-64, 13
 - Extended unique identifier, 13
 - Extension identifier, 13
 - FFD (full function device), 9, 12
 - Frame counter, 15
 - Full function device, 9, 12
 - ISM Band, 4, 6
 - Key identifier field, 15
 - Key lookup, 16
 - Link Quality Information, 7–8
 - LLC layer, 3
 - MAC layer, 3–4, 8–16
 - Non beacon-enabled access method, 11–12
 - Organizationally unique identifier, 13
 - OUI, 13
 - Page number, 6
 - PAN coordinator, 9–12, 20
 - PAN ID, 9–13
 - Physical header (PHR), 8
 - Physical layer, 3–8, 16–17
 - Physical Service Data Unit (PSDU), 8
 - Reduced Function Device (RFD), 9
 - RFD (Reduced Function Device), 9
 - Security, 14, 40
 - Security control field, 15
 - Security mode, 15
 - Slotted CSMA/CA, 10–11
 - Superframe, 10–11
 - Synchronous header (SHR), 8
 - U-NII Band, 4
 - Unslotted CSMA/CA, 11–12

- Access Right resource, 248
- Application association, 186–7
- Application Interworking Specification (AIS), 62
- Application Point of Contact, 244–8
- Area, 3, 87
- Area identifier, 87
- Association IC, 184–6
- Application Interworking Specification (AIS), 62

- Backbone Controller, 84
- BACnet, 45–6
 - Algorithmic reporting, 53
 - BACnet/WS, 59
 - COV reporting, 53
 - Network header (NPCI), 53, 59
 - Object, 4
 - Object properties, 53, 59
 - Object types, 60, 257
 - Object_Identifier, 49, 51–2
 - Services, 6
 - Who-Has / I-Have services, 55
- Balance responsible entities, 279, 283–5
- Base demand, 272
- Batibus, 46, 83
- B-Band, 27, 31–2
- Bidding mechanism, 281–3
 - Financial flows, 284, 296, 299
- Blue Book, 179–80, 184–6

- C.12 interworking, 262–4
- C.12 procedures table, 167–8
- C.12 PSEM, 168–71
- C.12 read service, 167
- C.12 write service, 167
- C.12, 165–6
- C.12.18, 165–6, 168–9
- C.12.19 decades, 166–8
- C.12.19 tables, 166–8
- C.12.19, 165–8
- C.12.21, 165–6, 169–71

- C.12.22, 165–6, 171–6
- CAN bus, 319–20
- Carrier sense multiple access (CSMA), 64
- C-Band, 27, 31–2
- CENELEC 50 065-1, 27, 31–2, 35–6
- CENELEC A-Band, 26–7, 31–2
- CENELEC B-Band, 27, 31–2
- CENELEC C-Band, 27, 31–2
- CENELEC D-Band, 27, 31–2
- CHAdEMO, 307–21
- Coexistence, 35–6
- Configuration, 9, 53, 68, 72, 92
- Container resource, 248–50
- COSEM IC, 181–2
- COSEM Interface class, 181–2, 184–6
- COSEM logical device, 181–2
- COSEM object, 181–2
- COSEM, 179–82

- D8QSK, 26–7
- DAO, 204, 206–7
- Data storage IC, 184–6
- Datapoints, 91–2
- D-Band, 27, 31–2
- DBPSK, 26–7
- DC charging, 307–21, 310–17
- Demand profile, 273, 276–7, 285, 297, 301
- Demand shedding, 285–6, 297, 302
- Device number, 87–8
- dIa, 238–9
- Differential Manchester encoding, 63
- DIO, 204–6
- DIS, 204
- DLMS, 26, 179–81
- DLMS/COSEM interworking, 264–6
- DNP 3.0, 79
- DODAG, 202–6
- DODAG repair, 204–6
- DQPSK, 26–7

- EHS, 46, 83
- EIB, 46, 83–4, 86

- EIBA, 83
- E-mode, 92
- EN13757, 160
 - EN13757-1:2002, 155
 - EN13757-2:2004, 155
 - EN13757-3 2004, 155
 - EN13757-4:2005, 155, 160, 162
 - EN13757-5:2007, 155
 - EN13757-6:2007, 155
- EN 13757, 160
- EN 1434, 155
- EN 65065-1, 24
- ETSI M2M, 237–8
- ETSI M2M architecture, 238–9
- EUI-64, 198–200
- European Directive 2005/32/EC, 32–3
- European Home Systems Association, 83
- European Installation Bus Association, 83
- EXI, 242–52

- FastInfoSet, 242–52
- FCC, 4
- FCC Part15, 27, 31–2
- Fragment header, 198
- French national consumption, 277
- Functional block, 89–90

- G1-PLC, 26
- G3-PLC, 26–7
- GA, 238–9
- Green Book, 179–80
- Group address, 66, 86, 88, 90, 95
- Group Object, 90
- Group resource, 250

- HC1, 198–200
- HC2, 198–200
- Header compression, 196–200
- Heating control, 304
- Homeplug AV, 24–5
- Homeplug AV2, 24–5
- Homeplug GreenPhy, 24–5

- Homeplug v1.0, 24–5
- Homeplug, 24–7

- IEC 15 118, 307–21, 313–17
- IEC 60870-2-1:1992, 162
- IEC 61 334, 26
- IEC 61 681:2001, 307–21
- IEC 61 851, 310–17, 311–13
- IEC 61 851:2010, 307–21
- IEC 62 056-42, 180–81
- IEC 62 056-46, 180–81
- IEC 62 056-47, 180–81
- IEC 62 056-53, 180–81
- IEC 62 056-61, 180–81
- IEC 62 056-62, 180–81, 184–6
- IEC 62 196.1, 307–21
- IEC 62 196.2, 307–21
- IEC 62 196.3, 307–21
- IEC 870-5-1, 155
- IEC 870-5-1, 155
- IEC 870-5-101, 79
- IEC 870-5-2, 155
- IEC 870-5-2, 155
- IEC TC.13, 180–81
- IEEE 1901, 35–6
- IEEE 1901.2, 35–6
- IID, 198–200
- Inputs, 89–90
- Interface objects, 90–91, 186–91
- Interworking proxy, 258–62
- IPHC, 200–202
- IPv6 compression header, 198–200
- ISM Band, 4, 6
- U-NII Band, 4
- ISO/IEC 14 908–3, 25
- ISO/IEC 14908, *see* LonWorks

- J.2847, 320–21
- JSON, 242–52

- KNX over RF, 86–7
- KNX/IP, 87

- KNXnet/IP, 83, 87
- KonCert group, 84

- L bit, 198–200
- LDN, 181–2
- Line couplers, 84, 88
- Line identifier, 87
- Lines, 84–5
- LOAD, 195–6
- Logical tag, 92
- Long-polling, 251–2
- Lonmark, *see* LonWorks
- LonWorks, 25, 60–61
 - EIA-709.1, 62
 - Authentication, 67
 - Backlog, 64–5
 - Beta-2 slots, 64
 - Channel, 63–4
 - Connections, 69
 - Delta backlog, 65
 - Domain, 65–7
 - FO-20, 63
 - Group, 66
 - IP-852, 63
 - LonMark, 61, 70–71
 - Media access control (MAC), 64
 - Network variable index, 68–70
 - Network Variable Messages, 68–9
 - Network variable selector, 68–70
 - Neuronid, 65
 - Node, 65
 - PL-20, 63
 - Pri bit, 65
 - Segment, 64
 - Standard Network Variables Type (SNVT), 70–71
 - Subnet, 65–6
 - TP/FT-10, 63
 - TP/RS485-39, 63
 - TP/XF-1250, 63
 - User network variable types (UNVTs), 71s

- LonTalk, *see* LonWorks
- LowPan_NHC header, 200–202

- Marginal production cost, 272
- M-Bus, 92, 155–63
 - Data Information Block (DIB), 158
 - FT 1.2, 156
 - Primary addresses, 157–8
 - Secondary address, 158
 - Security, 163
 - Segments, 157
 - Selected state, 157–8
 - Value Information Block (VIB), 158
 - Value Information Field (VIF), 158–9
 - Variable data block, 158
 - Wired link layer, 156, 160, 162
 - Wired physical layer, 160
 - Zones, 157
- Mesh header, 197–8
- Mesh network, 93, 98–9
- Mesh under, 195–6
- mIa, 238–9
- ModBus/TCP, 80–82
- Mode 1 charging, 310–17
- Mode 2 charging, 310–17
- Mode 3 charging, 310–17
- Mode 4 charging, 310–17

- NA, 238–9

- OBIS code, 182–4
- Objective function, 202–6
- OUI, 198–200
- Outputs, 89–90

- PAN-ID, 198–200
- Parameters, 89–91
- Peak demand, 273, 300–303
- Physical address, 88, 91
- Pilot wire, 311–13
- PL110, 86
- PLC, 23

- p-persistent CSMA, 64
- Primary adjustment mechanism, 280
- PRIME, 26–7
- Priority levels, 88
- Profibus, 79
- PSEM, 168–9

- Repeaters, 48, 84, 140, 142, 144–5, 157
- RFC 2460, 200–202
- RFC 2474, 200–202
- RFC 3168, 200–202
- RFC 3260, 200–202
- RFC 3697, 200–202
- RFC 4291, 198–200
- RFC 4919, 195–6
- RFC 4944, 195–6, 198–200
- RFC 6142, 176–7
- RFC 6206, 202–6
- ROLL, 33–5, 195–6, 202–6
- Route over, 195–6
- Routing counter, 88
- RPL, 33–5, 195–208, 202–6
- RPL header, 207–8
- RS 232, 80
- RS 442, 80
- RS 485, 47, 63, 80
- RTE (Réseau Transport Electricité), 277–85

- SAE J1772, 307–21, 317–18
- SCL, 244
- S-FSK, 26
- Slave, 80–82
- Smart grid, 271–306
- S-mode, 92, 160
- Standby power, 32–3
- Subscription and Notification Channel resource, 251–2

- Telegram, 87–9, 92
- Tertiary adjustment mechanism, 281

- Thermal preconditioning, 323–4
- Time and event-bound IC, 184–6
- TP1, 86, 88
- TR 102.966, 255–66
- Transmission system operators (TSO), 274–7
- Transport layer, 66–7, 89, 187–8, 213–14
- TS 102-690, 238–9
- Type 1 connector, 307–21
- Type 2 connector, 307–21
- Type 3 connector, 307–21

- U bit, 198–200

- White Book, 179–80
- Wireless M-Bus, 160–63
 - 169MHz, 160
 - 868MHz, 160
 - Security, 163
 - Transmission modes, 160–61
- wM-Bus, 155, 161

- X.10, 27
- X.891, 242–52
- X10, 27, 140
- xDLMS, 179–80

- Yellow Book, 179–80

- Z-wave, 139
 - Always listening nodes, 140
 - Association Set command, 149–50
 - Associations, 149–50
 - Auto inclusion, 144
 - Basic Command class, 148–9
 - Basic Device Class, 147
 - BASIC_GET command, 145, 149–50
 - BASIC_SET command, 145, 149–50
 - Battery powered nodes, 145
 - Bridge Controller, 141
 - Broadcast frame, 143

- Z-wave (*Continued*)
 - Collision avoidance, 143
 - Command class, 147–50
 - Command Configuration Set, 150
 - Command record, 150
 - COMMAND_CLASS_
 - ASSOCIATION, 149–50
 - COMMAND_CLASS_
 - ASSOCIATION_COMMAND_
 - CONFIGURATION, 150
 - Commands, 148–9
 - Controllers, 141
 - Device Class, 147
 - Enhanced Slave, 142
 - Explorer frame, 144
 - FLiRS, 142
 - Frame format, 143
 - Frequencies, 142
 - Frequently Listening Routing Slave, 142
 - Generic Device class, 147
 - Get node information command, 148
 - Grouping ID, 149–50
 - Home ID, 141, 146–7
 - Inclusion, 147–8
 - Inclusion Controllers, 141
 - Key Set security encapsulated
 - command, 151
 - MAC layer, 142–3
 - Multi Instance Association Set
 - command, 149–50
 - Multi Instance Command
 - Encapsulation command, 149
 - Multicast frame, 143
 - Multi-instance devices, 149
 - Network wide inclusion, 144
 - NIF, 147, 151
 - Node ID, 141, 146–7
 - Node information frame, 147, 151
 - Output power, 142
 - Portable Controller, 141
 - Power down mode, 140
 - Primary controller, 141, 146–7
 - Rediscovery procedure, 148
 - Response route, 142
 - Return route, 145
 - RF layer, 142–3
 - Route resolution, 144
 - Routed Singlecast, 145–6
 - Routing layer, 140, 145–8
 - Routing slave, 142
 - Scene Activation command class, 150
 - Scene Actuator Configuration
 - command class, 150
 - SCENE_ACTIVATION_SET
 - command, 150
 - Scenes, 150
 - SearchReply, 146
 - SearchRequest, 144–5
 - SearchResult, 144
 - SearchStop, 144, 146
 - Secondary controller, 141, 146–7
 - Security, 150–51
 - Security Command Class, 150
 - Security encapsulated secure command, 151
 - Singlecast frames, 143
 - SIS, 141
 - Slave, 141–2
 - Specific Device class, 147
 - Static Controller, 141
 - Static Update Controller, 141
 - SUC, 141
 - SUC ID Server, 141
 - Transfer layer, 143–5
 - Wake Up Interval Set command, 145
 - Wake Up No More Information, 145
 - Wake Up Notification command, 145
 - Wake-up, 140
 - Wake-Up Beam, 145
 - Wake-Up Command Class, 145
 - ZensorNet™, 142, 145
 - Z-wave alliance, 139

- ZigBee
 - 2006/2007, 94, 105
 - Address assignment, 100
 - Adopter status, 93
 - Advanced Ad-Hoc On-Demand
 - distance Vectoring, 103
 - AODV, 99, 103
 - Application profile, 95, 97, 101, 116–17, 119–29
 - Application support sublayer (APS), 94
 - Application Support Sub-Layer
 - Management Entity (APSME), 95
 - APS address map, 106
 - APS frame format, 108–9
 - APS layer, 94, 102, 105–9
 - APS security, 114
 - APSME (Application Support Sub-Layer Management Entity), 95
 - Association Request, 97
 - Association Response, 97–8
 - Attribute, 117
 - Authentication, 115
 - AUX header, 113
 - Beacon, 96–9
 - Beacon Request, 96–9
 - Binding, 69, 94
 - Binding table, 94, 107–8
 - Broadcast, 101–3
 - Broadcast Transaction Table, 101–2, 120
 - BTT, 101–2, 120
 - Child node, 97
 - Cluster, 116
 - Cluster Library, 95, 108, 116–19
 - Command, 117
 - Commissioning ZDP Cluster, 99
 - Coordinator, 96–100
 - Coordinator (ZC), 96
 - CSkip, 100, 105
 - Demande response and Load control
 - cluster, 124–6
 - Developers conference, 93
 - Device Depth, 97
 - Device ID, 106, 117, 119
 - Device Object, 95, 109, 112
 - Device Profile, 95, 106, 108–9, 130
 - Devices, 97–8, 106, 108–10
 - Dimmable light, 121
 - Dimmer Switch, 121
 - End-Device (ZED), 96
 - Endpoint, 106
 - Energy Service Portal (ESP), 123–4
 - EPID, 97
 - ESP (Energy Service Portal), 123–4
 - Extended PAN ID, 97, 99
 - Golden units, 93
 - Groups, 107
 - HA application profile, 119–22
 - Heating / cooling Unit, 122
 - High security mode, 114–15, 123
 - Home Automation, 119–22
 - In premise display service, 124
 - Indirect addressing, 107–8
 - Interop events, 93
 - Key load key, 114
 - Key transport key, 114
 - Keys, 113–14
 - Light Sensor, 121
 - Link keys, 114
 - Load Control Device, 124
 - Logo, 93
 - Low security mode, 114
 - Mains power outlet, 121
 - Manufacturer specific profiles, 108, 119
 - Master keys, 113
 - Messaging Cluster, 129
 - Metering Device, 124
 - Multicast, 101–3
 - Network keys, 114
 - Network layer, 99–105
 - Network management, 110

ZigBee (Continued)

- Node discovery, 95
- NWK frame format, 100
- NWK level security, 112
- Nwksecuritylevel, 98
- ON/OFF light, 121
- PAN ID, 94, 97, 99, 101
- Participant status, 93
- PCT (Programmable Communicating Thermostat), 124
- Permissions table, 116
- Permit joining, 98
- Price cluster, 128–9
- Pro, 8, 94, 103
- Profile 0x01, 94, 99–100, 120
- Profile 0x02, 94, 99–100
- Programmable Communicating Thermostat (PCT), 124
- Promoter status, 93
- Public application profile, 108, 111, 116–17, 119–22
- Radius, 101–3
- Range extender, 121
- Range Extender Device, 124
- Route discovery, 103, 107
- Router (ZR), 96
- Routing commands, 101–2
- Security, 111–16
- Security levels, 112
- Shade, 121
- Shade Controller, 121
- Short addresses, 99–100
- Simple descriptor, 106, 116–17
- Simple metering cluster, 126–8
- SKKE, 114–15
- Smart Appliance Device, 124
- Smart Energy, 123
- Smart Energy 1.0, 122–9
- Source routing, 105
- Stub APS, 122
- Temperature Sensor, 122
- Thermostat, 122
- Tree based routing, 105
- Trust center, 96, 114–15
- ZCL (ZigBee Cluster Library), 116–19
- ZCL frame format, 118
- ZCL general commands, 118–19
- ZCP certification, 93s
- ZDO (ZigBee Device Object), 95, 109–11
- ZDO permissions table, 116
- ZDP (ZigBee Device Profile), 95, 109–11
- ZED, 96
- ZigBee alliance, 93–4, 119
- ZigBee cluster Library (ZCL), 116–19
- ZigBee Compliant Platform (ZCP) certification, 93
- ZigBee interworking, 258–62
- ZigBee SEP 2.0, 307–21

COURSE OBJECTIVE

➤ To help students develop knowledge and competence in ethical management and decision making in organizational contexts.

UNIT I ETHICS AND SOCIETY 9

Ethical Management- Definition, Motivation, Advantages-Practical implications of ethical management. Managerial ethics, professional ethics, and social Responsibility-Role of culture and society's expectations- Individual and organizational responsibility to society and the community.

UNIT II ETHICAL DECISION MAKING AND MANAGEMENT IN A CRISIS 9

Managing in an ethical crisis, the nature of a crisis, ethics in crisis management, discuss case studies, analyze real-world scenarios, develop ethical management skills, knowledge, and competencies. Proactive crisis management.

UNIT III STAKEHOLDERS IN ETHICAL MANAGEMENT 9

Stakeholders in ethical management, identifying internal and external stakeholders, nature of stakeholders, ethical management of various kinds of stakeholders: customers (product and service issues), employees (leadership, fairness, justice, diversity) suppliers, collaborators, business, community, the natural environment (the sustainability imperative, green management, Contemporary issues).

UNIT IV INDIVIDUAL VARIABLES IN ETHICAL MANAGEMENT 9

Understanding individual variables in ethics, managerial ethics, concepts in ethical psychology-ethical awareness, ethical courage, ethical judgment, ethical foundations, ethical emotions/intuitions/intensity. Utilization of these concepts and competencies for ethical decision-making and management.

UNIT V PRACTICAL FIELD-GUIDE, TECHNIQUES AND SKILLS 9

Ethical management in practice, development of techniques and skills, navigating challenges and dilemmas, resolving issues and preventing unethical management proactively. Role modelling and creating a culture of ethical management and human flourishing.

TOTAL: 45 PERIODS

COURSE OUTCOMES

- CO1: Role modelling and influencing the ethical and cultural context.
- CO2: Respond to ethical crises and proactively address potential crises situations.
- CO3: Understand and implement stakeholder management decisions.
- CO4: Develop the ability, knowledge, and skills for ethical management.
- CO5: Develop practical skills to navigate, resolve and thrive in management situations

REFERENCES

1. Brad Agle, Aaron Miller, Bill O' Rourke, The Business Ethics Field Guide: the essential companion to leading your career and your company, 2016.
2. Steiner & Steiner, Business, Government & Society: A managerial Perspective, 2011.
3. Lawrence & Weber, Business and Society: Stakeholders, Ethics, Public Policy, 2020.

MAPPING OF POs AND COs

	PO1	PO2	PO3	PO4	PO5	PO6
CO1	3	3	2	3	2	3
CO2		3	2	3	1	3
CO3	3	3	3	3	2	3
CO4	3	3	3	2	1	3
CO5	3	3	3	2	2	3

ETHICAL MANAGEMENT NOTES (OBA434)

UNIT 1 :

Ethical Management- Definition, Motivation:

Ethical Management is the branch of philosophy that studies the values and behaviour of a person. Value study of a person is used to determine his positive and negative attitude towards life. Ethics studies concepts like good and evil, responsibility and right and wrong. Ethics can be distinguished in three categories: normative ethics, descriptive ethics and metaethics. Metaethics focuses on the issues of universal truths, ethical judgements and the meaning of ethical terms. Normative ethics can be used to regulate the right and wrong behaviour of individuals. Descriptive ethics, also called applied ethics, is used to consider controversial issues, such as abortion, animal rights, capital punishment and nuclear war.

Meaning of Ethical Management

The term 'ethics' defines the standards that bear on right and wrong issues of society. Business ethics is thus a set of professional standards, which emphasize principles of honesty and duty to the business and the general public. The other significant principles included in business ethics are:

- Fairness
- Integrity
- Commitment to agreements
- Broad-mindedness
- Considerateness
- Importance given to human esteem and self-respect
- Responsible citizenship
- Attempt to excel
- Accountability

These principles, if strictly pursued, lead to a decent business environment and create healthy relationships in the organization. However, deviations from these principles can occur due to the following factors:

- Ignorance and indifference to issues
- Selfishness
- Imperfect reasoning

Advantages- of ethical management.:

- **ethical management** helps in improving society by establishing government agencies, unions, laws and regulations in the society.
- **ethical management** helps an organization maintain ethical values during times of crisis. Business ethics programmes guide leaders about the right or wrong ways of dealing with complex dilemmas and how they should act during that time.
- **ethical management** helps employees behave according to the ethical values that are preferred by the top management of an organization. An organization discovers many differences between the values that reflect in the actions of the employees and the values preferred. Employees experience a relationship that is strong between the values of the organization and their values. Ethical values induce teamwork and increase the efficiency of the employees
- **ethical management** supports employee growth. When an employee pays attention to

ethics, it induces confidence in the employee to deal with reality and face both good and bad circumstances. Bennett, in his article 'Unethical Behaviour, Stress Appear Linked', explained that the more an employee is emotionally healthy, the more ethical he is.

- **ethical management** have become legal instruments. These days, there are several lawsuits regarding personnel matters and the influence of the services of the organization on the investors and customers. Major ethical principles that are applied in the organization are the laws that are made by the government. A greater attention on ethical issues on the part of the government ensures high ethical procedures and policies in the workplace. An employee, for example, is subject to breach of contract on non-compliance of the terms and conditions of the contract.
- **ethical management** helps to avoid criminal acts of 'omission' and it also helps in lowering the fines. Ethics helps in ascertaining the violation of ethical issues and helps in rectifying the violation that is committed by the organization. The guidelines set by an organization about ethical values helps to lower fines. An organization, for example, that has knowingly violated a contract is considered to have committed a criminal act and the organization is subject to penalty.
- **ethical management** helps to identify and manage the values associated with quality management, strategic management and diversity management. For managing these values, ethical programmes record the values, develop policies and procedures and then provide training to the employees on these policies and procedures. These ethical programmes manage certain values of quality management, such as reliability, performance, measurement and feedback. Similarly, these programmes also manage various strategic values, such as reducing cost and increasing market share.
- **ethical management** helps in building a strong and positive public image of an organization. Ethical values enable an organization to increase their goodwill in the market. Those organizations that value their customers have a positive influence in the market. Ethical values are the milestones that enable the establishing of a successful and socially responsible business.
- **ethical management** strengthens organizational culture. Ethical values improve relationships between an organization and its customers. They strengthen the organization by ensuring consistency in the standard and quality of the product.
- **ethical management** makes sure that the right activities are performed in an organization.

Managerial and Professional Ethics:

Ethical subjectivism: This concept emphasizes that the ethical choice of the individual decides the rightness or wrongness of his behaviour.

Ethical relativism: According to this concept, no principle is universally applicable and so it would be inaccurate to measure the behaviour of one society with another's principles or standards. Relativism overlooks the fact that there may be enough evidence to believe that an ethical practice is based on false belief, illogical reasoning, and so on.

Consequentialism: Consequentialism is based on two ideas: the concept of value and the maximization of value. If, for example, honesty is considered a value, an act is considered ethical only if it maximizes this value. An act, which does not maximize the said value, is not ethically permissible.

Deontological ethics: This concept stresses that ethical values can be developed from the concepts of reason as all rational individuals possess the ability to reason. We may, for example, end up causing pain unknowingly while trying to create happiness. Therefore, the ethical value of an action cannot be determined by its consequences. Instead, it is in the motive that lies behind the particular action.

Ethics of virtue: This concept emphasizes those traits that give the individual a sense of satisfaction from ethical point of view. Virtuous acts like courage, honesty, tolerance and generosity are done as a way of living and not by chance.

Whistle blowing: Whistle blowing refers to the attempt of an employee to disclose what he or she believes to be illegal behaviour in or by the organization. From one point of view, this seems to deceive the principle of honesty in business ethics, as it is taken for granted that the employees of an organization need to be loyal to its workings. However, when loyalty to one's organization in particular is perceived to be harming one's general loyalty to mankind, the act of whistle blowing is justified. Failure on the part of the management of the organization to fulfil its social obligations calls for whistle blowing. It is the responsibility of the whistle blower to be careful about revealing the organization's secrets and to consider the harm it may cause to his colleagues and shareholders. The steps that should be taken into consideration by the whistle blower are:

- Ascertain the gravity of the situation before whistle blowing
- Scrutinize the purpose
- Authenticate and keep a record of the concerned information
- Determine the type of offence and to whom it should be reported
- Assert your claim in a proper way
- Stick to the facts
- Determine if the whistle blowing need be external or internal
- Decide if the whistle blowing should be anonymous or otherwise
- Make sure to follow proper rules in reporting the offence
- Consult a lawyer (if required)
- Anticipate and document vengeance

Organizational ethics and responsibility to society and the community:

Organizational ethics is used to consider the issues of morality and rationality in organizations. Organizational ethics is completely different from management ethics. Management ethics focuses on the ethical quality of the decisions and actions taken by managers of an organization. Thus, management ethics deals with the individuals in the organization and organizational ethics deals with all the activities of an organization. Therefore, organizational ethics is collective in scope. Organizational ethical issues can be handled at three levels. These levels are:

- Corporate mission
- Constituency relations
- Policies and practices

Organizational ethics can also be used to evaluate the policies and practices of the organizations. Public commitment to ethical principles can give way to business and administrative practices.

Organizational ethics also depends on the type of the organization. Organizations can be classified by considering their economic and ethical concerns. Organizations can be classified into four types. These are:

Exploitative: Organizations with low economic and ethical concerns are called exploitative organizations. These organizations utilize child labour and use rivers for dumping wastes to maximize their profits.

Manipulative: Organizations with high economic performance concerns and low ethical concerns are called manipulative organizations. These organizations use tax laws, labour laws and union leaders to maximize profit.

Holistic: Organizations with high ethical concerns and low economic concerns are called holistic organizations. These organizations spend their money in social and environmental purposes.

Balanced: Balanced organizations have high ethical and economic concerns. These types of organizations gain profit as well as work for social and environmental purposes.

Practical Implications of Ethical Management:

- Employee rights
- Ethical business conducts
- Environmental protection
- Child labour in business
- Discrepancies in the wages of women employees
- Bonded labour
- Exploitation of unorganized labour
- Minimum wages
- Obligations of large and multinational corporations

Individual Roles and Responsibilities in Ethics Management to society:

Each individual in the organization should be provided a specific role in managing ethics in the organization. However, the role assigned to each individual depends on the size and nature of the organization. The roles can also be full-time or part-time. The following responsibilities can be assigned in ethics management:

The chief executive officer of the organization must support the ethics management programme.

A committee or group should be developed in the organization to control the development and operation of the ethics management programme.

A committee or group should be developed which should be responsible for training the employees on the policies and procedures of the ethics management programme and should resolve any ethical dilemmas that may arise. This committee can contain senior officers.

A person in the organization should be designated as ombudsperson who has the responsibility of investigating or resolving complaints from the employees of the organization against the ethics management programme.

Each person of the organization is responsible for the implementation of the ethics management programme.

Professional Ethics:

Professional ethics maintains moral values and ensures that the behaviour of employees is aligned with these values. Still, there are certain myths regarding business ethics, which are as follows:

Professional ethics is a matter of belief than management.

Professional ethics constitutes the principles propounded by philosophers and theologians. Many people are of the opinion that business ethics is a religion or a theoretical debate. In the day-to-day issues of the organization, business ethics has very little to contribute. However, ethics is a management discipline that requires a planned approach and several management programmes.

Professional ethics only states the obvious do-good situations. Several people are of the opinion that ethics represents the values that a person should naturally aspire to have and, therefore, establishing codes of ethics is unnecessary. However, importance should be given to the ethical values of an organization. For instance, it is understood that everybody should be honest. If the workers of an organization are not honest then the code of ethics of that organization should have honesty listed in it. Code of ethics changes with the change in the society and the needs of an organization.

Professional ethics is an opinion. Many people believe that stress and confusion may inspire good people to behave unethically. Ethics in an organization can be managed by helping each other to stay ethical and to work together through confusing and stressful ethical dilemmas.

Professional ethics is the new trend. Many people believe that business ethics is a recent phenomenon and has recently gained attention. However, it is an old phenomenon that has received importance now.

Professional ethics is unmanageable by an organization. Actually, ethics is not directly 'managed' by an organization, but the behaviour of the team leader has a strong moral influence on the employees. The objectives of an organization, such as maximizing profit and minimizing costs also have a strong impact on the ethics of an organization. Even the laws, regulations and rules have a good impact on the ethical values of the employees and hence minimize the harm to the business. But still, some believe that business ethics cannot be managed by an organization.

Social Responsibility-Role of culture and society's expectations:

Social responsibility of business involves the consideration of general public interest by businessmen while taking business decisions and actions.

The concept of social responsibility has emerged due to several economic, social, political and legal influences. These forces, which have obliged, persuaded and helped businessmen to become aware of their responsibility to society, are as follows:

Public opinion: Public interference with the help of the government has instilled a fear in the heart of businessmen. The threat of public regulation and public ownership has compelled them to acknowledge the fact that responsible behaviour is essential on their part for survival in the private sector.

Trade union movement: The recent development of socialism that boosted the strength of labour unions has forced businessmen to give a fair share to workers. Human relations and labour legislation have facilitated trade unions to increase their influence.

Consumerism: Consumer organizations have encouraged awareness about consumer rights. Consequently, businesses have become more responsive to consumer needs and stress the dictum of 'consumer is the king'. Businessmen can no longer adopt the approach of 'let the buyer beware'.

Education: Extensive education has made businessmen conscious about the quality of life, moral values and social standards. Liberal business leaders have been pressing the business community to acknowledge its social obligations.

Public relations: Modern businessmen are aware that a good public image contributes to their growth. There is a greater alertness in their hearts that business is a construction of society and hence, it should consider and react positively to the expectations of society.

Managerial revolution: Separation of ownership from control in large corporations has resulted in professionalism in management. A professional manager is fairly aware of the society's expectations and attempts to meet the demands of all social components, like customers, employees, shareholders and the government, in a well-adjusted manner.

Role of culture and society's expectations Culture can be considered as a constellation of factors that are learned through our interaction with the environment and during our developmental and growth years. A growing baby learns a basic set of values, ideas, perceptions, preferences, concept of morality, code of conduct, and so on, through family and cultural socialization and such prevailing culture with which the member of the family is associated determines many of the responses that an individual makes in a given situation.

The organizational culture is a system of shared beliefs and attitudes that develop within an organization and guides the behaviour of its members. It is also known as 'corporate culture', and has a major impact on the performance of organizations, especially on the quality of work life experienced by the employees at all levels of the organizational hierarchy. The corporate culture 'consists of the norms, values and unwritten rules of conduct of an organization as well as management styles, priorities, beliefs and interpersonal behaviour that prevail. Together, they create a climate that influences the way people communicate, plan and make decisions. Strong corporate values let people know what is expected of them. There are clear guidelines as to how employees are to behave generally within the organization and their expected code of conduct outside the organization. Also, if the employees understand the basic philosophy of the organization, then they are more likely to make decisions that will support these standards set by the organization and reinforce corporate values.

The word 'culture' has been derived metaphorically from the idea of 'cultivation', the process of tilling and developing land. When we talk about culture, we are typically referring to the pattern of development reflected in a society's system of knowledge, ideology, values, laws, social norms and day-to-day rituals. Since the pattern of development differs from society to society, the cultural phenomenon varies according to a given society's stage of development. Accordingly, culture varies from one society to another requiring a study of cross-national and cross-cultural phenomenon within organizations.

While culture has been a continuous development of values and attitudes over many generations, at least the organizational culture can be partially traced back to the values held by the founders of the organization. Such founders were usually dynamic personalities with strong values and a clear vision as to where they wanted to take their organizations. These founders usually selected their associates and their employees who had a similar value system so that these values became an integral part of the organization.

Second, the organizational culture is influenced by the external environment and the interaction between the organization and the external environment. For example, one organization may create a niche for itself for an extremely high quality defect-free product as a result of competitive forces and customer demand, while another organization may opt for moderate quality but lower prices. The work cultures of these two types of organizations would accordingly differ and would be influenced by external forces such as customer demand.

Third, work culture is also a function of the nature of the work and the mission and goals of the organization. For example, in a professional, research-oriented, small organization, the workers may be more informal at all hierarchical levels of the organization, the dress code may not be strictly observed and the employees may be encouraged to be independent and innovative. In contrast, other organizations may have a strictly enforced formal classical hierarchical structure with clearly established channels of communication and strict adherence to work rules.

UNIT-II

Ethical Decision Making:

The management of an organization is responsible for making effective decisions. Managers are responsible for all business operations and they also make all the important decisions. To make decisions ethically correct, various models are considered for the purpose of good decision-making. There are various frameworks of decision-making based on factors such as duty, consequences and virtue. Managerial decision-making involves defining problems and then structuring them for positive results. There are certain steps to be followed during decision-making.

Managers affect the behaviour and decision-making capability of individuals. The individuals in an organization are responsible for conducting business operations. Management is defined as a decision-making process. Ethical decision-making is a method of evaluating and choosing the alternatives decided by ethics management. The following should be kept in mind while making ethical decisions:

Identify and eliminate unethical options in the alternatives.

Identify complex, ambiguous and incomplete facts and try to avoid them.

- Determine the ethical dilemma and resolve it.
- Select the best ethical alternative.

Organizations need to perform a set of activities and take various decisions to achieve organizational goals. These are known as the business strategies of the organization. Business strategies are an important part of businesses, firms and industries. To make a business strategy,

all businesses, firms and industries need to develop a strategic plan once a year. Managers of the firms are given the responsibility to achieve the goals stated in the strategic plan. Business strategies are used for the following purposes:

- They help determine the products and services that an organization needs to provide.
- They help determine the various industries in which the organization competes.
- They help identify the competitors, suppliers and customers of the organization.
- They help analyse the long-term goals of the firm.

Types of Decisions:

Organizational decisions can differ in different ways, which initiates development of different types of decisions from which organizations can choose the appropriate decisions. Organizational decisions are primarily classified on the basis of the purpose of decision-making. Knowledge of outcomes is another approach for classifying organizational decisions. An outcome defines what is going to happen if a particular decision is taken or a particular course of action is taken.

Organizational decisions involve selecting the best alternative from amongst the available alternatives. Organizational decisions are classified into the following categories:

Strategic planning decisions: These are the decisions in which a decision-maker develops objectives and allocates resources for achieving these objectives. Decisions under this category are used for a long period of time and involve a large investment. For example, introducing new products and the acquisition of another organization are strategic planning decisions.

Management control decisions: These are the decisions taken by the management control-level managers, who are at the middle level of the management hierarchy in an organization. These managers deal with the use of resources in the organization. The management control decisions include the analysis of variance, product mix and planning decisions.

Operational control decisions: These are the decisions that deal with the day-to-day problems that affect the operation of an organization. For example, decisions such as production scheduling and inventory control fall under this category. Decisions under this category are taken by the operational-level managers, who are at the bottom-level of the management hierarchy in an organization.

Structured decisions: These are the decisions that are well defined and require application and implementation of some specified procedure or decision rule in order to reach a decision. Such decisions require less time for developing alternatives in the design phase. Structured decisions are made by operating procedures or by using other accepted tools. More modern techniques for making such decisions involve operations research (OR), mathematical analysis, modelling and simulation.

Unstructured decisions: These are the decisions which are not well defined and have no pre-specified procedure or decision rule. These decisions may range from one-time decisions relating to a crisis to decisions relating to recurring problems. The unstructured decisions usually consume much time in the design phase of the decision-making process. These decisions could be solved using judgement and intuition. Modern approaches to such decisions include special data analysis on computers and heuristic techniques. Such decisions are usually handled by strategic planning level managers because of their unstructured nature.

Semi-structured decisions: These are the decisions that are neither structured nor unstructured. These decisions fall somewhere between the structured and unstructured decisions. For example, the introduction of a new product is a semi-structured decision.

Characteristics of Good Decision-making

characteristics of a good decision-making are:

- Decision problems should be grabbed by the management both in space and time. This means, the decision problem should be analysed thoroughly by the management.
- The decision made by the decision-maker should keep him in a state of calm.
- Decisions made by the management should contribute to harmony in the organization.
- Self-interest and self-orientation should not come in the way of decision-making.

Problems in the Decision-making Process

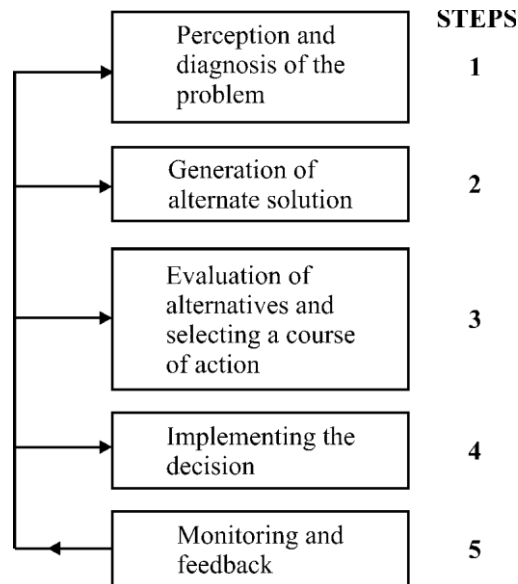
There are various problems faced by a management in the decision-making process. These problems are:

- **Insufficient information:** It refers to the lack of information which affects the performance and quality of the management in an organization.
- **Insufficient knowledge:** It refers to the difference between available knowledge and the required information for the management to take a decision.
- **Lack of time:** It refers to the pressure on the management to make decisions. If time is limited, then the management needs to take hasty decisions.
- **Poor communications:** It leads to the problem that arises due to improper communication of information.

Other limitations of any management in the decision-making process are with respect to the inability of the human mind to handle available knowledge as also human behaviour.

Steps in Decision-making

All decisions involve a series of sequential steps that lead to a particular result. These steps are generally followed to make systematic, objective, analytical and unemotional decisions and some management scholars have called this process a 'rational decision-making process.' Figure 3.1 shows the steps in decision-making.



Factors affecting Decision-making:

Programmed versus non-programmed decisions: As discussed earlier, in the types of problems that managers face, programmed decisions are made in predictable circumstances and managers have clear parameters and criteria. Problems are well structured and alternatives are well defined. The problems are solved and decisions are implemented through established policy directives, rules and procedures.

Information inputs: It is very important to have adequate and accurate information about the situation for decision-making; otherwise, the merit of the decision will suffer. It must be recognized, however that an individual has certain mental constraints, which limit the amount of information that he can adequately handle. Less information is as dangerous as too much information. Some highly authoritative individuals do make decisions on the basis of comparatively less information when compared to more conservative decision-makers.

Prejudice: Prejudice and bias are introduced in our decisions by our perceptual processes and may cause us to make ineffective decisions. First, perception is highly selective, which means that we only accept what we want to accept and hence, only such type of information filters down to our senses. Second, perception is highly subjective, meaning that information gets distorted in order to be consistent with our pre-established beliefs, attitudes and values. For example, a preconceived idea that a given person or an organization is honest or deceptive, good or poor source of information, late or prompt delivery, and so on, can have a considerable effect on the objective ability of the decision-maker and the quality of the decision.

Cognitive constraints: A human brain, which is the source of thinking, creativity and decision-making, is limited in capacity in a number of ways. For example, except for some unique circumstances, our memory is short term, having the capacity of only a few ideas, words and symbols. Second, we cannot perform more than a limited number of calculations in our heads and it is tough to compare all the possible alternatives and make a choice. Finally, psychologically, we are always uncomfortable with making decisions. We are never really sure if our choice of the alternative was correct and optimal until the impact of the implication of the decision has been felt. This makes us feel insecure.

Nature of Crisis and Ethics in Crisis Management :

Crisis problems develop suddenly and are totally unexpected at a given time. These may develop within the general framework of expectations that the management has prepared to some extent to handle these crisis situations. For example, a forest fire will create a crisis problem but the government and the community is generally prepared to fight the forest fire. Similarly, a major strike at the plant may not have been expected, but the management has generally made provisions to handle the situation. Solving crisis problems is reactive in nature and requires reacting quickly and aggressively to solve the problem. It may be achieved through task forces, which may try to mould crisis situations into familiar problems for which the solutions are known to exist.

The opportunity problems are more challenging. These must be exploited for the betterment of the organization, For example, if an opportunity of a highly beneficial merger arises, and the organization fails to recognize the potential, it would be considered a lost opportunity. Similarly, a slightly increased rate of employee absenteeism may mean some deeper organizational problem and if the management does not recognise this opportunity to deal with the problem, this missed opportunity may blow up into a crisis. The central management handles both the crisis problems as well as the opportunity problems.

:

Proactive Crisis:

Proactive Crisis are unique, unpredicted and unprecedented situations. These problems are ambiguous and poorly understood and defy any cut-and-dry solution. These are generally 'one-shot' occurrences for which standard responses are not available and hence, require a creative process of problem solving which is specifically tailored to meet the requirements of the situation at hand. Such problems may involve the closing of a plant, buying or merging into a new company, starting a new business, and so on. Because the Proactive Crisis do not have well-structured solutions, such solutions generally rely upon skill, intuition, creativity, experience and considered judgement and carry with them the consequences of diverse ramifications. The top-level management generally faces these problems because their environment is complex and is involved with high-level policy decisions.

Proactive Crisis, on the other hand, are clearly defined, routine, and repetitive and respond to standardized responses. They are familiar, complete and easily defined and analysed. These problems are generally faced by lower-level and middle-level managers who have, at their disposal, a set of rules, policies and procedures which can be used to solve these problems, so that such problems do not have to be referred to superiors for solutions. For example, if a professor cuts too many classes, the chairperson of the department can use the prescribed rules to discipline him and the issue does not have to be referred to the president of the college. Similarly, if you buy some merchandise and it turns out to be defective, you can take it back for a refund. The management of the company has a well-structured set of rules and procedures to deal with the problem of making refunds for defective merchandise.

Managing Ethical Crisis:

Operating level versus strategic level problems: Operating-level problems are generally well-structured problems encountered by the organization on a daily basis. For example, a newspaper shop owner has the problem of reordering the newspapers and magazines every day and he knows when to order and how much to order. Similarly, daily or weekly production levels, inventory levels or sales levels are set and known and standard solutions exist to solve any problems in these areas when they arise. These situations are not new or unique and do not involve any changes in organizational policies or procedures.

On the other hand, strategic-level problems are unique and demand high-level managerial attention. These problems may involve changes in policies and are important in terms of actions taken or resources committed. While operating-level problems do not affect the survival of an organization, strategic-level problems do. Sometimes, if the operating level problems are left unattended, they may become strategic-level problems. For example, if no action is taken against a professor who habitually miss classes, this may affect other professors thus making it a morale problem for the college, which then would be considered a strategic-level problem.

UNIT-III

Stakeholders:

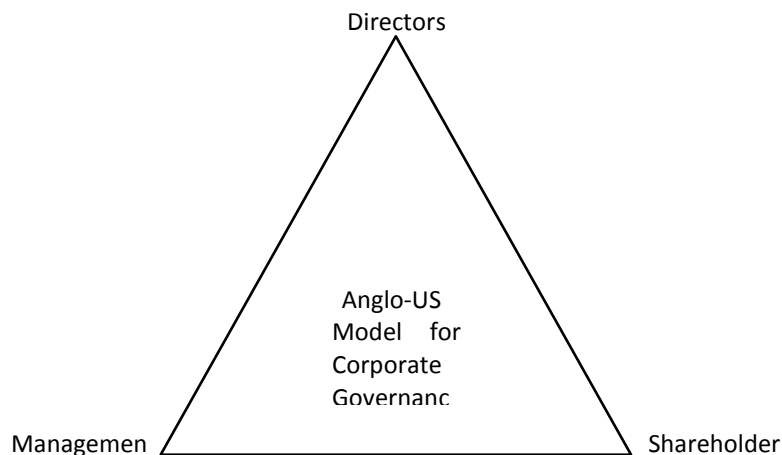
Stakeholders also take care about social causes and the commercial value of society in relation to different interest areas of the stakeholders. The success of different corporations is dependant on the operations of the organizations, which are performed to provide profit to the stakeholders of the corporation. Stakeholders of a corporation are of four types:

Primary social stakeholder: These are those stakeholders who have direct relations with the corporation. The presence of these stakeholders in the corporation can affect the progress of the corporation.

Primary non-social stakeholder: These are also directly related to the corporation, but they are never present in the corporation as primarysocial stakeholders.

Secondary social stakeholder: These are those stakeholders who are not directly related to the corporation but changes in these stakeholders can affect the corporation.

Secondary non-social stakeholder: These are those stakeholders who are not related to the corporation and can rarely affect the corporation



stakeholders of the corporation hold ownership and control of the corporation. The main bank system provides most of the finance forthe company requirements. In other words, it is the largest creditor of the corporation. It provides cross shareholding arrangement to the corporation that make the bank one of the important shareholders of the corporation. The main bank system is responsible for monitoring investment decisions and various clients of the corporation. It is also concerned with the running progress of the corporation. It provides financial help to the corporation. The main bank system is responsible for handling the client of the corporation. It provides the following services to the clients for handling clients of the corporation:

Enabling long-term investment of clients

Securing lender of the corporation

Providing stability to the shareholders of the corporation

Providing solution to the problem of irregularity of information

Managing the financial position by gathering rents from excessivedeposits

. There are two major groups of stakeholders – internal stakeholders and external stakeholders.

Internal Stakeholders:

Internal Stakeholders:

- Internal stakeholders are those people who are actively involved in the activities of a business or own shares in the company. These include owners, employees and investors of a company. External stakeholders are those outside parties that are connected to a company due to their shared interests. Examples of these stakeholders include customers, suppliers, competitors, government, etc.
- Internal stakeholders are the individuals or parties that are directly involved in the management of the business. These stakeholders have a vested interest in the business and hence, they can directly affect or be affected by the successes or failures experienced by the business.
- Internal stakeholders are also known as primary stakeholders. These stakeholders offer services to the organization and are significantly influenced by the outcomes, decisions, and performance of the company. In addition, they are aware of all the internal issues of the company. Internal stakeholders consist of all those who work for the organization, i.e. the employees, the individual or groups who have the ownership of the organization, all those who are involved in the management of the organization, the board of directors and the investors. All of these have a direct stake in the activities in the organization and are critical for the survival of a company.
- Internal stakeholders usually have a significant impact on the operations of an organization. For instance, owners are the ones who take critical business decisions. In addition, the managers and employees are actively involved in the routine operations of a company and make various decisions on a daily basis regarding various business activities.
- **Internal stakeholders are those who have a direct relationship with the business, for example, in terms of ownership, employment or investment. They either own the company's shares or are otherwise actively involved in the affairs of the company.** Due to their vested interest in the business, they are directly impacted by the ups and downs in the business's performance and by its policies & decisions. Internal stakeholders, among others, include the board of directors, senior executives, owners, shareholders, and employees.
- Employees, for example, are impacted a lot in terms of their economic well-being. Employees have a common concern about how much and how frequently they are paid by the company. Management decisions affecting such concerns are extremely significant to these stakeholders. Their financial well-being is directly impacted by how well the company is doing and how profitable it is.
- Further, in this regard, if the business owner decides to provide perks and other compensation packages to employees, it too has an impact on employees and they are affected by such

decisions. As a result, one of the primary concerns of employees is the company's long-term financial viability.

- Similarly, investors and shareholders are internal stakeholders as they benefit directly when the company makes a profit. Their investment decisions also depend upon how well or badly the company performs. In addition, they also examine the company's financial data to ensure that the company is performing well and that their investment is not losing money. They may also be in charge of voting on specific fund allocations.
- Not only are internal stakeholders impacted by the company's operations, but they also play an influential role in determining its performance. For example, managers and owners are the ones who take important strategic decisions in a business for its growth and survival. Similarly, the sincerity and hard work with which employees carry out their duties can have an impact on the degree of profitability of the business.

External stakeholders:

- External stakeholders are all those individuals, groups, firms and organizations that are not directly influenced by the performance of the business. These stakeholders might be interested in the performance and success of the organization, but they are not directly affected by it. They are also known as the secondary stakeholders of an organization.
- These external parties constitute the business environment of the organization. They use the financial information and other publicly available information about the company to become aware of its profitability and performance.
- External stakeholders are not involved in the everyday operations of an organization; however, the organizational activities do have an impact on them. They are not aware of the internal issues of the company and deal with it from the outside. Some of the external stakeholders are the customers, the suppliers who provide raw materials, clients, creditors, competitors, intermediaries, the general public as well as the government.
- **An external stakeholder, on the other hand, is someone who is not directly engaged with the business but may or shall be influenced by it at some point in time. External stakeholders are outside parties that do not work directly with the company but are affected by the activities of the business.** In today's corporate organizations, debt holders, trade creditors, suppliers, customers, the government, and society are all a part of external stakeholders.
- For example, if the government imposes any new regulation with regard to environmental

pollution, businesses must comply with it to ensure that their production activities do not harm the environment. Similarly, if a manufacturing unit pollutes the environment, social groups may protest, and hence, they fall under the category of external stakeholders.

Difference between internal stakeholders and external stakeholders:

The key points of difference between internal stakeholders and external stakeholders are listed below:

1. Meaning

Internal stakeholders are the people or entities that have a vested interest in the organization and are directly affected by its activities. External stakeholders, in contrast, are those people, groups or parties that are not directly affected by the success or failure of an organization.

2. Kind of influence

There is a direct impact of organizational activities on the internal stakeholders. However, external stakeholders are not directly influenced by organizational activities.

3. What do they do?

Internal stakeholders offer their services to the organization, whereas external stakeholders deal with the organization from the outside.

4. Information available

Internal stakeholders are aware of the internal problems and matters of the organization. In contrast, external stakeholders are not aware of the internal issues. Rather, they use financial information and any other information that is publicly available for different objectives.

5. Kind of stakeholders

Internal stakeholders are considered as the primary stakeholders whereas external stakeholders are considered as the secondary stakeholders.

6. Parties included

Internal stakeholders include the owners, managers, employees and investors of a company. External stakeholders comprise of the customers, competitors, suppliers, creditors, public and the government.

Customers (product and service issues), Employees (leadership, fairness, justice, diversity) :

Stakeholders are all those individuals, groups or entities that are interested in the performance of a company. There is direct involvement of internal stakeholders in the operations of a company, and they are directly affected by the way the organization performs. External stakeholders are, however, indirectly affected by the organizational operations and performance.

Both types of stakeholders are important part of the organization. Internal stakeholders are critical for the functioning of an organization. For example, in the absence of employees and managers, an organization

cannot carry out its day to day functions. In a similar way, external stakeholders are also very important. Customers are very important external stakeholders as they are the ones who will buy and use the product/service. Similarly, creditors are important as they offer companies the finances they need to carry out their operations. In addition, a company is supposed to adhere to the rules and laws put forward by the government and to pay taxes. Therefore, it is evident that like internal stakeholders, external stakeholders are also very significant. Companies, hence, need to establish good relationships with all of their stakeholders. This can be done when they align their objectives with those of their stakeholders.

Customers are external stakeholders as well, particularly if they rely substantially on the company's products or services to purchase products that they use on a daily basis. Likewise, they may also affect the business. For instance, the customers' likes and dislikes or demand patterns affect the nature of production/provisioning of services to be done by the company.

Unlike internal stakeholders who work closely within an organization, external stakeholders are indirectly influenced by the performance of the business and its successes or failures. For example, before doing business, suppliers who provide raw materials to a company rely on its market reputation and external reports.

Stakeholders role as suppliers, collaborators, business, community :

Stakeholders are essential for a variety of reasons. Internal stakeholders are crucial since the business's operations rely on their ability to collaborate & work efficiently to achieve the company's goals. External stakeholders, on the other hand, can have an indirect impact on the firm. Customers, for example, can change their purchasing patterns, suppliers can change their manufacturing and distribution techniques, and governments can change their laws and regulations.

Successful relationships with stakeholders are critical to the success of any business. In today's world, it is critical for every business organization not only to meet stakeholders' expectations but also to make genuine efforts to provide them with more than they are entitled to or desire because a business cannot survive in a vacuum without stakeholders.

Nature of Stakeholder:

A stakeholder is any individual, organization, social group, or society who has a stake in the business or has something to gain or lose. A stake implies a vital interest.

To put it in another way, a stakeholder is any individual or group whose interests are affected by the operations of the business. The key stakeholders in a typical company are its investors, employees, shareholders, and suppliers. However, in recent years, the definition of stakeholders has been expanded to include customers, communities, and governments as well. These are the people who are affected by the decisions of the business in some form or the other.

There are many types of stakeholders and not all stakeholders are equal. That is to say, the consumers of a company are entitled to fair trading practices but they are not entitled to the same amount of consideration as the company's employees.

Stakeholders can be internal and external. Both have some sort of interest in the company. Internal stakeholders are usually easy to identify because they have a financial stake in the business. External stakeholders are more difficult to define because they are not directly involved in the company's operations or decisions. Even though an external stakeholder does not have a direct financial interest in the organization, they do care about its success, failure, and direction.

UNIT –IV

- INDIVIDUAL VARIABLES IN ETHICAL MANAGEMENT

ETHICAL COURAGE:

Courage is the tendency to accept and face risks and difficult tasks in rational ways.

Selfconfidence is the basic requirement to nurture courage.

Courage is classified into three types, based on the types of risks, namely

(a) Physical

Courage

(b) Social courage, and

(c) Intellectual courage.

(a) **Physical courage:** In physical courage, the thrust is on the adequacy of the physical strength, including the muscle power and armaments. People with high adrenalin, may be prepared to face challenges for the mere ‘thrill’ or driven by a decision to ‘excel’.

(b) **Social courage:** The social courage involves the decisions and actions to change the order, based on the conviction for or against certain social behaviors. This requires leadership abilities, including empathy and sacrifice, to mobilize and motivate the followers, for the social cause.

(c) **Intellectual courage:** The intellectual courage is inculcated in people through acquired knowledge, experience, games, tactics, education, and training.

In professional ethics, courage is applicable to the employers, employees, public, and the press.

ETHICAL JUDGEMENT :

It is the ability to mutually experience the thoughts, emotions, and direct experience of others. The ability to understand another person’s circumstances, point of view, thoughts, and feelings is empathy. When experiencing empathy, you are able to understand someone else’s internal experiences.

Empathy is social radar. Sensing what others feel about, without their open talk, is the essence of empathy. Empathy begins with showing concern, and then obtaining and understanding the feelings of others, from others’ point of view. It is also defined as the ability to put one’s self into the psychological frame of reference or point of view of another, to know what the other person feels. It includes the imaginative projection into other’s feelings and understanding of other’s background such as parentage, physical and mental state, economic situation, and

association.

This is an essential ingredient for good human relations and transactions.

To practice 'Empathy', a leader must have or develop in him, the following characteristics:

1. **Understanding others:** It means sensing others feelings and perspectives, and taking active interest in their welfare.
2. **Service orientation:** It is anticipation, recognition and meeting the needs of the clients or customers.
3. **Developing others:** This means identification of their needs and bolstering their abilities. In developing others, the one should inculcate in him the 'listening skill' first.

Communication = 22% reading and writing + 23% speaking + 55% listening

One should get the feedback, acknowledge the strength and accomplishments, and then coach the individual, by informing about what was wrong, and giving correct feedback and positive expectation of the subject's abilities and the resulting performance.

4. **Leveraging diversity** (opportunities through diverse people): This leads to enhanced organizational learning, flexibility, and profitability.
5. **Political awareness:** It is the ability to read political and social currents in an organization.

The benefits of empathy include:

1. Good customer relations (in sales and service, in partnering).
2. Harmonious labor relations (in manufacturing).
3. Good vendor-producer relationship (in partnering.)

Through the above three, we can maximize the output and profit, as well as minimizing the loss. While dealing with customer complaints, empathy is very effective in realizing the unbiased views of others and in admitting one's own limitations and failures. According to Peter Drucker, purpose of the business is not to make a sale, but to make and keep a customer. Empathy assists one in developing courage leading to success.

ETHICAL AWARENESS

is defined as, decisions and actions exercised on the basis of moral concern for other people and recognition of good moral reasons. Alternatively, moral autonomy means 'self determinant or independent'.

The autonomous people hold moral beliefs and attitudes based on their critical reflection rather than on passive adoption of the conventions of the society or profession. Moral autonomy may also be defined as a skill and habit of thinking rationally about the ethical issues, on the basis of moral concern.

Viewing management as social experimentation will promote autonomous participation and retain one's professional identity. Periodical performance appraisals, tight-time schedules and fear of foreign competition threatens this autonomy. The attitude of the management should allow latitude in the judgments of their managers on moral issues. If management views profitability is more important than consistent quality and retention of the customers that discourage the moral autonomy, managers are compelled to seek the support from their professional societies and outside organizations for moral support. It appears that the blue-collar workers with the support of the union can adopt better autonomy than the employed professionals. Only recently the legal support has been obtained by the professional societies in exhibiting moral autonomy by professionals in this country as well as in the West.

ETHICAL FOUNDATION: It is defined as a set of attitudes concerned with the value of work, which forms the motivational orientation. It is a set of values based on hard work and diligence. It is also a belief in the moral benefit of work and its ability to enhance character. A work ethic may include being reliable, having initiative, or pursuing new skills. The 'work ethics' is aimed at ensuring the economy (get job, create wealth, earn salary), productivity (wealth, profit), safety (in workplace), health and hygiene (working conditions), privacy (raise family), security (permanence against contractual, pension, and retirement benefits), cultural and social development (leisure, hobby, and happiness), welfare (social work), environment (anti-pollution activities), and offer opportunities for all, according to their abilities, but without discrimination. Workers exhibiting a good work ethic in theory should be selected for better positions, more responsibility and ultimately promotion. Workers who fail to exhibit a good work ethic may be regarded as failing to provide fair value for the wage the employer is paying them and should not be promoted or placed in positions of greater responsibility. Work ethic is not just hard work but also a set of accompanying virtues, whose crucial role in the development and sustaining of free markets.

ETHICAL PSYCHOLOGY :

- Psychological hedonism
 - Psychological hedonism is the view that humans are psychologically constructed in such a way that we exclusively desire pleasure.
- Ethical hedonism
 - Ethical hedonism is the view that our fundamental moral obligation is to maximize pleasure or happiness.

Utilitarianism

Utilitarianism is a tradition of ethical philosophy that is associated with Jeremy Bentham and John Stuart Mill (1806 – 1873) British philosophers. Utilitarianism is a normative ethical theory that places the locus of right and wrong solely on the outcomes (consequences) of choosing one actions.

Three Basic Principles of Utilitarianism

Pleasure or Happiness Is the Only Thing That Truly Has Intrinsic Value

Actions Are Right as they Promote Happiness, Wrong as they Produce Unhappiness

Everyone's Happiness Counts Equally.

Virtue ethics

Virtue ethics is developed by the philosopher Aristotle. Virtue ethics mainly deals with the honesty and morality of a person. It states that practicing good habits such as courage, honesty, ambitious, truthfulness and patience makes a moral and virtuous person. Virtues are admirable qualities that lead to moral excellence.

ETHICS IN DECISION MAKING MANAGEMENT :

- Avoidance of conflict of interest.
- Accurate and timely disclosure in reports and documents that the company files before

Government agencies, as well as in Company's other communications.

- Compliance of applicable laws, rules and regulations including Insider Trading Regulations.
- Maintaining confidentiality of Company affairs.
- Non-competition with Company and maintaining fair dealings with the Company.
- Standards of business conduct for Company's customers, communities, suppliers, shareholders, competitors, employees.
- Prohibition of Directors and senior management from taking corporate opportunities for themselves or their families.
- Review of the adequacy of the Code annually by the Board.
- No authority of waiver of the Code for anyone should be given.

UNIT V

PRACTICAL FIELD, GUIDE , TECHNIQUES AND SKILLS: **ETHICAL DILEMMA**

An ethical dilemma (moral dilemma) is a problem in the decision-making process between two possible options, neither of which is absolutely acceptable from an ethical perspective. Ethical dilemmas can be solved in various ways, for example by showing that the claimed situation is only apparent and does not really exist, or that the solution to the ethical dilemma involves choosing the greater good and lesser evil, or that the whole framing of the problem omits creative alternatives that situational ethics or situated ethics must apply because the case cannot be removed from context and still be understood.

A popular ethical conflict is that between an imperative or injunction not to steal and one to care for a family that you cannot afford to feed without stolen money. Debates on this often revolve around the availability of alternate means of income or support such as a social safety net, charity, etc. The debate is in its starkest form when framed as stealing food. Under an ethical system in which stealing is always wrong and letting one's family die from starvation is always wrong, a person in such a situation would be forced to commit one wrong to avoid committing another, and be in constant conflict with those whose view of the acts varied.

However, there are no legitimate ethical systems in which stealing is more wrong than letting one's family die. Ethical systems do in fact allow for, and sometimes outline, trade-offs or priorities in decisions. Resolving ethical dilemmas is rarely simple or clear cut and very often involves revisiting similar dilemmas that recur within societies.

HOW TO SOLVE AN ETHICAL DILEMMA

The biggest challenge of an ethical dilemma is that it does not offer a solution that would comply with ethical norms. Throughout the history of humanity, people have faced such dilemmas and philosophers aimed and worked to find solution

The following approaches to solve an ethical dilemma were deducted:

Refute the paradox (dilemma): The situation must be carefully analysed. In some cases, the existence of the dilemma can be logically refuted.

Value theory approach: Choose the alternative that offers the greater good or the lesser evil.

the problem can be reconsidered and new alternative solutions may arise.

ETHICS IN PRACTICE

Human dignity, human rights and justice, which refers to the duty to promote universal respect for the human person. In the context of fisheries, this principle relates, for example, to fishers' self-determination, access to fishing resources and the right to food. It is best represented by a rights-based approach in ethics that emphasizes the protection of the personal domain of each individual. It may require, however, the establishment of individual or community rights, the exact nature of which will depend on local conditions.

Beneficence, which concerns human welfare, reducing the harms and optimizing the benefits of social practices. In the context of fisheries, this principle needs to be observed when the effects of policies and practices upon the livelihoods of fishing communities are evaluated. The principle relates to working conditions (safety on board), as well as food quality and safety. The issue of genetically modified organisms should also be addressed in this context (FAO, 2001b). This principle invites an ethical approach to fisheries that puts consequences to general welfare in focus.

Cultural diversity, pluralism and tolerance, which relates to the need to take different value systems into account within the limits of other moral principles. The pressing moral issues in fisheries take different shapes across different cultures, and it is an important moral demand that people themselves define how their interests are best served in a particular cultural setting. This principle squares well with dialogical ethics, which stresses the actual participation of those concerned.

Solidarity, equity and cooperation, which refers to the importance of collaborative action, sharing scientific and other forms of knowledge, and non-discrimination. In the context of fisheries, this principle underpins the moral imperative to eradicate poverty in developing countries and ensure equity within fisheries and between sectors. It also requires transparent

policies and stresses the need to reduce the gap between producers and consumers.

This principle is relevant at the level of policy as well as at the individual level of virtues and professional duties to further trust and tolerance among stakeholders. Responsibility for the biosphere, which concerns the interconnections of all life forms and the protection of biodiversity. This principle stresses that ecosystem well-being is a sine qua non condition of sustainable fisheries providing for the needs of future generations, as well as for the lives of those who currently rely on the natural environment and are responsible for its use.

UNETHICAL BEHAVIOUR

The Civil Service Commission of Philippines defined an unethical behaviour as any behaviour prohibited by law. An unethical behaviour would therefore be defined as one that is not morally honourable or one that is prohibited by the law. Many behaviours will fall in the classification including corruption, mail and wire fraud, discrimination and harassment, insider trading, conflicts of interest, improper use of company assets, bribery

CAUSES OF UNETHICAL BEHAVIOUR IN WORKPLACE

Misusing Company Time : One of the most regularly revealed “bad behaviours” in the workplace is the misuse of company time. This category includes knowing that one of your colleagues is directing personal business on company time, staff appearing late, extra breaks or fake timesheets. These negative behaviour patterns can rapidly spread to different workers. It can also cultivate hatred amongst colleagues, severely influencing One of the most regularly revealed “bad behaviours” in the workplace is the misuse of company time. This category

includes knowing that one of your colleagues is directing personal business on company time, staff appearing late, extra breaks or fake timesheets. These negative behaviour patterns can rapidly spread to different workers. It can also cultivate hatred amongst colleagues, severely influencing morale and efficiency.

1. Unethical Leadership

Having a personal issue with your boss or manager is a certain thing, yet reporting to a person who is acting dishonestly is another. This may come in a clear form, such as manipulating numbers in a report or sending company money on improper activities; nonetheless, it can also happen more subtly, through bullying, accepting inadequate gifts from suppliers, or requesting that you avoid a standard system just once. With studies demonstrating that managers are responsible for 60 percent of workplace wrongdoing, the abuse of leadership authority is a disastrous reality.

2. Lying to Employees

The quickest way to lose the trust of your employees is to lie to them, but managers do it constantly. One out of every five workers report that their supervisor or manager has lied to them within the previous year.

3. Harassment and Discrimination

Laws require associations to be equivalent to business opportunity employers. Organizations must select a various workplace, authorize policies and training that help an equivalent open-door program, and encourage a situation that is respectful of a wide range of people. Unfortunately, there are still numerous people whose practices break with EEOC rules and regulations. When harassment and discrimination of employees based on ethnicity, race, gender, handicap or age occur, has a moral line been crossed as well as a legitimate one also. Most companies are attentive to maintain a strategic distance from the costly legal and public implications of harassment and discrimination, so you may experience this ethical problem in more delicate ways, from apparently “harmless” offensive jokes by a manager to a more unavoidable “group think” mindset that can be a symptom of a toxic culture. This could be a group mindset toward an “other” group. Your best reaction is to keep up your qualities and

repel such intolerant, illegal or unethical group standards by offering an option, inclusive aspect as the best decision for the group and the company.

4. Violating Company Internet Policy

Cyberloafers and Cybershackers are terms used to recognize people who surf the web when they ought to work. It's a huge, multi-billion-dollar issue for organizations. Every day at least 64 percent of employers visit sites that have nothing to do with their work.

CREATING CULTURE OF ETHICAL MANAGEMENT :

Character

Those with a good work ethic often also possess generally strong character. This means they are self-disciplined, pushing themselves to complete work tasks instead of requiring others to intervene. They are also often very honest and trustworthy, as they view these traits as befitting the high-quality employees they seek to become. To demonstrate their strong character, these workers embody these positive traits daily, likely distinguishing themselves from the rest.

CODE OF CONDUCT

Code of conduct or what is popularly known as Code of Business Conduct contains standards of business conduct that must guide actions of the Board and senior management of the Company.

The Code may include the following:

- Company Values.
- Avoidance of conflict of interest.
- Accurate and timely disclosure in reports and documents that the company files before Government agencies, as well as in Company's other communications.
- Compliance of applicable laws, rules and regulations including Insider Trading Regulations.
- Maintaining confidentiality of Company affairs.
- Non-competition with Company and maintaining fair dealings with the Company.
- Standards of business conduct for Company's customers, communities, suppliers,

shareholders, competitor employees.

- Prohibition of Directors and senior management from taking corporate opportunities for themselves or their families.
- Review of the adequacy of the Code annually by the Board.
- No authority of waiver of the Code for anyone should be given.

The Code of Conduct for each Company summarises its philosophy of doing business. Although the exact details of this code are a matter of discretion, the following principles have been found to occur in most of the companies:

- Use of company's assets;
- Avoidance of actions involving conflict of interest;
- Avoidance of compromising on commercial relationship;
- Avoidance of unlawful agreements;
- Avoidance of offering or receiving monetary or other inducements;

DEVELOPING TECHNIQUES AND SKILLS IN ETHICAL MANAGEMENT:

Reliability

Reliability goes hand in hand with a good work ethic. If individuals with a good work ethic say they are going to attend a work function or arrive at a certain time, they do, as they value punctuality. Individuals with a strong work ethic often want to appear dependable, showing their employers that they are workers to whom they can turn. Because of this, they put effort into portraying -- and proving -- this dependability by being reliable and performing consistently.

Dedication

Those with a good work ethic are dedicated to their jobs and will do anything they can to ensure that they perform well. Often this dedication leads them to change jobs less frequently, as they become committed to the positions in which they work and are not eager to abandon these posts. They also often put in extra hours beyond what is expected, making it easy for their employers to see that they are workers who go beyond the rest of the workforce and truly dedicate themselves to their positions.

Productivity

Because they work at a consistently fast pace, individuals with a good work ethic are often highly productive. They commonly get large amounts of work done more quickly than others who lack their work ethic, as they don't quit until they've completed the tasks with which they were presented. This high level of productivity is also due, at least in part, to the fact that these individuals want to appear to be strong workers. The more productive they are, the more beneficial to the company they appear to those managing them.

Cooperation

Cooperative work can be highly beneficial in the business environment, something that individuals with a strong work ethic know well. Because they recognize the usefulness of cooperative practices - such as teamwork -- they often put an extensive amount of effort into working well with others. These individuals commonly respect their bosses enough to work with any individuals with whom they are paired in a productive and polite manner, even if they do not enjoy working with the individuals in question.

Character

Those with a good work ethic often also possess generally strong character. This means they are self-disciplined, pushing themselves to complete work tasks instead of requiring others to intervene.

- problem Compare their efficiency and accuracy
5. Try to implement a Bio inspired computing in Networks/Biomedical/Cloud computing applications to obtain an optimal solution

COURSE OUTCOMES:

Upon completion of the course, the students should be able to

- CO1:**Implement and apply bio-inspired algorithms
CO2:Explain random walk and simulated annealing
CO3:Implement and apply genetic algorithms
CO4:Explain swarm intelligence and ant colony for feature selection
CO5:Apply bio-inspired techniques in various fields

REFERENCES

1. Eiben,A.E.Smith,James E, "Introduction to Evolutionary Computing", Springer 2ndEdition2015.
2. Helio J.C. Barbosa, "Ant Colony Optimization - Techniques and Applications", IntechFirstEdition,2013
3. Xin-She Yang , Joao Paulo papa, "Bio-Inspired Computing and Applications in Image Processing",ElsevierFirst Edition, 2016
4. Xin-She Yang, "Nature Inspired Optimization Algorithm",Elsevier First Edition 2014
5. Yang ,Cui,Xiao,Gandomi,Karamanoglu,"Swarm Intelligence and Bio-Inspired Computing", Elsevier First Edition 2013

MC4015

DIGITAL MARKETING

L T P C
3 0 0 3

COURSE OBJECTIVES:

- To understand the difference between Traditional Marketing and digital Marketing
- To understand and analyze the search engine functions
- To develop a deep knowledge about the Digital marketing platforms and the theoretical aspects of creating a website
- To analyze inbuilt tools for digital Marketing

UNIT I INTRODUCTION TO DIGITAL MARKETING

9

What is Digital Marketing- Need of Digital Marketing-Digital Marketing Platforms – Understanding digital marketing process- Difference between Traditional Marketing and digital Marketing- tools of Digital marketing - Advantage of Digital Marketing-Digital Marketing Manager Role and functions - How we use both Digital & Traditional Marketing

UNIT II WEBSITE & SEARCH ENGINE

9

Website –Hosting and Domain– Different platforms for website creation- Introduction to SERP- What are search engines- How search engines work- Major functions of a search engine- What are keywords -Different types of keywords- Google keyword planner tool.

UNIT III MISC TOOLS- GOOGLE WEBMASTER TOOLS

9

Site Map Creators- Browser-based analysis tools-Page Rank tools-pinging & indexing tools-

Dead links identification tools- Open site explorer Domain information/ whois tools- Quick sprout

UNIT IV LEAD MANAGEMENT & DIGITAL MARKETING

9

Web to lead forms- Web to case forms- Lead generation techniques- Leads are everywhere- Social media and lead gen Inbuilt tools for Digital Marketing-Ip Tracker- CPC reduction (in case of paid ads) Group posting on Social Media platforms

UNIT V TRENDING DIGITAL MARKETING SKILLS

9

Search Engine Optimization(SEO)-Search Engine Marketing(SEM).-Social Media Marketing/Optimization- Email Marketing. Website :Product Marketing- Content Writing. Marketing the created content online Copywriting- Blogging- Local Marketing. Google Ad Words - Campaign Management- PPC Advertising- Affiliate Marketing. Mobile and SMS Marketing- Marketing Automation-Web Analytics- Growth Hacking

SUGGESTED ACTIVITIES:

1. Subscribe to a weekly/quarterly newsletter and analyze how it's content and structure aid with the branding of the company and how it aids its potential customer segments.
2. Perform keyword search for a skincare hospital website based on search volume and competition using Google keyword planner tool.
3. Demonstrate how to use the Google WebMasters Indexing API
4. Discuss an interesting case study regarding how an insurance company manages leads.
5. Discuss negative and positive impacts and ethical implications of using social media for political advertising.
6. Discuss how Predictive analytics is impacting marketing automation.

TOTAL: 45 PERIODS

COURSE OUTCOMES:

CO1:To gain insight on the concept of digital marketing and the role of a digital manager.

CO2:To understand and administer the website and the search engines.

CO3:To understand how to use MISC and Google Webmaster tools.

CO4:To understand the concepts of lead management and digital marketing.

CO5:To gain knowledge on the latest digital marketing trends

REFERENCES

1. Chaffey, D. (2019). Digital marketing strategy, Implementation and Practice. Pearson
2. Chaffey, D., & Smith, P. R. (2017). Digital marketing excellence: planning, optimizing and integrating online marketing. Taylor & Francis. ·
3. Kaufman, I., & Horton, C. (2014). Digital marketing: Integrating strategy and tactics with values, a guidebook for executives, managers, and students. Routledge.
4. Royle, J., & Laing, A. (2014). The digital marketing skills gap: Developing a Digital Marketer Model for the communication industries. International Journal of Information Management, 34(2), 65-73.
5. Dodson, I. (2016). The art of digital marketing: the definitive guide to creating strategic, targeted, and measurable online campaigns. John Wiley & Sons.

Syllabus:

**Site Map Creators- Browser-based analysis tools-Page Rank tools-
pinging & indexing tools-Dead links identification tools- Open site explorer
Domain information/ whois tools- Quick sprout**

Site Map Creators:

This page describes how to build a sitemap and make it available to Google. Learn more about sitemaps here.

- Decide which sitemap format you want to use.
- Create the sitemap, either automatically or manually.
- Make your sitemap available to Google.

Sitemap formats:

- Google supports several sitemap formats
- XML
- RSS, mRSS, and Atom 1.0
- Text

Google accepts the standard sitemap protocol in all formats. Google does not currently consume the <priority> tag included in sitemaps.

All formats limit a single sitemap to 50MB (uncompressed) or 50,000 URLs. If you have a larger file or more URLs, you will have to break your list into multiple sitemaps. You can optionally create a sitemap index file (a file that points to a list of sitemaps) and submit that single index file to Google. You can submit multiple sitemaps and/or sitemap index files to Google

XML:

Here is a very basic XML sitemap that includes the location of a single URL:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>https://www.example.com/foo.html</loc>
    <lastmod>2022-06-04</lastmod>
  </url>
</urlset>
```

You can find more complex examples and full documentation at sitemaps.org.

You can see examples of sitemaps that specify alternate language pages and sitemaps for news, image, or video files.

RSS, mRSS, and Atom 1.0:

If you have a blog with an RSS or Atom feed, you can submit the feed's URL as a sitemap. Most blog software is able to create a feed for you, but recognize that this feed only provides information on recent URLs.

Google accepts RSS 2.0 and Atom 1.0 feeds.

You can use an mRSS (media RSS) feed to provide Google details about video content on your site.

Text:

If your sitemap includes only web page URLs, you can provide Google with a simple text file that contains one URL per line. For example:

<https://www.example.com/file1.html>

<https://www.example.com/file2.html>

Guidelines for text file sitemaps :

Encode your file using UTF-8 encoding.

Don't put anything other than URLs in the sitemap file.

You can name the text file anything you wish, provided it has a .txt extension (for instance, `sitemap.txt`).

Sitemap extensions for additional media types :

Google supports extended sitemap syntax for the following media types. Use these extensions to describe video files, images, and other hard-to-parse content on your site to improve indexing.

- Video
- Images
- Google News

General sitemap guidelines :

- Use consistent, fully-qualified URLs. Google will crawl your URLs exactly as listed. For instance, if your site is at <https://www.example.com/>, don't specify a URL as <https://example.com/> (missing www) or [./mypage.html](/mypage.html) (a relative URL).
- A sitemap can be posted anywhere on your site, but a sitemap affects only descendants of the parent directory. Therefore, a sitemap posted at the site root can affect all files on the site, which is where we recommend posting your sitemaps.
- Don't include session IDs and other user-dependent identifiers from URLs in your sitemap. This reduces duplicate crawling of those URLs.
- Tell Google about alternate language versions of a URL using hreflang annotations.
- Sitemap files must be UTF-8 encoded, and URLs escaped appropriately.
- Break up large sitemaps into smaller sitemaps: a sitemap can contain up to 50,000 URLs and must not exceed 50MB uncompressed. Use a sitemap index file to list all the individual sitemaps and submit this single file to Google rather than submitting individual sitemaps.
- List only canonical URLs in your sitemaps. If you have multiple versions of a page, list in the sitemap only the one you prefer to appear in search results. If you have multiple versions of your site (for example, www and non-www), decide which is your preferred site, and put the sitemap there, and add rel=canonical or redirects on the other site.
- If you have different URLs for mobile and desktop versions of a page, we recommend pointing to only one version in a sitemap. However, if you want to point to both URLs, annotate your URLs to indicate the desktop and mobile versions.
- Use sitemap extensions for pointing to additional media types such as video, images, and news.
- If you have alternate pages for different languages or regions, you can use hreflang in either a sitemap or html tags to indicate the alternate URLs.
- Non-alphanumeric and non-latin characters. We require your sitemap file to be UTF-8 encoded (you can generally do this when you save the file). As with all XML files, any data values (including URLs) must use entity escape codes for the characters listed in the following table. A sitemap can contain only ASCII characters; it can't contain extended ASCII characters or certain control codes or

special characters such as * and {}. If your sitemap URL contains these characters, you'll receive an error when you try to add it.

Character	Symbol	Escape Code
Ampersand	&	&
Single Quote	'	'
Double Quote	"	"
Greater Than	>	>
Less Than	<	<

In addition, all URLs (including the URL of your sitemap) must be encoded for readability by the web server on which they are located and URL-escaped. However, if you are using any sort of script, tool, or log file to generate your URLs (anything except typing them in by hand), this is usually already done for you. If you submit your sitemap and you receive an error that Google is unable to find some of your URLs, check to make sure that your URLs follow the RFC-3986 standard for URIs, the RFC-3987 standard for IRIs, and the XML standard.

Here is an example of a URL that uses a non-ASCII character (ü), as well as a character that requires entity escaping (&):

<https://www.example.com/ümlat.html&q=name>

Here is that same URL encoded using ISO-8859 encoding, and with the entity escaped:

<https://www.example.com/%FCmlat.html&q=name>

Here is that same URL using UTF-8 encoding, and with the entity escaped:

<https://www.example.com/%C3%BCmlat.html&q=name>

Remember that sitemaps are a recommendation to Google about which pages you think are important; Google does not pledge to crawl every URL in a sitemap.

Google ignores <priority> and <changefreq> values.

Google uses the <lastmod> value if it's consistently and verifiably (for example by comparing to the last modification of the page) accurate.

The position of a URL in a sitemap is not important; Google does not crawl URLs in the order in which they appear in your sitemap

How to create a sitemap:

When creating a sitemap, you're telling search engines about which URLs you prefer to show in search results. These are the canonical URLs. If you have the same content accessible under different URLs, choose the URL you prefer and include that in the sitemap instead of all URLs that lead to the same content.

Once you've decided which URLs to include in the sitemap, pick one of the following ways to create a sitemap, depending on your site architecture and size:

- Let your CMS generate a sitemap for you.
- For sitemaps with less than a few dozen URLs, you can manually create a sitemap.
- For sitemaps with more than a few dozen URLs, automatically generate a sitemap.

Let your CMS generate a sitemap for you

If you're using a CMS such as WordPress, Wix, or Blogger, it's likely that your CMS has already made a sitemap available to search engines. Try searching for information about how your CMS generates sitemaps, or how to create a sitemap if your CMS doesn't generate a sitemap automatically. For example, in case of Wix, search for "wix sitemap".

For all other site setups, you will need to generate the sitemap yourself.

Manually create a sitemap

For sitemaps with less than a few dozen URLs, you may be able to manually create a sitemap. For this, open a text editor such as Windows Notepad or Nano (Linux, MacOS), and follow a syntax described in the Sitemap Formats section. You can name the file anything you like as long as the characters are allowed in a URL.

You can manually create larger sitemaps, but it's a tedious process and hard to maintain long term.

Automatically generate a sitemap with tools

For sitemaps with more than a few dozen URLs, you will need to generate the sitemap. There are various tools that can generate a sitemap. However, the best way is to have your website software generate it for you. For example, you can extract your site's URLs from your website's database and then export the URLs to either the screen or actual file on your web server. Talk to your developers or server manager about this solution. If you need inspiration for the code, check out our old collection of third-party sitemap generators.

Keep in mind the size requirements for sitemaps. Learn more about managing large sitemaps.

Submit your sitemap to Google

Google doesn't check a sitemap every time a site is crawled; a sitemap is checked only the first time that we notice it, and thereafter only when you ping us to let us know that it's changed. Alert Google about a sitemap only when it's new or updated; don't submit or ping unchanged sitemaps multiple times.

If you have updated pages in the sitemap, mark them with the <lastmod> field. Other XML files have a similar field, such as <updated> for Atom XML. You can also learn how to compute this date.

There are a few different ways to make your sitemap available to Google:

Submit a sitemap in Search Console using the Sitemaps report. This will allow you to see when Googlebot accessed the sitemap and also potential processing errors.

Use the Search Console API to programmatically submit a sitemap.

Use the ping tool. Send a GET request in your browser or the command line to this address, specifying the full URL of the sitemap. Be sure that the sitemap file is accessible:

https://www.google.com/ping?sitemap=FULL_URL_OF_SITEMAP

Example:

<https://www.google.com/ping?sitemap=https://example.com/sitemap.xml>

Insert the following line anywhere in your robots.txt file, specifying the path to your sitemap. We will find it the next time we crawl your robots.txt file:

Sitemap: https://example.com/my_sitemap.xml

Use Web Sub if you use Atom/RSS for your sitemap and want to broadcast your changes to other search engines in addition to Google.

Submitting a sitemap is merely a hint: it doesn't guarantee that Google will download the sitemap or use the sitemap for crawling URLs on the site.

Browser-based analysis tools :

8 free website analytics tools :

- Smartlook
- Google Analytics
- Clicky
- Matomo
- Hotjar
- Woopra
- Open Web Analytics
- Clarity

3 complementary digital marketing tools :

- Similarweb
- Semrush
- HubSpot

8 free website analytics tools :

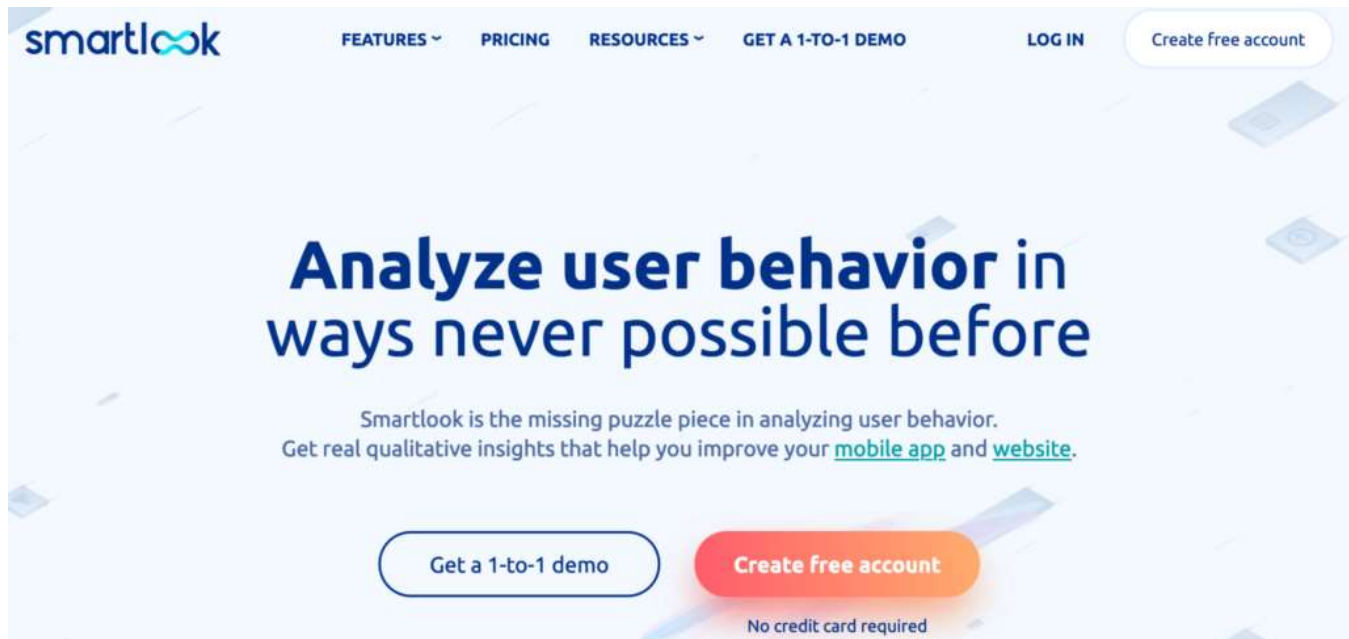
Before we dive in, note that we've only included tools that either:

Have a free plan (which you can use without a time constraint and without providing a credit card — like Smartlook).

Are completely free (like Google Analytics).

We won't be discussing analytics tools that only have a free trial or demo (like Crazy Egg or Kissmetrics), as they eventually force you to upgrade or stop using their service.

1. Smartlook



- **Free plan or free trial:** Both. The free plan lets you record up to 1,500 user sessions/month. All paid plans start with a free 10-day trial (no credit card required).
- **Self-hosted or SaaS tool:** SaaS tool.
- **Quantitative or qualitative analytics:** Both. Smartlook combines the power of quantitative and qualitative data, so you can understand what your users do and why they do it.
- **User interaction (event) tracking:** User interactions are collected automatically and you can define events without coding.

Smartlook is our tool that helps more than 2800+ paying customers analyze user behavior on their **websites as well as Android and iOS native apps**.

Our tool comes with four key features — *session recordings, events, funnels,* and *heatmaps* — all of which are available on the free plan. You can use this unique

feature set to do both **quantitative and qualitative analytics**, which helps you uncover what users do and why they do it.

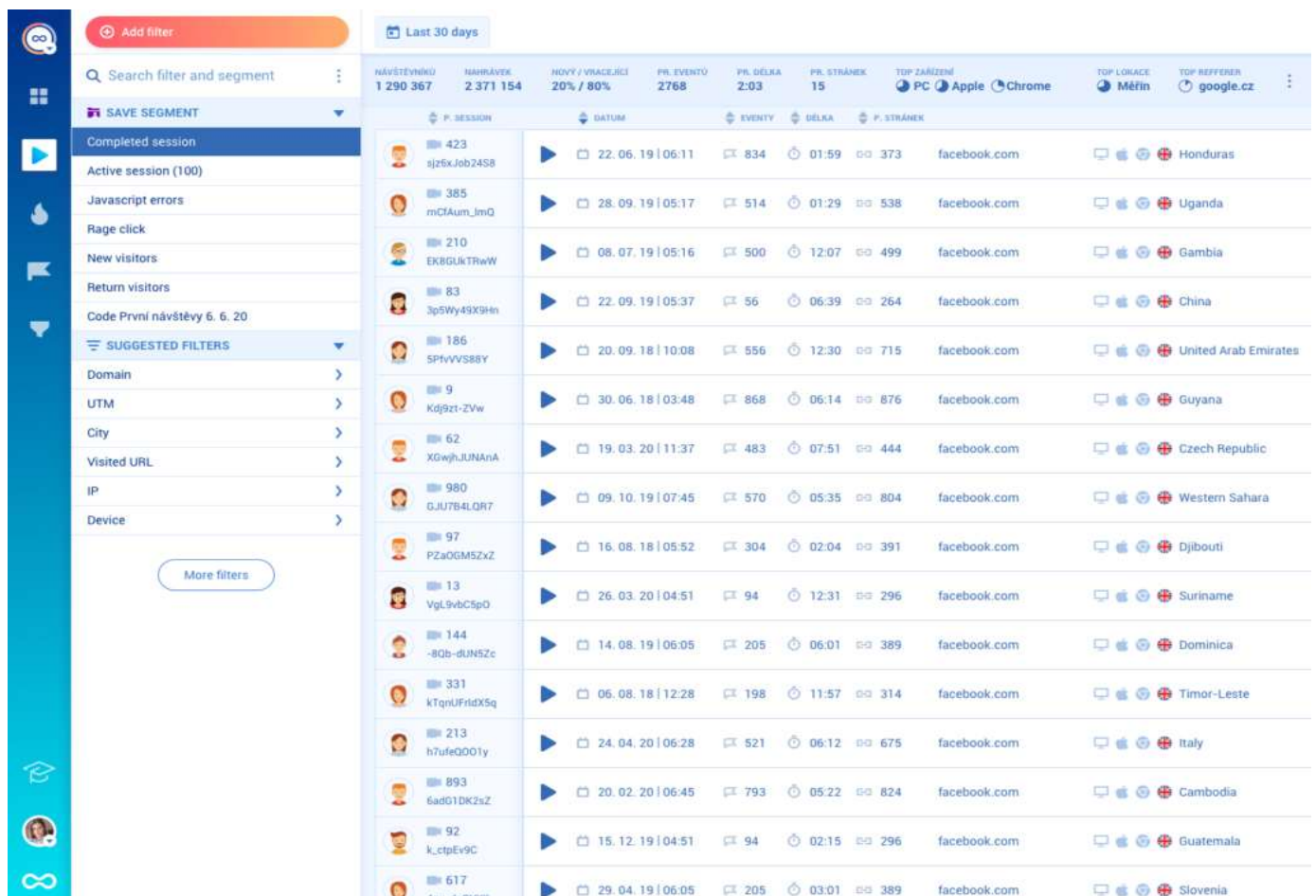
Let's look at three popular Smartlook use cases.

#1 Record your users' sessions automatically

Session recordings (sometimes called session replays) capture a user's entire session on your site and every interaction.

Once our code snippet is installed, Smartlook automatically starts recording your website's visitors, so you don't have to manually start and stop the session recorder.

You can find the *session replays* in the **"Recordings"** tab of your Smartlook *dashboard*.



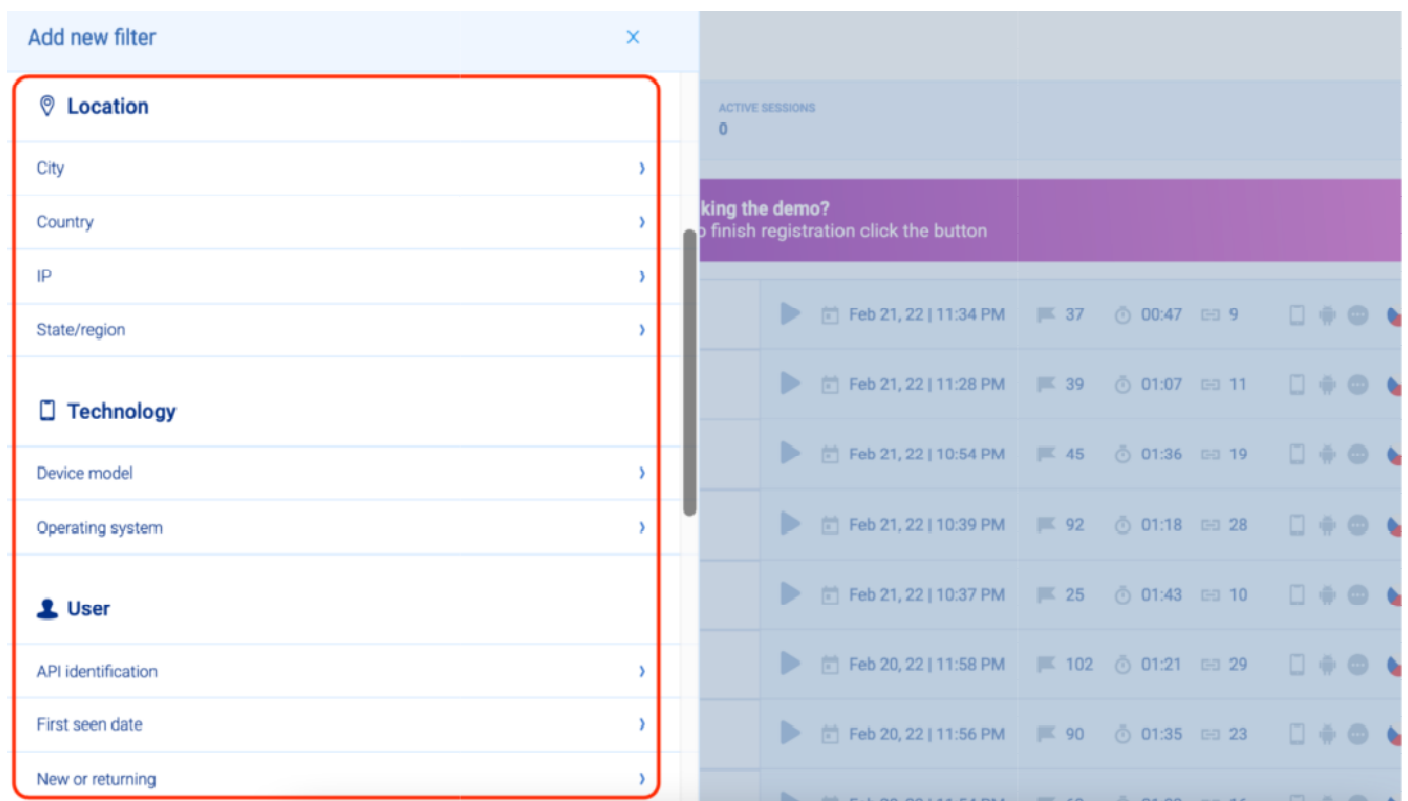
Add filter		Last 30 days		NÁVŠTĚVNÍKU		NOVÝ / VRAČEJÍCÍ		PR. EVENTŮ		PR. DÉLKA		PR. STRÁNEK		TOP ZAŘÍZENÍ		TOP LOKACE		TOP REFERRER			
1 290 367		2 371 154		20% / 80%		2768		2:03		15		PC Apple Chrome		Měřin		google.cz					
SESSION	DATUM	EVENTY	DÉLKA	P. STRÁNEK																	
423 sj25xJob2458	22. 06. 19 06:11	834	01:59	373	facebook.com	Honduras															
385 mCfAum_lm0	28. 09. 19 05:17	514	01:29	538	facebook.com	Uganda															
210 EK8GUKTrwW	08. 07. 19 05:16	500	12:07	499	facebook.com	Gambia															
83 3p5Wy49X9Hn	22. 09. 19 05:37	56	06:39	264	facebook.com	China															
186 SPfvVS88Y	20. 09. 18 10:08	556	12:30	715	facebook.com	United Arab Emirates															
9 Kdj9zi-ZVw	30. 06. 18 03:48	868	06:14	876	facebook.com	Guyana															
62 XGwjhJUNAnA	19. 03. 20 11:37	483	07:51	444	facebook.com	Czech Republic															
980 GJU7B4LQR7	09. 10. 19 07:45	570	05:35	804	facebook.com	Western Sahara															
97 PZaOGM5ZxZ	16. 08. 18 05:52	304	02:04	391	facebook.com	Djibouti															
13 VgLS9vbc5p0	26. 03. 20 04:51	94	12:31	296	facebook.com	Suriname															
144 -80b-dUN5Zc	14. 08. 19 06:05	205	06:01	389	facebook.com	Dominica															
331 kTqnlUfridX5q	06. 08. 18 12:28	198	11:57	314	facebook.com	Timor-Leste															
213 h7ufe0001y	24. 04. 20 06:28	521	06:12	675	facebook.com	Italy															
893 6adG1DK2sZ	20. 02. 20 06:45	793	05:22	824	facebook.com	Cambodia															
92 k_ctpEv9C	15. 12. 19 04:51	94	02:15	296	facebook.com	Guatemala															
617 4au-da0VXE	29. 04. 19 06:05	205	03:01	389	facebook.com	Slovenia															

With Smartlook's free plan, you can capture up to 1,500 sessions per month, while recording up to 10 users simultaneously.

If 15 visitors are on your site simultaneously, the first 10 will be recorded, but the last five won't be. However, **with a paid plan (starting at \$39/month), Smartlook automatically records *all* of your users' sessions.**

Paid plans also include our *Session Vault* feature, which lets you save selected recordings longer than your plan's data retention period. That way, you can store useful recordings for long periods and refer back to them when you need to.

To find relevant recordings fast, you can use 30+ filters, like device, operating system, country, IP address, and more.



On the free plan, you can apply up to two *filters* simultaneously. Paid plans let you apply even more *filters*, so you can conduct a more detailed analysis of your traffic.

Additionally, Smartlook tracks JavaScript errors and rage clicks by default. As a result, you can quickly find *session replays* of users who were frustrated or experienced an error. In fact, one of our clients used *session recordings* to find 15 bugs on their website in three hours.

Lastly, Smartlook is GDPR-compliant by design and you can also set up masks on any user input forms to exclude sensitive information so it never gets sent to our servers.

#2 Track events and build funnels without coding

As we said, tracking specific interactions (i.e., *events*) is vital for understanding your users' behavior.

However, most analytics tools require you to manually set up tracking on each element (buttons, forms, links, etc.) before they start collecting user interaction data. This issue has two important consequences:

- There's always a delay between realizing you need to analyze a particular action and having the data to do so.
- Setting up event tracking is difficult and time-consuming, so you usually need help from a developer.

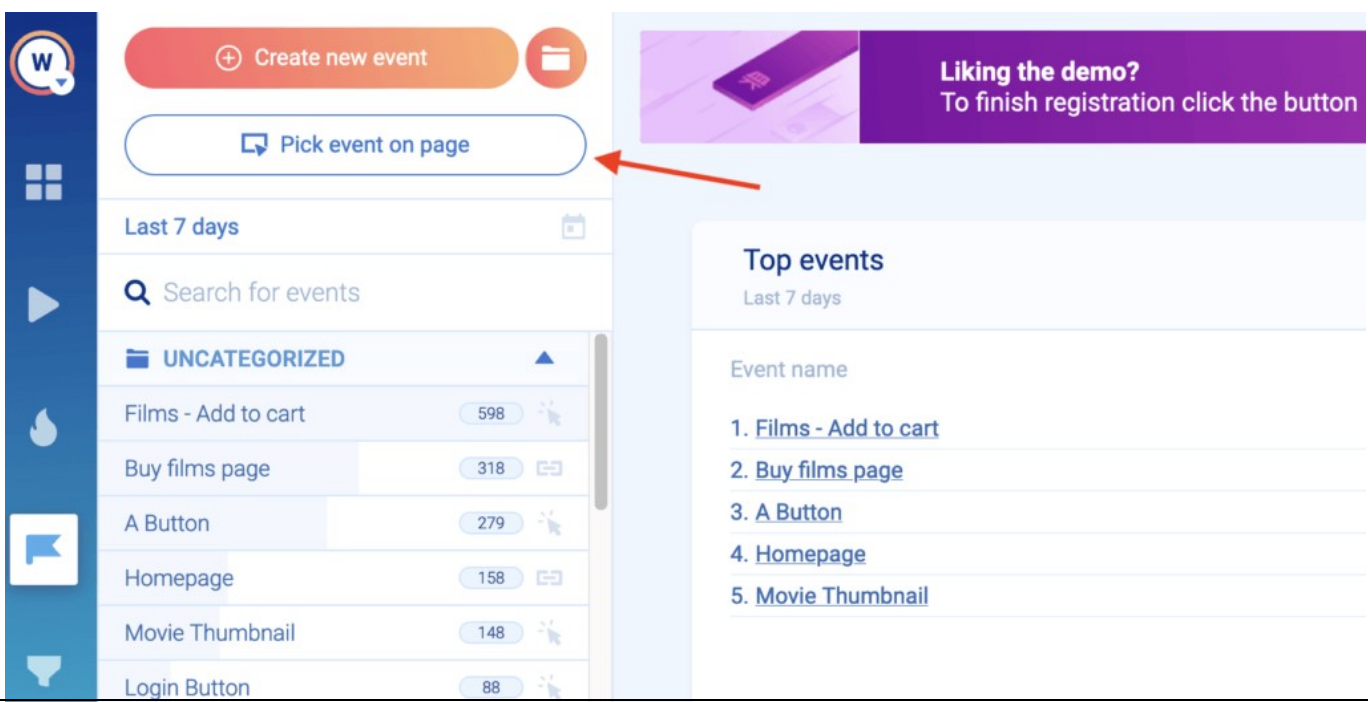
In contrast, once Smartlook is set up, it automatically starts recording all user interactions and lets you track *events* without coding.

You don't have to implement tracking on every element or get help from a developer each time you want to analyze a new interaction. Instead, the user interaction data is collected automatically.

All you have to do is select which interactions you want to track as *events* in your *dashboard*. The process by which you select these interactions is called defining an *event*.

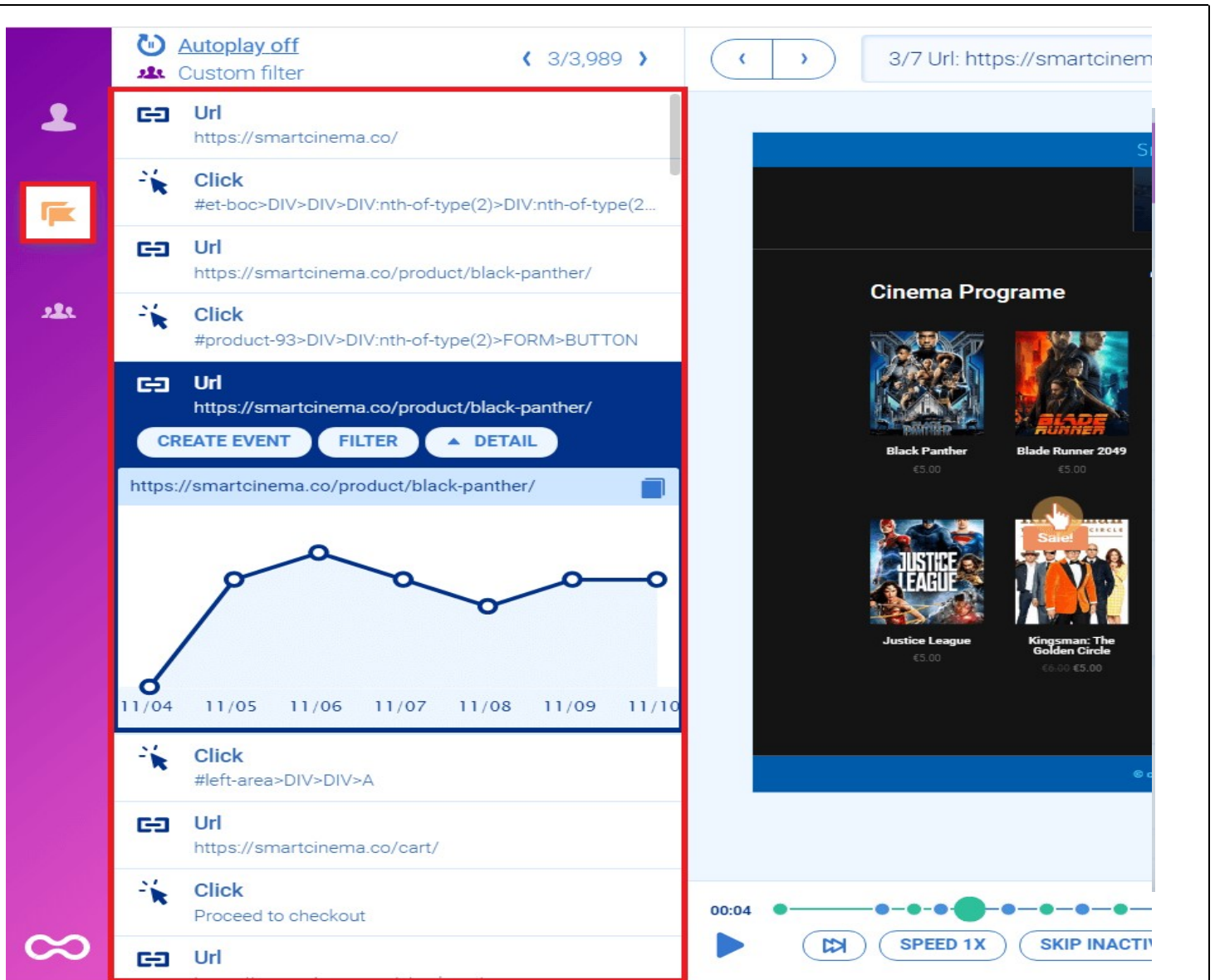
You can define *events* in three ways without coding:

1. **Choose from a list of standard *events***, including clicked on URL, clicked on text, typed text, clicked-on CSS selector (allowing you to select any element on the page).



The screenshot displays the Smartlook dashboard interface. On the left, a sidebar contains navigation icons. The main content area features a top navigation bar with a 'Create new event' button and a 'Pick event on page' button, the latter of which is highlighted with a red arrow. Below this is a search bar and a list of events under the 'UNCATEGORIZED' filter. The 'Top events' section for the last 7 days lists the following events:

Event name	Count
1. Films - Add to cart	598
2. Buy films page	318
3. A Button	279
4. Homepage	158
5. Movie Thumbnail	148
Login Button	88



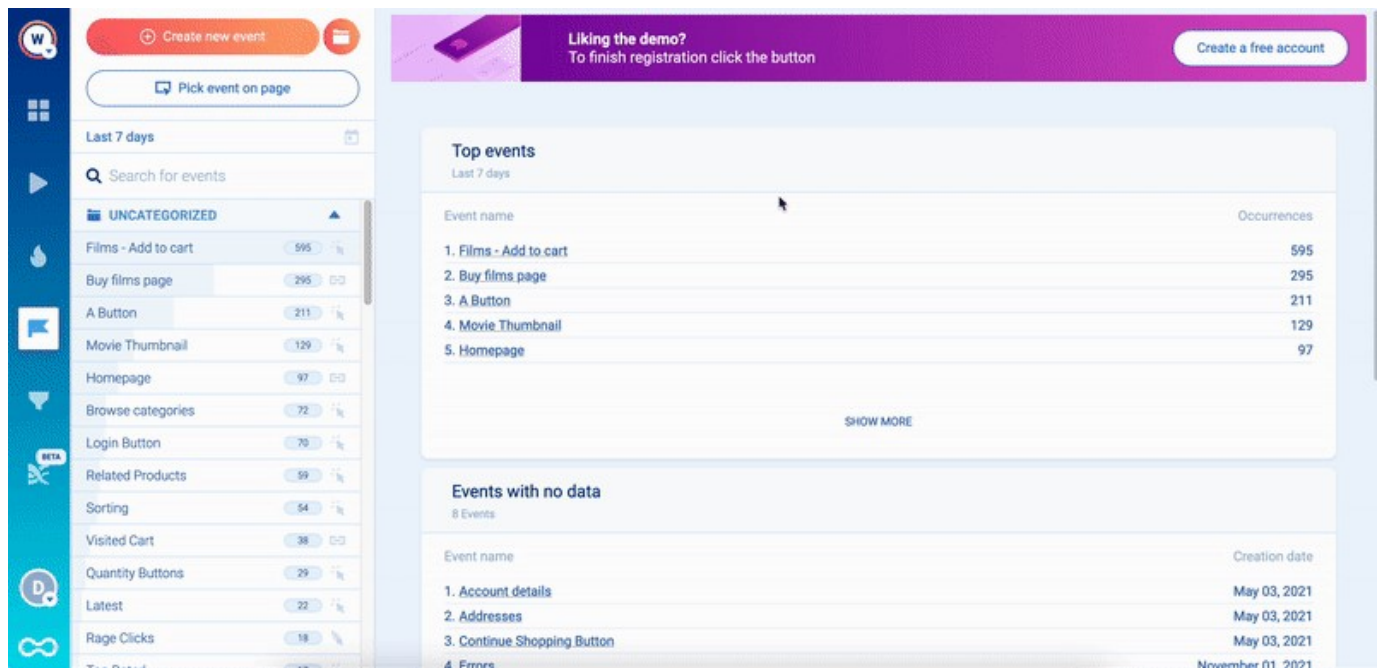
2. Use the *no-code event picker*. When you click on the button that says “Pick event on page” (see screenshot below), you’ll be taken to your web page where you can define an *event* by clicking on elements in the user interface of your website or web app.

3. Define an *event from a recording*. Again, Smartlook’s *session recordings* capture every action visitors do. If you see an action that would be useful to track, you can pause the recording and make that action into an *event* without leaving the replay.

You can also create *custom events* with JavaScript to track pretty much everything else outside of the standard events. For mobile apps, most *events* are typically custom.

On that note, if you have a mobile app and want to see how Smartlook stacks up to other app analytics tools like Adobe Analytics and Mixpanel, check out our review of 11 free and paid mobile app analytics tools.

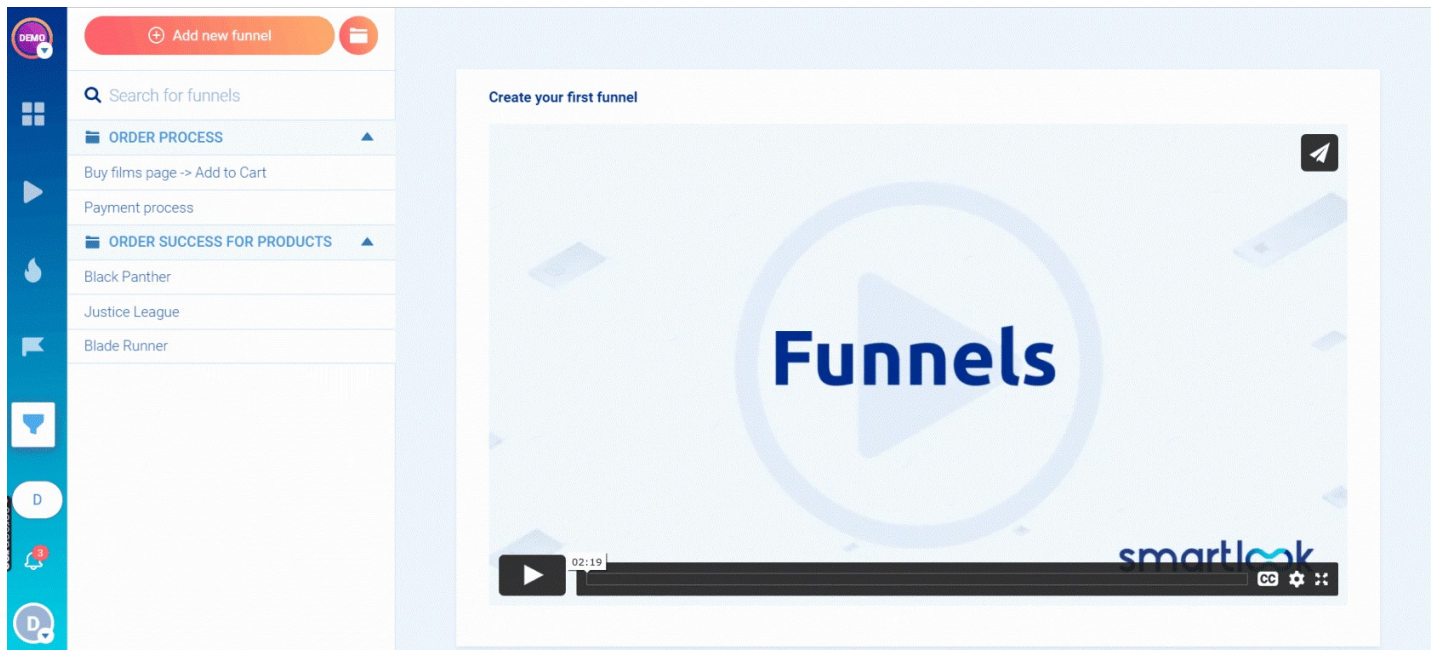
When you define an *event*, Smartlook immediately identifies every instance of that *event*, going back as far as your data retention plan goes. As a result, the *event tracking* visualization appears instantly, like in the GIF below.



With the free plan, you can define up to two *events*. Paid plans allow you to define more *events*, so you can **keep track of all business-critical user actions on your site.**

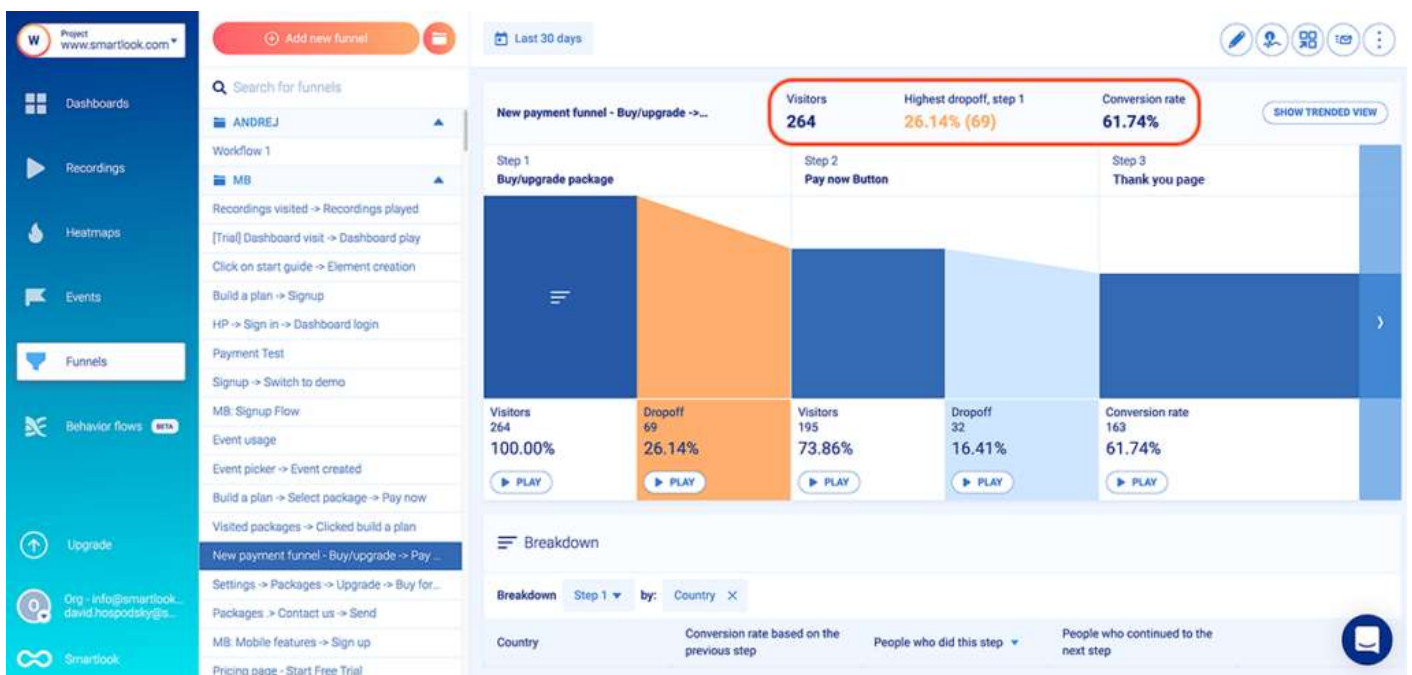
One of the most important use cases for *events* is building *funnels*. *Funnels* are sequences of steps users go through to complete a goal, like purchasing a product or signing up for a newsletter. Building *funnels* lets you analyze users' behavior through these steps, find out where they drop off, and calculate conversion rates.

In Smartlook, you can create a *funnel* by placing two or more *events* in the order you believe your users follow. Then, just like with *event tracking*, the *funnel visualization* appears instantly.



Once you have a *funnel*, you can see its overall conversion rate, as well as the conversion rates and drop-offs between each step.

On the free plan, you can build and monitor one *funnel*. Paid plans allow you to build many more *funnels*, so you can **analyze all the journeys users take to complete different goals on your site.**



For example, an e-commerce store may want to track the customer journey from landing on the homepage, through searching for an item and completing the purchase.

In Smartlook, you can build such a *funnel* with five events, using our *no-code event picker* or our drop-down *event list* (without coding):

- **Event 1:** A homepage visitor clicks on the shop (select the button with our no-code picker to define the *event*).
- **Event 2:** The visitor searches for an item (use the “Typed text” *event*).
- **Event 3:** The visitor clicks “Add to cart” (select the button with our no-code picker to define the *event*).
- **Event 4:** They click on the “Pay Now” button (again, select that button with the *no-code event picker*).
- **Event 5:** They land on the “Thank you” page (use the “Visited URL” standard *event* and enter the page URL).

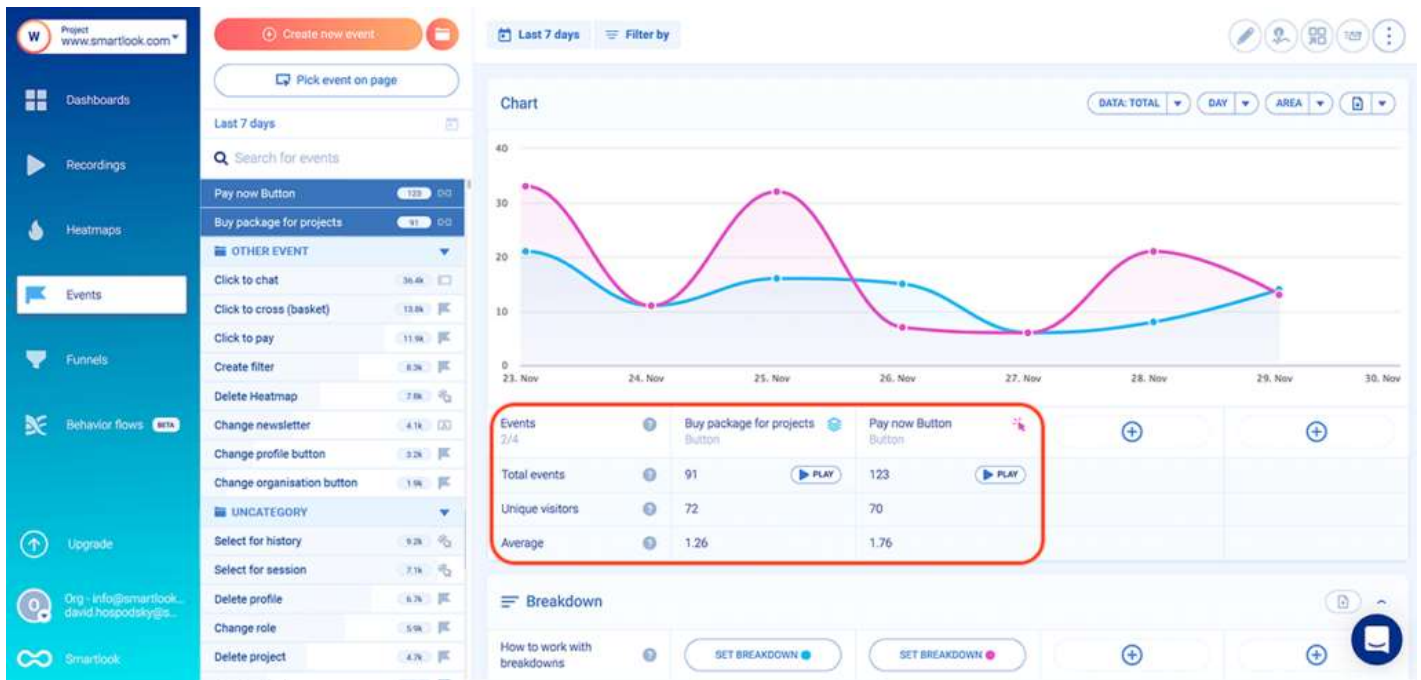
You can see how event tracking and funnel analysis work in detail with Smartlook’s live public demo (no registration or credit card required).

#3 Pair event tracking and funnel analysis with session recordings to find actionable insights fast

So far, we’ve covered use cases that employ either qualitative analytics (*session recordings*) or quantitative analytics (*events* and *funnels*). However, our tool also lets you combine the power of quantitative and qualitative data to uncover even more insights into your users’ behavior.

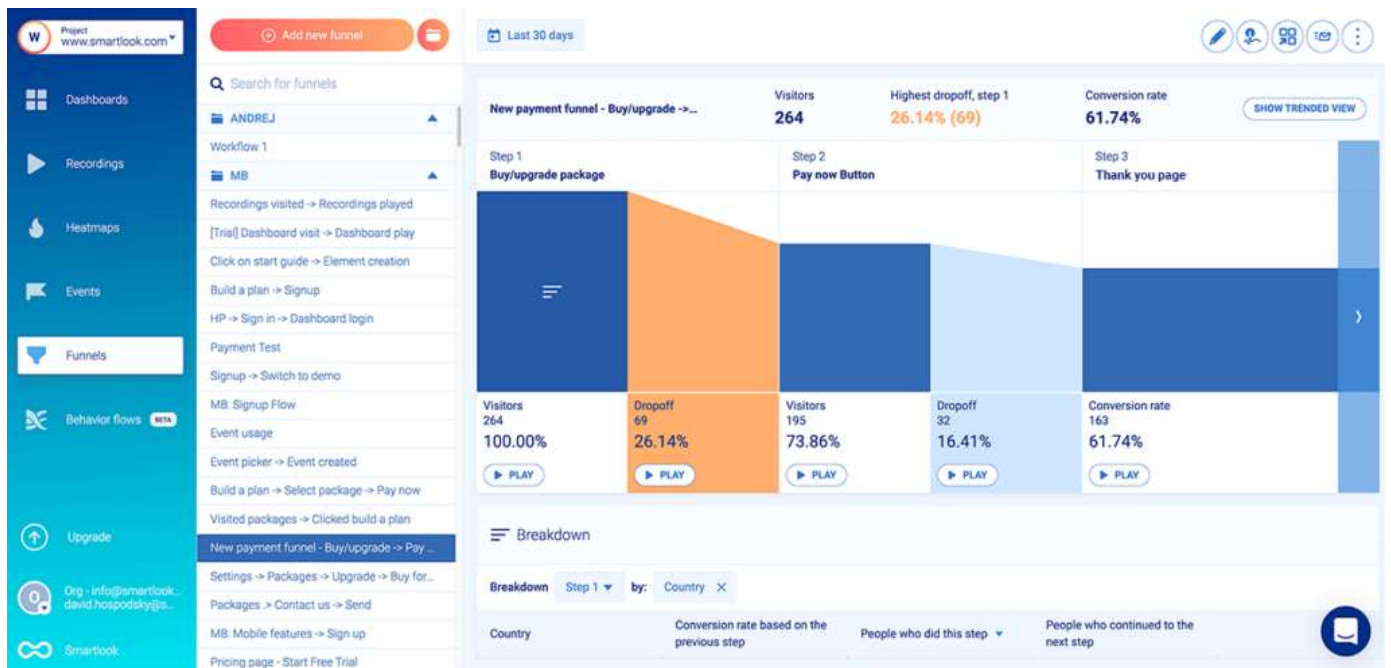
With Smartlook, you can combine *event tracking* and *funnel analysis* with *session recordings*.

When you open an *event*, you can see a “Play” button under the data visualization. Clicking it takes you **directly to all *session recordings* where that *event* took place.**



As a result, you can:

- Instantly find all recordings where a specific *event* took place.
- **Get the full context behind specific user actions**, which isn't possible with quantitative analytics tools. For example, you can watch all sessions where users clicked on "Pay now" and see what they did before and after that.



You can also **combine *funnel analysis* with *session recordings* to see why users drop off.**

The screenshot above shows a funnel where 16.41% (32) of users drop off between clicking “Pay now” and landing on the thank you page. Similar to the events section, there’s a “Play” button under each step of the funnel, including the drop-off stages.

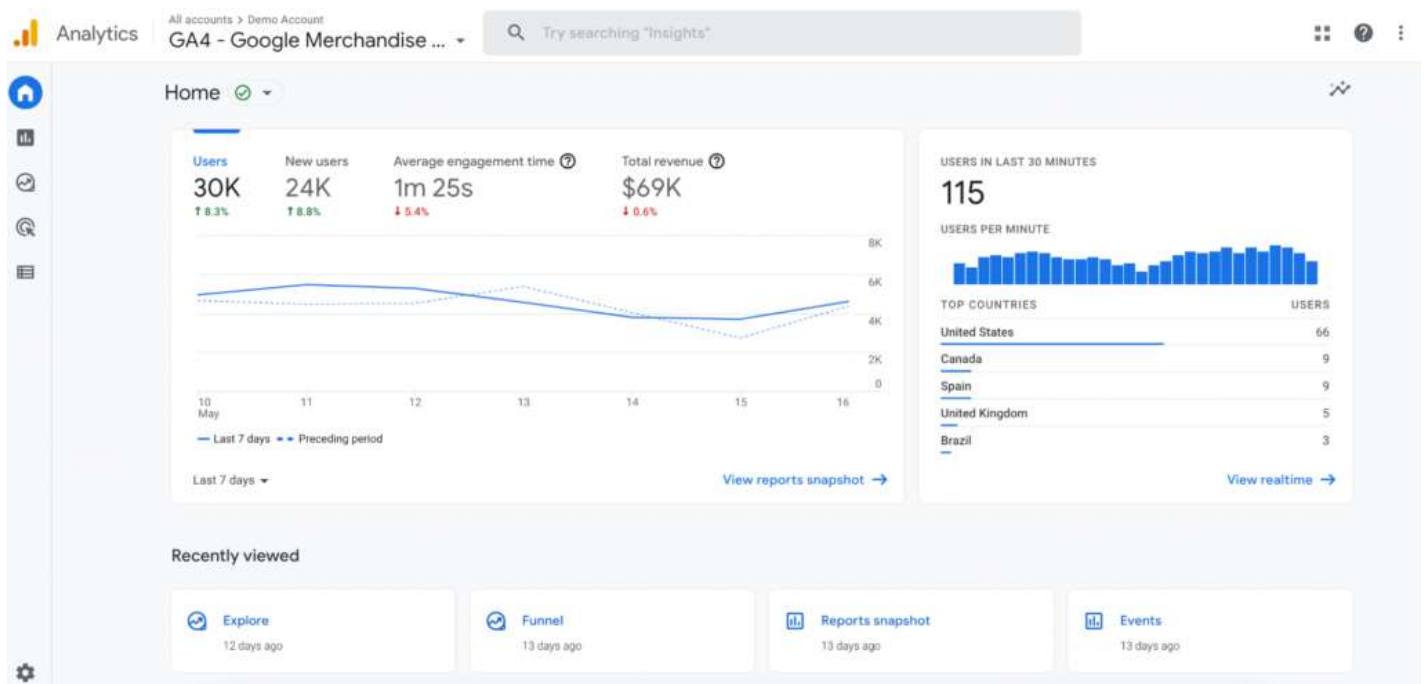
Clicking the “Play” button under the drop-off stage between steps two and three (or any drop-off stage) takes you to the session recordings of these 32 users, so you can find out *why* they dropped off. You don’t have to sift through all 264 session recordings of users who entered the funnel, or even through the 195 sessions of users who clicked “Pay now”.

One of our clients used this process to discover that many users were abandoning their purchase *funnel* right at the last step.

By combining *funnel analysis* with *session recordings*, they were able to deduce that shipping prices were the biggest reason for this drop-off. Once they implemented changes based on their findings, orders increased by 161%, bringing in half a million dollars more in yearly revenue.

To start analyzing the behavior of your website’s users today, sign up for a free Smartlook account (no credit card required).

2. Google Analytics



- **Free plan or free trial:** Google Analytics is free.
- **Self-hosted or SaaS tool:** SaaS tool.
- **Quantitative or qualitative analytics:** Quantitative only.
- **User interaction (event) tracking:** Some events are tracked automatically, but others must be set up manually via code changes or Google Tag Manager.

Google Analytics is the most popular free web analytics tool out there. It's a traditional analytics solution, meaning it provides real-time data about your site's traffic like pageviews, sessions, time on page, bounce rates, and other stats and metrics.

3. Clicky

- **Free plan or free trial:** Both. The free plan works for websites with up to 3,000 daily

clicky Login Register Help

Privacy-friendly Website Analytics

Clicky's [privacy-friendly, GDPR-compliant website analytics](#) service is used by more than one million websites around the globe.

[Register now](#)

[Login](#) [Demo](#) [Learn more](#) [Compare](#)

"It is very nice that Clicky offers real-time tracking. This means that instead of waiting an entire day to see if my goals are being completed according to plan, I can instead look at the stats in real-time and make changes as needed until I am satisfied. Clicky is fast, efficient, affordable, and has great support."

-- Joshua Goring -- [More testimonials >](#)

Monitor, analyze, and react to your traffic in real time

The Basics

	Summary	Visitors	Actions	Uniques	Time
👤 Visitors	Expand	130	+2%		
🔗 Actions	Expand	188	-1%		
📄 Average actions		1.4	-7%		
🕒 Total time		5:04:37	-12%		
🕒 Average time per visit		02:23	-17%		
🚫 Bounce rate		48%	+5%		

Visitors

Compare... ▾ Last 90 days ▾

— Visitors — Previous period

pageviews. All plans start with a free 21-day trial.

- **Self-hosted or SaaS tool:** SaaS tool.
- **Quantitative or qualitative analytics:** Quantitative only.

- **User interaction (event) tracking:** Downloads and outbound link clicks are tracked automatically, but all others must be tracked via JavaScript.

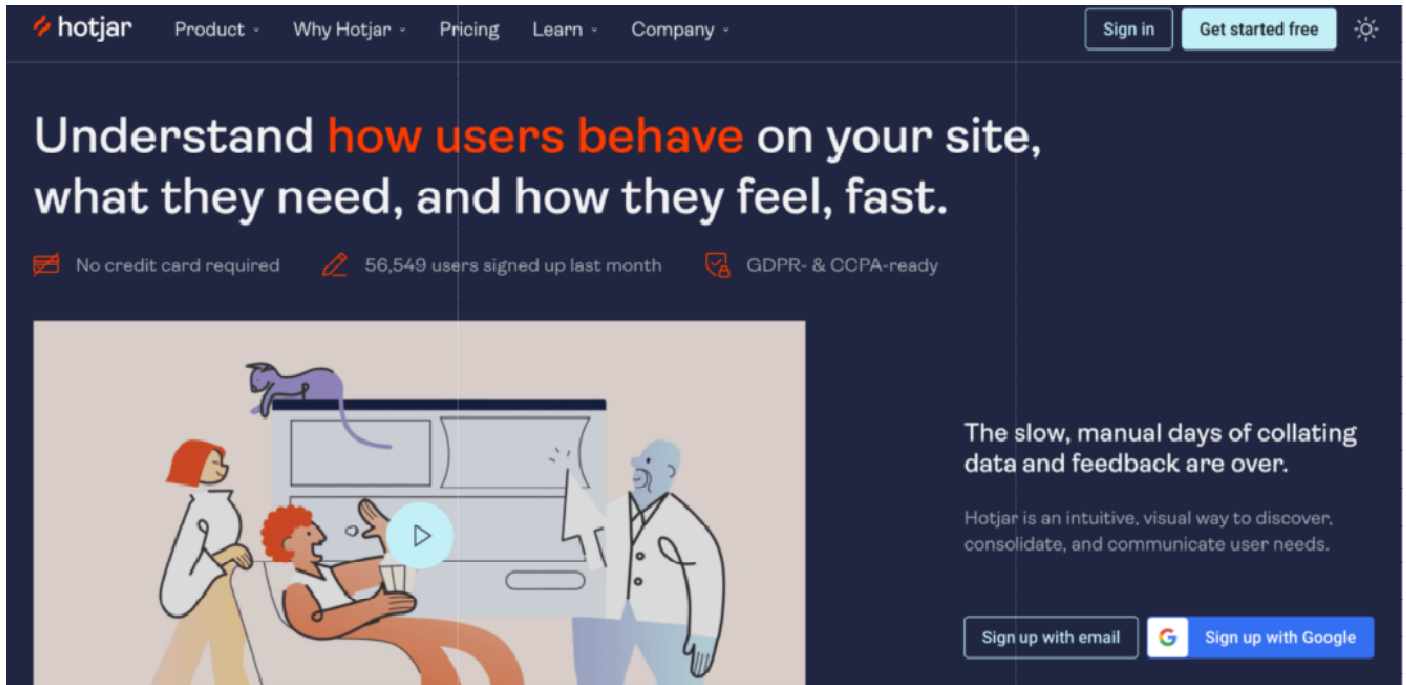
Clicky is a privacy-friendly web analytics software. Similar to Google Analytics, it also provides real-time analytics about your website's traffic.

4. Matomo (formerly Piwik)

- **Free plan or free trial:** Both.
- **Self-hosted or SaaS tool:** Free version is self-hosted, while the paid version is delivered as a SaaS tool and offers more advanced features and plugins.
- **Quantitative or qualitative analytics:** Quantitative only.
- **User interaction (event) tracking:** Not automatic, has to be set up manually.

Matomo advertises itself as a privacy-friendly web analytics platform. Matomo is built to replace Google Analytics, so it comes with similar quantitative analytics features, as well as limitations in the qualitative analytics department.

5. Hotjar

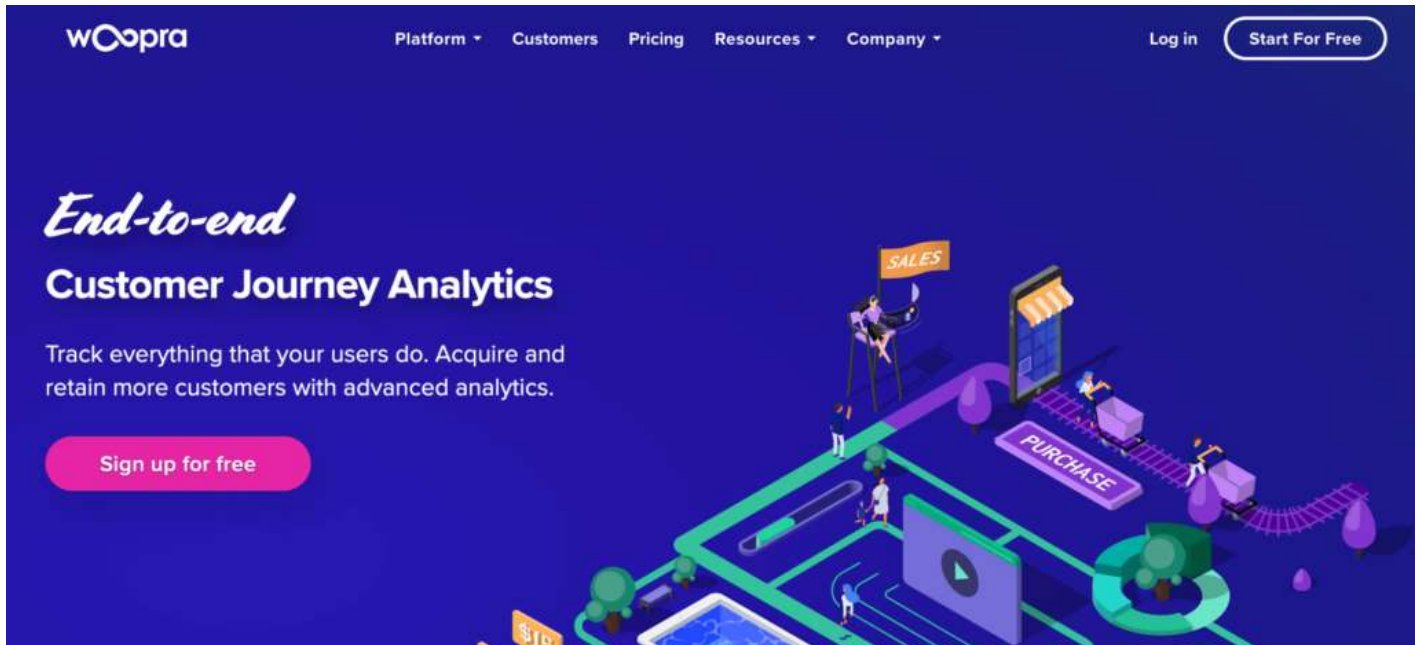


- **Free plan or free trial:** Both. The free plan lets you record up to 35 user sessions/day. All paid plans also start with a free 15-day trial.
- **Self-hosted or SaaS tool:** SaaS tool.
- **Quantitative or qualitative analytics:** Qualitative, with limited quantitative capabilities.
- **User interaction (event) tracking:** Not automatic. You need to write JavaScript code and use their API to track events.

Hotjar is best known for its session recordings and heatmaps tools, but it also has user feedback widgets as well as survey features. Two downsides to this tool, which we've discussed in our Hotjar alternatives article, are its manual event tracking setup and limited quantitative analytics capabilities.

6. Woopra

- **Free plan or free trial:** Free plan for tracking up to 500,000 user actions per month.
- **Self-hosted or SaaS tool:** SaaS tool.



The image shows the Woopra website landing page. The header includes the Woopra logo, navigation links for Platform, Customers, Pricing, Resources, and Company, and buttons for Log in and Start For Free. The main content area features the headline "End-to-end Customer Journey Analytics" and a sub-headline "Track everything that your users do. Acquire and retain more customers with advanced analytics." A prominent pink button says "Sign up for free". The background is a dark blue with a colorful illustration of a customer journey path, including a person climbing a ladder labeled "SALES", a person at a computer labeled "PURCHASE", and various icons representing user interaction and analytics.

- **Quantitative or qualitative analytics:** Quantitative only.
- **User interaction (event) tracking:** Pageviews are tracked automatically, but all other events need to be set up manually.

Woopra is a quantitative analytics tool for tracking your customers' journeys end-to-end. It's more versatile than traditional analytics tools, as it comes with features for product, marketing, sales, and customer support teams.

7. Open Web Analytics



Home About OWA Download Help & Support Screenshots



Control your data.

Web Analytics. Open Sourced.

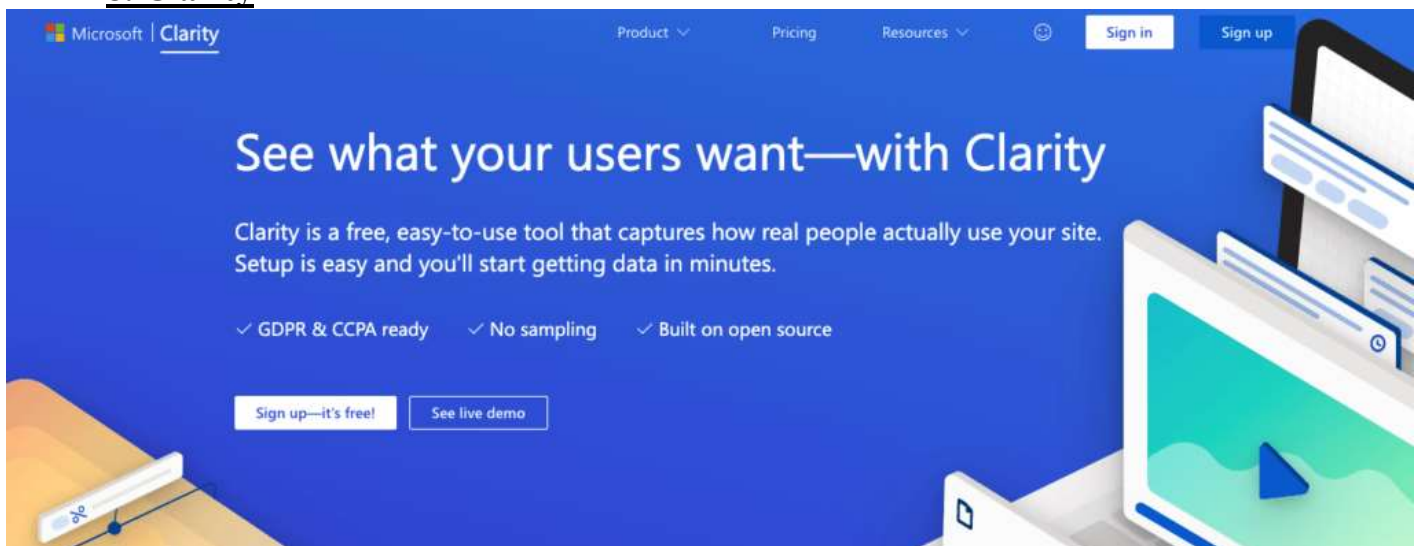
Open Web Analytics is the free and open source web analytics framework that lets you stay in control of how you instrument and analyze the use of your websites and application.

DOWNLOAD FROM GITHUB

- **Free plan or free trial:** Open Web Analytics is a free framework.
- **Self-hosted or SaaS tool:** Self-hosted framework.
- **Quantitative or qualitative analytics:** Quantitative only.
- **User interaction (event) tracking:** Being a framework, Open Web Analytics requires you to set up everything manually, including event tracking.

Open Web Analytics is an open-source framework, which gives you granular control over how you collect and analyze user behavior data. It requires programming skills to set up and run, but it's also very versatile, as you can use it under your own domain or as part of a web app.

8. Clarity



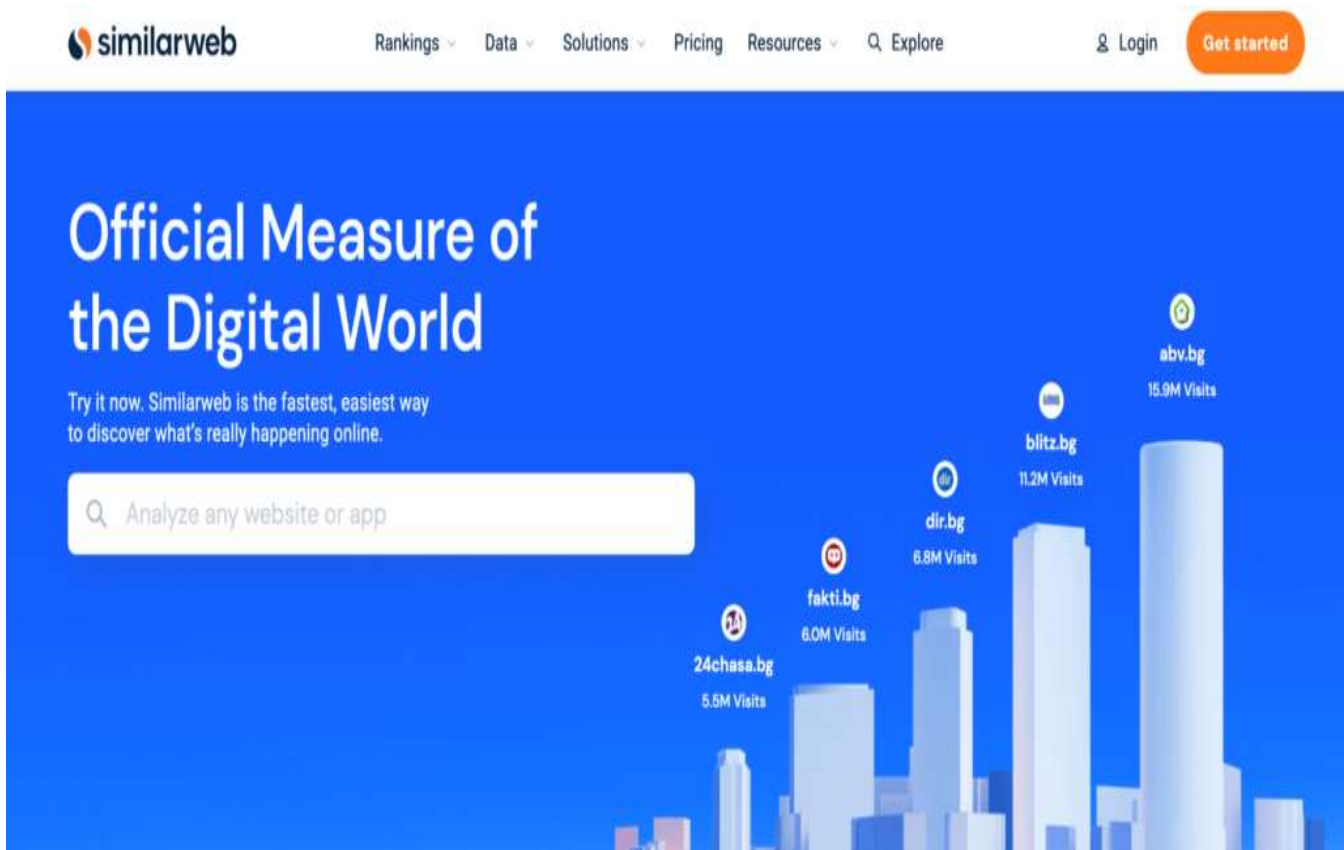
- **Free plan or free trial:** Clarity is a free tool.
- **Self-hosted or SaaS tool:** SaaS tool
- **Quantitative or qualitative analytics:** Qualitative, with limited quantitative capabilities.
- **User interaction (event) tracking:** Clarity's heatmaps tool tracks some user events, like rage clicks and errors, but it doesn't allow you to define your own events.

Clarity is a free user behavior analytics tool from Microsoft that offers session recordings, heatmaps, automated insights, and an integration with Google Analytics.

3 complementary digital marketing tools

The tools in this second category can be a nice addition to your analytics solution, as they provide valuable data about your website's performance. Additionally, they also come with various features and capabilities that many website owners, as well as marketing, product, and customer service teams may find useful.

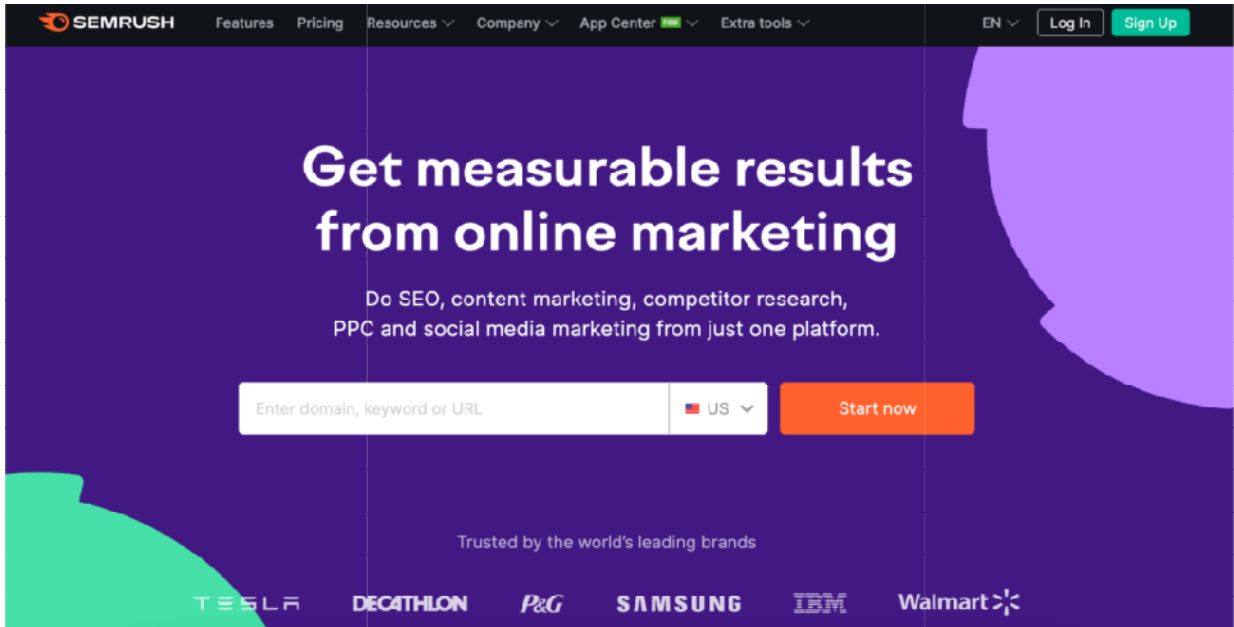
1. Similarweb



- **Free plan or free trial:** A limited version of Similarweb is available for free. Paid plans start with a free 7-day trial.

Similarweb is a popular tool for getting data about web traffic and performance. The tool is really simple to use — you just enter a website URL and Similarweb gives you information about the site's organic rankings, competitors, marketing channels, and more.

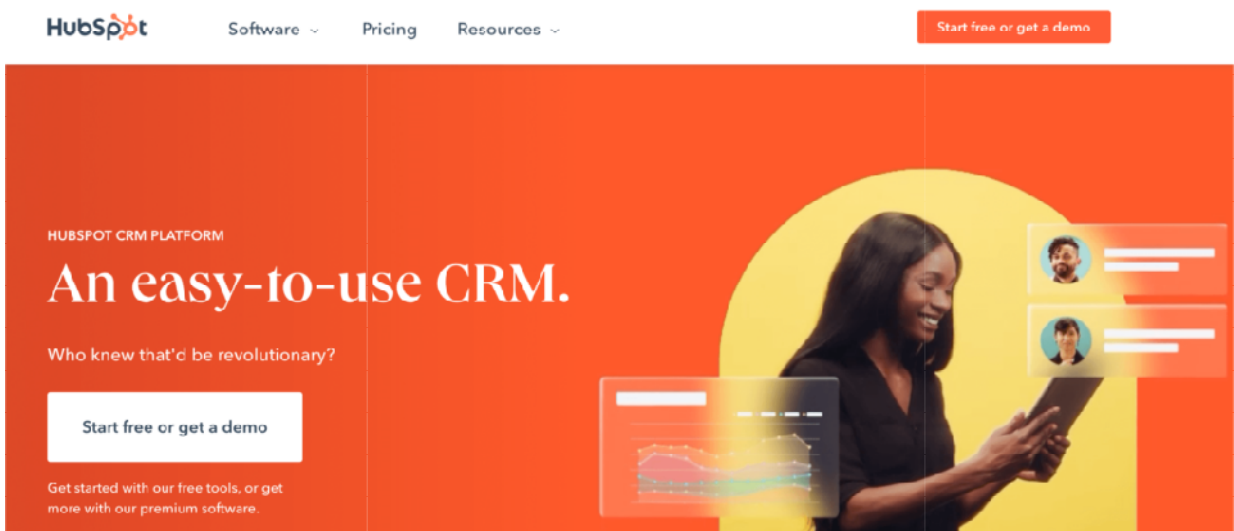
2. Semrush



- **Free plan or free trial:** Free account is available with limited functionalities. All paid plans start with a free 7-day trial.

Semrush is one of the most popular tools for search engine optimization (SEO) and search engine marketing (SEM). The tool specializes in keyword research, backlink and competitor analysis, as well as PPC campaign (Google AdWords) analysis.

3. HubSpot



- **Free plan or free trial:** HubSpot offers various free tools for marketing, sales, customer service, and operations professionals.

HubSpot is the most versatile platform on this list. It became famous for its customer relationship management (CRM) tool, but today, HubSpot's portfolio of solutions has expanded to cover use cases like marketing automation, content management, operations, and more.

Page Rank tools :

What is Google PageRank?

What is Google PageRank? PageRank, if you are involved with Google or Search in a way or you do SEO, you are most likely to come across this term at some point in your life. You are also very likely to get confused about what exactly is a PageRank.

Let's begin with what Google has to say about that. To Google, links are like votes. In addition, it also considers that some votes tend to be more important than the others. PageRank is Google's way of counting the link votes and determine which pages are most important by link votes. Later these scores are used along with various others factors to evaluate if a page will rank well in a search engine or nor.

How Does PageRank Work?

A link analysis algorithm is what PageRank is. It counts the links pointing toward a website or a web page and more importantly the quality of the links or websites. PageRank uses a numerical scale of 0-10 where 0 is the lowest PR while 10 is the highest. It often happens that some websites in their attempt to "cheat the system" end up purchasing links that point to their page hoping that it will make them achieve a high PageRank. However, such unethical approach and low-quality links can result in a negative impact and low Google PageRank. Additionally, a website can be blocked or penalized by the search engine. On the contrary, preference is given to the quality links and content. Therefore, it is important to avoid any such practice and remember that search engines are smart enough to catch you.

Significance of Google PageRank Checker

Why is PageRank important? Because it is one of the factors that search engines like Google will consider when deciding which results to show in search engines listings. In fact, PageRank is a trademark of Google; however, other search engines also use similar methods.

It's not the only or the most crucial factor. To begin with, a page has to be relevant to whatever is a search query, for example, your page about online banking isn't going to show if someone is searching for cooking. Regardless of everything, PageRank does matter in search engine rankings.

It is one of the techniques that Google uses for determining the relevance of a page or its importance. Important pages will get a higher PageRank, and they are also more likely to appear in top results. Google PageRank or Google PR is a scale of 0-10, and it is based on backlinks. The more quality backlinks will result in a higher Google PageRank.

To enhance your PageRank, it is important first to know it. How to find page rank is no longer an issue. To determine your PR, there are many Google PageRank Checker tools available. You can use one of many Google PageRank Checker or PR Checker tools. Most of these tools are free to use, requiring no sign-up or registration.

Google PageRank Checker: How does PageRank Checker Work

How does PageRank work? What is the algorithm behind it? These are the questions that we feel our users are most interested in. As per Google, algorithms are computer programs that look out for clues to give users back what exactly they want. They can be defined as the computer processes or formulas that take questions and return them back as answers; while they generally facilitate the process of entering a query and quickly receive and accurate answer in return.

Google PageRank Checker also uses an algorithm. Google develops its own algorithms to make sure that their users receive SERPs that correctly correlate with the search terms that they input into the search engine. If it isn't for these rigid algorithms, Google will look similar to an older version, when people used questionable SEO techniques for manipulating their place in the SERPs.

Use Google PageRank Checker: Measure the Importance of a Web Page

What is the significance of web PageRank and how to determine it? Many people define the web page importance as the overall visibility of the entire website, in addition to its content. To Google, however, various other factors contribute to the importance of a web page.

PageRank Checker or Google PageRank uses a link analysis algorithm utilized by Google's search engine. This gives a score to each element within a hyperlinked documents set, which in turn, helps to measure the relative importance within a given set. In other words, web pages are ranked by PageRank Checker on the basis of the web pages linking back to a particular web page. That is why the number of incoming links to a webpage is important for its PageRank.

Web pages with a high PageRank help other web pages to achieve a high PageRank, while lower PageRank doesn't help much.

Google PageRank Checker by DupliChecker

Every website owner or service provider in the internet marketing industry must have a favorite PageRank Checker, or PR Checker tool bookmarked because it can serve many purposes. You can use a PR Checker to research the Google Page Rank for each page on your website (because it will vary for each page) or to research competition. You can also use Google PageRank Checker to evaluate the Google PR of a page before advertising or publishing a backlink.

To use Google PageRank Checker offered by DupliChecker, simply enter the URL of a particular web page into the box, hit Enter, and pour yourself a cup of coffee. By the time you get back, you will have your results.

You can only test one page at a time, but it is important to remember that the Google Page Rank for the home page of a website will almost always be higher than the PR of a specific web page on that site. (Sometimes, the reverse is true.) Since PR is based on the number of authority links, you will want to look closely at the PR of a specific page that you want to advertise on (or, other purposes for identifying PR) for accuracy.

pinging & indexing tools :**Ping Bulk/Mass URLs to Search Engines**

Prepostseo ping website tool is widely used by webmaster to submit site to search engines. Pinging your websites URLs to the search engine is very important in many cases. It helps you submit site to google quickly. Google or any other search engine won't be able to know itself about the changes or updates, you have made to your website. So this step is very important. This not only goes for the alteration but as well as, a new URL or web page can also be pinged for the Google to update the database. Search engines normally take the time to recognize and index your data without ping. For example, You have written an article and updated it in your link but didn't submitted it to google. Now, what can happen? It can be stolen by someone else before Google even reach to it and he or she can post it as their own unique content and index it from google. Now where this leads you? Well, your content will be treated as plagiarized if you try to ping it later. So ping your content or new URLs because it's extremely important.

100% Free: With all premium features, our ping urls tool allows you to submit unlimited web pages

65+ Search Engines: Our submit url to google tool submit site to google and 65 others search engines with just one click

Bulk Submission: Bulk ping sites means you can add up to 10 URLs at once to add those web pages to search engines

Don't wait for the search engines to recognize the changes in your website. You must submit online web pages to google, Bing and Yahoo ASAP. Reading following reasons why you should google ping blog urls quickly

To index your new content asap, before someone else grab that content and use it in any other website.

Fresh content improves your ranking in the search engines, so why wait.

This tool can send 650 (10x65) ping url requests at one time hence a lot of your time saved.

Why is crawling and indexing is important?

Crawling and indexing is Google's way of welcoming you to the club. This means that once you are on Google's database, it will start to show you in its search engine results. If your content quality is great and you align with what is the general definition of a decent website, you can get paid through AdSense. But this all is only possible if you appear on Google's search engine. And indexing is the only way to appear on Google's search engine.

How to use Google indexer Tool:

PrePostSEO Google Index Tool is facilitating you widely in this matter. Simply copy/paste the URLs and press the button ping blog to get them instantaneously pinged. This online ping tool can send requests up to the number 650 (65*10) at one time. The use of this tool is very easy and hence it ends up saving you a lot of time. Plus you can paste up to 10 URLs at a single time.

How Online Ping Tool Works:

This tool will fetch urls one by one and then ping those links to the 65+ search engines. Title of the url is used as the name of the pinged url.

Ping backlinks: How it helps in Google Ranking?

If you got a quality backlink from a high authority website, first thing that must be done is to submit that webpage to search engines. Pinging backlinks also become important when a specific webpage is not being crawled by Google, Bing and Yahoo.

After you submitted url for indexation in google, you can check if that specific post listed in search engines or not by using our google indexed pages checker tool. Please note sometimes Google took 24-48 hours to add new URLs in its database

Dead links identification tools:**Websites Broken Link Checker**

Use this Websites Broken Link Checker to identify the broken links on your website easily.

When you click on a link that is supposed to take you to a particular page, yet instead it takes you to another page that shows a 404 error message - this is called a broken link. Listed below are some of the most common causes for a broken link:

- The website is temporarily or permanently unavailable
- The web page has been deleted
- The web page Permalink was modified or changed
- The web page was blocked by firewall or other similar software

Broken links make the user experience very unpleasant, and can damage the reputation of your website. It makes sense that having several broken links on a website is often referred to as “link rot” - because it is as bad as the phrase sounds.

Using this broken link checker will save you the trouble, and you will be able to keep the credibility of your website.

This broken link finder tool will quickly locate any broken links on your website. This way you can correct any errors immediately. It doesn't require you to be SEO expert or webmaster to use this tool because it is very user-friendly.

Anybody can use it countless times, and it comes to you for free!

What is a Broken Link?

A broken link is also often referred to as a dead link. It is a link on a particular page that is already malfunctioning.

You will know if it is a broken link if:

- the website site is always unavailable
- the web page is outdated
- it relocates to a new domain
- it has been removed

Having several broken links to your page is not good especially if you have an online business. If you are a website owner, it is very important to always please your site visitors. You should also make sure that all the links on your website are working so that people will trust your site.

This broken link checker will be of great help if you want to track all broken links on your website and in keeping the links on your web pages up to date.

Seeing broken links on a website can be frustrating to the end user because people are coming to your site with a purpose and if you cannot provide whatever the visitor needs, they will likely move on to the next website and will no longer return to your website because of the bad experience with broken links.

When the website has not been updated for a long period, it can lead to having “link rot” it means that the website contains many broken links. That is why you must find these links using this website broken link checker so you can identify and fix all errors.

Why should you use our Websites Broken Link Checker?

We at Small SEO Tools want to provide you with the best tool that can help you check broken links on your website.

This free online broken link checker is very efficient in identifying link problems. This tool is very user-friendly, it lets you check all broken links on your website so you can correct them.

With this broken link checker tool, we make it easier for you to find all the dead links on your site!

How to use this Broken Link Checker?

Overtime, you get to increase the number of pages on your site which may contain hyperlinks; it will be difficult to keep track on all of them. So, the easiest way to check your website for broken links is by using this broken link checker tool.

To check your website for broken links, all you need to do is to enter the URL in the space provided, and then click on the “Check” button. Our system uses a unique algorithm that will process your request, and it will show the results right away.

This broken link checker is very easy to use, there are no special skills required, and anybody can use it. Very useful tool for website owners, webmasters, and SEO professionals because there is no limit for searches. This online tool is totally free of charge and no registration required.

Why are Broken Links bad for your website?

Having broken links or dead links on a website is not only frustrating, but could affect your website’s reputation as well.

Some webmasters and website owners who don’t update their website regularly might not be even aware that they have these broken links on their site unless a user tells them so. On the other hand, web visitors who are not familiar with broken links would likely think that the problem is on their end; like internet connection problem or faulty system. For mobile users, who may have experienced this broken links would just normally click the back button and move to another site feeling disappointed because they couldn’t load the page on your website.

And with the popularity of social media, people now have the opportunity to leave their comments. So, if they had a bad experience with your website they will most likely say so, and this can further damage the credibility of your website.

We know that traffic is one of the major components of having a good page ranking on search engines that is why you must ensure that all broken links on your websites are cleared as this can help increase traffic on your site.

It is important to update your website regularly especially when using external links because your partner website won't always notify you if they have made any changes or move the link to another location. There is also the possibility of external servers being brought down temporarily or permanently as well as domains could be expired or sold. These are some of the things that you cannot control, but you can make the precautionary action by checking your website regularly with this broken link checker.

All you have to do is to write the URL in the given field then click on the "Check" button. In a matter of seconds, you will get the results.

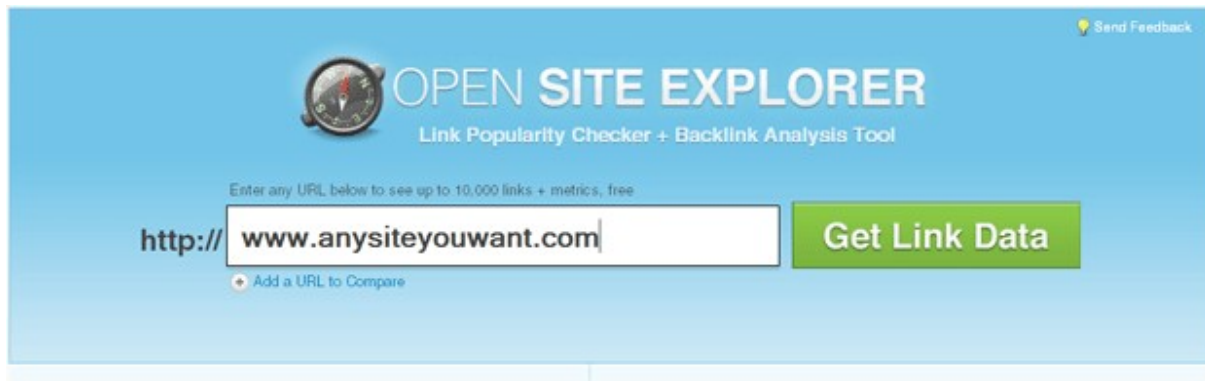
Open site explorer Domain information/ whois tools- Quick sprout :

Introducing New Features for Open Site Explorer

SEO Tools

The author's views are entirely his or her own (excluding the unlikely event of hypnosis) and may not always reflect the views of Moz.

Today I am proud to announce the launch of the second version of Open Site Explorer. Since SEOMoz has officially moved out of consulting, we are now able to put our full resources into building fantastic SEO software. We want to thank all of you who provided feedback on the first version of the tool for your guidance and we look forward to hearing more from you in the future.



Now enough with the chit chat, on to the new features!

New Features:

- Top Pages on a Domain
- Target URL
- Comprehensive CSV Export
- Usability Enhancements (The end of page reloads when applying filters!)
- Improved Filtering

Top Pages on a Domain

With the new version of Open Site Explorer you can get a sorted listed of the top 10,000 pages on a domain. This is essential for viewing your own site and for doing competitive analysis.

-

Page Title URL	HTTP Status	Linking Root Domains	Page Authority
1 [No Title] www.microsoft.com/	302	104,699	99
2 Internet Explorer 8: Home page www.microsoft.com/windows/ie/ie8/default.aspx	200	30,632	96
3 [No Title] www.microsoft.com/windows/ie/ie8/download-ie.aspx	301	22,067	97
4 [No Title] www.microsoft.com/ie/	301	17,100	97
5 [No Title] www.microsoft.com/windows/windowsmedia/download/default.aspx	302	15,945	97

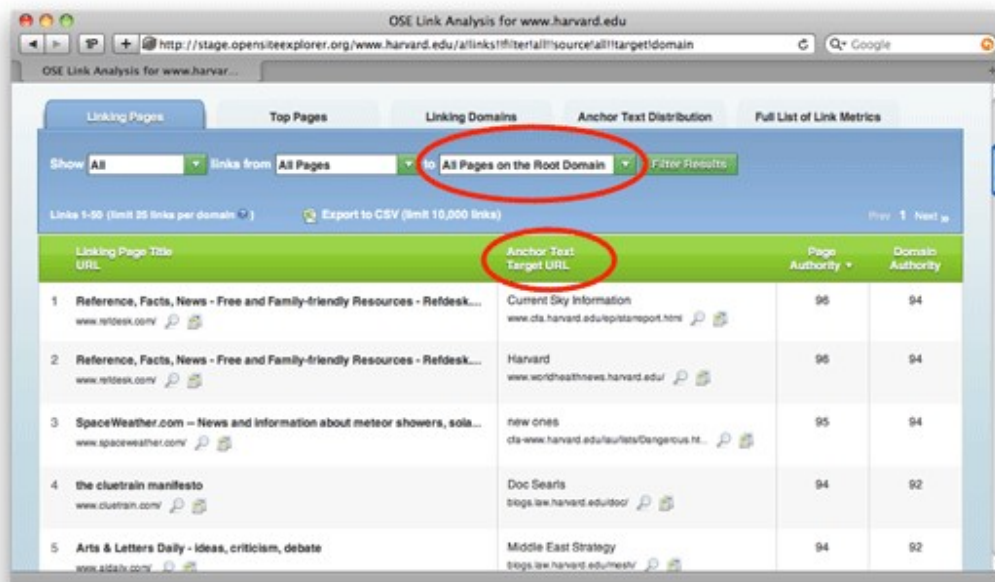
With this new feature, we can see that Microsoft is unwisely_302 redirecting their homepage!
Doh!

Page Title URL	HTTP Status	Linking Root Domains	Page Authority
1 xkcd - A webcomic of romance, sarcasm, math, and language - By Randall Munroe www.xkcd.com/	200	4,923	86
2 xkcd - A webcomic of romance, sarcasm, math, and language - By Randall Munroe www.xkcd.com/256/	200	154	61
3 xkcd - A webcomic of romance, sarcasm, math, and language - By Randall Munroe www.xkcd.com/386/	200	135	66
4 [No Title] www.xkcd.com/256.html	301	96	71
5 [No Title] www.xkcd.com/386.html	No Data	95	59

You can also see which content is drawing the most links on your competitors websites. In this example we see that that these are the most linked to comics on XKCD.

Target URL

The new version of Open Site Explorer shows you which URL a given link is targeting when you sort by sub or root domains so you can see exactly where the given link is helping you. (This is also available for all links when the data is exported as a CSV)



With this new feature you can see which link is most important to Harvard.edu's domain and which page it is linking to.

Comprehensive CSV Export

After lots of input, we are now offering more robust CSV exports.

-

links-page-www.microsoft.com.numbers

Open Site Explorer Report
www.OpenSiteExplorer.org/www.microsoft.com/links/follow/page_authority/g?11&external?page
for the URL:
www.microsoft.com

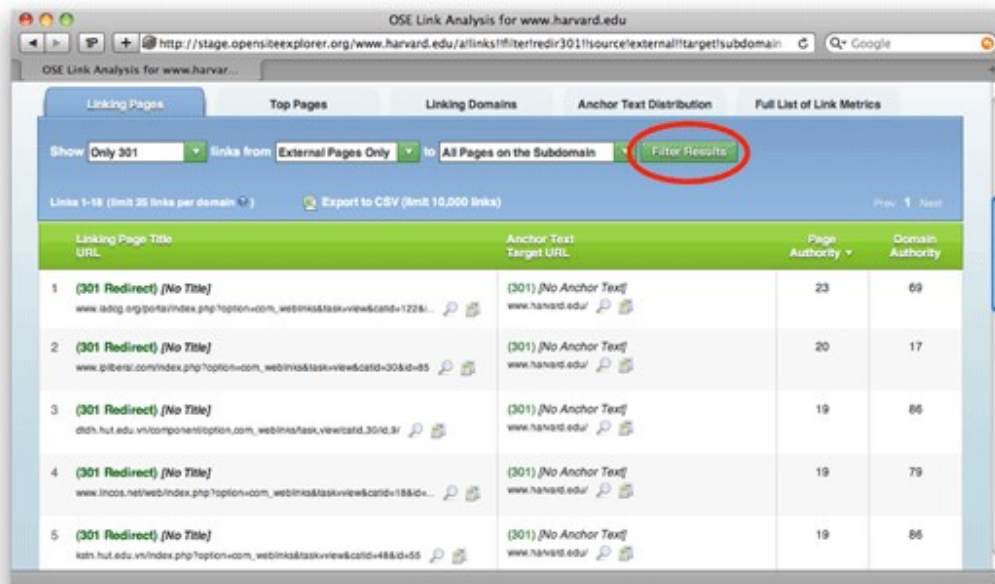
URL	Title	Anchor Text	Page Authority	Domain Authority	Number of Link Followable	301	Origin	Target URL
http://www.aspx Open Archives	MS Logo		97	94	3334 Yes	No	External	http://www.ms
http://www.ig? Award winning	Microsoft.com		96	82	381 Yes	No	External	http://www.ms
http://www.w3. Web Accessibility	Microsoft Corp		94	100	13155 Yes	No	External	http://www.ms
http://www.why Welcome OIG	Microsoft		94	82	835 Yes	No	External	http://www.ms
http://www.w3. Cascading Style	Microsoft		93	100	8605 Yes	No	External	http://www.ms
http://www.nyx Linux OS SUS	Microsoft		92	97	3125 Yes	No	External	http://www.ms
http://www.e2c Enterprise 2.0	Microsoft		92	87	782 Yes	No	External	http://www.ms
http://www.ig? ICANN ICANN	Microsoft Corp		92	96	1910 Yes	No	External	http://www.ms
http://www.ig? Welcome to Site	No Anchor Text		91	92	2212 Yes	No	External	http://www.ms
http://www.org Creative eWorld IE	5.5		91	74	326 Yes	No	External	http://www.ms
http://www.w3. CSS3 module	Microsoft Corp		91	100	1643 Yes	No	External	http://www.ms
http://www.ark Jerkoff	Microsoft		91	86	408 Yes	No	External	http://www.ms
http://www.igs OmegaDrivers	Microsoft Corp		91	84	1180 Yes	No	External	http://www.ms
http://www.als Sheldon Brown	Microsoft		91	93	994 Yes	No	External	http://www.ms
http://www.myl MyLife Support	Microsoft		91	92	1487 Yes	No	External	http://www.ms
http://www.3dn 3DNews - Daily	Microsoft		90	92	2964 Yes	No	External	http://www.ms
http://www.igs Goethe H	Microsoft		90	87	1101 Yes	No	External	http://www.ms
http://www.igs DVD FAQ	Microsoft		90	87	1427 Yes	No	External	http://www.ms
http://www.igs eLearning Africa	Microsoft		90	86	511 Yes	No	External	http://www.ms
http://www.igs Stephen Hawk	Internet Explorer		90	100	776 Yes	No	External	http://www.ms
http://www.igs Frequently Ask	Microsoft		90	100	374 Yes	No	External	http://www.ms
http://www.igs Home - Contents	Microsoft.com		90	89	1879 Yes	No	External	http://www.ms
http://www.igs Search.com -	Microsoft		90	84	422 Yes	No	External	http://www.ms
http://www.igs Franklin - Site	Microsoft		90	81	280 Yes	No	External	http://www.ms
http://www.igs Swinick Bible	Microsoft Explic		90	84	423 Yes	No	External	http://www.ms
http://www.igs The CoverPage	Microsoft Corp		89	94	1216 Yes	No	External	http://www.ms

The new CSV exports offer:

- The Target URL of the given link
- Numbers of links to the given source page
- Indication of whether or not the linked is followed
- Indication of whether the link is internal or external

Usability Enhancements

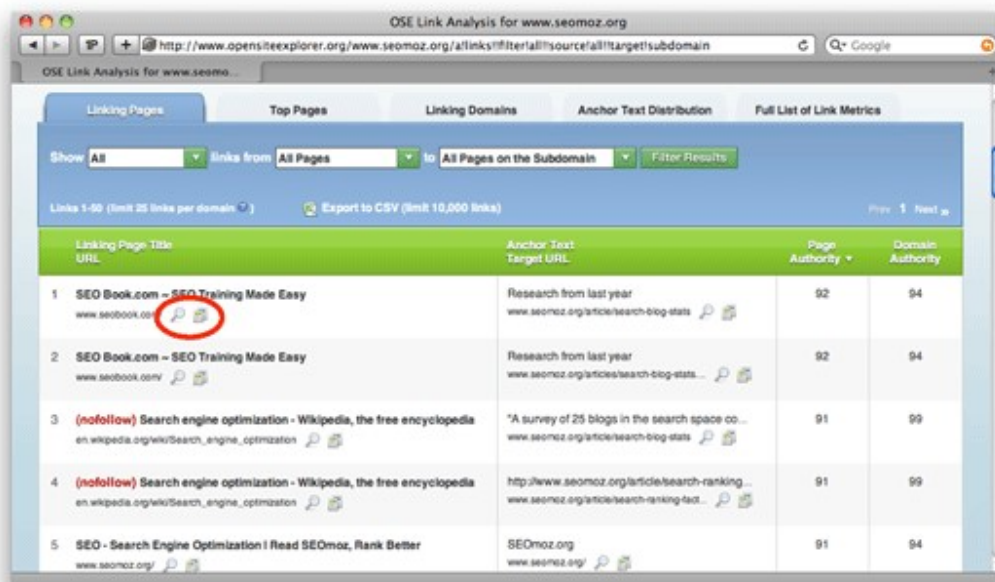
Remember how you used to have to reload the page every time you applied a filter in Yahoo! Site Explorer?



With the addition of the Filter Results button, these needless page reloads are a thing of the past.

Common Tasks are Easier to Perform

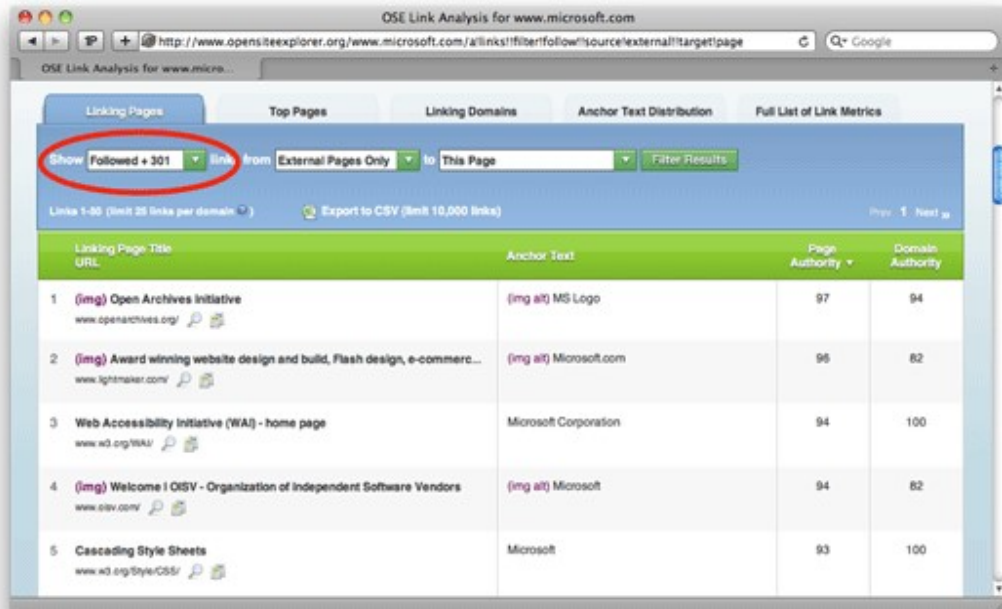
New buttons make performing common tasks easier and faster to do.



The new Explore and Compare buttons make it easier to get more information about any links you find interesting.

Improved Filtering

With the new version of this tool you can do even more filtering to drill down into what you think is important.



In this example, we filtered the data to show only followed (dofollow) and 301 redirecting external links to the specific page.

Open Site Explorer vs. Linkscape vs. Yahoo! Site Explorer

Throughout this process, we also heard a lot of questions about the differences between Yahoo! Site Explorer, Open Site Explorer and Linkscape. The chart below lays out the similarities and differences.

Feature	Yahoo! Site Explorer	Open Site Explorer	Linkscape
Links Available per Report	1,000	10,000	3,000
Filter Followed, Nofollowed & 301s	No	Yes	Yes
See Anchor Text Passed	No	Yes	Yes
Refine by Linking Domain TLD	Yes <small>(via search queries)</small>	No	Yes
Side-by-Side Comparison Reports	No	Yes	Yes
Link Metrics Available	None	DA / PA	Many
Search/Sort by Keywords in URL	Yes <small>(via search queries)</small>	No	Yes
Search/Sort by Anchor Text KWs	No	No	Yes
Search/Sort by Keywords in Title	Yes <small>(via search queries)</small>	No	Yes

SYLLABUS:

Web to lead forms- Web to case forms- Lead generation techniques- Leads are everywhere- Social media and lead gen Inbuilt tools for Digital Marketing-ip Tracker- CPC reduction (in case of paid ads) Group posting on Social Media platforms

Web-to-lead

A web-to-lead form is an essential component of marketing and sales automation. The aim is to capture data presented by website visitors, such as contact information and product interest, and store it as a “lead” record in a CRM product, in this case Salesforce.

The use cases are many, but the concept is similar. A visitor to your website is presenting contact information on your website and this submission creates a record as a lead in Salesforce. It is a way to get feedback on your product and services or to grow your marketing database.

Steps to Create Web-to-lead

Go to Setup -> Search for Lead -> Click web-to-Lead -> then click Create Web-to-Lead form

Web-to-Lead Setup

[Help for this Page](#)

Using pre-existing pages on your company's website, you can capture contact and profile information from users and automatically generate new leads in salesforce.com, enabling you to respond in real-time to customer requests.



Web-to-Lead Settings Edit Create Web-to-Lead Form

Web-to-Lead Enabled

Require reCAPTCHA Verification

Default Lead Creator Sapna Chandani

Default Response Template

Get Info Before You Start

- [What is the maximum number of leads I can capture?](#)
- [How do I specify which information to capture?](#)
- [Can I capture leads from multiple web pages?](#)
- [What status is assigned to web-generated leads?](#)

Select fields and Enter Return URL. If you want Captcha Enable **“Include reCAPTCHA Key Pair”** otherwise deselect it.

Web-to-Lead Setup

[Help for this Page](#)

Easily set up a page on your website to capture new leads.

Create a Web-to-Lead Form

Select the fields to include on your Web-to-lead form:

<p>Available Fields</p> <ul style="list-style-type: none"> Salutation Title Website Phone Mobile Fax Street Zip Country 	<p>Add</p> <p>▶</p> <p>Remove</p> <p>◀</p>	<p>Selected Fields</p> <ul style="list-style-type: none"> First Name Last Name Email Company City State/Province 	<p>Up</p> <p>▲</p> <p>Down</p> <p>▼</p>
---	--	---	---

NOTE: Would you like to add custom fields that you do not see listed under Available Fields? You can set up custom lead fields to gather additional information from your website. [Tell me more.](#)

After users submit the Web-to-Lead form, they will be taken to the specified return URL on your website, such as a "thank you" page.

Return URL

Include reCAPTCHA in HTML [i](#)

For reCAPTCHA you need to go to this link

<https://developers.google.com/recaptcha/docs/settings>

By logging into your Gmail account you can create key value pairs for reCAPTCHA

Here, I am not including recaptcha.

Create a Web-to-Lead Form

Copy and paste the sample HTML below and send it to your webmaster.

```

<!-- -----> -->
<!-- NOTE: Please add the following <META> element to your page <HEAD>. -->
<!-- If necessary, please modify the charset parameter to specify the -->
<!-- character set of your HTML page. -->
<!-- -----> -->

<META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=UTF-8">

<!-- -----> -->
<!-- NOTE: Please add the following <FORM> element to your page. -->
<!-- -----> -->

<form action="https://webto.salesforce.com/servlet/servlet.WebToLead?encoding=UTF-8"
method="POST">

<input type="hidden" name="oid" value="00D7F000003ggss">
<input type="hidden" name="retURL" value="http://www.google.com">
        
```

Finished

Create a Web-to-Lead Form

Copy and paste the sample HTML below and send it to your webmaster.

```
size="20" type="text" /><br>  
<label for="email">Email</label><input id="email" maxlength="80" name="email" size="20"  
type="text" /><br>  
<label for="company">Company</label><input id="company" maxlength="40" name="company"  
size="20" type="text" /><br>  
<label for="city">City</label><input id="city" maxlength="40" name="city" size="20"  
type="text" /><br>  
<label for="state">State/Province</label><input id="state" maxlength="20" name="state"  
size="20" type="text" /><br>  
<input type="submit" name="submit">  
</form>
```

Finished

This is the complete code:

Create this form in Your PC and save it in **.html** format.

```
<!-- ----- -->  
<!-- NOTE: Please add the following <META> element to your page <HEAD>. -->  
<!-- If necessary, please modify the charset parameter to specify the -->  
<!-- character set of your HTML page. -->  
<!-- ----- -->  
  
<META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=UTF-8">  
  
<!-- ----- -->  
<!-- NOTE: Please add the following <FORM> element to your page. -->  
<!-- ----- -->
```

```
<form action="https://webto.salesforce.com/servlet/servlet.WebToLead?encoding=UTF-8"
method="POST">
```

//Here Oid is Organization Id or org id, To Navigate Go to Setup > Administration Setup > Company Profile > Company Information – you’ll see your OID listed as a field on that page as well.

```
<input type=hidden name="oid" value="00D7F000003ggss">
```

```
<input type=hidden name="retURL" value="http://www.google.com">
```

//If the debug value is set to “1” you will see a confirmation after you submit your lead like the one below. You will want to remove this once you go live – it’s just for testing.

```
<!-- ----- -->
```

```
<!-- NOTE: These fields are optional debugging elements. Please uncomment -->
```

```
<!-- these lines if you wish to test in debug mode. -->
```

```
<!-- <input type="hidden" name="debug" value=1> -->
```

```
<!-- <input type="hidden" name="debugEmail" -->
```

```
<!-- value="sapnagulabchandani23@gmail.com"> -->
```

```
<!-- ----- -->
```

```
<label for="first_name">First Name</label><input id="first_name" maxlength="40"
name="first_name" size="20" type="text" /><br>
```

```
<label for="last_name">Last Name</label><input id="last_name" maxlength="80"
name="last_name" size="20" type="text" /><br>
```

```
<label for="email">Email</label><input id="email" maxlength="80" name="email" size="20"
type="text" /><br>
```

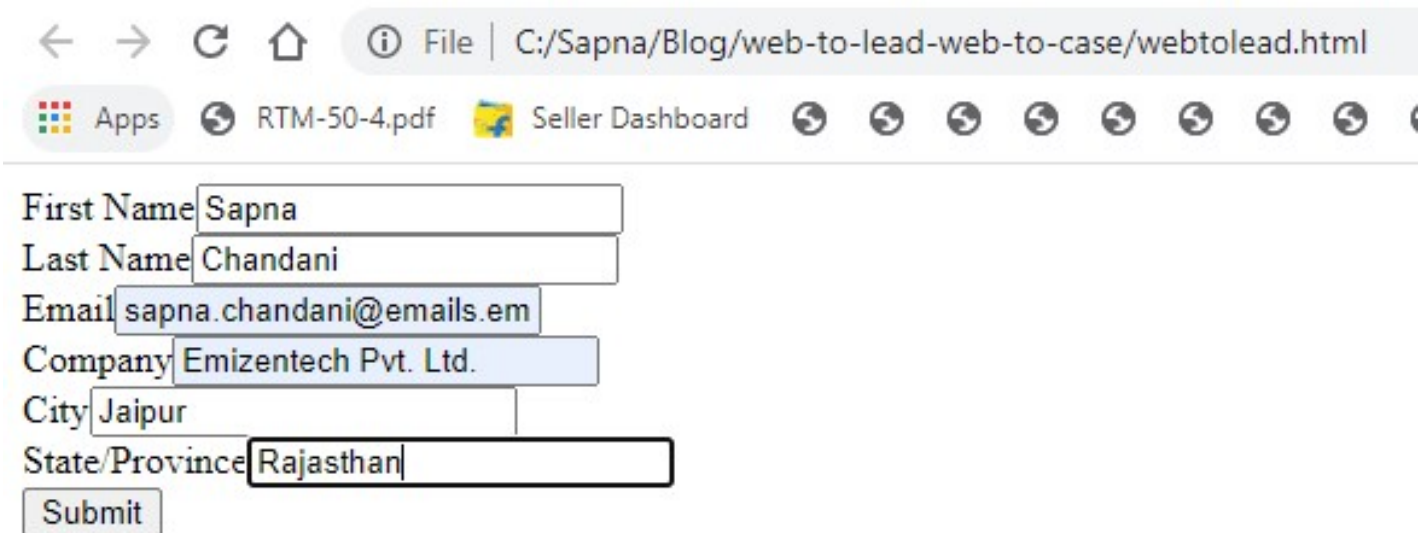
```
<label for="company">Company</label><input id="company" maxlength="40" name="company" size="20" type="text" /><br>
```

```
<label for="city">City</label><input id="city" maxlength="40" name="city" size="20" type="text" /><br>
```

```
<label for="state">State/Province</label><input id="state" maxlength="20" name="state" size="20" type="text" /><br>
```

```
<input type="submit" name="submit">
```

```
</form>
```



Here you can find the Lead in Salesforce after submission of Lead.

There are no updates.

[Open Activities \(0\)](#) | [Activity History \(0\)](#) | [Campaign History \(0\)](#) | [HTML Email Status \(0\)](#)

Lead Detail

[Edit](#) [Delete](#) [Convert](#) [Clone](#) [Find Duplicates](#)

Lead Owner	Sapna Chandani (Change)	Phone	8522256355
Name	Sapna Chandani	Mobile	8765454658
Company	Emizentech	Fax	
Title		Email	sapna.chandani@emizentech.com
Lead Source		Website	
Industry		Lead Status	Open - Not Contacted
Annual Revenue		Rating	
		No. of Employees	
Address			
Product Interest		Current Generator(s)	
SIC Code		Primary	
Number of Locations			
Created By	Sapna Chandani , 5/26/2020 7:09 AM	Last Modified By	Sapna Chandani , 5/26/2020 7:09 AM
Description			

[Edit](#) [Delete](#) [Convert](#) [Clone](#) [Find Duplicates](#)

Also Read: How To Do Component Communication In LWC In Salesforce

Web-to-Lead Setup

Easily set up a page on your website to capture new leads.

Create a Web-to-Lead Form

Select the fields to include on your Web-to-lead form:

Available Fields		Selected Fields	
Salutation	Add 	First Name	Up
Title		Last Name	
Website	Remove 	Email	Down
Phone		Company	
Mobile		City	
Fax		State/Province	
Street			
Zip			
Country			

NOTE: Would you like to add custom fields that you do gather additional information from your website. [Tell me](#)

After users submit the Web-to-Lead form, they will be taken to the specified return URL on your website, such as a "thank you" page.

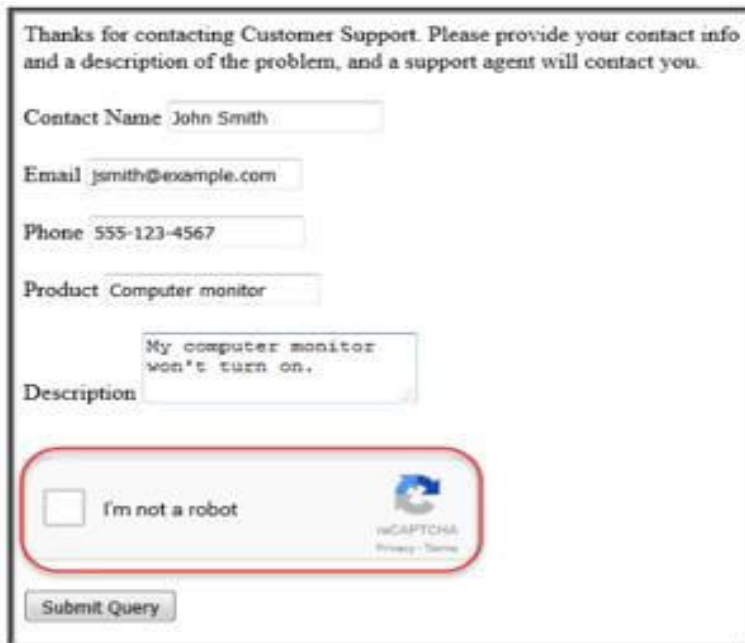
Return URL

Enable spam filtering (recommended)

reCAPTCHA API Key Pair

Enable server fallback

[Generate](#) [Cancel](#)

If Added reCaptcha**You will find this form**

Thanks for contacting Customer Support. Please provide your contact info and a description of the problem, and a support agent will contact you.


Contact Name

Email

Phone

Product

Description

I'm not a robot 

Web-to-case

Web to Case is a tool through which you can post a simple, random web page that allows your customers. To submit a case directly to your Salesforce.com instance. This means that you can present a public case

Page on your website with your own branding and style.

When your customer posts a case via Web to Case, a few fields are generally required, most importantly the name and the email address.

Prerequisites for Web-to-case are

- 1. Create a custom case field if necessary.
- 2. To create a default email template for automatic notification that will be sent to your customers upon submission matter.

- 3. Cases Create case queues if you want to assign incoming cases to queues as well as individual users.
- 4. The Customize support settings to select the default owner of cases that do not meet the criteria in your assignment Rule.
- 5. Generated Create an active case assignment rule to ensure how web-generated cases are assigned to or inserted into users
- 6. Queues. If you do not set an active assignment rule, all web-generated cases are assigned to the default owner
- 7. You specify in the support settings.

For Web-to-case Settings

Go to Setup -> Quick Find Box -> web-to-case

Web-to-Case Settings

[Help for this Page](#)

Use a simple web form or a self-service customer community to make it easy for customers to submit cases directly to your customer support group. For detailed information on setting up Web-to-Case or Self-Service Community templates, see the Salesforce help.

Basic Settings

Enable Web-to-Case [i](#)

Require reCAPTCHA Verification [i](#)

Default Case Origin Web [i](#)

Auto-Response Email Settings

Default Response Template Support: Case Created (Wet) [i](#)

Hide Record Information [i](#)

Email Signature [i](#)

Save Cancel

Auto-Response Email Settings

Default Response Template

--None--
Phone
Email
Web
Support: Case Created (Wet) [i](#)

Here you can require the Captcha, you can Select Case Origin from Phone, Email or Web.

Capture Cases

Select the fields to include:

Available Fields		Selected	
Company	Add ▶	Contact Name	Up ▲
Type		Email	
Status	Remove ◀	Phone	Down ▼
Case Reason		Subject	
Priority		Description	
Engineering Req Number			
SLA Violation			
Product			
Potential Liability			

NOTE: Would you like to add custom fields that you do not see listed under Available Fields? You can set up custom lead fields to gather additional information from your website. [Tell me more.](#)

Visible in Self-Service Portal

Enter the URL that the user will be returned to:

URL

Include reCAPTCHA in HTML [i](#)

To create Web-to-case html form

Go to Setup -> Self-Service or Search for Web-to-Case-HTML-Generator -> Then to create form Select Fields for Case -> Return Url -> reCAPTCHA Key Pair if needed.

Then Click on Generate. It will Generate this form:

```

<!-- ----- -->

<!-- NOTE: Please add the following <META> element to your page <HEAD>. -->

<!-- If necessary, please modify the charset parameter to specify the -->
<!-- character set of your HTML page. -->

<!-- ----- -->

<META HTTP-EQUIV="Content-type" CONTENT="text/html; charset=UTF-8">

<!-- ----- -->

<!-- NOTE: Please add the following <FORM> element to your page. -->

<!-- ----- -->

<form action="https://webto.salesforce.com/servlet/servlet.WebToCase?encoding=UTF-8"
method="POST">

<input type="hidden" name="orgid" value="00D7F000003ggs">
    
```

```
<input type=hidden name="retURL" value="http://www.google.com">
<!-- ----- -->
<!-- NOTE: These fields are optional debugging elements. Please uncomment -->
<!-- these lines if you wish to test in debug mode. -->
<!-- <input type="hidden" name="debug" value=1> -->
<!-- <input type="hidden" name="debugEmail" -->
<!-- value="sapnagulabchandani23@gmail.com"> -->
<!-- ----- -->
<label for="name">Contact Name</label><input id="name" maxlength="80" name="name"
size="20" type="text" /><br>
<label for="email">Email</label><input id="email" maxlength="80" name="email" size="20"
type="text" /><br>
<label for="phone">Phone</label><input id="phone" maxlength="40" name="phone"
size="20" type="text" /><br>
<label for="subject">Subject</label><input id="subject" maxlength="80" name="subject"
size="20" type="text" /><br>
<label for="description">Description</label><textarea name="description"></textarea><br>
<input type="submit" name="submit">
</form>
```

← → ↻ 🏠 ⓘ File | C:/Sapna/Blog/web-to-lead-web-to-case/webtocase.html

📱 Apps 📄 RTM-50-4.pdf 📂 Seller Dashboard 🔄 🔄 🔄 🔄 🔄 🔄 🔄 🔄

Contact Name


Email


Phone

Subject

Description

Here is the case, created from Web



 00001031

Subject: Integration with Revel System

Priority:

Status: New

Case Number: 00001031

[Comment](#) · [Like](#) · Today at 5:10 AM

If you ever require assistance for a complicated project based on salesforce then our team of professional and adroit sales force and developers will be perfect. We have been providing salesforce development services to clients all over the world for a long time and have been able to help them achieve their goals.

Lead generation techniques:

What is Lead Generation?

Lead generation is the process of cultivating a qualified buyer’s interest for your product or service with the goal of converting that person into a customer. Most often, that is done by collecting their information through a form on your website so you can follow up with them after they leave your site.



You can generate leads at any stage of your sales funnel. In fact, you should make sure not to focus too intensely on any one stage. Ideally, you want to have ways to reach customers who are in the Awareness, Evaluation, *and* Conversion stages.

If you only focus on building awareness of your brand, you won't be able to convert as many of those leads into customers. Likewise, if you optimize your Conversion strategies but neglect Awareness and Evaluation, you won't have any leads to convert at all.

Creative Ways to Generate Leads

The strategies below work for leads at any stage of the customer journey. You can also adapt these for almost any industry. Read this list with an open mind and imagine how you could use each strategy for your business.

Since this is a long list, we've broken it into sections for you based on overall strategy type. Please use the table of contents to navigate to the section that interests you most:

Table of Contents

- Compliments
- Offline Lead Generation Ideas
- Gifts and Giveaways
- Time and Attention
- Generosity

Compliments

People love hearing good things about themselves. One of the best compliments you can give is to show that your business values your customers' opinions. Here are some ways to use kind words to generate leads:

1. Ask for a Referral

Referrals are an easy way to make your customers feel like you care about them. When you ask for a referral, tell them how much you enjoyed working with them. Then ask if they know anybody who might also benefit from your product or service. Add a little extra flattery by saying how much you'd love to work with someone else like them.

You can offer a reward for referrals such as an extra product, complimentary service, or credit toward future purchases. If they are genuinely pleased with their experience, many people are happy to give referrals without any promised reward.

2. Request a Testimonial or Case Study

Another easy way to compliment your customers is ask them for a testimonial or case study. This shows that you take their opinion seriously. People enjoy talking about positive experiences they've had. Quoting them on your website can make them feel influential. This may motivate them to share about your business. Their network will see the results they got and perhaps be interested in working with you too.

Learn how to write an epic case study or check out our case study library to see how it's done.

3. Respond to Social Media and Customer Support

Everyone wants to feel seen and heard. When you respond positively to public comments on social media, you show that you're paying attention to your customers. Likewise, timely and helpful customer support can set high expectations for working with you. Keeping your existing customers happy is a great way to generate referrals and repeat purchases.

4. Feature Real Users in Your Marketing

Give your best customers a claim to fame by featuring their words and images in your marketing. This shows potential leads how they might benefit from your service or product. Plus, people will be excited to share their feature with family and friends. Remember, always ask permission to share testimonials and photos.

Offline Lead Generation Ideas



While digital marketing is effective and widespread, don't forget about your customers offline! It might even be easier to stand out in real life than online. These strategies are especially important and effective for brick and mortar businesses that rely on customers actually coming in to purchase. Here are some ways to use physical objects to bring in new leads:

5. Throw a Party

Parties and events are a great way to help customers associate your business with having fun. Gather potential customers at your physical location or offsite and treat them to a good time. You can demonstrate your products and services, or simply build connections. You can also sponsor other events to build your brand recognition.

6. Send Physical Mail

In the age of flooded email inboxes, a physical envelope can really get people's attention.

Consider sending "lumpy mail," which is any kind of mailer that contains something besides a flat letter or postcard. You might send a branded pen or magnet with your contact information, or something more unusual like a balloon or key. People are often curious enough about an oddly shaped package to open it instead of tossing it in the recycling bin.

You can also send different sized mail than a standard envelope. This can also catch people's eye. Be aware that "lumpy" or off-size mail may need additional postage.

Make sure to include a way to contact you or purchase your product. It can be fun to connect the item you send with the act of purchasing or visiting. For example, a car dealership might send a fake car key with the invitation to bring the key when they come in for a test drive.

7. Offer Samples

Another way to generate leads for your business is to offer samples or trials. Costco is famous for its food samples, but you can give away sample products or mini services. Respectfully ask for an email address or phone number before they claim their freebie. You'll have a whole list of people who are interested in hearing from you and have already tried out your product or service.

This strategy also works well with digital offers. Use OptinMonster to create popup campaigns that ask for an email address to send a free eBook sample, limited membership access, or trial software code.

8. Create an Experience

We spend so much of our time online that offline experiences can take on a new meaning. Turn your marketing or sales process into an enjoyable experience for your customer.

Put a photo-friendly prop by your storefront to encourage people to take and share photos of your location. You could even offer a discount or other promotion if they tag you on social media.

You can also create a game or activity related to your business and take it to local events. Collect contact information from participants to send them marketing materials later.

The experience doesn't even have to relate directly to your product or service. You're just trying to generate interest and awareness of your work.

Gifts and Giveaways

Everyone loves getting gifts, and some people will even spend money to get something for free! Here are some ways to use gifts to generate new leads:

9. Send Referral Gifts

In addition to a sincere thank-you note, you can send a gift to people who refer you to new leads. This shows that you appreciate your existing customers. Referral gifts can also motivate new leads to become customers since they see how well you will treat them. Some businesses only send referral gifts for leads that actually purchase, while others do so for all referrals. It's up to you!

10. Offer a Gift With Purchase

Adding a bonus item or upgrade can help people decide to purchase your product or service. Combine this strategy with a time limit to turn leads into customers quickly.

11. Start a Loyalty Program

A great way to encourage repeat business and referrals is to offer a loyalty program. Choose a way to track a customer's purchases or referrals and give them a gift when they reach a certain number.

For example, if you sell consumable items, you could give a free product after the customer has purchased ten.

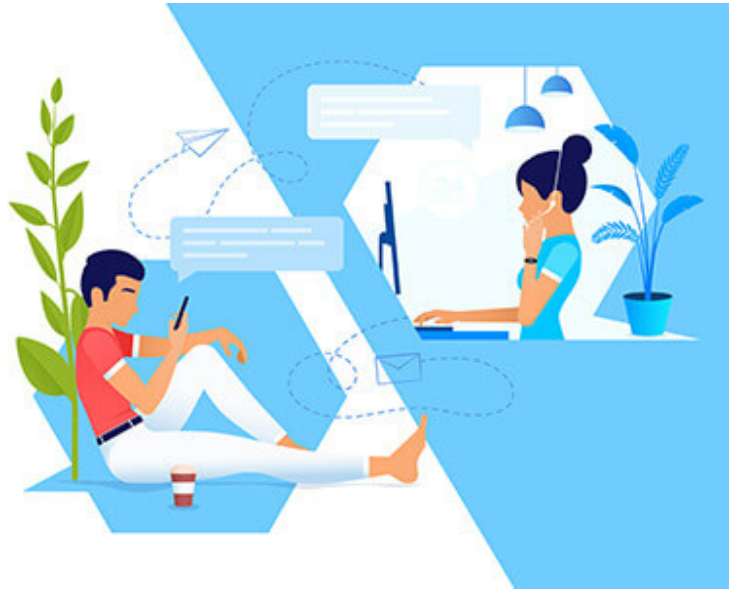
You can also keep track of a customer's referrals and send them a reward after a certain number of referrals have become customers.

Another fun variation on the loyalty program is to offer a discount for bringing a friend in for a service or purchasing two products at once.

12. Host a Contest or Giveaway

People enjoy the feeling of winning. Many are happy to hand over their contact information for the chance to win a prize. You can collect entries on an ongoing basis or for a limited time.

With OptinMonster's gamified coupon wheel campaigns, you can even give every visitor to your website a chance to win a free or discounted product.

Time and Attention

With everyone in a hurry, giving people some extra personalized attention can help you attract leads for your business. You don't necessarily have to spend a lot of time to make someone feel special. Here are some ways to use attention to find new leads:

13. Give Them a Call

Outbound calling can still reach new customers for your business. Keep the call from feeling too "cold" and impersonal by learning about their needs before reaching out. You can collect information through a form or survey on your website. You can also use [email segmentation](#) to track your subscribers' activity and customize your call accordingly.

14. Provide Awesome Customer Service

Make both your existing and potential customers feel cared for and like they are "more than just a number." Take time to answer questions thoughtfully and compassionately.

Hire extra help or invest in tools to help you do this at scale. Happy customers will reward you by purchasing again and telling their friends all about your business.

15. Host a Customer Appreciation Event

Show your existing customers how much they mean to you with a customer appreciation event.

This should be a sales-free space and focus on building positive relationships. Create a memorable experience for your guests and they'll have fond memories of your brand.

You could even invite customers to bring their friends who could then become new leads for your business.

16. Teach a Group Workshop.

You can connect with lots of potential customers by teaching a workshop or class.

Choose a topic related to your product or service and tell participants about your business at the beginning or end. You are teaching them something valuable and building a relationship with new leads.

The workshop participants get a chance to learn something and see that your business can help them.

Generosity

Most people appreciate a good deed done for them or in their name. Here are a few ways that generosity can help you attract new customers who value kindness and helpfulness

17. Participate in Community Service

Give your employees and leadership time to help on community service projects. It's a more personal way to make an impact than financial sponsorship, though donations are helpful too.

Volunteer for causes together, then share authentically about the experience on your website and marketing. Branded clothing can lend your business some brand recognition and credibility.

Find reputable nonprofit or NGO partners and follow their guidance on volunteering to avoid being performative or exploitative.

18. Make Charitable Donations

Many customers want to feel good about their purchase beyond the thrill of receiving the item or service. Consider making a donation to a charity on behalf of your customers.

You can even let them pick from a list. This lets them feel like their purchase has a positive impact and may encourage them to work with you instead of another business.

For example, you could include a field in your checkout form where customers can choose a charity to which you will donate a portion of their purchase.

19. Offer Free Setup or Migration

Some leads might feel nervous or overwhelmed by the idea of starting with a new product or service provider. Help them over the initial hesitation by offering free setup or migration from their existing solution. You can adapt this strategy and offer free webinars or workshops showing how easy it is to get started.

20. Provide a Service Upgrade

Another to get more leads is to provide a free service upgrade in return for giving their email or signing up for a trial.

This strategy costs very little as you will only provide the upgrade to paying customers.

Creative Ways to Generate Leads for Real Estate

Let's take a look at some industry specific examples, starting with real estate. The housing market is highly competitive. Finding new leads for real estate can be difficult even if lots of people are buying and selling. Here are five creative ways to generate leads for realtors from each of the categories above:

- 21. Compliments: Send thank-you card and ask for referral after closing.**
- 22. Offline: Send a physical mailer about new listings or closings in the neighborhood.**
- 23. Gifts: Cover the client's home warranty or other closing costs.**
- 24. Attention: Host open houses to meet new buyer or seller leads.**
- 25. Generosity: Volunteer with Habitat for Humanity to connect with others interested in housing.**

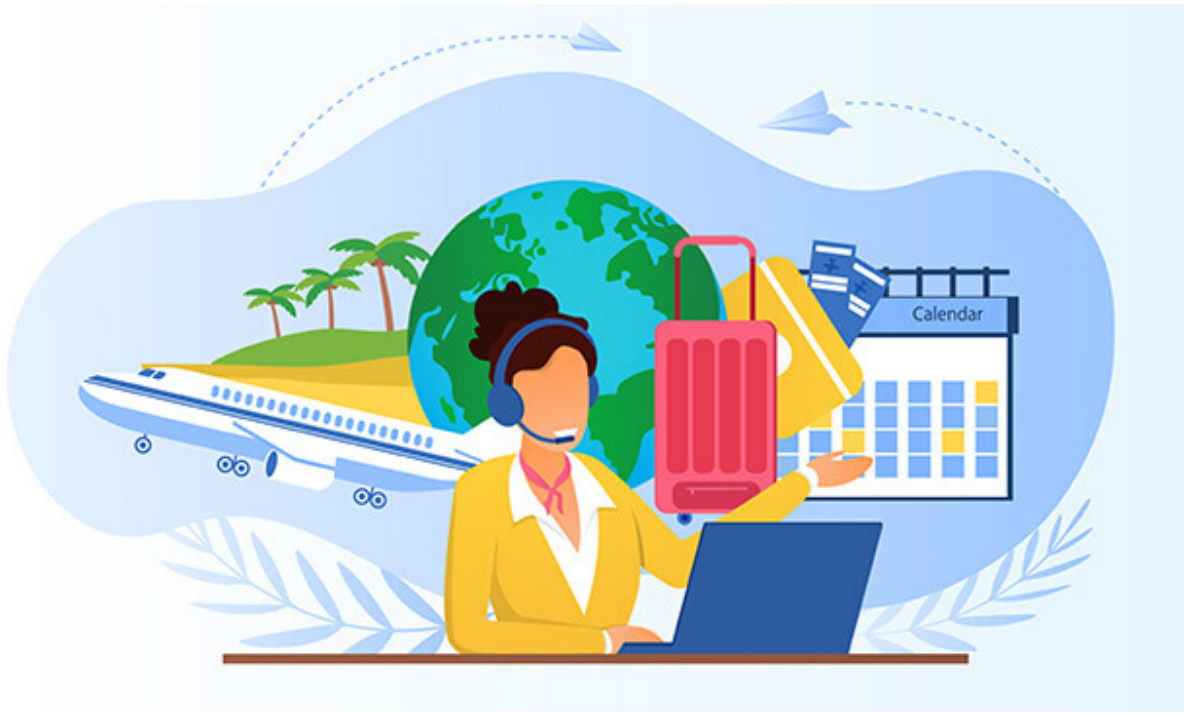
Creative Ways to Generate Leads for Local Businesses

Local businesses remain key parts of many communities. But it's getting harder to compete with large global corporations. Here are five creative ways to generate leads for local brick and mortar businesses.

- 26. Compliments: Feature local celebrities or community leaders as testimonials.**
- 27. Offline: Host an event where people can sample products or services.**
- 28. Gifts: Offer a loyalty program for frequent customers.**
- 29. Attention: Make personal phone calls reaching out to customers who haven't been in for awhile.**

30. Generosity: Offer free delivery within certain distance of your location to encourage the local community to shop with you.

Creative Ways to Generate Leads for Travel



31. Compliments: Post on social media about different destinations and respond to questions and comments personally.

32. Offline: Offer destination-themed refreshments during customer consultations.

33. Gifts: Sponsor a giveaway contest on your website to collect email addresses in exchange for the chance to win a travel voucher or upgrade.

34. Attention: Host customer appreciation mixer to connect customers with similar

35. Generosity: Donate a portion of each booking to a local charity at the customer's destination.

There you go! Here are the new and creative ways to generate leads for any kind of business. You can adapt these strategies across many industries and business models.

What is lead management?

Definition: Lead Management is the process of acquiring and managing leads (potential customers) until the point where they make a purchase. This is a more involved process than standard advertising, and is most applicable to ecommerce stores that generate individual relationships with customers.

Lead management should not be confused with lead nurturing, which is a specific part of lead management that takes place towards the end of the process.

Why efficient lead management is important for online businesses

Giving customers the information they need to continue through the funnel is the primary objective of lead management. When different parts of a business' marketing organization are out of step, or leads are not properly qualified, customers can receive duplicate or non-relevant information - resulting in the death sentence for an otherwise on-track conversion. Simply managing leads in an efficient manner, whether it's a CRM or other B2B lead generation strategy, reduces manual work for an online business and improves the customer experience.

Managing leads in 9 steps

1. Lead Generation

The first part of the lead management process is advertising and actually acquiring leads. Nothing can be done until potential customers have been reached.

2. Customer Inquiry

The management process truly begins when customers respond somehow, signaling that they are interested in something being offered. In most cases, this happens when a customer clicks on a link.

3. Identity Capture

The next part of lead management is understanding who the customer actually is. Some of this information is available through Google Analytics, while other information can be obtained by getting the customer to send it to the company.

4. Inquiry Filtering

Once identities have been captured, they need to be verified for accuracy. This part of lead management helps the company get a better sense for the truth behind any information entered.

5. Lead Grading

After their unique identity is known, leads should be filtered based on their estimated value to the company. Leads more likely to result in a sale — or that offer more value to the company — should be prioritized over casual users.

6. Lead Distribution

Qualifying leads are distributed to the marketing and sales personnel of the company, often with specific instructions and information. In general, customers with the highest potential value should be given to the sales personnel most likely to convert them into a customer.

7. Sales Contact

This is when the sales process truly gets underway. Sales personnel need to structure their contact in a way that encourages a response from the lead, and exactly how that happens should be dictated based on the lead's behavior to this point.

8. Lead Nurturing

Leads who respond to the Sales Contact should be entered into the lead nurturing process, which uses both automated and personal follow-ups to help convince them of the value of making a purchase.

9. Sales Result

Finally, the management process comes to an end when a lead makes a purchase. If repeat sales are desired, return to Step Seven.

What is lead generation in digital marketing? 6 Proven strategies to generate more leads

Today, digital marketing can drive tremendous results in everything and for every business. Gone is the time when all businesses thrived through radio, tv, or print media advertisement tactics.

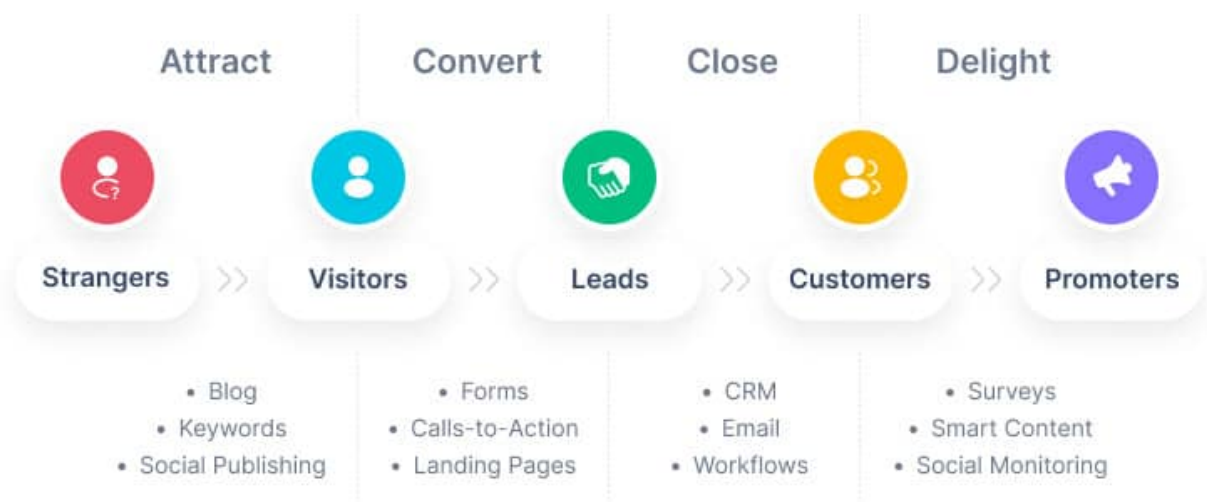
Digital marketing is currently the most preferred approach, enabling companies to reach a vast audience in less time and build lasting connections.

In fact, digital lead generation spending in the US is expected to reach 3.155 billion US dollars by the end of 2022.

Marketers can execute successful campaigns to attract their prospects and generate more targeted leads that are more likely to purchase your product or opt for your services. Digital marketing relies upon real-time data and analytics; hence the leads generated are qualified.

Before we discuss all the strategies to generate leads with digital marketing campaigns, let's understand the term **digital marketing lead generation**.

Digital marketing lead generation



What is lead generation in digital marketing?

Lead generation in digital marketing **involves digital practices that turn the online audience into a paying customer of a business.**

It is simply a process of finding, attracting, and capturing the most suitable prospect for selling a product or service.

Digital marketing lead generation mediums are online channels such as websites, blogs, social media, email, etc.

These digital mediums generate leads via various strategies such as paid ads, email campaigns, gated content, referral coupons, landing pages, etc.

How is lead generation effective with the help of digital marketing?

Digital marketing offers various technologies to capture leads and nurture them while monitoring their engagement until they transform into your customers.

Lead generation digital marketing tactics are efficacious as these techniques rely on customers' interests.

With advanced digital marketing lead generation tools, businesses can track customers' online journeys and analyze their habits to target them with the right content at the right time. Also, paid social media advertising has broadened customer reach.

Several key factors make digital marketing in lead generation effective. Let's quickly go through the insights.

1. It involves content marketing that encourages brand awareness.
2. It helps to understand their buyer persona with content engagement analytics.
3. Paid ads enable the business to be on top of SERPs results.
4. Get more brand awareness with word-of-mouth promotion on social media.

Businesses procure success due to their loyal customers, and digital marketing lead generation is the top way to grow their loyal customer list.

How do you create a digital marketing strategy for lead generation?

Businesses rely on digital marketing techniques to drive traffic to their website and social media pages or to generate leads.

Digital marketing strategies for lead generation



 Salesmate

For creating a solid digital marketing lead generation strategy, you can check out the below image. It explains how you can connect with your leads when they're in the consideration stage and later convert them into customers.

The ultimate mission of any business is to have a continuous flow of qualified leads, and digital marketing makes this possible with its various channels.

Let's find out all the ways through which lead generation is possible by utilizing digital marketing.

Top 6 digital marketing lead generation strategies

There are various ways to generate leads through digital marketing, and below we have shared a comprehensive list of effective lead generation strategies that can be implemented.

There are various ways to generate leads through digital marketing, and below we have shared a comprehensive list of effective lead generation strategies that can be implemented.

1. Content marketing for lead generation

Content marketing is the key element as all organic digital marketing practices require content.

Whether email marketing, SEO, or social media, every medium thrives due to content.

With the help of original, relevant, informative content and customer-targeted content, a business can capture more leads and improve its conversion rate.

With lead generation, content marketing builds trust among your target audience by offering helpful information through different forms of content.

Let's look at strategies helping lead generation with the support of content marketing.

Strategies for lead generation with content marketing

- Use gated content such as free downloadable guides, statistic reports, and eBooks.
- Produce well-researched informative blogs that have a high search volume.
- Write content such as articles, and blogs targeting TOFU (top of the funnel), MOFU (middle of the funnel), and BOFU (bottom of the funnel) audiences.
- Create engaging and relevant content to be in the eyes of your prospects.
- You can build a resource library on LMS about your product/services to help the sales team.
- Write FAQs to resolve common queries of your website visitors.

Remember, the highest quality of content drives high numbers of the audience, so make sure your business has an excellent content marketing team.

2. Search engine optimization for lead generation

Search engine optimization for lead generation



 Salesmate

SEO or search engine optimization enables businesses to generate leads as it allows them to reach their target audience once they search anything on Google, Yahoo, Bing, or any other search engines.

Search engine optimization is an organic digital marketing strategy that helps optimize your website, blogs, and landing pages to rank higher on search engine results pages (SERPs).

For example, if anyone searches for “best CRM for B2B businesses.” And if you have written an optimized article on “10 best CRM for B2B businesses,” then it will appear on top of the SERPs if it’s well executed.

To give you more insights, below are some search engine optimization strategies for lead generation.

Strategies for lead generation with SEO

- Ensure your blog fulfills the intent of the search query
- Work on your blog meta title and meta description
- Add relevant internal links to your article
- Do thorough keyword research
- Build good quality backlinks
- Improve the website speed for a better user experience

SEO enables the business to get the top spot in SERPs and thus enhances the chances of getting more and more leads.

For online businesses, it is now crucial to stay in the top spots on search engine results as out of all, 75% of clicks get by the top; first three google results.

3. Paid advertising for lead generation

Businesses with a high budget can go for paid advertising and start generating leads. It is the easiest, fastest, and result-driven digital marketing lead generation tactic that can be used to test your MVP (Minimum Viable Product).

Paid advertising is a digital marketing technique that uses a bidding strategy to participate and win the real-time auction to be on the top among their competitors on a specific channel like SERPs and social media.

With paid advertising, you can advertise on various channels such as Google SERPs, social media, websites, etc.

The tactic is to get the attention of your target audience towards your business. It offers more value than traditional marketing and is the fastest way to grow.

Strategies for lead generation with paid advertising

Execute a search ad with gated content such as free eBooks, programs, guides, etc.

- You can create ads on social media promoting a sales offer or an event organized by your brand.
- Also, build a remarketing ad highlighting a special limited offer for targeting marketing qualified customer lists.
- You can create hyper-targeted Google ads to target sales-qualified leads.

When it comes to the cost of paid advertisement, companies usually get charged as CPC (cost-per-click) or PPC (pay-per-click), depending upon your ad paying option.

You can show your ads to the most prospective audience for business when they search relating to your product, services, brand, or business.

4. Social media and advertising for lead generation

Comment, like, and share have become the most common words we hear on social media daily. It is one of the powerful channels for lead generation.

47% of marketers mention that they still struggle to manage their social media channels.

Still, it offers tremendous opportunities to businesses to leverage social platforms such as LinkedIn, Facebook, Instagram, etc.

61% of B2B marketers use social media channels, such as LinkedIn, to increase lead generation.

When it comes to lead generation, social media is a great source to identify, capture or engage your likely prospects through posts, videos, polls, reels, etc.

You can set targeted social media ads to enrich customer engagements with more followers, likes, comments, or shares.

Strategies for lead generation with social media and advertising

- Timely post relevant content to enhance brand awareness and boost website traffic.
- Always engage with your audience through comments or messages.
- Leverage social media platforms to stand out with engaging content.
- Provide solutions regarding your customers' pain points.

Social media works best to reach an audience that comes at the top of the funnel for a business. It is a great medium to showcase how you are better than your competitors. Your social media activity will help your customer to identify the brand.

5. Email marketing for lead generation

Did you know that 85% of marketers choose email marketing for lead generation?

With email marketing, you can easily track your email marketing KPIs like email open rates, click-through rate, campaign performance reports, engagement statistics, etc.

Email marketing for lead generation

59% of consumers agree that marketing emails influence their purchase decisions.

Not only this, you can target personalized content that captures more attention which is good for business growth or building great customer relationships.

Businesses mainly use email marketing to generate qualified leads, nurture customers, engage, update, or make a target audience about services or products offered by your company.

Email marketing for lead generation is highly effective due to its high return on investment; according to a report, \$1 spending drives a \$42 business.

To help you more, we have mentioned a few lead generation strategies through email marketing.

Strategies for lead generation with email marketing

- Build a strong customer relationship with marketing qualified leads by offering valuable content.
- Offer value-driven bottom of the funnel content to sales qualified leads.
- You can send catchy offers, updates, and content to your prospecting audience.
- Also, convince your existing customers to promote with a bait of referral discounts/coupons.

Moreover, you must create content relevant to your audience, not the brand's promotional content. If you offer value to your audience, the reciprocity magic will give you promising results.

Execute email marketing campaigns like a pro!

Generate more leads with Salesmate's email automation, personalization & email builder.

Start your free trial

6. Landing page and website optimization for lead generation

Lead generation marketing through a website or landing page is one of the most effective ways to generate high-quality leads.

A responsive business website is not enough to generate ideal leads for your sales pipeline.

Moreover, an optimized website and landing pages help to drive potential prospects and user behaviors.

Optimized landing pages on your website are crucial as these purpose-driven landing pages give most of the leads.

The design and usability of the landing pages should be top-notch as per their subject.

Strategies to optimize your landing page and website

- Use testimonials on your website as social proof to build brand trust.
- Check if your landing pages are easy to navigate.
- Make your website design captivating with the right color contrast that matches your logo.
- Loading time should be fast enough to save from a high bounce rate.
- Check your links to navigate correctly to the desired pages.
- Your web forms and CTA buttons should be in perfect color contrast with your brand. If a potential prospect lands on your website, it must be relevant enough to retain them and convert them as the lead. Your website must have enticing lead generation forms or Live Chat pop-ups to capture your ideal leads.

You can use Salesmate eye catchy **Web Forms** and **Live Chat** pop-up systems to make your website more effective in lead capturing.

Build real-time connections with your leads!

Seamlessly connect with your website leads with Salesmate's Live Chat & Chat Journeys.

Start your free trial**Tools for lead generation in digital marketing**

Lead generation in digital marketing is incomplete without the support of the right tool. Mainly two robust lead generation tools are needed by your sales and marketing team.

- **Email marketing tools**
- **Customer relationship management tools**

These tools enhance business by setting up successful lead generation campaigns with the support of web forms & email builders.

To save more money, it would be best if you found a unified customer management platform that provides extensive support in lead generation, nurturing, segmentation, distribution, and conversions.

Salesmate for lead generation in digital marketing

Salesmate is a unified customer relationship platform helping all sizes of businesses to grow. You can go through the following domains where salesmate is providing great support for your marketing or sales executive.

Once a lead is generated, it is usually passed to the sales team to nurture it and hopefully convert it into a buying customer.

- Send bulk or mass emails to generate leads through salesmate.
- Track email campaign performance, like click-through rates and open rates.
- Perform split testing to draw better results through email campaigns.
- Use personalized, customizable templates for targeting through emails.
- Set automated email campaign.
- You can access social media and your leads' emails with profile enrichment.
- Use web forms or live chat features on your website to get leads through gated content start.
- With mobile CRM, you manage lead generation or any task from anywhere.
- You can sync calendars to set timely follow-ups and automate them as well.

Salesmate takes care of every business need, and thus, it offers integration with various purposeful apps so that you are not required to go anywhere. You can use Salesmate as a lead generation CRM to resolve all your lead generation hassles.

With Salesmate, you can easily generate leads, track their journey, and convert them. Along with your website, you can leverage various lead sources, such as social media, through contact management and profile enrichment features.

Salesmate offers true support to marketers or salespeople with its marketing automation and sales pipeline management feature.

ip Tracker :**Digital Marketing with Web Cookies and IP addresses: Part 2**

In the previous article, we discussed how you can utilize web cookies to aid some of your targeted marketing campaigns especially those that run on Google Ad platforms. It is no doubt that web cookie make audience targeting a lot easier, however, IP address, another type of tracking is also a good variable to understand and utilize. IP address tracking another tool that businesses can utilize rightfully, as these may cause an infringement to laws.

In part two of this series, we'll discuss what IP addresses are, what information you could potentially harvest and the pros and cons of using it to market, let's start!



IP Addresses, what are they?

An IP address is a unique number that gets linked to your physical location. Other than helping the network determine where to send information to, IP addresses can also act as a form of tracking for those who know how to utilize it for digital marketing.

What information could ip address tracking harvest?

Since IP address are linked to physical locations, once you know your visitor's location, you can easily cross reference other data about them to refine your targeting. With IP address tracking, you are able to identify information that is more demographic in nature as compared to the user's purchasing habits with web cookies.

Things like:

- Household Income
- of Children in the home
- Education levels

However, you have to be careful when taking the data into consideration because most households have more than one device and each device can have more than one user. Hence, data harvested might be super generalized or even downright incorrect with ip address tracking due to the nature of devices and household digital items.

If you are in the B2B industry, most office also share the same IP address for all company devices or even route it through a IT hub making the information appear as if the physical location seem incorrect.



Lastly, with the rising trend of using VPNs, IP addresses are getting harder and harder to accurately track, making it a slightly less unreliable tool. Thus, ip address tracking might not be ultimately the best solution to retarget or remarket your audiences.

So what is the best approach? IP addresses or Web cookies? Well, we say both! If you utilize both web cookies and IP addresses simultaneously, you will be able to know which companies are visiting our website and what kind of people are visiting. Targeting your market audience using both can help you to maximize your ROI by marketing to your exact market directly.

If you would like to talk strategy or need help ensuring you are reaching and influencing your target consumers effectively, on the right devices, at the right time, with the right message, contact Adssential Marketing today for a full analysis of your existing digital marketing campaigns!

CPC reduction (in case of paid ads) Group posting on Social Media platforms:**PPC vs. CPC: What Sets Apart PPC from CPC in Digital Marketing?**

It is common for PPC and CPC to be used interchangeably, which leads to confusion among new advertisers. In spite of their similarities, both have their differences that advertisers need to be aware of to be successful. Thus, in the world of digital marketing, knowing what PPC (pay-per-click) and CPC (cost-per-click) are is essential, since these are tools that are indispensable for growing your business. Cost-per-click marketing and pay-per-click marketing are both effective ways to reach millions of people in minutes.



PPC vs. CPC:
What Sets Apart PPC from CPC in Digital Marketing?

Although the reach of this opportunity is vast, and the investment may yield a healthy return (potentially around \$2 or more for every \$1 invested), if you don't know what you're doing, you may lose money quickly. We will explore the differences between PPC vs CPC in this post to help you grow your business and move forward in the right direction.

Let's dive in.



The graphic is a promotional banner for 'PPC Signal'. On the left, the text reads 'Control your Google Ads Cost' in large blue font, followed by 'Make Google Ads smarter, simpler and more productive' in black. Below this is an orange button with the text 'Learn More'. On the right, there is a 'PPC Signal' logo with a signal icon. A central graphic features a blue dollar sign inside a circle, with a numbered list: '1 Get actionable signals daily', '2 Take necessary actions on time', and '3 Keep your campaign cost under control'. At the bottom right, it says 'For single accounts or massive MCCs'.

What is PPC?

PPC, or pay-per-click, is a type of online advertising where businesses can purchase ads that will appear on search engines and other websites. Using pay-per-click marketing to establish a presence on search engines and increase sales conversions is a popular method of online marketing today. Advertisers use this type of advertising model to place ads on third-party websites, social media platforms, and search engines. You only pay for PPC marketing when people click on your ads.

In addition, you can also guarantee that your ads are only shown to people with the same demographic as your average client, thus increasing your sales. With more than \$134 billion generated in ad revenue, Google is the market leader in PPC marketing.

How Does PPC Work?

PPC covers several different types of advertising platforms since it is a global marketing channel. Google Ads is the most common platform. There are several types of Google Ads, including Search Ads, Video Ads, Shopping Ads, and Discovery Ads. In most cases, businesses start with Google Ads, because it is the most effective way to reach the largest audience possible.

There are some common processes irrespective of the platform or the ad format. Whenever a user performs a Google search, an auction determines which ads appear in the SERPs and how much each of the displayed ads will cost. An auction is influenced by a number of variables.

- An important factor in the cost-per-click is the maximum CPC that a business specifies in their Google Ads account. Essentially, it represents the maximum amount per keyword/group of ads that you are willing to pay per click.
- The Quality Score, referring to an ad's predicted clickthrough rate (CTR) and relevance to the keyword, is another key component. Quality Score is based on the end-user experience on the web page that ultimately receives traffic from the advertisement.

What is CPC?

CPC advertising is actually a part of PPC advertising, not something completely different. A cost-per-click, therefore, is a financial metric used to measure how much you pay for every click on your PPC campaign. You can apply the CPC metric to a single keyword. It can also be used to determine your ad's performance.

How Does CPC Work?

In order to maintain and measure the effectiveness and relevance of your pay-per-click campaign as it runs, you will use your cost-per-click number. In any case, you don't want to keep running advertisements that aren't proving successful. You should see a decrease in cost-per-click as you receive more clicks on your ad. That is the ultimate indication of a successful campaign. Your cost per click can vary depending on many factors.

- The rank of an ad is one of the factors that determine the amount you will pay each time it is clicked.
- Along with the ad's rank, factors such as the keyword's competitiveness, search volume, and clickthrough rate are also taken into consideration.
- These factors are applied across the board regardless of what industry you operate in or what type of business you run.
- Every platform, however, has its own set of standards that it adheres to. This means that the costs per click for keywords will vary from platform to platform. In other words, while you may pay \$0.50 for each click on Google, you may pay twice that on Bing. Also, the more desirable a keyword is, the higher its cost-per-click will be.



PPC vs. CPC: What's the Difference?

PPC and CPC are two terms that are often used in online marketing. PPC stands for pay-per-click, while CPC stands for cost-per-click.

So, what's the difference between PPC vs CPC?

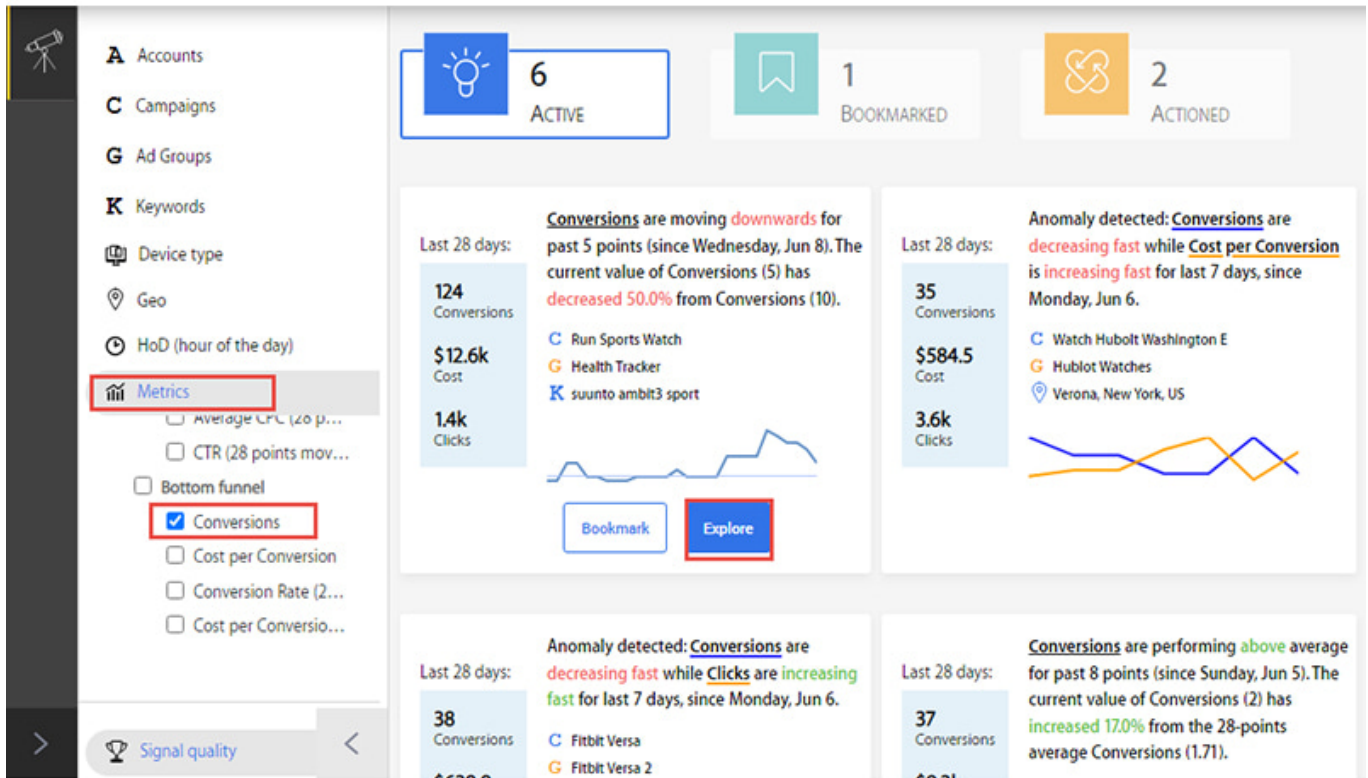
PPC is an advertising model in which advertisers pay a fee each time one of their ads is clicked. CPC, on the other hand, is a pricing model used to calculate the cost of a PPC campaign.

How PPC Signal Can Help in optimization of PPC Campaigns?

The majority of businesses use pay-per-click campaigns, but these campaigns require much effort to run. Because you've spent a lot of money on paid ads, it's crucial that these campaigns are managed and optimized to reach the desired results, especially when you have multiple campaigns running in the same account.

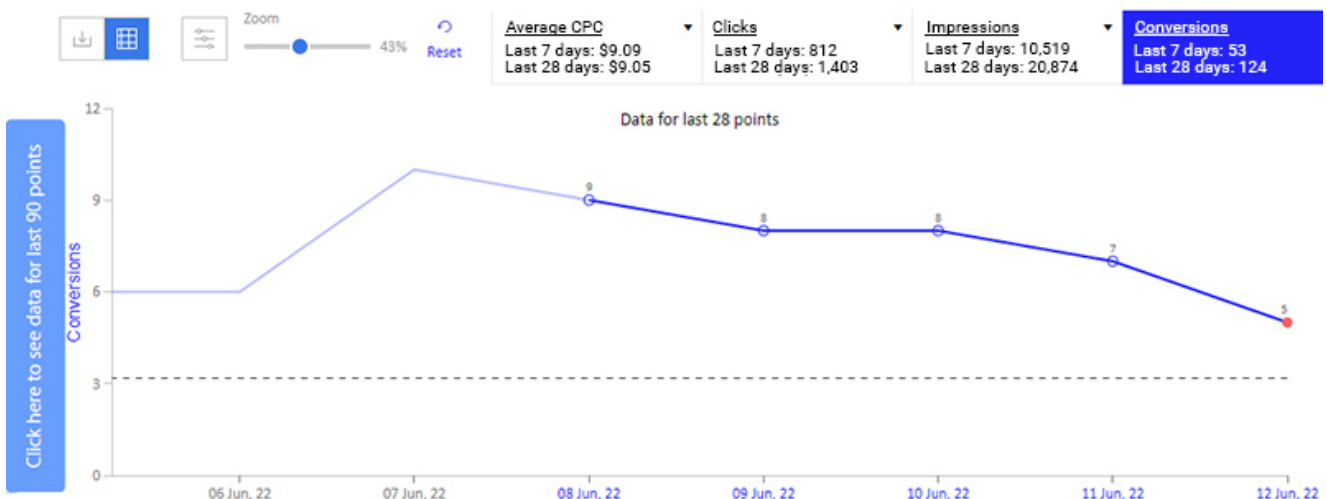
It is here that tools like PPC Signal can help you manage your PPC campaign data and tell you before time what is happening wrong in your campaign so you can correct it before it drains all your budget.

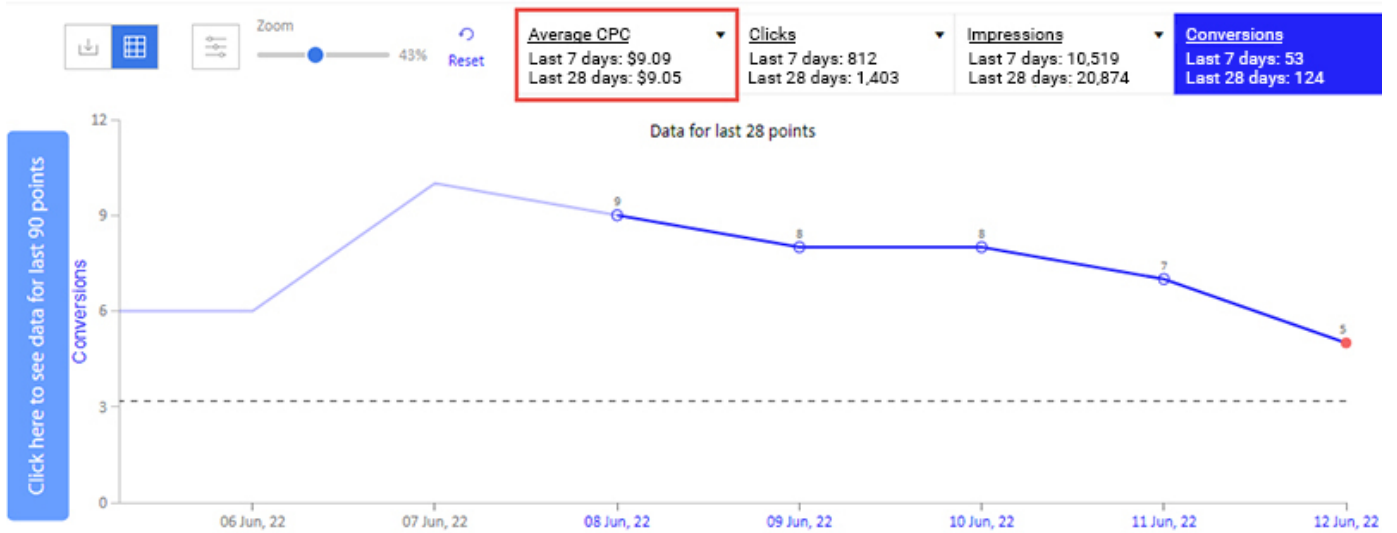
As an example, assume you are running a campaign for your online business and you'd like to optimize the conversions. PPC Signal's dashboard allows you to choose metrics and then conversions. You receive an automated signal that shows how your conversions are performing. The signal can also be explored for more insight.



If you click on the explore button you can access graphical information for your campaign that can help you check in detail how fast your conversions are decreasing.

This signal can also be used to check the CPC of your campaign.





Additionally, you can look at the data in tabular form, which helps you to identify other campaign metrics that may affect your campaign’s conversion rates.

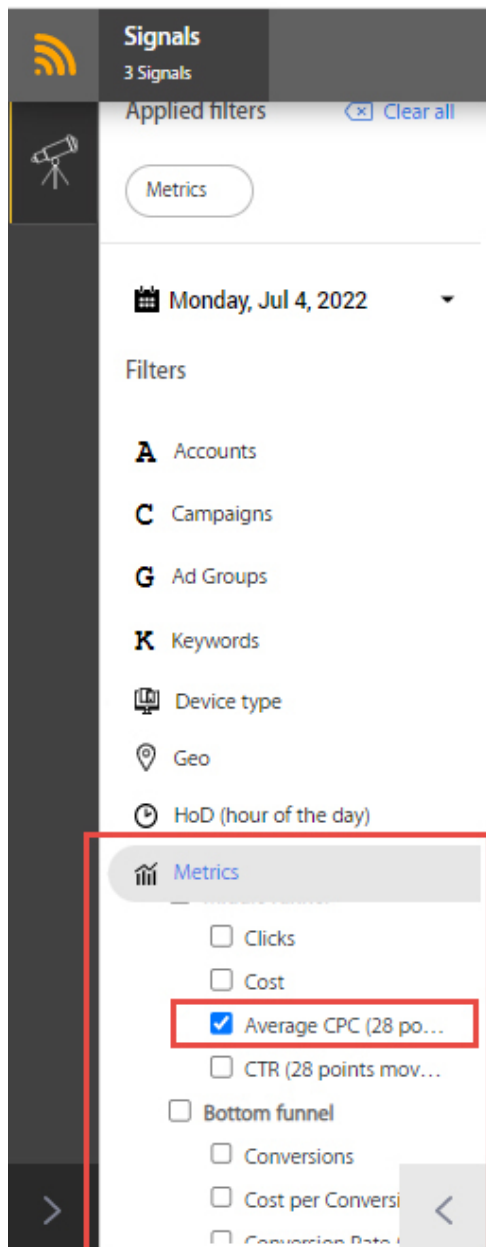


Date ↓	Conversions	Impressions ▼	Clicks ▼	Average CPC ▼
Sunday, Jun 12, 2022	5	1,978	165	\$9.29
Saturday, Jun 11, 2022	7	1,442	129	\$9.30
Friday, Jun 10, 2022	8	1,697	130	\$8.83
Thursday, Jun 9, 2022	8	1,458	120	\$9.08
Wednesday, Jun 8, 2022	9	1,699	117	\$7.44
Tuesday, Jun 7, 2022	10	1,338	91	\$8.85
Monday, Jun 6, 2022	6	907	60	\$10.86
Sunday, Jun 5, 2022	6	696	46	\$10.16
Saturday, Jun 4, 2022	6	1,056	60	\$9.06
Friday, Jun 3, 2022	6	415	26	\$9.83

Signals such as these help you understand how conversions are performing and how other campaign metrics are affecting them. By doing this, you can go one step further and take action, while using less effort and getting better results.

As a result, you will be able to make faster decisions and take action to stop your campaign from killing your budget. The analyzed data gives you actionable insights from your data, so you can take action to increase your sales. This tool allows you to improve your PPC campaign faster, so you can boost your sales.

You can filter your result based on Average CPC as well from left window under metrics section.



What is Digital Marketing- Need of Digital Marketing-Digital Marketing Platforms – Understanding digital marketing process- Difference between Traditional Marketing and digital Marketing- tools of Digital marketing - Advantage of Digital Marketing-Digital Marketing Manager Role and functions - How we use both Digital & Traditional Marketing

What is Digital Marketing?



Digital Marketing refers to the marketing of products and services of a company or business through digital channels such as search engines, websites, email, social media, mobile apps, etc. It involves the use of electronic devices and the internet

Digital marketing mainly comprises Search Engine Optimization (SEO), Social Media Optimization (SMO), and Search Engine Marketing (SEM). We can say that it can be divided into three parts SEO, SMO, and SEM. However, Email Marketing and Affiliate Marketing have also become important components of digital marketing over the past few years. So, in digital marketing, we mainly deal with the following components:

- SEO
- SMO
- SEM
- Email Marketing
- Affiliate Marketing



1) Search Engine Optimization (SEO):

It is a process of improving the structure and content of your site and doing promotional activities to increase the traffic, and thus ranking on search engine result pages.

On-Page SEO:

It refers to all the measures or methods used by website owners within their websites to increase the traffic and ranking of a website on search engine result pages. Within the website means you deal with such elements of SEO that are in your control, such as meta tags, technical tags, content quality, etc. So, there should not be any delay in resolving on page SEO issues to maintain and improve the ranking.

Some of the important On-Page SEO factors are as follows:



1) Meta Tags: Meta Tags are HTML tags that contain meta data and provide information about the content of a webpage. They tell what the page is about when it was updated, and who has created it. This information is very important in terms of SEO as it helps search engine crawlers understand and index the page.

These tags are placed inside the head section of a HTML page, e.g., <head> meta tag</head>. The users who visit your site can't see these tags, but the search engine can see them for indexing and deciding the ranking of your site.

There are mainly three types of meta tags:

- i. **Meta Title:** It is the title tag which is also your page title. It appears on the title bar of the browser window in search engine result pages.
 - ii. **Meta Description Tag:** It is the summary of the information contained in your page. It is displayed below the URL of your page when your URL appears on the search engine result pages in response to the search query made by a user.
 - iii. **Meta Keywords tag:** This meta tag contains all of your key keywords related to the content of your page.
- 2) Page Length:** The search engine prefers long pages to rank higher than short pages. It knows that users do not get satisfied with basic information. Instead, they expect a full explanation
- 3) Outbound Links:** You can give links of other sites on your page that provide similar information. It may act as a trust factor for Google.
- 4) Internal Links:** Interlinks your popular pages to new pages so that traffic from one page may be diverted to other pages.
- 5) Canonical Tag:** This tag is used to prevent the duplicate issues that arise when you have two URLs with similar content. It tells Google that two or more pages with similar content are equivalent to one another and belong to the original page.
- 6) Image Optimization:** Image is required to be optimized using alt text, description, etc. Additionally, instead of naming your image as 'image1.jpg' use descriptive filenames, for example, 'boy-playing-in-the-park.jpg.'
- 7) Sitemap:** A sitemap is created for a site. It helps search engines in indexing pages of your site.
- 8) Content:** The content of your pages should be unique, relevant, and the latest and should be related to highly searched topics, keywords, etc.
- 9) URL Optimization:** Keep your URL less than 255 characters, and use hyphens '-' to separate different parts of the URL. Additionally, it should be short, descriptive, and contain your main keywords. For example, www.javatpoint.com/smo-tutorial-for-beginners. Also, optimize the structure of URL by making categories that help search engines and users to find the content with ease. For example, [Homepage](#)>[Social Media](#)>[Facebook](#)>[Post](#)
- 10) Mobile Friendliness:** Around 60% searches in Google are made through mobile phones and other such devices. So, make sure your website is mobile-friendly.

Off-page SEO:

Although Off-page optimization has the same objective of increasing traffic, it is different from On-page optimization. In On-Page SEO, we deal with the factors that are in our control,

i.e., within the website, but in Off-page SEO, the measures are taken outside the site, which is not in the control of a website owner such as blog submission, article submission, forum posting, etc.

Some of the important Off-Page SEO factors are as follows:



Off-Page SEO techniques mainly deal with increasing the links to a site which is called Link popularity. These links can be internal or external. The internal links come from your own webpages, and external links come from other websites or webpages. High link popularity indicates that you have more connections to your site, which is a plus for SEO. Some important off page SEO techniques to increase link popularity are as follows:

1) Influencer Outreach: If your content is unique, relevant, and the latest, you should share it with influencers in your industry.

2) Guest Posting: There are many authors or blogs that allow you to submit your post or content as a guest post on their sites. If you have written quality content, you can post it there to get backlinks from them.

3) Social Bookmark Submission: There are many social bookmark submission sites where you can upload your webpage or blog post containing a link to your site to drive traffic to your site.

4) Forum Submission: In this method, you participate in forums related to your business, websites. Here, you can reply to threads, answer questions and queries, and provide feedback and suggestions. For better results use, "Do-Follow" forums.

5) Directory Submission: Here, you can submit your pages in directories to build backlinks. You should choose relevant directories and categories.

6) Article Submission: There are also many article submission sites where you can submit articles, again choose relevant categories to submit articles.

7) Video Submission: You can create videos with proper title, description, tags, and reference links and submit them to video submission sites to get backlinks.

8) Image Submission: You can share your images in various image submission sites. But, don't forget to optimize your images with the relevant title tag, URL, alt tag, description, etc

9) Infographics submission: Infographic is a visual representation of information or data such as graphs, charts, etc. You can submit it to infographic submission sites with links to your website.

10) Web2.0 Submission: This off page SEO technique allows you to create a subdomain in high domain authority websites, such as blogger, wordpress, medium, and more.

2) Social Media Optimization (SMO):

In Social Media Optimization, we increase traffic, and thus ranking of sites through social media sites such as Facebook, Twitter, LinkedIn, and Google+. These sites offer an online platform to interact with other people and build a social network throughout the world. Each of the social media sites has different features and offers you lots of ways to drive traffic to your website and promote your business and services. Let us know about some of the popular social media sites:



i) Facebook:

Facebook is an online social media platform that offers you an online platform to invite and connect with other people, which can be your family members, friends, colleagues, etc. This social media platform was created by Mark Zuckerberg in 2004. Initially, it was open to the students of some colleges, but after a few years, anyone who was more than 13 years old was allowed to join it using an email address. Facebook kept developing over the years to offer users new features to interact and share.

Today, it is not only a platform to interact with others, but has become a powerful marketing tool.

What is Facebook Marketing: Facebook marketing is a new form of marketing that allows you to promote your business, product, services, etc., on Facebook.

Facebook offers you plenty of ways to promote your business. Some of the commonly used Facebook features for marketing are as follows:

Facebook Business Page: How to create a Facebook Business Page [Click Here](#)

Facebook Group: How to create a Facebook group [Click Here](#)

Facebook Group: How to join a Facebook group [Click Here](#)

ii) Twitter:

Twitter is another online media platform. It is a micro-blogging tool that allows users to read, write, and share messages that are up to 140 characters long. These short messages are called tweets. Twitter was created by Jack Dorsey in 2006. Today, it has become a popular social media site with a huge user base.

What is Twitter marketing: Twitter marketing refers to using twitter to promote or advertise your business, product, services, and drive traffic to your website.

Some Twitter features that can be used for marketing are as follows:

Images and Videos: You can add images or videos to your tweets to drive more traffic to your tweet. You can also tag users to improve the exposure of your tweets.

Hashtags: You can use it to highlight specific keywords or phrases in your tweets. Up to 3 hashtags can be used per tweet.

Twitter Chat: It allows you to host your chat or participate in an existing chat.

Twitter Alert: It allows you to get notifications when someone tweets. It helps you understand your users and make them interested in your business.

Twitter Analytics: You can use it to check how your tweets are performing, who is retweeting, liking your tweets, etc. You can track your followers' activity over time, including their demographics and interests.

Twitter Moments: This feature was introduced in 2015. It is a list of curated stories or big events happening around the world. It allows you to create your own story to attract visitors to your site.

iii) LinkedIn Marketing:

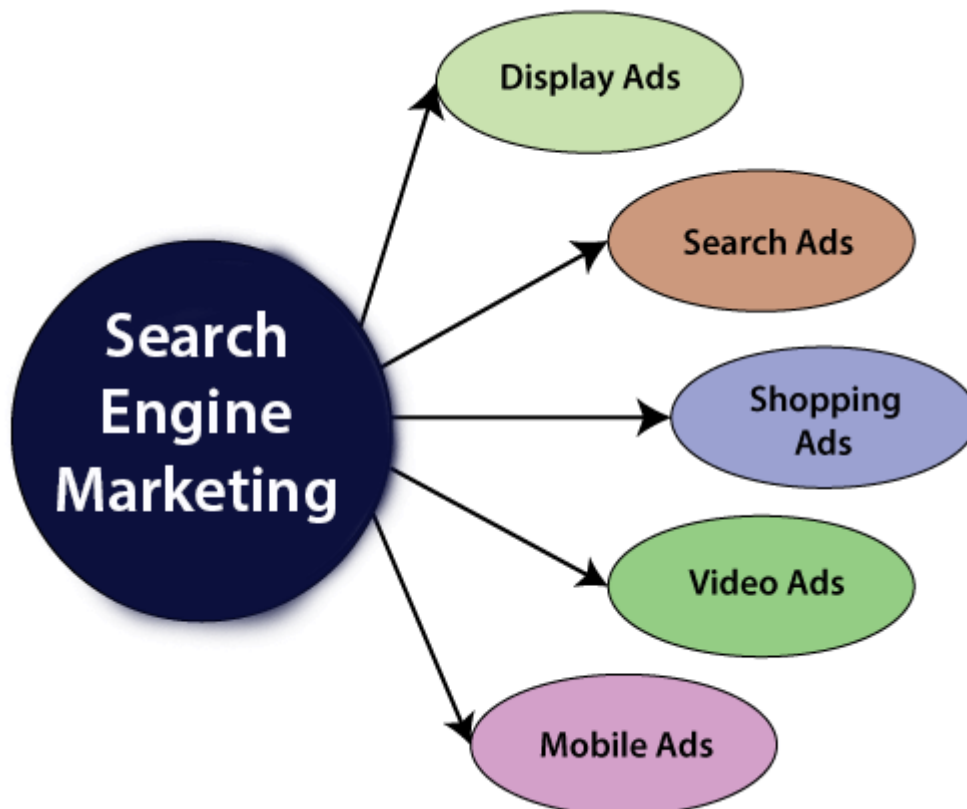
LinkedIn is a professional networking site that allows you to interact with professionals and build a professional network. You can share industry news, your profession or job-related ideas, and requirements like job openings. Today, it is widely used to promote businesses, products, brands, which is called LinkedIn marketing.

Some LinkedIn features that can be used for marketing are as follows:

iv) Pinterest:

Pinterest is a social media network that acts as a virtual online pinboard. It allows you to create your own virtual online boards, where you can pin images, videos, and share them with other users. The users interact through visuals, they can't share ideas, thoughts, etc., without using an image or video. When you post visuals like images or videos to your own or other boards (a collection of pins), it is called pinning on Pinterest.

3) Search Engine Marketing (SEM):



SEM is a digital marketing strategy that is different from SEO and SMO, as here you have to pay the search engines like Google for the marketing of your products and services on Search Engine Result Pages. The more you pay, the more are the chances of your ads to appear on the top of the search engine listings. So, it is the practice of marketing a business through paid advertisements such as Pay per Click ads (PPC) that appear on search engine result pages above the organic results.

In SEO, you don't pay Google for traffic and clicks; rather, you get a free slot in the search results based on the quality and relevancy of your content for a given keyword search.

SEM is also known by many names based on the Search Engine you are using for marketing. For example, Google ads.

What is Google Ads:

Google Ads, which was earlier known as Google Adwords, is a paid search engine marketing platform of Google. It is also known as PPC advertising or pay-per-click advertising. It offers PPC advertising, including banner, text, and rich media ads.

It allows advertisers to bid on certain keywords for their clickable ads to appear in Google's search results. Advertisers pay for these clicks, and this is how Google makes money. It is called PPC (Pay per Click) as you only pay once when someone clicks on your ad.

Google Ads comprises two networks Google Search Network and Google Display Network.

Google Search Network: In this network, the ads appear on the search engine result pages. When users make queries using keywords related to product or service, your ad is displayed on the search engine result pages above the organic results.

Display Network: In the display network, your ads are displayed on websites that have given space to Google to run advertisements. Here, the users don't search for your products or services. However, when they visit a website, they see a banner ad. If they like it, they click it and thus diverted to your site. For example, Text Ads on websites, Image Ads on websites, Video Ads on websites, and Ads on mobile websites.

Points to remember before creating a PPC campaign are as follows:

1. **Determine your goals:** Decide whether you want to increase sales, drive traffic to your site, want to increase subscribers or inquiries, etc.
2. **Set a budget:** Decide how much you can invest in starting a campaign. You can run a small campaign with a small investment to test the market and its functioning. Furthermore, you can set limits on how much you will pay per click and how much you will pay per day to keep your cost in line.
3. **Make a keyword list:** Choose the right keywords that are relevant to your target market, and within your budget, for that, you can use Google AdWords Keyword Planner.
4. **Bid on keywords:** This step is like an auction where you bid on chosen keywords. If you have a sufficient budget, you can bid more than your competitors for a keyword. If you do so, your ad will appear higher on search engine result pages than your competitors for that keyword. The more you bid, the higher your ad appears on Google Search Engine.
5. **Set up different keyword campaigns:** If you have different products and services, you may create separate campaigns to make more profits.
6. **Use keyword-optimized headlines and ad copy:** Incorporate keywords in your headlines and body copy. Users use keywords as search terms while searching for products and services that you sell.

7. **Create unique landing pages:** A landing page is a page that a customer lands on after clicking on your PPC ad. Avoid sending users to your homepage, instead send them to unique landing pages that contain the relevant information and keywords as searched by the users. **How to set up Google AdWords Campaign / How to run Google advertisement or PPC advertising campaign:**

Create a Google Account: A Google account is needed to use AdWords. If you already have one, you can use it.

Go to Adwords.Google.com and sign in: Follow the steps one by one and provide details to complete the form to set up your campaign. Some of the basic steps are as follows:

- Set your budget based on the daily rate or expenses that you can bear.
- Choose your target audience by defining the location, network options (Google Search or Display) and keywords
- Select the bid amount, or Google sets the bid amount automatically in a way that you get the most clicks for your budget.
- Write the text ad, here you are supposed to insert the URL of the landing page, two headlines, and a short description, and option to enable phone calls from the ad. After completing the form, click "Save and Continue."
- Add your payment info by providing the payment info on the billing page. After this, click "Finish and create ad" to complete the setup. Now, your Google ad campaign is ready to go.

Some Benefits of PPC:

It immediately brings people to your site who are searching for your products and services on search engines. Thus, it gives you immediate results. For example, during festival seasons, if you want to sell the product that you particularly introduced for this season, you have to use PPC else you may have to suffer loss. Thus, running a paid google advertisement (PPC) is a great option for businesses who want new leads fast or looking for immediate response or profit.

4) Email Marketing:



Email marketing is also one of the most profitable means of marketing, like SMO, PPC, etc. It refers to sending a commercial message to a group of people, usually potential customers using mail. In simple words, it is the use of emails for promoting products or services as well as developing and maintaining relationships with the clients.

Common Goals of an Email Marketing:

- To increase new signups for your product and services
- To generate new leads for the sales department
- To reach to maximum attendees for your event, to attend the event
- To get donations for your cause

Benefits of Email Marketing:

- It helps increase Brand Awareness.
- It is a cost-efficient way of digital marketing.
- It allows you to create highly personalized messages or ads based on previous sales and purchases.
- It also provides you metrics to evaluate the performance of your Email Campaign.
- It has a larger reach.

Some Email Marketing Tools:

- Mail Chimp
- Litmus
- Reach Mail
- Active Campaign

- Campaign Monitor
- Get Response
- Infusion Soft

5) Affiliate Marketing:

Affiliate marketing is a type of digital marketing in which an affiliate earns a commission for marketing the products or services of a company or a seller, etc. The affiliate gets a part of the profit from each sale. So, the company compensates the third-party publishers, the affiliates, to generate traffic to company products and services.

It is a relationship between three parties: Advertiser, Publisher, and Consumer. Affiliate marketers join affiliate programs that relate to their websites or blogs and have reputable promotes to promote. The affiliate shares these products with their audiences and earns a commission when a product is purchased.

The basic step involved in an Affiliate Marketing / How Affiliate Marketing Works:



1. The affiliate joins the merchant's program and gets a unique ID and a specific URL for promoting the company's product.
2. The affiliate incorporates the link in their blog or post and via their marketing efforts attracts readers to visit their blog and click on the link if they link the product.
3. When a potential buyer clicks the link and lands on the affiliate's site, a cookie that identifies the affiliate is placed on his or her computer. The cookie makes sure that the publisher gets its commission even if the sale occurs after a few days or weeks. The affiliate gets the commission at the end of each payment period.

4. When the sale process is completed, the merchants check the sales process, and if they find a cookie with an affiliate ID, they credit the affiliate with the sale.
5. The merchant shares the reports with affiliates so that they can see their contribution to the sales and accordingly can get their share.

Need of Digital marketing

- Digital marketing plays a valuable role in shaping consumer behavior in today's world, but how does it positively impact businesses? And more importantly, why? Here are a few reasons why every business should prioritize digital marketing over traditional methods of advertising.
- Digital marketing levels the online playing field. When it comes to a business's visibility on the web, just like opening a store, location is everything. Being easily noticeable on the web, getting a business's name out there, and updating information frequently will bring customers to its door.
- Digital marketing helps businesses stay a step ahead of their competition. Digital marketing is the best way for a brand to get a leg up on its competition. Think SEO, organic search, local search, Google Adwords, social media, and blogs. Businesses want to reach as many people as possible, and this is significantly easier to achieve on the web than it is in person.
- Digital marketing is less expensive than traditional advertising. Traditional advertising can cost large sums of money, whether it be via television, radio, newspaper, magazine, or direct mail. Now business owners can find a cheap equivalent online. Think Youtube instead of television, blogs instead of magazines, social media instead of flyers, and podcasts instead of the radio. Some of the digital equivalents to traditional advertising are free, and all can be much cheaper than their traditional counterparts if the business has someone to manage and develop its strategies.
- Digital marketing delivers analytics. The good news about digital marketing is that an ad's creators can find out how it is pulling through using analytics that can't be executed with traditional methods of advertising. Analytic reports can quickly be pulled up to test ad campaigns and find out what is getting read, looked at, or bought.
- Digital marketing reaches mobile users. Digital marketing can be formatted to mobile devices in order to reach customers no matter where they may be. Once a business's local search and digital presence have been optimized, it can rest assured that it will be found. Having a digital presence and local search optimization is vital as people increasingly rely on their phones' web browsing capabilities.
- Digital marketing builds brand recognition. Since there are so many channels on the web with marketing potential, getting the message out about new brands is easier than ever before. Shotgun marketing approaches will confuse potential customers. It is much better to invest in a succinct campaign and build onto it from there, maintaining a distinct voice and brand style.
- Digital marketing allows businesses to monitor their brand. The great thing about digital marketing is that those who use it can easily check their reputation and engage with unsatisfied customers, making it possible for them to address negative press before it circulates too far and too wide. Just because a business doesn't take advantage of the web and social media doesn't mean that its customers don't. It is far better to know what is going on as it happens than to find out the hard way along with everyone else.
- Digital marketing can help brands develop trust with their audience. If a business is following its customers on social media and reaching out, commenting, and showing them that they care, it will gain trust, camaraderie, and friendships, in addition to more customers. There is nothing greater than engaging with a customer online and seeing them come into the store later to say how much it meant to them that the business took the time to connect. It is the small things a business can do that often make the most difference.

- Digital marketing provides businesses with additional sales channels. Think of digital marketing as branches off of a brick and mortar business. One branch of the business contains social media posts with tips, advice, local news, and events involving the brand. Another branch hosts the business's blog, which serves to educate, inspire, and entertain its audience. Branching off from these main channels are landing pages and blog posts for different products. Add in Youtube videos, and the business can be everywhere on the web, delivering its message and brand, and promoting its products and services.
- Digital marketing can educate customers. One of the best features of digital marketing is its ability to help brands educate their customers and share information that makes their lives better. One of the greatest reasons to use digital marketing is the potential that it has to improve others' lives. By sharing expertise, business models, and positive lifestyles, businesses truly can make the world a better place for the next generation and beyond.
- Digital marketing brings a brand's story to more people. Lastly, digital marketing is about telling a unique story to the world. Digital marketing is a business's best chance to stand up and be heard, bringing its brand to the doorsteps of people that need the product or service, and those that might like to learn more about it.

Digital Marketing Platforms:

What is Digital Marketing?

Digital marketing is any marketing initiative that leverages online media and the internet through connected devices such as mobile phones, home computers, or the Internet of Things (IoT). Common digital marketing initiatives center around distributing a brand message through search engines, social media, applications, email, and websites.

Today, digital marketing often focuses on reaching a customer with increasingly conversion-oriented messages across multiple channels as they move down the sales funnel. Ideally, marketing teams will be able to track the role each of these messages and/or channels plays in reaching their ultimate goal of gaining a customer.

10 Types of Digital Marketing Platforms

Ten common types of digital marketing platforms include:

- Social Media
- Influencer Marketing
- Content Marketing
- Email Marketing
- Search Engine Optimization (SEO)
- Pay-per-click (PPC)
- Affiliate Marketing
- Mobile marketing
- Marketing Automation Platforms
- Marketing Analytics Platforms

Digital Marketing Process.

There are various situations in which Digital Marketing Process would be required. They could be:

1 – You are an Entrepreneur & want to start marketing your business on Digital Media.

2 – You are a Digital Marketing Agency or a Freelancer & want to Market your client's business on Digital Media.

3- You are a Digital Marketing Executive working for an Agency, & you are told to create a Digital Marketing campaign for the client that your agency has just acquired.

Sit back & relax. With Digital Marketing Process given below, your job is going to be simple. Understanding Digital Marketing Process & fitting your project in that will lead to complete control of a project.

Let's start to understand Digital Marketing Process.

Following is a 5 Step Digital Marketing Process, that can be used for marketing anything on Digital Media.

Step 1: Research

Step 2: Create

Step 3: Promote

Step 4: Analyze

Step 5: Optimize

Detailed Step By Step

Digital Marketing Process

Step 1: Research

At this stage, you will collect all the information that will be required for decision making in the next stages. Information collected during the research will become your raw material to strategize & create your digital marketing campaign. This stage can also be called as *Digital Marketing Research*. At this stage, you will research 4 sets of information:

1. About Business
2. About Your Target Customers
3. About The Product That You Want To Market
4. About Online Competition

Each set is unique & equally important. You will require multiple sources to collect the information.

Step 2: Create

Once you collect information at the research stage, you can now start creating:

1 – Digital Marketing Objectives / Goals: These are the ultimate goals that you want to achieve through your Digital Marketing Campaign. Every business is unique, therefore their goals will also be unique. Campaigns without clear goals will end up spending money without the assurance of achieving goals. What goals you should set, can be answered after looking at information collected at the Digital Marketing Research stage. *Learn more about [Digital Marketing Objectives / Goals](#).*

2 – Digital Marketing Strategy: After you set the goals, it's time to create a strategy to achieve those goals. Your Digital Marketing Strategy will include Positioning Strategy, Branding Strategy, Content Strategy, Digital Marketing Channels Strategy. What strategy should be adapted/created, will be answered from the information collected at the Digital Marketing Research stage.

3 – Digital Marketing Plan: At this stage, you will lay down a documented plan that will include all your detailed Digital Marketing activities with timelines.

4 – Creating Primary Digital Identities: The 3 primary Digital Identities of business are Website, Blog & App. These are like your online office, shops, or showrooms. These are the places where you want your target customer to reach & ultimately buy your products & services. Before you move on to the next stage, i.e promote, your primary digital identities must be fully ready. For businesses that want to sell their products through major ECommerce portals, creating digital identities can be optional. But, it's better to at least have a website for establishing some credibility of your business.

Step 3: Promote

After your primary digital identities are fully ready, you will start promoting them. That means you want relevant people to start coming to your primary digital identities. This is also called as generating relevant traffic. Relevant traffic is an important word here.

The more you get relevant traffic to your website, the more the conversion you can expect. Your options to promote your website/blog / app will be:

1. Search Engines
2. Display Network
3. Ecommerce Portals
4. Social Media
5. Email
6. Messaging
7. Affiliate

The above are also known as Digital Marketing Channels, which you need to promote your Primary Digital Identities (Website / Blog / App). There are sub-channels & networks within some of the channels mentioned above.

Which channels, subchannels, networks to go for & whether to do organic or inorganic promotions, these questions will already be answered at the Digital Marketing Strategy creation stage.

Step 4: Analyze

Once you create your primary digital identities & start promoting them through various digital marketing channels, it's time to start monitoring your performance. Analyzing is like looking at the outcome of your digital marketing work. You will receive analytics for your primary digital identities, as well as the channels through which you have done the promotions.

The most important & ultimate analytics for any business is the analytics of your website/blog / app. Google Analytics is widely popular to generate analytics of your primary digital identities. The 4 major sections of Google Analytics are:

1. Audiences
2. Acquisition
3. Behavior
4. Conversion

Difference between Traditional Marketing and digital Marketing:

Traditional Marketing	Digital Marketing
The promotion of products and services through TV, Telephone, Banner, Broadcast, Door to Door, Sponsorship, etc.	The promotion of products and services through digital media or electronic mediums like SEO, sem, PPC, etc.
Traditional Marketing is not cost-effective.	Digital Marketing is more cost-effective-promoting.
It is not so good for Brand building.	It is efficient and fast for brand building.
Traditional Marketing is difficult to Measure.	Digital Marketing is easy to Measure with the help of analytics tools.
It is difficult to quantify the return on investment in traditional marketing.	It is simple to calculate in the case of digital marketing.
After the posting of the advertisement, it cannot be altered.	Even after the posting of an advertisement, it can be amended.

<p>traditional Marketing includes.</p> <ul style="list-style-type: none"> • T.V. advertisement • Radio. • Banner Ads. • Broadcast. • Sponsorship. • print Ads. 	<p>Digital Marketing includes..</p> <ul style="list-style-type: none"> • Search engine optimization (SEO) • Pay-per-click advertising (PPC) • Web design. • Content marketing. • Social media marketing. • Email marketing.
Users have no option except to watch the ads.	Users can even skip the ads if they lack interest.
The traditional type of marketing has local reach.	The digital type of marketing has carried a global reach.
There are standardized ways of targeting users.	The targeting here is customized and relies on the type of user.
The methods opt in traditional marketing for market analysis by a company leads to waiting for weeks or months to get results.	Digital marketing gives quick results and thus helps in getting real-time marketing results easily.
No real-time results are obtained in traditional marketing so there is a need to draft a marketing strategy beforehand as it relies on marketing results.	The improvement in marketing strategy is quite flexible as it can be changed according to marketing results.
One-way communication occurs in traditional marketing because of its rigid means to carry out the process of marketing.	Two-way communication occurs that leads to more customer satisfaction.

Tools of Digital marketing:

Digital marketing tools by strategy

We've put together a comprehensive list of digital marketing tools that can help you regardless of your goals, including a few free and freemium tools for those on a tight budget.

Whether it's managing customer relationships, winning over leads or uncovering new marketing opportunities, this list can serve as the foundation of a powerful marketing stack.

Check out the categories below to get started:

- [Social media marketing tools](#)
- [Email marketing tools](#)
- [SEO \(search engine optimization\) tools](#)
- [Conversion optimization tools](#)
- [Lead enrichment tools](#)
- [Landing page and lead capture tools](#)
- [Graphic creation tools](#)

Social media marketing tools

We've seen firsthand how social media has evolved into a priority marketing channel for businesses today.

A prime place to nurture leads and build business relationships, social is ideal for gathering valuable data when it comes to what your customers want.

Managing all of the moving pieces of social media "by hand" is a recipe for burnout. Instead, consider how dedicated software allows you to publish more meaningful content and use social media to meet your big-picture business goals.

1. Sprout Social

Hey, you can't blame us for putting ourselves on the top of the list!

That's because Sprout Social is the ultimate control center for any business looking to translate its social presence into actual results.

As a **social media management** platform, Sprout helps businesses organize their content calendar and assets in one place. This allows you to publish and schedule your content across multiple platforms, all timed to perfection based on when your followers are most active.

Beyond social media management features, Sprout makes it easy to collaborate with colleagues and customers alike. Our Smart Inbox gives you a bird's-eye view of all of your social messages so you can respond thoughtfully and in a timely manner.

Oh, and don't forget about Sprout's full suite of social analytics.

Optimizing your content's performance doesn't have to be a guessing game as Sprout identifies your top-performing posts and the success of your social

campaigns. Coupled with robust social listening features, you can uncover trends, hashtags and opportunities to engage with new customers.

2. Loomly

Loomly's self-described "brand success platform" is a tool that's ideal for smaller social teams looking to organize and collaborate on content.

Built-in calendars, deadlines and workflows make both scheduling and brainstorming content a breeze. As an added bonus, Loomly actually curates fresh content ideas for users based on trending topics and Twitter conversations.

The clean, no-frills interface is easy to navigate and friendly to users who might not be the most tech-savvy. The affordability of the platform is enticing for solo businesses and smaller agencies looking to wrangle their social presence.

3. Audiense

We've talked time and time again about the importance of social listening for identifying trends and potential customers.

Tools like Audiense take listening to the next level by helping companies both identify and segment their social media audiences. Doing so makes it easier to run laser-targeted ad campaigns and likewise dive deeper into your customer personas. Digging into demographics, personality straight and beyond, you'd be surprised at what you can learn from social alone.

Audiense's platform focuses primarily on Twitter, making it great for B2B brands interested in prospecting *and* B2C companies looking to understand more about their target audience.

Email marketing tools

Email represents arguably the most tried-and-tested, scalable marketing channel available to modern companies. As such, email solutions are a staple of Internet marketing software for businesses big and small.

From list-building and improving deliverability to coming up with awesome offer campaigns, let's look at some tools that can break down your customer data and put a good chunk of your marketing efforts on autopilot.

4. HubSpot

HubSpot, a company well known for its CRM and inbound marketing software, recently launched an email marketing product. HubSpot Email Marketing is easy to use, boasts impressive deliverability, and of course, comes natively integrated with all other HubSpot products, such as the free-forever **CRM**, as well as hundreds of other popular marketing tools. The tool comes with a Free Plan which includes up to 2,000 email sends/month, contact lists, a drag-and-drop email builder, and **ready-made templates** so you can get started right away.

The best part? HubSpot reduces complexity with integrations and brings tools together. One of the integrations available is Sprout Social. As a part of Sprout's care features, you can create and delegate tasks for your customer care team. By connecting Sprout and HubSpot, your team can create, track, manage and resolve issues without leaving the app.

5. SendGrid

SendGrid offers a full suite of email marketing services, many of which are totally friendly to novices and email veterans alike.

For example, the platform offers flexible design options via visual, drag-and-drop editing, coding or a combination of both. In-depth deliverability and performance analytics are also built into SendGrid, cluing marketers in on which messages are performing and what needs work.

We know: there are more email marketing solutions out there than we can count. That said, a big upside of SendGrid is the forever-free plan for up-and-coming businesses as well as scalable pricing that matches your needs as you grow your list.

6. lemlist

lemlist is unique among our email marketing tools because the platform is primarily focused on deliverability. Emphasizing the best times to send your messages (and how often) for the sake of more opens and clicks, lemlist is awesome for optimizing your existing campaigns.

Rather than second-guessing your marketing pushes, lemlist can be an eye-opener in terms of how to warm up your list.

Additional features include personalization tools to make your outreach emails sound less spammy and follow-up email sequences to encourage more replies from cold prospects.

7. Moosend

Moosend is among the most user-friendly and affordable of our digital marketing tools, representing a surprisingly robust email solution for those just beginning to build their list.

Codeless campaigns, simple automations and easy-to-read reporting are all built into the platform. Paid users can also take advantage of landing page features, including mobile popups and countdown timers.

SEO (search engine optimization) tools

Fact: **over half of all website traffic** comes from search queries.

As the digital landscape becomes increasingly competitive, anything companies can do to increase their search presence is a plus. Although SEO might be daunting, especially to up-and-coming businesses, there are plenty of online marketing tools to help you uncover search opportunities and optimize your existing search engine efforts.

8. Ahrefs

Ahrefs is the gold standard when it comes to brainstorming keyword ideas and opportunities to rank.

The platform's site explorer lets you check any URL's top organic keywords, while also estimating how much traffic a competitor receives for any given search term. You can also identify a site's top-performing content and sources of backlinks.

In short, Ahrefs is a fantastic tool not only for competitive analysis but also for making sure that your existing content is up to snuff for search.

9. Clearscope

Now, let's say you're interested in capitalizing on keyword opportunities and optimizing your existing content. That's where Clearscope comes in.

Particularly useful for content writing, the platform provides a detailed editor to recommend keywords, headers and readability to help you write (or rewrite) high-ranking, well-balanced blog posts.

Whether you're creating a content strategy from scratch or refreshing your existing blogs, Clearscope covers every nook and cranny of search optimization.

10. SEMrush

Another staple SEO tool, SEMrush allows you to track the position of your priority keywords and likewise explore new terms to rank for.

The tool's breakdown of keyword ideas, difficulty and variations are great for brainstorming content ideas and identifying search intent, too.

Conversion optimization tools

The smallest changes can make the biggest difference when it comes to getting people to convert on-site.

Just changing the color of your call-to-action button can spell the difference between scoring a free trial sign-up and someone bouncing. Among the digital marketing tools in your toolbox, conversion optimization software can clue you in on low-hanging opportunities to increase your revenue.

11. Unbounce

Unbounce is an amazing tool for quickly building, tweaking and publishing new landing pages to test. Built-in A/B testing and variant analytics can answer directly which creatives, calls-to-action and additional page elements are working (and which aren't).

Even if you're not much of a designer, you can use Unbounce's proven landing page templates as a jumping-off point and then modify them to fit your style. The platform's analytics spell out clearly which landing page variants are your top performers.

12. Optimizely

Emphasizing landing page experiments, Optimizely combines tools for visual creation and audience targeting to quickly run tests on different segments of your audience.

A no-code platform that allows you to test both major and minor edits to your pages, fine-tuning your site for performance doesn't have to be a huge undertaking.

13. Hotjar

Hotjar's platform provides a real-time visual record of your visitors' actions and behaviors on-site.

Through heatmaps that clue you in on where people are (or aren't) clicking. Synced to actual video recordings of your visitor's journey, you can quite literally see what needs to be tweaked at a glance.

Lead enrichment tools

It's well-documented that the majority of website visitors fail to provide enough data after leaving your site.

Thankfully, there are business intelligence tools out there to help you better understand your leads and highlight key information to reach out to them once they've shown interest. Particularly powerful for **B2B marketing**, lead enrichment tools supercharge your prospecting and outbound marketing pushes by giving you a more holistic view of your traffic.

14. Clearbit

Using 100+ sources, including Salesforce and additional marketing platform data, Clearbit creates an up-to-date profile of your leads to make your outreach efforts go more smoothly. The details gathered include company, role and company size, just to name a few.

Rather than dig for details or rely on outdated information, the platform regularly updates itself every 30 days to ensure that your data stays fresh. This allows you to prospect with confidence and save serious time in the process.

15. Datanyze

Similar to Clearbit, Datanyze also uncovers crucial contact information about your on-site leads to fill out your digital rolodex. The platform also works brilliantly for prospecting on LinkedIn by pulling social data on decision-makers, too.

Landing page and lead capture tools

As our attention spans shrink, making a conscious effort to reel in visitors once they land on-site is crucial. Marketing platforms focused on lead capture ensure that your traffic doesn't go to waste and visitors are more likely to take action. When done right, the end result is more leads and conversions.

16. OptiMonk

OptiMonk's platform allows businesses to grab the attention of customers and prevent them from bouncing through personalized pop-ups.

With an emphasis on lead capture and exit intent messages, the platform's behavior-based targeting means that your pop-ups don't have to be disruptive.

For example, OptiMonk encourages users to segment their marketing messages and serve them only when they make sense. From returning buyers to first-time window shoppers, the platform empowers you to create campaigns that speak to all of your customers rather than treat them as one-size-fits-all.

The upside of OptiMonk is that it's known for its ease of use. Boasting a ton of templates with established average click-through rates, brands can customize their messages based on pop-ups that are proven to be effective.

17. Typeform

Typeform is a sleek tool for marketers looking to create attractive, minimalist forms.

Unlike more traditional pop-ups, Typeform's intake forms are seriously stylish and don't feel like ads at all. The platform's simple editor and easy embeds are a nice added bonus, as is the ability to create quizzes and interactive forms.

18. MailMunch

A hybrid email marketing and landing page tool focused on list-building, MailMunch offers several engaging form types and emails to send to leads once they've opted-in.

The platform lets you segment your audience based on factors such as how often they buy and customer demographics. Additionally, their goal-based form-builder is straightforward and lets you work from a variety of templates.

Graphic creation tools

Infographics. Memes. Graphs and graphics. The list goes on and on.

Visuals are the cornerstone of social marketing and branding at-large. If you don't have the budget for a designer or are running a DIY business, digital marketing tools such as **Canva** have become the go-to for producing eye-catching visuals.

That said, there are a few other graphic creation tools you should consider so your creatives don't grow stale.

19. Creatopy (formerly Bannersnack)

Creatopy is a graphic creation tool akin to Canva, but emphasizes quite a few features specific to marketers.

For example, the platform's design sets and brand kits allow you to seamlessly work with fellow marketers and keep your brand creatives organized. This is particularly useful for agencies managing multiple clients or social accounts.

Meanwhile, the ability to edit the same design within multiple formats (think: desktop banner versus mobile) in a single click is a huge time-saver.

20. Visme

Visme's platform is focused primarily on creating presentations and data visualization.

Because infographics and fresh data are among the most-shared types of content on social media, Visme is ideal for anyone frequently publishing research to platforms like Twitter or LinkedIn.

Beyond straight-up graphic creation, the platform lets you pull data from external sources (think: spreadsheets) to make presentation creation a snap.

21. Vennengage

Vennengage is another graphic creator with an emphasis on infographics. With spreadsheet imports and hundreds of chart configurations, you can customize your infographics however you want. Customize any infographics based on your branding with tons of built-in graphics to choose from.

Which digital marketing tools are part of your stack?

With the right tools at your disposal, you can streamline your marketing campaigns and automate a ton of tasks in the process.

Any combination of the tools above can serve as the foundation for a solid digital marketing stack. Don't be shy about test-driving and going through trials to find what works for your business and budget.

Once you've sorted out your digital marketing tools, you can move forward with your campaigns with a sense of confidence and likewise keep a better pulse on your marketing efforts.

And if you haven't already, make sure you check out our variety of **social media templates** to help organize and measure your social marketing campaigns so they align with your business' goals.

Advantages of Digital Marketing:

1. Global Reach

Traditional marketing is restricted by geography and creating an international marketing campaign can be hard, expensive, as well as labor-intensive. However, digital marketing happens on the Internet, which means that the reach you can achieve with it is immense. Even a very small local business owner has the ability to reach an international audience with an online store. This would never be possible with traditional marketing or would cost a whole lot of money to do so. This online accessibility has opened many growth opportunities for businesses to explore. The combination of global reach and visibility is a great opportunity for any business.

2. Local Reach

While global reach is a significant advantage of digital marketing, it also improves local visibility, which is especially important if your business relies on nearby customers. Local SEO and locally targeted ads can be beneficial for companies trying to bring more customers to their doors. Think of the reach you can get to a whole neighborhood with digital marketing versus the reach it would take you to print out flyers and distribute them around.

3. Lower Cost

Whether you want to promote your business locally or internationally, digital marketing provides you with cost-effective solutions. It allows even the smallest companies to compete with larger companies using highly targeted strategies. Most of these strategies won't even cost anything at all to start with (such as SEO, social media, and content marketing). However, not every form of digital marketing is suitable for every business and some may even have more costs than others. A business can find appropriate solutions based on its marketing goals.

4. Easy to Learn

While there are many aspects of digital marketing that you need to learn, it is fairly easy to get started with. It gets more complex from the nature of the goals and the scale of the campaigns. However, it is all a matter of finding the right strategy that works for your business.

Professional Certificate in Digital Marketing

Designed With Meta Blueprint & IMT Ghaziabad [APPLY NOW](#)



5. Effective Targeting

Even if you don't have a clear idea of your target audience, digital marketing enables you to extract data to see which audiences will work best for you and optimize your campaign around them. There are many different options of targeting such as through keywords for search engine optimization (SEO), pay-per-click (PPC), or through demographic information on social media. This enormous amount of targeting elements at your disposal makes sure that every campaign reaches the right audience. It also helps you to analyze the changing behaviors of customers and modify campaigns for those changes. This ability to understand customers' changing needs quickly is a sure way of success for any company.

6. Multiple Strategies

There are different strategies of digital marketing that can be used by different types of businesses. A B2B business that is interested in gaining international leads may have a totally different strategy than a B2C local business selling clothes. While some companies can benefit more easily with content marketing and SEO, others can benefit from conversion-based ad campaigns. The key is to always analyze the results and develop better tactics and methods with time. A well-executed digital marketing strategy is one that changes and adapts quickly as the needs of the business transform.

Here are some of the most common types of digital marketing you can choose from:

- SEO-based content creation
- Search engine marketing
- Social paid ads
- Video marketing
- Forum engagement
- Social media marketing

- Email marketing
- Local search
- Remarketing
- Influencer marketing

7. Multiple Content Types

Another crucial advantage of digital marketing is the different content types available to showcase your brand online. For a lot of platforms, there is a wide range of content types you can choose from to keep your brand fresh and build effective online campaigns. Unlike traditional marketing, you can more easily reproduce one content to fit as many platforms as you want.

Here are some of the most common types of content that you can choose from:

- Blogs
- Podcasts
- Emailers
- Ebooks
- Visual content
- Infographics
- Whitepapers
- Quizzes
- Social media posts
- Webinars

Go From Beginner to Expert in No Time!

Purdue PCP in Digital Marketing [EXPLORE COURSE](#)



8. Increased Engagement

One of the most important advantages of digital marketing is increased engagement. Digital marketing is designed to be highly engaging by default. Users can share a blog post, like a photo, save a video, or engage with your website via a paid ad click. The best part is that all of these actions can be measured. This enables you to create even more engaging posts to increase brand awareness or boost sales. The more you engage online, the more loyal customers you can get. Businesses that use engaging formats effectively in their online strategies have an easier time converting cold traffic to loyal customers.

9. Analytics and Optimization

Another important advantage of digital marketing is web analytics which measures the result of digital marketing campaigns in real-time. This helps to optimize future campaigns and fix any possible mistakes quickly. Analyzing your digital marketing campaigns also enables you to have the ability to pinpoint every source of traffic and take total control of your sales funnels.

Digital Marketing Manager Responsibilities:

- Designing and overseeing all aspects of our digital marketing department including our marketing database, email, and display advertising campaigns.
- Developing and monitoring campaign budgets.
- Planning and managing our social media platforms.
- Preparing accurate reports on our marketing campaign's overall performance.
- Coordinating with advertising and media experts to improve marketing results.
- Identifying the latest trends and technologies affecting our industry.
- Evaluating important metrics that affect our website traffic, service quotas, and target audience.
- Working with your team to brainstorm new and innovative growth strategies.
- Overseeing and managing all contests, giveaways, and other digital projects.

Digital Marketing Manager Requirements:

- Bachelor's degree in marketing or relevant field.
- A minimum of 5 years experience in a digital marketing or advertising position.
- In-depth knowledge of various social media platforms, best practices, and website analytics.
- Solid understanding of HTML, CSS, and JavaScript is required.
- Highly creative with excellent analytical abilities.
- Outstanding communication and interpersonal skills.
- Up-to-date on the latest trends and technologies in digital marketing.

Customers now interact with your brand in a multitude of ways from reading a magazine to coming across a sponsored ad on Instagram.

The secret to driving success for your brand is to create a seamless online to offline experience from one point to another using an omnichannel approach. What this involves is integrating new channels with traditional assets. Examples include:

- **Events and Account-Based Marketing** – Your sales team can use in-person events to target key clients (insights garnered from your ABM research and insights). For example, you could target a CMO in a high-profile tech company by attending an event they sponsor and knowing that person is on a panel.
- **Mail-outs/brochures and online discounts** – Including leaflets in print media or mailing out brochures could include a discount code that drives people online to a custom landing page to claim the offer.

- **Outdoor advertising and geotargeting** – Billboards and banners are great for brand awareness in specific areas. You can use geotargeting from online platforms such as social media to narrow down the locations of prospects so you can place advertising in that area to drive engagement.
- **TV and QR codes** – QR codes are a great marketing tool for driving mobile users from traditional ads to online channels. Superbowl LVI for example just used a QR code on its ad to offer people \$15 in free bitcoin. This offer was then used on social media to drive engagement. QR codes can also be used in print and outdoor media to drive prospects online.
- **Print and website links** – Magazines and newspapers are wise to the fact that people search online. That's why print media include URLs as part of a page or offer to drive readers online. ASOS, for example, has a magazine that profiles products and celebrities and includes each item in a 'Shop the ASOS magazine' which drives traffic to a web page.
- **In-store and apps** – According to Mercator Advisory Group, 52 percent of shoppers used a mobile app to purchase while shopping in-store in 2020. The reason for this was either to find or redeem coupons or find sale items to purchase that could not be located in-store. For retailers, this shows the value of including a mobile experience in any in-store experience to build customer loyalty and drive purchases.

Examples of great traditional and digital marketing

There are lots of great examples of brands being innovative and using all touchpoints in a customer journey to drive engagement. Let's look at a few of the most innovative.

Adidas

Using a QR code located on the tongue of a sneaker to launch the new Pulseboost HD range, Adidas drove customers to a Spotify playlist, These playlists contained music based on their current location to drive geotargeting and personalize the experience for the customer.

This partnership with Spotify has been expanded to add QR codes to t-shirts and hoodies while the playlists are also growing.

Website -Hosting and Domain- Different platforms for website creation- Introduction to SERP- What are search engines- How search engines work- Major functions of a search engine- What are keywords -Different types of keywords- Google keyword planner tool.

Website:

A website is a collection of many web pages, and web pages are digital files that are written using HTML(HyperText Markup Language). To make your website available to every person in the world, it must be stored or hosted on a computer connected to the Internet round a clock. Such computers are known as a **Web Server**.

The website's web pages are linked with hyperlinks and hypertext and share a common interface and design. The website might also contain some additional documents and files such as images, videos, or other digital assets.

With the Internet invading every sphere, we see websites for all kinds of causes and purposes. So, we can also say that a website can also be thought of as a digital environment capable of delivering information and solutions and promoting interaction between people, places, and things to support the goals of the organization it was created for.

Components of a Website: We know that a website is a collection of a webpages hosted on a web-server. These are the components for making a website.

- **Webhost:** Hosting is the location where the website is physically located. Group of webpages (linked webpages) licensed to be called a website only when the webpage is hosted on the webserver. The webserver is a set of files transmitted to user computers when they specify the website's address..
- **Address:** Address of a website also knows as the URL of a website. When a user wants to open a website then they need to put the address or URL of the website into the web browser, and the asked website is delivered by the webserver.
- **Homepage :** Home page is a very common and important part of a webpage. It is the first webpage that appears when a visitor visits the website. The home page of a website is very important as it sets the look and feel of the website and directs viewers to the rest of the pages on the website.
- **Design :** It is the final and overall look and feel of the website that has a result of proper use and integration elements like navigation menus, graphics, layout, navigation menus etc.
- **Content :** Every web pages contained on the website together make up the content of the website. Good content on the webpages makes the website more effective and attractive.
- **The Navigation Structure:** The navigation structure of a website is the order of the pages, the collection of what links to what. Usually, it is held together by at least one navigation menu.

How to access Websites?

When we type a certain URL in a browser search bar, the browser requests the page from the Web server and the Web server returns the required web page and its content to the browser. Now, it differs from how the server returns the information required in the case of static and dynamic websites.

Types of Website:

- Static Website
- Dynamic Website

Static Website: In Static Websites, Web pages are returned by the server which are prebuilt source code files built using simple languages such as HTML, CSS, or JavaScript. There is no processing of content on the server (according to the user) in Static Websites. Web pages are returned by the server with no change therefore, static Websites are fast. There is no interaction with databases. Also, they are less costly as the host does not need to support server-side processing with different languages.

Dynamic Website: In Dynamic Websites, Web pages are returned by the server which is processed during runtime means they are not prebuilt web pages, but they are built during runtime according to the user's demand with the help of server-side scripting languages such as PHP, Node.js, ASP.NET and many more supported by the server. So, they are slower than static websites but updates and interaction with databases are possible. Dynamic Websites are used over Static Websites as updates can be done very easily as compared to static websites (Where altering in every page is required) but in Dynamic Websites, it is possible to do a common change once, and it will reflect in all the web pages.

There are different types of websites on the whole internet, we had chosen some most common categories to give you a brief idea –

- **Blogs:** These types of websites are managed by an individual or a small group of persons, they can cover any topics — they can give you fashion tips, music tips, travel tips, fitness tips. Nowadays professional blogging has become an external popular way of earning money online.
- **E-commerce:** These websites are well known as online shops. These websites allow us to make purchasing products and online payments for products and services. Stores can be handled as standalone websites.
- **Portfolio:** These types of websites acts as an extension of a freelancer resume. It provides a convenient way for potential clients to view your work while also allowing you to expand on your skills or services.
- **Brochure:** These types of websites are mainly used by small businesses, these types of websites act as a digital business card, and used to display contact information, and to advertise services, with just a few pages.
- **News and Magazines:** These websites needs less explanation, the main purpose of these types of websites is to keep their readers up-to-date from current affairs whereas magazines focus on the entertainment.
- **Social Media:** We all know about some famous social media websites like Facebook, Twitter, Reddit, and many more. These websites are usually created to let people share their thoughts, images, videos, and other useful components.
- **Educational:** Educational websites are quite simple to understand as their name itself explains it. These websites are designed to display information via audio or videos or images.

- **Portal:** These types of websites are used for internal purposes within the school, institute, or any business. These websites often contain a login process allowing students to access their credential information or allows employees to access their emails and alerts.

Hosting and Domain:

Web hosting

A web host provides the space where you display your site's content, like text, images, and videos. A web host doesn't necessarily provide the address visitors use to reach your site, like www.yourdomain.com.

When you build a site with Squarespace, Squarespace is your web host. This means that in addition to providing tools for creating and managing your content, we provide a place on the internet to display your content. Every Squarespace site is stored on our servers, similar to how physical stores rent space in a shopping mall.

Domain hosting

A domain host provides a domain name, like www.yourdomain.com, that visitors can use to find you. A domain name is like a street address that directs people to your website's location, but it's not the content that visitors see when they visit your site.

Domain hosts store domain names and facilitate their registration. If you're using a domain registered through a third-party provider, like GoDaddy or Hover, they're your domain host, and you'll manage your domain through them.

When you create a Squarespace site, you're automatically assigned a built-in domain, like yoursite.squarespace.com. You can also transfer supported domains to Squarespace or register a new custom domain.

How both hosts work together

You can connect one or more custom domains to your site by registering a Squarespace domain, transferring a domain to Squarespace, or connecting a domain from a different provider.

No matter who hosts your domain, visitors will see your Squarespace hosted web content while your domain is linked to your Squarespace site. Even if you use multiple domains, they'll all forward to a single primary domain. This is similar to how you can forward an email from one email address to another.

Paying for hosting

When you upgrade to a paid Squarespace plan, your plan only includes web hosting, but you can also register domains through Squarespace for a fee. If you sign up for an annual billing plan, you may also be eligible for one free custom domain for your first year.

Different platforms for website creation

The Indeed Editorial Team comprises a diverse and talented team of writers, researchers and subject matter experts equipped with Indeed's data and insights to deliver useful tips to help guide your career journey.

Web design is the creation of websites and pages to reflect a company's brand and information and ensure a user-friendly experience. Appearance and design are incorporated as vital elements whether you're designing a website, mobile app or maintaining content on a web page. Gaining web design skills can help you in applying for roles where your creativity could help a business improve their brand, their message and their bottom line.

In this article, we take a look at web design, including what web designers do and the common web-designing elements they work with and use.

Get the latest trending stories, job search tips, career advice and more!Subscribe

What do web designers do?

Web design identifies the goals of a website or webpage and promotes accessibility for all potential users. This process involves organizing content and images across a series of pages and integrating applications and other interactive elements.

The professionals who perform this process are called web designers, and their job includes the following duties:

- Selecting easy-to-read fonts
- Choosing attractive color schemes that also enable easy-to-read fonts
- Implementing a brand's identity into the colors, fonts and layout
- Creating a map of the website's structure to ensure intuitive navigation
- Placing images, logos, text, videos, applications and other elements
- Using coding languages, such as HTML and CSS, to create layouts and to style pages
- Making optimized versions of websites and pages both for desktop and mobile viewing

There are two common web design methods: adaptive and responsive design. In adaptive design, the website content is created using standard screen sizes as the frame for the layout.

In responsive design, content moves dynamically according to the screen size. Web designers use the various steps of the general web design process to employ these design methods depending on their client or employer's preferences and goals for the site.

What are the elements of web design?

The web design process allows designers to adjust to any preferences and provide effective solutions. There are many standard components of every web design, including:

Layout

The layout of the website is how the material is displayed on a page. Choosing the layout is an essential task for the designer. It should be simple, intuitive and accessible. Web designers can use blank areas called white spaces to organize the elements of the site with grid-based designs to keep them in order.

Designers can create specialized layouts for desktop screens and mobile devices. Mobile-friendly websites are a necessity because many visitors access websites on their cellphones or tablets.

To ensure a website is ready for mobile visitors, the designer can use a responsive template that adapts to different screen sizes or a mobile-only look that will activate when a non-desktop device connects to the website. A consistent layout between supports contributes to the visitors' trust.

Images

Images are illustrations, graphics, photographs, icons and others used to provide supplementary information to the text. To create the effect desired, designers can pick images that complement each other and the brand that the website represents.

Visual hierarchy

Visual hierarchy is the order in which the user will process the information on the site. The designer creates it by applying a visual pattern to the website. The visual pattern is the way the design directs visitors' eyes and behaviors.

For example, F-Patterns or Z-Patterns emphasize the top horizontal section of your site, where most designers place navigation and the brand's logo and sometimes a search box. These are elements that inspire user interaction and brand recognition.

Color scheme

The color scheme is a combination of colors that is in harmony with the brand and industry it represents. To achieve this, they will pick a dominant color and a few others to create a palette. A color palette can be monochromatic (different shades of the same color), analogous (colors close to each other) or complementary. Designers also account for what colors users are more likely to be attracted to.

Typography

The typography is the style or font of the written content. Web designers pick one or a combination that is attractive and easy to read. To make the best choice, they should choose a font that corresponds to the target audience. Some sites may be better in serif fonts while others can use non-serif fonts, depending on the site's industry, purpose and typical user.

Readability

Readability is when the text of content is easy to see and read on a webpage. The text on the website should be readable because visitors usually spend little time on it and should find information quickly. The designers can achieve this by selecting an appropriate size and pixel for the text. The contrast between the text and the site's background colors also improves readability.

Navigation

The navigational elements are the tools allowing users to choose where they want to go within a website. They may be present in the header, body and footer of the website, depending on the site's layout and structure. These elements are essential as they direct visitors to the information they want as quickly as possible.

Designers can choose a variety of navigation designs and layouts, such as using a button that hides and reveal navigation menus. They can also incorporate one-click arrows and other buttons that direct users back to the top of a page, to a specific area of a page or another page entirely.

Content

Content is all of the information available on the website. It is a pivotal element because visitors want to get information quickly. When the website communicates clearly and grabs the readers' attention, it is more likely to convert them into consumers. The designer can achieve this by using the appropriate tone and provide the right information on the entire website, including the "About" and "Contact" pages.

What is the use of web design?

Web design is used to achieve various tasks and goals, including:

- **Search engine optimization:** Search engine optimization (SEO) is a method for improving the chances for a website to be found by search engines. Web design codes information in a way that search engines can read it. It can boost business because the site shows up on the top search result pages, helping people to find it.
- **Customer satisfaction:** A professional web design impacts clients' satisfaction positively as it provides them the information they are looking for quickly. It helps the company build a positive relationship

with the visitors by ensuring the navigation on its website is easy to understand, predictable and consistent.

- **Mobile responsiveness:** Mobile responsiveness is the feature of a website that allows it to display on a mobile device and adapt its layout and proportions to be legible. Web design ensures sites are easy to view and navigate from mobile devices. When a website is well-designed and mobile-responsive, customers can reach the business with ease.
- **Consistent branding:** Branding refers to the promotion of a product with a unique design. Web design helps companies build or maintain a clear brand for their business. When a website expresses a business's brand consistently, it makes it easier to navigate and helps customers more clearly identify the visual elements of a brand as a specific company and its products or services.
- **Technical efficiency:** This term refers to how productive a website can be in making a comfortable experience on a website. Designers can achieve this with clean coding that allows for quick loading times, functioning links and dynamic images and graphics. Web design services also fix those eventual glitches when they occur.
- **User experience optimization:** Web designers run reports to understand the way people are interacting with a website all over the world. They determine which pages have more or less traffic and adapt the web design to optimize the user experience.
- **Conversion:** Conversion happens when a visitor completes a desired action on the website. Attractive web design encourages visitors to stay long enough to be converted into consumers. They will click on a call-to-action button, exchange valuable information and subscribe or buy a product.
- **Improved sales:** Increasing the number of items sold or acquiring more active customers are objectives of a compelling website. As web design reaches targeted customers and search engines, it helps the business make conversions on their site and improve its sales.

1. [Wix](#)
2. [HubSpot](#)

3. [Framer](#)
4. [Pixpa](#)
5. [Squarespace](#)
6. [Hostgator](#)
7. [Weebly](#)
8. [Shopify](#)
9. [BigCommerce](#)
10. [Duda](#)
11. [Wordpress](#)

Introduction to SERP

SERP stands for Search Engine Result Page. The main role of SERP is listing a set of pages over the web-based on the keywords research. In other words, when you open the browser and search for the information over the internet through search engines, then the list of web pages comes as a result. These results are nothing but SERPs.

SERP results can be organic search results or can be paid advertisements. Organic search results are significant than the paid advertisements as many times users ignore the paid advertisements. The rank of the web pages on the SERP is page rank. High page ranking helps organizations to get more visitors to the site and increases the number of impressions.

As per the statistics, 75% of the users don't navigate after the first or second page. So it is essential to ranking at the top position, i.e. on the first page.

How Does SERP Works?

As we already discussed, web pages the user will get when they search for a particular keyword using a search engine. Users enter the keyword, i.e. query he wants to know. Based on that keyword search engine will present the results to the user.

Search engine result pages change as time passes. Their appearance changes as search engines like google, yahoo, etc., updates regularly to give the best services for their users. In other words, because of the emerging of new technologies in the search engine appearance of search engine result pages of today may differ from the appearance of search engine results emerging of new pages of previous.

Every search query shows unique results even if you use the same keyword and same search engine. The reason behind their unique feature is all search engines customize their strategies to provide the best results based on a certain factor like user browsing history, location, social settings, etc.

SERP Types of Results

Search engine result pages show two types of results – organic results and paid results. Let's discuss those in detail.

1. Organic Results

Organic results show the list of web pages that the search engine gives by performing some searching algorithms. SEO specialist search engine optimization professionals know how to optimize content and rank the website at the top position in organic results.

To understand organic results better, refer to the below image, which shows the list of organic results.

I have searched for educba; it gives the SERP, i.e. list of web pages as shown in the figure.

You can see a box on the right-hand side of the SERP; this box is called a knowledge graph or knowledge box. Google introduced a knowledge graph in 2012. The goal behind this feature is to fetch the data for the queries from the available source across the world wide web to give answers to the queries at particular locations on the SERP

In the above-mentioned image, you can see the information about educba, such as an address, phone number, office timing. Every fact has its own links to other web pages. Some of them give significant organic results than others. This happens because of the different meanings of various searches.

Internet searches can be of three types of informational searches, Navigational searches, and transactional searches.

- **Informational Searches:** Users find the information for a given topic, e.g., educba. It will give the result of educba on SERP rather than placing ads.
- **Navigational Searches:** Navigational searches are those in which users visit the specific site through search. E.g., if you don't remember the educba site URL, you will type educba in search engines. This will give a SERP related to educba, and when you visit the site through the SERP.
- **Transactional Searches:** In transactional searches, paid results are shown on the SERP.

2. Paid Results

Paid results are those that an advertiser has paid to display their content at the top position on SERP. In earlier days, paid results were limited to text-based ads that had a small duration of time. Those ads were displayed in the corner of the screen above organic search results. But nowadays, paid results are in various forms. There are so many different formats in which paid results are displayed.

To understand paid results better, refer to the below image, which shows the list of paid results.

In the above-mentioned example, you can see I have searched for Books; it has given the list of web pages on SERP; these all are paid results. Also, there is a list of nearby book stores and maps based on the SERP of the keyword. These are the Only findings on this SERP that are not explicitly laid out: the map and company listing. This map is displayed based on a user's place and listings of features for local companies setting up their free Google My Business listing. Google My Business is a free company directory that can help smaller local businesses improve their visibility to Geolocation-based search engines.

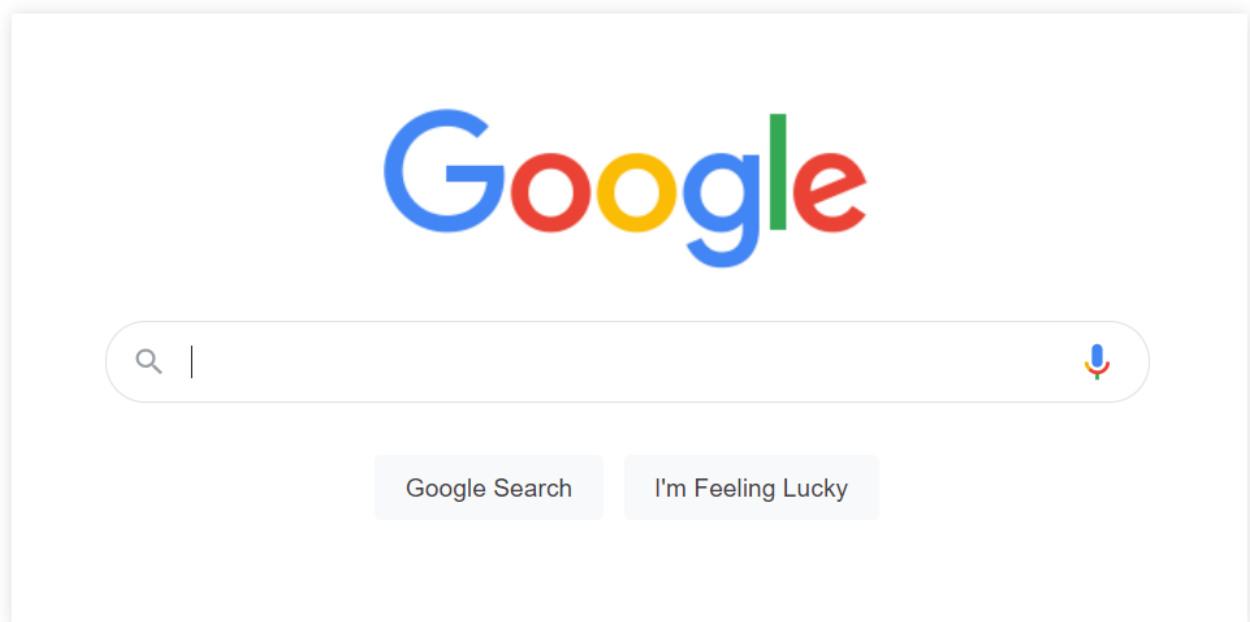
You can see there is a large text-based ad in the image given below.

These paid ads appear on top of the SERP viz. prime position for advertisers. These are PPC (Pay Per Click) ads. Shopping advertisements are the image-based advertisements on the page's right, a feature provided on the Google AdWords platform that enables product data for e-commerce distributors to be displayed along with other SERP outcomes. These advertisements consist of a wide variety of data, including specifications, accessibility of products, detailed customer reviews and rankings, offers, and more.

What are search engine

A search engine is an online tool that is designed to search for websites on the internet based on the user's search query.

It looks for the results in its own database, sorts them and makes an ordered list of these results using unique search algorithms. This list is called a search engine results page (SERP).



Although there are various search engines in the world (e.g. Google, Bing, Yahoo, etc.), the general principles of searching and providing answers are the same across all of them.

The beginnings of search engines

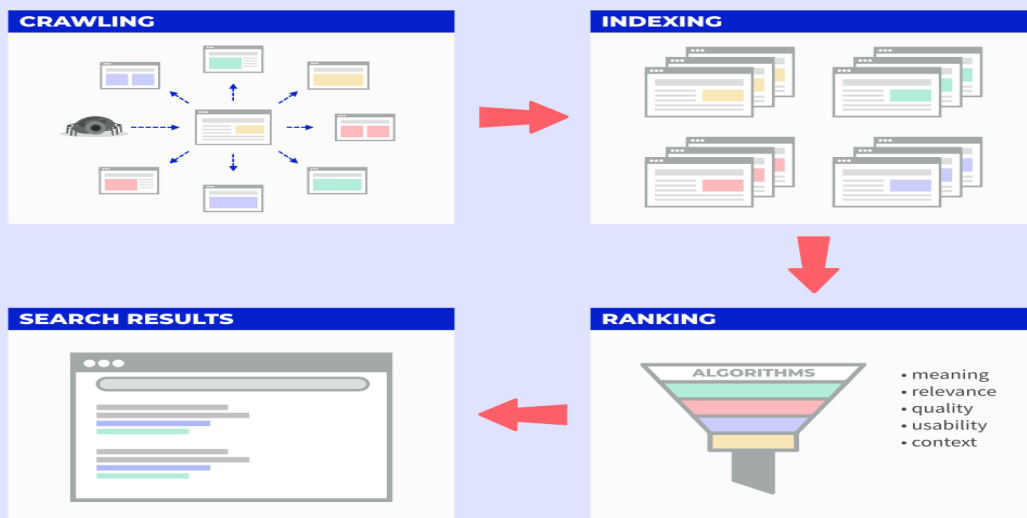
- The first internet search engine was **Archie Query Form** introduced in 1990 – it was a simple tool that searched for FTP sites and provided them in the form of a list. It did not include any further content.
- **Google's predecessor – BackRub** was introduced in 1996. This search engine brought the basic principle of backlinks and laid the foundation for the PageRank algorithm, which is used to this day.
- **Google** was launched in 1998 as a successor of BackRub and over the years became the most popular and dominant search engine in the world.

How search engines work

Search engines can differ from one to another in their ways of providing the answers to the user but all of them are built on the 3 fundamental principles:

1. Crawling
2. Indexing
3. Ranking

HOW SEARCH ENGINES WORK



1. Crawling

The actual discovery of new webpages on the internet starts with the process called crawling.

Search engines use small programs called web crawlers (sometimes called bots or spiderbots) that follow links from already known pages to the new ones that need to be discovered.

Every time a web crawler finds a new webpage through a link, it scans and passes its content for further processing (called indexing) and continues in the discovery of new webpages.

2. Indexing

Once the bots crawl the data, it's time for indexing – the process of validating and storing the content from the webpages in the search engine's database called "index". It is basically a big library of all the websites.

Your website has to be indexed in order to be displayed on the search engine results page. Keep in mind that both crawling and indexing are continuous processes that take place over and over again to keep the database fresh.

Once the webpage is analyzed and saved in the index, it can be used as a search result for a potential search query.

A quick tip: To find all your indexed pages, you can use Google Search Console or you can do a quick check by using the *site search operator*: “*site:domain.com*”

3. Ranking

The last step includes picking the best results and creating a list of pages that will appear on the result page.

Every search engine uses dozens of ranking signals and most of them are kept as a secret, unavailable to the public.

What is a search engine algorithm?

The search engine algorithm is a term used to define a complex system of several algorithms that evaluates all the indexed pages and determines which of them should appear in the search results for a given query.

For example, the Google algorithm uses dozens of factors (many of them are well-known, while some of them are kept a secret) in several areas such as:

- **Meaning of the query** (understanding what the user means by using the exact words they used, what is the search intent, etc.)
- **Page relevance** (the search engine needs to find out whether the page answers the search query)
- **Content quality** (the algorithms determine whether the webpages are an excellent source of information based on internal and external factors; number and quality of backlinks are important factors here)
- **Page usability** (considers the quality of webpage from the technical standpoint – responsiveness, page speed, security, etc.)

Search engine optimization

Besides providing useful information for their users, search engines can also help brands to promote their websites.

Optimizing your website for the relevant search queries is an important part of any online marketing strategy since it can drive more traffic to your webpages.

The sum of all the practices and techniques the website owners do to improve their search rankings is called Search Engine Optimization (SEO).

If we wanted to simplify SEO, we could say it all revolves around the 3 most important factors:

- Technical optimization
- Great content
- Quality backlinks

What are the most popular search engines?

Although there are hundreds of search engines in the world, only a few of them dominate the overall search engine market and remain popular thanks to their quality, usefulness, etc.

In terms of worldwide popularity, Google has been #1 for years. This is the list of top 5 most popular search engines:

1. Google

Google is the biggest and most popular search engine in the world.

Owned by its parent company Alphabet, Google dominates the search engine market with over 90 percent market share worldwide.

With all its features that include sophisticated algorithms, effective crawling, indexing, and ranking, Google provides excellent search results not only within its own search engine, it powers some other search engines as well (e.g. ask.com).

2. Microsoft Bing

Bing is the second largest search engine. It was launched in 2009 and it's owned by Microsoft.

Although it is impossible to compare Bing as a real opponent to Google with only 2 – 3 percent of the overall search engine market share, it is still a great alternative for those who would like to try something different.

Microsoft Bing is in many ways similar to Google, providing search result types like images, videos, places, maps or news.

Although Bing uses the fundamental principles of search engines (crawling, indexing, ranking) it also uses a special algorithm called Space Partition Tree And Graph which is based on the vectors for categorization of information and for answering the search queries.

3. Yahoo!

Yahoo is a popular website, email provider and the third biggest search engine in the world with almost 2% of the overall search engine market share.

Once a very popular and dominant search engine, Yahoo was dropping in value over the years and became somewhat overshadowed by Google.

Nowadays, Yahoo competes with smaller search engines such as Bing or DuckDuckGo.

4. Yandex

Yandex (from the term “Yet Another iNDEXer”) is a search engine that is popular mostly in the eastern countries.

Although it has less than 1 percent of the overall search engine market share, it is one of the most popular search engines in countries like Russia (with over 60 percent of all searches in the country), Turkey, Ukraine or Belarus.

Similar to Google, Yandex provides various types of services including Maps, Translator, Yandex Money or even Yandex Music.

5. Baidu

Baidu is the most dominant search engine in China. Even though its overall worldwide market share is barely 1 percent, it represents over 80 percent of the market share in China with billions of searches every day.

Baidu is similar to Google in many ways. It provides classic blue links with green URLs and shows rich results the same way as Google does.

Keyword

In general, when we talk about **keywords** in the context of digital marketing, we are talking about SEO Keywords. Keywords are the most fundamental and significant element of Search Engine Optimization. Users use keywords on search engines when they are looking for something on the internet. That is why it is really important to make your website SEO optimized for important keywords for your business. SEO helps your website's pages to be ranked higher with search engines and makes it easier for people to find your website.

Keywords are the core of SEO.

There are many tools and programs such as Google Trends, Ad Words keywords planner that can help you find the most effective keywords for your website. Reevaluating your keywords list periodically is always a wise thing to do because the value keywords are constantly changing. One of the important goals of SEO keywords is to bring the right users to your website. To achieve this goal, you may need more specific keywords or phrases.

How do I use SEO keywords?

Placing keywords in a proper manner that optimizes ranking on your pages is the key to improving your page rank on search engines. Always remember, the most important thing when you optimize the website page is keywords relevance, not keywords density. Low keywords relevance may cause a higher bounce rate and hurt the website ranking.

Here are some tips for placing keywords

- Placing keywords in the title of the page. Ex: What is Total Link? Digital Marketing Terms
- Placing keywords in the meta tags: Bounce Rate is a percentage that shows how many visitors click on a single website page and shut it without click any other elements.
- Placing keywords in the URL. EX:
<https://digitalshiftmedia.com/marketing-terms/total-links-linking-root-domains/>

- Placing keywords in the image file paths: Digital Marketing Terms_Bounce Rate

Different types of keywords

1. Market segment keywords

These keywords are generic words associated with a specific brand or industry. They target audiences searching for general information, though they can be more specific for niche marketing needs. For example, someone looking to buy shoes for running might search for the general phrase "running shoes" rather than a more specific brand.

2. Customer-defining keywords

These keywords are intended for a specific category of customers. For example, you might consider the age of your target audience while using these keywords. You can then research their gender, occupation and place of residence in order to target a specific group for advertising. Your customer-defining keywords can address your target audience. If you deal with sportswear products, for instance, a customer-defining keyword to use could be "adult sports enthusiast." Try to find customer-defining keywords that reflect your brand's target market demographics.

3. Product-defining keywords

These keywords describe and explain a product. Customers use product-defining keywords for particular search results, such as specific items. Your brand needs to use product-defining keywords to outline the business's exact products or services. Buyers search for product-defining keywords when they are at the initial stage of making a purchase.

The best way to use these keywords is by first analyzing your product list and then coming up with a thorough explanation of every product on your list. After that, check the descriptions of your products and select at least two relevant keywords. Use these keywords as your product-defining keywords.

Related: Everything You Need To Know About Job Keywords

4. Product keywords

Product keywords are keywords that relate to specific brands' offerings. These keywords are phrases or terms that directly refer to a company's services or products. Every brand needs to identify product keywords for all its services and products to help its existing and prospective clients find its products via search. If you search for a word such as "copier," for example, you will most likely get results from a reputable brand. Whatever phrase you type, you will get results from brands that offer those products or services in the industry.

The sports industry, for another example, usually leverages product keywords since companies in this industry link to important sporting events and sportspersons. Someone searching for a prominent person is likely to come across a wide array of products from their sponsor on the first search page, making their name function as the product keyword.

5. Competitor keywords

These are the keywords that your competitor uses in their marketing strategy to get high search engine rankings. Perform keyword research to uncover the competitor keywords other businesses are using to generate traffic to their websites. Identifying the right competitor keywords helps you understand the specific keywords that are working for your competitors. It further gives you opportunities to draft new content, ultimately boosting your brand's search engine rankings.

6. Long-tail keywords

These are usually the longest search keywords, targeting a specific audience or topic. These keywords have low-competing keywords. The keywords also have limited search traffic, making it easier to rank them. Since long-tail keywords are more specific than other keywords, they may have higher conversion rates than most keywords. A good example of a long-tail keyword might be "best running shoes for injured knees."

7. Short-tail keywords

These keywords are also known as generic keywords. The popular, broad search keywords lead to tons of search traffic. This type of keyword comprises less than two words. Furthermore, they rank competitively compared to most keywords. Short-tail keywords are brief and contain one or two phrases. A good example of a short-tail keyword may be “running shoes.”

8. Mid-tail keywords

These keywords fall between short-tail keywords and long-tail keywords. Although mid-tail keywords have a relatively more minor traffic volume, they have higher conversion rates and little competition than other keywords.

9. Intent targeting keywords

These keywords match the intention of the user while they are searching for a particular phrase. These keywords are an essential part of paid search. Marketers can use it to conduct intent-driven marketing. Intent targeting keywords help marketers drive more traffic to their websites, generate more leads, and attract better prospective customers.

10. LSI keywords

Latent Semantic Indexing (LSI) keywords are conceptual phrases that search engines use to understand a website's content. For instance, you could write an article about "The Benefits of Eating Eggs." From the topic, you are writing for an audience that wants to know more about the specific benefits of eating eggs. However, you might forget to mention the phrase " food" somewhere in your article. Many search engines will still be able to identify and rank your article as a food-related article.

11. Phrase match keywords

These keywords look for precise matches within a search engine's search parameter to start an ad. For example, you may search for a site with content such as "dentists who install dental crowns." Some ads might pop up showing

you dental products. These ads are a result of a phrasal match which usually happens in the background. Phrase match usually contains multiple variations to help account for synonyms, misspellings, paraphrases, and implied terms.

Related: [How To Use Resume Keywords To Get an Interview \(Includes Tips and Examples\)](#)

12. Exact match keywords

Exact match keywords are most similar to short-tail keywords. Marketers usually use these keywords to target advertisers whose adverts open up when an internet user searches for a specific phrase on a search engine. Advertisers typically bid on these keywords, and search engines use them to target specific audiences with particular ads. Your brand can use these keywords to target people who search for particular terms. Ultimately, these keywords can increase your chances of getting conversion. Exact match keywords form part of some paid search services.

13. Negative keywords

These keywords are the opposite of exact match keywords. They prevent ads from popping up once the user searches for a particular term, usually referred to as negative matches. Some search engines consider words like “free” as negative keywords. This means if a user performs a search using this negative word, they might not see certain business results.

14. Related vertical keywords

These keywords usually offer a more detailed perspective into your business' content. Let's say you have a firm that specializes in selling computer hardware, for example. "Computer hardware dealer" could be a horizontal keyword in this context. The related vertical keywords, in this case, might be something like "selling printers" or "RAM for sale."

15. Locational keywords

These keywords cover anything that relates to a specific location. Locational keywords are instrumental for locational-based businesses. They might be something like "towing services North Carolina." Locational keywords can include words or phrases that aim to show ads that have businesses near the person who is searching. In this case, the locational keyword might be something such as "towing firm near me."

16. Long-term evergreen keyword

These are keywords that remain relevant indefinitely. While search volume might fluctuate, it won't affect these keywords. Long-term evergreen keywords remain relevant months and even years after publishing because people will search for content related to these keywords for a long time. Long-term evergreen keywords require updating very rarely, perhaps annually.

17. Informational keywords

Informational keywords are keywords that clients use while searching for general information about a particular topic, product, or service. Buyers usually use these keywords in the awareness stage of the buying process. Buyers are aware they want a specific product or to solve a particular problem. Thus they need relevant information before they make a purchasing decision. One excellent example of informational keywords might be "what are the best fishing rods?"

18. Navigational keywords

These keywords are also referred to as "go" keywords. People use these keywords when they want to navigate to a specific brand's website. People using these keywords already know why they need to buy a product and where they will get those products. Thus, they use specific buying keywords to find the right place to buy what they want. For example, a user might enter the navigational keywords "[brand] running shoes," using the specific name of the brand of shoe they want to buy.

In this case, the person searching for these keywords wants running shoes, and they have decided to get them from a particular company. They are using navigational keywords to navigate to a site that will help them find exactly what they need.

19. Transactional keywords

Transactional keywords are also referred to as "do" keywords. These are the keywords buyers use when they have already decided to purchase a particular product or service. Buyers use transactional keywords at the conversation stage of the purchasing process. For example, a user might search "buy running shoes online" when they are ready to make a purchase.

Google keyword planner tool

Keyword Planner helps you research keywords for your Search campaigns.

You can use this free tool to discover new keywords related to your business and see estimates of the searches they receive and the cost to target them.

Keyword Planner also provides another way to create Search campaigns that's centered around in-depth keyword research.

This article shows you how to use Keyword Planner to lay the groundwork for a successful campaign.

Benefits

- **Discover new keywords:** Get suggestions for keywords related to your products, services, or website.
- **See monthly searches:** See estimates on the number of searches a keyword gets each month.
- **Determine cost:** See the average cost for your ad to show on searches for a keyword.
- **Organize keywords:** See how your keywords fit into different categories related to your brand.
- **Create new campaigns:** Use your keyword plan to create new campaigns centered on in-depth keyword research.

Keep in mind that while Keyword Planner can provide insights into keyword targeting, campaign performance depends on a variety of factors. For example, your bid, budget, product, and customer behavior in your industry can all influence the success of your campaigns.

Instructions

To access Keyword Planner:

- **Your account must be using Expert Mode. You won't be able to access Keyword Planner if your account is using Smart Mode.**
- **You must complete your account setup by entering your billing information and creating a campaign. If you're not yet ready to spend money, you can choose to pause your campaigns.**

1. Create a keyword plan

Once you open Keyword Planner, there are 2 ways to create your keyword plan:

- 1. Search for new keywords by clicking Discover new keywords.**
- 2. Upload existing keywords by clicking Get search volume and forecasts.**

Discover new keywords

A. Get ideas for new keywords

B. Edit your list of keyword ideas

C. Add keywords to your plan and see a performance forecast

2. Understand your keyword forecast

Your plan forecast shows you how many conversions, clicks, or impressions you're likely to get for your keywords based on your spend. Learn more About Keyword Planner forecasts

Understand your forecast

3. Organize keywords into ad groups (English only)

UNIT V

TRENDING DIGITAL MARKETING

SKILLS Search Engine Optimization (SEO)-Search Engine Marketing (SEM).-Social MediaMarketing/Optimization- Email Marketing. Website Product Marketing- Content Writing. Marketing the created content online Copywriting- Blogging- Local Marketing. Google Ad Words - Campaign Management- PPC Advertising- Affiliate Marketing. Mobile and SMS Marketing- Marketing Automation-Web Analytics- Growth Hacking

Search Engine Optimization (SEO)

Search engine optimization is the science of improving a website to increase its visibility when people search for products or services. The more visibility a website has on search engines, the more likely it is that brand captures business.

Website visibility is commonly measured by the placement -- or ranking -- of the site on search engine results pages (SERPs). And companies always vie for the first page, where they are most likely to garner the most attention.

Using Google as an example, SERPs often feature ads at the top of the page. These are positions that businesses are willing to pay for to ensure placement on that first page. Following ads are the regular search listings, which marketers and search engines refer to as organic search results. The SEO process aims to increase a business's organic search results, driving organic search traffic to the site. This enables data marketers to distinguish between traffic that comes to a website from other channels -- such as paid search, social media, referrals and direct -- and the organic search traffic.

Organic search traffic is usually higher-quality traffic because users are actively searching for a specific topic, product or service for which a site might rank. If a user finds that site through the search engine, it can lead to better brand engagement.

How does SEO work?

While there is a way to maximize results, it is almost impossible to fully manipulate search algorithms. Businesses often look to the shortest path toward ideal results with the least amount of effort, but SEO requires a lot of action and time. There is no SEO strategy where something can be changed today with the expectation of clear results tomorrow. SEO is a long-term project, with daily action and constant activity.

Search engines use bots to crawl all website pages, downloading and storing that information into a collection known as an index. This index is like a library and when someone searches for something in it, the search engine acts as the librarian. The search engine pulls and displays relevant information from the search query and shows users content related to what they were looking for. Search engine algorithms analyze webpages in the index to determine the order those pages should be displayed on the SERP.

Here is a brief description of how search engine optimization works.

What algorithms evaluate for search engine optimization?

There are hundreds of factors that go into what content from the index gets displayed into a SERP. However, they bubble up into five key factors that help determine which results are returned for a search query.

1. **Meaning of the query.** To return relevant results, the algorithm needs to first establish what information the user is searching for. This is known as intent. To understand intent, the algorithm is looking to understand language. Interpreting spelling mistakes, synonyms and that some words mean different things in different contexts all play into the algorithm understanding searcher intent. For example, search engines would need to be able to distinguish between "bass" as a fish and "bass" as an instrument. Intent would be based on additional search terms, historical search, location search and more to display the correct information.
2. **Relevance of webpages.** The algorithm analyzes webpage content to assess whether the sites contain information relevant to what a user is looking for. This comes after intent is established. A basic signal for relevance would be if the webpage includes the keywords used in the search. This includes showing up in the body copy or page headings. But beyond keyword matching, search engines use aggregated

interaction data to determine if the page is relevant to the search query. This looks at anonymized data from previous searches to match the page with the query.

3. **Quality of the content.** Search engines' aim is to prioritize the most reliable sources available. The intelligence built into the algorithms can identify which pages demonstrate expertise, authoritativeness and trustworthiness in relation to the intent.
4. **Usability of webpages.** Web design and accessibility play a big part in search rankings. The algorithm looks for how the site appears in different browsers, if it's designed for different device types -- such as desktops, tablets and phones -- and if the page loading times work well for users with slower internet connections.
5. **Context and settings.** Search engines and their algorithms use information from past search history and search settings to help determine which results are most useful to a user at that moment. Country and location can be used to deliver content relevant to the area from which someone is searching. For example, someone searching for "football" in New England would get different results than someone entering for the same query in England.

SEO benefits

Search engine optimization is an essential marketing activity to make a website or business visible on the web. But it also provides several other benefits to companies.

Builds trust and credibility

Sites that rank high on SERPs are considered to be of the highest quality and most trustworthy. Results shown on the first page are the most relevant, resulting in more credibility for the business or website. Having the right content on the site and a good user experience will help the website rank higher.

Provides a competitive advantage

When good SEO is deployed consistently, those that do it more and better will outrank the competition. Many businesses feel they cannot afford to not be on the first page of a search result. But if a team works toward that goal and shows ahead of the competition, they will have a competitive edge.

Reaches more people

SEO helps attract any user with intent at any time, regardless of phase of the customer journey that user is in. It relies on keywords and phrases to attract audiences to specific products and services. Businesses can create a list of keywords for which they would like to rank, then build content around those keywords.

Supports content marketing

By having a list of keywords to rank for and building content around those keywords, users are more likely to find the information they seek. Content and SEO work in harmony with each other. A site will rank better by creating useful, high-quality content that is optimized for those keywords. Ensuring the keywords are present in headings, meta descriptions and the body of the content will improve rankings for those terms.

Ranks better in local searches

The use of local searches are becoming more common, with users looking for products or services "near me." To improve listings in these searches, a company can create a [Google My Business](#) account and optimize the listing for local searches. Along with that and the localized content on the website, a user will be more likely to see local search results in their queries.

Understand web environment

Users that stay up to date on the everchanging internet will be better able to execute the ongoing SEO needs for a website. By staying up to date, businesses can better understand how search works and make more informed decisions on how to change and adapt their strategies.

Relatively inexpensive

To have an effective SEO strategy, companies need to invest in the time and resources to be effective. There are companies that can be hired as SEO experts to manage the strategies, but companies with the right team in place can do it themselves.

Get quantifiable results

There are tools and analytics data that can be tapped into to measure the effectiveness of SEO efforts. Google Analytics can provide comprehensive data around organic traffic. Data includes pages that customers engaged with and keywords used in search. That data can then be cross-referenced with intended actions taken to see how SEO played a role in customer engagement or acquisition.

SEO techniques

There are three main components -- or pillars -- to SEO that go into building effective SEO strategies:

1. **Technical optimization.** This is the process of completing activities when building or maintaining a website to improve SEO. It often has nothing to do with the content that is on the page itself. Some ways to manipulate technical optimization include having an XML sitemap, structuring content in a way that is intuitive to user experience and improving site performance -- such as page load times, correct image sizing and hosting environment.
2. **On-page optimization.** This is the process for ensuring the content on a website is relevant to the users. This content includes the right keywords or phrases in the headings and body of the copy; and ensuring that each page includes meta descriptions, internal links within the site and external links to other reputable sites, and a good URL with the focus keyword. Website administrators use content management systems (CMS) to maintain on-page content.
3. **Off-page optimization.** This technique is deployed to enhance a site's rankings through activities outside of the website. This type of activity is driven largely by backlinks. Businesses can generate these from partnerships, social media marketing and guest blogging on other sites.

These three components help marketers focus on the activities and techniques to build strong rankings for their websites.

Here are some additional techniques to use within each of the three pillars:

- **Keyword research and selection.** Perform keyword research on the terms for which are most desirable to rank. Businesses should focus on keywords that receive high search volume to be relevant for search engines. Looking at competitors' top performing keywords also gives an opportunity to build a strategy to compete with them.
- **Create quality content.** When the keyword strategy is in place, the content strategy follows. Pages are more likely to appear higher on SERPs by creating quality content that is relevant to the readers and their search queries.
- **Develop unique page titles and meta descriptions.** Page titles should include the page's focus keyword. And meta descriptions should be brief summaries of what a user should expect to learn on the page. These elements are displayed in SERPs and will likely be what people use to inform their clicks.
- **Use pay per click to supplement organic traffic.** Paid advertising can help improve organic click-through rates by giving marketers an outlet to test title tags and meta descriptions that are shown in SERPs. Ensuring a first-page placement, these ads mimic organic search results to see what copy entices users to click. That can be used to adjust page titles and meta descriptions with organic results.
- **Use alt text with images.** Alt text is used to describe an image on a webpage. This is critical for bots crawling the site to understand what the image represents. It also verbally describes what the image is to people who are visually impaired. This is also another opportunity to input keywords.
- **URL slug.** A URL slug is the portion of the URL that is unique to a specific page. This is also an area where it is important to put the focus keyword as it relates to what is on the page.

SEO tools

There are hundreds of tools to improve, manage and report on the effectiveness of SEO. Some are free and some aren't. But they all help marketers research keywords, develop strategies and measure the results of ranking higher on SERPs. Here are some common SEO tools used by marketers today:

- **Semrush.** This platform is used for keyword research and online ranking data. It gives marketers insight into developing and maintaining keyword strategies. Semrush also can be used to measure SEO performance over time.
- **Google Analytics and Search Console.** This platform provides real-time data on the effectiveness of SEO. Combined with Google's Search Console, marketers can monitor website traffic, optimize rankings and make informed decisions about the appearance of a site's search results.
- **Yoast SEO.** For websites using WordPress as their CMS, Yoast SEO is a plugin used to improve on-page optimization. Users can define the URL slug, meta description and page title. They can also see how the content on their page may perform in searches. A checklist of items is available so users can ensure their page is as optimized as possible.
- **Ahrefs.** This tool is used to audit websites and provide keyword, link and ranking profiles. It can also identify which pages perform the best and which ones need improvement.
- **SpyFu.** This is a competitor keyword research tool for Google Ads. In addition to the keyword research and data it can produce, it gives detailed insight into competitor SEO and pay per click data.
- **HubSpot Website Grader.** This is a free tool that delivers report cards with actionable insights about SEO performance. It can show if a website is optimized for mobile, measures website performance and gives security recommendations.
- **Google Trends.** This tool looks for content trends in countries or regions of the world. It finds popular topics and long-tail keywords related to them. It also compares those trends over time.
- **Bing Webmaster.** This tool enables marketers to see backlink profiles and keyword research. It also has a built-in site scanning feature.
- **Consultants.** While not exactly a tool to purchase or use for free, consultants have the SEO expertise that some teams lack internally. Good consultants can help develop the right strategies, execute or recommend the long-term plan. They can also report on the metrics to determine success.

Search Engine Marketing (SEM)

Search engine marketing, or SEM, is one of the most effective ways to grow your business in an increasingly competitive marketplace. With millions of businesses out there all vying for the same eyeballs, it's never been more important to advertise online, and search engine marketing is the most effective way to promote your products and grow your business.

In this guide, you'll learn an overview of search engine marketing basics as well as some tips and strategies for doing search engine marketing right.

Search Engine Marketing – An Overview

Search engine marketing is the practice of marketing a business using paid advertisements that appear on search engine results pages (or SERPs). Advertisers bid on keywords that users of services such as Google and Bing might enter when looking for certain products or services, which gives the advertiser the opportunity for their ads to appear alongside results for those search queries.

These ads, often known by the term pay-per-click ads, come in a variety of formats. Some are small, text-based ads, whereas others, such as product listing ads (PLAs, also known as Shopping ads) are more visual, product-based advertisements that allow consumers to see important information at-a-glance, such as price and reviews.

Search engine marketing's greatest strength is that **it offers advertisers the opportunity to put their ads in front of motivated customers who are ready to buy at the precise moment they're ready to make a purchase**. No other advertising medium can do this, which is why search engine marketing is so effective and such an amazingly powerful way to grow your business.

SEM vs. SEO

SEM versus SEO: What's the difference?

Generally, "search engine marketing" refers to paid search marketing, a system where businesses pay Google to show their ads in the search results.

Search engine optimization, or SEO, is different because businesses don't pay Google for traffic and clicks; rather, they earn a free spot in in the search results by having the most relevant content for a given keyword search.

Both SEO and SEM should be fundamental parts of your online marketing strategy. SEO is a powerful way to drive evergreen traffic at the top of the funnel, while search engine advertisements are a highly cost-effective way to drive conversions at the bottom of the funnel.

Keywords: The Foundation of Search Engine Marketing

Keywords are the foundation of search engine marketing. As users enter keywords (as part of *search queries*) into search engines to find what they're looking for, it should come as little surprise that keywords form the basis of search engine marketing as an advertising strategy.

SEM Keyword Research

Before you can choose which keywords to use in your search engine marketing campaigns, you need to conduct comprehensive research as part of your keyword management strategy.

First, you need to identify keywords that are relevant to your business and that prospective customers are likely to use when searching for your products and services. One way to accomplish this is by using WordStream's Free Keyword Tool.

Simply enter a keyword that's relevant to your business or service, and see related keyword suggestion ideas that can form the basis of various search engine marketing campaigns.

WordStream's Free Keyword Tool provides you with a range of valuable information, such as search volume for each individual keyword in Google and its general competitiveness.

Social Media Marketing (SMM)

Social media marketing (SMM) (also known as digital marketing and e-marketing) is the use of social media—the platforms on which users build social networks and share information—to build a company's brand, increase sales, and drive website traffic. In addition to providing companies with a way to engage with existing customers and reach new ones, social media marketing (SMM) has purpose-built data analytics that allow marketers to track the success of their efforts and identify even more ways to engage.

Why Is Social Media Marketing So Powerful?

The power of social media marketing (SMM) is driven by the unparalleled capacity of social media in three core marketing areas: connection, interaction, and customer data.

Connection: Not only does social media enable businesses to connect with customers in ways that were previously impossible, but there is also an extraordinary range of avenues to connect with target audiences—from content platforms (like YouTube) and social sites (like Facebook) to microblogging services (like Twitter).

Interaction: The dynamic nature of the interaction on social media—whether direct communication or passive “liking”—enables businesses to leverage free advertising opportunities from eWOM (electronic word-of-mouth) recommendations between existing and potential customers. Not only is the positive contagion effect from eWOM a valuable driver of consumer decisions, but the fact that these interactions happen on the social network makes them measurable. For example, businesses can measure their “social equity”—a term for the return on investment (ROI) from their social media marketing (SMM) campaigns.

Customer Data: A well-designed social media marketing (SMM) plan delivers another invaluable resource to boost marketing outcomes: customer data. Rather than being overwhelmed by the 3Vs of big data (volume, variety, and velocity), SMM tools have the capacity not only to extract customer data but also to turn this gold into actionable market analysis—or even to use the data to crowdsource new strategies.

How Social Media Marketing Works

As platforms like Facebook, Twitter, and Instagram took off, social media transformed not only the way we connect with one another but also the way businesses are able to influence consumer behavior—from promoting content that drives engagement to extracting geographic, demographic, and personal information that makes messaging resonate with users.

SMM Action Plan: The more targeted your social media marketing (SMM) strategy is, the more effective it will be. Hootsuite, a leading software provider in the social media management space, recommends the following action plan to build an SMM campaign that has an execution framework as well as performance metrics:⁵

- Align SMM goals to clear business objectives
- Learn your target customer (age, location, income, job title, industry, interests)
- Conduct a competitive analysis on your competition (successes and failures)
- Audit your current SMM (successes and failures)
- Create a calendar for SMM content delivery
- Create best-in-class content
- Track performance and adjust SMM strategy as needed

Customer Relationship Management (CRM): Compared to traditional marketing, social media marketing has several distinct advantages, including the fact that SMM has two kinds of interaction that enable targeted customer relationship management (CRM) tools: both customer-to-customer and firm-to-customer. In other words, while traditional marketing tracks customer value primarily by capturing purchase activity, SMM can track customer value both directly (through purchases) and indirectly (through product referrals).

Shareable Content: Businesses can also convert the amplified interconnectedness of SMM into the creation of "sticky" content, the marketing term for attractive content that engages customers at first glance, gets them to purchase products, and then makes them want to share the content. This kind of word-of-mouth advertising not only reaches an otherwise inaccessible audience, but also carries the implicit endorsement of someone the recipient knows and

trusts—which makes the creation of shareable content one of the most important ways that social media marketing drives growth.

Earned Media: Social media marketing (SMM) is also the most efficient way for a business to reap the benefits of another kind of earned media (a term for brand exposure from any method other than paid advertising): customer-created product reviews and recommendations.

Viral Marketing: Another SMM strategy that relies on the audience to generate the message is viral marketing, a sales technique that attempts to trigger the rapid spread of word-of-mouth product information. Once a marketing message is being shared with the general public far beyond the original target audience, it is considered viral—a very simple and inexpensive way to promote sales.

Customer Segmentation: Because customer segmentation is much more refined on social media marketing (SMM) than on traditional marketing channels, companies can ensure they focus their marketing resources on their exact target audiences.

Tracking Metrics

According to Sprout Social, the most important social media marketing (SMM) metrics to track are focused on the customer: engagement (likes, comments, shares, clicks); impressions (how many times a post shows up); reach/virality (how many unique views an SMM post has); share of voice (how far a brand reaches in the online sphere); referrals (how a user lands on a site); and conversions (when a user makes a purchase on a site). However, another very important metric is focused on the business: response rate/time (how often and how fast the business responds to customer messages).

When a business is trying to determine which metrics to track in the sea of data that social media generates, the rule is always to align each business goal to a relevant metric. If your business goal is to grow conversions from an SMM campaign by 15% within three months, then use a social media analytics tool that measures the effectiveness of your campaign against that specific target.

Even in the digital age, people appreciate the human touch, so don't rely only on social media to get the word out.

Advantages and Disadvantages of Social Media Marketing

Tailored social media marketing (SMM) campaigns that instantly reach a range of target audiences are clearly advantageous to any business.

But—like any social media content—SMM campaigns can leave a company open to attack. For example, a viral video claiming that a product causes illness or injury must be addressed immediately—whether the claim is true or false. Even if a company can set the record straight, false viral content can make consumers less likely to purchase in the future.

What Is Sticky Content in Social Media Marketing?

Sticky content is the marketing term for attractive content that engages customers at first glance and then influences them not only to purchase products but also to share the content.

What Is Viral Marketing in Social Media Marketing?

Viral marketing is an SMM strategy that attempts to trigger the rapid spread of word-of-mouth product information—a very simple and inexpensive way to promote sales.

What Is Earned Media in Social Media Marketing?

Earned media is a marketing term for brand exposure from any method other than paid advertising, e.g., customer-created content ranging from product reviews and recommendations to shares, reposts, and mentions.

What Are Some Examples of Social Media Marketing Strategies?

Social media marketing has grown to include several techniques and strategies to engage users and market products and services. These include audience-targeted advertising, the use of interactive chatbots, creating personalized experiences for customers online, the use of social media influencers, building an online audience, and so on.

How Can One Get Started in Social Media Marketing?

To start working in social media marketing, it is good to have at least a bachelors degree in marketing or a related field. Then, it's critical to gain a good understanding of how marketing campaigns work on platforms like Facebook, Twitter, and Instagram. After that, showcase your talents by creating compelling and effective content. Follow influencers and other social media marketers to learn what they are doing well and where they fall flat. Together,

use these steps to create a personal brand that you can use to sell yourself and your work.

The Bottom Line

Social media marketing (SMM) is the use of social media platforms to interact with customers to build brands, increase sales, and drive website traffic. As social media usage grows around the world, both via computer and mobile devices, the ability to drive sales from certain user populations is a growing business, rife with competition for views and clicks.

SPONSORED

A Digital Wallet for All Your Web3 Needs

From crypto to NFTs and beyond, accessing a wealth of DeFi platforms is simpler than you might think. With OKX, a leading digital asset financial service provider, you can access world-class security as you trade and store assets. You can also connect existing wallets and win up to \$10,000 when you complete a deposit of more than \$50 through a crypto purchase or top-up within 30 days of registration. Learn more and sign up today.

Email Marketing

The use of email within your marketing efforts to promote a business's products and services, as well as incentivize customer loyalty. Email marketing is a form of marketing that can make the customers on your email list aware of new products, discounts, and other services. It can also be a softer sell to educate your audience on the value of your brand or keep them engaged between purchases. It can also be anything in between. Mailchimp can help you design, build, and optimize your email marketing to get the best ROI in your marketing program.

Level up your email marketing game

Use Mailchimp's email tools and all-in-one marketing platform to grow your brand and sell more stuff.

Sign up

When you want to grow your brand or sell your stuff, email marketing is one of the most popular—and effective—tools around for marketing campaigns. In this article we'll discuss how email marketing - and the usage of promotional emails - can help you to grow your business, and we'll give you a few tips to help you get started with a successful email marketing campaign.

What is email marketing?

Email marketing is a powerful marketing channel, a form of direct marketing as well as digital marketing, that uses email to promote your business's products or services. It can help make your customers aware of your latest items or offers by integrating it into your marketing automation efforts. It can also play a pivotal role in your marketing strategy with lead generation, brand awareness, building relationships or keeping customers engaged between purchases through different types of marketing emails.

A brief history of email

The very first email was sent in 1971 by a computer engineer named Ray Tomlinson. The message he sent was just a string of numbers and letters, but it was the beginning of a new era of communication. Tomlinson was also the person who introduced the usage of the "@" symbol in email addresses.

In 1978, a marketing manager at Digital Equipment Corp named Gary Thuerk used this new method of direct communication to send out the first commercial email to let people know about a new product.

By the '90s, the internet had become commercially available to the masses. The way people communicated with one another began to change dramatically, and marketers discovered that email could be an effective way to advertise. The emergence of marketing emails also ushered in the need for regulatory updates; the U.K.'s Data Protection Act, for example, was adjusted to require an "opt out" option for all marketing emails.

Advantages of email marketing

Email has become such a popular marketing tool for businesses partly because it forces the user to take some kind of action; an email will sit in the inbox until it's read, deleted, or archived.

Email marketing can help you build a relationship with your audience while also driving traffic to your blog, social media, or anywhere else you'd like folks to visit. You can even segment your emails and target users by demographic so you're only sending people the messages they want to see most.

Email marketing also allows you to run A/B tests of a subject line or call to action to identify the best performing message by using email marketing software that can also be configured to easily send out emails. Check out Mailchimp's email templates to see more of what you can do with email marketing.

Disadvantages of email marketing

While email marketing seems like the perfect way to reach out to customers, create new prospects, and grow important business relationships, there are some drawbacks. In fact, many businesses are opting to use EZ Texting as another form of communication.

Here are some of the significant downsides to email marketing campaigns.

Spam

It seems like our inboxes are filled with worthless information. "Lose 25 pounds in two weeks," "Click here for a big discount." We all get them and nearly instantly hit delete. In addition, we never even see many of these emails because they end up in our junk or spam folders. Unless you are

actively avoiding spam filters, these are messages are often just a waste of time for the company that sent them.

Size

If your email is too large, it might take a long time to load—or even not load at all. In that time it takes to download, a potential customer has just lost interest, costing you business.

Competition

Disadvantages aside, email marketing is a popular form of marketing, which means that your email isn't going to be the only one flooding users' inboxes. This means that to stand out from competitors, you might need to invest in strong copywriters or offer additional promotions to capture your audience's attention.

Engagement

Frequently, a customer sees an ad and signs up for emails based on that 1 instance or offer. They may or may not use it. In any case, they are now in the clients' database, but that doesn't mean they will keep opening up your emails and clicking through to your site. You have to continuously find ways to engage your audience, or you might find yourself with high unopened rates or a lot of people unsubscribing.

Design

Today, you can access an email across a range of devices, such as phones, tablets, and computers. This means that unless you're designing an email for each platform, your customers might see a less than ideal version of your email.

Email marketers don't know what type of operating system the recipient is using. In many cases, what was once a visually appealing email, can have odd breaks, missing visuals, and logos. These are annoying to the recipient and are quickly deleted—especially if the recipient mistakes it for spam or a scam. These emails are hard to read in most cases and are of very little value.

Cost

While many email services purport to be free, many still charge fees for additional actions such as adding images or exceeding a word count. Make sure that you know exactly what the guidelines are for free emails or understand what additional charges you may incur. When you have someone design an email template, help build a database of relevant contacts, and the dissemination of the email may start stressing the budget.

Email marketing types and examples

There are many different types of email marketing. Each one serves a different purpose and takes a different avenue to engage with your audience. We are going to look at some of the many different types, so you can create the best email marketing campaign for your company.

Welcome emails

This type of email welcomes customers and encourages them to learn more about your product or service. They often offer a trial or other bonus. It is used to introduce a potential new customer to the business.

Newsletter emails

Newsletter emails are very popular, and they often highlight new products and services. They may also include articles, blogs, and customer reviews. Usually, there will be a call to action to move the reader to do something, whether that is reading a new blog post or checking out a new product.

Lead nurturing emails

This type of email targets a specific audience through a series of emails in the hope of eventually converting them. Typically, lead nurturing emails focus on a group that is interested in a specific product or service and then build their interest through more emails that offer additional information or relevant promotions. The goal is to push users from the consideration stage to the purchasing stage.

Confirmation emails

Those that have recently signed up for emails or newsletters, or have purchased an item online for the first time may get a confirmation email. This ensures the prospect that the information has been received and they are on the list to receive additional information. These are also a way to let users know that their purchase has been received or that their sign-up was successful and can include more actions for them to take.

Dedicated emails

If you want to reach out to only a portion of your email list, this is called a dedicated email. Its list may be based on recent purchases, inactive clients, new members, and other specific types of criteria.

Invite emails

These types of emails often announce upcoming events, new product launches, and seminars. Most companies use these types of emails when there is something special going on to gain attention and increase awareness about special events.

Promotional emails

These types of marketing emails are very common and tend to be generic and go out to a large audience. They are usually used to maintain awareness and may tease new products and services.

Survey email

Feedback from customers is one of the best tools for a business. Sending out these emails communicates to your customers that you value their opinion and want to create an experience, product, or whatever you're offering that they'll enjoy. Businesses can also take the feedback from these surveys and apply them to their offerings, creating what is hopefully a better product.

Seasonal marketing emails

Many companies take advantage of the holiday season or special occasions to reach out to their customers and prospects with information on upcoming sales and promotions. They are often tied to holidays like Christmas, Valentine's Day, Mother's, and Father's Day.

Tips for building your email marketing list

But how do you build an audience of people to send email to as part of your internet marketing efforts in the first place? There are a few ways, and all of them have to do with treating your customers right, taking into consideration marketing best practices.

Don't buy email lists. Many email marketing companies (including Mailchimp) have a strict, permission-based policy when it comes to email addresses, which means that sending to purchased lists is prohibited. Instead, concentrate on encouraging folks to opt into receiving messages from you by using lead magnets. You could offer a discount on your customers' first orders when they sign up for your email list via a custom signup form. Or maybe you can offer new subscribers free shipping on their next order—or give them a chance to win a prize when they join your list. Here are some more tips to help you build an email list.

Be aware of national (and international) email regulations. Make sure you adhere to any legal requirements and applicable laws in your area when sending automated emails, like the CAN-SPAM Act in the United States, the Canadian Anti-Spam Law (CASL), or the General Data Protection Regulation (GDPR) in the European Union for the treatment of personal information. The regulations are based on both your location and the location of your subscribers, and it's your responsibility to know which laws apply to you.

Use email to have a conversation with your customers. Email is a great marketing tool, but it can help your business in other ways, too. Consider taking the occasional break from your regular marketing content to send out surveys, tell you customers how much you appreciate them after buying from you, following up after an abandoned cart, or just say hello. Not only does it give your audience a chance to provide you with valuable feedback, but it also allows them to get more insight into the person behind the business.

Only send when you really need to. Once someone has trusted you with their email address, don't abuse that trust. Flooding your audience's inbox with superfluous emails will cause them to lose interest or unsubscribe entirely. Focus on sending them relevant, engaging messages about the stuff they like, and they'll be loyal for a long time to come.

Website Product Marketing

Product marketing is the process of bringing a product to market. This includes deciding the product's positioning and messaging, launching the product, and ensuring salespeople and customers understand it. Product marketing aims to drive the demand and usage of the product.

Product marketing doesn't stop once the product has gone to market (if it did, well, product marketers at a one-product company wouldn't have much to do after the product's launch). The process of marketing a product as the final step is to ensure the right people are aware of the product. Those people who know how to use it, according to the needs and feedback of customers are being listened to over the product's lifecycle.

Let's talk about where to start in product marketing and what other aspects of your business can support this product as it grows.

A good way to begin brainstorming your campaign is through implementing inbound marketing methodology into your strategic plan. We mentioned before that product marketing is continual, and your approach should be the same. Inbound marketing is a strategy that focuses on attracting your audience and turning them into loyal customers that advocate for your product.

This is demonstrated in our “Attract, Engage, Delight” model below.

attract engage delight inbound methodology model for product marketing

You can attract, engage, and delight your customers with other aspects of your business including strategies that identify your target audience, provide a clear positioning or marketing message, and countless other ideas. But in short, starting your product marketing plan with this model and an understanding of inbound methodology can set your business up for success.

Now that we have a sturdy foundation to build upon, let's get into it.

What does a product marketing process look like before, during, and after a product is launched?

Product Marketing Starts With Your Customer

HubSpot's early years faced a challenge that many small businesses face: product ambiguity. Except for the slight majority of people who perceived HubSpot as "marketing services" — which is indeed part of our product stack — our perception consisted of numerous other terms that our audience used to describe us.

This is a primary reason businesses implement a formal product marketing operation, and it starts with your buyer persona.

A great product means nothing if it doesn't get the attention of the people who would benefit from it. So, who's your audience for this product? How (and where) are you reaching them, and what's the story you're telling to present this product to them? When preparing to launch a product, working with the rest of your marketing team to identify your customer and develop the messaging is critical.

Seven Critical Steps of Product Marketing

When product marketers know exactly whom their product caters to, the marketing can begin. Here are seven things product marketers may do before, during, and after their product enters the market:

1. **Product Research:** A helpful and well-made product isn't made in a vacuum, and it also isn't marketed in one. In the weeks and months before a product launch, product marketers work with the product's developers to test the product both internally and externally through controlled beta environments.
2. **Product Story:** Products are also brought to market in the form of a story. What problem does the product solve? Who's facing this problem? How does it solve this problem? What does it do that competitors don't?
3. **Product-Focused Content:** Product marketing's next stop is at the desks of the content creators. Here, product marketers may create and A/B test various marketing copy, blog content, case studies, and landing pages on their website — all dedicated to describing the product.
4. **Product Launch Plan:** No product marketing team is complete without a written launch plan, spelling out every last stage of the marketing process and who's responsible at each point.

5. **Product Launch Meeting:** When the product is launched, everyone involved meets the day it's rolled out. Much like a rocket launch, this is the product marketer's finest hour — it's the climax of a product marketing campaign.

6. **Community Engagement:** As product marketing generates enough buzz around the product within the industry, it's common for the marketing team to capitalize on what the market is saying about them. This includes reaching out to partners, influencers, and existing customers for commentary.

7. **Sales Enablement:** As a product is being prepared for the marketplace, the sales team is waiting in the wings to develop a sales strategy around this new business opportunity. It's the product marketing team's job to meet with sales staff before, during, and after the product is rolled out to the public. This ensures the messaging created for this product is consistent through to the first sales call.

With all of this in mind, you may be wondering what exactly a product marketer has to do to see these projects to completion. Let's dive into it.

Product Marketer Job Description

A Product Marketer, or Product Marketing Manager, promotes products and their features to an organization's target audience. Their duties include studying the company's products, highlighting key features to attract customers and creating marketing campaigns for products.

Product Marketer Responsibilities

A product marketer's main responsibility is to promote a product's value to the target audience. This goal is achieved through a combination of strategy and ideation such as:

Determining the mix of marketing content for creation and distribution

Creating and managing budgets for marketing campaigns

Working with content creators to make content that reflects the product and brand image

Managing a calendar of content and creating the schedule

Product Marketer Salary

A product marketer, or product marketing manager's salary in the United States varies greatly depending on the experience and tier. According to 2021 industry averages, the median salary of different tiers are as follows:

Entry-Level Product Marketer or Product Marketing Assistant: \$43,630

Product Marketer or Product Marketing Manager: \$111,890

Director of Product Marketing: \$166,928

Promote Your Product with a Plan

As you develop your product marketing team and strategy, think about how the elements above might take shape, and who you'll need to work with to make it a success. Take these questions into consideration in your next great product marketing plan.

Content Writing:

Content is a general term which is used to define the texts, graphics, audio, video or any informative element in the website. Until and unless you have valuable and interactive content no visitor is going to engage on the website. Your site must have the valuable, concise and appropriate content to engage the readers on the website. Content Writing is simply creating content for the website, blogs, social networks, e-commerce sites etc. Though every writer has their own voice but for content writer along with their own voice they have to be the voice of the brand for which they are working for.

The main job for content writer is to fill the site with information about business, products, services, industry, employees etc. in order to earn new customers. For lead based industry, the good content writer earns customer by getting website visitor to contact you. And if content writer works in retail, he can earn new customer by making additional sales on your sites.

The above listed both methods helps in growing the business, especially when you know your target audience. So content is most important part for your site to initialize digital marketing. But why it is so important? What is its need?

Importance of Content Writing in digital marketing:

Any person visits your site than they have reason, without reason there may be a visitor but very minute chance of becoming a customer. So a good content gives the visitor a reason to visit your site leads them to engage on your site and become your customer.

· First of all good content describes the visitor about the kind of your business or industry. If content writer has well knowledge about his target audience and he is able to answer the basic questions of visitors and your content has enough capability to convince the visitor about the product or services provided, then there are lots of chances for his conversion from visitor to a customer.

· A good content writing is useful for search engine optimisation that is, making rank of the site higher. Creating content and posting to your website is the best way to reach your potential customers. In the initial phase the content writer has to search the keywords that will help him to reach his audience. Once they well set keywords they have traffic to their sites, the writer could create content around those content and optimize it for search engines like Google.

· Content Writers could also mention links in their contents for their website and bring traffic to their sites. Links are one of the most essential key factor for SEO ranking. If in any content links to your website is mentioned and user clicks on it he will be redirected to your website and consequently, increasing the traffic to your site.

· A good content is useful for conversion of visitors to a customers. Good content engages the visitor and every time visitor decides to take action and becomes the customer it is called as conversion.

· In order to launching the digital marketing campaign the content will be needed. The well specified content is the one of the most important SEO parameter because it consist of the pages that includes the search results and ranks your site.

· A content is always shareable on social media. Irrespective of the kind of business every business or organisations have the social media accounts. Those social networks are used to promote the business. Whether it is on Facebook, LinkedIn, Instagram, Google+ etc. All these social networks promotes the content from you site.

If running any business is the mind then creating the content is the heart. If businessmen are the king then content writers are kingmakers. They just don't write five hundred words instead they understands the importance of the headlines, keywords that are going to increase the search rankings, and knows the best practice and effective usage of SEO.\

Copywriting is the art of producing text for promotional, marketing, instructional, or advertising purposes. The goal of this content is typically to persuade, inspire, entertain, educate, or inform a reader in order to increase brand awareness or authority, sell a product, describe a course of action, or to convince a person or group of people to take a specific step.

This text is produced by a copywriter. Copywriters can work as members of in-house marketing, editorial, or product teams, for copywriting agencies, or they can work for themselves as freelancers. Both agencies and freelance copywriters typically work for a variety of different clients fulfilling each company's different copywriting needs. Although in-house copywriters will only have one employer, they may often be asked to produce copy for a multitude of departments or teams within that organization.

Although copywriting isn't exclusive to online platforms, the explosion of online content and the increase in users consuming content online means that today a large majority of copywriters work in digital marketing teams. In this context, the copywriter works closely with the SEO specialist and marketing manager to produce the text needed to boost the online presence of a brand, grow an audience, and convert users into customers.

2. What is the difference between copywriting and content marketing?

While both copywriters and content writers require a deep understanding of language and communication, a passion for storytelling, and a love of prose, the stand out difference between the two roles is the goal of the content. While a copywriter is typically writing to persuade a reader to perform a certain action, a content writer seeks to educate or inform the readers with their text.

It's for this reason that copywriting tends to be more concise. Copywriters might produce the copy for an advertisement, a slogan, or a tagline which excites, stimulates, and persuades in a matter of moments. In contrast, content writing is longer, with a content writer seeking to educate their readers via articles, blog posts, ebooks, reports, or white papers, with extensive resources cited to support their arguments and add authority to their content. When this is done successfully, a reader is persuaded over the long term of the trustworthiness of the company and its expertise.

Both copywriters and content writers have the same overall goal: to convert a reader into a customer. However, a copywriter is seeking to do that with urgency, over a much shorter period of time, and therefore with a much shorter piece of writing. A content writer is playing the long game: building trust and demonstrating authority and knowledge in the field, via a long and well-researched piece of content.

Although a copywriter will likely have some understanding of SEO, a content writer will need to be well-versed in SEO best practices, and will typically work closely to the briefs set by the SEO specialist. A copywriter is

not needed to have such a deep understanding of SEO, as they will typically be working with very short texts that will not rank in search engine results.

Although there are clear distinctions between the two roles, many of the skills of the two positions overlap, and you'll often find companies advertising for one of these roles but actually seeking the skills and flexibility of both. Depending on the size of the company and the demands of the marketing and product teams, a writer will frequently work as both copywriter and content writer for the same organization.

3. What are the different types of copywriting?

Here are some of the most popular types of copywriting and content writing that in-house and freelance copywriters will produce in order to:

- Reach new audiences
- Educate users on an industry
- Communicate the brand's vision
- Persuade a potential customer of the benefits of a service
- Build trust

We'll go over these types in detail over the next few sections.

Blogs

A company blog publishes posts to educate, inform, and inspire readers on topics that are relevant to the product or service the company offers in order to increase the brand's visibility and attract more users to their site. In addition, a blog might be used for company and corporate updates and industry news.

Blogs are also a great opportunity for a company to produce informative how-tos or answer FAQs in detail which teach users about a product, explain a service, or demonstrate specifically and with easy-to-follow steps on how a feature works.

Ebooks

An ebook is an opportunity for a company to cover a topic in depth, demonstrating its expertise in a particular area, while passing on useful information, techniques, or advice to readers to help solve a problem or answer commonly-asked questions. The brand therefore becomes established as a trusted source for users when looking to solve related problems in the future.

Another purpose behind ebooks is that they can be a way for a company to collect leads. An ebook can be 'gated', meaning it is only available for download once a user has provided an email address. With high-quality pieces of downloadable content such as ebooks, a company can collect user email addresses to be used later for email marketing purposes or newsletters promoting other products or services.

Social media posts

Social media is an opportunity for businesses to directly engage with their existing customers, reach new audiences, and promote their vision, product, or service in a recognizable voice and tone. The copywriter's role here is usually to provide a relatively short entertaining, informative, or persuasive text that accompanies and corresponds with a posted image.

A seasoned copywriter will go above and beyond a descriptive two-liner, and instead seek to additionally involve the readers in a conversation, encourage audience participation, trigger an emotional response from the brand's followers, and include a powerful call to action (CTA) that prompts the reader to engage further with the brand.

Depending on which social media platform the text is being provided for, different styles of writing might be employed. A copywriter is expected to have knowledge of the styles popular with each network, and how best to engage with the different audiences those networks attract.

Newsletters

A newsletter is a customizable and trackable email that brings high-quality content directly to the user's inbox.

A company can use newsletters in a variety of ways, including:

- Sharing company news or events
- Highlighting products or features

- Drawing the user's attention to content on the company blog or social media pages
- Tracking user interest and engagement

A newsletter is an opportunity for the copywriter to use more personable, friendly language than a more formal article or blog post and to speak directly to the user about their challenges. In addition, the copywriter can customize the content of automated emails to make them appear to have been written to a specific individual, which encourages interaction and builds trust.

4. What are the qualities of good copywriting?

Having a good understanding of the key qualities of great copywriting will be a huge help as you prepare for a career in this exciting content production role. With this knowledge, you'll be able to focus on what's really important to the site's readership, and practice the techniques that guarantee your copy makes the right kind of impact on potential customers.

Let's take a look at what makes for great copywriting.

Communicates clearly

Ultimately, all copy is trying to tell us something. However, learning how to communicate your message clearly and succinctly takes time and practice. When you're starting out it can be tempting to show off your writing skills, but unless throwing in that Latin phrase really adds something to the message you're sending, it's best to stick to terminology that speaks to everyone. Essentially, your role is to enable a reader to take away what they need to know quickly and with ease—and, in the best cases, it should be fun to read, too.

Puts the reader first

It's easy to forget sometimes who we're writing copy for, especially when a client's strict deadline is looming, or your CEO is keen to share their wisdom. However, irrespective of who has requested the content, it's crucial to keep the person on the other end of your text in your mind at all times: the reader. Without the reader's attention, the goal of the copy—to reach new users, build trust, or generate interest—will be much harder, if not impossible, to reach.

Sounds natural

There's a reason why search engines rank high-quality content above content that simply contains the right keywords. It's because writing that sounds like it has been written by a robot does not connect with audiences, and therefore does not build trust, convert customers, or demonstrate expertise or authority. One of your advantages as a human is that you can speak the language of your audience. You can convey empathy for their problems and, because you've been there yourself, you can also help solve them.

Tells a story

Storytelling has gained increasing attention in marketing circles over the years, as marketing teams realize how well this works as a tool to influence, teach, and inspire. Storytelling builds connections between people and ideas: it helps a writer demonstrate empathy, and makes a brand seem relatable.

The great thing about storytelling is that it demonstrates to the reader that the brand not only understands their issues inside out, but it also knows how to solve them. When done skillfully and sincerely, storytelling builds trust with users which enables a marketing team to grow their readership and customer

Bloggging:

Blog Marketing is a World Wide Web evolution of business successes. Blog-style content has proved to engage humans on many psychological levels. This marketing strategy derived its purpose from original web loggers. The term "web log" became "weblog" which became today's "blog".

Blog Marketing strategy entered business marketing arenas because of the onset of both technological advances and evolutionary trends. These two events stemmed from personal online journals written by people who could program their own websites. Prior to 1999, only website programmers could customize their blogging platforms.

When blogging entered the business arena, it traveled through a series of trials beginning with pop culture, politics, and news events. The first media and business blogs were popular because they produced more in-depth coverage than mainstream media websites.

In 2017, businesses use Blog Marketing to engage their target market, potential customers, and retain to current customers.

Blog Marketing Production

In 2017, Blog Marketing has shown to increase lead flow by 60% on average and in some cases as much as 700%. Blog Marketing is an integrated part of content marketing and a strategy for acquiring leads and/or prospects. Predictably, these leads will buy a product or service from the company employing this tactic. There are differing opinions on the definition of leads and prospects. But, Blog Marketing is a proven, effective way to develop leads and prospects.

Also, as of 2017, Blog Marketing is easily managed and produced on the blog pages of a commercial website. The revolutionary ease of use with website formats like **WordPress** and **Hubspot** do not require webmaster services to move, add, delete, and update content on a Website.

Blog Marketing content relies on personas, also called customer profiles and buyer personas. Personas are defined by the use of search habits of website visitors' keyword searches as well as website behavior. Blog Marketing utilizes several other information-mining systems that record visitors' interactive behavior on a website. This website tracks persona actions and retrieves their data. This data then becomes the purpose for the content of the business blog.

Elements and Components of Blog Marketing

Sprout Content, an industry leader in "the business of conversation", offers the following reasons for businesses to incorporate Blog Marketing:

- Blogs use simple platforms that allow for connecting with and sharing timely and meaningful information for a direct information channel to customers.
- Search engine optimization (SEO): Search engines reward valuable content that is regularly published.
- Companies do not need the services of a webmaster to implement Blog Marketing because of the development of revolutionary blogging platforms like **WordPress** and **Hubspot**.
- Businesses that blog can accumulate up to 60% more customers
- A blog is the purpose for all content marketing goals. Businesses can use their blog's content on Facebook, Twitter, LinkedIn, and for ebooks and email newsletters or anywhere they distribute content.
- Blogs give a company a voice and connect its audience to new products and services, news, timely information, market trends, and to what happens at the company beyond its website.
- Blogs showcase a company's expertise and influencers and establish industry credibility.
- Blogs establish two-way conversations with customers, industry associates, and potential buyers by creating a comment section below the blog. This can build confidence in the company and can increase sales.
- One of the elements of Blog Marketing includes creating a blog calendar similar to an editorial calendar. Blogging keeps a company centered on its content marketing goals.
- Blog Marketing is a public relations tool. News outlets and journalists pay attention to high profile blogs and bloggers.
- A business must define who its audience is and why it is blogging to that demographic. This tightens up the purpose of the company's marketing strategies.
- Companies can promote their brand stories, relay company philosophies, highlight employees, and reveal innovative ideas through blogging. This lets customers and prospects understand how that business can help solve customer problems with its products and services.
- Blogs are long-term marketing assets and extremely cost-effective. Blogs do not go away. In 2017, many websites maintain a content curation page on their website which drives SEO.
- Companies can take advantage of blog analytics which keep a log of click-through rates, trending topics, social media shares, comments, and what days and times of the day the blog receives the most visitors.
- Sprout Content declares that blogging is inspiring. Employees involved in their company's Blog Marketing process begin to see new ideas in everyday events of life wherever they go.

Local Marketing:

Local marketing is particularly vital for small businesses and brands producing perishable or fragile goods. It allows them to generate more leads and ensure that consumers will visit their store. Such a strategy is often cheaper because business owners don't need to spend much money to reach a wide audience. Local marketing focuses only on consumers nearby that are more likely to buy your products.

Also, the local promotion allows companies to discover more about their clients and interact with them better. You can gain more insights into the needs, preferences, and behavior of people who live

in a certain area. Moreover, concentrating on a specific audience helps businesses provide a better experience, leading to higher customer satisfaction and increasing loyalty.

Finally, the local promotion allows companies to get more word-of-mouth referrals. Semrush Blog states that 90% of consumers are more likely to trust a recommended brand; consequently, positive reviews can help you save money on advertising and significantly increase the number of clients.

Now you understand why local marketing is a powerful tool to boost your company and generate more sales. Let's go further to discover ideas for small business promotion.

6 Local Marketing Ideas for Your Small Business

According to Nectafy, 88% of people who search for a local business visit it within 24 hours. This percentage is really impressive and means that discovering a local company is an incredibly vital step in the customer journey. Therefore, the main aim of local marketing is to increase the brand awareness of consumers who live nearby. Below we provide 6 ideas for small businesses to attract more potential clients to their company.

1. **Partner with other companies in the area.** Building strong networks is especially important for small businesses and can help them promote each other. For example, if you bake cakes and cookies, you can partner with coffee-to-go points in the area to sell your products to their clients. This strategy is extremely powerful when companies have similar values, buyer personas, and aims for future development.
2. **Encourage people to create UGC and leave feedback.** According to Forrester, 48% of people believe that user-generated content helps them discover new products. These statistics prove the importance of UGC in increasing the audience's brand awareness and generating more potential clients. Collecting feedback can boost your company even further because 58% of people are ready to pay more to a company with good reviews. Moreover, testimonials help companies rank higher in search results and increase the click-through rate. That is why 88% of small companies actively track their online reputation.
3. **Optimize your website and use local SEO.** To attract more clients in your region, your company should have a user-friendly website that loads quickly. It is better to make it handy for mobile consumers because 61% of them are more likely to contact a local business with a convenient website. SEO is also essential as 'Near me' searches have grown more than 5 times in the last two years. To make your website rank higher in the results, you should add local-based keywords, provide data about your working hours, give contact information, and collect positive reviews.
4. **Focus on the local audience on social media.** Firstly, you should make your content more specific. You can set a location in your profile, tag it in each post, write more about local events and share photos from popular places nearby. Collaborate with local influencers whose audience may include your potential customers. Secondly, you can use hyperlocal advertising on social media to target people within a small area. Narrow your audience according to their place of living to save money and reach more consumers nearby.
5. **Become a sponsor of events nearby.** This way of offline marketing allows small brands to increase citizens' awareness of their products. You can finance local concerts, sports competitions, or charity events to attract the attention of local media and make consumers more loyal to your company.
6. **Try email marketing.** According to Mailigen, 89% of marketers use email campaigns as the main channel to generate leads. It proves the effectiveness of email marketing, particularly for small companies. You can collect consumers' email addresses by offering some discounts or providing free online materials after purchase. Staying in touch with such consumers will encourage them to come back again and again, which is extremely important for small businesses. With SendPulse, you can launch bulk email campaigns for free to help your business save money and make product promotions more effective.

Use these ideas to attract more clients and create a strong competitive advantage in the local market. Continue reading to discover examples of effective local advertising campaigns and gain more insights on how to promote your business successfully.

Effective Local Marketing Campaigns

Launching local marketing campaigns can be beneficial for both global and small companies. In this section, we provide examples of successful local marketing to give you some ideas about your brand's promotion.

McDonald's

The world-famous fast-food cafe is not only about the standardized menu of hamburgers and fries. To increase the loyalty of consumers worldwide, McDonald's has launched a range of localized menus in different countries. For example, in Argentina, the restaurant offers a McFiesta burger with mayonnaise instead of ketchup. In India, you can buy a Vegetable Pizza McPuff — this position is not offered anywhere else in the world.

There is a McBaguette combo in the picture — an original McDonald's set from France. It consists of a sandwich with a breaded chicken, cheese, and ham, or a hamburger with two hash browns. It is a great example of localized dishes because McBaguette is created to satisfy the high demands of the French people.

Google Ad Words:

Google AdWords is a marketplace where companies pay to have their website ranked right with the top organic search results, based on keywords.

The basic gist is, you select to promote your brand based on keywords. A **keyword** is a word or phrase the user searches for, who then sees your ad. Your ads will only show up for the keywords you pick.

Google counts the clicks on your ads and charges you for each click. They also count impressions, which is simply the number that tells you how often your ad has already been shown when the users searched for that keyword.

If you divide clicks by impressions, you get the **click-through-rate** or CTR. This is the percentage of users who land on your advertised page, because they clicked on your ad.

Consider Google AdWords to be an auction house. You set a budget and a bid. The bid sets how much you are willing to pay per click. If your maximum bid is \$2, Google will only show your ad to people, if other aren't bidding more on average.

Google doesn't just want to show people the ads by the highest bidder – they could still be horrible ads. They care about their users so much that they'd rather show them a more relevant and better ad by someone who pays less.

Therefore – **Quality ads + good bid = win!**

Create a Google AdWords Account

To create a Google AdWords account, visit – www.adwords.google.com/. From there you'll create your account, and set up your first campaign. Here are the steps –

Step 1

Select your campaign type and name.

Step 2

Choose the geographic location where you'd like ads to show.

Step 3

Choose your "bid strategy," and set your daily budget. Change the default "Bid strategy" to "I'll manually set my bids for clicks". This gives you more control and will help you learn AdWords at a greater level of understanding.

Step 4

Create your first ad group, and write your first ad. More people click on ads when the headline includes the keyword they're searching on. So use your keywords in your headline when you can.

You're limited to 25 characters here, so for some search terms, you'll need to use abbreviations or shorter synonyms. Here's the short version of your ad template –

- Headline: Up to 25 characters of text
- 2nd line: Up to 35 characters
- 3rd line: Up to 35 characters
- 4th line: Your display URL

Step 5

Insert your keywords into the keyword field in your account. Paste in your keywords. Start with just one set, and add plus signs (+), brackets ([]), and quotes (“ ”) to see precisely how many searches of each type you'll get.

Step 6

Set your maximum cost-per-click. Set your maximum price-per-click (called your “default bid”). However, realize this: Every keyword is theoretically a different market, which means that each of your major keywords will need a bid price of its own. Google will let you set individual bids for each keyword later.

Step 7

Enter your billing information and Voila!

Campaign Management:

Digital marketing campaign management is an integral part of a marketing team's responsibilities, consisting of planning, implementing, and analyzing marketing campaigns.

The fast-changing nature of digital marketing makes it essential to review and gain insights from every digital marketing campaign and customer touchpoint. You can improve your digital marketing strategy by noting what works and what doesn't. Set clear goals and use digital marketing project management software to make marketing workflows and collaboration with external collaborators and creative agencies efficient.

What is a digital marketing campaign?

Brands launch digital marketing campaigns to achieve specific business and marketing goals such as promoting products, creating brand awareness, and acquiring new customers. A digital marketing campaign conveys a message to the public through various digital channels. A campaign is considered successful when it reaches the right audience and brings in new customers or converts leads.

The types of digital marketing campaigns available to marketers today are increasing and include text, photos, video, email, and experiential (e.g., VR) digital marketing campaigns. Hybrid digital marketing campaigns are a mix of all or some of the different types of digital marketing campaigns. The best digital marketing campaign strategy evokes positive emotions in customers.

What is digital marketing campaign management?

Digital marketing campaign management consists of a sequence of marketing-related activities planned and executed to achieve the best results for your business.

Digital marketing campaign plans break down the business marketing goals into smaller, actionable tasks. They aim to get their message to prospects and customers and build a trusted and engaged brand. The three main reasons for creating marketing campaigns are:

- Brand awareness
- Lead conversion
- Customer acquisition

Digital marketing campaign examples

It is easy to be overwhelmed when deciding on a suitable digital marketing campaign for your business needs. Different campaigns require different teams, resources, and operations to bring them to life. It's the marketing team's responsibility to decide on the campaign most likely to bring the best results. What worked two years ago may not work today.

Take video content marketing, for example. For many years, Youtube solely dominated the space with little competition from Vimeo and Facebook. Fast forward to 2020, when TikTok, a new social app focused on short-form video content, quickly gained global popularity during the coronavirus pandemic. Now, nimble marketing teams have TikTok for Business accounts and actively market to a growing audience on the platform, which boasts an estimated 1 billion monthly active users.

Consider your marketing goals and audience before going all-in on the different digital marketing campaign options available. Digital marketing campaign examples include:

Content marketing campaigns

Content marketing campaigns aim to educate the customers and share information about the company, product, or services. They provide value to the customers and answer questions about the product. A content marketing digital marketing campaign includes:

- Creating a content calendar
- Conducting keyword research
- Writing and designing content, e.g., blog posts, eBooks, newsletters, technical guides, white papers, and case studies
- Distributing content
- Measuring performance

Social media marketing campaigns

Most brands aspire to create social media campaigns that go viral and grab the attention of many potential customers. Social media marketing campaigns involve:

- Creating a social content strategy and schedule
- Publishing posts
- Measuring the response to know what works

To get better results, brands may choose to promote their campaigns with ad marketing spend.

Search engine optimization and marketing campaigns

Search engine marketing campaigns create content optimized or paid to get ranked higher on search engines. A critical part of the digital marketing campaign strategy is searching for keywords that align with your brand.

Search engine optimization and marketing campaigns include:

- Conducting keyword research
- Creating a blog content strategy and identifying content marketing topics

- Optimizing your website and content to rank for the right keywords
- Running paid search ad campaigns

How to set up a successful digital marketing campaign

Marketing teams must create repeatable systems, workflows, and processes to facilitate faster and more predictable marketing project success. Planning, creativity, and careful analysis of past campaigns are essential steps in setting up successful digital marketing campaigns.

Below, you will find tips for creating successful digital marketing campaign plans:

- **Know your audience:** Get to know your audience so you can create relatable and engaging content.
- **Find your platform:** Identify which channels are used by your target market and align with your brand.
- **Create content:** Study the format and type of content that matches your audience's digital habits. Create content consistently.
- **Set digital marketing KPIs:** Every campaign must have specific metrics to measure its success.
- **Monitor the digital landscape:** Observe your marketing environment and leave some room for flexibility and innovation in your digital marketing campaign strategy.
- **Analyze digital marketing campaign metrics:** Monitor critical KPIs to discover how your digital marketing campaigns are functioning.

PPC Advertising :

PPC or pay-per-click is a type of internet marketing which involves advertisers paying a fee each time one of their ads is clicked. Simply, you only pay for advertising if your ad is actually clicked on. It's essentially a method of 'buying' visits to your site, in addition to driving website visits organically.

One of the most popular forms of PPC is search engine advertising, which allows advertisers to pay for ad placement in a search engine's sponsored links. This works when someone searches for a keyword related to their business offering. For example, if we bid on the keyword 'Google Shopping Management' our ad might show up at the top of the Google results page.

Google Shopping, what we do best, is an example of how PPC advertising can be used to boost both visibility and profitability.

The benefits of using PPC

PPC has many advantages that will benefit your business, including the following:

- Optimised PPC is cost-effective
- Your ad will be displayed on the first results page when someone searches for a related term
- More exposure and brand visibility as targeted traffic is directed to your site
- You can run multiple ad campaigns for each keyword
- It results in higher click-through rates (CTR), more conversions, and increased sales revenue
- Provides instant traffic

If PPC is working as it should, the return on ad spend (ROAS) should be high, as a visit to your site is worth more than what you pay for it. However, it's not as simple as just paying for the clicks and gaining traffic, a lot goes into putting together a strong PPC campaign. It consists of choosing the right keywords, arranging those keywords into well-organised campaigns and ad groups, and setting up PPC landing pages that are optimised for conversions.

It's important to learn the best ways to conduct a PPC campaign, as the more relevant, well-targeted PPC campaigns will be rewarded by search engines with lower-costing ad clicks. Google will reduce your cost per click if your ads are satisfying and useful to users, ultimately, earning you a higher profit.

Who should use PPC?

PPC advertising is ideal for small businesses as you get to decide how much you're willing to spend on each keyword, so you can manage how much you will be spending when your ads are displayed. In essence, this ensures that the money you spend on ads isn't wasted, as you're more likely to attract more people who are interested in your product.

Google Ads

Google Ads is the most commonly used PPC advertising network. The platform enables you to create campaigns that will appear on all Google services. Choosing the right ad format and keywords is important, as Google Ads ranks potential advertisers based on their ad quality, relevancy and size, as well as the price of the bid.

Since Google is the most used search engine, using Google Ads will get you the most impressions, so take into consideration the following factors when creating your PPC campaign through Google Ads:

- Ensure your keywords are relevant, popular, and likely to be searched
- Have a high-quality landing page that looks appealing and has a clear message
- Better quality scores get more ad clicks at lower costs

In summary, PPC advertising is a great marketing option for your business. It's a simple yet effective digital marketing technique to promote your products, drive traffic to your website, and ultimately, increase your sales.

Google Shopping

Google Shopping is another excellent example of a PPC channel. It runs through Google Ads and can be used by retailers to get their products in front of interested shoppers. Since Google is the most used search engine, having your products shown as part of the search engine results page (SERP) can hugely increase visibility, impressions, and help drive traffic to your website.

On Google Shopping, advertisers place bids to secure the best possible spot on the Shopping carousel, with the first spot being the most highly sought-after. The reason is that 65% of shoppers click the first ad on the Google Shopping carousel regardless of price, perhaps due to beliefs it's from the highest quality retailer or sold at the best price available.

Google Shopping management can be a complex beast to tame. From different automation options to manual management or third-party agencies, retailers have different methods of making the channel profitable.

Everything you need to know about Google Shopping

Bidnamic's technology platform works 24/7 to ensure that our clients pay the most profitable bid price for the most visibility on the Shopping carousel. For our clients that have a lot of product SKUs in their catalogue, Bidnamic's automated solution takes the time-consuming task of manual bidding and gives that time back to our clients to use in other areas of their business.

Affiliate Marketing:

Affiliate marketing is a process in which an affiliate will promote a business's products and services and receives a commission if the affiliate achieves a sale.

An affiliate is a person that utilizes its networks and online marketing capabilities to promote products and services.

Affiliate marketing and the use of affiliate marketers is a sales and performance-based advertising method that offers many benefits to both the business and the marketer.

When a product is excellent, and the marketers' online marketing skills match, both will achieve higher sales and brand awareness.

Affiliate marketing is a beneficial method for any business that wants to grow. It's sales without spending on a traditional advertising budget.

Growing your affiliate marketing strategy is a long process that involves individuals and giving them the right tools and incentives to promote your brand and products.

Most commonly, companies and marketers use affiliate marketing in the online marketing world. It can also include the contribution of offline and physical sales if that's what the company desires.

Business spending on affiliate marketing is growing fast, and spending is estimated to grow up to \$8.2 Billion in 2022, a growth of ~52% from 2017. If your business isn't on par with the growth of affiliate marketing, you might be losing on valuable online real estate that is only achievable through the use of affiliate marketers.

In this article, we will cover the benefits of affiliate marketing benefits for businesses and also how affiliate partnerships help every party involved in the process.

The combined benefits to both parties make affiliate marketing an essential part of any digital marketing strategy. But first, let's look at how affiliate marketing works.

How Does Affiliate Marketing Work In A Nutshell?

At its core affiliate marketing is simple, and here it is in six simple steps:

1. The affiliate and the business get into a relationship either through a third-party (an affiliate network) or directly.
2. An Affiliate shares the product and service to their network of choice (typically a link or a coupon)
3. Potential customer engages with the link.
4. Customer lands on the businesses landing page
5. The customer makes a purchase.
6. The business gets the sales, and the affiliate marketer receives a commission.

Affiliate marketing, therefore, consists of three (four if you're utilizing a network) stakeholders: The marketer (also called a publisher), the product provider (or the one who distributes the products and the services), and the customer.

The relationship between the marketer and product producer depends on the nature of the deals in place between them. A customer will purchase from a business, and the customer's first interaction will be with the marketer.

A business can streamline its affiliate process and won't need to manage the process themselves, using an affiliate network, although increased costs are something to consider.

A perfect affiliate marketing campaign is beneficial for every stakeholder. The business and the marketer gains revenue, and customers get more personal recommendations from a trusted source.

Affiliate Marketing Benefits For Businesses

1. A New Profitable Sales Channel
2. Benefits A Variety Of Products
3. Cost-Effective
4. Increase Competitive Advantage
5. New Audiences Through Influencers
6. Affiliate Marketers And Access To Their Channels
7. Performance-Based Advertising
8. Knowledge From Affiliates
9. Analytics and Campaign Performance
10. Easy To Start And Manage

A New Profitable Sales Channel

For a business, affiliate marketing offers a much-needed scale without accessing large advertising budgets. Even one successful affiliate can bring much more traffic, leads, and sales than a traditional marketing campaign or even a highly effective online advertising campaign.

When comparing the benefits of advertising with publishers, or allowing the publisher to become an affiliate, becoming an affiliate will create better and long-lasting relationships that benefit both parties.

When done correctly, affiliate marketing might become your most important and influential sales channel that can grow sales and brand awareness over time.

Unlike a sales team, affiliate marketing works for you 24/7, making money for both parties at all times. A passive sales funnel for a business is crucial if the company wants to expand and scale faster.

If an affiliate campaign has been built the proper way with great products and services with enticing opportunities for marketers, a business can enjoy the benefits of passive income and promotion.

By its core as affiliate marketing is result-based, and the cost for upkeep will be lower than managing your sales teams. In most cases, not having a sales team wouldn't be entirely realistic, but it enables the specialization of your sales teams.

For example, affiliates would bring cold traffic or leads that your sales team can nurture further into customers. In a model like this, the compensation model might need readjustment. If your sales team is the one making the sales, then affiliates will require some other form of result-based compensation, like commission per lead.

The convenience factor for a business is the fact that you can diversify your revenue and promotional sources.

Benefits A Variety Of Products

Products that are perfect for affiliate marketing has a couple of the following characteristics:

- High end or high margin
- Scalable
- Online (or online distribution)
- Fits a niche
- Growing trend

- Social traction
- Shareability for virality

Amazon has shown that low margin products are still a viable option for affiliate marketing, but small percentages for affiliates might affect more extensive opportunities. But as customers would trust the supplying company and the affiliate, it's a combination hard to miss.

Eventually, low commissions will drive the best affiliates into other venues, which means when optimizing commissions, you need to think ahead, what else is in the marketplace in your industry, how much they pay in commissions, and on which terms.

A product that has a considerable commission potential will drive more affiliates to you. Still, if the product also matches the other criteria of the perfect product, you're more likely to drive more affiliates into your products and services and to your program.

If your company produces a range of products and services, it might be worthwhile to pick the best suitable one for affiliate marketing rather than a general link for everything. The benefit of the methods is that it's easier to build more effective campaigns when you have a clear product for promotion.

General links can work (Like Amazon), though, but might require additional effort to create better marketing resources or affiliates need to create a more specific story around a company.

In general, a large variety of products is viable for affiliate marketing, which means most businesses can utilize it to grow their brand awareness and sales.

Sometimes it's better to find your influencers first and work together with an affiliate deal with a product that matches the most to their audiences.

Access To Influencers And Their Networks

Affiliate marketing allows a business to target particular niches that match with a business's core audience.

Every industry and customer segment has its influencers that can connect with your target audiences in a way that most commonly is almost impossible for a company selling products.

While not impossible to gain a natural following as a product company, not reaching out and utilizing influencers, you might be losing out on valuable customer data that is essential for scaling your business effectively.

In most cases, the audiences of influencers might bring new data that can widen your audience related opportunities.

For example, if we look at an online clothing company, through testing and targeting online advertising found a perfect audience for their products. Then the company would reach out to an influencer whose audience matches the description, only to find out that the influencer's audience is broader than initially thought. And if the audience performs better than the average, the data leads to better decisions about data-driven audiences.

Another benefit is the discovery of entirely new audiences. In some industries, their target audiences might be harder to reach through traditional methods. In this case, offering an influencer an affiliate deal could potentially help you achieve this new audience.

Mobile and SMS Marketing

What is SMS marketing?

SMS (short message service) marketing is the use of a mobile phone to send an SMS or text message that contains promotional content, such as special deals, product updates, customer loyalty perks, or any other marketing information directly to customers.

How it works

An SMS campaign operates on a permission-based model, meaning customers opt in to receive marketing offers from you via SMS message. By texting a designated keyword to a shortcode, which is usually a five-digit number, customers give your company permission to add them to your SMS campaign and database. You could also get consumer approval to collect their number when you're on the phone with them, through a sign-up form on your website, or at your physical store's register.

What is SMS marketing used for?

SMS and text marketing are generally used to issue reminders, delivery notices, limited-time or exclusive offers, and redeemable coupons to your consumer base. SMS marketing is also an effective method for driving customer engagement through polls or survey-type messages.

What SMS marketing tools are available?

These are some of the many SMS marketing tools available:

- [Twilio](#), which includes options for SMS messaging and Facebook Messenger, plus tools for traditional email marketing campaigns

- Textedly, which includes easy shortcode creation, list opt-in tools and text message scheduling
- EZ Texting, which includes keyword options and text-to-landline capabilities
- Mobiniti, which includes free unlimited support and integrates with many email and content marketing platforms
- Avochato, which includes contact segmentation tools and integrates with many internal database and communication platforms.

Advantages of SMS marketing

Since text messages are generally opened within five minutes upon receipt, SMS is an instant way of marketing to your audience and building customer loyalty. Additionally, some SMS marketing software options allow you to personalize text messages for recipients of your messages. When you choose an SMS marketing tool that allows you to insert a customer's first name in your message, you're more likely to capture their interest and attention – in addition to generating a good response and conversion rate.

Another benefit of text message marketing is its automated response function. Depending on the action your subscribers take, you can send messages automatically to their phones that would trigger a series of follow-up replies or actions from your SMS marketing software to engage customers further.

Last but not least, your SMS marketing tool will usually give you options for easy segmentation and management of your contact database. This makes SMS marketing a cost-effective business strategy that helps your team hit your SMS metrics.

SMS marketing tips

- Be clear in communicating what your SMS marketing campaign is all about, including what types of SMS messages your customers will receive, how often they will receive them, and how they will benefit when they opt in.
- Give subscribers the chance to opt out of your notifications anytime by including relevant instructions in your text campaigns from time to time.
- Don't text during irregular hours, and don't use shorthand texts – aim for a professional tone instead.
- Add a disclaimer in your initial text message about possibly incurring data and message charges if customers don't have unlimited plans for their subscription.

Mobile marketing

Mobile marketing is the practice of marketing your business on the mobile device platform, smartphones and tablets included, although current trends extend the concept of mobile marketing to other gadgets, such as smartwatches.

It's a multichannel digital marketing strategy that utilizes content marketing tools such as websites, social media, mobile apps and SMS/MMS (multimedia messaging service) to send mobile ads to audiences.

The key with mobile marketing is to use smartphones and other handheld devices to provide potential or existing customers personalized and time- and location-based information about your business.

Types of mobile marketing

Mobile marketing is a well-rounded technique in digital marketing because it gives you several options on how to format your ads. These are some of the most popular strategies for marketing on mobile:

- **Video ads:** As digital technologies make watching videos possible on mobile devices, you could create video ads to reach your target market. Contrary to finding them intrusive (as some people may consider email marketing campaigns to be), many consumers consider video ads entertaining and engaging. Your video ads can be part of your MMS messages that you send from your mobile messaging service provider or website to your mobile audience.
- **Mobile apps:** Mobile consumers spend 82% of their device time on apps, which makes app-based marketing or in-app advertising a popular strategy among marketers. Facebook is an example of a third-party mobile app that allows you to integrate your ads into the social media platform through sponsored posts.
- **In-game mobile marketing:** Mobile games can also host your mobile ads, which can be in the form of banner popups, full-page image ads or video ads that load in between game times.
- **Location-based marketing:** This allows you to send ads on users' mobile devices when they are within a specific distance from your business or location. Fast food or other specialty stores use this strategy to promote brand awareness among consumers.
- **Mobile search ads:** These are ads that customers see when they do a basic search on Google. These ads often have extensions like click-to-call or maps to help users find your business.

Mobile marketing tips

- Optimize your ad's message, format and design for mobile. With the small screen that you have to work with on mobile devices, every space and word should be used wisely.
- Your business's homepage must be responsive to mobile SEO so users can find your local business at the time of their search.
- Know your audience. You'd do better, for example, to use in-game ads to interact with gamers than sponsored posts on social networks like Twitter, Instagram or Facebook.

As users use mobile devices more and more, it's safe to say that the future of digital marketing is shifting more toward mobile, specifically SMS marketing and mobile marketing.

Mobile marketing tools

These are some of the most commonly used mobile marketing tools:

- [Google Analytics](#), which includes relevant metrics for mobile marketing campaigns such as traffic reports, keyword referrals, event tracking and attribution
- [AppsFlyer](#), which monitors app engagement, tracks app installation locations, and tracks and improves mobile marketing campaigns
- [App Annie](#), which can rank apps and keywords, estimate app usage and advertising, and offer SDK insights
- [Flurry](#), which provides tools for funnel analysis, user segmentation and event tracking

- **Braze**, which offers live customer views, journey building and real-time mobile marketing campaign optimization

When considering SMS versus mobile marketing, keep in mind that SMS marketing can help you build a more direct and personal relationship with your audience, while mobile marketing can help you improve customers' interaction with your brand. Either way, it's a must to incorporate these two strategies into your overall marketing plan.

Marketing Automation:

With consumers being bombarded with marketing messages from every direction, making a marketing campaign stand out has become more difficult than ever before. Marketing teams are under increasing pressure to come up with more innovative ideas that are bigger and better than anything their competitors are doing. But doing this, on top of day to day marketing activities, is not easy.

Marketing teams need more time to focus on the bigger picture and therein lies one of the major challenges being faced by many businesses across the UK. But technology, in the form of marketing automation software, could be the answer.

Marketing automation explained

Marketing automation helps you to identify potential customers, automating the process of nurturing those leads to sales-readiness. It automates actions that bring prospects to the point where they can be directly approached by the sales team with the aim of closing a sale and starting an ongoing relationship, and the information it gathers can drive your choice of marketing tactics.

Marketing automation does this by massively improving the efficiency of your sales funnel, quickly turning a broad base of leads into happy customers using a combination of tactics.

For example, early in the lead-nurturing process education and awareness might be the key task. Marketing automation can supply useful content that develops trust in and respect for your brand, quickly and easily helping leads to understand what it is they're getting. Further along, when prospects have narrowed down the types of products they're interested in, you can reach out with targeted messaging, specifically tailored to the groups that could best help to grow your brand. Finally, as activity tracked via the marketing automation system indicates yet more focused interest, a qualified, comprehensive and well-understood lead is automatically handed to the sales team.

Why is marketing automation valuable?

Lead generation is an extremely important step in any business's growth. By automating many steps in the process from marketing to sales, your team has more time to focus on overall strategy and nurturing the leads that show real promise. That means more prospects, and more customers.

Marketing automation is also able to give you a richer, more detailed picture of the behaviour of potential customers. Using behavioural tracking methods such as following a user's path through your website, marketing automation software can help your marketing team to understand a prospect's interests and where they are in the purchasing lifecycle. They can then customise any follow-up based around those action points.

For instance, let's say a particular customer is reading about a broad category of products. This might indicate that they are at the beginning of the purchasing process, researching and comparing while preparing a shortlist. If they later download white papers on a specific product, that could indicate a narrower focus and a readiness to be contacted by a salesperson.

Bringing together information from touchpoints including website visits and downloads, social media activity and direct marketing enables automatic scoring, qualifying and prioritising of leads. This in turn can then drive wider marketing campaigns, including:

trigger-based marketing messages

infrequent "drip-feed" emails to maintain interest

personalised emails

Facebook or Twitter messages

Prioritise leads and improve marketing ROI

Besides automating the lead-nurturing process, marketing automation software can also allow you to establish clear, objective measures of progress through the customer lifecycle. Whereas in the past a particular prospect's readiness for sales approaches might be down to an individual marketer's intuition, it can now be based on pre-defined and measured outcomes.

With such analytics at its heart, marketing automation can help prioritise sales staff time, measure the effectiveness of touch points with the prospect and establish overall campaign effectiveness.

Guesswork is also removed from the process: in particular when deciding the point at which an individual prospect is handed over from marketing to sales. Indeed, closed-loop reporting allows you to calculate accurately your cost per opportunity and the ROI of your marketing efforts.

Learn more about marketing automation from Salesforce:

B2B Marketing automation

The business-to-business arena has been the traditional home of marketing automation. That's because B2B prospects and customers form a small, focused target market, which is usually:

engaged in a multi-stage procurement process

part of an ongoing relationship of repeat business

For these reasons, B2B is a relationship-driven environment where product education and awareness building are vital. It's also an arena where purchases are not made on the spur of the moment by individuals; instead, they are considered and rational, and usually involve more than one person.

The buying decision in such circumstances will also take some time, with weeks, months or longer elapsing between the start of the process and its conclusion. This means that in general, business procurement necessitates a longer process of lead nurturing than B2C marketing, and in that process the prospect will have done a lot of research before they are ready to speak to a salesperson – let alone ready to buy.

These characteristics – a long marketing-to-sales evolution, combined with customer research generating multiple data points that can be analysed – makes marketing automation software supremely well-suited to the B2B environment.

Web analytics

Web analytics involves collecting, measuring, and analyzing website data. Web Analytics tools can provide lots of useful information about the origin of website traffic, how website users navigate and interact throughout a website, what content and web pages they're most engaged with, and how they exit the site. Marketers can then use this data to optimize the performance of their channels and websites by taking data-led decisions. These are actions we take to improve performance based on our understanding of the data we've recorded; in other words, the data is leading the decision. While it's best practice to follow data-led decisions, we must always remind ourselves to make sure we test and verify the data.

Web analytics tools

Web analytics tools provide information about the origin of website traffic, how users navigate and interact throughout a website, what content and webpages they're most engaged with, and if they take valuable actions on the site, known as conversions, these include purchases or contact requests. Using this data, marketers can optimize channel and website performance with data-led decisions.

A number of analytics tools are available on the market today. Some of them are free, and some of them require a paid subscription. Paid analytics products and free analytics products will differ in terms of support, features, and functionality.

Some examples of available analytics packages include:

- Google Analytics and Google Analytics 360, which are part of the Google Marketing Platform
- Adobe Analytics
- Woopra
- Kissmetrics
- Webtrends
- Piwik

The market leader and most commonly used analytics program is the free version of Google Analytics, or GA as it is sometimes known. The paid version of GA, called Google Analytics 360, offers some additional functionality in terms of report validity and sample sizes. But it's mainly for websites that receive more than 10 million page views per month. For the vast majority of websites, the free version of Google Analytics is perfectly fine.

Because of its functionality and widespread adoption in the market, Google Analytics is seen by many marketers as the single 'source of truth' for website traffic, engagement, and conversion data.

Advantages of web analytics

Web analytics is a valuable way to deduce the 'story' behind the data, in order to gain valuable insights and enhance business performance. Web analytics can help a digital marketer understand their customers better by providing:

- Insight into who the customers are and their interests

- Conversion challenges
- Enhanced appreciation of what consumers like or don't like
- Understanding of how to improve user experience for the consumer

One of the real values of web analytics is that it allows you to deduce the 'story' behind the data in order to gain valuable insights and enhance business performance. But how exactly does this lead to commercial returns? Web Analytics can help you understand your customers better. It tells you who they are, where they're coming from, and what their interests are. It tells you about their demographics and location. It also helps reveal any conversion challenges that might exist on your website. It helps you grasp what content and products your consumers like or don't like, and how they interact on your website. You can use all this information to improve the consumer experience on your site and to optimize the channels that consumers use to visit your website.

Growth Hacking:

What is a Growth Hacker?

A growth hacker is someone who uses creative, low-cost strategies to help businesses acquire and retain customers. Sometimes growth hackers are also called growth marketers, but growth hackers are not simply marketers. Anyone involved in a product or service, including product managers and engineers, can be a growth hacker.

Growth hackers tend to be obsessive, curious and analytical:

- Growth hackers focus solely on strategies related to growing the business.
- They hypothesize, prioritize and test innovative growth strategies.
- They analyze and test to see what's working.

The ideal growth hacker knows how to set growth priorities, identify channels for customer acquisition, measure success, and scale growth.

How Growth Hacking Works

So, how does growth hacking work? For each company, it's about figuring out why you grow, and looking for ways to make that happen on purpose.

Many startups use Dave McClure's "pirate funnel" as a recipe for growth. These are acquisition, activation, retention, referral, and revenue (AARRR). Others include raising awareness as a key part of growth hacking. Either way, the point is to get traffic and visitors, turn visitors into users, and retain those users as happy customers.

How to Start Growth Hacking

Here's how a company can get started with growth hacking. First of all, create your product and test to make sure people want it, and are willing to pay for it. This will help you gather data so you understand your key buyer personas and can target growth marketing tactics accordingly.

Update your product at regular intervals, and keep getting customer feedback so you always know if you're on the right track. At the same time, market your product to foster continued growth, and track the success of those results. A/B testing and other conversion optimization techniques are crucial for effective growth hacking.

Growth Hacking Strategies

Most growth hacking strategies fall into three main areas:

- Content marketing
- Product Marketing
- Advertising

Depending on the tactics used, content marketing can be a low-cost way to get the word out about your product. Typical content marketing activities include:

- Starting a blog and creating valuable, shareable content
- Guest blogging
- Creating social media content
- Writing ebooks and white papers
- Podcasting
- Running webinars
- Running contests and giveaways
- Getting bloggers to review your product
- Joining relevant forums, groups and subreddits
- Influencer marketing
- Using email marketing to build a stronger connection with users
- Improving content visibility with SEO
- Getting listed in relevant marketplaces and sites, such as Product Hunt

Product marketing includes techniques for making your product more appealing, and building the user base. They include:

- Leveraging the fear of missing out (FOMO) by using an invite-only signup system
- Gamifying the user onboarding process to make it more enjoyable, and offering rewards
- Offering incentives for referrals that benefit both the referrer and the new user

- Affiliate marketing, which will also use content marketing growth tactics

Growth hackers can also use social advertising and pay per click (PPC) advertising to promote their business.

Some well-known examples of successful growth hacking campaigns include:

Growth Hacking Examples

- Dropbox, which rewards existing users for inviting new ones with additional storage
- Hotmail, which appended a line to each outgoing email encouraging people to sign up for a new account
- AirBnB, which used Craigslist to find and market to people looking for affordable accommodation

COURSE OBJECTIVES:

- To enable the students to understand the Organizational Behavior
- To analyse various factors affecting Personality Organizational Change
- dynamic of groups
- To Understand various type of Group Behavior

UNIT I ORGANIZATIONAL BEHAVIOR INTRODUCTION 9

Organization Behaviour – Definition – Scope and Application in Management – Contributions of Other Disciplines to OB. Emerging Issues in Organizational Behaviour- Organizational behaviour models

UNIT II INDIVIDUAL PROCESSES 9

Personality – types – Factors influencing personality– Theories. Emotions - Theories – Emotional Intelligence- Learning – Types of learners – The learning process – Learning theories. Perceptions – Importance – Factors influencing perception- Attitudes – Nature of Attitudes Components of Attitudes Formation of Attitude Benefits of Positive Attitude Functions of Attitudes– Measurement-Motivation – Importance – Types – Theories.

UNIT III LEADERSHIP AND POWER 9

Meaning – Importance – Leadership styles – Theories – Leaders Vs Managers – Sources of power – Power centers – Power and Politics.

UNIT IV GROUP DYNAMICS 9

Meaning – Types of Groups – Functions of Small Groups – Group Size Status – Managerial Implications – Group Behaviour – Group Norms – Cohesiveness – Group Thinking

UNIT V ORGANIZATIONAL CHANGE AND DEVELOPMENT 9

Organizational Change: Meaning – Nature of Work Change – Need for Change – Change Process – Types of Change – Factors Influencing Change – Resistance to Change – Overcoming Resistance – Organizational Development: Meaning and Different Types of OD Interventions

SUGGESTED ACTIVITIES:

1. To analyze and understand the impact of various functional modules on the behaviour of individuals with real time examples like buying behavior of consumers in supermarkets.
2. To Analyze and understand the Perception of individuals and performance based on situations like an individual's effectiveness in the workplace(often depends on their personality, attitudes and values along with their motivation) to succeed.
3. Conduct a group discussion among 10 members on some topic and write a report on analysis of behaviour of team members in group decision making
4. Justify the selection of team members for executing a project with the analysis of various factors like domain expertise ,communication skill of members etc
5. To study the Performance of employees on organizational change with respect to environment

TOTAL: 45 PERIODS

COURSE OUTCOMES:

On completion of the course should be able to:

CO1:Students will have a better understanding of human behavior in organization.

CO2:They will know the framework for managing individual and group performance.

CO3:Characteristics of attitudes and components of attitudes — A brief discussion

CO4:List the determinants of personality

CO5:List the characteristics of various leadership styles.

REFERENCES

1. K. Aswathappa, "Organisational behaviour", Himalaya Publishing House Pvt. Ltd. 11th Edition.
2. Stephen P. Robbins, "Organizational Behavior", PHI Learning / Pearson Education, Edition 17, 2016 (Global edition)
3. Fred Luthans, "Organizational Behavior", McGraw Hill, 12th Edition
4. Nelson, Quick, Khandelwal. "ORGB – An innovative approach to learning and teaching". Cengage, 2nd edition 2012
5. Ivancevich, Konopaske Matteson, "Organizational Behaviour & Management", Tata McGraw Hill, 7th edition, 2008

MC4023

WEB DESIGN

L T P C
3 0 2 4

COURSE OBJECTIVES:

- To understand the concepts and architecture of the World Wide Web.
- To understand and practice markup languages
- To understand and practice embedded dynamic scripting on client-side Internet Programming
- To understand and practice web development techniques on client-side

UNIT I INTRODUCTION TO WWW

9+6

Understanding the working of Internet-Web Application Architecture-Brief history of Internet-Web Standards – W3C-Technologies involved in Web development – Protocols-Basic Principles involved in developing a website-Five Golden Rules of Web Designing

UNIT II UI DESIGN

9+6

SVG- Iframes - HTML5 Video and Audio tags - CSS Specificity - Box model - Margins, padding and border – Inline and block elements - Structuring pages using Semantic Tags - Positioning with CSS: Positions, Floats, z-index – CSS with CSS Preprocessors: SASS

UNIT III ADVANCED UI WITH CSS3

9+6

Layouts with CSS Grids Flexbox– Responsive web design with media queries - Advanced CSS Effects – Gradients, opacity, box-shadow - CSS3 Animations: Transforms and Transitions - CSS Frameworks: Bootstrap

UNIT IV JAVA SCRIPT

9+6

JavaScript Events - Modifying CSS of elements using JavaScript- Javascript Classes- Introduction to JQuery – JQuery Selectors - Using JQuery to add interactivity - JQuery Events-Modifying CSS

UNIT III LEADERSHIP AND POWER

Meaning – Importance – Leadership styles – Theories – Leaders Vs Managers – Sources of power – Power centers – Power and Politics.

Definitions And Meaning Of Leadership

According to Alford and Beatty "Leadership is the ability to secure desirable actions from a group of followers voluntarily, without the use of coercion".

According to Chester I Barnard, "It (leadership) refers to the quality of the behaviour of the individual whereby they guide people on their activities in organized efforts"

According to Koontz and O'Donnell - Managerial leadership is "the ability to exert interpersonal influence by means of communication, towards the achievement of a goal.

Since managers get things done through people, their success depends, to a considerable extent upon their ability to provide leadership".

In the words of Theo Haimann - "Leadership is the process by which an executive imaginatively directs, guides and influences the work of others in choosing and attaining specified goals by mediating between the individuals and the organization in such a manner that both will obtain maximum satisfaction".

Nature / Characteristic / Features Of Leadership

1. Leadership is the process of influencing the activities of an individual or a group towards the achievement of a goal.
2. An effective leader motivates the subordinates for higher level of performance.
3. Leadership promotes team - spirit and team - work which is quite essential for the success of any organization.
4. Leadership is an aid to authority. A leadership helps in the effective use of formal authority.
5. Leadership creates confidence in the subordinates by giving them proper guidance and advice.
6. Leadership involves an unequal distribution of authority among leaders and group members: Leaders can direct some of the activities of group members, i.e., the group members are compelled or are willing to obey most of the leader's directions.
7. Leadership is a process of Influence: Leadership implies that leaders can influence their followers or subordinates in addition to being able to give their followers or subordinates legitimate directions.
8. Leadership is the function of stimulation: Leadership is the function of motivating people to strive willingly to attain organizational objectives. A successful leader allows his subordinates (followers) to have their individual goals set up by themselves in such a way that they do not conflict with the organizational objectives.
9. A leader must be exemplary. "A Leader shows the way by his own example. He is not a pusher, he pulls rather than pushes".
From the above explanation it is clear that a leader must set an ideal before his followers. He must stimulate his followers for hard and sincere work by his personal behaviour.
10. A Leader ensures absolute justice: A leader must be objective and impartial. He should not follow unfair practices like favouritism and nepotism. He must show fair play and absolute justice in all his decisions and actions.

Leadership Skill

The leader is expected to play many roles and therefore, must be qualified to guide others to organizational achievement.

In a broad way the skills which are necessary for an industrial leader may be summarized under four heads:-

- (a) Human skill
- (b) Conceptual skill
- (c) Technical skill and
- (d) Personal skill.

a) Human Skill :

A good leader is considerate towards his followers because his success largely depends on the co-operation of his followers. He approaches various problems in terms of people involved more than in terms of technical aspects involved. A leader should have an understanding of human behaviour. He should know people; know their needs, sentiments, emotions, as also their actions and reactions to particular decisions, their motivations etc.

Thus, a successful leader possesses the human relations attitude.

The human skill involves the following:-

- (a) Empathy: A leader should be able to look at things objectively. He should respect the rights, belief and sentiments of others. He should equip himself to meet the challenges emanating from the actions and reactions of other people. The leader should be empathetic towards his followers so that he can carefully judge their strengths, weakness, and ambitions and give them the attention they deserve.
- (b) Objectivity: A good leader is fair and objective in dealing with subordinates. He must be free from bias and prejudice while becoming emotionally involved with the followers. His approach to any issue or problem should be objective and not based on any pressure, prejudice or preconceived notions. Objectivity is a vital aspect of analytical decision making. Honesty, fairplay, justice and integrity of character are expected of any good leader.
- (c) Communication Skill: A leader should have the ability to persuade, to inform, stimulate, direct and convince his subordinates. To achieve this, a leader should have good communication skill. Good communications seem to find all responsibilities easier to perform because they relate to others more easily and can better utilize the available resources.
- (d) Teaching Skill: A leader should have the ability to demonstrate how to accomplish a particular task.
- (e) Social Skill: A leader should understand his followers. He should be helpful, sympathetic and friendly. He should have the ability to win his followers confidence and loyalty.

b) Conceptual Skill

In the words of Chester Barnard -"the essential aspect of the executive process is the sensing of the organization as a whole and the total situation relevant to it". Conceptual skills include -

- (a) The understanding of the organization behaviour
- (b) Understanding the competitors of the firm, and
- (c) Knowing the financial status of the firm.

c) Technical Skill

A leader should have a thorough knowledge of, and competence in, the principles, procedures and operations of a job. Technical skill involves specialized knowledge, analytical skill and a facility in the use of the tools and techniques of a specific discipline. Technical competence is an essential quality of leadership.

d) Personal Skill

The most important task of the leader is to get the best from others. This is possible only if he possesses certain qualities. These personal skills include

(a) Intelligence: Intellectual capacity is an essential quality of leadership. Leaders generally have somewhat higher level of intelligence than the average of their followers.

(b) Emotional Maturity: A leader should act with self-coincidence, avoid anger, take decisions on a rational basis and think clearly and maturely. A leader should also have high frustration tolerance. According to Koontz and O'Donnell - "Leaders cannot afford to become panicky, unsure of themselves in the face of conflicting forces, doubtful of their principles when challenged, or amenable to influence".

(c) Personal Motivation: This involves the creation of enthusiasm within the leader himself to get a job done. It is only through enthusiasm that one can achieve what one wants. Leaders have relatively intense achievement type motivational drive.

(d) Integrity: In the words of F.W Taylor - "integrity is the straight forward honesty of purpose which makes a man truthful, not only to others but to himself; which makes a man high-minded, and gives him high aspirations and high ideals".

(e) Flexibility of Mind: A leader must be prepared to accommodate other's viewpoints and modify his decisions, if need be. A leader should have a flexible mind, so that he may change in obedience to the change in circumstances. Thomas Carle has said - "A foolish consistency is the hobgoblin of a little mind".

FUNCTIONS OF A LEADER/ LEADERSHIP

1. To take the initiative: A leader initiates all the measures that are necessary for the purpose of ensuring the health and progress of the undertaking in a competitive economy. He should not expect others to guide or direct him. He should lay down the aims and objectives, commence their implementation and see that the goals are achieved according the predetermined targets.

2. He identifies group goals: A leader must always help the group identify and attain their goals. Thus, a leader is a goal setter.

3. He represents the organization: A leader represents the organization and its purpose, ideals, philosophy and problems to those working for it and to the outside world .In other words, leaders is true representative of the entire organization.

4. He acts as a arbitrator: When groups experience internal difference, whether based on emotional or intellectual clashes, a leader can often resolve the differences.

He acts as an arbitrator to prevent serious group difference.

5. To assign reasons for his action: It is a delicate task of leaders to assigns reason to his every command. He has to instruct things in such a way that they are intelligible to all concerned and their co-operation is readily forthcoming.

6. To interpret: He interprets the objectives of the organization and the means to be followed to achieve them; he appraises his followers, convinces them, and creates confidence among them.

7. To guide and direct: It is the primary function of the leader to guide and direct the organization. He should issue the necessary instructions and see that they are properly communicated.

8. To encourage team work: A leader must try to win the confidence of his subordinates. He must act like the capital of a team.

9. He manages the organization: Last, but not the least, he administers the undertaking by arranging for the forecast, planning, organization, direction, co-ordination and control of its activities.

TYPES OF LEADERS

1. Autocratic / Task management leadership

They are the authoritarian leaders. They are the leaders by authority. The authoritarian leader directs his subordinates to perform the requisite task in accordance with the dictates given to them.

The autocratic Leader gives order which he insists shall be obeyed. He determines policies for the group without consulting them, and does not give detailed information about future plans, but simply tells the group what steps must they take. In other words, an autocratic leader is one who centralizes the authority in himself and does not delegate authority to his subordinates. He is dictatorial by nature, and has no regard for the subordinates. He drives himself and his subordinates with one thought uppermost in his mind- action must produce results. An autocratic close the entire planning and cells upon his subordinates to execute what he has planned. An Autocratic leader operates on the following assumptions:-

- (a) An average human being has inherent dislikes of work and will avoid it if he can.
- (b) His assumption is that if his subordinate was intelligent enough, he would not be in that subordinate position.
- (c) He assumes that unintelligent subordinates are immature, unreliable and irresponsible persons. Therefore, they should be constantly watched in the course of their work.
- (d) As he has no regard for his subordinates, he gets the work done by his subordinates through negative motivation i.e. through threats of penalty and punishment.

Types of autocratic leadership

Strict autocratic leaders: A strict autocratic relies on negative influence and gives orders which the subordinates must accept. He may also use his power to disperse rewards to his group.

Benevolent Autocrat: The benevolent is effected in getting high productivity in many situations and he can develop effective human relationship. His motivational style is usually positive.

Manipulative Autocrat: A manipulative autocratic leader is one who makes the subordinates feel that they are participating in decision making process even though he has already taken the decision.

2 Participative / Democratic leaders

A democratic leader is one who consults and invites his subordinates to participate in the decision making process. He gives orders only after consulting the group; sees to it that policies are worked out in group decisions and with the acceptance of group. The manager largely avoids the use of power to get a job done. He behaves that a desired organizational behaviour can be obtained if employees' needs and wants are satisfied. Therefore, he not only issues orders but interprets them and sees to it that the employees have the necessary skill and tool to carry out their assignments. He assigns a fair work load to his personal and recognizes the job that is well done; there is a team approach to the attainment of organizational goals. He recognizes human value for greater concern for his subordinates. A participative leader operates on the following assumptions:-

- (a) Subordinates are capable of doing work and assuming the responsibility if they are given opportunities and incentives.
- (b) Subordinates are supervised, guided and aided rather than threatened and commanded to work.
- (c) Mistakes are not viewed seriously. The assumption is that disciplinary action breeds discontent and frustration among employees and creates an unhealthy work environment.

3. Laissez Faire or Free-rein Leadership

A free-rein leader does not lead, but leaves the group entirely to itself. The leader avoids using power and interest the decision making authority to his subordinates. He does not direct his subordinates and there is complete freedom for the subordinates. Group of members work themselves and provide their own motivation. The manager exits as a contact man with outsiders to bring for his group the information and resources it needs to accomplish its job. A free-rain leadership operates on the following assumption:-

- (a) He follows the rule of minimum exposure to accountability.
- (b) He relieves himself of responsibilities and is ready to blame his subordinates if something goes wrong.
- (c) He has no clear idea of the goals to be attained.
- (d) He is more security conscious than status conscious.

This mode of direction can produce good and quick results if the subordinates are highly educated and brilliant people who have a will to go ahead and perform their responsibility.

4. Intellectual leaders

They are the leaders by intellect. Such persons are recognized as leaders on the basis of intellectual work of great importance and relevance done by them for the good of the people. Whether they were scientists, doctors, engineers, poets or philosophers, all have made significant contribution to the good of humanity. This brought to them the status of intellectual leaders.

5. Institutional leaders

They are the leaders by position. Generally, the head of a particular institution is recognized as a leader.

6. Persuasive leaders

They are the leaders by personality. Such leaders fall in the category of charismatic (magnetic / heroism) leadership, The charismatic leader attracts followers on the basis of the qualities of persuasiveness he possesses.

7. Creative or Innovative leaders

They are accepted as leaders on the basis of the contribution made by them in their branch of knowledge. Their contribution is generally of great relevance to human upliftment whether they are scientists, engineers, architects or business experts.

LEADERSHIP THEORIES

More recently the situation in which the leader operates has been given much importance.

It is believed that the leadership effectiveness depends on the situation in which the Leader operates.

We shall discuss a few important theories on leadership with an assertion that any theory will be complete only when it covers three important dimensions of leadership, namely:

1. The leader and his or her psychological attributes;
2. The followers with his or her problems, and needs;
3. The group situation in which followers and leaders relate with one another.

TRAIT THEORY

The trait theories of leadership focus on the **individual characteristic** of successful leaders. According to the theories, leaders possess a set of traits which make them distinct from followers. An attempt must, therefore, be made to identify and measure these traits. Attempts were indeed made in the past to identify such qualities. Ralph Stogdill, for instance, surveyed more than 5000 **leadership** studies and concluded that successful leaders tend to have the following qualities.

- (i) A strong desire for accomplishment
- (ii) Persistent pursuit of goals
- (iii) Creativity and intelligence used to solve problems
- (iv) Initiative applied to social situations
- (v) Self-assumed personality
- (vi) Willingness to accept behavioural consequences
- (vii) Low susceptibility to interpersonal stress
- (viii) High tolerance of ambiguity
- (ix) Ability to influence other people
- (x) Ability to structure social interactions

Evaluation of the Trait Theory: The trait approach to leadership has been severely criticized by many. Some of the limitations of the theory are the following :

- (i) The list of personality traits of successful leaders is too long and there seems to be no finality about it. Although hundreds of traits have been identified, no consistent pattern has emerged.
- (ii) How much of which react a successful leader must have is not clear- Furthermore, certain, particularly psychological, cannot be quantified.
- (iii) The theory assumes that a leader is born and not trained. This assumption is not acceptable to the contemporary thinkers on the subject.
- (iv) Contrary to what the theory assumes, leadership effectiveness does not depend upon the personality of the leader alone. Other variables like the situation, the task, the organization and the characteristics of followers will equally determine the effectiveness of leaders.
- (v) It is well known that people who fail as leaders and people who never achieve positions of leadership often possess some of the same traits as successful leaders.

The Behavioral Theory of Leadership

According to behavioral theory, a leader is as a leader does, so the focus is on the common behaviors of leaders. In that case, there are many types of behaviors exhibited by leaders all around the world and throughout history. There are leaders whose word is law, and there are those that prefer to allow the people to have a hand in the decision-making process.

Autocratic Leadership: These are the leaders who do not consult their subordinates when making decisions in the workplace. Once the decisions have been made the subordinates are expected to cooperate with them with no objections. This type of leadership certainly has an environment where it is highly effective. When decisions have to be made fast, and the leader has extensive knowledge and experience, needing little input, then they can use autocratic leadership to their advantage.

Democratic Leadership: A democratic leader seeks the input of their subordinates before making a decision. The exact degree of input that the leader wants from their team will vary with the leader.

Democratic leadership works in situations where the agreement of the team is necessary for a successful outcome. It also works when the team is cohesive and well-aligned with its goals.

Laissez-Faire Leadership: This type of leader does not involve themselves in the dealings of their subordinates. They give their subordinates the leeway to make their own decisions and direct their own work. To be sure, this type of leadership can work in certain situations, such as where a team is composed of highly skilled and experienced individuals who are competent, motivated, and capable of taking initiative, therefore not requiring any kind of supervision.

The Functional Theory of Leadership

According to this theory, the leader has one main responsibility: to assess the needs of their followers and then meet those needs. They are also tasked with other functions that relate to this one main responsibility:

- ❖ To monitor the environment within which their subordinates work.
- ❖ To organize activities for their followers so that everyone always has something to do.
- ❖ To train their subordinate_s and increase their knowledge and skill sets.
- ❖ To motivate and inspire their followers.
- ❖ To participate in the activities of the group. This is important as it forces them to have skin in the game and builds trust in them among their followers.

The Transformational Theory of Leadership

According to this theory, the leader is tasked with seeing the bigger picture in every situation and motivating their followers to attain greater goals and execute the group's vision. This type of leadership demands that the leader be clearly visible to followers and that they are accessible at all times. They should constantly look for new ideas and ways to realize the goals of the group.

The Transactional Theory of Leadership

According to this theory, a leader is defined by an ability to reward those who perform well and to punish those who do not. A leader should have a specific goal for followers to work toward. A leader should also have the ability to train followers to give them the ability to work towards that goal. From there they should evaluate their followers' performance and determine whether it is satisfactory.

Situational Contingency Theories

Situational contingency theories maintain that the situation is the ultimate factor in the leadership style adopted by a leader. With that in mind, there is no single ultimate leadership style. Autocratic leadership works in times of intense crisis, whereas democratic leadership works in times of relaxation. Situations dictate the most appropriate type of leadership style for other leadership styles, as well.

Difference between Leadership and Management:

Leadership is different from management. The main differences between these two terms are:-

1. A manager is required to plan, organize, direct and control. But a leader is one who gets others to follow him.
2. A manager depends on his authority. But a leader depends on his confidence and goodwill. He inspires enthusiasm.
3. Management is concerned with the formulation of broad policies to guide the operations of an enterprise. But leadership is concerned with the initiation of action for the accomplishment of the goals.

4. An individual is a leader in the true sense if he is accepted as a leader by the group. A manager is appointed and he derives his authority by virtue of his office.
5. Management is associated with the organized structure. But leadership may be associated with unorganised groups.

POWER

Power is the ability to influence other people. It refers to the capacity to affect the behaviour of the subordinate with the control of resources. It is an exchange relationship that occurs in transactions between an agent and a target. The agent is the person who uses the power and target is the receipt of the attempt to use power.

Sources (Base) of Power

According to French and Raven, a manager derives power from five sources : Reward, Coercive, Legitimate, Referent and Expert power.

Reward Power

It is based on the agent's/manager's ability to control rewards the target/employee wants. The common, e.g., of it are managers control rewards of salary increases, bonuses and promotions. This power is based on old saying that 'wealth is power'.

Coercive Power

It is opposite of reward power. It is based on a manager's ability to cause an unpleasant experience for its people. In organizational situation, it may be in the form of action for or threat for dismissal, suspension, or demotion, for the people working in organization.

Legitimate Power

It is based on position and mutual agreement. Both the agent and target agree that the agent has the right to influence the employees. It is in the form of authority which is delegated to the positions of organizational members.

Referent Power

It is an elusive power that is based on interpersonal attraction. Charismatic individuals are often thought to have referent power. Here, people take somebody as ideal and behave accordingly upto a certain stage.

Expert Power

It exists when the agent has information or knowledge that the target needs. It is based on the proverb, "knowledge in power". Three conditions to be fulfilled are :

- (1) The target must trust that the information given by the agent is accurate and correct.
- (2) The information should be relevant and useful to the target.
- (3) The target must consider the agent as an expert.

Acquisition of Power

Some people enjoy more power than others because :

1. Extraordinary Works : Doing things in a non-routine or extraordinary works contribute to power. For example, negotiating a new contract, developing a new product, or formulating a new programme.
2. Visible Activities : Even extraordinary activities not known to others do not generate much power. Therefore, activities need to be visible or known to others. Activities announced and appreciated by the people of higher echelons bring more power.

3. **Cultivate Right People** : Individuals can also increase their personal power by developing their interpersonal relationships with their superiors, subordinates and peers.

4. **Coalitions** : Coalescing is yet another way to earn power. The philosophy behind joining together is gaining increased capability to influence others.

5. **Co-opt** : Individuals can increase their personal power by co-opting people or groups. Co-opting, seeks to eliminate threats and opposition to an individual's base of power.

Meaning of Organizational Politics

It means the use of power and influences in organizations. Actions not officially sanctioned or acceptable by an organization that are taken to influence others in order to meet personal goals refer to politics.

Reasons for Organizational Politics

There are many reasons that contribute to political behaviour in organizations. Some of them are

1. **Clear Goals** : Organizations are human groups work for achieving certain goals. The more unclear and complex the goals are, the more politics will be.

2. **Discretionary Authority** : Organizations provide position with discretionary authority that is used based on individual judgement.

3. **Autocratic Decisions** : The leader dictates the decisions or orders and the subordinates have no right to disobey. This leads to low employee morale and doubts about what the manager-leader decides. Therefore, in order to safeguard their interests, workers involve in politics by forming coalitions and associations.

4. **Power Politics** : Power is also a limited in supply. Hence, there is a competition among managers/executives to acquire more and more power. They try to acquire more power and resources than their competitors. Managers' such behaviour becomes quite dysfunctional.

5. **Saturation in Promotion** : Some people reach maximum level of promotion. They feel dissatisfaction and resort to the organizational politics. Some people may like work performance more than positional achievement and therefore, may not resort to politics.

6. **Biased Performance Appraisal** : When the job performance of a personnel cannot be measured quantitatively, performance appraisal is made on the basis of the judgement of the superior. As such the performance appraisal is likely to be subjective and biased. This may force the subordinates into dysfunctional political behaviour.

UNIT IV GROUP DYNAMICS

Meaning – Types of Groups – Functions of Small Groups – Group Size Status – Managerial Implications – Group Behaviour – Group Norms – Cohesiveness – Group Thinking

Group : A Group is a collection of two or more individuals, interacting and interdependent, which have come together to achieve particular objectives.

A group is, thus, an aggregation of people who interact with each other, are aware of one another, have a common objective, and perceive themselves to be a group.

GROUP DYNAMICS

A group can be defined as several individuals who come together to accomplish a particular task or goal. Group dynamics refers to the attitudinal and behavioral characteristics of a group. Group dynamics concern how groups form, their structure and process, and how they function. Group dynamics are relevant in both formal and informal groups of all types.

Characteristics of group can be listed:

1) Two or More People : A single individual cannot form a group. For group formation, at least two persons are needed. There is no specific limit on the maximum number of persons to form a group.

2) Collective Identity : Each group member knows one another. Each member of the group perceives that he/she is a part of group.

3) Interaction : There is an interaction among the members of the group.

Each member shares his ideas with each other through different communication methods such as face-to-face communication, in writing, over the telephone, and across a computer network etc.

4) Common Purpose : The members of the group work to achieve some common objective or purpose. In fact, it is the common purpose that binds the group members together.

Features of Group Dynamics

The important features of group dynamics are perception, motivation, groups goals, group organization, interdependency, interaction.

1. *Perception.* Group dynamics as defined by perception implies that every member of the group is aware of his respective relationship with others.

The group consists of organisms or agents. The members or agents are engaged in interaction with one another. They have face to face meetings.

They develop some impression or perception about each other and give their reactions to each other. Each member perceives the group differently, which he reveals at some situations. The members perceive the role of the group based on their learning and background.

2. *Motivation.* Members join groups because they expect that the group will solve their problems. They want progress and promotion which are achieved through group performance. The pressures and problems are jointly met by them. Group norms emerge to guide individual behaviour. Cooperative feelings are increased for helping each other. The group is developed taking into consideration individual interests.

3. *Group goals.* Group goals are targets towards which input, process and output are directed. Group goal is the essential component of group formation, although it is not the only condition for forming a group. A goal is used for motivating the employees. The path goal relationship produces a higher responsibility for attaining the goals.

4. *Group organization.* Group is an organization which is composed of different organs to attain certain objectives. A group has the structural elements of an effective organization.

A socio-psychological group is evolved wherein two or more individuals are interrelated. It has a set standard of relationship among its members. Similarly, it has a set of norms that regulate the functions of the group. A number of individuals in the group have definite status, role relationship, set of values and own regulating behaviour. The group structure has power relations, effective relations and well-defined jobs.

5. *Interdependency.* The main feature of a group is the members' interdependence. The members of a group may have a common goal but they may not be a part of the group because they are not interdependent.

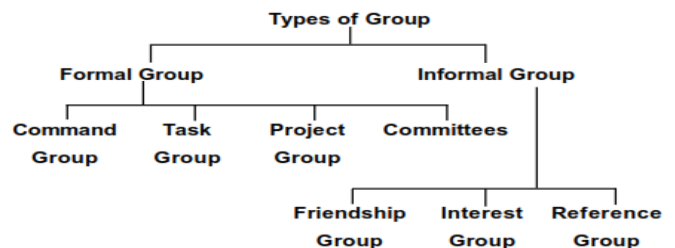
6. *Interaction.* Members of a group must interact with each other. If they are interdependent but do not interact, the group's goals are not achieved. Members have an interpersonal problem-solving mode. If any problem arises, the interaction of all the members is needed to solve the problem.

Each person must communicate with others when the need arises. Interaction differentiates the group from a mere collection of people. Interaction between the members of a group may take different forms, e.g. verbal interaction, physical interaction, emotional interaction and so on.

7. *Entitativity.* A group has its own identity. It has similarity and proximity. It is felt and realised but cannot be seen. The collection of individual experiences become the guidelines for the members.

TYPES OF GROUPS

There are mainly two ways of classifying groups into types: formal and informal groups. Different kinds of formal and informal groups are listed in the table below :



Formal Groups

Groups established by the organization to achieve organizational goals are called formal groups. In formal groups, the behaviour that a member should exhibit is conditioned by organization and directed towards organizational goals. It is possible to subclassify formal groups into the following ones:

1. Command Group : A command group is composed of a supervisor (manager) and the subordinates who report directly to that supervisor. A command group is determined by the organizational chart. For example, in the Department of Business Administration of XYZ University, e.g., the Head of the department and the other faculty members in the department would comprise a command group.

2) Task Group : A task group comprises persons working together to complete a common task. However, a task group can cross command relationship. In a University, for instance, if a student is accused of a campus crime, it may involve interaction among the Head of the Department, the Dean of the school, the Dean of the student's welfare, the *Proctor*, the Registrar of the University. Here it should be noted that all command groups are task groups, but task groups can cut across the organizational boundary. The reverse need not be true.

3) Project Group : Likewise, project groups are formed to complete a specific project. The life of the project group normally coincides with the length of the project. Assigning a research project to a university Professor by the University Grants Commission is an example of a project group.

4) Committees : Committees are usually created outside the usual command group structure to solve recurring problems.

The life of a committee may be relatively long or short. An example of committees is a University's Examination Discipline Committee created to solve discipline problems relating to examination.

Informal Groups

Groups, which are not formal, are informal. In other words, these are the groups that are neither formally created nor controlled by the organization. These groups are natural formations in the work environment that appear in response to the need for social contact.

Four employees belonging to four different departments taking their lunch together represent an example of an informal group.

The various kinds of informal groups are :

1) Friendship Groups : Friendship groups are associations of people who like each other and who like to be together. Such groups are formed because members have one or more common characteristics, such as age or ethnic heritage, political beliefs, religious values, and other bonds of attraction.

2) Interest Groups : Interest groups are composed of individuals who may be members of the same organization (command or task groups), but they are united by their interest in a common issue. Examples of interest groups may include a group of University Professors who organize a seminar on Socio- Economic Problems in the North-Eastern Region of India.

3) Reference Group : A reference group is a special type of informal group that people use to evaluate themselves. A reference group may or may not be an actual one that meets together; it can be an imaginary group. The reference group for a new group for a new university Lecturer, for example, may be other scholars in the same discipline at other universities.

Functions of Small group

A small group is generally defined as a group that consists of at least three members and at the maximum around twelve to fifteen members. A group that has just two members or more than fifteen members would not come in the category of a small group.

This small group is generally formed to reach a common goal which could be to solve a particular problem, make decisions, determine policies and submit reports. Groups involved in regular meeting such as personnel committee, audit committee, report committee, grievance committee are the example of small groups.

Since it may be an informal network of people communicating, the group may or may not have any centralized structure. Every member can influence and can be influenced for performing their task. This generally takes place in a context that mixes interpersonal interactions with social clustering.

- Know skills to encourage meaningful interactions and member contribution, maximize individual participation, enhance motivation and assure commitment to the decisions reached.
- Understand whether to hold a meeting, when to schedule it, how to arrange the meeting rooms and how to develop discussion content (agendas) that keeps meetings on track and get results.
- Any decisions taken must be imparted by influential mentors to all the members.
- Handle problem behaviors and problems effectively.

- Manage the complex dynamics of small groups such as communication, hidden agendas, consensus decision making, coordination difficulties, change and conflict.
- Every competent team has team members with specific skills and knowledge that must be utilized and imparted to other members in the course of the work.
- Any questions or issues about the project must be broached and shared in order to resolve them. This gives a powerful advantage as a group.

Characteristics of Small Groups of communication

So, small-group communication is the process in which information are exchanged among the members (Three to Fifteen) of the same group to ensure interdependent goal accomplishment of the organization.

Advantages or even benefits of **little group communication** usually are as follows:

- **Brainstorming:** Brainstorming can be a form of communication that is designed to aid friends creates ideas. It involves communication between task-oriented groups. While in brainstorming times, associates from the group at first develop numerous ideas as it can be. Following a person in ideas has become introduced towards group, group members examine these ideas along with determining the ones usually are the most likely for their goals.
- **Details discussing:** Little group created to explain available data get a major contributor inside the group to teach some other associates. Now and again, these kinds of group may possibly consist associated with college students with better understanding regarding assessments. When little groups come together to explain data, they will do distinct dialogue designs based on the matter associated with discussion tasks.
- **Difficulty handling:** Every time a little group participates in difficulty handling, it takes to achieve a goal in regards to a distinct problem. With these, associates from the group outline the condition, discover and examine new courses of action (achievable solutions) and pick the ideal solutions for the difficult class discussion.

CONCEPT OF GROUP BEHAVIOUR

Groups are composed of individuals. Hence, the group behavior means behaviour of its members and how in turn it is also affected by them.

The nature and patterns of reinforcement the members receive through interaction with one another is also determined by the group itself. This is because the behaviour of individual members in a group becomes different than their behaviour outside the group situation. Therefore while studying group behaviour, the factors that should be considered are group norms, group cohesion, group decision-making etc.

Group Norms

Group norm is a standard of behaviour. In other words, group norm is a rule that tells the individual how to behave in a particular group. Thus, group norms identify the standards against which the behaviour of group members will be evaluated and help the group members to know what they should or should not do. Norms could be formal or could be informal.

Thus, the group norms have following characteristics;

- 1) As personality reveals an individual, so group norms do for groups.
- 2) Norms serve as the basis for behaviour of group members.
- 3) They predict and control the behaviour of members in groups.
- 4) Norms are applied to all members of the group, though not uniformly.

According to Hackman, norms have the following characteristics:

- (i) Norms summarize and simplify group influence processes. They resolve impersonal differences in a group and ensure uniformity of action.
- (ii) Norms apply only to behaviour, not to private thoughts and feelings.

(iii) Norms are usually developed gradually, but the process can be shortened if members so desire.

(iv) Not all norms apply to everyone. High-status members often enjoy more freedom to deviate from the 'letter of the law' than do other members.

Types of Norms

Norms are unique to each work group. Yet, there are some common classes of norms that appear in most work groups.

(i) *Performance-related processes*: Work groups typically provide their members with explicit cues on how hard they should work, how to get the job done, their level of output, etc. These norms deal with performance related processes and are extremely powerful in affecting an individual employee's performance.

(ii) *Appearance factors*: Some organizations have formal dress codes.

However, even in their absence, norms frequently develop to dictate the kind of clothes that should be worn to work.

(iii) *Allocation of resources*: These norms cover pay, assignment of difficult jobs, and allocation of new tools and equipment.

(iv) *Informal social arrangement*: These norms can originate in the group or the organization and cover pay assignment of difficult jobs, and allocation of new tools and equipment.

Factors Influencing Conformance to Norms

(i) *Personality factors*: Research on personality factors suggests that the more intelligent are less likely to conform than the less intelligent. Again, in unusual situations where decisions must be taken on unclear items, there is a greater tendency to conform to the group's norms. Under conditions of crisis, conformity to group norms is highly probable.

(ii) *Situational factors*: Group size, communication patterns, degree of group unanimity, etc., are the situational factors influencing conformity to norms.

(iii) *Intragroup relationships*: A group that is seen as being creditable will evoke more compliance than a group that is not.

(iv) *Compatible goals*: When individual goals coincide with group goals, people are more willing to adhere to group norms.

Group Cohesion

Group cohesion means the degree to which the group members are attracted to each other and remain within the group. It is usually reflected by its *resiliency* to disruption by outside forces.

Group cohesion develops out of the activities, interactions and sentiments of the members. Cohesiveness binds all the group members to work as one man to attain the set goals.

Factors to increase Group Cohesiveness :There are various factors that determine group cohesiveness.

Factors increasing Group Cohesion are as follows :

- 1) Inducing agreement on group goals
- 2) Increasing membership homogeneity
- 3) Increasing interactions among group members
- 4) Down-sizing of the group
- 5) Encouraging competition with other rival groups.
- 6) Allocating rewards to the group not to the members
- 7) Keeping the members isolated from other groups.

GROUP DECISION-MAKING

Decision-making is the process whereby a final but best choice is made among the alternatives available. When a group makes decision, it can be either through the *consensus* mode or through majority vote. When all members of the group agree to the decision arrived at, it is called 'consensus'. If majority of the group members agree to the decision arrived at, it is called majority vote. Whether the decision arrived at will be consensus or majority mode depends mainly on the size of the group.

Components of Group-Decision Making : The following are the components of group-decision making that should be taken into consideration :

1) Group Size : Research indicates that as the number of members in problem-solving group increases beyond a certain point, the quality of decisions made by the group tends to decrease. This is because the group pressure tends to increase the influence on the decision.

Though the size of an ideal group have not been determined, groups consisting of five to seven members have been found to be effective for decision-making. This is because the members of a group of this size get adequate opportunities to express their opinions, listen to others, seek clarifications on points that are not clear, and reach to a *unanimous* decision. This does not happen in case of large groups.

Nonetheless, larger group may be necessary where variety of skills, knowledge experience and expertise from different functional areas are required for making decisions on critical issue like developing a new product.

2) Group Composition : The qualifications of group members also influence the group decision. Group members with higher status, either due to their background or expertise, are likely to exercise *subtle* pressure, manipulate force or otherwise alter or shake the thinking of the other group member in a particular direction. Minority group members tend to be highly influenced by such group pressures.

3) Unanimity of Group Consensus : It is found that a united group exerts greater pressure if the group is divided by disagreement. It is worth noticing that consensus doesn't require unanimity but requires no outright *dissent* either.

4) The Risky Shift : Research findings are that people tend to make risky decisions when they are engaged in-group decision-making than when the same members make decisions individually.

The members involved in group decision making tend to make risky decision as the members shall be collectively responsible for the consequences of such decisions and the same will be shared by all the group members together rather than one individual member shouldering the entire burden. Such phenomenon for groups to make risky decisions is known as the risky shift. Individuals tend to make conservative decisions because the consequences of the decisions will be accepted by the individual alone.

ADVANTAGES OF GROUP DECISION-MAKING :

1) Compared to an individual, the groups usually have a greater knowledge, expertise, and skill base to make better decisions.

2) Larger number of members provides more perspectives of the problem. As such, the narrow vision of a single perspective is avoided in making decisions.

3) With larger number of group members, the participation also increases that helps reach at a quality decision.

4) Following are increased group participation, comprehension of final decision arrived at is usually high.

DISADVANTAGES OF GROUP-DECISION MAKING :

All is not always good with group decision-making. It suffers from the following disadvantages also:

- 1) Group decision-making is a time consuming process.
- 2) Influential members manipulate the group decision in the direction of their liking and interest.
- 3) Sometimes decisions made by the group members are simply a compromise between the various views and options offered by the group members.

In view of the above disadvantages, there is a need to improve group decisions.

BUILDING AND MANAGING EFFECTIVE TEAM / GROUP

Team building consists of activities designed to construct, develop and sustain groups of people who are working together to achieve common goals with a commitment to take collective responsibility.

- 1) **Balanced Roles** : People with different work preferences must gain entry into teams rather than like-minded people.
 - 2) **Open Communication** : Communication should be open, flexible and capable of building trust between people.
 - 3) **Handling Stress** : Working with others in close proximity can itself be stressful. Additionally, we tend to react differently to various work pressures. The key skill which effective teams develop is the ability to recognize when either the individual or the collective stress is becoming a problem and to reduce it altogether.
 - 4) **Team Choices** : These choices must be made explicit. Once the range is known, alternatives could be ascertained. Suppose someone from the team says, "I think teams are good in principle, but in practice they can be a complete waste of time", Once a response like this come out, the team can begin to deal with it.
 - 5) **Team Goals** : Teams make sense only when there is a common goal, which requires collective action. Defining the team's purpose is vital. This, then, gives the team members a focus for their energy and action. It is also helpful to set some short-term goals to create gains, which sustain the team as it pursues long-term goals.
 - 6) **Review Mechanism** : For a team's success, there must be proper control over team's activities and outcomes. A periodical review is needed to keep everything on track. If it is missing, teamwork will become just another 'flavor of the mouth'.
 - 7) **Shared Leadership** : Teams need different forms of leadership at different times. Both the leader and the members must be willing to exchange roles, depending on the situation.
 - 8) **Facilitation Skills** : Ensure that team meetings are well organized, allowing enough room for all to express their feelings and thoughts openly.
 - 9) **Consensus**: Team decisions need to be based on consensus, so that all members can agree with and be committed to implementing important decisions. Consensus building does not imply 100 percent agreement on the part of the members. It only indicates the willingness of a member to support the decisions on reaching a certain stage or point.
-