MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COMMON FOR: DEPARTMENT OF INFORMATION TECHNOLOGY

GE6757 – TOTAL QUALITY MANAGEMENT

R – 2017

LECTURE NOTES

UNIT I INTRODUCTION

Introduction – Need for quality – Evolution of quality – Definitions of quality – Dimensions of product and service quality – Basic concepts of TQM – TQM Framework – Contributions of Deming, Juran and Crosby – Barriers to TQM – Quality statements – Customer focus – Customer orientation, Customer satisfaction, Customer complaints, and Customer retention – Costs of quality.

INTRODUCTION:

Quality does not mean an expensive product. On contrary it is fitness for use of the product.

NEED FOR QUALITY:

The need for quality was felt, during World War II due to the unprecedented need for manufacture goods. From them on me hodologies for assuring quality in products and services evolved continuously fin lly lead to TQM.

EVOLUTION OF QUALITY/ CONTRIBUTION OF QUALITY GURU:

SHEWHART - Control chart theory
PDCA Cycle

DEMING - Statistical Process Control

JURAN - Concepts of SHEWHART
- Return on Investment (ROI)

FEIGANBAUM - Total Quality Control
- Management involvement
- Employee involvement

ISHIKAWA - Cause and Effect Diagram
- Quality Circle concept

CROSBY - "Quality is Free"
- Conformance to requirements

TAGUCHI - Loss Function concept
- Design of Experiments

DEFINITION OF QUALITY:

1. Quality = Performance / Expectations
2. Quality is defined as the predictable degree of uniformity and dependability, at low cost Suited to the market. (Deming).
3. Quality is defined as fitness for use (Juan).
4. Quality is defined as conformance to requirements (Crosby).
5. Quality is totality of the characteristics of entity that bear on its ability to satisfy stated and implied needs (ISO).

DIMENSIONS OF PRODUCT AND SERVICE QUALITY: PRODUCTQUALITY:

1. Performance - FulfillmentCivildatasofprimaryrequirement

2. Features - Additional things that enhance performance

3. Conformance - Meeting specific st nd rds set by the industry

4. Reliability - Consistence performance over period of time

5. Durability - Long life and less maintenance

6. Service - Ease of repair, guarantee, and warranty

7. Response - Dealer customer relationship, human interface

8. Aesthetics - exteriors, packages, appearance

9. Reputation - Past performance, ranking, branding

SERVICE QUALITY:

1. **Reliability -** Refers to the dependability of the service providers and their ability to keep their promises.
2. **Responsiveness -** Refers to the reaction time of the service.
3. **Assurance -** Refers the level of certainty a customer has regarding the quality of the service provided.
4. **Empathy - Being** able to understand the needs of the customer as an individual.

5. **Tangibles -** Similar to the physical characteristics of quality of products.

6. **Other Dimensions - Time**, Courtesy, Timeliness, consistency, accuracy, credibility and security.

TOTAL QUALITY MANAGEMENT (TQM):

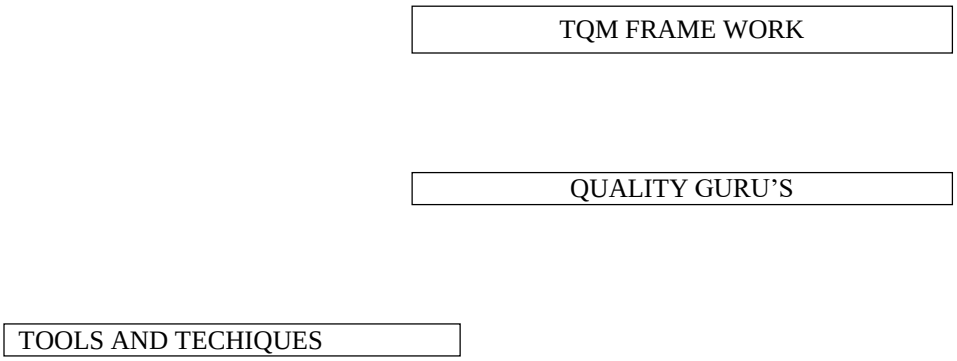
TQM is defined as both philosophy and a set of guiding principles that represent the foundation of continuously improving organization. It is the application of quantitative methods and human resources to i prove all the process within the organization and exceed customer needs now and in the future.

Total Quality Management is an effective system for integrating the quality development, quality maintenanceandqualityimprovement efforts of various groups in an organization continuously, so s to enable marketing, engineering, production and service at the most economic levels which allow for full customer satisfaction.

BASIC CONCEPTS OF TQM:

1. **Management Involvement** – Participate in quality program, develop quality council, direct participat on.
2. **Focus on customer** – who is the customer – internal and external, voice of the customer, do r ght first time and every time.
3. **Involvement and utilisation of entire work force** – All levels of Management
4. **Continuous improvement** – Quality never stops, placing orders, bill errors, delivery, minimize wastage and scrap etc.
5. **Treating suppliers as partners** – no business exists without suppliers.
6. **Performance measures** – creating accountability in all levels.

TQM FRAME WORK:



CONTRIBUTIONS OF DEMING:

1. Create and publish the Aims and Purposes of the organization.
2. Learn the New Philosophy.
3. Understand the purpose of Inspection.
4. Stop awarding business based on price alone.
5. Improve constantly and forever the System.
6. Institute Training.
7. Teach and Institute Leadership.
8. Drive out Fear, Create Trust and Create a climate for innovation.
9. Optimize the efforts of Teams, Groups and Staff areas.
10. Eliminate exhortations for the Work force.

11a. Eliminate numerical quotas for the work force.

11b. Eliminate Management by objectives.

12. Remove Barriers that rob people of pride of workmanship.
13. Encourage Education and Self-improvement for everyone.

14. Take action to accomplish the transformation.

CONTRIBUTIONS OF JURAN:

THE JURAN TRILOGY

Juran views quality as fitness for use.

Juran Trilogy is designed to reduce the cost of quality over time.

1. QUALITY PLANNING

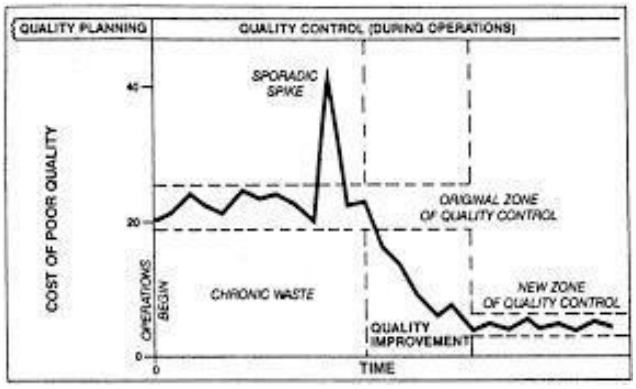
1. Determine internal & external customers.
2. Their needs are discovered.
3. Develop product / service features.
4. Develop the processes able to produce the product / service features.
5. Transfer plans to operations.

2. QUALITY CONTROL

1. Determine items to be controlled.
2. Set goals for the controls.
3. Measure actual performance.
4. Compare actual performance to goals.
5. Act on the difference.

3. QUALITY IMPROVEMENT

1. Establishment of quality council.
2. Identify the improvement projects.
3. Establish the project teams with a project leader.
4. Provide the team with the resources.



CONTRIBUTIONS OF CROSBY:

The Four absolutes of quality are

1. Quality is defined asCivildatasconformancetorequirements
2. The system for causing Quality is prevention.
3. The performance standard must be zero defec .
4. The measurement of Quality is the Price of Nonconformance

Crosby's Fourteen Points:

1. Management Commitment
2. Quality Improvement Team
3. Quality Measurement
4. Cost of Quality Evaluation
5. Quality Awareness
6. Corrective Action
7. Establish an Ad Hoc Committee for the Zero Defects Program
8. Supervisor Training
9. Zero Defects Day
10. Goal Setting
11. Error Cause Removal
12. Recognition
13. Quality Councils
14. Do It Over Again

OBSTACLES (BARRIERS) IN IMPLEMENTING TQM:

1. Lack of Management Commitment
2. Inability to change Organizational culture
3. Improper planning
4. Lack of continuous training and education
5. Incompatible organizational structure and isolated individuals and departments.

6. Ineffective measurement techniquesandlackof access to data and results.

7. Paying inadequate attention to intern nd external customers
8. Inadequate use of empowerment nd te mwork
9. Failure to continually improve

BENEFITS OF TQM

Customer satisfaction or ented benefits:

1. Improvement in product qual ty
2. Improvement in product design
3. Improvement in production flow
4. Improvement in employee morale and quality consciousness
5. Improvement in product service
6. Improvement in market place acceptance

Economic improvement oriented benefits:

1. Reduction in operating costs
2. Reduction in operating losses
3. Reduction in field service costs
4. Reduction in liability exposure

QUALITY STATEMENTS

a. Vision statement,
b. Mission statement, and
c.Qualitypolicystatement

1. ***The vision statement*** Civil datas is a short declaraion of wht on organization as pirto be tomorrow.
2. It is the ideal state that might never be reached; but on which one will work hard continuously to achieve. Successful visions provide a brief guideline for decision making.
3. The vision statement should be coined in such way that the leaders and the employees working in the organization should work towards the achievements of the vision statement.

a) **The mission statement** describes the function of the organization. It provides a clear statement of purpose for employees, customers, and suppliers.
b) The mission statement answers the following questions: who we are? Who are our customers? ; What we do? and how we do it?

i. ***The quality policy*** is a guide for everyone in the organization as to how they provide products and service to the customers.

ii. It should be written by the CEO with feedback from the workforce and be approved by the quality council.
iii. A quality policy is a important requirement of ISO 9000 quality systems.

CUSTOMER FOCUS:

Customer is the King.
"Quality what the customer wants" It emphasis on the custo er. Customer satisfaction must be the primary goal of any organization, therefore it is essential that every employee in the organization understands the importance of the customer. A satisfied customer will led to increased profits.

CUSTOMER SATISFACTIONMODEL:

Customer satisfaction is not an objective but feeling or attitude. Since it is subjective it is not easy to measure. There re so many facets to a customer experience with a product and service th need to be measured individually to get the accurate picture of customer satisfaction. Customer Satisfaction Model – Teboul

Types of Customers

1. Internal customers
2. External customers

Internal Customers:

1. The customers inside the organization
2. The flow of work, product and service in the organizati
   dependent on one and another.
3. Every person in a
   processCivildatasisconsideredthecutomer operation.

External Customers:

1. Uses the product or service
2. Who purchase the product.
3. Who influence the sale of the Product or services.

Kano Model

CUSTOMER COMPLAINTS:

Customer Satisfaction analysis helps the organization in the following ways:

1. A totally satisfied customer contributes to revenue of the company.
2. A totally dissatisfied customer decrease revenue.

CUSTOMER FEEDBACK:

Customer feedback is required for the following reasons.

1. To discover customer dissatisfaction
2. To identify the customer needs
3. To discover relative priories of quality
4. To compare performance with competition
5. To determine opportunities for improvement.

TOOLS OF CUSTOMER COMPLAINTS:

a. Comment card
b. Customer Questionnaire
c. Focus Groups
d. Toll Free telephone
e. Customer Visit
f. Report Card
g. Internet & Computers
h. Employee Feedback.
i. Mass customization

CUSTOMER RETENTION

It means "retaining the customer" to support the business. It is more powerful and effective than customer satisfaction.
For Customer Retention, we need to have both "Customer satisfaction & Customer loyalty".

The following steps are important for customer retention.

1. Top management commitment to the customer satisfacti n.
2. Identify and understand the customers what they like and dislike about the organization.
3. Develop standards of quality service and performance.
4. Recruit, train and reward good staff.
5. Always stay in touch with customer.
6. Work towards continuous improvement of customer service and customer

retention.

7. Reward service accomplishments by the front-line staff.

8. Customer Retention mo es customer satisfaction to the next level by determining what is truly mportant to the customers.
9. Customer satisfaction is the connection between customer satisfaction and bottom line.

COST OF QUALITY:

The Value of Quality must be based on its ability to contribute to profits. Quality related cost is the cost incurred by an organization to ensure that the products / services it provides conform to customer requirements.

DEFINITION:

Quality cost is defined as those costs associated with the non-achievement of products / service quality as defined by the requirements established by the organization and its contract with the customer. Quality cost is the cost of poor products or services. When Quality Cost is too high, it is sign of management ineffectiveness, which affects the organization competitive position.

PREVENTION COST:

"The cost that are incurred on preventing a quality problem from arising."

a. Marketing / Customer/ User
b. Product / Service / Design Development
c. Purchasing
d. Operations
e. Quality Administration

APPRAISAL COST:

"The Cost incurred in assessing that the products / services conform to the requirements"

a. Purchasing Appra sal Cost
b. Operation (Manufacturing or Service) Appraisal Cost
c. External Appra sal Cost
d. Review of test and inspection data:
e. Miscellaneous quality evaluation:

INTERNAL FAILURE COST:

"Cost arises due to internal failures."

a. Product or Service Design Failure Cost:
b. Purchasing failure cost
c. Operations cost

EXTERNAL FAILURE COST:
The cost incurred due to the non conformance of the products or services after delivery of products to the customer.

Quality Improvement Strategy:

1. Project team:
2. Reduce the Failure cost:
3. Prevention of quality cost.
4. Reducing appraisal cost.

UNIT 2 - TQM PRINCIPLES

SYLLABUS: Leadership - Strategic quality planning, Quality Councils - Employee involvement - Motivation, Empowerment, Team and Teamwork, Quality circles Recognition and Reward, Performance appraisal - Continuous process improvement - PDCA cycle, 5S, Kaizen – Supplier partnership - Partnering, Supplier selection, Supplier Rating.

LEADERSHIP:

The success of quality management is to a greater extent is influenced by the quality of the leadership. Peter Drucker, the eminent management thinker and writer quotes: "LeadershipCivildatasisliftingofman'sviions to higher sights, the raising of man's performance to higher standard, the building of man's personality beyond its normal limitations".

Leadership is the process of influencing others towards the accomplishment of goals. Leader triggers the will to o, show the direction and guide the group members towards the accompl shment of the company's goal.

CHARACTERISTICS OF QUALITY LEADERS:

1. They give priority attent on to external and internal customers and their needs.
2. They empower, rather than control, subordinates.
3. They emphasis improvement rather than maintenance.
4. They emphasis prevention.
5. They emphasis collaboration rather than competition.
6. They train and coach, rather than direct and supervise.
7. They learn from the problems.
8. They continually try to improve communications.
9. They continually demonstrate their commitment to quality.
10. They choose suppliers on the basis of quality, not price.
11. They establish organizational systems to support the quality effort.
12. They encourage and recognize team effort.

THE 7 HABITS OF HIGHLY EFFECTIVE PEOPLE:

1. Be Proactive
2. Begin with the End in mind
3. Put First Things First
4. Think Win – Win
5. Seek First to Understand, then to Be Understood
6. Synergy
7. Sharpen the Saw (Renewal)

LEADERSHIP CONCEPTS

A leader should have the following concep s

1. People, Paradoxically, need security nd independence at the same time.

2. People are sensitive to external and punishments and yet are also strongly self – motivated.
3. People like to hear a kind word of praise. Catch people doing something right, so you can pat them on the back.
4. People can process only few facts at a time; thus, a leader needs to keep things simple.
5. People trust their gut reaction more than statistical data.
6. People distrust a leader's rhetoric if the words are inconsistent with the leader's actions.

STRATEGIC QUALITY PLANNING:

Strategic quality planning (SQP) is a systematic approach to defining long-term business goals, including goals to improve quality and the means (i.e., the plans) to achieve them.
Goals should:

- Improve customer satisfaction, employee satisfaction and process
- Be based on statistical evidence
- Be measurable
- Have a plan or method for its achievement
- Have a time frameforachievingthegoal
- Finally, it should be challenging yet achiev ble

SEVEN STEPS TO STRATEGIC QUALITY PLANNING:

1. **Customer needs** - Discover the future needs of the customer.

2. **Customer positioning** - Planners etermine where the organization wants to be in relation to the customers.
3. **Predict the future** – Demographics, economic forecasts, and technical assessments or projection tools for predicting the future.
4. **Gap Analysis** – Ident fy the gaps between current state and the future state of the organization. An analysis of core values and concepts are excellent techniques for pinpointing the gaps
5. **Closing the Gap** – A plan has to be developed to close the gap by establishing goals and responsibilities.
6. **Alignment** – Once a plan is developed it must be aligned with the vision, mission, and core valuesand concepts of the organization.
7. **Implementation** – Resources must be allocated to collecting data, designing

changes, and overcoming resistance to change.

Employee Involvement:

Employee involvement is one approach to improve quality and productivity. It is a means to better meet the organization's goals for quality and productivity.

MOTIVATION:

"Motivation means a process of stimulating people to accomplish desired goals."

Motivation is the process of inducing people inner drives and action towards certain goals and committing his energies to achieve these g als.

IMPORTANCE OF MOTIVATION:

a. Motivation improvesemployeeinvolvement.
b. Motivation promotes job satisfaction and thus reduces absenteeism and turnover. c. Motivation helps in securing high level of performance and hence enhances efficiency and productivity.

d. Motivation creates a congenial working atmosphere in the organization and thus promotes interpersonal cooperation.

THEORIES OF MOTIVATION:

Though there are many theories of motivation, the Maslow's hierarchy of needs theory and Herzberg's two factor theory are more important from our subect of view.

MASLOW'S HIERARCHY OF NEEDS:

Maslow has set up hierarchy of five levels of basic needs. Beyond these needs, higher levels of needs exist. These include needs for understanding, esthetic appreciation and purely spiritual needs. In the levels of the five basic needs, the person does not feel the second need until the demands of the first have been satisfied, nor the third until the second has been satisfied, and so on. Maslow's basic needs are as follows:

Physiological Needs

These are biological needs. They consist of needs for oxygen, food, water, and a relatively constant body temperature. They are the strongest needs because if a person were deprived of all needs, the physiological ones would come first in the person's search for satisfaction.

Safety Needs

When all physiological needs are satisfied and are no longer controlling thoughts and behaviors, the needs for security can become active. Adults have little awareness of their security needs except in times of e ergency or periods of disorganization in the social structure (such as widespread rioting). Children often display the signs of insecurity and the need to be safe.

Needs of Love, Affection and Belongingness

When the needs for safety and for physiological well-being are satisfied, the next class of needs for love, affection and belongingness can emerge. Maslow states that people seek to overcome feelings of loneliness and alienation. This involves both giving and receiving love, ffection and the sense of belonging.

Needs for Esteem

When the first three classes of needs are satisfied, the needs for esteem can become dominant. These involve needs for both self-esteem and for the esteem a person gets from others. Humans have a need for a stable, firmly based, high level of self-respect, and respect from others. When these needs are satisfied, the person feels self-confident and valuable a person in the world. When these needs are frustrated, the person feels inferior, weak, helpless and worthless.

Needs for Self-Actualization

All of the foregoing needs are satisfied, then and only then are the needs for selfactualization activated. Maslow describes self-actualization as a person's need to be and do that which the person was "born to do." "A musician must make music, an artist must paint, and a poet must write." These needs make themselves felt in signs of restlessness. The person feels on edge, tense, lacking something, in short, restless. If a person is hungry, unsafe, not loved or accepted, or lacking self-

esteem, it is very easy to know what the person is restless about. It is not always clear what a person wants when there is a need for self-actualization.
The hierarchic theory is often represented as a pyramid, with the larger, lower levels representing the lower needs, and the upper point representing the need for self-actualization. Maslow believes that the only reason that people would not move well in direction of self-actualization is because of hindrances placed in their way by society. He states that education is one of these hindrances.

HERZBERG'S TWO FACTOR THEORY:

This theory is also called ***motivation-hygiene theory***. This theory is based on two factors: 1. Motivation factors or sati fier , and 2. Hygiene factors or dissatisfiers.

Motivational factors:

1. Achievement
2. Recognition
3. Work itself
4. Responsibility
5. Advancement and growth

Hygiene factors:

1. Supervisors
2. Working conditions
3. Interpersonal relationship
4. Pay and security
5. Company policy and administration

According to Herzberg, maintenance or hygiene factors are necessary to maintain a reasonable level of satisfaction among employees. These factors do not provide satisfaction to the employees but their absence will dissatisfy them. Therefore these factors are called dissatisfiers.

On the other hand, motivational factors create satisfaction to the workers at the time of presence but their absence does not cause dissatisfaction. It can be noted that Herzberg's dissatisfiers are roughly equivalent to Maslow's lower levels, and the motivators are similar to the Maslow's upper levels.

EMPOWERMENT:

Empowerment is investing people with authority. Its purpose is to tap the enormous reservoir of potential contribution that lies within every worker. The principles of empowering people are given below:

1. Tell people what their responsibilities are
2. Give authority
3. Set standards for excellence.
4. Render training.
5. Provide knowledge and inform tion.
6. Trust them.
7. Allow them to commit mistakes.
8. Treat them with dign ty and respect.

CHARACTERISTICS OF EMPOWERED EMPLOYEES:

1. They feel respons ble for their own task.
2. They are given free hand in their work.
3. They balance their own goals with those of the organization.
4. They are ell trained, equipped, creative, and customer oriented.
5. They are critical, have self-esteem, and are motivated.
6. They are challenged and encouraged.
7. They monitor and improve their work continuously.
8. They find new goals and change challenges.

TEAM:

A team is defined as a group of people working together to achieve common objectives or goals.

TEAMWORK:

Teamwork is the cumulative actions of the team during which each member of the team subordinates his individual interests and opinions to fulfill the objectives or goals of the group.

NEED FOR TEAMWORK:

1. Many heads are more knowledgeable than one.
2. The whole is greater than the sum of its members
3. Team members develop rapport which each other.
4. Teams provide the vehicle for improved communication.

TYPES OF TEAMS:

1. Process improvement team.
2. Cross – functional team.
3. Natural work teams.
4. Self – Directed / Self – Managed work teams.

CHARACTERISTICS OF SUCCESSFUL TEAMS:

1. Sponsor
2. Team Charter
3. Team Composition
4. Training
5. Ground Rules
6. Clear Objectives
7. Accountability
8. Well-Defined decision procedure
9. Resources
10. Trust
11. Effective Problem Solving
12. Open Communication
13. Appropriate Leadership
14. Balanced Participation
15. Cohesiveness

1. Sponsor: In order to have effective liason with the quality council, there should be a sponsor. The sponsor is a person from the quality uncil; he is to provide support to the organization.

2. Team Charter: A Civildatasteamcharterisdocumentthat defines the team's mission, boundaries, the background of the problem, the te m's authority and duties, and resources. It also identifies the members and heir ssigned roles – leader, recorder, time keeper and facilitator.

3. Team Composition: The size of the team should not exceed ten members except in the case of natural work teams or self-directed teams. Teams should be diversed by having members w th different skills, perspective and potential. Wherever needed, the nternal and external customers and suppliers should be included as a team member.

4.Training: The team members should be trained in the problem-solving techniques, team dynamics and communication skills.

4. Ground Rules: The team should have separate rules of operation and conduct. Ground rules should be discussed with the members, whenever needed it should be revie ed and revised.

5. Clear Objectives: The objective of the team should be stated clearly. Without the clear objective, the team functions are not to be effective.

6. Accountability: The team performance is accountable. Periodic status report of the team should be given to the quality council. The team should review its performance to determine possible team process weaknesses and make improvements.

8. Well-defined Decision Procedures: The decision should be made clearly at the right time by the team.

9. Resources: The adequate information should be given to the team wherever needed. The team cannot be expected to perform successfully without the necessary tools.

10. Trust: Management must trust the team to perform the task effectively. There must also be trust among the members and a belief in ea h other.

11. Effective Problem-Solving: Problem-solvingmethods are used to make the effective decision.

12. Open Communication: Open communic ion should be encouraged i.e., everyone feels free to speak in the te m wh tever they are thinking, without any interruptions.

13. Appropriate Leadership: Leadership is important in all the team. Leader is a person who leads the team, mot vates the team and guides the team in a proper direction.

14. Balanced Participat on: Everyone in team should be involved in the team's activities by voicing their opinions, lending their knowledge and encouraging other members to take part

15. Cohesiveness: Members should be comfortable working with each other and act as a single unit, not as individuals or subgroups.

ELEMENTS OF EFFECTIVE TEAM WORK:

1. Purpose
2. Role and responsibilities
3. Activities
4. Effectiveness
5. Decisions
6. Results, and
7. Recognition.

STAGES OF TEAM DEVELOPMENT:

Each term takes someCivildatastimetostartfunctioning effectively towards problem solving. Each team goes through six distinct st ges in its development. These are **farming, storming, norming, performing m in enance and evaluating.**

1. Farming stage: When team is cre ted, it consists of group of individuals and team work does not exist at this stage. Team's purpose, members' roles, acceptance of roles, authority and process of functioning are learnt in the formation process.

2. Storming stage: Init al agreements and role allocations are challenged and re-established at this stage of team development. At this stage, hostilities and personal needs often emerge which may be resolved.

3. Norming stage: During norming stage of team development, formal and informal relationships get established among team members. Openness and cooperation have been observed as signs of team's behaviour.

4. Performing stage: At this stage, the team starts operating in successful manner. Trust, openness, healthy conflict and decisiveness of a group's performance can be reached at this stage.

5. Maintenance stage: Functioning of team does not deteriorate overtime. At this stage, the performance of teamwork at the earlier stage will be maintained for some period of time.

6. Evaluating stage: At this stage, team's performance is to be evaluated in view of the set targets. Both self-evaluation and management-based evaluation form this stage of team development.

COMMON BARRIERS TO TEAM PROGRESS:

1. Insufficient training.
2. Incompatible rewards and compensation.
3. First-line supervisorresistance.
4. Lack of planning.
5. Lack of management support.
6. Access to information systems.
7. Lack of Union support.
8. Project scope too large.
9. Project objectives are not s gn ficant.
10. No clear measures of success.
11. No time to do improvement work.

RECOGNITION AND REWARD:

Recognition is process whereby management shows acknowledgement of an employee's outstanding performance. Recognition is a form of employee positive motivation. Recognition of employees is highly essential as people find themselves in a accepted and winning role. To sustain employee's interest and to propel them towards continuous improvement, it is essential to recognize the people. This acknowledgement may be of financial, psychological or both in nature.

Reward is a tangible one, such as increased salaries, commissions, cash bonus, gain sharing, etc; to promote desirable behavior.

METHODS TO RECOGNIZE PEOPLE:

1. Develop a behind the scenes awards specifically for those whose actions are not usually in the lime light, make sure such awards are in the lime light.
2. Create best ideas of the year booklet and include everyone's picture name and description of their best ideas.
3. Feature the quality team of the month and put their picture in a prominent place.
4. Honor peers who have helped you by recognizing them at your staff meetings.
5. Let people attend meetings, commi ees e c; in your place when you are not available.
6. Involve teams with external customers and suppliers, sending them on appropriate visits to solve prob ems and look for opportunities.
7. Invite a team for coffee or lunch any time, not necessarily when you need them for someth ng.

8. Create a visibility wall to display information, posters, and pictures, thanking individual employees and their teams**,** and describing their contributions
9. When you are discussing an individual or group ideas with other people, peers, or higher management make sure that you give them credit.

NEED FOR RECOGNIZATION:

1. Improve employee's morale
2. Show the company's appreciation for better performance
3. Create satisfied workplace
4. Create highly motivated workplace.
5. Reinforce behavioral patterns.
6. Stimulate creative efforts.

TYPES OF REWARDS:

1. Intrinsic rewards
2. Extrinsic rewards

Intrinsic rewards are related to feelings of accomplishment of self-worth.

Extrinsic reward are related to pay or compension issues.

EFFECTS OF RECOGNITION AND REWARD SYSTEM:

1. Recognition and reward go together for letting people know that they are valuable members for the organization.
2. Employee involvement can be achieved by recognition and reward system.
3. Recognition and reward system reveals that the organization considers quality and product v ty as important.
4. It provides the organization an opportunity to thank high achievers.
5. It provides employees specific goal to achieve.
6. It motivates employees to improve the process.
7. It increases the morale of the workers.

PERFORMANCE APPRAISAL:

The performance appraisal is used to let employees know how they are performing. The performance appraisal becomes a basis for promotions, increase in salaries, counseling and other purposes related to an employee's future.

IMPORTANCE OF PERFORMANCE APPRAISALS:

1. It is necessary to prevail a good relationship between the employee and the appraiser.
2. Employee should be informed about how they are performing on a continuous basis, not just at appraisal time.
3. The appraisal should highlight strength and weakness and how to improve the performance.
4. Employee should be allowed to comment on the evaluati n and protest if necessary.
5. Everyone shouldCivildatasunderstandthatthepurpoeof performance appraisal is to have employee involvement.
6. Errors in performance evaluations should be voided.
7. Unfair and biased evaluation will render poor rating and hence should be eliminated.

BENEFITS OF EMPLOYEE INVOLVEMENT:

1. Employees make better dec ons using their expert knowledge of the process
2. Employees are better able to spot and pin-point areas for improvement.
3. Employees are better able to take immediate corrective action.
4. Employee involvement reduces labour / management friction.
5. Employee involvement increases morale.
6. Employees have an increased commitment to goals because they are involved.

CONTINUOUS PROCESS IMPROVEMENT:

Continuous process improvement is designed to utilize the resources of the organization to achieve a quality-driven culture.

PDCA (plan-do-check-act)

PDCA (plan-do-check-act, sometimes seen as plan-do-check-adjust) is a repetitive four-stage model for continuous improvement in business process management.

The PDCA model is also known as the Deming circle/cycle/wheel, Shewhart cycle, control circle/cycle, or plan–do–study–act (PDSA).

PDCA was popularized by Dr. W. Edwards Deming, an American engineer, statistician and management consultant. Deming is often considered the father of modern quality control.

TQM processes are often divided into the four sequential categories: plan, do, check, and act.

Plan*:* Define the problem to be addressed, collect relevant data, and ascertain the problem's root cause.

Do: Develop and implement a solution; decide upon a measurement to gauge its effectiveness.

Check: Confirm the results through before-and-after data comparison.

Act: Document the results, inform others about process changes, and make recommendations for the problem to be addressed in the next PDCA cycle.

5S Principles:

The 5S framework was originally developed by just-in-time expert and international consultant Hiroyuki Hirano. The 5S framework is an extension of Hirano's earlier works on justin-time production systems. The 5Ss represent a simple "good housekeeping" approach to improving the work environment consistent with the tenets of Lean Manufacturing System. It pro otes daily activity for continuous improvement. It fosters efficiency and productivity while improving work flow. It encourages a proactive approach that prevents pr blems and waste before they occur. It provides a practical method for dealing with the real problems

that workers face every day.

SEIRI / SORT / CLEANUP:

The first step of the "5S" process, Seiri, refers to the act of throwing away all unwanted, unnecessary, and unrelated m teri ls in the workplace. People involved in Seiri must not feel sorry about having to throw away things. The idea is to ensure that everything left in the workp ace is related to work. Even the number of necessary items in the workplace must be kept to its absolute minimum. In performing SEIRI, th s mple guideline is a must:

1. Separate needed tems from unneeded items.
2. Remove unneeded items from working areas.
3. Discard the items never used.
4. Store items not Item not needed now.
5. Remove all excess items from working areas, including work pieces, supplies, personal items, tools, instruments, and equipment.
6. Use red tag to get rid of unneeded items.
7. Store items needed by most people in a common storage area.
8. Store items only needed by each individual in his/her own working area.
9. Organize working / storage area.

SEITON / SET IN ORDER / ARRANGING:

SEITON, or orderliness, is all about efficiency. This step consists of putting everything in an assigned place so that it can be accessed or retrieved quickly, as well as returned in that same place quickly. If everyone has quick access to an item or materials, work flow becomes efficient, and the worker beco es productive. Every single item must be allocated its own place for safekeeping, and each location must be labelled for easy identification of what it's for. Its objective includes; the needed items can be easily found, stored and retrieved, supports efficiency and productivity, First-in first-out (FIFO), and save space and time.

In performing SEITON, follow these guidelines:

1. A place for everything and everything in its place.
2. Place tools and instructional manual close to the point of use.
3. Store similar items together. Different items in separate rows.
4. Don't stack items together. Use rack or shelf if possible.
5. Use small bins to organ ze small items.
6. Use color for quickly dentifying items.
7. Clearly label each item and its storage areas (lead to visibility).
8. Use see-through cover or door for visibility.
9. Use special designed cart to organize tools, jigs, measuring devices, etc., that are needed for each particular machine.

SEISO / SHINE / NEATNESS

SEISO, the third step in "5S", says that 'everyone is a janitor.' SEISO consists of cleaning up the workplace and giving it a 'shine'. Cleaning must be done by everyone in the organization, from operators to managers. It would be a good idea to have every area of the workplace assigned to a person or group of Persons for cleaning. SEISO is not just cleaning, but a whole attitude that includes ensuring everything is in perfect condition. Everyone should see the 'workplace' through the eyes of a visitor - always thinking if it is clean enough to make a good impression. Its objective includes; cleanliness ensures a more comfortable and safe working place, cleanliness will lead to visibility so as to reduce search ti e and cleanliness ensures a higher quality of work and products.

Follow these guidelines in performing SEISO:

1. Use dust collecting covers or devices to prevent possible dirt or reduce the amount of dirt.

2. Investigating the causes of dirtiness nd implement a plan to eliminate the sources of dirt.
3. Cover around cords, legs of m chines nd tables such that dirt can be easily and quickly removed.
4. Operators clean their own equipment and working area and perform basic preventive maintenance.
5. Keep everything clean for constant state of readiness.

SEIKETSU / SYSTEMIZE / DISCIPLINE

The fourth step of "5S", or SEIKETSU, more or less translates to 'standardized clean-up' It consists of defining the standards by which personnel must measure and maintain 'cleanliness'. SEIKETSU encompasses both personal and environmental cleanliness. Personnel must therefore practice 'SEIKETSU' starting ith their personal tidiness. Visual management is an important ingredient of SEIKETSU. Color-coding and standardized coloration of surroundings are used for easier visual identification of anomalies in the surroundings. Personnel are trained to detect abnormalities using their five senses and to correct such abnormalities immediately.

The guidelines include:

1. Removing used, broken, or surplus items from the work area
2. Making safety a prime requirement by paying attention to noise, fumes, lighting,

1. cables, spills, and other aspects of the workplace environ ent
2. Checking that items are where they should be

3. Listening to the "voice" of the process and being alert to things such as unusual noises

4. Ensuring that there Civildatas is a place for everything and that everything is in its place
5. Wearing safe working apparel and using s fe equipment
6. Minimizing all waste and the use of valuable resources such as oil, air, steam, water, and electricity

SHITSUKE / SUSTAIN / ON-GOING IMPROVEMENT:

The last step of "5S", SHITSUKE, means 'Discipline.' It denotes commitment to mainta n orderliness and to practice the first 4 S as a way of life. The emphasis of SHITSUKE elimination of bad habits and constant practice of good ones.

Once true SHITSUKE is achieved, personnel voluntarily observe cleanliness and orderliness at all times, without having to be reminded by management. The characteristic of 5S tends to overlap significantly rather than cover very different subjects. Rather than worry about what fits into SEIRI and what fits into Seiton, use them to reinforce each other and implement the whole thing.

KAIZEN: [KAI =CHANGE, ZEN = GOOD]

Kaizen is the practice of continuous improvement. Kaizen was originally introduced to the West by Masaaki Imai in his book Kaizen: The Key to Japan's Competitive Success in 1986. Kaizen is continuous improvement that is based on certain guiding principles:

1. Good processes bring good results
2. Go see for yourself to grasp the current situation
3. Speak with data, manage by facts
4. Take action to contain and correct root causes of problems
5. Work as a team
6. Kaizen is everybody's business

KAIZEN WHEEL:

The Kaizen improvementCivildatasfocusesontheuseof:

1. Value – added and non – value work actives.
2. Muda, which refers to the seven cl sses of waste – over-production, delay, transportation, processing, inventory, wasted motion, and defective parts.
3. Principles of materials hand ing and use of one – piece flow.
4. Documentation of standard operating procedures.
5. The five S's for workplace organization.
6. Visual management.
7. Just – in – time principles.
8. Principles of motion study and the use of cell technology.
9. Poka – Yoke.
10. Team dynamics.

SUPPLIER PARTNERSHIP:

What is Supplier Partnering?

Partnering is a defined as a continuing relationship, between a buying firm and supplying firm, involving a commitment over an extended time period, an exchange of information, and acknowledgement of the risks and rewards of the relationship. The relationship between customer and supplier should be based upon trust, dedication to common goals and objectives, and an understanding of each party's expectations and values.

Benefits of Partnering:

a. Improved quality;

b. reduced cost;

c. Increased productivity;

d. Increased efficiency;

e. Increased market share;

f. Increased opportun ty for innovation; and

g. Continuous improvement of products / services.

The three key elements to partnership relationship are

1. Long term commitment
2. Trust
3. Shared Vision

SOURCING:

The three types of sourcing are

1. Sole sourcing
2. Multiple sourcing
3. Single sourcing

SUPPLIER SELECTION

The suppliers should be selected with the f ll wing ten conditions

1. The supplier should understand clearly the management philosophy of the organization.
2. The supplier should have stable management system.
3. The supplier should maintain high echnic st ndards.
4. The supplier should provide the r w m terials and parts which meet quality specifications required by the purchaser.
5. The supplier should have the required capability in terms of production.
6. The supplier should not leak out the corporate secrets.
7. The supplier should quote right price and should meet the delivery schedule. The supplier should be accessible with respect to transportation and communication.
8. The supplier should be sincere in implementing the contract provisions.
9. The supplier should have an effective quality system such as ISO / QS 9000.

10.The supplier should be renowned for customer satisfaction.

SUPPLIER CERTIFICATION:

A certified supplier is one which, after extensive investigation, is found to supply material of such quality that is not necessary to perform routine testing.

The Eight criteria for supplier certification are

1. No product related lot rejections for at least 1 yearcom.
2. No non-product related rejections for atleast 6 months.
3. No production related negative incidents for atleast 6 nths.
4. Should have passed a recent on-site quality system evaluation.
5. Having a fully agreed specifications.
6. Fully documented process and quality sy tem.
7. Timely copies of inspection and test d .
8. Process that is stable and in control

SUPPLIER RATING:

Supplier Rating is done

1. To obtain an overall rating of supplier performance.
2. To communicate w th suppliers regarding their performance.
3. To provide each supplier with detailed and true record of problems for corrective action
4. To enhance the relationship between the buyer and the supplier.

UNIT III - TQM TOOLS & TECHNIQUES I

SYLLABUS: The seven traditional tools of quality - New management tools - Six sigma: Concepts, Methodology, applications to manufacturing, service sector including IT - Bench marking – Reason to bench mark, Bench marking process - FMEA - Stages, Types.

The seven traditional tools of quality:

1. Pareto diagram
2. Flow diagram
3. Cause and effect diagram
4. Check sheets
5. Histogram
6. Control charts
7. Scatter diagram

PARETO DIAGRAM:

Pareto charts are used for identifying set of priorities. You can chart any number of issues/variables related to specific concern and record the number of occurrences. This way you can f gure out the parameters that have the highest impact on the specific concern. This helps you to work on the propriety issues in order to get the condition under control.

FLOW CHARTS:

This is one of the basic quality tools that can be used for analyzing a sequence of events. The tool maps out a sequence of events that take place sequentially or in parallel. The flow chart can be used to understand a complex process in order to find the relationships and dependencies between events. You can also get a brief idea about the critical path of the process and the events involved in the critical path. Flow charts can be used for any field and to illustrate events involving processes of any complexity. There are specific software tools developed for drawing flow charts, such as MS Vision

CAUSE AND EFFECT DIAGRAM:

Cause and effect diagrams (Ishikawa Diagram) are used for understanding organizational or business problem causes. Organizations face problems everyday and it is required to understand the causes of these problems in order to solve them effectively. Cause and effect diagrams exercise is usually teamwork. A brainstorming session is required in order to come up with an effective cause and effect diagram. All the main components of a problem area are listed and possible causes from each area is listed. Then, most likely causes f the problems are identified to carry out further analysis.

CHECK SHEET:

A check sheet can be introduced as the most basic tool for quality. A check sheet is basically used for gathering and organizing data. When this is done with the help of software packages such as Microsoft Excel, you can derive further analysis graphs and automate through macros available. Therefore, it is always a good idea to use a software check sheet for information gathering and organizing needs. One can always use a paper-based check sheet when the information gathered is only used for backup or storing purposes other than further processing. Types of check sheet

1. Process distribution check sheets.
2. Defective item check sheets.
3. Defect location check sheet.
4. Defect factor check sheet.

HISTOGRAM:

Histogram is used for illustrating the frequency and the extent in the context of two variables. Histogram is a chart with columns. This represents the distribution by mean. If the histogram is normal, the graph takes the shape of a bell curve. If it is not normal, it may take different shapes based on the condition of the distribution. Histogram can be used to measure something against another thing. Always, it should be two variables. Consider the following exa ple: The following histogram shows morning attendance of a class. The X-axis is the number of students and the Y-axis the time of the day.

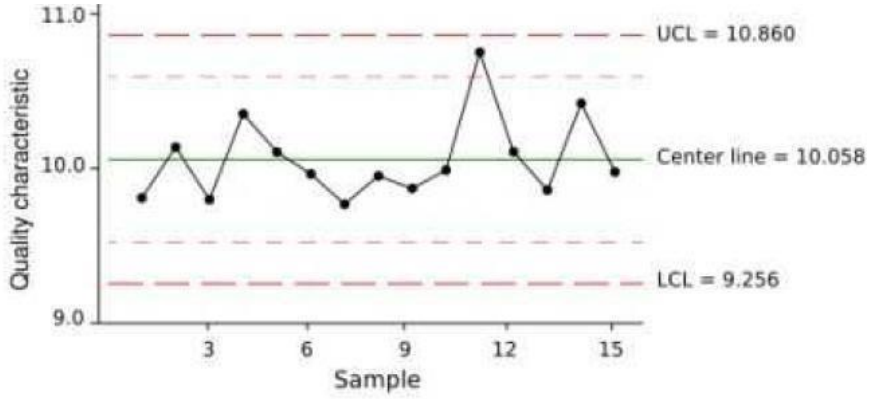
SCATTER DIAGRAM:

When it comes to the values of two variables, scatter diagrams are the best way to present. Scatter diagrams present the relationship between two variables and illustrate the results on a Cartesian plane. Then, further analysis, such as trend analysis can be performed on the values. In these diagrams, one variable denotes one axis and another variable denotes the other axis.

CONTROL CHARTS:

Control chart is the best tool for moni oring the performance of a process. These types of charts can be used for monitoring any processes related to function of the organization. These charts allow you to identify the following conditions related to the process that has been monitored.

- Stability of the process
- Predictability of the process
- Identification of common cause of variation
- Special conditions here the monitoring party needs to react



NEW SEVEN MANAGEMENT TOOLS:

These tools, unlike SPC tools are qualitative tools. Most of these tools do not involve the use of numerical data. . Like all management tools these are judgmental tools. Managers are often called upon to make decisions based on their judgement with help of incomplete information or on subjective issues. Team work and techniques like brainstorming are very essential for best results with such tools.

The seven tools we will see are :

1. Affinity diagram
2. Relations diagram
3. Tree diagram
4. Matrix diagram
5. Matrix data analysis diagram .
6 .Process decision programme chart
7. Arrow diagram

AFFINITY DIAGRAM

The purpose of an affin ty d agram is to provide a visual representation of grouping of a large number of ideas or factors or requirements into logical sets of related items to help one organise action plans in systematic manner. The steps in the procedure for preparing an affinity diagram are :

1. Decide the subject or the topic
2. Generate a large number of ideas through brainstorming
3. Decide the number of groups and their titles. Create a card for each group. Enter the title of the group at the top of the card.
4. Distribute all the ideas among the cards. If necessary, create new cards for additional groups.
5. Arrange the cards according to the relationship between the groups.
6. Give a name to the affinity diagram.

Relations Diagram:

The purpose of relations diagram is to generate a visual representation of the relations between an effect and its causes as well as the interrelationship between the causes in complex problems.

The steps in the preparation of relations diagram are:

1. Decide the ‘effect’ or the problem for which causes are to be found. Write is in the centre of the fl p chart or a board and enclose it in a dark bordered rectangle. Discuss the subject and confirm the ‘effect’.
2. Brainstorm to identify the immediate causes for the effect first. Enter these in rectangles around the central dark rectangle. Take care to place causes likely to be related to one another in adjacent positions. It is quite possible that the locations of the causes may have to be changed as one progresses. Hence a hite board is preferable to a flip chart for this exercise. If a flip chart is used, the causes may be written on post-it pads and stuck on the chart so that their location can be changed easily.

3. Connect these immediate causes to the effect by connecting the rectangles of the causes to that of the effect with a line with an arrow pointing towards the effect. Explore the cause and effect relationship among the immediate causes and connect them, keeping in mind that the arrow always points to an effect.
4. Taking each of these immediate causes as an effect, brainstorm to find causes for them one by one. The key question for identifying causes is "why ?". Keep asking the question till the root causes are identified for the immediate, secondary and tertiary causes.
5. Explore the relationship between all the causes and onnect the rectangles as in step-3. Show as many relations among different causes as possible. A large number of routes leading to the s me root causes provides an indication that the root cause may be n important contributor to the problem.

6. Brainstorm to find the more important root causes and more prominent links leading to the effect. Mark these by making the rectangles and the connecting lines darker.
7. If necessary, rearrange the rectangles in such a way that the connecting lines are short and the agram compact.
8. Provide a suitable title to the diagram.

TREE DIAGRAM:

The purpose of the tree diagram is to explore ways and means to achieve an objective, develop a list of alternate means to reach the desired situation in a sequential order and to present them in a visual form.
The steps in the procedure to develop a tree diagram are:

1. Identify a high priority problem that needs to be solved at the earliest.

2. Prepare an objective statement describing the desired situation or the target solution.
3. Decide the appropriate form of the diagram - as ade or tree as well as direction of flowCivildatasafterabriefdiscussion.Place the target solution in the dark rectangle.
4. Brainstorm to identify the primary means o achieve the objective. Arrange them in an appropriate order keeping in mind the likely interrelations between them and place them in rectangles the first level.
5. For each of the primary means, identify secondary means which would be necessary to attain those means. Arrange them in next level boxes.
6. Identify tertiary means required to attain each of the secondary means and place them in a proper order in the next level boxes.
7. Continue the process till the group feels that the end of the line has been reached.
8. If a lo er level means is required to attain two higher level means, it may be connected to both. Rearrange the boxes if necessary to make this possible. Use of POST-IT pads can make such a rearrangement simple.
9. Brainstorm to reach a consensus on the relative importance of the last level means to priorities action.
10. Give a suitable title to the diagram. Application The most important application of the tree diagram is for devising solutions for problems. It helps one to develop a systematic step by step strategy to achieve an objective. It is also useful in monitoring the implementation of solutions by taking care of accomplishment of means at different levels.

MATRIX DIAGRAM:

The purpose of a matrix diagram is to explore the existence and the extent of relations bet een individual items in two sets of factors or features or characteristics and express them in a symbolic form that is easy to understand. The purpose for which the tool is most frequently used is to understand the relation between customer expectations as expressed by the customers and product characteristics as designed, manufactured and tested by the manufacturer.

The steps in the procedure to prepare a matrix diagram are :

1. Decide the two sets of factors for which relations are required to be clarified. Call the set of the main factors ‘features’ and the set of factors dependent on it counterpart ‘characteristics’ or characteristics.

2. Divide the features into primary, secondary and tertiary features.

3. Divide the characteristics into primary, secondary and tertiary characteristics.

4. Place the features vertically on the left hand side f the matrix and characteristics horizontally on top of the matrix.

5. Enter the importance of the features on the column after that for the tertiary features.

6. In the main body of the matrix, place symbols at the squares denoting the relationship between the feature and the characteristic meeting the intersection. The symbols to be used are :
- Strong relationship
- Medium relationship
- Weak relationship

In case there is no relation between the concerned feature and characteristic, leave the square blank to indicate ‘no relation’. The relationship should be based on data available ith the team or on the results of a brainstorming session which must be confirmed by collecting necessary data.

7. Title the diagram suitably.

APPLICATIONS:

Matrix diagram, being a very simple table showing relations between individual items in two sets of factors, can be put to a wide variety of uses. The symbolic representation of the relationship makes the diagram so much easier to understand as compared to a table with a lot of figures. Let us see some of the possible applications of a matrix diagram. Matrix diagram can be used to solve problems by arranging data in such a way that the relations between relevant factors are brought into sharp focus. It can be used to understand relations between customer satisfaction and product characteristics, between omplaints and product groups, between complaints and geographical regions, between a product's performance in the market and promotion inputs on it and so on. Once the relations between individual items in se s of f ctors are clearly understood and agreed upon, it becomes easy to solve problems and to plan and implement solutions systematically.

MATRIX DATA ANALYSIS DIAGRAM:

The purpose of matrix data analysis diagram is to present numerical data about two sets of factors in a matrix form and analyse it to get numerical output. The factors most often are products and product characteristics. The purpose then is to analyse the data on several characteristics for a number of products and use the information to arrive at optimum values for the characteristics for a new product or to decide the strong points of a product and use the information for designing a strategy for the promotion of the product.

The procedure for creating a matrix data analysis diagram onsists of the following steps

1. Decide the two factorsCivildataswhoserelationsaretobeanalyses.
2. Check the number of individual items in he wo f ctors.
3. Prepare a matrix to accommodate all the ems of the two factors.
4. Enter numerical data in the matrix.
5. Give the diagram a suitable title.

PROCESS DECISION PROGRAMME CHART:

The purpose of process decision programme chart is to prepare for abnormal occurrences with low probability which may otherwise be overlooked and to present the occurrences as well as the necessary countermeasures to guard against such occurrences in the form of a visual chart. The toolcomforcesone to think of the possible obstacles in the smooth progress of a process or a project and then find ways and means to surmount those obstacles to ensure the successful and timely completion of the process or the project. Thus the t l helps one to prepare a contingency plan to achieve the objective if adverse events cur.

The steps in the preparationCivildatasofprocessdecisionprogramme chart are :

1. Prepare a ‘normal’ flowchart of the process with all expected events as steps in the chart.
2. Consider the possibility of the process not going as per the plan due to any abnormal, though less probab e, occurrences.
3. Show these occurrences on the flowchart through branching at appropriate locations.
4. Consider how the abnormal occurrence will affect the process and search for ways and means to counter the effect.
5. Show these countermeasures in rectangles connecting the corresponding abnormal occurrence on one side and the process objective or the goal on the other.
6. Give a suitable title to the diagram.

ARROW DIAGRAM:

The purpose of an arrow diagram is to cre te a visual presentation of the steps of a process or tasks necessary to comple e a project with special emphasis on the time taken for these activities. The di gram provides a clear understanding of the schedule of various steps in the process which helps one to monitor the process for ensuring its complet on on time. The steps for preparing an arrow diagram are:

1. List all tasks or act ties that need to be accomplished before the completion of the process or the project.
2. Decide which steps are undertaken in series and which steps can be run in parallel.
3. Arrange the activities in a proper sequence.
4. Prepare 'Event Nodes' at the completion of steps and number them. Where the process is bifurcating into two or more parallel streams, more lines will flow from a node and where the parallel streams are merging, two or more steps will lead to a node.
5. Write the description of the step on top of the line or to the left of the line. Decide the time required for completing each step and write it under or to the right of the line.
6. Calculate the earliest time to reach an event node for the start of the process. Where more than one streams are combining, the maximum time taken by a stream is taken into consideration. This time is entered on the top half of the rectangle. This time is related to the starting time of the process which is taken as zero.
7. After the time for all event nodes including the ompletion of the process or the project is available,Civildatasonecalculatesthelatet time by which an event node must be reached. This is done by st rting t the time of completion and going back step by step. The time is entered on the bottom half of the rectangle. The time indication t ll event nodes will appear as : X Y where X is the earliest time by which the event can be completed and Y is the latest time by which the event shou be completed.
8. Give a title to the diagram. As the calculation of the time indications is extremely important n the construction of an arrow diagram it is necessary that we understand the procedure well. Let us understand the concept through diagram.

SIX SIGMA:

Six sigma stands for six standard deviation from mean (sigma is the Greek letter used to represent standard deviation in statistics). The objective of six sigma principle is to achieve zero defects products/process. It allows 3.4 defects per million opportunities.

DMAIC – It is used for improving existing processes/products.

DMADV – It is applied to new processes/products.

SIX SIGMA PROJECT METHODOLOGY:

DMAIC (Define)

- Define (What is important?)
- Base-lining and benchmarking processes
- Decomposing processes into sub-processes
- Specifying customer satisfaction goals/sub-goals

(requirements) Support tools for Define step:

- Benchmarking
- Baseline
- Voice of Customer (Win Win)
- Voice of Business (Win Win)
- Quality Function Deployment & etc.

DMAIC (Measure)

- Measure (How are we do ng?)

1. Identifying relevant metrics based on engineering principles and models
2. Performance measurement: throughput, quality (statistically, mean and variation)
3. Cost (currency, time, and resource)
4. Other example of measurement: response times, cycle times, transaction rates, access frequencies, and user defined thresholds

Support tools for Measure step:

Basic tools : Flow chart, Check Sheets, Pareto diagrams, Cause/Effect diagrams, Histograms, and Statistical Process Control (SPC).
Defect Metrics
Data Collection Forms, Plan, Logistics

DMAIC (Analyze)

- Analyze (What‟s wrong?)

Evaluate the data/information for trends, patterns, causal relationships and "root cause"
Example: Defect analysis, and Analysis of variance Determine candidate

improvements

- Support tools for Analyze step:

  Cause/Effect diagram
  Failure Modes & Effects
  Analysis Decision & Risk
  Analysis Statistical Inference
  Control Charts
  Capability Analysis and etc.

- DMAIC (Improve)
- Improve (What needs to be done?)

Making prototype or initial improvement
Measure and compare the results with the simulation results

Iterations taken between Measure-Analyze-Improve steps to achieve the target level of performance

Support tools for Improve step:

Design of Experiments
Modeling
Tolerancing
Robust Design
DMAIC (Control)
Control (How do we guarantee performance?)

Ensuring measurements are put into place to maintain improvementsSupport tools for Control step:
Statistical Controls: Control Charts, Time Series methods Non-Statistical Controls: Procedural adherence, Performance Mgmt., Preventive Activities.

BENCHMARKING:

Benchmarking is a systematic method by which organizations can measure themselves against the best industry practices. Benchmarking is a systematic search for the best practices, innovative ideas, and highly effective operating procedures.

BENCHMARKING CONCEPT:

REASONS TO BENCHMARK:

1. It is a tool to achieve business and competitive objectives.
2. It can inspire managers (and Organizations) to compete.
3. It is time and cost effective.
4. It constantly scans the external environment to improve the process.
5. Potential and useful technological breakthroughs can be located and adopted early.

PROCESS OF BENCHMARKING:

1. Decide what to benchmark
2. Understand current performance
3. Plan
4. Types of benchmarking
5. Study Others
6. Learn from the Data.

Decide what to benchmark:

1. Benchmarking can be applied to any business or produ tion process.
2. The strategy is Civildatasusuallyexpressedintermsofmi ion and vision statements.
3. Best to begin with the mission and critical factors.
4. Choosing the scope of the Benchmarking s udy.
5. Pareto analysis – what process to investigate.
6. Cause and Effect diagram – for tracing outputs back.

Understand current performance:

1. Understand and document the current process.
2. Those working n the process are the most capable of identifying and correcting problems.
3. While documenting, it is important to quantify.
4. Care should be taken during accounting information.

Plan:

1. A benchmarking team should be chosen.
2. Organizations to serve as the benchmark need to be identified.
3. Time frame should be agreed upon for each of the benchmarking tasks.

Types of benchmarking:

1. Internal
2. Competitive
3. Process

Study Others:

Benchmarking studies look for two types of information

- How best the processes are practiced
- Measurable results of these practices

Three techniques for conducting the research are

- Questionnaires
- Site visits
- Focus groups

Learn from the data:

- What is the gap? How much is it?
- Why is there a gap? What does the best-in-class do differently that is better?
- If best-in-class practices were adopted, what would be the resulting improvement?

Benchmarking studies can reveal three different outcomes

- Negative gap
- Parity
-

Positive gap

SIGNIFICANCE:

1. Benchmarking is a systematic method by which organizations can measure themselves against the best Industry practices
2. It promotes superior performance by providing an organized framework through which organization learn how the " best in class" do things.
3. It helps for continuous improvement.
4. Benchmarking inspire managers (and organization) to compete.
5. Through Benchmark proces organization can borrow ideas, adopt and refine them to gain competitive advantages.

FAILURE MODE AND EFFECTS ANALYSIS:

Failure mode and effect analysis also known as risk analysis is a preventive measure to systematically display the causes, effects, and possible actions regarding observed failures.

OBJECTIVES OF FEMA:

1. The objective of FEMA is to anticipate failures and prevent them from occurring. FEMA prioritizes failures and attempts to eliminate their causes.
2. FEMA is an engineering technique is used to define, identify and eliminate known and or potential failures, problems, errors which occur in the system, design, process and service before they reach the customer.
3. FEMA is a before the event action and is done when existing systems products processes are changed or redesigned.
4. FEMA is a never ending process improvement tool.

TYPES OF FEMA:

1. System FEMA
2. Design FEMA
3. Process FEMA
4. Service FEMA
5. Equipment FEMA
6. Maintenance FEMA
7. Concept FEMA
8. Environmental FEMA

BENEFITS OF FEMA:

1. Improve product/process reliability and quality.
2. Increase customer satisfaction.
3. Early identification and elimination of potential product/process failure modes.
4. Prioritize product or process deficiencies
5. Capture engineering/organization knowledge
6. Document and track the actions taken to reduce risk
7. Provide focus for improved testing and development.
8. Minimize late changesCivildatasandassociatedco.
9. Act as catalyst for teamwork and idea exch nge between functions.

STAGES OF FEMA:

1. Specifying possibilities
    a. functions
    b. possible failure modes
    c. root causes
    d. effects
    e. detection/prevent on

2. Quantifying risk
    a. probability of cause
    b. severity of effect
    c. effectiveness of control to prevent cause.
    d. risk priority number.

3. Correcting high risk causes
    A. prioritizing work
    B. detailing action
    C. assigning action responsibility.
    D. checks points on completion.
4. Re-evaluation of risk
5.Recalculation of risk priority number

UNIT IV - TQM TOOLS & TECHNIQUES II

SYLLABUS: Control Charts - Process Capability - Concepts of Six Sigma - Quality Function Development (QFD) - Taguchi quality loss function - TPM - Concepts, improvement needs - Performance measures.

QUALITY FUNCTION DEPLOYMENT:

It is kind of conceptual map that provides a means of interfunctional planning and communication.

Ultimately the goal of QFD is to translate often subjective quality criteria

into objective ones. That can be quantified and measured and which can then be used to design and manufactureCivildatastheproduct.Itiscomplimentary method for determining how and where priorities are to be ssigned in product development.

BENEFITS OF QFD:

1. Improves Customer satisfaction
2. Reduces Implementation Time
3. Promotes Team Work
4. Provides Documentation

HOUSE OF QUALITY:

THE STEPS IN BUILDING A HOUSE OF QUALITY ARE:

1. List Customer Requirements (WHAT's)
2. List Technical Descriptors (HOW's)
3. Develop a Relationship Matrix Between WHAT's and HOW's
4. Develop an Inter-relationship Matrix between HOW's
5. Competitive Assessments
    a. Customer Competitive Assessments
    b. Technical Competitive Assessments

6. Develop Prioritized Customer Requirements
7. Develop Prioritized Technical Descriptors

Phase 1: product planning

- Step1: list customer requirements
- Step2: List technical descriptors
- Step3: Develop a relationship between WHATS AND HOWS
- Step4: Develop a interrelationship matrix between HOWS
- Step5: Do competitive assessments
- Step6: Develop prioritized customer requirements
- Step7: Develop prioritized technical descriptors.

Phase 2: part development

- Step8: Deploy QFD process down to sub-components level both in terms of requirements and characteristics.
- Step9: Deploy the component deployment ch rt. Relate the critical sub-component control characteristics.

Phase 3: process planning

- Step10: Develop the relationship between the critical characteristics and process used to create the characteristics
- Step11: Develop the control plan relating critical control to critical processes.

Phase 4: production planning

- Step 12: Tabulate operating instructions from process requirements
- Step13: develop prototype and do testing
- Step14: Launch the final product to the market.

TAGUCHI'S QUALITY LOSS FUNCTIONS:

Taguchi has defined quality as the loss imparted to society from the time a product is shipped. Societal losses include failure to meet customer requirements, failure to meet ideal performance and harmful side effects.

TAGUCHI LOSS FUNCTION CURVE

TAGUCHI LOSS FUNCTION CURVE

There are three common quality loss functions.

1. Nominal - the - best.
2. Smaller - the - better.
3. Larger - the - better.

Nominal the best:

Although Taguchi developed so many loss functions, any situations are approximated by the quadratic function which is called the **No inal – the – best** type.

The quadratic function is shown in figure. In this situation, the loss occurs as soon as the performance characteristic, y, departs from the target τ. At τ, the loss is Rs 0
At LSL (or) USL, the loss is Rs. A.

The quadratic loss function is described by the equation $L = k (y - \tau) 2$.
Where,
L = cost incurred as quality deviates from the target.

y = Performance characteristic

τ = target

k = Quality loss coefficient.

The loss coefficient is determined by setting $\Delta = (y - \tau)$, the deviation from the target. When Δ is the USL (or) LSL, the loss to the customer of repairing (or) discarding the product is Rs. A.

Thus, $K = A / (y - \tau)2 = A / \Delta 2$.

Smaller – the – better:

The following figure shows the smaller – the – better concepts. The target value for **smaller – the – better** is 0. There are no negative values for the performance characteristic.

Larger – the – better:

In the Larger – the – better concept, the target value is ∞ (infinity), which gives a **zero loss**. There are no negative values and the worst case is at $y = 0$. Actually, larger – the – better is the reciprocal of smaller – the – better. The performance characteristics in Larger – the – better are bond strength of adhesives, welding strength etc.

TOTAL PRODUCTIVE MAINTENANCE:

Total Productive Maintenance (TPM) is an important and effective tool for the excellence. Total productive maintenance (TPM) is keeping the current plant and equipment at its highest productivity level through cooperation of all areas of the organization.

PRINCIPLES OF TPM:

1. Use overall equipment effectiveness as a compass for success.
2. Improve existing planned maintenance system.
3. Work towards zero loss.
4. Provide training to upgrade operation and maintenance skills.
5. Involve everyone and use cross-functional teams.

OBJECTIVES OF TPM:

1. To maintain and improve equipment capacity.
2. To maintain equipment for life.
3. To use support from all areas of the operation
4. To encourage input from all employees.
5. To use teams for continuous improvement.

TPM PHILOSOPHY – CONCEPT OF TPM:

Total Productive Maintenance (TPM) is an extension of the Total Quality Management (TQM) philosophy to the maintenance function.

TPM has the following steps:

1) Management should learn the new philosophy of TPM.
2) Management should promote the new philosophy of TPM.
3) Training should be funded and developed for everyone in the organization.
4) Areas of needed improvement should be identified.
   Loss measurements to identify improvement needs are
   Down time losses
   Reduced speed losses
   Poor quality losses
5) Performance goals should be formulated.
6) An implementation plan should be developed.
7) Autonomous worth groups should be established.

TPM PILLAR:

PILLAR -1 - JISHU HOZEN (Autonomous maintenance) :

This pillar is geared towards developing operators to be able to take care of small maintenance tasks, thus free ng up the skilled maintenance people to spend time on more value added acti ity and technical repairs. The operators are responsible for upkeep of the r equipment to prevent it from deteriorating.

PILLAR -2 – KOBETSU KAIZEN :

"Kai" means change, and "Zen" means good (for the better). Basically kaizen is for small improvements, but carried out on a continual basis and involve all people in the organization. Kaizen is opposite to big spectacular innovations. Kaizen requires no or little investment. The principle behind is that "a very large number of small improvements are move effective in an organizational environment than a few improvements of large value. This pillar is aimed at reducing losses in the workplace that affect our efficiencies. By using a detailed and thorough procedure we eliminate losses in a systematic method using various Kaizen tools.

PILLAR -3 - PLANNED MAINTENANCE :

It is aimed to have trouble free machines and equipments producing defect free products for total customer satisfaction. This breaks maintenance down into 4 "families" or groups which was defined earlier.

1. Preventive Maintenance
2. Breakdown Maintenance
3. Corrective Maintenance
4. Maintenance Prevention

With Planned Maintenance we evolve our efforts from a reactive to a proactive method and use trained maintenance aff o help train the operators to better maintain their equipment.

PILLAR -4 – Hinshitsu Hozen or QUALITY MAINTENANCE :

It is aimed towards customer de ight through highest quality through defect free manufacturing. Focus is on el minating non conformances in a systematic manner, much like Focused Impro ement. We gain understanding of what parts of the equipment affect product quality and begin to eliminate current quality concerns, then move to potential quality concerns. Transition is from reactive to proactive (Quality Control to Quality Assurance). QM activities is to set equipment conditions that preclude quality defects, based on the basic concept of maintaining perfect equipment to maintain perfect quality of products. The condition are checked and measure in time series to very that measure values are within standard values to prevent defects. The transition of measured values is watched to predict possibilities of defects occurring and to take counter measures before hand.

PILLAR – 5: Development Management / Early Management:

Early management or development management helps in drastically reducing the time taken to receive, install, and set – up newly purchased equipments (known as vertical start – up). Early management can also be used for reducing the time to manufacture a new product in the factory.

PILLAR 6 – TRAINING and EDUCATION:

It is aimed to have multi-skilled revitalized employees whose orale is high and who has eager to come to work and perform all required functions effectively and independently. Education is given to operators to upgrade their skill. It is not sufficient know onlyCivildatas"Know-How"bytheyhould also learn "Know-why". By experience they gain, "Know-How" to overcome problem what to be done.
This they do without knowing the root cause of the problem and why they are

doing so. Hence it become necessary to tr in them on knowing "Know-why".

PILLAR- 7: SAFETY, HEALTH AND ENVIRONMENT
Target :

1. Zero accident,
2. Zero health damage
3. Zero fires.

In this area focus is on to create safe workplace and a surrounding area that is not damaged by our process or procedures. This pillar will play an active role in each of the other pillars on a regular basis. A committee is constituted for this pillar which comprises representative of officers as well as workers. The committee is headed by Senior vice President (Technical). Utmost importance to Safety is given in the plant. Manager (Safety) is looking after functions related to safety. To create a areness among employees various competitions like safety slogans, Quiz, Drama, Posters, etc. related to safety can be organized at regular intervals.

PILLAR -8 : OFFICE TPM

Office TPM should be started after activating four other pillars of TPM (JH, KK, QM, PM). Office TPM must be followed to improve productivity, efficiency in the administrative functions and identify and eliminate losses. This includes analyzing processes and procedures towards increased office automation. Office TPM addresses twelve major losses. They are

1. Processing loss
2. Cost loss including in areas such as pr curement, accounts, marketing, sales leading to high inventories
3. Communication loss
4. Idle loss
5. Set-up loss
6. Accuracy loss
7. Office equipment break own
8. Communication channel breakdown, telephone and fax lines
9. Time spent on retr eval of information
10. Non ava lab ty of correct on line stock status
11. Customer complaints due to logistics
12. Expenses on emergency dispatches/purchases

PERFORMANCE MEASURES:
Performance measures are required for the managers for managing an organization perfectly.

Performance measures are used to achieve the following objectives.

To establish performance measures and reveal trend.

1. To identify the processes to be improved.
2. To determine the process gains and losses.
3. To compare the actual performance with standard performance.
4. To provide information for individual and team evaluation.
5. To determine overall performance of the organization.
6. To provide information for making proper decisions.

WHAT SHOULD BE MEASURED?

Human resources

1. Lost time due to accidents, absenteeism.
2. Employee turnover.
3. Employee satisfaction index.
4. Training cost per employee.
5. Number of grievances.

Customers

1. Number of complaints from customers.
2. Number of on-time deliveries.
3. Warranty data.
4. Dealer satisfaction.

Production

1. Inventory.
2. SPC Charts.
3. Amount of scrap / rework.
4. Machine do n time.

Research and Development

a. New product time to market.
b. Design change orders.
c. Cost estimating errors.

Suppliers

1. On-time delivery.
2. Service rating.
3. Quality performance.
4. Average lead time.

Marketing / Sales

1. Sales expense to revenue.
2. New product sales to total sales.
3. New customers.

Administration

1. Revenue per employee.
2. Purchase order error.
3. Billing accuracy.
4. Cost of poor quality.

PERFORMANCE MEASURE PRESENTATION:

There are six basic techniques for presenting performance measures. They are

1. Time series graph.
2. Control charts.
3. Capability Index.
4. Taguchi's loss function.
5. Cost of poor quality.
6. Malcolm Baldrige National Quality Award.

UNIT-V QUALITY SYSTEMS

SYLLABUS: Need for ISO 9000 - ISO 9001-2008 Quality System - Elements, Documentation, Quality Auditing - QS 9000 - ISO 14000 - Concepts, Requirements and Benefits – TQM Implementation in manufacturing and service sectors.

QUALITY SYSTEM:

In order to assure the quality of a product, the manufacturer ust ensure its quality. So, to ensure this quality it is necessary to make a systematic study and

control check at every stage of production. It is also essential to take critical review of efforts and achievementsCivildatasofthecompanywith re pect to the quality of the product. Thus it is necessary to develop a st nd rd quality system.

ISO 9000 STANDARDS:

The ISO 9000 system is quality m n gement system that can be adopted by all types of organizations belonging to government, public, private, (or) joint sectors. The ISO 9000 system shows the way in creating products by preventing deficiencies, instead of conduct ng expensive post product inspections and rework.

ISO 9000

a. ISO 9001
b. ISO 9002
c. ISO 9003

ISO 9001

Design, Development, Production, Installation & Servicing

ISO 9002

Production, Installation & Servicing

ISO 9003

Inspection & Testing

ISO 9004

Provides guidelines on the technical, administrative and human factors affecting the product or services.

BENEFITS OF ISO 9000 STANDARDS:

a. Achievement of international standard of quality.
b. Value for money.
c. Customer satisfaction.
d. Higher productivity.
e. profitability
f. Improved corporate image
g. Access to global market
h. Growth of the organization
i. Higher morale of employees

CLAUSES (ELEMENTS) OF ISO 9000:

1. Scope
2. Normative Reference
3. Terms and Definitions
4. Quality Management System (QMS)
    4.1 General Requirements
    4.2 Documentation

5. Management Responsibility

    5.1 Management Commitment
    5.2 Customer Focus
    5.3 Quality Policy
    5.4 Planning
    5.5 Responsibility, Authority and Communication
    5.6 Management Review

6. Resource Management
    6.1 Provision of Resources
    6.2 Human Resources
    6.3 Infrastructure
    6.4 Work Environment

7. Product Realization
    7.1 Planning of Product Realization
    7.2 Customer related processes
    7.3 Design and Development
    7.4 Purchasing
    7.5 Production and Service Provision

7.6 Control of Monitoring and Measuring devices

8. Monitoring and Measurement

8.1 General
8.2 Monitoring and Measurement
8.3 Control of Non-Conforming Product
8.4 Analysis of Data
8.5 Improvement

ISO 9000:2000 Quality Systems:

The term I S O 9000 refers to a set of quality management standards. ISO 9000 currently includes three quality standard : ISO 9000:2000, ISO 9001:2000, and ISO 9004:2000. ISO 9001:2000 presents requirements, while ISO 9000:2000 and ISO 9004:2000 present guidelines. ISO's purpose is to facilitate international trade by providing a single set of st nd rds that people everywhere would recognize and respect. The ISO 9000 2000 Standards apply to all kinds of organizations in all kinds of areas. Some of these areas include manufacturing, processing, servicing, printing, forestry, electronics, steel, computing, legal services, financial serv ces, accounting, trucking, banking, retailing, drilling, recycling, aerospace, construction, exploration, textiles, pharmaceuticals, oil and gas, pulp and paper, petrochemicals, publishing, shipping, energy, telecommunications, plastics, metals, research, health care, hospitality, utilities, pest control, aviation, machine tools, food processing, agriculture, government, education, recreation, fabrication, sanitation, software development, consumer products, transportation, design, instrumentation, tourism, communications, biotechnology, chemicals, engineering, farming, entertainment, horticulture, consulting, insurance, and so on.

ISO 9000 is important because of its orientation. While the content itself is useful and important, the content alone does not account for its widespread appeal. ISO 9000 is important because of its international orientation. Currently, ISO 9000 is supported by national standards bodies from more than 120 countries. This makes it the logical choice for any organization that does business internationally or that serves customers who demand an international standard of quality. ISO is also important because of its systemic orientation. We think this is crucial. Many people in this field wrongly emphasize motivational and attitudinal factors. The assumption is that quality can only be created if workers are motivated and have the right attitude. This is fine, but it doe n't go far enough. Unless you institutionalize the right attitude by supporting it with the right policies, procedures, records, technologies, resources, nd structures, you will never achieve the standards of quality that other org niz tions seem to be able to achieve.

ISO 9000 DOCUMENTATION

STRUCTURE

The documentation created for ISO 9000 registration is submitted to the company's 3rd-party reg strar prior to them visiting the site to conduct the actual audit. In fact, one type of documentation is used by the registrar to develop the audit plan for your company. Structuring your ISO 9000 documentation to facilitate the audit process only serves to enhance the potential for a successful audit. This structuring will also make it easy for you to plan and monitor your documentation efforts, both for the registration audit and all subsequent maintenance audits.

DOCUMENT CONTROL AND ISO 9000

Once the documentation structure has been defined and the documentation written, a strategy for controlling it must be put in place. ISO 9000 requires that documentation must be readily available to those who need it, be of current issue, and that all obsolete material be completely removed from the system. The control of documentation, from creation of new material through to the destruction of obsolete material, presents one of ISO 9000's biggest challenges. It is also one of the elements audited by your 3rd-party registrar.

DOCUMENTING ISO 9000

A thorough analysis of each element prior to writing ensures the resulting documentation will meet ISO 9000's criteria. Specific characteristics exist for robust Quality Systems, and these must be clearly established within the organization. Since ISO9000 registration is not a one time occurrence, clearly documented procedures for maintaining compliant Quality System must be in place. Historically, companies have produced policy and procedure manuals which, because they contained corpor te policies, where often not made available to all employees.

As a result, the procedures were also not readily available. ISO 9000's requirement that procedures be read ly available to all persons performing the work usually necessitate the separation of these procedures from the policy manual Perhaps the biggest stumbl ng block for North American businesses is the requirement to clearly define and document the processes that it uses. Developing documentation that tells HOW we do something is not new to us, but accurately describing WHAT it is we do is far less common. Most of our existing documentation is product or department based. ISO looks only at the processes used to create products, and these generally run across many areas of an

organization. We can no longer write documentation in isolation, the whole organization must be considered when writing ISO compliant documentation.

WHEN IS ENOUGH, ENOUGH?

One of the complaints often heard about ISO 9000 refers to the large amount of documentation that is perceived to be required. While procedural documentation is important to the proper functioning of an effective Quality System, many companies tend to over document. First and foremost, you must remember that it is your company and the documentation must fit the company, not the standard.
The ISO 9000 series of Quality Standards does indicate key characteristics of a properly functioning Quality System, but how they are implemented is the responsibility of the organization. ISO documentation must reflect what the company does, not what it thinks the ISO audit or will want to hear. In determining whether procedural documentation is required, look at the skill sets of the people performing the task as well as any unique requirements the company may have for completing the task. In many cases, documen a ion will not be required because there is no unique process and/or the person has been trained in how to complete the task.

CLAUSES IN ISO 9001:

ISO 9001 defines 20 elements necessary for quality management system, as listed below:

Management Responsibility (Element 1)

The company has to define its commitment to a quality policy, which is understood, implemented and maintained at all levels of the organization, and to define its quality goals. Responsibilities and authorities have to be defined and documented. The company must provide adequate resources and appoint a member of the management as a representative for quality management. At least once a year, a management review must be held and recorded to evaluate the quality system.

Quality System (Element 2)

A quality manual, covering all elements of the ISO standard, has to be prepared to document the quality system. Procedures must be documented and controlled. The company has to prepare a quality plan to ensure that quality requirements are understood and fulfilled.

Contract Review (Element 3)

The company has to establish and maintain docu ented procedures for contract review, to document the customers' requirements and ensure the capability to fulfill the contract or order requirements. Records of ontract review shall be maintained.

Design Control (Element 4)

The company has to establish and m int in documented procedures to control and verify the design of new product or service to fulfill customers' requirements. The requirements must be identified and there must be design reviews, design verification and design validation. Design changes shall be documented, reviewed and author zed.

Document Control (Element 5)

All documents relevant for quality have to be controlled to ensure that the pertinent issues of appropriate documents are available at all locations. When necessary, they are to be replaced by updated versions. Changes shall be reviewed and approved by the same organization/person that performed the original review or approval.

Purchasing (Element 6)

The company must monitor the flow of purchasing and evaluate the subcontractor's ability to fulfill specified requirements.

Purchaser Supplied Product (Element 7)

Goods supplied by the customer have to be recorded. It must be ensured that they are separately controlled and stored to prevent loss or damage.

Product Identification And Traceability (Element 8)

Where appropriate, purchased and delivered products or services must be made traceable through documentation or batches.

Process Control (Element 9)

All processes of production or service that directly affect quality must be documented and planned and carried out under ontrolled conditions to add consistency to the process. Control of proce parameters and product characteristics must ensure that the specified requirements are met.

Inspection And Testing (Element 10)

The company must ensure receiving inspection and testing, in-process inspection and testing, and final inspection and testing. These inspections and tests must be recorded. Control of inspection, measuring and

Test Equipment (Element 11)

The items of equ pment used for inspection, measuring and testing must be identified and recorded. They must be controlled, calibrated and checked at prescribed intervals.

Inspection And Test Status (Element 12)

The status of the product or service must be identified at all stages as conforming or nonconforming. This is to ensure that only conforming products or services are dispatched or used

Control Of Nonconforming Product (Element 13)

The company must establish procedures to ensure that nonconforming products or services are prevented from unintended use. The disposal of nonconforming products must be determined and recorded.

Correctional Prevention (Element 14)

Procedures must be established to ensure effective handling of customer complaints and corrective actions after identifying nonconformities. The cause of nonconformities is to be investigated in order to prevent recurrence. The corrective action shall be monitored to ensure its long-term effectiveness. Preventive actions are to be initiated to eliminate potential causes of nonconfor ance. Handling, storage, packaging and

Delivery (Element 15)

Documented procedures must be established to ensure that products are not damaged and reach the customer in the required condition

Control Of Quality Records (Element 16)

All records related to the quality sys em must be identified, collected and stored together. The quality records demonstrate conformity with specified requirements and verify effective operation of the quality system.

Internal Quality Audits (Element 17)

The company must establ sh and maintain documented procedures for planning and implement ng nternal quality audits to determine the effectiveness of the quality system. The comments made by internal auditors must be recorded and brought to the attention of the personnel having responsibility in the area audited. Follow-up audit activities shall verify and record the implementation and effectiveness of the corrective action taken.

Training (Element 18)

The company shall establish and maintain documented procedures for identifying training needs and must have a training record for each employee.

Servicing (Element 19)

Where servicing is a specific requirement, the company must establish and maintain documented procedures for performing, verifying and reporting that the servicing meets the specified requirements.

Statistical Techniques (Element 20)

The company must establish and maintain docu ented procedures to implement and control the application of statistical techniques which have been identified as necessary for performance information. This structure looks very theoretical at first glance, but this is because ISO 9000 stipulates the elements of a quality management Civildatassystemforanyenterprie,irre pective of its branch of activity. "ISO 9000 is not a prescriptive st nd rd; it does not detail the how but rather the what. This allows each individu l comp ny to define how it intends to comply with the standard in way th t best suits that company's method of operation". It is possible that some of the elements are of no relevance or almost no relevance in specific sectors.

IMPLEMENTATION OF QUALITY MANAGEMENT SYSTEM:

1. Top Management Commitment
2. Appoint the Management Representative
3. Awareness
4. Appoint an Implementation Team
5. Training
6. Time Schedule
7. Select Element Owners
8. Review the Present System
9. Write the Documents
10. Install the New System
11. Internal Audit
12. Management Review
13. Pre-assessment
14. Registration

PITFALLS OF SUCCESSFULIMPLEMENTATION:

1. Using a generic documentation progr m or nother organization's documentation program
2. Over-documentation or document tion that is too complex
3. Using External Consultants without involvement
4. Neglecting to obtain top management's involvement
5. Developing a system that does not represent what actually occurs

QUALITY AUDITING:

The term Audit refers to regular examination and checking of accounts or financial records, settlement or adjustment of accounts. It also refers to checking, inspection and examination of Production Processes.

PURPOSE OF QUALITY AUDIT:

1. To establish the adequacy of the system.
2. To determine the effectiveness of the system.
3. To afford opportunities for system analysis.
4. To help in problem solving.
5. To make decision making easier etc.

PES OF QUALITY AUDIT:

1. First – Party Audit.
2. Second – Party Audit.
3. Third – Party Audit.

An internal audit (first – party audit) is conducted by personnel within the organization. An external audit is conducted by people from the organization such as the purchasing party (second – party audit) (or) a certified auditing agency (third – party audit).

ISO 14000 STANDARDS:

ISO 14000 standard gives the company background on which to base its Environmental Management System (EMS). This system can be joined with other quality standards and can be implemen ed oge her to achieve the organizations environmental targets. The overall aim of the system is to provide protection to environment and to prevent pollution.

REQUIREMENT OF ISO 14001

There are six elements

1. GENERAL REQUIREMENTS

EMS should include policy, planning implementation & operation, checking & corrective action, management review.

2. ENVIRONMENTAL POLICY (Should be based on mission)

1. The policy must be relevant to the organization's nature.
2. Management's Commitment (for continual improvement & preventing pollution).
3. Should be a framework (for Environmental objectives & Targets).
4. Must be Documented, Implemented, & Maintained.

3. PLANNING

1. Environmental Aspects
2. Legal & other Requirements
3. Objectives & Targets
4. Environmental Management Programs

4. IMPLEMENTATION & OPERATION

1. Structure & Responsibility
2. Training, Awareness & Competency
3. Communication
4. EMS Documentation
5. Document Control
6. Operational Control
7. Emergency Preparedness & Response

5. CHECKING & CORRECTIVE ACTION

1. Monitoring & Measuring
2. Nonconformance & Corrective & Preventive action
3. Records
4. EMS Audit

6. MANAGENMENT REVIEW

1. Review of objectives & targets
2. Review of Environmental performance against legal & other requirement
3. Effectiveness of EMS elements
4. Evaluation of the continuation of the policy

BENEFITS OF ENVIRONMENTAL MANAGEMENT SYSTEM :

GLOBAL BENEFITS

1. Facilitate trade & remove trade barrier
2. Improve environmental performance of planet earth
3. Build consensus that there is need for environmental management and a common terminology for EMS

ORGANIZATIONAL BENEFITS

1. Assuring customers of commitment o environmental management
2. Meeting customer requirement
3. Improve public relation
4. Increase investor satisfaction
5. Market share increase
6. Conserving input material & energy
7. Better industry/government relation
8. Low cost insurance, easy attainment of permits & authorization

TQM IN MANUFACTURING:

Quality assurance through statistical methods is a key component in a manufacturing organization, where TQM generally starts by sampling a random selection of the product. The sample can then be tested for things that matter most to the end users. The causes of any failures are isolated, secondary measures of the production process are designed, and then the causes of the failure are corrected. The statistical distributions of important measurements are tracked. When parts' measures drift into a defined "error band", the process is fixed. The error band is usually a tighter distribution than the "failure band", so that the production process is fixed before failing parts can be produced.

It is important to record not just the me surement ranges, but what failures caused them to be chosen. In that way, che per fixes can be substituted later (say, when the product is redesigned) with no loss of quality. After TQM has been in use, it is very common for parts to be re esigned so that critical measurements either cease to exist, or become much wider.

Often, a "TQMed" product s *cheaper* to produce because of efficiency/ performance improvements and because there's no need to repair dead-on-arrival products, which represents an mmensely more desirable product.

****

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COMMON FOR: DEPARTMENT OF INFORMATION TECHNOLOGY

CS8791 – CLOUD COMPUTING

R – 2017

LECTURE NOTES

CS8791 - CLOUD COMPUTING

SYLLABUS

UNIT I INTRODUCTION

Introduction to Cloud Computing –Definition of Cloud –Evolution of Cloud Computing –Underlying Principles of Parallel and Distributed Computing –Cloud Characteristics –Elasticity in Cloud –On-demand Provisioning.

UNIT II CLOUD ENABLING TECHNOLOGIES

Service Oriented Architecture – RESTful Systems – Web Services – Publish-Subscribe Model – Basics of Virtualization – Types of Virtualization – Implementation Levels of Virtualization – Virtualization Structures – Tools and Mechanisms – Virtualization of CPU – Memory – I/O Devices – Virtualization Support and Disaster Recovery.

UNIT III CLOUD ARCHITECTURE, SERVICES AND STORAGE

Layered Cloud Architecture Design – NIST Cloud Computing Reference Architecture – Public, Private and Hybrid Clouds - laaS – PaaS – SaaS – Architectural Design Challenges – Cloud Storage – Storage-as-a-Service – Advantages of Cloud Storage – Cloud Storage Providers – S3.

UNIT IV RESOURCE MANAGEMENT AND SECURITY IN CLOUD

Inter Cloud Resource Management – Resource Provisioning and Resource Provisioning Methods – Global Exchange of Cloud Resources – Security Overview – Cloud Security Challenges – Software-as-a-Service Security – Security Governance – Virtual Machine Security – IAM – Security Standards.

UNIT V CLOUD TECHNOLOGIESAND ADVANCEMENTS

Hadoop – MapReduce – Virtual Box -- Google App Engine – Programming Environment for Google App Engine – Open Stack – Federation in the Cloud – Four Levels of Federation – Federated Services and Applications – Future of Federation.

REFERENCES

1. Kai Hwang, Geoffrey C. Fox, Jack G. Dongarra, "Distributed and Cloud Computing, From Parallel Processing to the Internet of Things", Morgan Kaufmann Publishers, 2012
2. Rittinghouse, JohnW., and James F. Ransome, "Cloud Computing: Implementation, Management and Security", CRC Press, 2017.
3. Rajkumar Buyya, Christian Vecchiola, S. ThamaraiSelvi, "Mastering Cloud Computing", Tata Mcgraw Hill, 2013.
4. Toby Velte, Anthony Velte, Robert Elsenpeter, "Cloud Computing -A Practical Approach", Tata Mcgraw Hill, 2009.
5. George Reese, "Cloud Application Architectures: Building Applications and Infrastructure in the Cloud: Transactional Systems for EC2 and Beyond (Theory in Practice)", O'Reilly, 2009
6. Fang Liu, Jin Tong, Jian Mao, Robert Bohn, John Messina, Lee Badger and Dawn Leaf, "Recommendations of the National Institute of Standards and Technology", Special publication, NIST, U.S. Department of Commerce, 500-292.
7. Oracle Virtual Box official documentation
8. https://en.wikipedia.org/wiki/VirtualBox
9. https://docs.openstack.org/train/install/
10. https://en.wikipedia.org/wiki/OpenStack
11. https://cacoo.com/examples/network-diagram-software
12. https://www.supraits.com/infrastructure/managed-cloud/hybrid-cloud-3/cloud-computing/
13. https://www.researchgate.net/figure/Components-make-up-of-Cloud-Computing-Solution_fig1_289259494
14. https://www.telegraph.co.uk/technology/connecting-britain/colossus-bletchley-computer-broke-hitler-codes/
15. https://medium.com/penn-engineering/on-eniacs-anniversary-a-nod-to-its-female-computers-267c97a0a17
16. https://arstechnica.com/information-technology/2011/11/the-40th-birthday-ofmaybethe-first-microprocessor/
17. https://newatlas.com/anniversary-of-vannavar-bushs-famous-essay-describing-the-memex-machine/4303/
18. https://wiki.xenproject.org/wiki/Book/HelloXenProject/1-Chapter
19. https://www.safe.com/industry/natural-resources-solutions/
20. https://www.researchgate.net/figure/The-cases-of-over-provisioning-under-provisioning-and-delay-caused-by-under-provisioning_fig5_283948945
21. https://www.researchgate.net/figure/An-Enterprise-Inter-cloud-Architecture-Adapted-from-Dayananda-and-Kumar-2012_fig2_314057960
22. https://www.researchgate.net/figure/The-Hadoop-Master-Slave-Architecture-232-MapReduce-MapReduce-is-a-Hadoop-computational_fig1_274069405
23. https://mindmajix.com/hadoop-mapreduce
24. https://docs.openstack.org/train/install/
25. https://www.researchgate.net/figure/Results-of-IDC-ranking-security-challenges-3Q2009-n263_fig4_224162841

UNIT I INTRODUCTION

Introduction to Cloud Computing – Definition of Cloud – Evolution of Cloud Computing – Underlying Principles of Parallel and Distributed Computing – Cloud Characteristics – Elasticity in Cloud – On-demand Provisioning.

Introduction to Cloud Computing

- Over the last three decades, businesses that use computing resources have learned to face a vast array of buzzwords like grid computing, utility computing, autonomic computing, on-demand computing and so on.

- A new buzzword named cloud computing is presently in state-of-the-art and it is generating all sorts of confusion about what it actually means.

- In history, the term cloud has been used as a metaphor for the Internet.

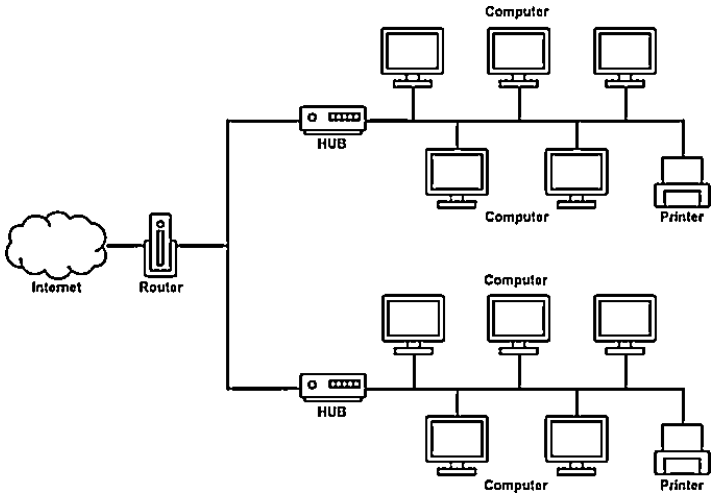


Figure 1.1 illustration of network diagram

- This usage of the term was originally derived from its common illustration in network diagrams as an outline of a cloud and the symbolic representation used to represent the transport of data across the network to an endpoint location on the other side of the network.

- Figure 1.1 illustrates the network diagram which includes the symbolic representation of cloud

- The cloud computing concepts were initiated in 1961, when Professor John McCarthy suggested that computer time-sharing technology might lead to a future where computing power and specific applications might be sold through a utility-type business model.

- This idea became very popular in the late 1960s, but in mid 1970s the idea vanished away when it became clear that the IT Industries of the day were unable to sustain such a innovative computing model. However, since the turn of the millennium, the concept has been restored.

- Utility computing is the provision of computational resources and storage resources as a metered service, similar to those provided by a traditional public utility company. This is not a new idea. This form of computing is growing in popularity, however, as companies have begun to extend the model to a cloud computing paradigm providing virtual servers that IT departments and users can access on demand.

- In early days, enterprises used the utility computing model primarily for non-mission-critical requirements, but that is quickly changing as trust and reliability issues are resolved.

- Research analysts and technology vendors are inclined to define cloud computing very closely, as a new type of utility computing that basically uses virtual servers that have been made available to third parties via the Internet.

- Others aimed to describe the term cloud computing using a very broad, all-inclusive application of the virtual computing platform. They confront that anything beyond the network firewall limit is in the cloud.

- A more softened view of cloud computing considers it the delivery of computational resources from a location other than the one from which the end users are computing.

- The cloud sees no borders and thus has made the world a much smaller place. Similar to that the Internet is also global in scope but respects only established communication paths.

- People from everywhere now have access to other people from anywhere else.

- Globalization of computing assets may be the major contribution the cloud has made to date. For this reason, the cloud is the subject of many complex geopolitical issues.

- Cloud computing is viewed as a resource available as a service for virtual data centers. Cloud computing and virtual data centers are different one.

- For example, Amazon's S3 is Simple Storage Service. This is a data storage service designed for use across the Internet. It is designed to create web scalable computing easier for developers.

- Another example is Google Apps. This provides online access via a web browser to the most common office and business applications used today. The Google server stores all the software and user data.

- Managed service providers (MSPs) offers one of the oldest form of cloud computing.

- A managed service is an application that is accessible to an organization's IT infrastructure rather than to end users which include virus scanning for email, anti spam services such as Postini, desktop management services offered by CenterBeam or Everdream, and application performance monitoring.

- Grid computing is often confused with cloud computing. Grid computing is a form of distributed computing model that implements a virtual supercomputer made up of a cluster of networked or Inter networked computers involved to perform very large tasks.

- Most of the cloud computing deployments in market today are powered by grid computing implementations and are billed like utilities, but cloud computing paradigm is evolved next step away from the grid utility model.

- The majority of cloud computing infrastructure consists of time tested and highly reliable services built on servers with varying levels of virtualized technologies, which are delivered via large scale data centers operating under various service level agreements that require 99.9999% uptime.

Definition of cloud

- Cloud computing is a model for delivering IT services in which resources are retrieved from the internet through web based tools and applications rather than a direct connection to the server.

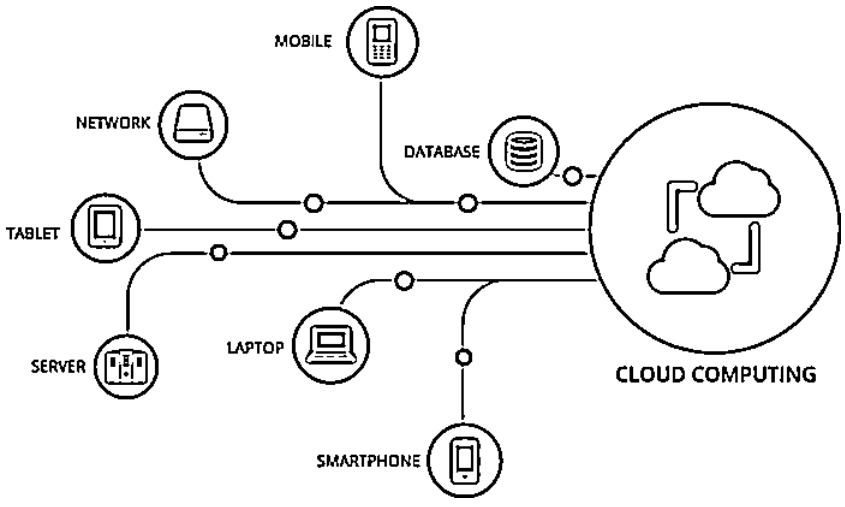


Figure 1.2 Cloud Computing Paradigm

- In other words, cloud computing is a distributed computing model over a network and means the ability to run a program on many connected components at a same time

- In the cloud computing environment, real server machines are replaced by virtual machines. Such virtual machines do not physically exist and can therefore be moved around and scaled up or down on the fly without affecting the cloud user as like a natural cloud.

- Cloud refers to software, platform, and Infrastructure that are sold as a service. The services accessed remotely through the Internet

- The cloud users can simply log on to the network without installing anything. They do not pay for hardware and maintenance. But the service providers pay for physical equipment and maintenance.

- The concept of cloud computing becomes much more understandable when one begins to think about what modern IT environments always require scalable capacity or additional capabilities to their infrastructure dynamically, without investing money in the purchase of new infrastructure, all the while without needing to conduct training for new personnel and without the need for licensing new software.

- The cloud model is composed of three components.

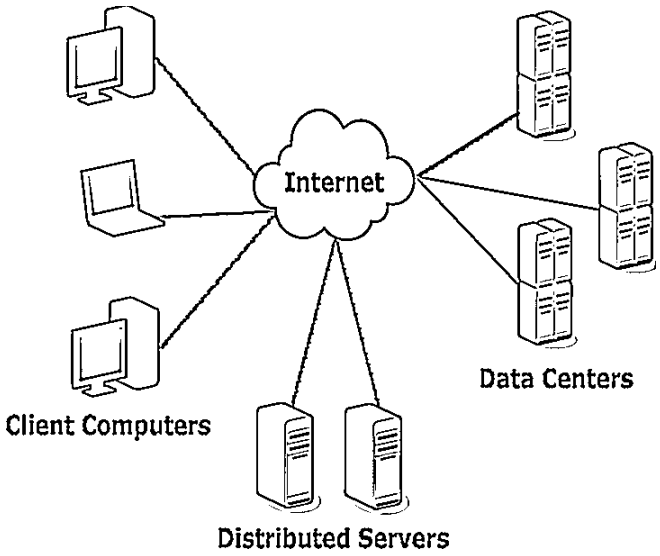


Figure 1.3 Cloud Components

- **Clients** are simple computers might be laptop, tablet, mobile phone.
- Categories of clients are Mobile clients, Thin clients and Thick clients.
- Mobile clients which includes smartphones and PDAs
- Thin clients which include servers without internal hardware. Usage of this type of clients leads to Low hardware cost, Low IT Cost, Less power consumption and less noise.
- Thick clients which includes regular computers.
- **Data Center** is a collection of servers and it contains clients requested applications.
- **Distributed Server** in which server is distributed in different geographical locations

Evolution of Cloud Computing

- It is important to understand the evolution of computing in order to get an appreciation of how IT based environments got into the cloud environment. Looking at the evolution of the computing hardware itself, from the first generation to the fourth generation of computers, shows how the IT industry's got from there to here.

- The hardware is a part of the evolutionary process. As hardware evolved, so did the software. As networking evolved, so did the rules for how computers communicate. The development of such rules or protocols, helped to drive the evolution of Internet software.

- Establishing a common protocol for the Internet led directly to rapid growth in the number of users online.

- Today, enterprises discuss about the uses of IPv6 (Internet Protocol version 6) to ease addressing concerns and for improving the methods used to communicate over the Internet.

- Usage of web browsers led to a stable migration away from the traditional data center model to a cloud computing based model. And also, impact of technologies such as

server virtualization, parallel processing, vector processing, symmetric multiprocessing, and massively parallel processing fueled radical change in IT era.

1.3.1 Hardware Evolution

- The first step along with the evolutionary path of computers was occurred in 1930, when the first binary arithmetic was developed and became the foundation of computer processing technology, terminology, and programming languages.

- Calculating devices date back to at least as early as 1642, when a device that could mechanically add numbers was invented.

- Adding devices were evolved from the abacus. This evolution was one of the most significant milestones in the history of computers.

- In 1939, the Berry brothers were invented an electronic computer that capable of operating digital aspects. The computations were performed using vacuum tube technology.

- In 1941, the introduction of Z3 at the German Laboratory for Aviation purpose in Berlin was one of the most significant events in the evolution of computers because Z3 machine supported both binary arithmetic and floating point computation. Because it was a "Turing complete" device, it is considered to be the very first computer that was fully operational.

1.3.1.1 First Generation Computers

- The first generation of modern computers traced to 1943, when the Mark I and Colossus computers were developed for fairly different purposes.

- With financial support from IBM, the Mark I was designed and developed at Harvard University. It was a general purpose electro, mechanical, programmable computer.

- Colossus is an electronic computer built in Britain at the end 1943. Colossus was the world's first programmable, digital, electronic, computing device.

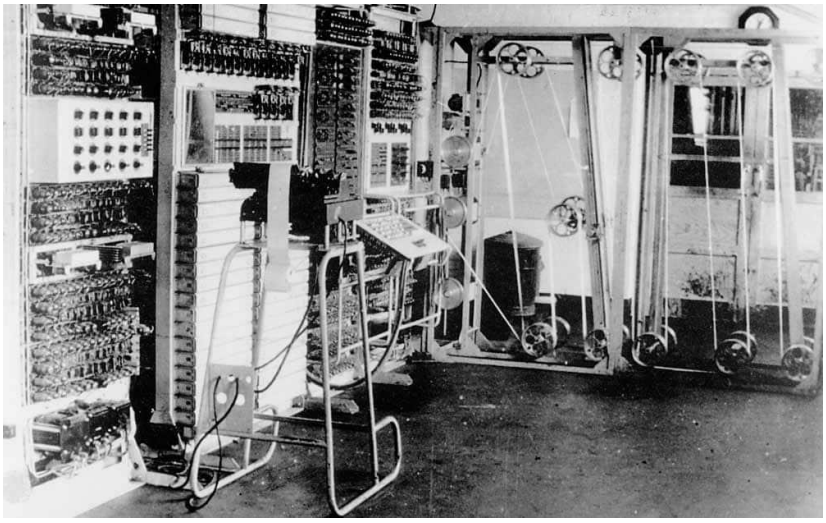


Figure 1.4 Colossus

- In general, First generation computers were built using hard-wired circuits and vacuum tubes.

- Data were stored using paper punch cards.

1.3.1.2 Second Generation Computers

- Another general-purpose computer of this era was ENIAC (Electronic Numerical Integrator and Computer), which was built in 1946. This was the first Turing complete, digital computer that capable of reprogramming to solve a full range of computing problems.

- ENIAC composed of 18,000 thermionic valves, weighed over 60,000 pounds, and consumed 25 kilowatts of electrical power per hour. ENIAC was capable of performing one lakh calculations a second.

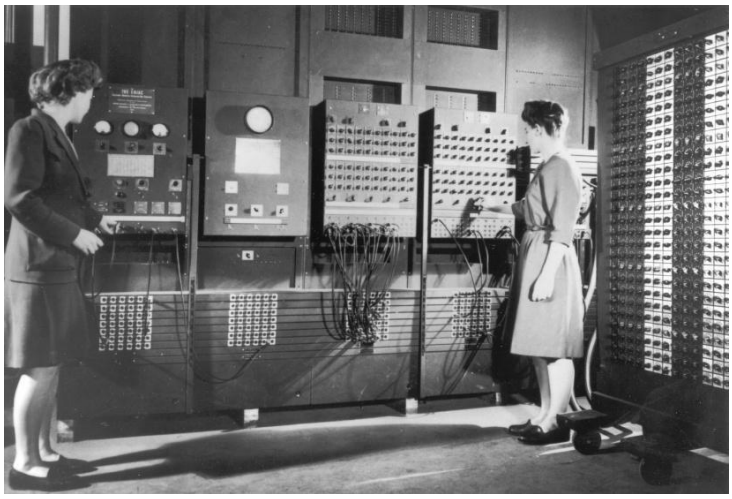


Figure 1.5 ENIAC

- Transistorized computers marked the initiation of second generation computers, which dominated in the late 1950s and early 1960s. The computers were used mainly by universities and government agencies.

- The integrated circuit or microchip was developed by Jack St. Claire Kilby, an achievement for which he received the Nobel Prize in Physics in 2000.

1.3.1.3 Third Generation Computers

- Claire Kilby's invention initiated an explosion in third generation computers. Even though the first integrated circuit was produced in 1958, microchips were not used in programmable computers until 1963.

- In 1971, Intel released the world's first commercial microprocessor called Intel 4004.

Figure 1.6 Intel 4004

- Intel 4004 was the first complete CPU on one chip and became the first commercially available microprocessor. It was possible because of the development of new silicon gate technology that enabled engineers to integrate a much greater number of transistors on a chip that would perform at a much faster speed.

1.3.1.4 Fourth Generation Computers

- The fourth generation computers that were being developed at this time utilized a microprocessor that put the computer's processing capabilities on a single integrated circuit chip.

- By combining random access memory, developed by Intel, fourth generation computers were faster than ever before and had much smaller footprints.

- The first commercially available personal computer was the MITS Altair 8800, released at the end of 1974. What followed was a flurry of other personal computers to market, such as the Apple I and II, the Commodore PET, the VIC-20, the Commodore 64, and eventually the original IBM PC in 1981. The PC era had begun in earnest by the mid-1980s.

- Even though microprocessing power, memory and data storage capacities have increased by many orders of magnitude since the invention of the 4004 processor, the technology for Large Scale Integration (LSI) or Very Large Scale Integration (VLSI) microchips has not changed all that much.

- For this reason, most of today's computers still fall into the category of fourth generation computers.

1.3.2 Internet Software Evolution

- The Internet is named after the evolution of Internet Protocol which is the standard communications protocol used by every computer on the Internet.

- Vannevar Bush was written a visionary description of the potential uses for information technology with his description of an automated library system called MEMEX.

- Bush introduced the concept of the MEMEX in late 1930s as a microfilm based device in which an individual can store all his books and records.

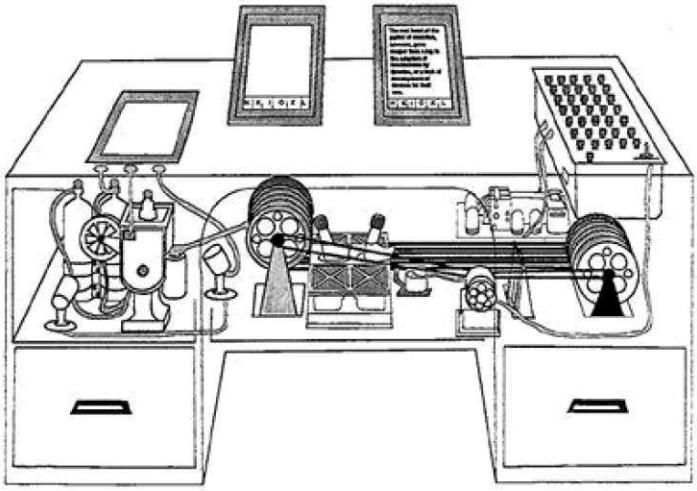


Figure 1.7 MEMEX Systems

- The second individual who has shaped the Internet was Norbert Wiener.

- Wiener was an early pioneer in the study of stochastic and noise processes. Norbert Wiener work in stochastic and noise processes was relevant to electronic engineering, communication, and control systems.

- SAGE refers Semi Automatic Ground Environment. SAGE was the most ambitious computer project and started in the mid 1950s and became operational by 1963. It remained in continuous operation for over 20 years, until 1983.

- A minicomputer was invented specifically to realize the design of the Interface Message Processor (IMP). This approach provided a system independent interface to the ARPANET.

- The IMP would handle the interface to the ARPANET network. The physical layer, the data link layer, and the network layer protocols used internally on the ARPANET were implemented using IMP.

- Using this approach, each site would only have to write one interface to the commonly deployed IMP.

- The first networking protocol that was used on the ARPANET was the Network Control Program (NCP). The NCP provided the middle layers of a protocol stack running on an ARPANET connected host computer.

- The lower-level protocol layers were provided by the IMP host interface, the NCP essentially provided a transport layer consisting of the ARPANET Host-to-Host Protocol (AHHP) and the Initial Connection Protocol (ICP).

- The AHHP defines how to transmit a unidirectional and flow controlled stream of data between two hosts.

- The ICP specifies how to establish a bidirectional pair of data streams between a pair of connected host processes.

- Robert Kahn and Vinton Cerf who built on what was learned with NCP to develop the TCP/IP networking protocol commonly used nowadays. TCP/IP quickly became the most widely used network protocol in the world.

- Over time, there evolved four increasingly better versions of TCP/IP (TCP v1, TCP v2, a split into TCP v3 and IP v3, and TCP v4 and IPv4). Now, IPv4 is the standard protocol, but it is in the process of being replaced by IPv6.

- The amazing growth of the Internet throughout the 1990s caused a huge reduction in the number of free IP addresses available under IPv4. IPv4 was never designed to scale to global levels. To increase available address space, it had to process data packets that were larger.

- After examining a number of proposals, the Internet Engineering Task Force (IETF) settled on IPv6, which was released in early 1995 as RFC 1752. IPv6 is sometimes called the Next Generation Internet Protocol (IPNG) or TCP/IP v6.

1.3.3 Server Virtualization

- Virtualization is a method of running multiple independent virtual operating systems on a single physical computer. This approach maximizes the return on investment for the computer.

- The creation and management of virtual machines has often been called platform virtualization.

- Platform virtualization is performed on a given computer (hardware platform) by software called a control program.

- Parallel processing is performed by the simultaneous execution of multiple program instructions that have been allocated across multiple processors with the objective of running a program in less time.

- The next advancement in parallel processing was multiprogramming.

- In a multiprogramming system, multiple programs submitted by users are allowed to use the processor for a short time, each taking turns and having exclusive time with the processor in order to execute instructions.

- This approach is called as round robin scheduling (RR scheduling). It is one of the oldest, simplest, fairest, and most widely used scheduling algorithms, designed especially for time-sharing systems.

- Vector processing was developed to increase processing performance by operating in a multitasking manner.

- Matrix operations were added to computers to allow a single instruction to manipulate two arrays of numbers performing arithmetic operations. This was valuable in certain types of applications in which data occurred in the form of vectors or matrices.

- The next advancement was the development of symmetric multiprocessing systems (SMP) to address the problem of resource management in master or slave models. In SMP systems, each processor is equally capable and responsible for managing the workflow as it passes through the system.

- Massive parallel processing (MPP) is used in computer architecture circles to refer to a computer system with many independent arithmetic units or entire microprocessors, which run in parallel.

1.2 Principles of Parallel and Distributed Computing

- The two fundamental and dominant models of computing environment are sequential and parallel. The sequential computing era was begun in the 1940s. The parallel and distributed computing era was followed it within a decade.

- The four key elements of computing developed during these eras are architectures, compilers, applications, and problem solving environments.

- Every aspect of this era will undergo a three phase process.

    - Research and Development (R&D)
    - Commercialization
    - Commoditization

1.4.1 Parallel vs distributed computing

- The terms parallel computing and distributed computing are often used interchangeably, even though which meant somewhat different things.

- The term parallel implies a tightly coupled system, whereas distributed refers to a wider class of system which includes tightly coupled systems.

- More specifically, the term parallel computing refers to a model in which the computation is divided among several processors which sharing the same memory.

- The architecture of a parallel computing system is often characterized by the homogeneity of components.

- In parallel computing paradigm, each processor is of the same type and it has the same capability. The shared memory has a single address space, which is accessible to all the processors.

- Processing of multiple tasks simultaneously on multiple processors is called as parallel processing.

- The parallel program consists of multiple active processes or tasks simultaneously solving a given problem.

- A given task is divided into multiple subtasks using a divide and conquer technique, and each subtask is processed on a different Central Processing Unit (CPU).

- Programming on a multiprocessor system using the divide and conquer technique is called parallel programming.

- The term distributed computing encompasses any architecture or system that allows the computation to be broken down into units and executed concurrently on different computing elements, whether these are processors on different nodes, processors on the same computer, or cores within the same processor.

- Therefore, distributed computing includes a wider range of systems and applications than parallel computing and is often considered a most common term.

1.4.2 Elements of parallel computing

- The core elements of parallel processing are CPUs. Based on the number of instruction streams and data streams that can be processed simultaneously, computing systems are classified into four categories proposed by Michael J. Flynn in 1966.

- Single Instruction Single Data systems (SISD)
- Single Instruction Multiple Data systems (SIMD)
- Multiple Instruction Single Data systems (MISD)
- Multiple Instruction, Multiple Data systems (MIMD)

- An SISD computing system is a uniprocessor system capable of executing a single instruction, which operates on a single data stream.

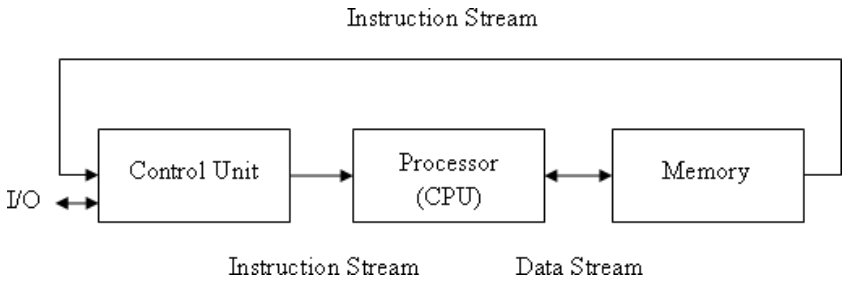


Figure 1.8 SISD

- An SIMD computing system is a multiprocessor system capable of executing the single instruction on all the CPUs but operating on different data streams.

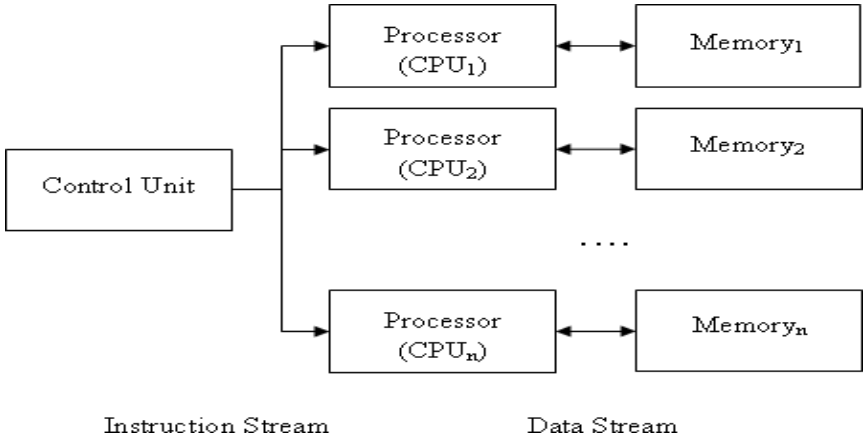


Figure 1.9 SIMD

- An MISD computing system is a multiprocessor system capable of executing different instructions on different processing elements but all of them operating on the same data streams.

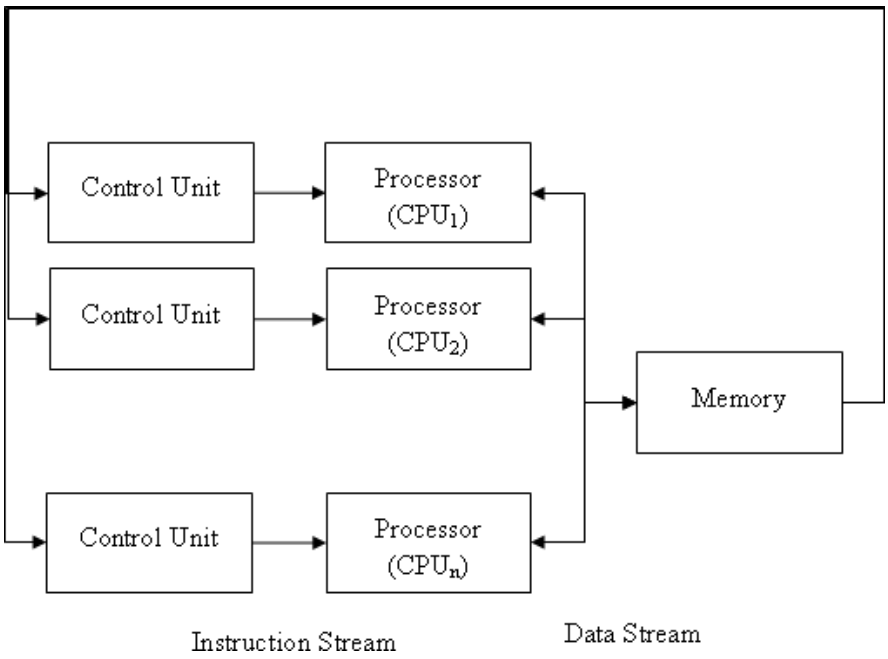


Figure 1.10 MISD

- An MIMD computing system is a multiprocessor system capable of executing multiple instructions on multiple data streams.

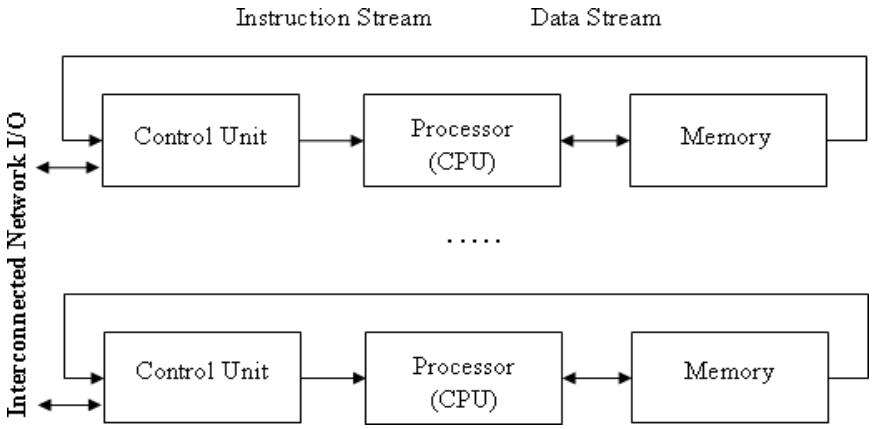


Figure 1.11 MIMD

- MIMD systems are broadly categorized into shared memory MIMD and distributed memory MIMD based on the way processing elements are coupled to the main memory.

- In the shared memory MIMD model, all the processing elements are connected to a single global memory and they all have access to it.

- In the distributed memory MIMD model, all processing elements have a local memory. Systems based on this model are also called loosely coupled multiprocessor systems.

- In general, Failures in a shared memory MIMD affects the entire system, where as this is not the case of the distributed model, in which each of the processing elements can be easily isolated.

- A wide variety of parallel programming approaches are available in computing environment. The most prominent among them are the following:

    - Data parallelism
    - Process parallelism
    - Farmer-and-worker model

- In data parallelism, the divide and conquer methodology is used to split data into multiple sets, and each data set is processed on different processing elements using the same instruction.

- In process parallelism, a given operation has multiple distinct tasks that can be processed on multiple processors.

- In farmer and worker model, a job distribution approach is used in which one processor is configured as master and all other remaining processing elements are designated as slaves. The master assigns jobs to slave processing elements and, on completion, they inform the master, which in turn collects results.

- Parallelism within an application can be detected at several levels such as Large grain (or task level), Medium grain (or control level), Fine grain (data level), Very fine grain (multiple-instruction issue)

- Speed of computation is never increase linearly. It is proportional to the square root of system cost. Therefore, the faster a system becomes, the more expensive it is to increase its speed.

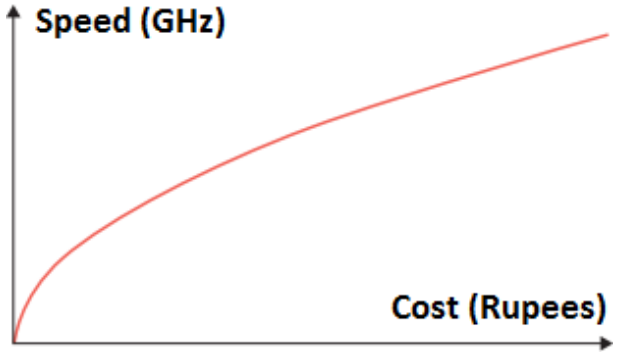


Figure 1.12 Cost versus Speed

- Speed by a parallel computer increases as the logarithm of the number of processors (i.e.,y= klog(N)).

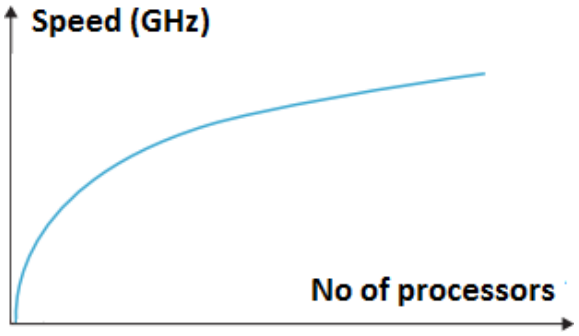


Figure 1.13 No of processors versus Speed

1.4.3 Elements of distributed computing

- A distributed system is the collection of independent computers that appears to its users as a single coherent system.

- A distributed system is the result of the interaction of several components that pass through the entire computing stack from hardware to software.

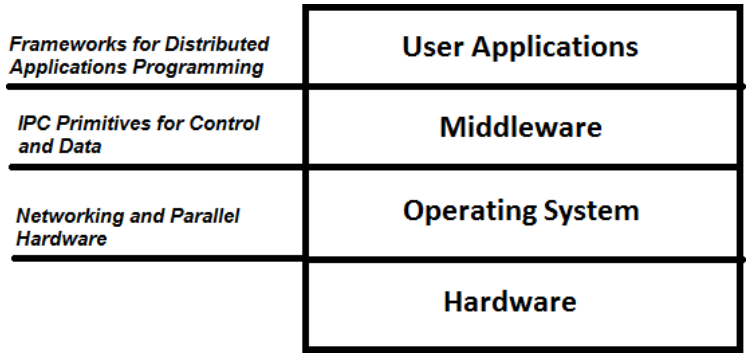


Figure 1.14 A layered view of a distributed system

- At the very bottom layer, computer and network hardware constitute the physical infrastructure.

- The hardware components are directly managed by the operating system, which provides the basic services for inter process communication (IPC), process scheduling and management, and resource management in terms of file system and local devices.

- The use of well-known standards at the operating system level and even more at the hardware and network levels allows easy harnessing of heterogeneous components and their organization into a coherent and uniform system.

- The middleware layer leverages such services to build a uniform environment for the development and deployment of distributed applications.

- The top of the distributed system stack is represented by the applications and services designed and developed to use the middleware.

- In distributed computing, Architectural styles are mainly used to determine the vocabulary of components and connectors that are used as instances of the style together with a set of constraints on how they can be combined.

- Architectural styles are classified into two major classes.

- Software architectural styles
- System architectural styles

- The first class relates to the logical organization of the software.

- The second class includes all those styles that describe the physical organization of distributed software systems in terms of their major components.

- A component represents a unit of software that encapsulates a function or a feature of the system. Examples of components can be programs, objects, processes, pipes, and filters.

- A connector is a communication mechanism that allows cooperation and coordination among components. Differently from components, connectors are not encapsulated in a single entity, but they are implemented in a distributed manner over many system components.

- Software architectural styles are based on the logical arrangement of software components.

- According to Garlan and Shaw, architectural styles are classified as shown in Table 1.1

Category	Most Common Architectural Styles
Data-centered	• Repository • Blackboard
Data flow	• Pipe and filter • Batch sequential
Virtual machine	• Rule-based system • Interpreter

Call and return	• Top down systems • Object oriented systems • Layered systems
Independent components	• Communicating processes • Event system

Table 1.1 Software Architectural Styles

- The repository architectural style is the most relevant reference model in this category. It is characterized by two main components: the central data structure, which represents the current state of the system, and a collection of independent components, which operate on the central data.

- The batch sequential style is characterized by an ordered sequence of separate programs executing one after the other. These programs are chained together by providing as input for the next program the output generated by the last program after its completion, which is most likely in the form of a file.

- The pipe and filter style is a variation of the previous style for expressing the activity of a software system as a sequence of data transformations. Each component of the processing chain is called a filter, and the connection between one filter and the next is represented by a data stream.

- Rule-Based Style architecture is characterized by representing the abstract execution environment as an inference engine. Programs are expressed in the form of rules or predicates that hold true.

- The core feature of the interpreter style is the presence of an engine that is used to interpret a pseudo code expressed in a format acceptable for the interpreter. The interpretation of the pseudo-program constitutes the execution of the program itself.

- Top Down Style is quite representative of systems developed with imperative programming, which leads to a divide and conquer approach to problem resolution.

- Object Oriented Style encompasses a wide range of systems that have been designed and implemented by leveraging the abstractions of object oriented programming

- The layered system style allows the design and implementation of software systems in terms of layers, which provide a different level of abstraction of the system.

- Each layer generally operates with at most two layers: the one that provides a lower abstraction level and the one that provides a higher abstraction layer.

- In Communicating Processes architectural style, components are represented by independent processes that leverage IPC facilities for coordination management.

- On the other hand, Event Systems architectural style where the components of the system are loosely coupled and connected.

- System architectural styles cover the physical organization of components and processes over a distributed infrastructure. They provide two fundamental reference styles: client/server and peer-to-peer.

- The client/server model features two major components: a server and a client. These two components interact with each other through a network connection using a given protocol. The communication is unidirectional. The client issues a request to the server, and after processing the request the server returns a response.

- The important operations in the client-server paradigm are request, accept (client side), and listen and response (server side).

- The client/server model is suitable in many-to-one scenarios.

- In general, multiple clients are interested in such services and the server must be appropriately designed to efficiently serve requests coming from different clients. This consideration has implications on both client design and server design.

- For the client design, there are two models: Thin client model and Fat client model.

- Thin client model, the load of data processing and transformation is put on the server side, and the client has a light implementation that is mostly concerned with retrieving and returning the data it is being asked for, with no considerable further processing.

- Fat client model, the client component is also responsible for processing and transforming the data before returning it to the user, whereas the server features a fairly light implementation that is mostly concerned with the management of access to the data.

- The three major components in the client-server model are presentation, application logic, and data storage.

- Presentation, application logic, and data maintenance can be seen as conceptual layers, which are more appropriately called tiers.

- The mapping between the conceptual layers and their physical implementation in modules and components allows differentiating among several types of architectures, which go under the name of multi-tiered architectures.

- Two major classes are Two-tier architecture and Three-tier architecture.

- Two-tier architecture partitions the systems into two tiers, which are located one in the client component and the other on the server. The client is responsible for the presentation tier by providing a user interface. The server concentrates the application logic and the data store into a single tier.

- Three-tier architecture separates the presentation of data, the application logic, and the data storage into three tiers. This architecture is generalized into an N-tier model in case it is necessary to further divide the stages composing the application logic and storage tiers.

- The peer-to-peer model introduces a symmetric architecture in which all the components are called as peers, play the same role and incorporate both client and server capabilities of the client/server model.

- The most relevant example of peer-to-peer systems is constituted by file sharing applications such as Gnutella, BitTorrent, and Kazaa.

1.4.4 Models for inter process communication

- There are several different models in which processes can interact with each other; these map to different abstractions for IPC. Among the most relevant models are shared memory, remote procedure call (RPC), and message passing.

- Message passing introduces the concept of a message as the main abstraction of the model. The entities exchanging information explicitly encode in the form of a message the data to be exchanged. The structure and the content of a message vary according to the model. Examples of this model are the Message-Passing Interface (MPI) and OpenMP.

- Remote procedure call paradigm extends the concept of procedure call beyond the boundaries of a single process, thus triggering the execution of code in remote processes. In this case, underlying client/server architecture is implied. A remote process hosts a server component, thus allowing client processes to request the invocation of methods, and returns the result of the execution.

1.4.5 Models for message-based communication

Point-to-point message model

- This model organizes the communication among single components. Each message is sent from one component to another, and there is a direct addressing to identify the message receiver. In a point-to-point communication model it is necessary to know the location of or how to address another component in the system.

Publish-and-subscribe message model

- This model introduces a different strategy, one that is based on notification among components.

- There are two major roles: the publisher and the subscriber.

- There are two major strategies for dispatching the event to the subscribers:
    - Push strategy. In this case it is the responsibility of the publisher to notify all the subscribers. For example, with a method invocation.
    - Pull strategy. In this case the publisher simply makes available the message for a specific event, and it is the responsibility of the subscribers to check whether there are messages on the events that are registered.

Request-reply message model

- The request-reply message model identifies all communication models in which, for each message sent by a process, there is a reply.

- This model is quite popular and provides a different classification that does not focus on the number of the components involved in the communication but rather on how the dynamic of the interaction evolves.

1.3 Technologies for distributed computing

Remote procedure call

- RPC is the fundamental abstraction enabling the execution of procedures on client's request.

- RPC allows extending the concept of a procedure call beyond the boundaries of a process and a single memory address space.

- The called procedure and calling procedure may be on the same system or they may be on different systems in a network.

- An important aspect of RPC is marshaling, which identifies the process of converting parameters and return values into a form that is more suitable to be transported over a network through a sequence of bytes. The term unmarshaling refers to the opposite procedure.

Distributed object frameworks

- Distributed object frameworks extend object-oriented programming systems by allowing objects to be distributed across a heterogeneous network and provide facilities so that they can coherently act as though they were in the same address space.

Service-oriented computing

- Service-oriented computing organizes distributed systems in terms of services, which represent the major abstraction for building systems.

- Service orientation expresses applications and software systems as aggregations of services that are coordinated within a service-oriented architecture (SOA).

- SOA is an architectural style supporting service orientation. It organizes a software system into a collection of interacting services.

- SOA encompasses a set of design principles that structure system development and provide means for integrating components into a coherent and decentralized system.

- SOA-based computing packages functionalities into a set of interoperable services, which can be integrated into different software systems belonging to separate business domains.

- There are two major roles within SOA: the service provider and the service consumer.

1.5 Cloud Characteristics

From the cloud computing's various definitions; a certain set of key characteristics emerges. Figure 1.15 illustrates various key characteristics related to cloud computing paradigm.

1.5.1 On-demand Provisioning

- On-demand provisioning is the single most important characteristic of cloud computing, it allows the users to request or release resources whenever they want.

- These demands are thereafter automatically granted by a cloud provider's service and the users are only charged for their usage, i.e., the time they were in possession of the resources.

- The reactivity of a cloud solution, with regard to resource provisioning is indeed of prime importance as it is closely related to the cloud's pay-as-you-go business model.

- It is one of the important and valuable features of Cloud Computing as the user can continuously monitor the server uptime, capabilities, and allotted network storage. With this feature, the user can also monitor the computing capabilities.

1.5.2 Universal Access

- Resources in the cloud need not only be provisioned rapidly but also accessed and managed universally, using standard Internet protocols, typically via RESTful web services.

- This enables the users to access their cloud resources using any type of devices, provided they have an Internet connection.

- Universal access is a key feature behind the cloud's widespread adoption, not only by professional actors but also by the general public that is nowadays familiar with cloud based solutions such as cloud storage or media streaming.

- Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms such as mobile phones, tablets, laptops, and workstations.

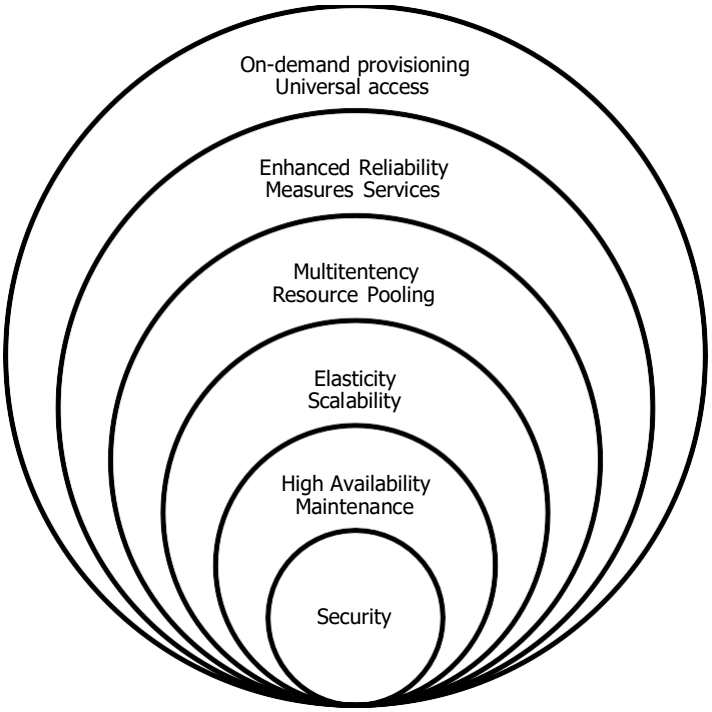


Figure 1.15 Cloud Characteristics

1.5.3 Enhanced Reliability

- Cloud computing enables the users to enhance the reliability of their applications.
- Reliability is already built in many cloud solutions via storage redundancy.
- Cloud providers usually have more than one data center and further reliability can be achieved by backing data up in different locations.
- This can also be used to ensure service availability, in the case of routine maintenance operations or the rarer case of a natural disaster.
- The user can achieve further reliability using the services of different cloud providers.

1.5.4 Measured Services

- Cloud computing refers generally to paid services.
- The customers are entitled to a certain quality of service, guaranteed by the Service Level Agreement that they should be able to supervise.
- Therefore, cloud providers offer monitoring tools, either using a graphical interface or via an API.
- These tools also help the providers themselves for billing and management purposes.

1.5.5 Multitenancy

- As the grid before, the cloud's resources are shared by different simultaneous users. These users had to reserve in advance a fixed number of physical machines for a fixed amount of time.
- In virtualized data centers, a user's provisioned resources no longer correspond to the physical infrastructure and can be dispatched over multiple physical machines.
- They can also run alongside other users' provisioned resources thus requiring a lesser amount of physical resources. Consequently, important energy savings can be made by shutting down the unused resources or putting them in energy saving mode.

1.5.6 Resource pooling

- The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand.

- There is a sense of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter).

- Examples of resources include storage, processing, memory, and network bandwidth.

1.5.7 Rapid elasticity and Scalability

- Elasticity is the ability of a system to include and exclude resources like CPU cores, memory, Virtual Machine and container instances to adapt to the load variation in real time.

- Elasticity is a dynamic property for cloud computing. There are two types of elasticity. Horizontal and Vertical.

- Horizontal elasticity consists in adding or removing instances of computing resources associated with an application.

- Vertical elasticity consists in increasing or decreasing characteristics of computing resources, such as CPU time, cores, memory, and network bandwidth.

- There are other terms such as scalability and efficiency, which are associated with elasticity but their meaning is different from elasticity while they are used interchangeably in some cases.

- Scalability is the ability of the system to sustain increasing workloads by making use of additional resources, it is time independent and it is similar to the provisioning state in elasticity but the time has no effect on the system (static property).

- The following equation that summarizes the elasticity concept in cloud computing.

  Auto scaling = Scalability +Automation
  Elasticity = Auto scaling + Optimization

- It means that the elasticity is built on top of scalability. It can be considered as an automation of the concept of scalability, however, it aims to optimize at best and as quickly as possible the resources at a given time.

- Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward commensurate with demand.

- To the consumer, the capabilities available for provisioning often appear to be unlimited and can be appropriated in any quantity at any time.

1.5.8 Easy Maintenance

- The servers are easily maintained and the downtime is very low and even in some cases, there is no downtime.

- Cloud Computing comes up with an update every time by gradually making it better. The updates are more compatible with the devices and perform faster than older ones along with the bugs which are fixed.

1.5.9 High Availability

- The capabilities of the Cloud can be modified as per the use and can be extended a lot. It analyzes the storage usage and allows the user to buy extra Cloud storage if needed for a very small amount.

1.5.10 Security

- Cloud Security is one of the best features of cloud computing. It creates a snapshot of the data stored so that the data may not get lost even if one of the servers gets damaged.

- The data is stored within the storage devices, which cannot be hacked and utilized by any other person. The storage service is quick and reliable.

TWO MARK QUESTIONS

1. Define utility computing.

   - Utility computing is the provision of computational resources and storage resources as a metered service, similar to those provided by a traditional public utility company.
   - This is not a new idea.
   - This form of computing is growing in popularity, however, as companies have begun to extend the model to a cloud computing paradigm providing virtual servers that IT departments and users can access on demand.

2. What is Grid Computing?

   - Grid computing is often confused with cloud computing.
   - Grid computing is a form of distributed computing model that implements a virtual supercomputer made up of a cluster of networked or Inter networked computers involved to perform very large tasks.

3. Define Cloud computing.

   - Cloud computing is a model for delivering IT services in which resources are retrieved from the internet through web based tools and applications rather than a direct connection to the server.

4. Define Cloud.

   - Cloud refers to software, platform, and Infrastructure that are sold as a service. The services accessed remotely through the Internet

   - The cloud users can simply log on to the network without installing anything. They do not pay for hardware and maintenance. But the service providers pay for physical equipment and maintenance.

5. What is the purpose of NCP?

   - The first networking protocol that was used on the ARPANET was the Network Control Program (NCP).
   - The NCP provided the middle layers of a protocol stack running on an ARPANET connected host computer.

6. How to increase the performance using multiprogramming?

   - In a multiprogramming system, multiple programs submitted by users are allowed to use the processor for a short time, each taking turns and having exclusive time with the processor in order to execute instructions.
   - This approach is called as round robin scheduling

7. Differentiate between Vector processing and Massive parallel processing

   - Vector processing was developed to increase processing performance by operating in a multitasking manner.
   - Massive parallel processing (MPP) is used in computer architecture circles to refer to a computer system with many independent arithmetic units or entire microprocessors, which run in parallel.

8. List the four key elements in parallel and distributed computing.

   - The four key elements of computing developed during these eras are architectures, compilers, applications, and problem solving environments.

9. Differentiate between parallel and distributed computing.

   - The terms parallel computing and distributed computing are often used interchangeably, even though which meant somewhat different things. Parallel implies a tightly coupled system, whereas distributed refers to a wider class of system which includes tightly coupled systems.

- The term distributed computing encompasses any architecture or system that allows the computation to be broken down into units and executed concurrently on different computing elements, whether these are processors on different nodes, processors on the same computer, or cores within the same processor.

10. Categorize computing systems base on Flynn's classification.

- Single Instruction Single Data systems (SISD)
- Single Instruction Multiple Data systems (SIMD)
- Multiple Instruction Single Data systems (MISD)
- Multiple Instruction, Multiple Data systems (MIMD)

11. List the most prominent parallel programming approaches.

- Data parallelism
- Process parallelism
- Farmer-and-worker model

12. What is farmer and worker model?

- A job distribution approach is used in which one processor is configured as master and all other remaining processing elements are designated as slaves.
- The master assigns jobs to slave processing elements and, on completion, they inform the master, which in turn collects results.

13. Differentiate between component and connector.

- A component represents a unit of software that encapsulates a function or a feature of the system.
  - A connector is a communication mechanism that allows cooperation and coordination among components.

14. Classify architectural styles according to Garlan and Shaw.

- Data-centered
- Data flow
- Virtual machine
- Call and return
- Independent components

15. What is repository architectural style?

- The repository architectural style is the most relevant reference model in this category.
- It is characterized by two main components: the central data structure, which represents the current state of the system, and a collection of independent components, which operate on the central data.

16. When the computing paradigm adapt client / server mode?

- The client/server model is suitable in many-to-one scenarios.
- The client/server model features two major components: a server and a client.
- These two components interact with each other through a network connection using a given protocol.
- The communication is unidirectional.

17. Differentiate between Thin client and Fat client model.

- Thin client model, the load of data processing and transformation is put on the server side, and the client has a light implementation that is mostly concerned with retrieving and returning the data it is being asked for, with no considerable further processing.

- Fat client model, the client component is also responsible for processing and transforming the data before returning it to the user.

18. Differentiate between two tier and three tier architecture.

- Two-tier architecture partitions the systems into two tiers, which are located one in the client component and the other on the server.

- Three-tier architecture separates the presentation of data, the application logic, and the data storage into three tiers.

19. What is point-to-point model?

- This model organizes the communication among single components.
- Each message is sent from one component to another, and there is a direct addressing to identify the message receiver.
- In a point-to-point communication model it is necessary to know the location of or how to address another component in the system.

20. List the strategies for dispatching the event to the subscribers

- Push strategy. In this case it is the responsibility of the publisher to notify all the subscribers.
- Pull strategy. In this case the publisher simply makes available the message for a specific event.

21. What is request-reply model?

- The request-reply message model identifies all communication models in which, for each message sent by a process, there is a reply.

- This model is quite popular and provides a different classification that does not focus on the number of the components involved in the communication but rather on how the dynamic of the interaction evolves.

22. What is the purpose of Distributed object frameworks?

- Distributed object frameworks extend object-oriented programming systems by allowing objects to be distributed across a heterogeneous network and provide facilities so that they can coherently act as though they were in the same address space.

23. List the key characteristics of cloud.

- On-demand provisioning Universal access
- Enhanced Reliability Measures Services
- Multitentency Resource Pooling
- Elasticity Scalability
- High Availability Maintenance
- Security

UNIT II CLOUD ENABLING TECHNOLOGIES

Service Oriented Architecture –REST and Systems of Systems –Web Services –Publish-Subscribe Model –Basics of Virtualization –Types of Virtualization –Implementation Levels of Virtualization –Virtualization Structures –Tools and Mechanisms –Virtualization of CPU –Memory –I/O Devices –Virtualization Support and Disaster Recovery.

2.1 Service Oriented Architecture

- A service encapsulates a software component that gives a set of coherent and related functionalities that can be reused and integrated into larger and more complex applications.

- The term service is a general abstraction that encompasses several different implementations using different technologies and protocols.

- Don Box identifies four major characteristics with the intention of identify a service.

- *Boundaries are explicit*

    - A service oriented applications are generally composed of services that are spread across different domains, trust authorities and execution environments.

- *Services are autonomous*

    - Services are components that exist to offer functionality.
    - Services are aggregated and coordinated to build more complex system.
    - Services are not designed to be part of a specific system but they can be integrated in several software systems.
    - The notion of autonomy also affects the way services handle failures.

- *Services share schema and contracts*

    - Services never share class and interface definitions.
    - In object oriented systems, services are not expressed in terms of classes or interfaces but they define in terms of schemas and contracts.
    - Technologies such as XML and SOAP provide the appropriate tools to support such features rather than class definition and an interface declaration.

- *Services compatibility is determined based on policy*

    - Service orientation separates structural compatibility from semantic compatibility.
    - Structural compatibility is based on contracts and schema and can be validated by machine based techniques.
    - Semantic compatibility is expressed in the form of policies that define the capabilities and requirements for a service.

- Service Oriented architecture is an architectural style supporting service orientation.

- This architectural style organizes a software system into a collection of interacting services.

- SOA encompasses a set of design principles that structure system development and provide means for integrating components into a coherent and decentralized system.

- SOA based computing packages functionalities into a set of interoperable services, which can be integrated into different software systems belonging to separate business domains.

- There are two major roles exist in SOA

    - Service provider
    - Service consumer

- First, the service provider is the maintainer of the service and the organization that makes available one or more services for others to use.

- To advertise services, the provider can publish them in a registry along with a service contract that specifies the nature of the service, how to use the service, the requirements for the service and the fees charged.

- Second, the service consumer can locate the service metadata in the registry and develop the required client components to bind and use the service.

- Service providers and consumers can belong to different organization bodies.

- It is very common in SOA based computing systems that components play the roles of both service provider and service consumer.

- Services might aggregate information and data retrieved from other services or create workflows of services to satisfy the request of a given service consumer. This practice is called as service orchestration, which more generally describes the automated arrangement, coordination and management of more complex computer systems, middleware and services.

- Another important interaction pattern is service composition is the coordinated interaction of services without a single point of control.

- SOA provides a reference model for architecting several software systems primarily for enterprise business applications and systems.

- Interoperability, standards and service contracts plays a fundamental role.

- In particular, the following list of guiding principles characterize SOA platforms:

    - *Standardized service contract*

        - Services adhere to a given communication agreement, which is specified through one or more service description documents.

- *Loose coupling*

    - Services are designed as self-contained components, maintain relationships that minimize dependencies and only require being aware of each other.
    - Service contracts will enforce the required interaction among services.
    - This simplifies the flexible aggregation of services and enables a more agile design strategy that supports the evolution of the enterprise business.

- *Abstraction*

    - A service is completely defined by service contracts and description documents.
    - Abstraction hiding the logic, which is encapsulated within their implementation.
    - The use of service description documents and contracts removes the need to consider the technical implementation details.
    - It provides a more intuitive framework to define software systems within a business context.

- *Reusability*

    - Designed as components, services can be reused more efficiently, thus reducing development time and the associated costs.
    - Reusability allows for a more agile design and cost effective system implementation and deployment.

- *Autonomy*

    - Services have control over the logic they encapsulate and do not know about their implementation.

- *Lack of state*

- By providing a stateless interaction pattern, services increase the chance of being reused and aggregated, particularly in a scenario in which a single service is used by multiple consumers that belong to different administrative and business domains.

Discoverability

- Services are defined by description documents that constitute supplemental metadata through which they can be effectively discovered.
- Service discovery provides an effective means for utilizing third party resources.

- *Composability*

- Using services as building blocks, difficult operations can be implemented.
- Service orchestration and choreography provide a solid support for composing services and achieving desired business goals.

- Together with these principles, other resources guide the use of SOA for enterprise application integration (EAI).

- The SOA manifest integrates the previously described principles with general considerations about the overall goals of a service oriented approach to enterprise application software design and what is valued in SOA.

- Modeling frameworks and methodologies, such as the Service Oriented Modeling Framework (SOMF) and reference architectures introduced by the Organization for Advancement of Structured Information Standards (OASIS), provide means for effectively realizing service oriented architectures.

- SOA can be realized through several technologies.

- The first implementations of SOA have leveraged distributed object programming technologies such as CORBA and DCOM.

- CORBA has been a suitable platform for realizing SOA systems because it provides interoperability among different implementations and has been designed as a specification supporting the development of industrial applications.

- Nowadays, SOA is mostly realized through Web services technology, which provides an interoperable platform for connecting systems and applications.

2.2 Web Services

- Web services are the prominent technology for implementing SOA systems and applications.

- They leverage Internet technologies and standards for building distributed systems. Several aspects make Web services the technology of choice for SOA.

    - First, they allow for interoperability across different platforms and programming languages.
    - Second, they are based on well-known and vendor independent standards such as HTTP, SOAP, XML and WSDL.
    - Third, they provide an intuitive and simple way to connect heterogeneous software systems, enabling the quick composition of services in a distributed environment.
    - Finally, they provide the features required by enterprise business applications to be used in an industrial environment.

- They define facilities for enabling service discovery, which allows the system architect to more efficiently compose SOA applications and service metering to assess whether a specific service complies with the contract between the service provider and the service consumer.

- The concept behind a Web service is very simple.

- Using as a basis the object oriented abstraction, a Web service exposes a set of operations that can be invoked by leveraging Internet based protocols.

- The semantics for invoking Web service methods is expressed through interoperable standards such as XML and WSDL, which also provide a complete framework for expressing simple and complex types in a platform independent manner.

- Web services are made accessible by being hosted in a Web server

- HTTP is the most popular transport protocol used for interacting with Web services.

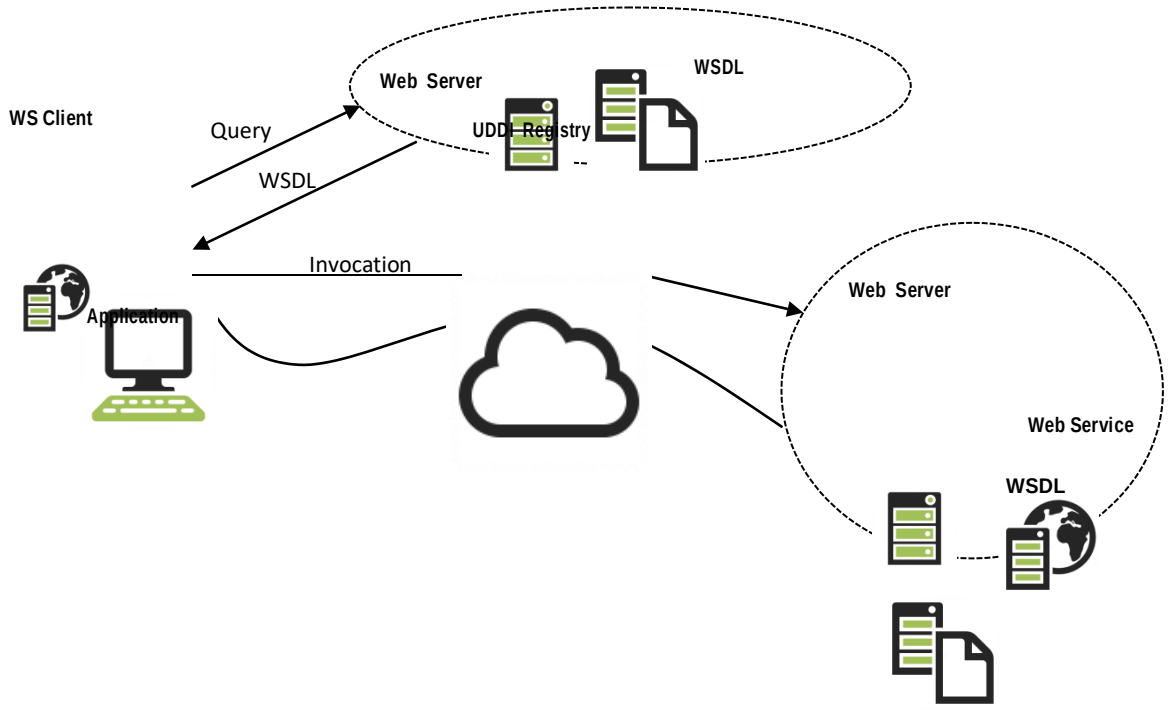


Figure 2.1 Reference scenario for Web Services

- Figure 2.1 describes the common use case scenarios for Web services.

- System architects develop a Web service with their technology of choice and deploy it in compatible Web or application servers.

- The service description document is expressed by means of Web Service Definition Language (WSDL), can be either uploaded to a global registry or attached as a metadata to the service itself.

- Service consumers can look up and discover services in global catalogs using Universal Description Discovery and Integration (UDDI).

- The Web service description document allows service consumers to automatically generate clients for the given service and embed them in their existing application.

- Web services are now extremely popular, so bindings exist for any mainstream programming language in the form of libraries or development support tools.

- This makes the use of Web services seamless and straightforward with respect to technologies such as CORBA that require much more integration effort.

- Moreover, being interoperable, Web services constitute a better solution for SOA with respect to several distributed object frameworks, such as .NET Remoting, Java RMI, and DCOM/COM1, which limit their applicability to a single platform or environment.

- Besides the main function of enabling remote method invocation by using Web based and interoperable standards, Web services encompass several technologies that put together and facilitate the integration of heterogeneous applications and enable service oriented computing.

- Figure 2.2 shows the Web service technologies stack that lists all the components of the conceptual framework describing and enabling the Web services abstraction.

- These technologies cover all the aspects that allow Web services to operate in a distributed environment, from the specific requirements for the networking to the discovery of services.

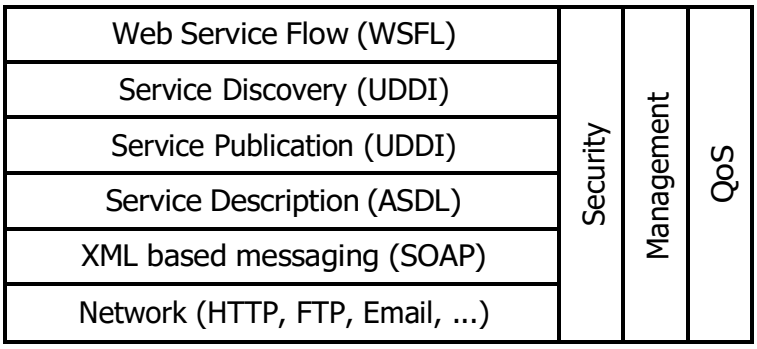


Figure 2.2 Web services technologies stack

- The backbone of all these technologies is XML, which is also one of the causes of Web service's popularity and ease of use.

- XML based languages are used to manage the low level interaction for Web service method calls (SOAP), for providing metadata about the services (WSDL), for discovery services (UDDI), and other core operations.

- In practice, the core components that enable Web services are SOAP and WSDL.

- Simple Object Access Protocol (SOAP) is an XML based language for exchanging structured information in a platform-independent manner, constitutes the protocol used for Web service method invocation.

- Within a distributed context leveraging the Internet, SOAP is considered an application layer protocol that leverages the transport level, most commonly HTTP, for IPC.

- SOAP structures the interaction in terms of messages that are XML documents mimicking the structure of a letter, with an envelope, a header, and a body.

- The envelope defines the boundaries of the SOAP message.

- The header is optional and contains relevant information on how to process the message.

```
Host : www.sample.com
Content-Type: application/soap+xml; charsetutf-8
Content-Length: <Size>

<?xml version= "1.0">
<soap: Envelope xmlns:soap= "http://www.w3.org/2001/12/soap-envelope"
soap:encodingStyle= "http://www.w3.org/2001/12/soap-enoding" >
<soap:Header></soap:Header>
<soap:Body xmlns=http://www.sample.com/stock>
<m:GetPrice>
<m: StockName>DELL</m:StockName>
</m:GetPrice>
</soap:Body>
</soap: Envelope>

POST /StockPrice HTTP/1.1
Host : www.sample.com
Content-Type: application/soap+xml; charsetutf-8
Content-Length: <Size>

<?xml version= "1.0">
<soap: Envelope xmlns:soap= "http://www.w3.org/2001/12/soap-envelope"
soap:encodingStyle= "http://www.w3.org/2001/12/soap-enoding" >
<soap:Header></soap:Header>
<soap:Body xmlns=http://www.sample.com/stock>
<m:GetPriceResponse>
<m: Price>58.5</m:Price>
</m:GetPriceResponse>
</soap:Body>
</soap: Envelope>
```

Figure 2.3 SOAP Message

- In addition to that it contains information such as routing and delivery settings, authentication, transaction contexts and authorization assertions.

- The body contains the actual message to be processed.

- The main uses of SOAP messages are method invocation and result retrieval.

- Figure 2.3 shows an example of a SOAP message used to invoke a Web service method that retrieves the price of a given stock and the corresponding reply.

- Despite the fact that XML documents are easy to produce and process in any platform or programming language, SOAP has often been considered quite inefficient because of the excessive use of markup that XML imposes for organizing the information into a well-formed document.

- Therefore, lightweight alternatives to the SOAP/XML pair have been proposed to support Web services.

2.3 REST and Systems of Systems

- The most relevant alternative to SOAP/XML pair is Representational State Transfer (REST), which provides a model for designing network based software systems utilizing the client / server model and leverages the facilities provided by HTTP for IPC without additional burden.

- In a RESTful system, a client sends a request over HTTP using the standard HTTP methods (PUT, GET, POST, and DELETE) and the server issues a response that includes the representation of the resource.

- By relying on this minimal support, it is possible to provide whatever it needed to replace the basic and most important functionality provided by SOAP, which is method invocation.

- The GET, PUT, POST, and DELETE methods constitute a minimal set of operations for retrieving, adding, modifying and deleting the data.

- Together with an appropriate URI organization to identify resources, all the atomic operations required by a Web service are implemented.

- The content of data is still transmitted using XML as part of the HTTP content, but the additional markup required by SOAP is removed.

- For this reason, REST represents a lightweight alternative to SOAP, which works effectively in contexts where additional aspects beyond those manageable through HTTP are absent.

- RESTful Web services operate in an environment where no additional security beyond the one supported by HTTP is required.

- This is not a great limitation, and RESTful Web services are quite popular and used to deliver functionalities at enterprise scale:

    - Twitter
    - Yahoo! (search APIs, maps, photos, etc)
    - Flickr
    - Amazon.com

- Web Service Description Language (WSDL) is an XML based language for the description of Web services.

- It is used to define the interface of a Web service in terms of methods to be called and types and structures of the required parameters and return values.

- In Figure 2.3 we notice that the SOAP messages for invoking the GetPrice method and receiving the result do not have any information about the type and structure of the parameters and the return values.

- This information is stored within the WSDL document attached to the Web service.

- Therefore, Web service consumer applications already know which types of parameters are required and how to interpret results.

- As an XML based language, WSDL allows for the automatic generation of Web service clients that can be easily embedded into existing applications.

- Moreover, XML is a platform and language independent specification, so clients for web services can be generated for any language that is capable of interpreting XML data.

- This is a fundamental feature that enables Web service interoperability and one of the reasons that make such technology a solution of choice for SOA.

- Besides those directly supporting Web services, other technologies that characterize Web 2.0 and contribute to enrich and empower Web applications and then SOA based systems.

- These fall under the names of Asynchronous JavaScript and XML (AJAX), JavaScript Standard Object Notation (JSON) and others.

- AJAX is a conceptual framework based on JavaScript and XML that enables asynchronous behavior in Web applications by leveraging the computing capabilities of modern Web browsers.

- This transforms simple Web pages in complete applications and used to enrich the user experience.

- AJAX uses XML to exchange data with Web services and applications

- An alternative to XML is JSON, which allows representing objects and collections of objects in a platform independent manner.

- Often it is preferred to transmit data in an AJAX context because compared to XML, it is a lighter notation and therefore allows transmitting the same amount of information in a more concise form.

2.4 Publish-Subscribe Model

- Publish-and-subscribe message model introduces a different message passing strategy, one that is based on notification among components.

- There are two major roles:

    - The publisher and the subscriber

        - The publisher provides facilities for the subscriber to register its interest in a specific topic or event.
        - Specific conditions holding true on the publisher side can trigger the creation of messages that are attached to a specific event.
        - A message will be available to all the subscribers that registered for the corresponding event.

- There are two major strategies for dispatching the event to the subscribers:

    - Push strategy

        - In this case it is the responsibility of the publisher to notify all the subscribers using method invocation.
    - Pull strategy

        - In this case the publisher simply makes available the message for a specific event and it is responsibility of the subscribers to check whether there are messages on the events that are registered.

- Publish and subscribe model is very suitable for implementing systems based on the one to many communication model and simplifies the implementation of indirect communication patterns.

- It is, in fact, not necessary for the publisher to know the identity of the subscribers to make the communication happen.

2.5 Basics of Virtualization

- Virtualization technology is one of the fundamental components of cloud computing, especially in regard to infrastructure based services.

- Virtualization allows the creation of a secure, customizable and isolated execution environment for running application.

- Virtualization is a large umbrella of technologies and concepts that are meant to provide an abstract environment whether virtual hardware or an operating system to run applications.

- The term virtualization is often synonymous with hardware virtualization, which plays a fundamental role in efficiently delivering Infrastructure as a Service (IaaS) solutions for cloud computing.

- Virtualization technologies have gained renewed interested recently due to the confluence of several phenomena:

    - Increased performance and computing capacity.
    - Underutilized hardware and software resources

- Lack of space
- Greening initiatives
- Rise of administrative costs

- Virtualization is a broad concept that refers to the creation of a virtual version of something, whether hardware, a software environment, storage and a network.

- In a virtualized environment, there are three major components:

    - Guest
    - Host
    - Virtualization layer

- The guest represents the system component that interacts with the virtualization layer rather than with the host, as would normally happen.

- The host represents the original environment where the guest is supposed to be managed.

- The virtualization layer is responsible for recreating the same or a different environment where the guest will operate.

2.5.1 Characteristics of virtualized environments

- Increased security

    - The ability to control the execution of a guest in a completely transparent manner opens new possibilities for delivering a secure, controlled execution environment.
    - The virtual machine represents an emulated environment in which the guest is executed.
    - This level of indirection allows the virtual machine manager to control and filter the activity of the guest, thus preventing some harmful operations from being performed.

- Managed execution Virtualization of the execution environment not only allows increased security, but a wider range of features also can be implemented.

- In particular, sharing, aggregation, emulation, and isolation are the most relevant features

- Sharing

    - Virtualization allows the creation of a separate computing environment within the same host.
    - In this way it is possible to fully exploit the capabilities of a powerful guest, which would otherwise be underutilized.

- Aggregation

    - Not only is it possible to share physical resource among several guests but virtualization also allows aggregation, which is the opposite process.
    - A group of separate hosts can be tied together and represented to guests as a single virtual host.

- Emulation

    - Guest programs are executed within an environment that is controlled by the virtualization layer, which ultimately is a program.
    - This allows for controlling and tuning the environment that is exposed to guests.

- Isolation

    - Virtualization allows providing guests whether they are operating systems, applications, or other entities with a completely separate environment, in which they are executed.
    - The guest program performs its activity by interacting with an abstraction layer, which provides access to the underlying resources.

- Benefits of Isolation
  - First it allows multiple guests to run on the same host without interfering with each other.
  - Second, it provides a separation between the host and the guest.

- Another important capability enabled by virtualization is performance tuning.

- This feature is a reality at present, given the considerable advances in hardware and software supporting virtualization.

- It becomes easier to control the performance of the guest by finely tuning the properties of the resources exposed through the virtual environment.

- This capability provides a means to effectively implement a quality of service (QoS) infrastructure that more easily fulfills the service level agreement (SLA) established for the guest.

- Portability

  - The concept of portability applies in different ways according to the specific type of virtualization considered.
  - In the case of a hardware virtualization solution, the guest is packaged into a virtual image that, in most cases, can be safely moved and executed on top of different virtual machines

2.6 Types of Virtualization

- Virtualization is mainly used to emulate execution environments, storage and networks.

- Execution virtualization techniques into two major categories by considering the type of host they require.

- Process level techniques are implemented on top of an existing operating system, which has full control of the hardware.

- System level techniques are implemented directly on hardware and do not require or require a minimum of support from existing operating system.

- Within these two categories we can list various techniques that offer the guest a different type of virtual computation environment:

    - Bare hardware
    - Operating system resources
    - Low level programming language
    - Application libraries

- Execution virtualization includes all techniques that aim to emulate an execution environment that is separate from the one hosting the virtualization layer.

- All these techniques concentrate their interest on providing support for the execution of programs, whether these are the operating system, a binary specification of a program compiled against an abstract machine model or an application.

- Therefore, execution virtualization can be implemented directly on top of the hardware by the operating system, an application and libraries (dynamically or statically) linked to an application image.

- Modern computing systems can be expressed in terms of the reference model described in Figure 2.4.

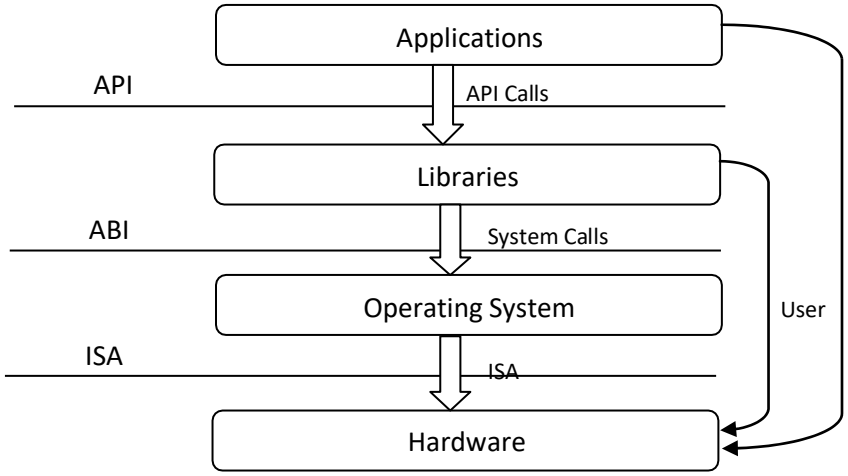


Figure 2.4 Machine reference model

- At the bottom layer, the model for the hardware is expressed in terms of the Instruction Set Architecture (ISA), which defines the instruction set for the processor, registers, memory and an interrupt management.

- ISA is the interface between hardware and software.

- ISA is important to the operating system (OS) developer (System ISA) and developers of applications that directly manage the underlying hardware (User ISA).

- The application binary interface (ABI) separates the operating system layer from the applications and libraries, which are managed by the OS.

- ABI covers details such as low level data types, alignment, call conventions and defines a format for executable programs.

- System calls are defined at this level.

- This interface allows portability of applications and libraries across operating systems that implement the same ABI.

- The highest level of abstraction is represented by the application programming interface (API), which interfaces applications to libraries and the underlying operating system.

- For this purpose, the instruction set exposed by the hardware has been divided into different security classes that define who can operate with them.

- The first distinction can be made between privileged and non privileged instructions.

    - Non privileged instructions are those instructions that can be used without interfering with other tasks because they do not access shared resources.
    - This category contains all the floating, fixed-point, and arithmetic instructions.

- Privileged instructions are those that are executed under specific restrictions and are mostly used for sensitive operations, which expose (behavior-sensitive) or modify (control-sensitive) the privileged state.

- Some types of architecture feature more than one class of privileged instructions and implement a finer control of how these instructions can be accessed.

- For instance, a possible implementation features a hierarchy of privileges illustrate in the figure 2.5 in the form of ring-based security: Ring 0, Ring 1, Ring 2, and Ring 3;

    - Ring 0 is in the most privileged level and Ring 3 in the least privileged level.
    - Ring 0 is used by the kernel of the OS, rings 1 and 2 are used by the OS level services, and Ring 3 is used by the user.
    - Recent systems support only two levels, with Ring 0 for supervisor mode and Ring 3 for user mode.

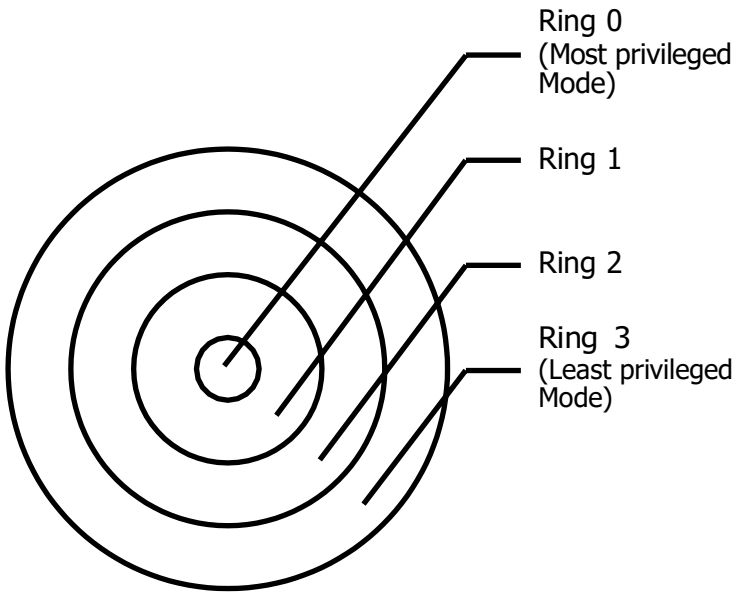


Figure 2.5 Security rings

- All the current systems support at least two different execution modes: supervisor mode and user mode.
    - The supervisor mode denotes an execution mode in which all the instructions (privileged and non privileged) can be executed without any restriction.
    - This mode, also called master mode or kernel mode, is generally used by the operating system (or the hypervisor) to perform sensitive operations on hardware level resources.
    - In user mode, there are restrictions to control the machine level resources.
- The distinction between user and supervisor mode allows us to understand the role of the hypervisor and why it is called that.
- Conceptually, the hypervisor runs above the supervisor mode and from here the prefix "hyper" is used.
- In reality, hypervisors are run in supervisor mode and the division between privileged and non privileged instructions has posed challenges in designing virtual machine managers.

2.6.1 Hardware level virtualization

- Hardware level virtualization is a virtualization technique that provides an abstract execution environment in terms of computer hardware on top of which a guest operating system can be run.

- In this model, the guest is represented by the operating system, the host by the physical computer hardware, the virtual machine by its emulation and the virtual machine manager by the hypervisor.

- The hypervisor is generally a program or a combination of software and hardware that allows the abstraction of the underlying physical hardware.

- Hardware level virtualization is also called system virtualization, since it provides ISA to virtual machines, which is the representation of the hardware interface of a system.

- This is to differentiate it from process virtual machines, which expose ABI to virtual machines.

- Hypervisors is a fundamental element of hardware virtualization is the hypervisor, or virtual machine manager (VMM).

- It recreates a hardware environment in which guest operating systems are installed.

- There are two major types of hypervisor: Type I and Type II. Figure 2.6 shows different type of hypervisors.

    - Type I hypervisors run directly on top of the hardware.
    - Type I hypervisor take the place of the operating systems and interact directly with the ISA interface exposed by the underlying hardware and they emulate this interface in order to allow the management of guest operating systems.

- This type of hypervisor is also called a native virtual machine since it runs natively on hardware.

- Type II hypervisors require the support of an operating system to provide virtualization services.
- This means that they are programs managed by the operating system, which interact with it through the ABI and emulate the ISA of virtual hardware for guest operating systems.
- This type of hypervisor is also called a hosted virtual machine since it is hosted within an operating system.

2.6.2 Hardware virtualization techniques

- Hardware virtualization provides an abstract execution environment by Hardware assisted virtualization, Full virtualization, Paravirtualization and Partial virtualization techniques.

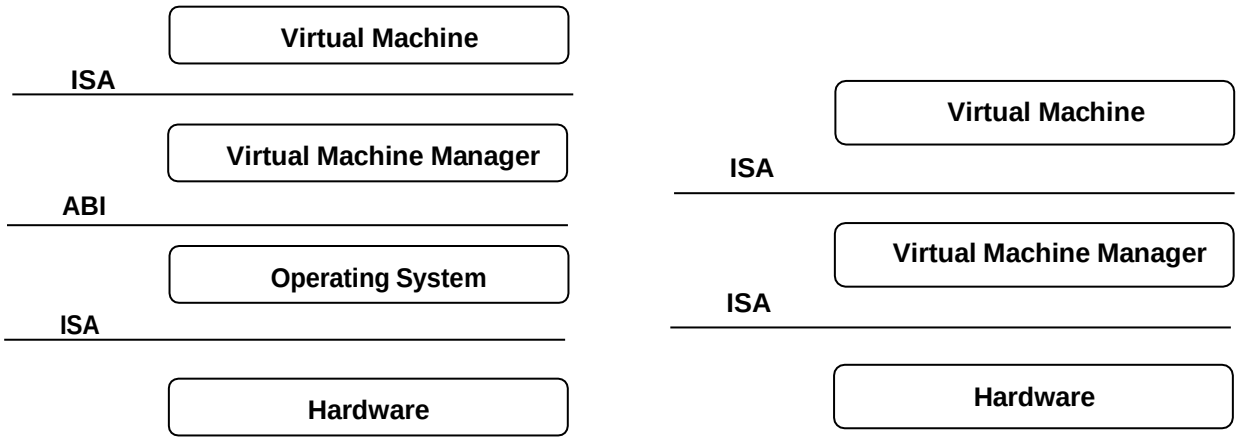


Figure 2.6 Hosted virtual machine and native virtual machine

2.6.2.1 Hardware assisted virtualization

- Hardware assisted virtualization refers to a scenario in which the hardware provides architectural support for building a virtual machine manager able to run a guest operating system in complete isolation.

- This technique was originally introduced in the IBM System/370.

- At present, examples of hardware assisted virtualization are the extensions to the x86 architecture introduced with Intel-VT (formerly known as Vanderpool) and AMD-V (formerly known as Pacifica).

- These extensions, which differ between the two vendors, are meant to reduce the performance penalties experienced by emulating x86 hardware with hypervisors.

- Before the introduction of hardware assisted virtualization, software emulation of x86 hardware was significantly costly from the performance point of view.

- The reason for this is that by design the x86 architecture did not meet the formal requirements introduced by Popek and Goldberg and early products were using binary translation to trap some sensitive instructions and provide an emulated version.

- Products such as VMware Virtual Platform, introduced in 1999 by VMware, which pioneered the field of x86 virtualization, were based on this technique.

- After 2006, Intel and AMD introduced processor extensions and a wide range of virtualization solutions took advantage of them: Kernel-based Virtual Machine (KVM), VirtualBox, Xen, VMware, Hyper-V, Sun xVM, Parallels, and others.

2.6.2.2 Full virtualization

- Full virtualization refers to the ability to run a program, most likely an operating system, directly on top of a virtual machine and without any modification, as though it were run on the raw hardware.

- To make this possible, virtual machine managers are required to provide a complete emulation of the entire underlying hardware.

- The principal advantage of full virtualization is complete isolation, which leads to enhanced security, ease of emulation of different architectures and coexistence of different systems on the same platform.

- Whereas it is a desired goal for many virtualization solutions, full virtualization poses important concerns related to performance and technical implementation.

- A key challenge is the interception of privileged instructions such as I/O instructions: Since they change the state of the resources exposed by the host, they have to be contained within the virtual machine manager.

- A simple solution to achieve full virtualization is to provide a virtual environment for all the instructions, thus posing some limits on performance.

- A successful and efficient implementation of full virtualization is obtained with a combination of hardware and software, not allowing potentially harmful instructions to be executed directly on the host.

2.6.2.3 Paravirtualization

- Paravirtualization is a not transparent virtualization solution that allows implementing thin virtual machine managers.

- Paravirtualization techniques expose a software interface to the virtual machine that is slightly modified from the host and, as a consequence, guests need to be modified.

- The aim of paravirtualization is to provide the capability to demand the execution of performance critical operations directly on the host, thus preventing performance losses that would otherwise be experienced in managed execution.

- This allows a simpler implementation of virtual machine managers that have to simply transfer the execution of these operations, which were hard to virtualize, directly to the host.

- To take advantage of such an opportunity, guest operating systems need to be modified and explicitly ported by remapping the performance critical operations through the virtual machine software interface.

- This is possible when the source code of the operating system is available, and this is the reason that paravirtualization was mostly explored in the opensource and academic environment.

- This technique has been successfully used by Xen for providing virtualization solutions for Linux-based operating systems specifically ported to run on Xen hypervisors.

- Operating systems that cannot be ported can still take advantage of para virtualization by using ad hoc device drivers that remap the execution of critical instructions to the paravirtualization APIs exposed by the hypervisor.

- Xen provides this solution for running Windows based operating systems on x86 architectures.

- Other solutions using paravirtualization include VMWare, Parallels, and some solutions for embedded and real-time environments such as TRANGO, Wind River, and XtratuM.

2.6.2.4 Partial virtualization

- Partial virtualization provides a partial emulation of the underlying hardware, thus not allowing the complete execution of the guest operating system in complete isolation.

- Partial virtualization allows many applications to run transparently, but not all the features of the operating system can be supported as happens with full virtualization.

- An example of partial virtualization is address space virtualization used in time sharing systems; this allows multiple applications and users to run concurrently in a separate memory space, but they still share the same hardware resources (disk, processor, and network).

- Historically, partial virtualization has been an important milestone for achieving full virtualization, and it was implemented on the experimental IBM M44/44X.

- Address space virtualization is a common feature of contemporary operating systems.

2.6.3 Operating system level virtualization

- Operating system level virtualization offers the opportunity to create different and separated execution environments for applications that are managed concurrently.

- Differently from hardware virtualization, there is no virtual machine manager or hypervisor and the virtualization is done within a single operating system where the OS kernel allows for multiple isolated user space instances.

- The kernel is also responsible for sharing the system resources among instances and for limiting the impact of instances on each other.

- A user space instance in general contains a proper view of the file system which is completely isolated and separate IP addresses, software configurations and access to devices.

- Operating systems supporting this type of virtualization are general purpose, timeshared operating systems with the capability to provide stronger namespace and resource isolation.

- This virtualization technique can be considered an evolution of the chroot mechanism in Unix systems.

- The chroot operation changes the file system root directory for a process and its children to a specific directory.

- As a result, the process and its children cannot have access to other portions of the file system than those accessible under the new root directory.

- Because Unix systems also expose devices as parts of the file system, by using this method it is possible to completely isolate a set of processes.

- Following the same principle, operating system level virtualization aims to provide separated and multiple execution containers for running applications.

- This technique is an efficient solution for server consolidation scenarios in which multiple application servers share the same technology: operating system, application server framework, and other components.

- Examples of operating system-level virtualizations are FreeBSD Jails, IBM Logical Partition (LPAR), SolarisZones and Containers, Parallels Virtuozzo Containers, OpenVZ, iCore Virtual Accounts, Free Virtual Private Server (FreeVPS), and others.

2.6.4 Programming language-level virtualization

- Programming language level virtualization is mostly used to achieve ease of deployment of applications, managed execution, portability across different platforms and operating systems.

- It consists of a virtual machine executing the byte code of a program which is the result of the compilation process.

- Compilers implemented and used this technology to produce a binary format representing the machine code for an abstract architecture.

- The characteristics of this architecture vary from implementation to implementation.

- Generally these virtual machines constitute a simplification of the underlying hardware instruction set and provide some high level instructions that map some of the features of the languages compiled for them.

- At runtime, the byte code can be either interpreted or compiled on the fly against the underlying hardware instruction set.

- Programming language level virtualization has a long trail in computer science history and originally was used in 1966 for the implementation of Basic Combined Programming Language (BCPL), a language for writing compilers and one of the ancestors of the C programming language.

- Other important examples of the use of this technology have been the UCSD Pascal and Smalltalk.

- Virtual machine programming languages become popular again with Sun's introduction of the Java platform in 1996.

- The Java virtual machine was originally designed for the execution of programs written in the Java language, but other languages such as Python, Pascal, Groovy and Ruby were made available.

- The ability to support multiple programming languages has been one of the key elements of the Common Language Infrastructure (CLI) which is the specification behind .NET Framework.

2.6.5 Application level virtualization

- Application level virtualization is a technique allowing applications to be run in runtime environments that do not natively support all the features required by such applications.

- In this scenario, applications are not installed in the expected runtime environment but are run as though they were.

- In general, these techniques are mostly concerned with partial file systems, libraries, and operating system component emulation.

- Such emulation is performed by a thin layer called a program or an operating system component that is in charge of executing the application.

- Emulation can also be used to execute program binaries compiled for different hardware architectures.

- In this case, one of the following strategies can be implemented:

- Interpretation: In this technique every source instruction is interpreted by an emulator for executing native ISA instructions, leading to poor performance. Interpretation has a minimal startup cost but a huge overhead, since each instruction is emulated.

- Binary translation: In this technique every source instruction is converted to native instructions with equivalent functions. After a block of instructions is translated, it is cached and reused.

- Application virtualization is a good solution in the case of missing libraries in the host operating system.

- In this case a replacement library can be linked with the application or library calls can be remapped to existing functions available in the host system.

- Another advantage is that in this case the virtual machine manager is much lighter since it provides a partial emulation of the runtime environment compared to hardware virtualization.

- Compared to programming level virtualization, which works across all the applications developed for that virtual machine, application level virtualization works for a specific environment.

- It supports all the applications that run on top of a specific environment.

- One of the most popular solutions implementing application virtualization is Wine, which is a software application allowing Unix-like operating systems to execute programs written for the Microsoft Windows platform.

- Wine features a software application acting as a container for the guest application and a set of libraries, called Winelib, that developers can use to compile applications to be ported on Unix systems.

- Wine takes its inspiration from a similar product from Sun, Windows Application Binary Interface (WABI) which implements the Win 16 API specifications on Solaris.

- A similar solution for the Mac OS X environment is CrossOver, which allows running Windows applications directly on the Mac OS X operating system.

- VMware ThinApp is another product in this area, allows capturing the setup of an installed application and packaging it into an executable image isolated from the hosting operating system.

2.6.6 Other types of virtualization

- Other than execution virtualization, other types of virtualization provide an abstract environment to interact with.

- These mainly cover storage, networking, and client/server interaction.

2.6.6.1 Storage virtualization

- Storage virtualization is a system administration practice that allows decoupling the physical organization of the hardware from its logical representation.

- Using this technique, users do not have to be worried about the specific location of their data, which can be identified using a logical path.

- Storage virtualization allows us to harness a wide range of storage facilities and represent them under a single logical file system.

- There are different techniques for storage virtualization, one of the most popular being network based virtualization by means of storage area networks (SANs).

- SANs use a network accessible device through a large bandwidth connection to provide storage facilities.

2.6.6.2 Network virtualization

- Network virtualization combines hardware appliances and specific software for the creation and management of a virtual network.

- Network virtualization can aggregate different physical networks into a single logical network (external network virtualization) or provide network like functionality to an operating system partition (internal network virtualization).

- The result of external network virtualization is generally a virtual LAN (VLAN).

- A VLAN is an aggregation of hosts that communicate with each other as though they were located under the same broadcasting domain.

- Internal network virtualization is generally applied together with hardware and operating system-level virtualization, in which the guests obtain a virtual network interface to communicate with.

- There are several options for implementing internal network virtualization:

    - The guest can share the same network interface of the host and use Network Address Translation (NAT) to access the network;
    - The virtual machine manager can emulate, and install on the host, an additional network device, together with the driver.
    - The guest can have a private network only with the guest.

2.6.6.3 Desktop virtualization

- Desktop virtualization abstracts the desktop environment available on a personal computer in order to provide access to it using a client/server approach.

- Desktop virtualization provides the same outcome of hardware virtualization but serves a different purpose.

- Similarly to hardware virtualization, desktop virtualization makes accessible a different system as though it were natively installed on the host but this system is remotely stored on a different host and accessed through a network connection.

- Moreover, desktop virtualization addresses the problem of making the same desktop environment accessible from everywhere.

- Although the term desktop virtualization strictly refers to the ability to remotely access a desktop environment, generally the desktop environment is stored in a remote server or a data center that provides a high availability infrastructure and ensures the accessibility and persistence of the data.

- In this scenario, an infrastructure supporting hardware virtualization is fundamental to provide access to multiple desktop environments hosted on the same server.

- A specific desktop environment is stored in a virtual machine image that is loaded and started on demand when a client connects to the desktop environment.

- This is a typical cloud computing scenario in which the user leverages the virtual infrastructure for performing the daily tasks on his computer.

- The advantages of desktop virtualization are high availability, persistence, accessibility, and ease of management.

- The basic services for remotely accessing a desktop environment are implemented in software components such as Windows Remote Services, VNC, and X Server.

- Infrastructures for desktop virtualization based on cloud computing solutions include Sun Virtual Desktop Infrastructure (VDI), Parallels Virtual Desktop Infrastructure (VDI), Citrix XenDesktop, and others.

2.6.6.4 Application server virtualization

- Application server virtualization abstracts a collection of application servers that provide the same services as a single virtual application server by using load balancing strategies and providing a high availability infrastructure for the services hosted in the application server.

- This is a particular form of virtualization and serves the same purpose of storage virtualization by providing a better quality of service rather than emulating a different environment.

2.7 Implementation Levels of Virtualization

- Virtualization is a computer architecture technology by which multiple virtual machines (VMs) are multiplexed in the same hardware machine.

- The purpose of a VM is to enhance resource sharing by many users and improve computer performance in terms of resource utilization and application flexibility.

- Hardware resources (CPU, memory, I/O devices) or software resources (operating system and software libraries) can be virtualized in various functional layers.

- The idea is to separate the hardware from the software to yield better system efficiency. For example, computer users gained access to much enlarged memory space when the concept of virtual memory was introduced.

- Similarly, virtualization techniques can be applied to enhance the use of compute engines, networks and storage.

- With sufficient storage, any computer platform can be installed in another host computer, even if they use processors with different instruction sets and run with distinct operating systems on the same hardware.

2.7.1 Levels of virtualization implementation

- A traditional computer runs with a host operating system specially tailored for its hardware architecture, as shown in Figure 2.7(a).

- After virtualization, different user applications managed by their own operating systems (guest OS) can run on the same hardware, independent of the host OS. This is often done by adding additional software, called a virtualization layer as shown in Figure 2.7(b).

- This virtualization layer is known as hypervisor or virtual machine monitor (VMM). The VMs are shown in the upper boxes, where applications run with their own guest OS over the virtualized CPU, memory, and I/O resources.

- The main function of the software layer for virtualization is to virtualize the physical hardware of a host machine into virtual resources to be used by the VMs, exclusively. This can be implemented at various operational levels, as we will discuss shortly.

- The virtualization software creates the abstraction of VMs by interposing a virtualization layer at various levels of a computer system.

- Common virtualization layers include the instruction set architecture (ISA) level, hardware level, operating system level, library support level, and application level.

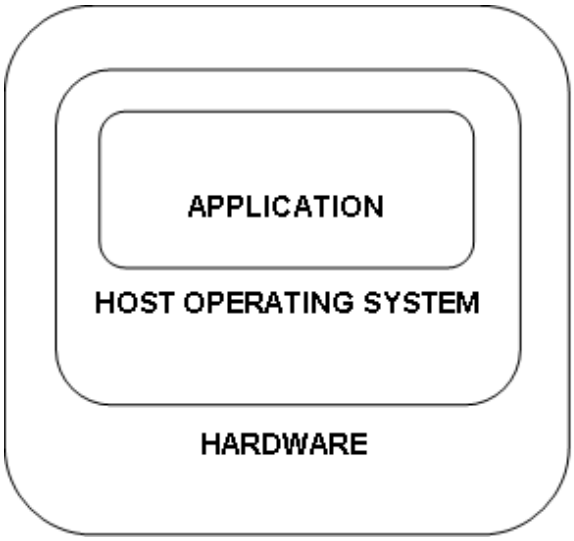


Figure 2.7 (a) Traditional Computer

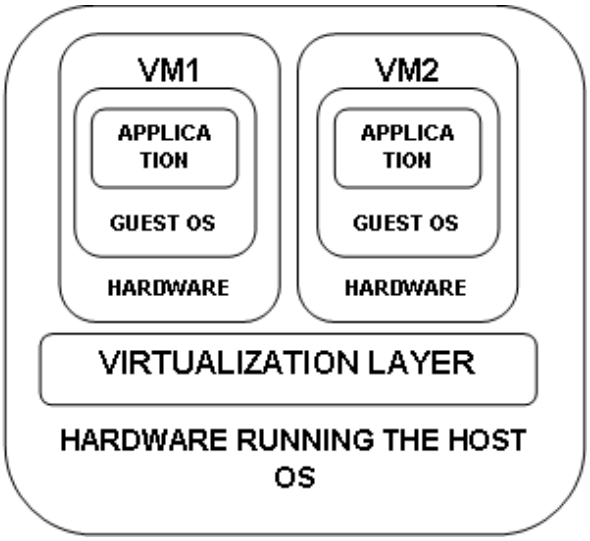


Figure (b) After Virtualization

2.7.2 Instruction set architecture level

- At the ISA level, virtualization is performed by emulating a given ISA by the ISA of the host machine. For example, MIPS binary code can run on an x86-based host machine with the help of ISA emulation.

- With this approach, it is possible to run a large amount of legacy binary code written for various processors on any given new hardware host machine. Instruction set emulation leads to virtual ISAs created on any hardware machine.

- The basic emulation method is through code interpretation. An interpreter program interprets the source instructions to target instructions one by one.

- One source instruction may require tens or hundreds of native target instructions to perform its function. This process is relatively slow.

- For better performance, dynamic binary translation is desired. This approach translates basic blocks of dynamic source instructions to target instructions.

- The basic blocks can also be extended to program traces or super blocks to increase translation efficiency.

- Instruction set emulation requires binary translation and optimization.

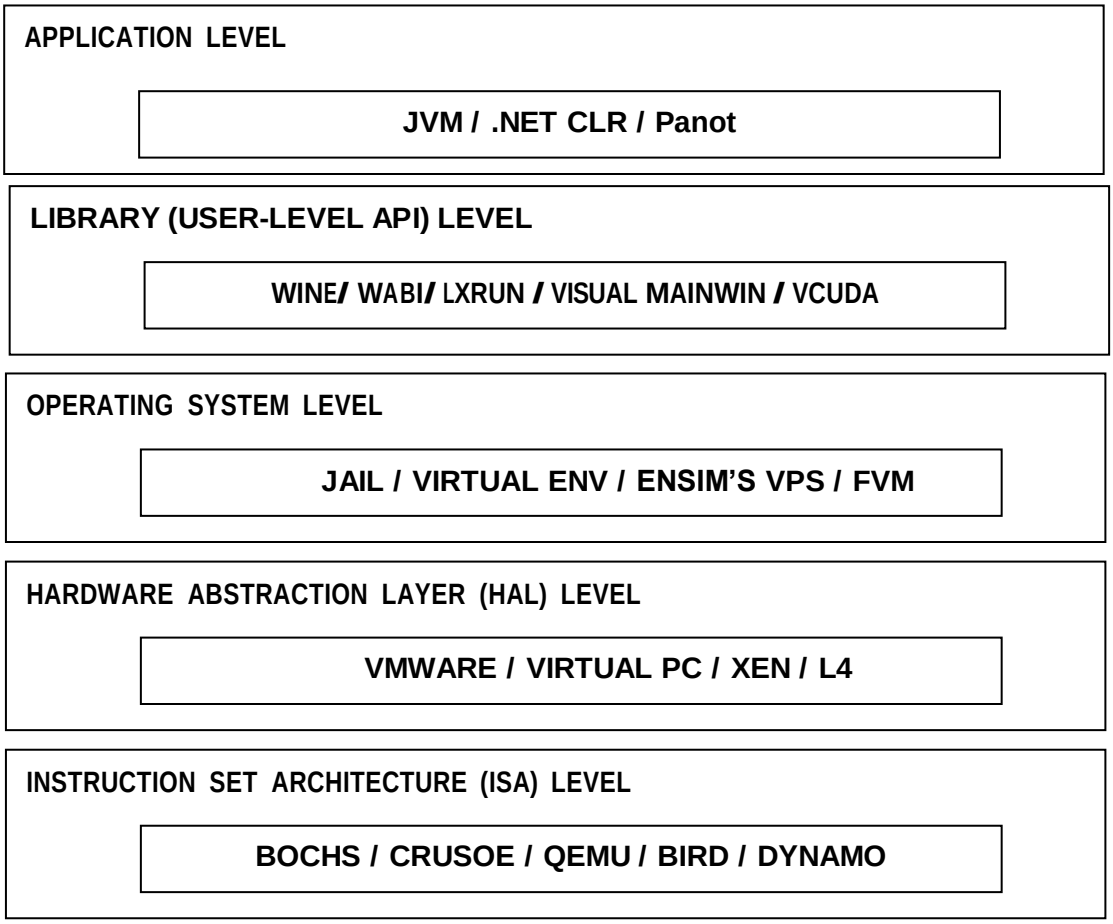


Figure 2.8 Virtualization ranging from hardware to applications in five abstraction levels

2.7.3 Hardware abstraction level

- Hardware-level virtualization is performed right on top of the bare hardware. On the one hand, this approach generates a virtual hardware environment for a VM. On the other hand, the process manages the underlying hardware through virtualization.

- The idea is to virtualize a computer's resources, such as its processors, memory, and I/O devices.

- The intention is to upgrade the hardware utilization rate bymultiple users concurrently.

- The idea was implemented in the IBM VM/370 in the 1960s. Xen hypervisor has been applied to virtualize x86-based machines to run Linux or other guest OS applications.

2.7.4 Operating system level

- This refers to an abstraction layer between traditional OS and user applications. OS-level virtualization creates isolated containers on a single physical server and the OS instances to utilize the hardware and software in data centers.

- The containers behave like real servers. OS-level virtualization is commonly used in creating virtual hosting environments to allocate hardware resources among a large number of mutually distrusting users.

- It is also used, to a lesser extent, in consolidating server hardware by moving services on separate hosts into containers or VMs on one server.

- Operating system virtualization inserts a virtualization layer inside an operating system to partition a machine's physical resources. It enables multiple isolated VMs within a single operating system kernel.

- This kind of VM is often called a virtual execution environment (VE), Virtual Private System (VPS), or simply container.

- Compared to hardware-level virtualization, the benefits of OS extensions are twofold: VMs at the operating system level have minimal startup/shutdown costs, low resource requirements, and high scalability

- For an OS-level VM, it is possible for a VM and its host environment to synchronize state changes when necessary.

- These benefits can be achieved via two mechanisms of OS-level virtualization:

- All OS-level VMs on the same physical machine share a single operating system kernel
    - The virtualization layer can be designed in a way that allows processes in VMs to access as many resources of the host machine as possible, but never to modify them.
- Virtualization Support for the Linux Platform
    - OpenVZ is an OS-level tool designed to support Linux platforms to create virtual environments for running VMs under different guest Operating Systems.
    - OpenVZ is an open source container-based virtualization solution built on Linux. To support virtualization and isolation of various subsystems, limited resource management, and check pointing, OpenVZ modifies the Linux kernel.

2.7.5 Library support level

- Most applications use APIs exported by user level libraries rather than using lengthy system calls by the OS. Since most systems provide well-documented APIs, such an interface becomes another candidate for virtualization.
- Virtualization with library interfaces is possible by controlling the communication link between applications and the rest of a system through API hooks.
- The software tool WINE has implemented this approach to support Windows applications on top of UNIX hosts. Another example is the vCUDA which allows applications executing within VMs to leverage GPU hardware acceleration.
- Library level virtualization is also known as user-level Application Binary Interface (ABI) or API emulation.
- This type of virtualization can create execution environments for running alien programs on a platform rather than creating a VM to run the entire operating system.

- API call interception and remapping are the key functions performed. The WABI offers middleware to convert Windows system calls to Solaris system calls.

- Lxrun is really a system call emulator that enables Linux applications written for x86 hosts to run on UNIX systems.

- Similarly, Wine offers library support for virtualizing x86 processors to run Windows applications on UNIX hosts.

- Visual MainWin offers a compiler support system to develop Windows applications using Visual Studio to run on some UNIX hosts.

- The vCUDA for Virtualization of General-Purpose GPUs. CUDA is a programming model and library for general-purpose GPUs. It leverages the high performance of GPUs to run compute-intensive applications on host operating systems. However, it is difficult to run CUDA applications on hardware-level VMs directly.

- vCUDA virtualizes the CUDA library and can be installed on guest OSes. When CUDA applications run on a guest OS and issue a call to the CUDA API, vCUDA intercepts the call and redirects it to the CUDA API running on the host OS.

2.7.6 User application level

- Virtualization at the application level virtualizes an application as a VM.

- On a traditional OS, an application often runs as a process. Therefore, application-level virtualization is also known as process level virtualization. The most popular approach is to deploy high level language (HLL) VMs.

- In this scenario, the virtualization layer sits as an application program on top of the operating system, and the layer exports an abstraction of a VM that can run programs written and compiled to a particular abstract machine definition.

- Any program written in the HLL and compiled for this VM will be able to run on it. The Microsoft .NET CLR and Java Virtual Machine (JVM) are two good examples of this class of VM.

- Other forms of application-level virtualization are known as application isolation, application sandboxing, or application streaming.

- The process involves wrapping the application in a layer that is isolated from the host OS and other applications. The result is an application that is much easier to distribute and remove from user workstations.

- An example is the LANDesk application virtualization platform which deploys software applications as self contained, executable files in an isolated environment without requiring installation, system modifications or elevated security privileges.

2.7.7 Relative merits of different approaches

Level of Implementation	Higher Performance	Application Flexibility	Implementation Complexity	Application Isolation
Instruction Set Architecture	Very Low	Very High	Moderate	Moderate
Hardware-level virtualization	Very High	Moderate	Very High	High
OS-level virtualization	Very High	Low	Moderate	Low
Library support level	Moderate	Low	Low	Low
User application level	Low	Low	Very High	Very High

Table 2.1 Relative Merits of Virtualization at Various Levels

2.8 Virtualization Structures, Tools and Mechanisms

- In general, there are three typical classes of VM architecture.
- The virtualization layer is responsible for converting portions of the real hardware into virtual hardware.
- Therefore, different operating systems such as Linux and Windows can run on the same physical machine, simultaneously.
- Depending on the position of the virtualization layer, there are several classes of VM architectures, namely the hypervisor architecture, paravirtualization and host based virtualization.
- The hypervisor is also known as the VMM (Virtual Machine Monitor). They both perform the same virtualization operations.

2.8.1 Hypervisor and Xen architecture

- The hypervisor supports hardware level virtualization on bare metal devices like CPU, memory, disk and network interfaces.
- The hypervisor software sits directly between the physical hardware and its OS. This virtualization layer is referred to as either the VMM or the hypervisor.
- The hypervisor provides hypercalls for the guest OSes and applications.
- Depending on the functionality, a hypervisor can assume micro kernel architecture like the Microsoft Hyper-V.
- It can assume monolithic hypervisor architecture like the VMware ESX for server virtualization.

- A micro kernel hypervisor includes only the basic and unchanging functions (such as physical memory management and processor scheduling).

- The device drivers and other changeable components are outside the hypervisor.

- A monolithic hypervisor implements all the aforementioned functions, including those of the device drivers. Therefore, the size of the hypervisor code of a micro-kernel hypervisor is smaller than that of a monolithic hypervisor.

- Essentially, a hypervisor must be able to convert physical devices into virtual resources dedicated for the deployed VM to use.

2.8.2 Xen architecture

- Xen is an open source hypervisor program developed by Cambridge University.

- Xen is a microkernel hypervisor, which separates the policy from the mechanism.

- The Xen hypervisor implements all the mechanisms, leaving the policy to be handled by Domain 0. Figure 2.9 shows architecture of Xen hypervisor.

- Xen does not include any device drivers natively. It just provides a mechanism by which a guest OS can have direct access to the physical devices.

- As a result, the size of the Xen hypervisor is kept rather small.

- Xen provides a virtual environment located between the hardware and the OS.

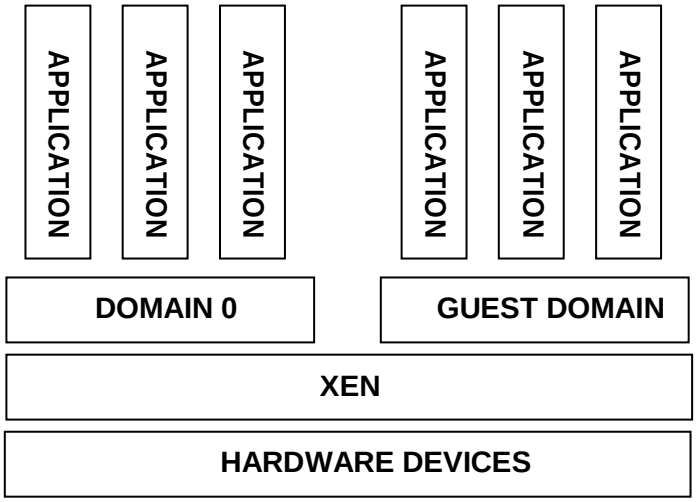


Figure 2.9 Xen domain 0 for control and I/O & guest domain for user applications.

- The core components of a Xen system are the hypervisor, kernel, and applications.
- The organization of the three components is important.
- Like other virtualization systems, many guest OSes can run on top of the hypervisor.
- However, not all guest OSes are created equal, and one in particular controls the others.
- The guest OS, which has control ability, is called Domain 0, and the others are called Domain U.
- Domain 0 is a privileged guest OS of Xen. It is first loaded when Xen boots without any file system drivers being available.
- Domain 0 is designed to access hardware directly and manage devices. Therefore, one of the responsibilities of Domain 0 is to allocate and map hardware resources for the guest domains (the Domain U domains).

- For example, Xen is based on Linux and its security level is C2. Its management VM is named Domain 0 which has the privilege to manage other VMs implemented on the same host.

- If Domain 0 is compromised, the hacker can control the entire system. So, in the VM system, security policies are needed to improve the security of Domain 0.

- Domain 0, behaving as a VMM, allows users to create, copy, save, read, modify, share, migrate and roll back VMs as easily as manipulating a file, which flexibly provides tremendous benefits for users.

2.8.3 Binary translation with full virtualization

- Depending on implementation technologies, hardware virtualization can be classified into two categories: full virtualization and host based virtualization.

- Full virtualization does not need to modify the host OS. It relies on binary translation to trap and to virtualize the execution of certain sensitive, non virtualizable instructions.

- The guest OSes and their applications consist of noncritical and critical instructions.

- In a host-based system, both a host OS and a guest OS are used.

- A virtualization software layer is built between the host OS and guest OS.

- With full virtualization, noncritical instructions run on the hardware directly while critical instructions are discovered and replaced with traps into the VMM to be emulated by software.

- Both the hypervisor and VMM approaches are considered full virtualization.

- The VMM scans the instruction stream and identifies the privileged, control and behavior sensitive instructions. When these instructions are identified, they are trapped into the VMM, which emulates the behavior of these instructions.

- The method used in this emulation is called binary translation.

- Full virtualization combines binary translation and direct execution.

- An alternative VM architecture is to install a virtualization layer on top of the host OS.

- This host OS is still responsible for managing the hardware.

- The guest OSes are installed and run on top of the virtualization layer. Dedicated applications may run on the VMs. Certainly, some other applications can also run with the host OS directly.

- Host based architecture has some distinct advantages, as enumerated next.

    - First, the user can install this VM architecture without modifying the host OS.
    - Second, the host-based approach appeals to many host machine configurations.

2.8.4 Paravirtualization with compiler support

- When x86 processor is virtualized, a virtualization layer is inserted between the hardware and the OS.

- According to the x86 ring definitions, the virtualization layer should also be installed at Ring 0. Different instructions at Ring 0 may cause some problems.

- Although paravirtualization reduces the overhead, it has incurred other problems.

    - First, its compatibility and portability may be in doubt, because it must support the unmodified OS as well.

    - Second, the cost of maintaining paravirtualized OSes is high, because they may require deep OS kernel modifications.
    - Finally, the performance advantage of paravirtualization varies greatly due to workload variations.

- Compared with full virtualization, paravirtualization is relatively easy and more practical. The main problem in full virtualization is its low performance in binary translation.

- KVM is a Linux paravirtualization system. It is a part of the Linux version 2.6.20 kernel.

- In KVM, Memory management and scheduling activities are carried out by the existing Linux kernel.

- The KVM does the rest, which makes it simpler than the hypervisor that controls the entire machine.

- KVM is a hardware assisted and paravirtualization tool, which improves performance and supports unmodified guest OSes such as Windows, Linux, Solaris, and other UNIX variants.

- Unlike the full virtualization architecture which intercepts and emulates privileged and sensitive instructions at runtime, paravirtualization handles these instructions at compile time.

- The guest OS kernel is modified to replace the privileged and sensitive instructions with hypercalls to the hypervisor or VMM. Xen assumes such paravirtualization architecture.

- The guest OS running in a guest domain may run at Ring 1 instead of at Ring 0. This implies that the guest OS may not be able to execute some privileged and sensitive instructions. The privileged instructions are implemented by hypercalls to the hypervisor.

2.9 Virtualization of CPU, Memory and I/O Devices

- To support virtualization, processors such as the x86 employ a special running mode and instructions known as hardware assisted virtualization.

- For the x86 architecture, Intel and AMD have proprietary technologies for hardware assisted virtualization.

2.9.1 Hardware support for virtualization

- Modern operating systems and processors permit multiple processes to run simultaneously. If there is no protection mechanism in a processor, all instructions from different processes will access the hardware directly and cause a system crash.

- All processors have at least two modes, user mode and supervisor mode, to ensure controlled access of critical hardware.

- Instructions running in supervisor mode are called privileged instructions. Other instructions are unprivileged instructions.

- In a virtualized environment, it is more difficult to make OSes and applications run correctly because there are more layers in the machine stack.

- At the time of this writing, many hardware virtualization products were available.

- The VMware Workstation is a VM software suite for x86 and x86-64 computers.

- This software suite allows users to set up multiple x86 and x86-64 virtual computers and to use one or more of these VMs simultaneously with the host operating system.

- The VMware Workstation assumes the host-based virtualization.

- Xen is a hypervisor for use in IA-32, x86-64, Itanium and PowerPC 970 hosts.

- One or more guest OS can run on top of the hypervisor.

- KVM is a Linux kernel virtualization infrastructure.

- KVM can support hardware assisted virtualization and paravirtualization by using the Intel VT-x or AMD-v and VirtIO framework, respectively.

- The VirtIO framework includes a paravirtual Ethernet card, a disk I/O controller and a balloon device for adjusting guest memory usage and a VGA graphics interface using VMware drivers.

2.9.2 CPU virtualization

- A VM is a duplicate of an existing computer system in which a majority of the VM instructions are executed on the host processor in native mode.

- The unprivileged instructions of VMs run directly on the host machine for higher efficiency.

- The critical instructions are divided into three categories: privileged instructions, control sensitive instructions, and behavior sensitive instructions.

- Privileged instructions execute in a privileged mode and will be trapped if executed outside this mode.

- Control sensitive instructions attempt to change the configuration of resources used.

- Behavior sensitive instructions have different behaviors depending on the configuration of resources, including the load and store operations over the virtual memory.

- CPU architecture is virtualizable if it supports the ability to run the VM's privileged and unprivileged instructions in the CPU's user mode while the VMM runs in supervisor mode.

- The privileged instructions including control and behavior sensitive instructions of a VM are executed; they are trapped in the VMM.

- RISC CPU architectures can be naturally virtualized because all control and behavior sensitive instructions are privileged instructions.

- The x86 CPU architectures are not primarily designed to support virtualization.

2.9.2.1 Hardware-assisted CPU virtualization

- This technique attempts to simplify virtualization because full or paravirtualization is complicated.

- Intel and AMD add an additional mode called privilege mode level (some people call it Ring-1) to x86 processors.

- Therefore, operating systems can still run at Ring 0 and hypervisor can run at Ring 1.

- All the privileged and sensitive instructions are trapped in the hypervisor automatically.

- This technique removes the difficulty of implementing binary translation of full virtualization.

- It also lets the operating system run in VMs without modification.

2.9.3 Memory virtualization

- Virtual memory virtualization is similar to the virtual memory support provided by modern operating systems.

- In a traditional execution environment, the operating system maintains mappings of virtual memory to machine memory using page tables, which is a one stage mapping from virtual memory to machine memory.

- All modern x86 CPUs include a memory management unit (MMU) and a translation lookaside buffer (TLB) to optimize virtual memory performance.

- However, in a virtual execution environment, virtual memory virtualization involves sharing the physical system memory in RAM and dynamically allocating it to the physical memory of the VMs.

- That means a two stage mapping process should be maintained by the guest OS and the VMM, respectively: virtual memory to physical memory and physical memory to machine memory.

- MMU virtualization should be supported, which is transparent to the guest OS.

- The guest OS continues to control the mapping of virtual addresses to the physical memory addresses of VMs.

- But the guest OS cannot directly access the actual machine memory.

- The VMM is responsible for mapping the guest physical memory to the actual machine memory.

- Since each page table of the guest OSes has a separate page table in the VMM corresponding to it, the VMM page table is called the shadow page table.

- Nested page tables add another layer of indirection to virtual memory.

- The MMU already handles virtual-to-physical translations as defined by the OS. Then the physical memory addresses are translated to machine addresses using another set of page tables defined by the hypervisor.

- VMware uses shadow page tables to perform virtual-memory-to-machine-memory address translation.

- Processors use TLB hardware to map the virtual memory directly to the machine memory to avoid the two levels of translation on every access.

- When the guest OS changes the virtual memory to a physical memory mapping, the VMM updates the shadow page tables to enable a direct lookup.

- The AMD Barcelona processor has featured hardware assisted memory virtualization since 2007.

- It provides hardware assistance to the two stage address translation in a virtual execution environment by using a technology called nested paging.

2.9.4 I/O virtualization

- I/O virtualization involves managing the routing of I/O requests between virtual devices and the shared physical hardware.

- There are three ways to implement I/O virtualization: full device emulation, paravirtualization, and direct I/O.

- Full device emulation is the first approach for I/O virtualization. Generally, this approach emulates well known and real world devices.

- All the functions of a device or bus infrastructure, such as device enumeration, identification, interrupts, and DMA are replicated in software.

- This software is located in the VMM and acts as a virtual device.

- The I/O access requests of the guest OS are trapped in the VMM which interacts with the I/O devices.

- A single hardware device can be shared by multiple VMs that run concurrently. However, software emulation runs much slower than the hardware it emulates.

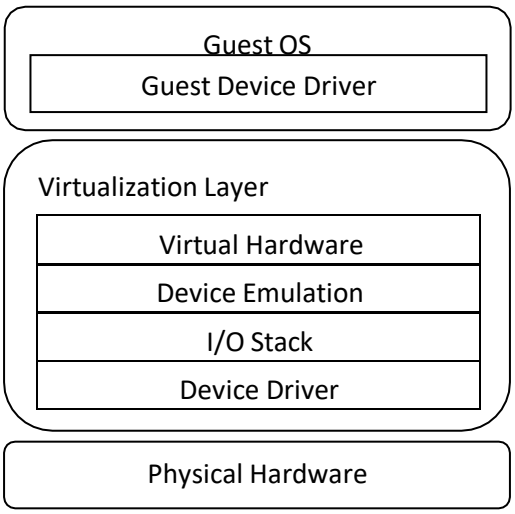


Figure 2.10 Device emulation for I/O Virtualization

- The paravirtualization method of I/O virtualization is typically used in Xen. It is also known as the split driver model consisting of a frontend driver and a backend driver.

- The frontend driver is running in Domain U and the backend driver is running in Domain 0. They interact with each other via a block of shared memory.

- The frontend driver manages the I/O requests of the guest OSes and the backend driver is responsible for managing the real I/O devices and multiplexing the I/O data of different VMs.

- Para I/O-virtualization achieves better device performance than full device emulation, it comes with a higher CPU overhead.

- Direct I/O virtualization lets the VM access devices directly. It can achieve close-to-native performance without high CPU costs.

- However, current direct I/O virtualization implementations focus on networking for mainframes. There are a lot of challenges for commodity hardware devices.

- For example, when a physical device is reclaimed (required by workload migration) for later reassignment, it may have been set to an arbitrary state (e.g., DMA to some arbitrary memory locations) that can function incorrectly or even crash the whole system.

- Since software based I/O virtualization requires a very high overhead of device emulation, hardware-assisted I/O virtualization is critical.

- Intel VT-d supports the remapping of I/O DMA transfers and device generated interrupts. The architecture of VT-d provides the flexibility to support multiple usage models that may run unmodified, special-purpose, or "virtualization-aware" guest OSes.

- Another way to help I/O virtualization is via self virtualized I/O (SV-IO).

- The key idea of SV-IO is to harness the rich resources of a multicore processor. All tasks associated with virtualizing an I/O device are encapsulated in SV-IO.

- It provides virtual devices and an associated access API to VMs and a management API to the VMM.

- SV-IO defines one virtual interface (VIF) for every kind of virtualized I/O device, such as virtual network interfaces, virtual block devices (disk), virtual camera devices,

2.9.4.1 Virtualization in multi-core processors

- Virtualizing a multi-core processor is relatively more complicated than virtualizing a uni-core processor.

- Muti-core virtualization has raised some new challenges to computer architects, compiler constructors, system designers, and application programmers.

- There are mainly two difficulties: Application programs must be parallelized to use all cores fully, and software must explicitly assign tasks to the cores, which is a very complex problem.

    - The first challenge, new programming models, languages, and libraries are needed to make parallel programming easier.
    - The second challenge has spawned research involving scheduling algorithms and resource management policies.

- Dynamic heterogeneity is emerging to mix the fat CPU core and thin GPU cores on the same chip, which further complicates the multi core or many core resource management.

- The dynamic heterogeneity of hardware infrastructure mainly comes from less reliable transistors and increased complexity in using the transistors.

2.9.4.2 Physical versus virtual processor cores

- A multicore virtualization method to allow hardware designers to get an abstraction of the low-level details of the processor cores.

- This technique alleviates the burden and inefficiency of managing hardware resources by software.

- It is located under the ISA and remains unmodified by the operating system or VMM (hypervisor).

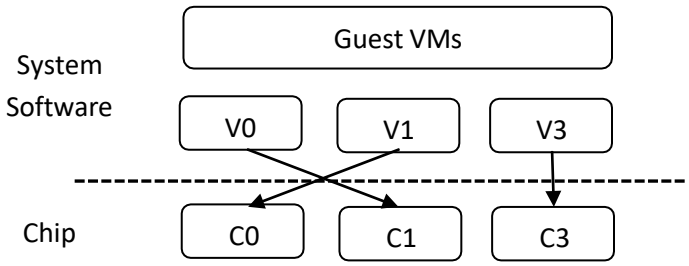


Figure 2.11 Multicore Virtualization method

- Figure 2.11 illustrates the technique of software visible VCPU moving from one core to another and temporarily suspending execution of a VCPU when there are no appropriate cores on which it can run.

2.9.4.3 Virtual hierarchy

- The emerging many core chip multiprocessors (CMPs) provide a new computing landscape.

- Instead of supporting time sharing jobs on one or a few cores, we can use the abundant cores in a space sharing, where single threaded or multithreaded jobs are simultaneously assigned to separate groups of cores for long time intervals.

- To optimize for space shared workloads, they propose using virtual hierarchies to overlay a coherence and caching hierarchy onto a physical processor.

- A virtual hierarchy is a cache hierarchy that can adapt to fit the workload or mix of workloads.

- The hierarchy's first level locates data blocks close to the cores needing them for faster access, establishes a shared-cache domain and establishes a point of coherence for faster communication.

- When a miss leaves a tile, it first attempts to locate the block (or sharers) within the first level. The first level can also provide isolation between independent workloads. A miss at the L1 cache can invoke the L2 access.

- Space sharing is applied to assign three workloads to three clusters of virtual cores:

    - Namely VM0 and VM3 for database workload, VM1 and VM2 for web server workload and VM4–VM7 for middleware workload.

- Each VM operates in a isolated fashion at the first level. This will minimize both miss access time and performance interference with other workloads or VMs.

- The shared resources of cache capacity, inter-connect links, and miss handling are mostly isolated between VMs. The second level maintains a globally shared memory.

- This facilitates dynamically repartitioning resources without costly cache flushes. A virtual hierarchy adapts to space-shared workloads like multiprogramming and server consolidation.

2.10 Virtualization Support and Disaster Recovery

- One very distinguishing feature of cloud computing infrastructure is the use of system virtualization and the modification to provisioning tools.

- Virtualization of servers on a shared cluster can consolidate web services.

- In cloud computing, virtualization also means the resources and fundamental infrastructure are virtualized.

- The user will not care about the computing resources that are used for providing the services.

- Cloud users do not need to know and have no way to discover physical resources that are involved while processing a service request.

- In addition, application developers do not care about some infrastructure issues such as scalability and fault tolerance. Application developers focus on service logic.

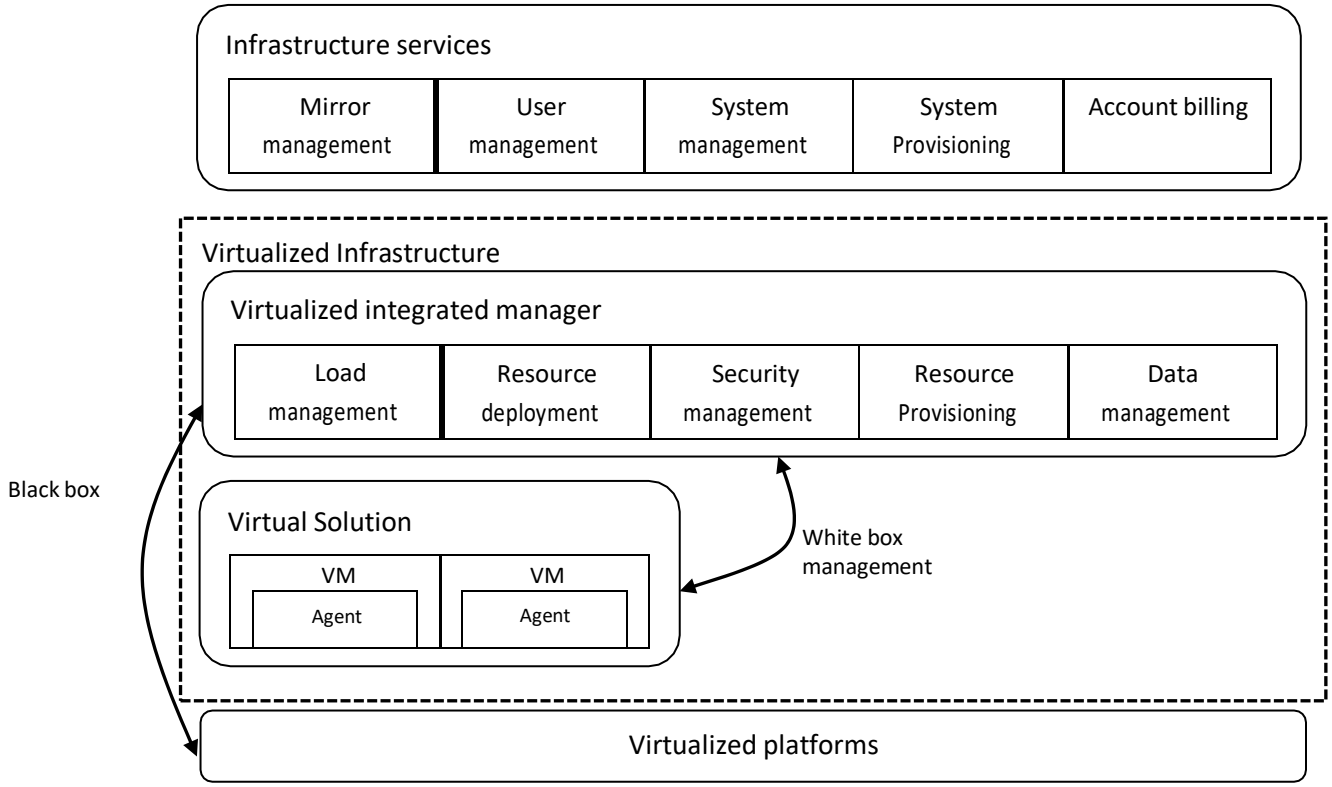


Figure 2.12 Virtualized servers, storage, and network for cloud platform construction

- In many cloud computing systems, virtualization software is used to virtualize the hardware.

- System virtualization software is a special kind of software which simulates the execution of hardware and runs even unmodified operating systems.

- Cloud computing systems use virtualization soware as the running environment for legacy software such as old operating systems and unusual applications.

2.10.1 Hardware Virtualization

- Virtualization software is also used as the platform for developing new cloud applications that enable developers to use any operating systems and programming environments they like.

- The development environment and deployment environment can now be the same, which eliminates some runtime problems.

- VMs provide flexible runtime services to free users from worrying about the system environment.

- Using VMs in a cloud computing platform ensures extreme flexibility for users. As the computing resources are shared by many users, a method is required to maximize the user's privileges and still keep them separated safely.

- Traditional sharing of cluster resources depends on the user and group mechanism on a system.

    - Such sharing is not flexible.
    - Users cannot customize the system for their special purposes.
    - Operating systems cannot be changed.
    - The separation is not complete.

- An environment that meets one user's requirements often cannot satisfy another user. Virtualization allows us to have full privileges while keeping them separate.

- Users have full access to their own VMs, which are completely separate from other user's VMs.

- Multiple VMs can be mounted on the same physical server. Different VMs may run with different OSes.

- The virtualized resources form a resource pool.

- The virtualization is carried out by special servers dedicated to generating the virtualized resource pool.

- The virtualized infrastructure (black box in the middle) is built with many virtualizing integration managers.

- These managers handle loads, resources, security, data, and provisioning functions. Figure 2.13 shows two VM platforms.

- Each platform carries out a virtual solution to a user job. All cloud services are managed in the boxes at the top.

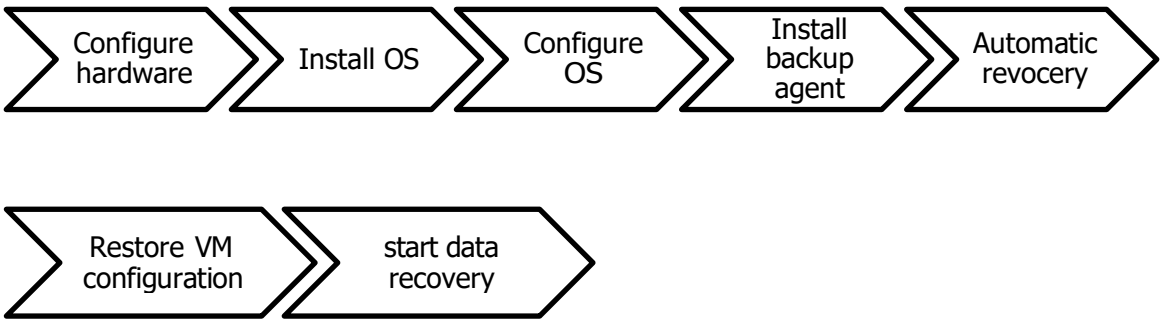


Figure 2.13 Conventional disaster recover scheme versus live migration of VMs

2.10.2 Virtualization Support in Public Clouds

- AWS provides extreme flexibility (VMs) for users to execute their own applications.
- GAE provides limited application level virtualization for users to build applications only based on the services that are created by Google.
- Microsoft provides programming level virtualization (.NET virtualization) for users to build their applications.
- The VMware tools apply to workstations, servers, and virtual infrastructure.
- The Microsoft tools are used on PCs and some special servers.
- The XenEnterprise tool applies only to Xen-based servers.

2.10.3 Virtualization for IaaS

- VM technology has increased in ubiquity.
- This has enabled users to create customized environments atop physical infrastructure for cloud computing.
- Use of VMs in clouds has the following distinct benefits:
    - System administrators consolidate workloads of underutilized servers in fewer servers
    - VMs have the ability to run legacy code without interfering with other APIs
    - VMs can be used to improve security through creation of sandboxes for running applications with questionable reliability

- Virtualized cloud platforms can apply performance isolation, letting providers offer some guarantees and better QoS to customer applications.

2.10.4 VM Cloning for Disaster Recovery

- VM technology requires an advanced disaster recovery scheme.
  - One scheme is to recover one physical machine by another physical machine.
  - The second scheme is to recover one VM by another VM.
- As shown in the top timeline of Figure 2.13, traditional disaster recovery from one physical machine to another is rather slow, complex, and expensive.
- Total recovery time is attributed to the hardware configuration, installing and configuring the OS, installing the backup agents and the long time to restart the physical machine.
- To recover a VM platform, the installation and configuration times for the OS and backup agents are eliminated.
- Virtualization aids in fast disaster recovery by VM encapsulation.
- The cloning of VMs offers an effective solution.
- The idea is to make a clone VM on a remote server for every running VM on a local server.
- Among the entire clone VMs, only one needs to be active.
- The remote VM should be in a suspended mode.

- A cloud control center should be able to activate this clone VM in case of failure of the original VM, taking a snapshot of the VM to enable live migration in a minimal amount of time.

- The migrated VM can run on a shared Internet connection. Only updated data and modified states are sent to the suspended VM to update its state.

- The Recovery Property Objective (RPO) and Recovery Time Objective (RTO) are affected by the number of snapshots taken.

- Security of the VMs should be enforced during live migration of VMs.

TWO MARK QUESTIONS

1. List the four major characteristics to identify a service.

   - Boundaries are explicit.
   - Services are autonomous.
   - Services share schema and contracts,
   - Services compatibility is determined based on policy.

2. Define SOA.

   - Service Oriented architecture is an architectural style supporting service orientation.
   - It organizes a software system into a collection of interacting services.
   - SOA encompasses a set of design principles that structure system development and provide means for integrating components into a coherent and decentralized system.

3. List the two major roles in SOA.

   - The service provider and the service consumer.
   - The service provider is the maintainer of the service and the organization that makes available one or more services for others to use.
   - The service consumer can locate the service metadata in the registry and develop the required client components to bind and use the service.

4. Characterize SOA platforms within an enterprise context.

   - Standardized service contract
   - Loose coupling
   - Abstraction
   - Reusability
   - Autonomy
   - Lack of state
   - Discoverability

5. Define Web services.

- Web services are the prominent technology for implementing SOA systems and applications.
- The concept behind a Web service is very simple.
- Using as a basis the object oriented abstraction, a Web service exposes a set of operations that can be invoked by leveraging Internet based protocols.

6. List the aspects that make Web services the technology of choice for SOA.

- First, they allow for interoperability across different platforms and programming languages.
- Second, they are based on well-known and vendor-independent standards such as HTTP, SOAP, XML, and WSDL.
- Third, they provide an intuitive and simple way to connect heterogeneous software systems
- Finally, they provide the features required by enterprise business applications to be used in an industrial environment.

7. What is the purpose of WSDL and UDDI?

- The service description document, expressed by means of Web Service Definition Language (WSDL), can be either uploaded to a global registry or attached as a metadata to the service itself.
- Service consumers can look up and discover services in global catalogs using Universal Description Discovery and Integration (UDDI) or, most likely, directly retrieve the service metadata by interrogating the Web service first.

8. What is SOAP?

- Simple Object Access Protocol (SOAP), an XML-based language for exchanging structured information in a platform-independent manner, constitutes the protocol used for Web service method invocation.

9. Write short note on RESTful systems.

- Representational State Transfer (REST) provides a model for designing network-based software systems utilizing the client/ server model and leverages the facilities provided by HTTP for IPC without additional burden.

10. What is Publish-and-subscribe message model?

- Publish-and-subscribe message model introduces a different message passing strategy, one that is based on notification among components.
- There are two major roles: The publisher and the subscriber.
- It is very suitable for implementing systems based on the one-to-many communication model

11. List the merits of Virtualization.

- Virtualization technology is one of the fundamental components of cloud computing, especially in regard to infrastructure-based services.
- Virtualization allows the creation of a secure, customizable, and isolated execution environment for running applications, even if they are untrusted, without affecting other users' applications.

12. List characteristics of virtualized environments.

- Increased security
    - Sharing
    - Aggregation
    - Emulation

    - Isolation
- Performance tuning.
- Portability

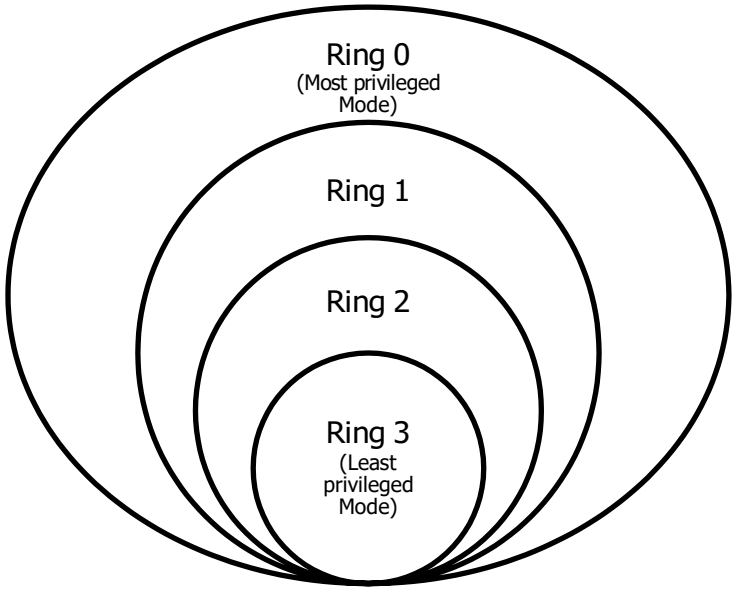
13. Categorize execution virtualization techniques.

- Process level techniques are implemented on top of an existing operating system, which has full control of the hardware.
- System level techniques are implemented directly on hardware and do not require or require a minimum of support from an existing operating system.

14. Differentiate between privileged and non privileged instructions.

- Non privileged instructions are those instructions that can be used without interfering with other tasks because they do not access shared resources.
- Privileged instructions are those that are executed under specific restrictions and are mostly used for sensitive operations, which expose (behavior-sensitive) or modify (control-sensitive) the privileged state.

15. Illustrate ring based security.



16. What is Hardware-level virtualization?

- Hardware-level virtualization is a virtualization technique that provides an abstract execution environment in terms of computer hardware on top of which a guest operating system can be run.

17. Define hypervisor.

- The hypervisor is generally a program or a combination of software and hardware that allows the abstraction of the underlying physical hardware.
- Hypervisors is a fundamental element of hardware virtualization is the hypervisor, or virtual machine manager (VMM).

18. List the types of hypervisor.

- Type I hypervisors run directly on top of the hardware.
- Type II hypervisors require the support of an operating system to provide virtualization services.

19. What is hardware assisted virtualization?

- This term refers to a scenario in which the hardware provides architectural support for building a virtual machine manager able to run a guest operating system in complete isolation.
- This technique was originally introduced in the IBM System/370.

20. Compare Full virtualization and Paravirtualization

- Full virtualization refers to the ability to run a program, most likely an operating system, directly on top of a virtual machine and without any modification, as though it were run on the raw hardware.
- Paravirtualization is a not-transparent virtualization solution that allows implementing thin virtual machine managers.

- Paravirtualization techniques expose a software interface to the virtual machine that is slightly modified from the host and, as a consequence, guests need to be modified.

21. How to virtualization implemented in partial virtualization?

- Partial virtualization provides a partial emulation of the underlying hardware, thus not allowing the complete execution of the guest operating system in complete isolation.
- Partial virtualization allows many applications to run transparently, but not all the features of the operating system can be supported, as happens with full virtualization.

22. What is Operating system-level?

- Operating system-level virtualization offers the opportunity to create different and separated execution environments for applications that are managed concurrently.
- Differently from hardware virtualization, there is no virtual machine manager or hypervisor, and the virtualization is done within a single operating system, where the OS kernel allows for multiple isolated user space instances.

23. What is storage virtualization?

- Storage virtualization is a system administration practice that allows decoupling the physical organization of the hardware from its logical representation.
- Using this technique, users do not have to be worried about the specific location of their data, which can be identified using a logical path.

24. Define Desktop virtualization.

- Desktop virtualization abstracts the desktop environment available on a personal computer in order to provide access to it using a client/server approach.
- Desktop virtualization provides the same outcome of hardware virtualization but serves a different purpose.

25. List the merits of Virtualization at various implementation levels.

Level of Implementation	Higher Performance	Application Flexibility	Implementation Complexity	Application Isolation
Instruction Set Architecture	Very Low	Very High	Moderate	Moderate
Hardware-level virtualization	Very High	Moderate	Very High	High
OS-level virtualization	Very High	Low	Moderate	Low
Library support level	Moderate	Low	Low	Low
User application level	Low	Low	Very High	Very High

26. Differentiate between micro-kernel and monolithic hypervisor.

- A micro-kernel hypervisor includes only the basic and unchanging functions (such as physical memory management and processor scheduling).
- A monolithic hypervisor implements all the aforementioned functions, including those of the device drivers.

UNIT III CLOUD ARCHITECTURE, SERVICES AND STORAGE

Layered Cloud Architecture Design –NIST Cloud Computing Reference Architecture –Public, Private and Hybrid Clouds -laaS –PaaS –SaaS –Architectural Design Challenges –Cloud Storage –Storage-as-a-Service –Advantages of Cloud Storage –Cloud Storage Providers –S3.

3.1 Layered Cloud Architecture Design

- The architecture of a cloud is developed at three layers: infrastructure, platform and application as demonstrated in Figure 3.1.
- These three development layers are implemented with virtualization and standardization of hardware and software resources provisioned in the cloud.
- The services to public, private and hybrid clouds are conveyed to users through networking support over the Internet and intranets involved.
- It is clear that the infrastructure layer is deployed first to support IaaS services.
- This infrastructure layer serves as the foundation for building the platform layer of the cloud for supporting PaaS services.
- In turn, the platform layer is a foundation for implementing the application layer for SaaS applications.
- Different types of cloud services demand application of these resources separately.

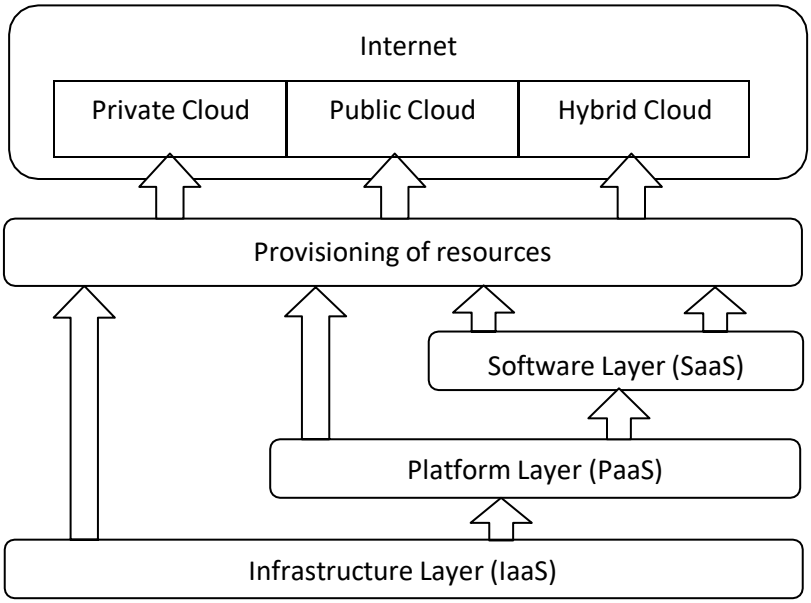


Figure 3.1 Layered architectural development

- The infrastructure layer is built with virtualized compute, storage and network resources.
- The abstraction of these hardware resources is meant to provide the flexibility demanded by users.
- Internally, virtualization realizes automated provisioning of resources and optimizes the infrastructure management process.
- The platform layer is for general purpose and repeated usage of the collection of software resources.
- This layer provides users with an environment to develop their applications, to test operation flows and to monitor execution results and performance.
- The platform should be able to assure users that they have scalability, dependability, and security protection.

- In a way, the virtualized cloud platform serves as a "system middleware" between the infrastructure and application layers of the cloud.

- The application layer is formed with a collection of all needed software modules for SaaS applications.

- Service applications in this layer include daily office management work such as information retrieval, document processing and calendar and authentication services.

- The application layer is also heavily used by enterprises in business marketing and sales, consumer relationship management (CRM), financial transactions and supply chain management.

- From the provider's perspective, the services at various layers demand different amounts of functionality support and resource management by providers.

- In general, SaaS demands the most work from the provider, PaaS is in the middle, and IaaS demands the least.

- For example, Amazon EC2 provides not only virtualized CPU resources to users but also management of these provisioned resources.

- Services at the application layer demand more work from providers.

- The best example of this is the Salesforce.com CRM service in which the provider supplies not only the hardware at the bottom layer and the software at the top layer but also the platform and software tools for user application development and monitoring.

- In Market Oriented Cloud Architecture, as consumers rely on cloud providers to meet more of their computing needs, they will require a specific level of QoS to be maintained by their providers, in order to meet their objectives and sustain their operations.

- Market-oriented resource management is necessary to regulate the supply and demand of cloud resources to achieve market equilibrium between supply and demand.

- This cloud is basically built with the following entities:

    - Users or brokers acting on user's behalf submit service requests from anywhere in the world to the data center and cloud to be processed.
    - The request examiner ensures that there is no overloading of resources whereby many service requests cannot be fulfilled successfully due to limited resources.
    - The Pricing mechanism decides how service requests are charged. For instance, requests can be charged based on submission time (peak/off-peak), pricing rates (fixed/changing), or availability of resources (supply/demand).
    - The VM Monitor mechanism keeps track of the availability of VMs and their resource entitlements.
    - The Accounting mechanism maintains the actual usage of resources by requests so that the final cost can be computed and charged to users.
    - In addition, the maintained historical usage information can be utilized by the Service Request Examiner and Admission Control mechanism to improve resource allocation decisions.
    - The Dispatcher mechanism starts the execution of accepted service requests on allocated VMs.
    - The Service Request Monitor mechanism keeps track of the execution progress of service requests.

3.2 NIST Cloud Computing Reference Architecture

- NIST stands for National Institute of Standards and Technology

- The goal is to achieve effective and secure cloud computing to reduce cost and improve services

- NIST composed for six major workgroups specific to cloud computing

- Cloud computing target business use cases work group
- Cloud computing Reference architecture and Taxonomy work group
- Cloud computing standards roadmap work group
- Cloud computing SAJACC (Standards Acceleration to Jumpstart Adoption of Cloud Computing) work group
- Cloud Computing security work group

- Objectives of NIST Cloud Computing reference architecture
  - Illustrate and understand the various level of services
  - To provide technical reference
  - Categorize and compare services of cloud computing
  - Analysis of security, interoperatability and portability

- In general, NIST generates report for future reference which includes survey, analysis of existing cloud computing reference model, vendors and federal agencies.

- The conceptual reference architecture shown in figure 3.2 involves five actors. Each actor as entity participates in cloud computing

- Cloud consumer: A person or an organization that maintains a business relationship with and uses a services from cloud providers

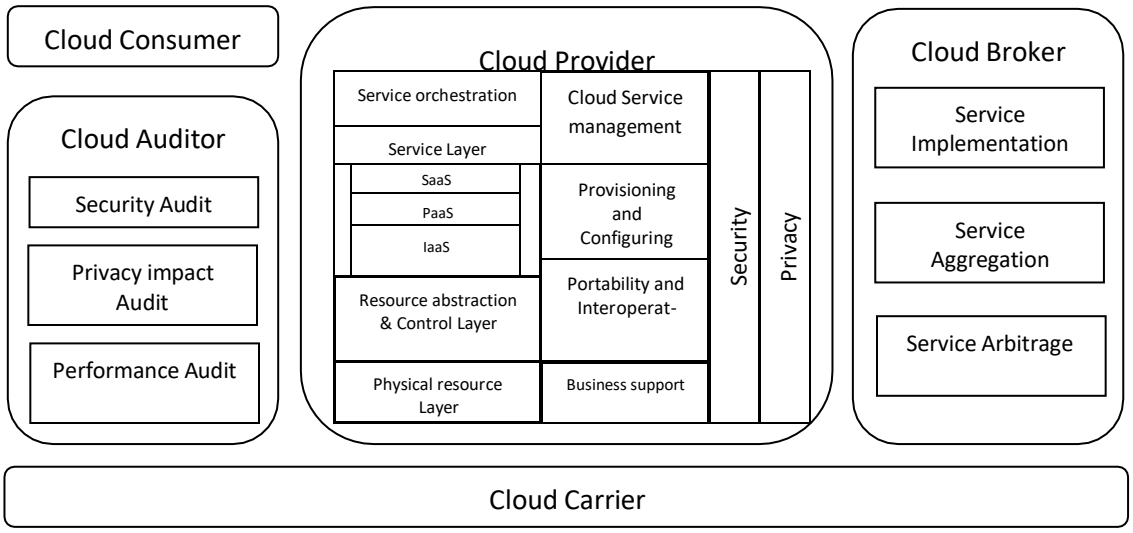


Figure 3.2 Conceptual reference model

- Cloud provider: A person, organization or entity responsible for making a service available to interested parties

- Cloud auditor: A party that conduct independent assessment of cloud services, information system operation, performance and security of cloud implementation

- Cloud broker: An entity that manages the performance and delivery of cloud services and negotiates relationship between cloud provider and consumer.

- Cloud carrier: An intermediary that provides connectivity and transport of cloud services from cloud providers to consumers.

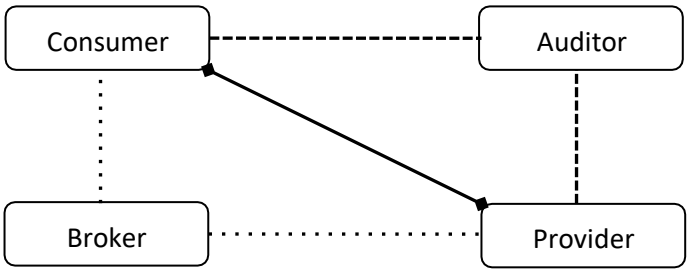


Figure 3.3 Interaction between actors

- Figure 3.3 illustrates the common interaction exist in between cloud consumer and provider where as the broker used to provide service to consumer and auditor collects the audit information.

- The interaction between the actors may lead to different use case scenario.

- Figure 3.4 shows one kind of scenario in which the Cloud consumer may request service from a cloud broker instead of contacting service provider directly. In this case, a cloud broker can create a new service by combining multiple services.

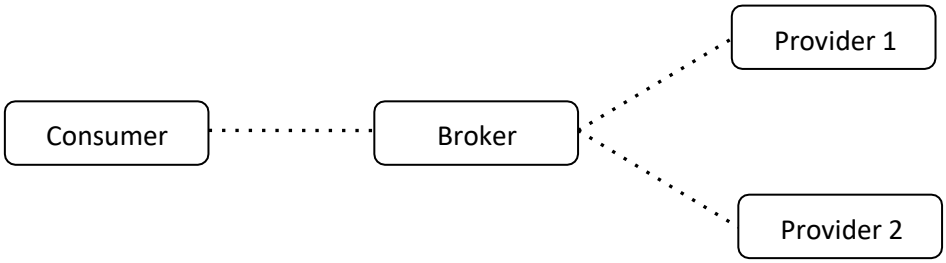


Figure 3.4 Service from Cloud Broker

- Figure 3.5 illustrates the usage of different kind of Service Level Agreement (SLA) between consumer, provider and carrier.

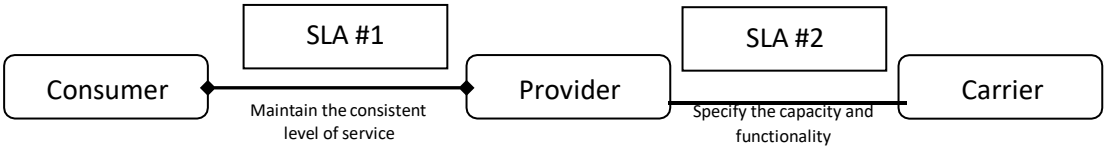


Figure 3.5 Multiple SLA between actors

- Figure 3.6 shows the scenario where the Cloud auditor conducts independent assessment of operation and security of the cloud service implementation.

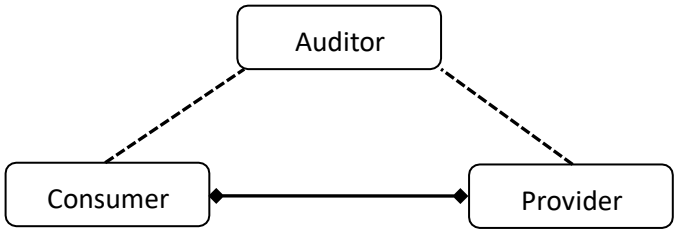


Figure 3.6 Independent assessments by cloud auditor

- Cloud consumer is a principal stake holder for the cloud computing service and requires service level agreements to specify the performance requirements fulfilled by a cloud provider.

- The service level agreement covers Quality of Service and Security aspects.

- Consumers have limited rights to access the software applications.

- There are three kinds of cloud consumers: SaaS consumers, PaaS Consumers and IaaS consumers.

- SaaS consumers are members directly access the software application. For example, document management, content management, social networks, financial billing and so on.

- PaaS consumers are used to deploy, test, develop and manage applications hosted in cloud environment. Database application deployment, development and testing is an example for these kind of consumer.

- IaaS Consumer can access the virtual computer, storage and network infrastructure. For example, usage of Amazon EC2 instance to deploy the web application.

- On the other hand, Cloud Providers have complete rights to access software applications.

- In Software as a Service model, cloud provider is allowed to configure, maintain and update the operations of software application.

- Management process is done by Integrated Development environment and Software Development Kit in Platform as a Service model.

- Infrastructure as a Service model covers Operating System and Networks.

- Normally, the service layer defines the interfaces for cloud consumers to access the computing services.

- Resource abstraction and control layer contains the system components that cloud provider use to provide and mange access to the physical computing resources through software abstraction.

- Resource abstraction covers virtual machine management and virtual storage management.

- Control layer focus on resource allocation, access control and usage monitoring.

- Physical resource layer includes physical computing resources such as CPU, Memory, Router, Switch, Firewalls and Hard Disk Drive.

- Service orchestration describes the automated arrangement, coordination and management of complex computing system.

- In cloud service management, business support entails the set of business related services dealing with consumer and supporting services which includes content management, contract management, inventory management, accounting service, reporting service and rating service.

- Provisioning of equipments, wiring and transmission is mandatory to setup a new service that provides a specific application to cloud consumer. Those details are described in Provisioning and Configuring management.

- Portability enforces the ability to work in more than one computing environment without major task. Similarly, Interoperatability means the ability of the system work with other system.

- Security factor is applicable to enterprise and Government. It may include privacy.

- Privacy is one applies to a cloud consumer's rights to safe guard his information from other consumers are parties.

- The main aim of Security and Privacy in cloud service management is to protect the system from vulnerable customers.

- Cloud auditor performs independent assessments among the services and cloud broker act as intermediate module.

- Service intermediation enhances a given service by improving some specific capability and providing value added services to cloud consumers,

- Service aggregation provides data integration. Cloud broker combines and integrate multiple service into one or more new services.

- Due to Service arbitrage, cloud broker has a flexibility to choose services from multiple providers.

- Cloud carrier is an intermediary that provides connectivity and transport of cloud service between cloud consumer and cloud provider.

- It provides access to cloud consumer with the help of network, telecommunication and other access devices where as distribution is done with transport agent,

- Transport agent is the business organization that provides physical transport of storage media.

3.3 Cloud Deployment Model

- As identified in the NIST cloud computing definition, a cloud infrastructure may be operated in one of the following deployment models: public cloud, private cloud, community cloud, or hybrid cloud.

- The differences are based on how exclusive the computing resources are made to a Cloud Consumer.

3.3.1 Public Cloud

- A public cloud is one in which the cloud infrastructure and computing resources are made available to the general public over a public network.

- A public cloud is owned by an organization selling cloud services, and serves a diverse pool of clients.

- Figure 4.7 presents a simple view of a public cloud and its customers.

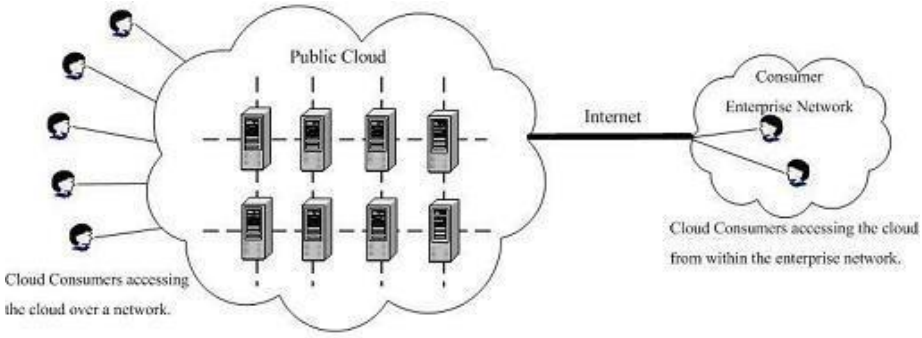


Figure 3.7 Public Cloud

3.3.1.1 Benefits of choosing a Public Cloud

- One of the main benefits that come with using public cloud services is near unlimited scalability.

- The resources are pretty much offered based on demand. So any changes in activity level can be handled very easily.

- This in turn brings with it cost effectiveness.

- Public cloud allows pooling of a large number of resources, users are benefiting from the savings of large scale operations.

- There are many services like Google Drive which are offered for free.

- Finally, the vast network of servers involved in public cloud services means that it can benefit from greater reliability.

- Even if one data center was to fail entirely, the network simply redistributes the load among the remaining enters making it highly unlikely that the public cloud would ever fail.

- In summary, the benefits of the public cloud are:

    - Easy scalability
    - Cost effectiveness
    - Increased reliability

3.3.1.2 Disadvantages of choosing a Public Cloud

- There are of course downsides to using public cloud services.

- At the top of the list is the fact that the security of data held within a public cloud is a cause for concern.

- It is often seen as an advantage that the public cloud has no geographical restrictions making access easy from everywhere, but on the flip side this could mean that the server is in a different country which is governed by an entirely different set of security and/or privacy regulations.

- This could mean that your data is not all that secure making it unwise to use public cloud services for sensitive data.

3.3.2 Private Cloud

- A private cloud gives a single Cloud Consumer's organization the exclusive access to and usage of the infrastructure and computational resources.

- It may be managed either by the Cloud Consumer organization or by a third party, and may be hosted on the organization's premises (i.e. on-site private clouds) or outsourced to a hosting company (i.e. outsourced private clouds).

- Figure 3.8 presents an on-site private cloud and an outsourced private cloud, respectively.

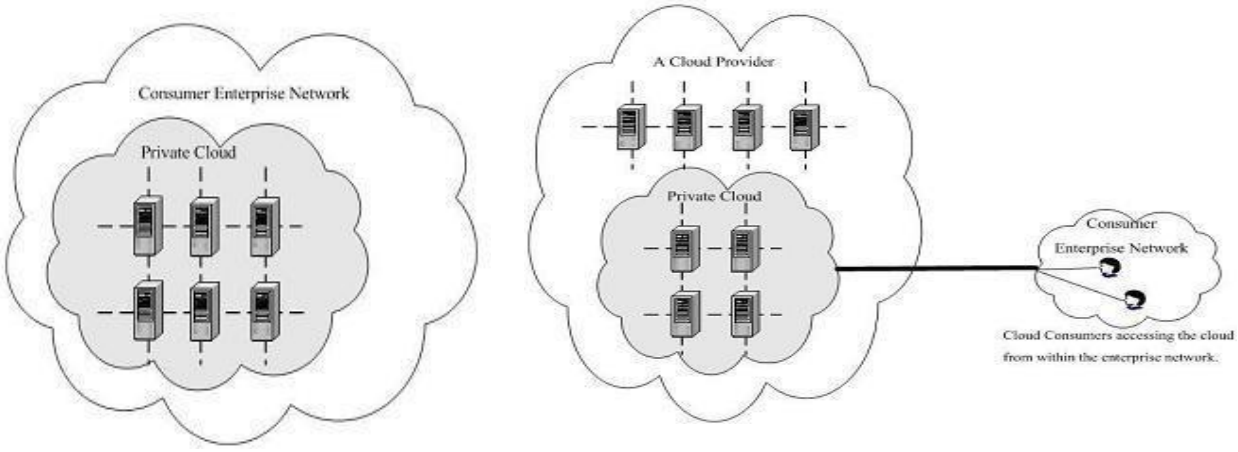


Figure 3.8 (a) On-site Private Cloud (b) Out-sourced Private Cloud

3.3.2.1 Benefits of choosing a Private Cloud

- The main benefit of choosing a private cloud is the greater level of security offered making it ideal for business users who need to store and/or process sensitive data.

- A good example is a company dealing with financial information such as bank or lender who is required by law to use secure internal storage to store consumer information.

- With a private cloud this can be achieved while still allowing the organization to benefit from cloud computing.

- Private cloud services also offer some other benefits for business users including more control over the server allowing it to be tailored to your own preferences and in house styles.

- While this can remove some of the scalability options, private cloud providers often offer what is known as cloud bursting which is when non sensitive data is switched to a public cloud to free up private cloud space in the event of a significant spike in demand until such times as the private cloud can be expanded.

- In summary, the main benefits of the private cloud are:

    - Improved security
    - Greater control over the server
    - Flexibility in the form of Cloud Bursting

3.3.2.2 Disadvantages of choosing a Private Cloud

- The downsides of private cloud services include a higher initial outlay, although in the long term many business owners find that this balances out and actual becomes more cost effective than public cloud use.

- It is also more difficult to access the data held in a private cloud from remote locations due to the increased security measures.

3.3.3 Community Cloud

- A community cloud serves a group of Cloud Consumers which have shared concerns such as mission objectives, security, privacy and compliance policy, rather than serving a single organization as does a private cloud.

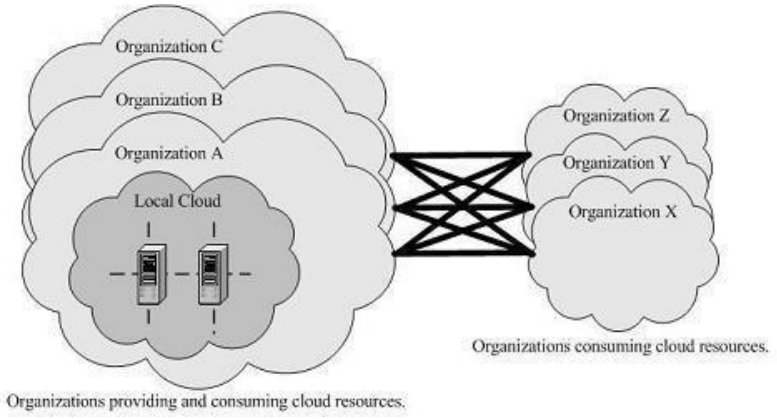
- Similar to private clouds, a community cloud may be managed by the organizations or by a third party and may be implemented on customer premise (i.e. on-site community cloud) or outsourced to a hosting company (i.e. outsourced community cloud).

- Figure 3.9 (a) depicts an on-site community cloud comprised of a number of participant organizations.

- A cloud consumer can access the local cloud resources, and also the resources of other participating organizations through the connections between the associated organizations.

- Figure 3.9 (b) shows an outsourced community cloud, where the server side is outsourced to a hosting company.

- In this case, an outsourced community cloud builds its infrastructure off premise, and serves a set of organizations that request and consume cloud services.



(a) On-site Community Cloud

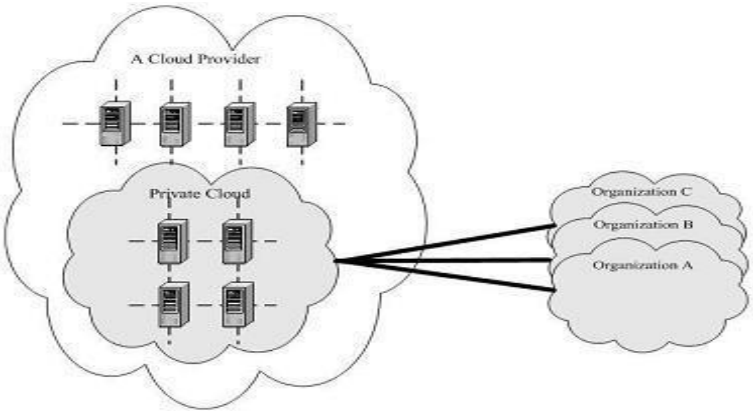


Figure 3.9 (b) Outsourced Community Cloud

3.3.3.1 Benefits of Choosing a Community Cloud

- Ability to easily share and collaborate
- Lower cost

3.3.3.2 Disadvantages of Choosing a Community Cloud

- Not the right choice for every organization
- Slow adoption to date

3.3.4 Hybrid Cloud

- A hybrid cloud is a composition of two or more clouds (on-site private, on-site community, off-site private, off-site community or public) that remain as distinct entities but are bound together by standardized or proprietary technology that enables data and application portability.

- Figure 3.10 illustrates a simple view of a hybrid cloud that could be built with a set of clouds in the five deployment model variants.

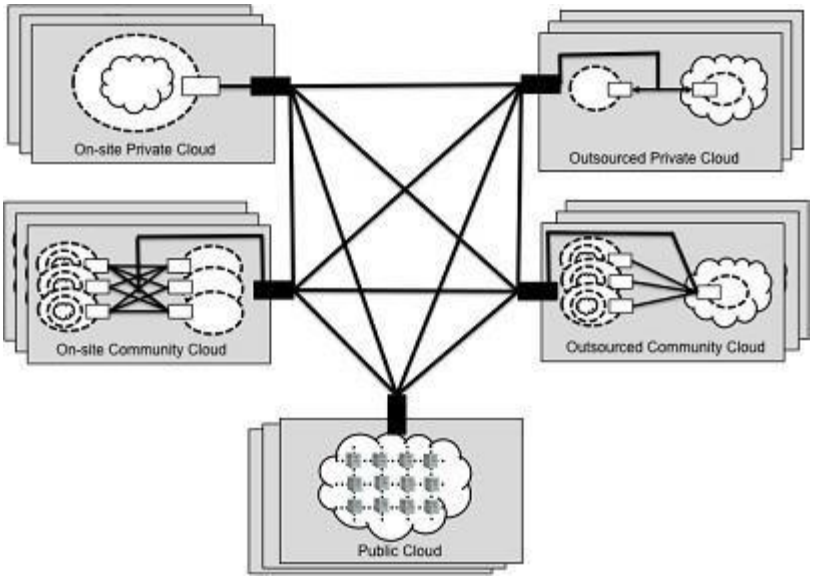


Figure 3.10 Hybrid Cloud

3.4 Cloud Service Model

- The development of cloud computing introduces the concept of everything as a Service (XaaS). This is one of the most important elements of cloud computing

- Cloud services from different providers can be combined to provide a completely integrated solution covering all the computing stack of a system.

- IaaS providers can offer the bare metal in terms of virtual machines where PaaS solutions are deployed.

- When there is no need for a PaaS layer, it is possible to directly customize the virtual infrastructure with the software stack needed to run applications.

- This is the case of virtual Web farms: a distributed system composed of Web servers, database servers and load balancers on top of which prepackaged software is installed to run Web applications.

- Other solutions provide prepackaged system images that already contain the software stack required for the most common uses: Web servers, database servers or LAMP stacks.

- Besides the basic virtual machine management capabilities, additional services can be provided, generally including the following:

    - SLA resource based allocation
    - Workload management
    - Support for infrastructure design through advanced Web interfaces
    - Integrate third party IaaS solutions

- Figure 3.11 provides an overall view of the components forming an Infrastructure as a Service solution.

- It is possible to distinguish three principal layers:

    - Physical infrastructure
    - Software management infrastructure
    - User interface

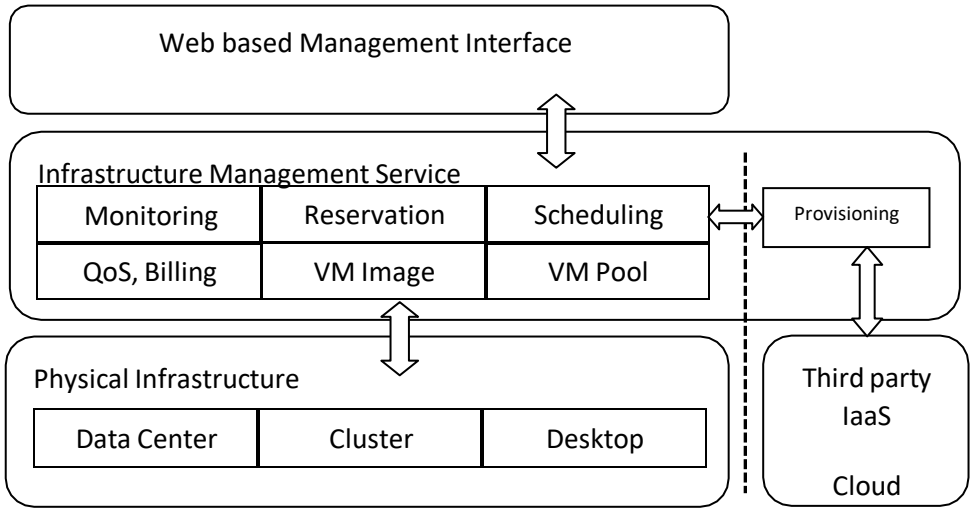


Figure 3.11 IaaS reference implementation

- At the top layer the user interface provides access to the services exposed by the software management infrastructure.

- Such an interface is generally based on Web 2.0 technologies: Web services, RESTful APIs and mash ups.

- Web services and RESTful APIs allow programs to interact with the service without human intervention, thus providing complete integration within a software system.

- The core features of an IaaS solution are implemented in the infrastructure management software layer.

- In particular, management of the virtual machines is the most important function performed by this layer.

- A central role is played by the scheduler, which is in charge of allocating the execution of virtual machine instances.

- The scheduler interacts with the other components such as

    - Pricing and billing component
    - Monitoring component
    - Reservation component
    - QoS/SLA management component
    - VM repository component
    - VM pool manager component
    - Provisioning component

- The bottom layer is composed of the physical infrastructure, on top of which the management layer operates.

- From an architectural point of view, the physical layer also includes the virtual resources that are rented from external IaaS providers.

- In the case of complete IaaS solutions, all three levels are offered as service.

- This is generally the case with public clouds vendors such as Amazon, GoGrid, Joyent, Rightscale, Terremark, Rackspace, ElasticHosts, and Flexiscale, which own large datacenters and give access to their computing infrastructures using an IaaS approach.

3.4.1 IaaS

- Infrastructure or Hardware as a Service (IaaS/HaaS) solutions are the most popular and developed market segment of cloud computing.

- They deliver customizable infrastructure on demand.

- The available options within the IaaS offering umbrella range from single servers to entire infrastructures, including network devices, load balancers, database servers and Web servers.

- The main technology used to deliver and implement these solutions is hardware virtualization: one or more virtual machines opportunely configured and interconnected define the distributed system on top of which applications are installed and deployed.

- Virtual machines also constitute the atomic components that are deployed and priced according to the specific features of the virtual hardware: memory, number of processors and disk storage.

- IaaS/HaaS solutions bring all the benefits of hardware virtualization: workload partitioning, application isolation, sandboxing and hardware tuning.

- From the perspective of the service provider, IaaS/HaaS allows better exploiting the IT infrastructure and provides a more secure environment where executing third party applications.

- From the perspective of the customer, it reduces the administration and maintenance cost as well as the capital costs allocated to purchase hardware.

- At the same time, users can take advantage of the full customization offered by virtualization to deploy their infrastructure in the cloud.

3.4.2 PaaS

- Platform as a Service (PaaS) solutions provide a development and deployment platform for running applications in the cloud.

- They constitute the middleware on top of which applications are built.

- A general overview of the features characterizing the PaaS approach is given in Figure 3.12.

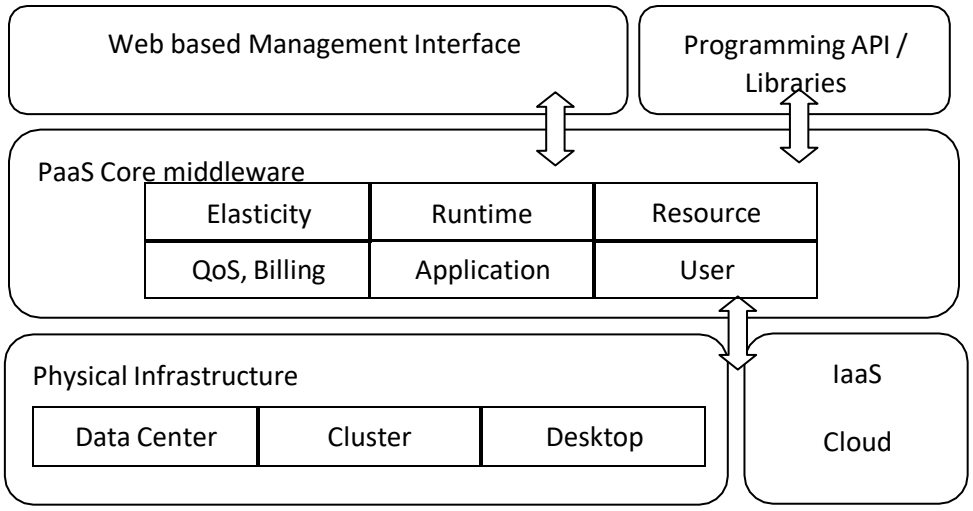


Figure 3.12 PaaS reference implementation

- Application management is the core functionality of the middleware.

- PaaS implementations provide applications with a runtime environment and do not expose any service for managing the underlying infrastructure.

- They automate the process of deploying applications to the infrastructure, configuring application components, provisioning and configuring supporting technologies such as load balancers and databases and managing system change based on policies set by the user.

- The core middleware is in charge of managing the resources and scaling applications on demand or automatically, according to the commitments made with users.

- From a user point of view, the core middleware exposes interfaces that allow programming and deploying applications on the cloud.

- Some implementations provide a completely Web based interface hosted in the cloud and offering a variety of services.

- It is possible to find integrated developed environments based on 4GL and visual programming concepts or rapid prototyping environments where applications are built by assembling mash ups and user defined components and successively customized.

- Other implementations of the PaaS model provide a complete object model for representing an application and provide a programming language-based approach.

- Developers generally have the full power of programming languages such as Java, .NET, Python and Ruby with some restrictions to provide better scalability and security.

- PaaS solutions can offer middleware for developing applications together with the infrastructure or simply provide users with the software that is installed on the user premises.

- In the first case, the PaaS provider also owns large datacenters where applications are executed

- In the second case, referred to in this book as Pure PaaS, the middleware constitutes the core value of the offering.

- PaaS implementation classified into three wide categories:

    - PaaS-I
    - PaaS-II
    - PaaS-III

- The first category identifies PaaS implementations that completely follow the cloud computing style for application development and deployment.

    - They offer an integrated development environment hosted within the Web browser where applications are designed, developed, composed, and deployed.
    - This is the case of Force.com and Longjump. Both deliver as platforms the combination of middleware and infrastructure.

- In the second class focused on providing a scalable infrastructure for Web application, mostly websites.

    - In this case, developers generally use the provider's APIs, which are built on top of industrial runtimes, to develop applications.
    - Google AppEngine is the most popular product in this category.
    - It provides a scalable runtime based on the Java and Python programming languages, which have been modified for providing a secure runtime environment and enriched with additional APIs and components to support scalability.

- AppScale, an open source implementation of Google AppEngine, provides interfacecompatible middleware that has to be installed on a physical infrastructure.

- The third category consists of all those solutions that provide a cloud programming platform for any kind of application, not only Web applications.

  - Among these, the most popular is Microsoft Windows Azure, which provides a comprehensive framework for building service oriented cloud applications on top of the .NET technology, hosted on Microsoft's datacenters.
  - Other solutions in the same category, such as Manjrasoft Aneka, Apprenda SaaSGrid, Appistry Cloud IQ Platform, DataSynapse, and GigaSpaces DataGrid, provide only middleware with different services.

- Some essential characteristics that identify a PaaS solution:

  - Runtime framework: This framework represents the software stack of the PaaS model and the most intuitive aspect that comes to people's minds when they refer to PaaS solutions.
  - Abstraction: PaaS solutions are distinguished by the higher level of abstraction that they provide.
  - Automation: PaaS environments automate the process of deploying applications to the infrastructure, scaling them by provisioning additional resources when needed.
  - Cloud services: PaaS offerings provide developers and architects with services and APIs, helping them to simplify the creation and delivery of elastic and highly available cloud application.

3.4.3 SaaS

- Software as a Service (SaaS) is a software delivery model that provides access to applications through the Internet as a Web based service.

- It provides a means to free users from complex hardware and software management by offloading such tasks to third parties, which build applications accessible to multiple users through a Web browser.

- On the provider side, the specific details and features of each customer's application are maintained in the infrastructure and made available on demand.

- The SaaS model is appealing for applications serving a wide range of users and that can be adapted to specific needs with little further customization.

- This requirement characterizes SaaS as a one-to-many software delivery model, whereby an application is shared across multiple users.

- This is the case of CRM and ERP applications that constitute common needs for almost all enterprises, from small to medium-sized and large business.

- Every enterprise will have the same requirements for the basic features concerning CRM and ERP and different needs can be satisfied with further customization.

- SaaS applications are naturally multitenant.

- Multitenancy, which is a feature of SaaS compared to traditional packaged software, allows providers to centralize and sustain the effort of managing large hardware infrastructures, maintaining as well as upgrading applications transparently to the users and optimizing resources by sharing the costs among the large user base.

- On the customer side, such costs constitute a minimal fraction of the usage fee paid for the software.

- The analysis carried out by Software Information and Industry Association (SIIA) was mainly oriented to cover application service providers (ASPs) and all their variations,

which capture the concept of software applications consumed as a service in a broader sense.

- ASPs already had some of the core characteristics of SaaS:
    - The product sold to customer is application access
    - The application is centrally managed
    - The service delivered is one-to-many
    - The service delivered is an integrated solution delivered on the contract, which means provided as promised.
- ASPs provided access to packaged software solutions that addressed the needs of a variety of customers.
- Initially this approach was affordable for service providers, but it later became inconvenient when the cost of customizations and specializations increased.
- The SaaS approach introduces a more flexible way of delivering application services that are fully customizable by the user by integrating new services, injecting their own components and designing the application and information workflows.
- Initially the SaaS model was of interest only for lead users and early adopters.
- The benefits delivered at that stage were the following:
    - Software cost reduction and total cost of ownership (TCO) were paramount
    - Service level improvements
    - Rapid implementation
    - Standalone and configurable applications
    - Rudimentary application and data integration
    - Subscription and pay as you go (PAYG) pricing

- With the advent of cloud computing there has been an increasing acceptance of SaaS as a viable software delivery model.

- This lead to transition into SaaS 2.0, which does not introduce a new technology but transforms the way in which SaaS is used.

- In particular, SaaS 2.0 is focused on providing a more robust infrastructure and application platforms driven by SLAs.

- SaaS 2.0 will focus on the rapid achievement of business objectives.

- Software as a Service based applications can serve different needs. CRM, ERP, and social networking applications are definitely the most popular ones.

- SalesForce.com is probably the most successful and popular example of a CRM service.

- It provides a wide range of services for applications: customer relationship and human resource management, enterprise resource planning, and many other features.

- SalesForce.com builds on top of the Force.com platform, which provides a fully featured environment for building applications.

- In particular, through AppExchange customers can publish, search and integrate new services and features into their existing applications.

- This makes SalesForce.com applications completely extensible and customizable.

- Similar solutions are offered by NetSuite and RightNow.

- NetSuite is an integrated software business suite featuring financials, CRM, inventory, and ecommerce functionalities integrated all together.

- RightNow is customer experience centered SaaS application that integrates together different features, from chat to Web communities, to support the common activity of an enterprise

- Another important class of popular SaaS applications comprises social networking applications such as Facebook and professional networking sites such as LinkedIn.

- Other than providing the basic features of networking, they allow incorporating and extending their capabilities by integrating third-party applications.

- Office automation applications are also an important representative for SaaS applications:

    - Google Documents and Zoho Office are examples of Web based applications that aim to address all user needs for documents, spreadsheets and presentation management.
    - These applications offer a Web based interface for creating, managing, and modifying documents that can be easily shared among users and made accessible from anywhere.

3.5 Architectural Design Challenges

3.5.1 Challenge 1: Service Availability and Data Lock-in Problem

- The management of a cloud service by a single company is often the source of single points of failure.

- To achieve HA, one can consider using multiple cloud providers.

- Even if a company has multiple data centers located in different geographic regions, it may have common software infrastructure and accounting systems.

- Therefore, using multiple cloud providers may provide more protection from failures.

- Another availability obstacle is distributed denial of service (DDoS) attacks.

- Criminals threaten to cut off the incomes of SaaS providers by making their services unavailable.

- Some utility computing services offer SaaS providers the opportunity to defend against DDoS attacks by using quick scale ups.

- Software stacks have improved interoperability among different cloud platforms, but the APIs itself are still proprietary. Thus, customers cannot easily extract their data and programs from one site to run on another.

- The obvious solution is to standardize the APIs so that a SaaS developer can deploy services and data across multiple cloud providers.

- This will rescue the loss of all data due to the failure of a single company.

- In addition to mitigating data lock-in concerns, standardization of APIs enables a new usage model in which the same software infrastructure can be used in both public and private clouds.

- Such an option could enable surge computing, in which the public cloud is used to capture the extra tasks that cannot be easily run in the data center of a private cloud.

3.5.2 Challenge 2: Data Privacy and Security Concerns

- Current cloud offerings are essentially public (rather than private) networks, exposing the system to more attacks.

- Many obstacles can be overcome immediately with well understood technologies such as encrypted storage, virtual LANs, and network middle boxes (e.g., firewalls, packet filters).

- For example, the end user could encrypt data before placing it in a cloud. Many nations have laws requiring SaaS providers to keep customer data and copyrighted material within national boundaries.

- Traditional network attacks include buffer overflows, DoS attacks, spyware, malware, rootkits, Trojan horses, and worms.

- In a cloud environment, newer attacks may result from hypervisor malware, guest hopping and hijacking or VM rootkits.

- Another type of attack is the man-in-the-middle attack for VM migrations.

- In general, passive attacks steal sensitive data or passwords.

- On the other hand, Active attacks may manipulate kernel data structures which will cause major damage to cloud servers.

3.5.3 Challenge 3: Unpredictable Performance and Bottlenecks

- Multiple VMs can share CPUs and main memory in cloud computing, but I/O sharing is problematic.

- For example, to run 75 EC2 instances with the STREAM benchmark requires a mean bandwidth of 1,355 MB/second.

- However, for each of the 75 EC2 instances to write 1 GB files to the local disk requires a mean disk write bandwidth of only 55 MB/second.

- This demonstrates the problem of I/O interference between VMs.

- One solution is to improve I/O architectures and operating systems to efficiently virtualize interrupts and I/O channels.

- Internet applications continue to become more data intensive.

- If we assume applications to be pulled apart across the boundaries of clouds, this may complicate data placement and transport.

- Cloud users and providers have to think about the implications of placement and traffic at every level of the system, if they want to minimize costs.

- This kind of reasoning can be seen in Amazon's development of its new CloudFront service.

- Therefore, data transfer bottlenecks must be removed, bottleneck links must be widened and weak servers should be removed.

3.5.4 Challenge 4: Distributed Storage and Widespread Software Bugs

- The database is always growing in cloud applications.

- The opportunity is to create a storage system that will not only meet this growth but also combine it with the cloud advantage of scaling arbitrarily up and down on demand.

- This demands the design of efficient distributed SANs.

- Data centers must meet programmer's expectations in terms of scalability, data durability and HA.

- Data consistence checking in SAN connected data centers is a major challenge in cloud computing.

- Large scale distributed bugs cannot be reproduced, so the debugging must occur at a scale in the production data centers.

- No data center will provide such a convenience. One solution may be a reliance on using VMs in cloud computing.

- The level of virtualization may make it possible to capture valuable information in ways that are impossible without using VMs.

- Debugging over simulators is another approach to attacking the problem, if the simulator is well designed.

3.5.5 Challenge 5: Cloud Scalability, Interoperability, and Standardization

- The pay as you go model applies to storage and network bandwidth; both are counted in terms of the number of bytes used.

- Computation is different depending on virtualization level.

- GAE automatically scales in response to load increases or decreases and the users are charged by the cycles used.

- AWS charges by the hour for the number of VM instances used, even if the machine is idle.

- The opportunity here is to scale quickly up and down in response to load variation, in order to save money, but without violating SLAs.

- Open Virtualization Format (OVF) describes an open, secure, portable, efficient and extensible format for the packaging and distribution of VMs.

- It also defines a format for distributing software to be deployed in VMs.

- This VM format does not rely on the use of a specific host platform, virtualization platform or guest operating system.

- The approach is to address virtual platform is agnostic packaging with certification and integrity of packaged software.

- The package supports virtual appliances to span more than one VM.

- OVF also defines a transport mechanism for VM templates and the format can apply to different virtualization platforms with different levels of virtualization.

- In terms of cloud standardization, the ability for virtual appliances to run on any virtual platform.

- The user is also need to enable VMs to run on heterogeneous hardware platform hypervisors.

- This requires hypervisor-agnostic VMs.

- And also the user need to realize cross platform live migration between x86 Intel and AMD technologies and support legacy hardware for load balancing.

- All these issues are wide open for further research.

3.5.6 Challenge 6: Software Licensing and Reputation Sharing

- Many cloud computing providers originally relied on open source software because the licensing model for commercial software is not ideal for utility computing.

- The primary opportunity is either for open source to remain popular or simply for commercial software companies to change their licensing structure to better fit cloud computing.

- One can consider using both pay for use and bulk use licensing schemes to widen the business coverage.

3.6 Cloud Storage

- Cloud storage means storing the data with a cloud service provider rather than on a local system.

- The end user can access the data stored on the cloud using an Internet link.

- Cloud storage has a number of advantages over traditional data storage.

- If the users stored some data on a cloud, they can get at it from any location that has Internet access.

- Workers do not need to use the same computer to access data nor do they have to carry around physical storage devices.

- Also, if any organization has branch offices, they can all access the data from the cloud provider.

- There are hundreds of different cloud storage systems, and some are very specific in what they do.

- Some are niche-oriented and store just email or digital pictures, while others store any type of data. Some providers are small, while others are huge and fill an entire warehouse.

- At the most rudimentary level, a cloud storage system just needs one data server connected to the Internet.

- A subscriber copies files to the server over the Internet, which then records the data. When a client wants to retrieve the data, the client accesses the data server with a web based interface and the server then either sends the files back to the client or allows the client to access and manipulate the data itself.

- More typically, however, cloud storage systems utilize dozens or hundreds of data servers.

- Because servers require maintenance or repair, it is necessary to store the saved data on multiple machines, providing redundancy.

- Without that redundancy, cloud storage systems could not assure clients that they could access their information at any given time.

3.6.1 Storage-as-a-Service

- The term Storage as a Service (another Software as a Service, or SaaS, acronym) means that a third-party provider rents space on their storage to end users who lack the budget or capital budget to pay for it on their own.

- Figure 3.13 illustrates the storage as a service where the data stored in cloud storage.

- It is also ideal when technical personnel are not available or have inadequate knowledge to implement and maintain that storage infrastructure.

- Storage service providers are nothing new, but given the complexity of current backup, replication, and disaster recovery needs, the service has become popular, especially among small and medium sized businesses.

- The biggest advantage to SaaS is cost savings.

- Storage is rented from the provider using a cost-per-gigabyte-stored or cost-per-data-transferred model.

- The end user does not have to pay for infrastructure. They simply pay for how much they transfer and save on the provider's servers.

Figure 3.13 Storage as a Service

- A customer uses client software to specify the backup set and then transfers data across a WAN.

- Examples:

    - Google Docs allows users to upload documents, spreadsheets, and presentations to Google's data servers. Those files can then be edited using a Google application.
    - Web email providers like Gmail, Hotmail, and Yahoo! Mail store email messages on their own servers. Users can access their email from computers and other devices connected to the Internet.
    - Flickr and Picasa host millions of digital photographs. Users can create their own online photo albums.

- YouTube hosts millions of user uploaded video files.
- Hostmonster and GoDaddy store files and data for many client web sites.
- Facebook and MySpace are social networking sites and allow members to post pictures and other content. That content is stored on the company's servers.
- MediaMax and Strongspace offer storage space for any kind of digital data.

- To secure data, most systems use a combination of the listed techniques:

- Encryption: A complex algorithm is used to encode information. To decode the encrypted files, a user needs the encryption key.

- Authentication processes: This requires a user to create a name and password.

- Authorization practices: The client lists the people who are authorized to access information stored on the cloud system. Many corporations have multiple levels of authorization.

- The other concern is reliability.

- If a cloud storage system is unreliable, it becomes a liability. No one wants to save data on an unstable system, nor would they trust a company that is financially unstable.

- Most cloud storage providers try to address the reliability concern through redundancy, but the possibility still exists that the system could crash and leave clients with no way to access their saved data.

3.6.2 Advantages of Cloud Storage

- Cloud storage is becoming an increasingly attractive solution for organizations.

- Cloud storage providers balance server loads and move data among various datacenters, ensuring that information is stored close and thereby available quickly while using the data.

- Storing data on the cloud is advantageous, because it allows the user to protect the data in case there's a disaster.

- Having the data stored off-site can be the difference between closing the door for good or being down for a few days or weeks.

- Which storage vendor to go with can be a complex issue, and how the end user technology interacts with the cloud can be complex.

- For instance, some products are agent based and the application automatically transfers information to the cloud via FTP.

- But others employ a web front end and the user has to select local files on their computer to transmit.

- Amazon S3 is the best known storage solution, but other vendors might be better for large enterprises.

- For instance, those who offer service level agreements and direct access to customer support are critical for a business moving storage to a service provider

3.6.3 Cloud Storage Providers

- There are hundreds of cloud store providers every day.

- This is simply a listing of what some of the big players in the game have to offer and anyone can use it as a starting guide to determine if their services match user's needs.

- Amazon and Nirvanix are the current industry top dogs, but many others are in the field, including some well known names.

- Google offers cloud storage solution called GDrive.

- EMC is readying a storage solution and IBM already has a number of cloud storage options called Blue Cloud.

3.6.4 S3

- The well known cloud storage service is Amazon's Simple Storage Service (S3), which is launched in 2006.

- Amazon S3 is designed to make web scale computing easier for developers.

- Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the Web.

- It gives any developer access to the same highly scalable data storage infrastructure that Amazon uses to run its own global network of web sites.

- The service aims to maximize benefits of scale and to pass those benefits on to developers.

- Amazon S3 is intentionally built with a minimal feature set that includes the following functionality:

    - Write, read, and delete objects containing from 1 byte to 5 gigabytes of data each. The number of objects that can be stored is unlimited.
    - Each object is stored and retrieved via a unique developer assigned key.
    - Objects can be made private or public and rights can be assigned to specific users.
    - Uses standards based REST and SOAP interfaces designed to work with any Internet development toolkit.

- Design Requirements Amazon built S3 to fulfill the following design requirements:

    - Scalable: Amazon S3 can scale in terms of storage, request rate and users to support an unlimited number of web-scale applications.
    - Reliable: Store data durably with 99.99 percent availability. Amazon says it does not allow any downtime.
    - Fast: Amazon S3 was designed to be fast enough to support high-performance applications. Server-side latency must be insignificant relative to Internet latency.
    - Inexpensive: Amazon S3 is built from inexpensive commodity hardware components.
    - Simple: Building highly scalable, reliable, fast and inexpensive storage is difficult.

- Design Principles Amazon used the following principles of distributed system design to meet Amazon S3 requirements:

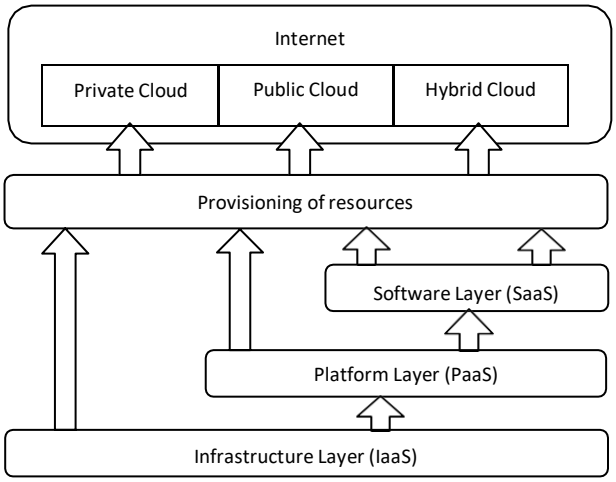
    - Decentralization: It uses fully decentralized techniques to remove scaling bottlenecks and single points of failure.
    - Autonomy: The system is designed such that individual components can make decisions based on local information.
    - Local responsibility: Each individual component is responsible for achieving its consistency. This is never the burden of its peers.
    - Controlled concurrency: Operations are designed such that no or limited concurrency control is required.
    - Failure toleration: The system considers the failure of components to be a normal mode of operation and continues operation with no or minimal interruption.
    - Controlled parallelism: Abstractions used in the system are of such granularity that parallelism can be used to improve performance and robustness of recovery or the introduction of new nodes.
    - Symmetry: Nodes in the system are identical in terms of functionality, and require no or minimal node specific configuration to function.
    - Simplicity: The system should be made as simple as possible, but no simpler.

- Amazon keeps its lips pretty tight about how S3 works, but according to Amazon, S3's design aims to provide scalability, high availability, and low latency at commodity costs.

- S3 stores arbitrary objects at up to 5GB in size, and each is accompanied by up to 2KB of metadata.
- Objects are organized by buckets.
- Each bucket is owned by an AWS account and the buckets are identified by a unique user assigned key.
- Buckets and objects are created, listed and retrieved using either a REST or SOAP interface.
- Objects can also be retrieved using the HTTP GET interface or via BitTorrent.
- An access control list restricts who can access the data in each bucket.
- Bucket names and keys are formulated so that they can be accessed using HTTP.
- Requests are authorized using an access control list associated with each bucket and object, for instance: http://s3.amazonaws.com/samplebucket/samplekey
- The Amazon AWS Authentication tools allow the bucket owner to create an authenticated URL with a set amount of time that the URL will be valid.
- Bucket items can also be accessed via a BitTorrent feed, enabling S3 to act as a seed for the client.
- Buckets can also be set up to save HTTP log information to another bucket.
- This information can be used for later data mining.

TWO MARK QUESTIONS

1. Illustrate architecture of a cloud is developed using three layers.



2. What is Market-Oriented Cloud Architecture?

   - As consumers rely on cloud providers to meet more of their computing needs, they will require a specific level of QoS to be maintained by their providers, in order to meet their objectives and sustain their operations.
   - Market-oriented resource management is necessary to regulate the supply and demand of cloud resources to achieve market equilibrium between supply and demand.

3. List the entities involved in the cloud platform.

   - Users or brokers and Request examiner
   - Pricing mechanism and VM Monitor mechanism
   - Accounting mechanism
   - Service Request Examiner and Admission Control mechanism
   - Dispatcher mechanism
   - Service Request Monitor mechanism

4. List the objectives of NIST Cloud Computing reference architecture

    - Illustrate and understand the various level of services
    - To provide technical reference
    - Categorize and compare services of cloud computing
    - Analysis of security, interoperatability and portability

5. Mention the major actors involved in NIST reference model.
    - Cloud consumer
    - Cloud provider
    - Cloud auditor
    - Cloud broker
    - Cloud carrier

6. Define service orchestration.

    - Service orchestration describes the automated arrangement, coordination and management of complex computing system.

7. Differentiate between Public cloud and Private Cloud.

    - A public cloud is one in which the cloud infrastructure and computing resources are made available to the general public over a public network.
    - A public cloud is owned by an organization selling cloud services, and serves a diverse pool of clients.
    - A private cloud gives a single Cloud Consumer's organization the exclusive access to and usage of the infrastructure and computational resources.
    - It may be managed either by the Cloud Consumer organization or by a third party, and may be hosted on the organization's premises (i.e. on-site private clouds) or outsourced to a hosting company (i.e. outsourced private clouds).

8. Tabulate the merits and demerits of Choosing a Community Cloud.

Merits	Demerits
• Ability to easily share and collaborate • Lower cost	• Not the right choice for every organization • Slow adoption to date

9. What is IaaS or HaaS?

- Infrastructure or Hardware-as-a-Service (IaaS/HaaS) solutions are the most popular and developed market segment of cloud computing.
- They deliver customizable infrastructure on demand.

10. What is PaaS?

- Platform-as-a-Service (PaaS) solutions provide a development and deployment platform for running applications in the cloud.
- They constitute the middleware on top of which applications are built.

11. Classify Pass Implementation

- PaaS implementation classified into three wide categories:
- PaaS-I, PaaS-II, and PaaS-III.

12. What is SaaS?

- Software-as-a-Service (SaaS) is a software delivery model that provides access to applications through the Internet as a Web-based service.
- It provides a means to free users from complex hardware and software management by offloading such tasks to third parties, which build applications accessible to multiple users through a Web browser.

13. What is SaaS 2.0?

- SaaS 2.0 is not a new technology but transforms the way in which SaaS is used.

- In particular, SaaS 2.0 is focused on providing a more robust infrastructure and application platforms driven by SLAs.
- SaaS 2.0 will focus on the rapid achievement of business objectives.

14. List the six architectural design challenges in cloud.

- Service Availability and Data Lock-in Problem
- Data Privacy and Security Concerns
- Unpredictable Performance and Bottlenecks
- Distributed Storage and Widespread Software Bugs
- Cloud Scalability, Interoperability, and Standardization
- Software Licensing and Reputation Sharing

15. What is cloud storage?

- Cloud storage means storing the data with a cloud service provider rather than on a local system. The end user can access the data stored on the cloud using an Internet link.
- Cloud storage has a number of advantages over traditional data storage.
- If the users stored some data on a cloud, they can get at it from any location that has Internet access.

16. What is Storage-as-a-Service?

- The term Storage as a Service means that a third-party provider rents space on their storage to end users who lack the budget or capital budget to pay for it on their own.
- It is also ideal when technical personnel are not available or have inadequate knowledge to implement and maintain that storage infrastructure.

17. List the real time examples for cloud storage.

- Google Docs allows users to upload documents, spreadsheets, and presentations to Google's data servers.

- Web email providers like Gmail, Hotmail, and Yahoo! Mail store email messages on their own servers.
- Flickr and Picasa host millions of digital photographs. YouTube hosts millions of user-uploaded video files.
- Hostmonster and GoDaddy store files and data for many client web sites.
- Facebook and MySpace are social networking sites and allow members to post pictures and other content.
- MediaMax and Strongspace offer storage space for any kind of digital data.

18. How to secure data in cloud storage?

- Most systems use a combination of following techniques:
    - Encryption
    - Authentication processes
    - Authorization practices

19. List the advantages of cloud storage.

- Storing data on the cloud is advantageous, because it allows you to protect your data in case there's a disaster.
- Having your data stored off-site can be the difference between closing your door for good or being down for a few days or weeks.
- Which storage vendor to go with can be a complex issue, and how the end user technology interacts with the cloud can be complex.

20. What is S3?

- The best-known cloud storage service is Amazon's Simple Storage Service (S3), which launched in 2006.
- Amazon S3 is designed to make web-scale computing easier for developers.
- Amazon S3 provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the Web.
- It gives any developer access to the same highly scalable data storage infrastructure that Amazon uses to run its own global network of web sites.

21. What are the design requirements considers by Amazon to build S3?

- Scalable
- Reliable
- Fast
- Inexpensive
- Simple

22. What are the design principles considers by Amazon to meet S3 requirements?

- Decentralization
- Autonomy
- Local responsibility
- Controlled concurrency
- Failure toleration
- Controlled parallelism
- Symmetry
- Simplicity

23. How the data stored in S3?

- S3 stores arbitrary objects at up to 5GB in size, and each is accompanied by up to 2KB of metadata.
- Objects are organized by buckets.
- Each bucket is owned by an AWS account and the buckets are identified by a unique, user-assigned key.
- Buckets and objects are created, listed, and retrieved using either a REST-style or SOAP interface.

UNIT IV RESOURCE MANAGEMENT AND SECURITY IN CLOUD

Inter Cloud Resource Management –Resource Provisioning and Resource Provisioning Methods –Global Exchange of Cloud Resources –Security Overview –Cloud Security Challenges –Software-as-a-Service Security –Security Governance –Virtual Machine Security – IAM –Security Standards.

4.1 Inter Cloud Resource Management

Cloud application (SaaS)
Cloud Software environment (PaaS)
Cloud software infrastructure (IaaS, DaaS, CaaS)
Collocation cloud services (LaaS)
Network cloud services (NaaS)
Hardware or Virtualization cloud Services (HaaS)

Figure 4.1 A stack of six layers of cloud services

- Figure 4.1 shows six layers of cloud services, ranging from hardware, network, and collocation to infrastructure, platform and software applications.
- The cloud platform provides PaaS, which sits on top of the IaaS infrastructure. The top layer offers SaaS.
- The bottom three layers are more related to physical requirements.
- The bottommost layer provides Hardware as a Service (HaaS).

- The next layer is for interconnecting all the hardware components and it is simply called Network as a Service (NaaS).

- Virtual LANs fall within the scope of NaaS.

- The next layer up offers Location as a Service (LaaS), which provides a collocation service to house, power and secure all the physical hardware as well as network resources.

- Some authors say this layer provides Security as a Service (SaaS).

- The cloud infrastructure layer can be further subdivided as Data as a Service (DaaS) and Communication as a Service (CaaS) in addition to compute and storage in IaaS.

- The three cloud models as viewed by different players.

- From the software vendor perspective, application performance on a given cloud platform is most important.

- From the provider perspective, cloud infrastructure performance is the primary concern.

- From the end users perspective, the quality of services, including security, is the most important.

- CRM offered the first SaaS on the cloud successfully.

- The approach is to widen market coverage by investigating customer behaviors and revealing opportunities by statistical analysis.

- SaaS tools also apply to distributed collaboration, financial and human resources management. These cloud services have been growing rapidly in recent years.

- PaaS is provided by Google, Salesforce.com, Facebook, and so on.
- IaaS is provided by Amazon, Windows Azure, RackRack, and so on.
- Based on the observations of some typical cloud computing instances, such as Google, Microsoft, and Yahoo!, the overall software stack structure of cloud computing software can be viewed as layers.
- Each layer has its own purpose and provides the interface for the upper layers just as the traditional soware stack does. However, the lower layers are not completely transparent to the upper layers.
- The platform for running cloud computing services can be either physical servers or virtual servers.
- By using VMs, the platform can be flexible; It means the running services are not bound to specific hardware platforms.
- The software layer on top of the platform is the layer for storing massive amounts of data.
- This layer acts like the file system in a traditional single machine. Other layers running on top of the file system are the layers for executing cloud computing applications.
- The next layers are the components in the software stack.

4.1.1 Runtime Support Services

- As in a cluster environment, there are also some runtime supporting services in the cloud computing environment.
- Cluster monitoring is used to collect the runtime status of the entire cluster.

- The scheduler queues the tasks submitted to the whole cluster and assigns the tasks to the processing nodes according to node availability.

- The distributed scheduler for the cloud application has special characteristics that can support cloud applications, such as scheduling the programs written in MapReduce style.

- The runtime support system keeps the cloud cluster working properly with high efficiency.

- Runtime support is software needed in browser initiated applications applied by thousands of cloud customers.

- The SaaS model provides the software applications as a service, rather than lifting users purchase the software.

- As a result, on the customer side, there is no upfront investment in servers or software licensing.

- On the provider side, costs are rather low, compared with conventional hosting of user applications.

- The customer data is stored in the cloud that is either vendor proprietary or a publicly hosted cloud supporting PaaS and IaaS.

4.2 Resource Provisioning

- Providers supply cloud services by signing SLAs with end users.

- The SLAs must commit sufficient resources such as CPU, memory and bandwidth that the user can use for a preset period.

- Under provisioning of resources will lead to broken SLAs and penalties.

- Over provisioning of resources will lead to resource underutilization, and consequently, a decrease in revenue for the provider.

- Deploying an autonomous system to efficiently provision resources to users is a challenging problem.

- Efficient VM provisioning depends on the cloud architecture and management of cloud infrastructures.

- Resource provisioning schemes also demand fast discovery of services and data in cloud computing infrastructures.

- In a virtualized cluster of servers, this demands efficient installation of VMs, live VM migration and fast recovery from failures.

- To deploy VMs, users treat them as physical hosts with customized operating systems for specific applications.

- For example, Amazon's EC2 uses Xen as the virtual machine monitor (VMM). The same VMM is used in IBM's Blue Cloud.

- In the EC2 platform, some predefined VM templates are also provided. Users can choose different kinds of VMs from the templates.

- IBM's Blue Cloud does not provide any VM templates.

4.3 Resource Provisioning Methods

- Figure 4.2 shows three cases of static cloud resource provisioning policies.

- In case (a), over provisioning with the peak load causes heavy resource waste (shaded area).

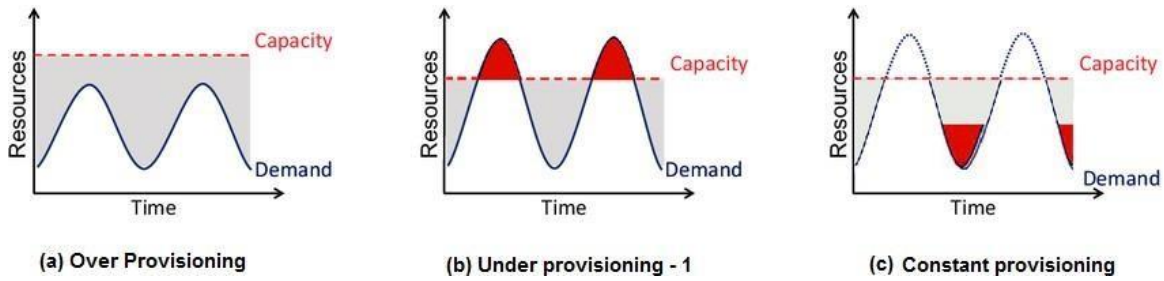


Figure 4.2 Three cases of resource provisioning

- In case (b), under provisioning (along the capacity line) of resources results in losses by both user and provider in that paid demand by the users (the shaded area above the capacity) is not served and wasted resources still exist for those demanded areas below the provisioned capacity.

- In case (c), the constant provisioning of resources with fixed capacity to a declining user demand could result in even worse resource waste.

- The user may give up the service by canceling the demand, resulting in reduced revenue for the provider.

- Both the user and provider may be losers in resource provisioning without elasticity.

- The demand-driven method provides static resources and has been used in grid computing for many years.

- The event-driven method is based on predicted workload by time.

- The popularity-driven method is based on Internet traffic monitored.

4.3.1 Demand-Driven Resource Provisioning

- This method adds or removes computing instances based on the current utilization level of the allocated resources.

- The demand-driven method automatically allocates two Xeon processors for the user application, when the user was using one Xeon processor more than 60 percent of the time for an extended period

- In general, when a resource has surpassed a threshold for a certain amount of time, the scheme increases that resource based on demand.

- When a resource is below a threshold for a certain amount of time, that resource could be decreased accordingly.

- Amazon implements such an auto-scale feature in its EC2 platform. This method is easy to implement.

- The scheme does not work out right if the workload changes abruptly.

4.3.2 Event-Driven Resource Provisioning

- This scheme adds or removes machine instances based on a specific time event.

- The scheme works better for seasonal or predicted events such as Christmastime in the West and the Lunar New Year in the East.

- During these events, the number of users grows before the event period and then decreases during the event period.

- This scheme anticipates peak traffic before it happens.

- The method results in a minimal loss of QoS, if the event is predicted correctly.

- Otherwise, wasted resources are even greater due to events that do not follow a fixed pattern.

4.3.3 Popularity-Driven Resource Provisioning

- In this method, the Internet searches for popularity of certain applications and creates the instances by popularity demand.
- The scheme anticipates increased traffic with popularity.
- Again, the scheme has a minimal loss of QoS, if the predicted popularity is correct.
- Resources may be wasted if traffic does not occur as expected.

4.4 Global Exchange of Cloud Resources

- In order to support a large number of application service consumers from around the world, cloud infrastructure providers (i.e., IaaS providers) have established data centers in multiple geographical locations to provide redundancy and ensure reliability in case of site failures.
- For example, Amazon has data centers in the United States (e.g., one on the East Coast and another on the West Coast) and Europe.
- However, currently Amazon expects its cloud customers (i.e., SaaS providers) to express a preference regarding where they want their application services to be hosted.
- Amazon does not provide seamless/automatic mechanisms for scaling its hosted services across multiple geographically distributed data centers.
- This approach has many shortcomings.
    - First, it is difficult for cloud customers to determine in advance the best location for hosting their services as they may not know the origin of consumers of their services.

- Second, SaaS providers may not be able to meet the QoS expectations of their service consumers originating from multiple geographical locations.

- Figure 4.3 shows the high-level components of the Melbourne group's proposed Inter Cloud architecture.

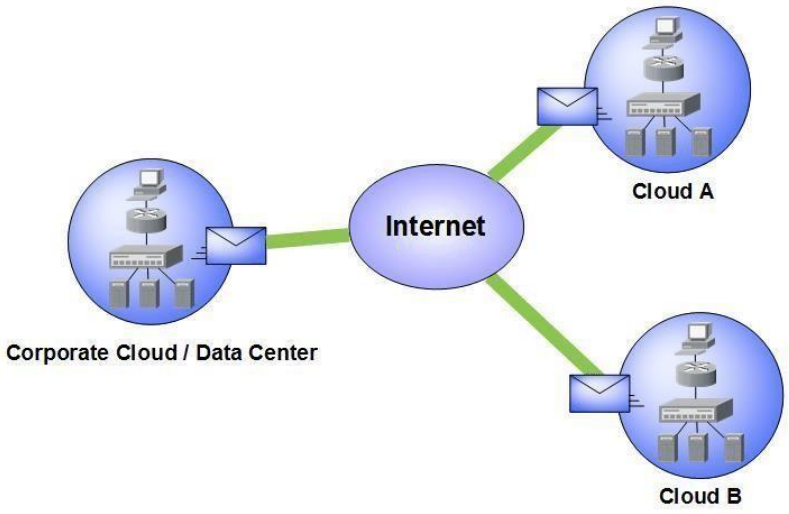


Figure 4.3 Inter cloud architecture

- In addition, no single cloud infrastructure provider will be able to establish its data centers at all possible locations throughout the world.

- As a result, cloud application service (SaaS) providers will have difficulty in meeting QoS expectations for all their consumers.

- The Cloudbus Project at the University of Melbourne has proposed InterCloud architecture supporting brokering and exchange of cloud resources for scaling applications across multiple clouds.

- By realizing InterCloud architectural principles in mechanisms in their offering,

    - Cloud providers will be able to dynamically expand or resize their provisioning capability based on sudden spikes in workload demands by leasing available computational and storage capabilities from other cloud service providers.

    - Operate as part of a market driven resource leasing federation, where application service providers such as Salesforce.com host their services based on negotiated SLA contracts driven by competitive market prices.
    - Deliver on-demand, reliable, cost-effective, and QoS-aware services based on virtualization technologies while ensuring high QoS standards and minimizing service costs.

- They need to be able to utilize market-based utility models as the basis for provisioning of virtualized software services and federated hardware infrastructure among users with heterogeneous applications.

- They consist of client brokering and coordinator services that support utility-driven federation of clouds:
    - Application scheduling
    - Resource allocation
    - Migration of workloads

- The architecture cohesively couples the administratively and topologically distributed storage and compute capabilities of clouds as part of a single resource leasing abstraction.

- The Cloud Exchange (CEx) acts as a market maker for bringing together service producers and consumers.

- It aggregates the infrastructure demands from application brokers and evaluates them against the available supply currently published by the cloud coordinators.

- It supports trading of cloud services based on competitive economic models such as commodity markets and auctions.

- An SLA specifies the details of the service to be provided in terms of metrics agreed upon by all parties, and incentives and penalties for meeting and violating the expectations, respectively.

- The availability of a banking system within the market ensures that financial transactions pertaining to SLAs between participants are carried out in a secure and dependable environment.

4.5 Security Overview

- Cloud service providers must learn from the managed service provider (MSP) model and ensure that their customer's applications and data are secure if they hope to retain their customer base and competitiveness.

- Today, enterprises are looking toward cloud computing horizons to expand their on-premises infrastructure, but most cannot afford the risk of compromising the security of their applications and data.

- For example, IDC recently conducted a survey1 (Figure 4.4) of 244 IT executives/CIOs and their line-of-business (LOB) colleagues to gauge their opinions and understand their companies' use of IT cloud services.

- Security ranked first as the greatest challenge or issue of cloud computing.

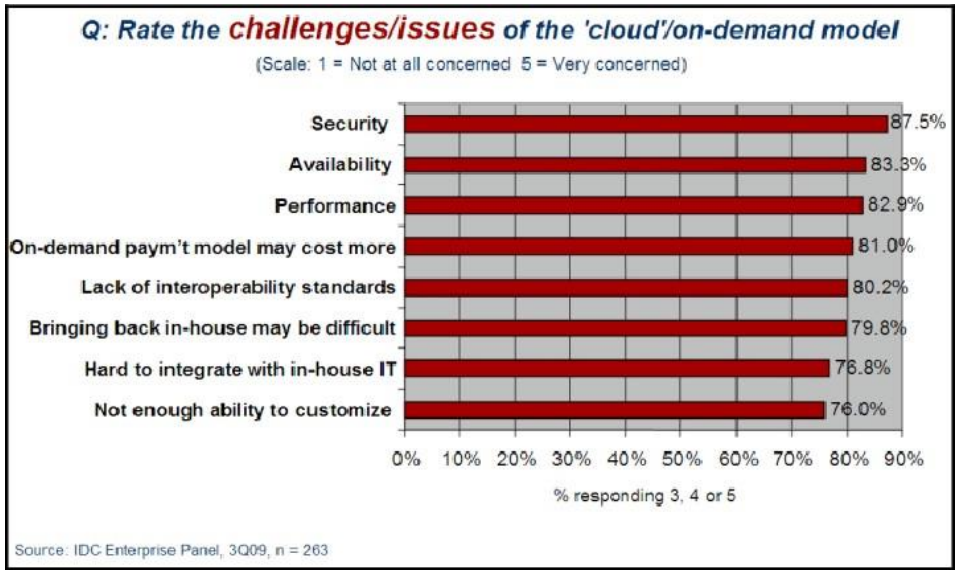


Figure 4.4 Results of IDC survey

- Moving critical applications and sensitive data to public and shared cloud environments is of great concern for those corporations that are moving beyond their data center's network perimeter defense.

- To alleviate these concerns, a cloud solution provider must ensure that customers will continue to have the same security and privacy controls over their applications and services.

- In addition, solution provider give evidence to customers that their organization and customers are secure and they can meet their service level agreements, and that they can prove compliance to auditors.

4.6 Cloud Security Challenges

- Although virtualization and cloud computing can help companies accomplish more by breaking the physical bonds between an IT infrastructure and its users, heightened security threats must be overcome in order to benefit fully from this new computing paradigm.

- Enterprise security is only as good as the least reliable partner, department and vendor.

- With the cloud model, the cloud consumer's loss control over physical security.

- In a public cloud, the consumers are sharing computing resources with other companies.

- In a shared pool outside the enterprise, users do not have any knowledge or control of where the resources run.

- Storage services provided by one cloud vendor may be incompatible with another vendor's services should you decide to move from one to the other.

- Ensuring the integrity of the data really means that it changes only in response to authorized transactions.

- The immature use of mash up technology (combinations of web services), which is fundamental to cloud applications, is inevitably going to cause unwitting security vulnerabilities in those applications.

- Since access to logs is required for Payment Card Industry Data Security Standard (PCI DSS) compliance and may be requested by auditors and regulators, security managers need to make sure to negotiate access to the provider's logs as part of any service agreement.

- Cloud applications undergo constant feature additions and users must keep up to date with application improvements to be sure they are protected.

- The speed at which applications will change in the cloud will affect both the SDLC and security.

- Security needs to move to the data level, so that enterprises can be sure their data is protected wherever it goes.

- Sensitive data is the domain of the enterprise, not the cloud computing provider.

- One of the key challenges in cloud computing is data level security.

- Most compliance standards do not envision compliance in a world of cloud computing.

- There is a huge body of standards that apply for IT security and compliance, governing most business interactions that will, over time, have to be translated to the cloud.

- SaaS makes the process of compliance more complicated, since it may be difficult for a customer to discern where its data resides on a network controlled by its SaaS provider, or a partner of that provider, which raises all sorts of compliance issues of data privacy, segregation, and security.

- Security managers will need to pay particular attention to systems that contain critical data such as corporate financial information or source code during the transition to server virtualization in production environments.

- Outsourcing means losing significant control over data, and while this is not a good idea from a security perspective, the business ease and financial savings will continue to increase the usage of these services.

- Security managers will need to work with their company's legal staff to ensure that appropriate contract terms are in place to protect corporate data and provide for acceptable service level agreements.

- Cloud based services will result in many mobile IT users accessing business data and services without traversing the corporate network.

- This will increase the need for enterprises to place security controls between mobile users and cloud based services.

- Although traditional data center security still applies in the cloud environment, physical segregation and hardware based security cannot protect against attacks between virtual machines on the same server.

- Administrative access is through the Internet rather than the controlled and restricted direct or on-premises connection that is adhered to in the traditional data center model.

- This increases risk and exposure and will require stringent monitoring for changes in system control and access control restriction.

- Proving the security state of a system and identifying the location of an insecure virtual machine will be challenging.

- The co-location of multiple virtual machines increases the attack surface and risk of virtual machine to virtual machine compromise.

- Localized virtual machines and physical servers use the same operating systems as well as enterprise and web applications in a cloud server environment, increasing the threat of an attacker or malware exploiting vulnerabilities in these systems and applications remotely.

- Virtual machines are vulnerable as they move between the private cloud and the public cloud.

- A fully or partially shared cloud environment is expected to have a greater attack surface and therefore can be considered to be at greater risk than a dedicated resources environment.

- Operating system and application files are on a shared physical infrastructure in a virtualized cloud environment and require system, file, and activity monitoring to provide confidence and auditable proof to enterprise customers that their resources have not been compromised or tampered with.

- In the cloud computing environment, the enterprise subscribes to cloud computing resources, and the responsibility for patching is the subscriber's rather than the cloud computing vendors.

- The need for patch maintenance vigilance is imperative.

- Data is fluid in cloud computing and may reside in on-premises physical servers, on-premises virtual machines, or off-premises virtual machines running on cloud computing resources and this will require some rethinking on the part of auditors and practitioners alike.

- To establish zones of trust in the cloud, the virtual machines must be self-defending, effectively moving the perimeter to the virtual machine itself.

- Enterprise perimeter security (i.e., firewalls, demilitarized zones [DMZs], network segmentation, intrusion detection and prevention systems [IDS/IPS], monitoring tools, and the associated security policies) only controls the data that resides and transits behind the perimeter.

- In the cloud computing world, the cloud computing provider is in charge of customer data security and privacy.

4.7 Software-as-a-Service Security

- Cloud computing models of the future will likely combine the use of SaaS (and other XaaS's as appropriate), utility computing and Web 2.0 collaboration technologies to leverage the Internet to satisfy their customer needs.

- New business models being developed as a result of the move to cloud computing are creating not only new technologies and business operational processes but also new security requirements and challenges as described previously.

- As the most recent evolutionary step in the cloud service model (Figure 4.5), SaaS will likely remain the dominant cloud service model for the predictable future and the area where the most critical need for security practices and oversight will reside.

- The technology analyst and consulting firm Gartner lists seven security issues which one should discuss with a cloud computing vendor.

- Privileged user access inquires about who has specialized access to data and about the hiring and management of such administrators.

- Regulatory compliance makes sure that the vendor is willing to undergo external audits and/or security certifications.

- Data location does the provider allow for any control over the location of data.

- Data segregation makes encryption is available at all stages and that these encryption schemes were designed and tested by experienced professionals.

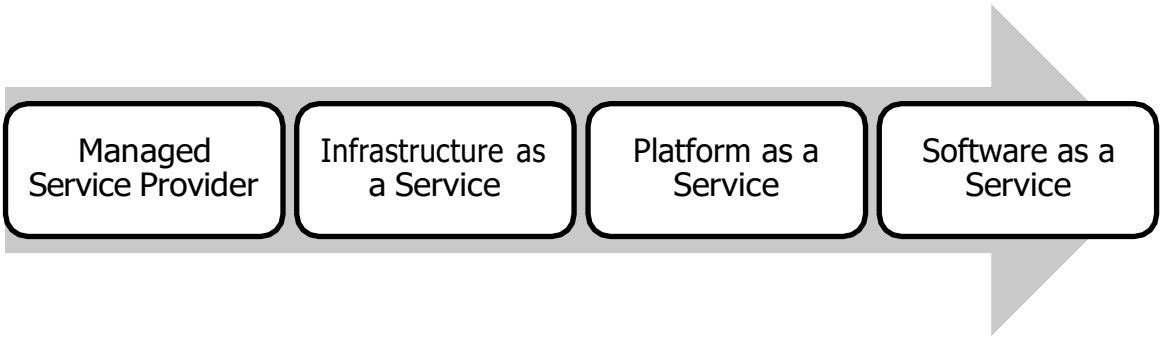


Figure 4.5 Evolution of cloud services

- Recovery is the way to find out what will happen to data in the case of a disaster. And also it covers the way to perform complete restoration.

- Investigative support does the vendor have the ability to investigate any inappropriate or illegal activity.

- Long-term viability focus on data if the company goes out of business and format and process behind the returned data.

- To address the security issues listed above, SaaS providers will need to incorporate and enhance security practices used by the managed service providers and develop new ones as the cloud computing environment evolves.

4.8 Security Governance

- A security steering committee should be developed whose objective is to focus on providing guidance about security initiatives and alignment with business and IT strategies.

- A charter for the security team is typically one of the first deliverables from the steering committee.

- This charter must clearly define the roles and responsibilities of the security team and other groups involved in performing information security functions.

- Lack of a formalized strategy can lead to an unsustainable operating model and security level as it evolves.

- In addition, lack of attention to security governance can result in key needs of the business not being met, including but not limited to, risk management, security monitoring, application security, and sales support.

- Lack of proper governance and management of duties can also result in potential security risks being left unaddressed and opportunities to improve the business being missed because the security team is not focused on the key security functions and activities that are critical to the business.

4.9 Virtual Machine Security

- In the cloud environment, physical servers are consolidated to multiple virtual machine instances on virtualized servers.

- Not only can data center security teams replicate typical security controls for the data center at large to secure the virtual machines, they can also advise their customers on how to prepare these machines for migration to a cloud environment when appropriate.

- Firewalls, intrusion detection and prevention, integrity monitoring and log inspection can all be deployed as software on virtual machines to increase protection as well as maintain compliance integrity of servers and applications as virtual resources move from on-premises to public cloud environments.

- By deploying this traditional line of defense to the virtual machine itself, the user can enable critical applications and data to be moved to the cloud securely.

- To facilitate the centralized management of a server firewall policy, the security software loaded onto a virtual machine should include a bidirectional stateful firewall that enables virtual machine isolation and location awareness, thereby enabling a tightened policy and the flexibility to move the virtual machine from on-premises to cloud resources.

- Integrity monitoring and log inspection software must be applied at the virtual machine level.

- This approach to virtual machine security, which connects the machine back to the mother ship, has some advantages in that the security software can be put into a single software agent that provides for consistent control and management throughout the cloud while integrating seamlessly back into existing security infrastructure investments, providing economies of scale, deployment, and cost savings for both the service provider and the enterprise.

4.10 IAM

- Identity and access management is a critical function for every organization and a fundamental expectation of SaaS customers is that the principle of least privilege is granted to their data.

- The principle of least privilege states that only the minimum access necessary to perform an operation should be granted, and that access should be granted only for the minimum amount of time necessary.

- However, business and IT groups will need and expect access to systems and applications.

- The advent of cloud services and services on demand is changing the identity management landscape.

- Most of the current identity management solutions are focused on the enterprise and typically are architected to work in a very controlled, static environment.

- User-centric identity management solutions such as federated identity management make some assumptions about the parties involved and their related services.

- In the cloud environment, where services are offered on demand and they can continuously evolve, aspects of current models such as trust assumptions, privacy implications, and operational aspects of authentication and authorization, will be challenged.

- Meeting these challenges will require a balancing act for SaaS providers as they evaluate new models and management processes for IAM to provide end-to-end trust and identity throughout the cloud and the enterprise.

- Another issue will be finding the right balance between usability and security. If a good balance is not achieved, both business and IT groups may be affected by barriers to completing their support and maintenance activities efficiently.

4.11 Security Standards

- Security standards define the processes, procedures, and practices necessary for implementing a security program.

- These standards also apply to cloud related IT activities and include specific steps that should be taken to ensure a secure environment is maintained that provides privacy and security of confidential information in a cloud environment.

- Security standards are based on a set of key principles intended to protect this type of trusted environment.

- Messaging standards, especially for security in the cloud, must also include nearly all the same considerations as any other IT security endeavor.

- Security (SAML OAuth, OpenID, SSL/TLS) A basic philosophy of security is to have layers of defense, a concept known as defense in depth.

- This means having overlapping systems designed to provide security even if one system fails. An example is a firewall working in conjunction with an intrusion-detection system (IDS).

- Defense in depth provides security because there is no single point of failure and no single entry vector at which an attack can occur.

- For this reason, a choice between implementing network security in the middle part of a network (i.e., in the cloud) or at the endpoints is a false dichotomy.

- No single security system is a solution by itself, so it is far better to secure all systems.

- This type of layered security is precisely what we are seeing develop in cloud computing.

- Traditionally, security was implemented at the endpoints, where the user controlled access.

- An organization had no choice except to put firewalls, IDSs, and antivirus software inside its own network.

- Today, with the advent of managed security services offered by cloud providers, additional security can be provided inside the cloud.

4.11.1 Security Assertion Markup Language (SAML)

- SAML is an XML-based standard for communicating authentication, authorization and attribute information among online partners.

- It allows businesses to securely send assertions between partner organizations regarding the identity and entitlements of a principal.

- The Organization for the Advancement of Structured Information Standards (OASIS) Security Services Technical Committee is in charge of defining, enhancing and maintaining the SAML specifications.

- SAML is built on a number of existing standards, namely, SOAP, HTTP and XML. SAML relies on HTTP as its communications protocol and specifies the use of SOAP (currently, version 1.1).

- Most SAML transactions are expressed in a standardized form of XML.

- SAML assertions and protocols are specified using XML schema.

- Both SAML 1.1 and SAML 2.0 use digital signatures (based on the XML Signature standard) for authentication and message integrity.

- XML encryption is supported in SAML 2.0, though SAML 1.1 does not have encryption capabilities.

- SAML defines XML based assertions and protocols, bindings and profiles.

- The term SAML Core refers to the general syntax and semantics of SAML assertions as well as the protocol used to request and transmit those assertions from one system entity to another.

- SAML protocol refers to what is transmitted, not how it is transmitted.

- A SAML binding determines how SAML requests and responses map to standard messaging protocols. An important (synchronous) binding is the SAML SOAP binding.

- SAML standardizes queries for, and responses that contain, user authentication, entitlements and attribute information in an XML format.

- This format can then be used to request security information about a principal from a SAML authority.

- A SAML authority, sometimes called the asserting party. It is a platform or application that can relay security information.

- The relying party (or assertion consumer or requesting party) is a partner site that receives the security information.

- The exchanged information deals with a subject's authentication status, access authorization, and attribute information.

- A subject is an entity in a particular domain.

- A person identified by an email address is a subject, as might be a printer.

- SAML assertions are usually transferred from identity providers to service providers.

- Assertions contain statements that service providers use to make access control decisions.

- Three types of statements are provided by SAML:

    - Authentication statements
    - Attribute statements
    - Authorization decision statements

- SAML assertions contain a packet of security information in this form:

<saml: Asssertion A>

<Authentication>

…

</Authentication>

<Attribute>

…

</Attribute>

<Authentication>

…

</Authentication>

</saml: Asssertion A>

- The assertion shown above is interpreted as follows:
  Assertion A, issued at time T by issuer I, regarding subject S, provided conditions C are valid.

- Authentication statements assert to a service provider that the principal did indeed authenticate with an identity provider at a particular time using a particular method of authentication.

- Other information about the authenticated principal (called the authentication context) may be disclosed in an authentication statement.

- An attribute statement asserts that a subject is associated with certain attributes.

- An attribute is simply a name-value pair.

- An authorization decision statement asserts that a subject is permitted to perform action A on resource R given evidence E.

- A SAML protocol describes how certain SAML elements (including assertions) are packaged within SAML request and response elements

- Generally, a SAML protocol is a simple request–response protocol.

- The most important type of SAML protocol request is a query.

- A service provider makes a query directly to an identity provider over a secure back channel. For this reason, query messages are typically bound to SOAP.

- Corresponding to the three types of statements, there are three types of SAML queries:
    - Authentication query
    - Attribute query
    - Authorization decision query.
- Of these, the attribute query is perhaps most important. The result of an attribute query is a SAML response containing an assertion, which itself contains an attribute statement.

4.11.2 Open Authentication (OAuth)

- OAuth is an open protocol, initiated by Blaine Cook and Chris Messina, to allow secure API authorization in a simple, standardized method for various types of web applications.
- Cook and Messina had concluded that there were no open standards for API access delegation.
- The OAuth discussion group was created in April 2007, for the small group of implementers to write the draft proposal for an open protocol.
- DeWitt Clinton of Google learned of the OAuth project and expressed interest in supporting the effort.
- In July 2007, the team drafted an initial specification and it was released in October of the same year.
- OAuth is a method for publishing and interacting with protected data.
- For developers, OAuth provides users access to their data while protecting account credentials.

- OAuth allows users to grant access to their information, which is shared by the service provider and consumers without sharing all of their identity.

- The Core designation is used to stress that this is the baseline, and other extensions and protocols can build on it.

- By design, OAuth Core 1.0 does not provide many desired features (e.g., automated discovery of endpoints, language support, support for XML-RPC and SOAP, standard definition of resource access, OpenID integration, signing algorithms, etc.).

- This intentional lack of feature support is viewed by the authors as a significant benefit.

- The Core deals with fundamental aspects of the protocol, namely, to establish a mechanism for exchanging a user name and password for a token with defined rights and to provide tools to protect the token.

- It is important to understand that security and privacy are not guaranteed by the protocol.

- In fact, OAuth by itself provides no privacy at all and depends on other protocols such as SSL to accomplish that.

- OAuth can be implemented in a secure manner.

- In fact, the specification includes substantial security considerations that must be taken into account when working with sensitive data.

- With Oauth, sites use tokens coupled with shared secrets to access resources.

- Secrets, just like passwords, must be protected.

4.11.3 OpenID

- OpenID is an open, decentralized standard for user authentication and access control that allows users to log onto many services using the same digital identity.

- It is a single-sign-on (SSO) method of access control. As such, it replaces the common log-in process (i.e., a log-in name and a password) by allowing users to log in once and gain access to resources across participating systems.

- The original OpenID authentication protocol was developed in May 2005 by Brad Fitzpatrick, creator of the popular community web site LiveJournal.

- In late June 2005, discussions began between OpenID developers and other developers from an enterprise software company named NetMesh.

- These discussions led to further collaboration on interoperability between OpenID and NetMesh's similar Light-Weight Identity (LID) protocol.

- The direct result of the collaboration was the Yadis discovery protocol, which was announced on October 24, 2005.

- The Yadis specification provides a general-purpose identifier for a person and any other entity, which can be used with a variety of services.

- It provides syntax for a resource description document identifying services available using that identifier and an interpretation of the elements of that document.

- Yadis discovery protocol is used for obtaining a resource description document, given that identifier.

- Together these enable coexistence and interoperability of a rich variety of services using a single identifier.

- The identifier uses a standard syntax and a well established namespace and requires no additional namespace administration infrastructure.

- An OpenID is in the form of a unique URL and is authenticated by the entity hosting the OpenID URL.

- The OpenID protocol does not rely on a central authority to authenticate a user's identity.

- Neither the OpenID protocol nor any web sites requiring identification can mandate that a specific type of authentication be used; nonstandard forms of authentication such as smart cards, biometrics, or ordinary passwords are allowed.

- A typical scenario for using OpenID might be something like this:

    - A user visits a web site that displays an OpenID log in form
    - Unlike a typical log in form, which has fields for user name and password, the OpenID log in form has only one field for the OpenID identifier (which is an OpenID URL).
    - This form is connected to an implementation of an OpenID client library.
    - A user will have previously registered an OpenID identifier with an OpenID identity provider.
    - The user types this OpenID identifier into the OpenID log-in form.
    - The relying party then requests the web page located at that URL and reads an HTML link tag to discover the identity provider service URL.

- With OpenID 2.0, the client discovers the identity provider service URL by requesting the XRDS document (also called the Yadis document) with the content type application/xrds+xml, which may be available at the target URL but is always available for a target XRI.

- There are two modes by which the relying party can communicate with the identity provider: checkid_immediate and checkid_setup.

- In checkid_immediate, the relying party requests that the provider not interact with the user. All communication is relayed through the user's browser without explicitly notifying the user.

- In checkid_setup, the user communicates with the provider server directly using the same web browser as is used to access the relying party site.

- OpenID does not provide its own authentication methods, but if an identity provider uses strong authentication, OpenID can be used for secure transactions.

- SSL/TLS Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), are cryptographically secure protocols designed to provide security and data integrity for communications over TCP/IP.

- TLS and SSL encrypt the segments of network connections at the transport layer.

- Several versions of the protocols are in general use in web browsers, email, instant messaging and Voice-over-IP (VoIP).

- TLS is an IETF standard protocol which was last updated in RFC 5246.

- The TLS protocol allows client/server applications to communicate across a network in a way specifically designed to prevent eavesdropping, tampering, and message forgery.

- TLS provides endpoint authentication and data confidentiality by using cryptography.

- TLS authentication is one way in which the server is authenticated, because the client already knows the server's identity. In this case, the client remains unauthenticated.

- TLS also supports a more secure bilateral connection mode whereby both ends of the connection can be assured that they are communicating with whom they believe they are connected.

- This is known as mutual authentication.

- Mutual authentication requires the TLS client side to also maintain a certificate.

- TLS involves three basic phases:

    - Peer negotiation for algorithm support
    - Key exchange and authentication
    - Symmetric cipher encryption and message authentication

TWO MARK QUESTIONS

1. List the runtime supporting services in the cloud computing environment.

    - Cluster monitoring is used to collect the runtime status of the entire cluster.
    - The scheduler queues the tasks submitted to the whole cluster and assigns the tasks to the processing nodes according to node availability.
    - The distributed scheduler for the cloud application has special characteristics that can support cloud applications, such as scheduling the programs written in MapReduce style.

2. Why inter cloud resource management requires runtime support system?

    - The runtime support system keeps the cloud cluster working properly with high efficiency.
    - Runtime support is software needed in browser-initiated applications applied by thousands of cloud customers.

3. Differentiate between over provisioning and under provisioning.

   - Under provisioning of resources will lead to broken SLAs and penalties.
   - Over provisioning of resources will lead to resource underutilization, and consequently, a decrease in revenue for the provider.
   - over provisioning with the peak load causes heavy resource waste (shaded area).
   - under provisioning (along the capacity line) of resources results in losses by both user and provider in that paid demand by the users (the shaded area above the capacity) is not served and wasted resources still exist for those demanded areas below the provisioned capacity.

4. List the various resource provisioning methods.

   - demand-driven resource provisioning
   - Event-Driven Resource Provisioning
   - Popularity-Driven Resource Provisioning
   - Global Exchange of Cloud Resources

5. What is demand-driven resource provisioning?

   - This method adds or removes computing instances based on the current utilization level of the allocated resources.
   - The demand-driven method automatically allocates two Xeon processors for the user application, when the user was using one Xeon processor more than 60 percent of the time for an extended period

6. Write short notes on cloud security.

   - Cloud service providers must learn from the managed service provider (MSP) model and ensure that their customer's applications and data are secure if they hope to retain their customer base and competitiveness.
   - Security ranked first as the greatest challenge or issue of cloud computing.

7. List the challenges in cloud security.

- Enterprise security is only as good as the least reliable partner, department, or vendor.
- With the cloud model, users lose control over physical security.
- In a public cloud, the users are sharing computing resources with other companies.
- In a shared pool outside the enterprise, users don't have any knowledge or control of where the resources run.
- Storage services provided by one cloud vendor may be incompatible with another vendor's services should you decide to move from one to the other.
- Ensuring the integrity of the data really means that it changes only in response to authorized transactions.

8. List the seven security issues with respect to cloud computing vendor.

- Privileged user access
- Regulatory compliance
- Data location
- Data segregation
- Recovery
- Investigative support
- Long-term viability

9. What is the purpose of security governance?

- A security steering committee should be developed whose objective is to focus on providing guidance about security initiatives and alignment with business and IT strategies.
- A charter for the security team is typically one of the first deliverables from the steering committee.

10. How to perform virtual machine security?

- Firewalls, intrusion detection and prevention, integrity monitoring, and log inspection can all be deployed as software on virtual machines to increase protection and maintain compliance integrity of servers and applications as virtual resources move from on-premises to public cloud environments.
- Integrity monitoring and log inspection software must be applied at the virtual machine level.

11. Define IAM.

- Identity and access management is a critical function for every organization, and a fundamental expectation of SaaS customers is that the principle of least privilege is granted to their data.

12. Why cloud requires security standards?

- Security standards define the processes, procedures, and practices necessary for implementing a security program.
- These standards also apply to cloud related IT activities and include specific steps that should be taken to ensure a secure environment is maintained that provides privacy and security of confidential information in a cloud environment.

13. What is SAML?

- Security Assertion Markup Language (SAML) is an XML-based standard for communicating authentication, authorization, and attribute information among online partners.
- It allows businesses to securely send assertions between partner organizations regarding the identity and entitlements of a principal.

14. List the types of statements are provided by SAML.

- Authentication statements
- Attribute statements
- Authorization decision statements

15. Describe about SAML protocol.

- A SAML protocol describes how certain SAML elements (including assertions) are packaged within SAML request and response elements
- SAML protocol is a simple request–response protocol.
- The most important type of SAML protocol request is a query.

16. List the types of SAML queries.

- Authentication query
- Attribute query
- Authorization decision query.

17. What is OAuth?

- OAuth (Open authentication) is an open protocol, initiated by Blaine Cook and Chris Messina, to allow secure API authorization in a simple, standardized method for various types of web applications.
- OAuth is a method for publishing and interacting with protected data.
- OAuth allows users to grant access to their information, which is shared by the service provider and consumers without sharing all of their identity.

18. What is the purpose of OpenID?

- OpenID is an open, decentralized standard for user authentication and access control that allows users to log onto many services using the same digital identity.
- It is a single-sign-on (SSO) method of access control.

- An OpenID is in the form of a unique URL and is authenticated by the entity hosting the OpenID URL.

19. Why cloud environment need SSL/TLS?

- SSL/TLS Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL), are cryptographically secure protocols designed to provide security and data integrity for communications over TCP/IP.
- TLS and SSL encrypt the segments of network connections at the transport layer.

20. What is mutual authentication?

- TLS also supports a more secure bilateral connection mode whereby both ends of the connection can be assured that they are communicating with whom they believe they are connected. This is known as mutual authentication.
- Mutual authentication requires the TLS client side to also maintain a certificate.

UNIT V CLOUD TECHNOLOGIES AND ADVANCEMENTS

Hadoop – MapReduce – Virtual Box -- Google App Engine – Programming Environment for Google App Engine – OpenStack – Federation in the Cloud – Four Levels of Federation – Federated Services and Applications – Future of Federation.

5.1 Hadoop

- Hadoop is an open source implementation of MapReduce coded and released in Java (rather than C) by Apache.

- The Hadoop implementation of MapReduce uses the Hadoop Distributed File System (HDFS) as its underlying layer rather than GFS.

- The Hadoop core is divided into two fundamental layers:

    - MapReduce engine
    - HDFS

- The MapReduce engine is the computation engine running on top of HDFS as its data storage manager.

- HDFS is a distributed file system inspired by GFS that organizes files and stores their data on a distributed computing system.

- HDFS Architecture: HDFS has a master/slave architecture containing a single NameNode as the master and a number of DataNodes as workers (slaves).

- To store a file in this architecture, HDFS splits the file into fixed-size blocks (e.g., 64 MB) and stores them on workers (DataNodes).

- The mapping of blocks to DataNodes is determined by the NameNode.

- The NameNode (master) also manages the file system's metadata and namespace.

- In such systems, the namespace is the area maintaining the metadata and metadata refers to all the information stored by a file system that is needed for overall management of all files.

- For example, NameNode in the metadata stores all information regarding the location of input splits/blocks in all DataNodes.

- Each DataNode, usually one per node in a cluster, manages the storage attached to the node. Each DataNode is responsible for storing and retrieving its file blocks.

- HDFS Features: Distributed file systems have special requirements, such as performance, scalability, concurrency control, fault tolerance and security requirements, to operate efficiently.

- However, because HDFS is not a general purpose file system, as it only executes specific types of applications, it does not need all the requirements of a general distributed file system.

- One of the main aspects of HDFS is its fault tolerance characteristic. Since Hadoop is designed to be deployed on low-cost hardware by default, a hardware failure in this system is considered to be common rather than an exception.

- Hadoop considers the following issues to fulfill reliability requirements of the file system

    - Block replication: To reliably store data in HDFS, file blocks are replicated in this system. The replication factor is set by the user and is three by default.

- Replica placement: The placement of replicas is another factor to fulfill the desired fault tolerance in HDFS.
- Heartbeat and Block report messages: Heartbeats and Block reports are periodic messages sent to the NameNode by each DataNode in a cluster.
- Applications run on HDFS typically have large data sets, individual files are broken into large blocks (e.g., 64 MB) to allow HDFS to decrease the amount of metadata storage required per file.
- This provides two advantages:

    - The list of blocks per file will shrink as the size of individual blocks increases.
    - Keeping large amounts of data sequentially within a block provides fast streaming reads of data.

- HDFS Operation: The control flow of HDFS operations such as write and read can properly highlight roles of the NameNode and DataNodes in the managing operations

    - To read a file in HDFS, a user sends an "open" request to the NameNode to get the location of file blocks.
    - For each file block, the NameNode returns the address of a set of DataNodes containing replica information for the requested file.
    - The number of addresses depends on the number of block replicas. Upon receiving such information, the user calls the read function to connect to the closest DataNode containing the first block of the file.
    - After the first block is streamed from the respective DataNode to the user, the established connection is terminated and the same process is repeated for all blocks of the requested file until the whole file is streamed to the user.
    - To write a file in HDFS, a user sends a "create" request to the NameNode to create a new file in the file system namespace.
    - If the file does not exist, the NameNode notifies the user and allows him to start writing data to the file by calling the write function.
    - The first block of the file is written to an internal queue termed the data queue while a data streamer monitors its writing into a DataNode.

- Since each file block needs to be replicated by a predefined factor, the data streamer first sends a request to the NameNode to get a list of suitable DataNodes to store replicas of the first block.
- The steamer then stores the block in the first allocated DataNode.
- Afterward, the block is forwarded to the second DataNode by the first DataNode.
- The process continues until all allocated DataNodes receive a replica of the first block from the previous DataNode.
- Once this replication process is finalized, the same process starts for the second block and continues until all blocks of the file are stored and replicated on the file system.

5.2 MapReduce

- The topmost layer of Hadoop is the MapReduce engine that manages the data flow and control flow of MapReduce jobs over distributed computing systems.

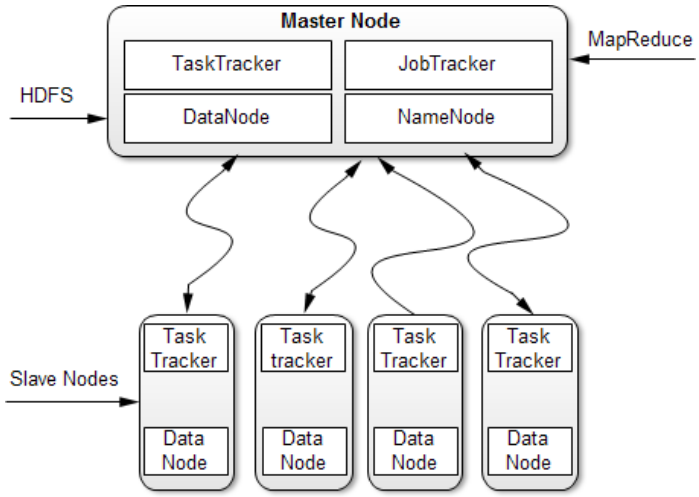


Figure 5.1 HDFS and MapReduce Architecture

- Figure 5.1 shows the MapReduce engine architecture cooperating with HDFS.

- Similar to HDFS, the MapReduce engine also has a master/slave architecture consisting of a single JobTracker as the master and a number of TaskTrackers as the slaves (workers).

- The JobTracker manages the MapReduce job over a cluster and is responsible for monitoring jobs and assigning tasks to TaskTrackers.

- The TaskTracker manages the execution of the map and/or reduce tasks on a single computation node in the cluster.

- Each TaskTracker node has a number of simultaneous execution slots, each executing either a map or a reduce task.

- Slots are defined as the number of simultaneous threads supported by CPUs of the TaskTracker node.

- For example, a TaskTracker node with N CPUs, each supporting M threads, has M * N simultaneous execution slots.

- It is worth noting that each data block is processed by one map task running on a single slot.

- Therefore, there is a one to one correspondence between map tasks in a TaskTracker and data blocks in the respective DataNode.

Running a Job in Hadoop

- Three components contribute in running a job in this system:

    - User node
    - JobTracker
    - TaskTrackers

- The data flow starts by calling the runJob (conf) function inside a user program running on the user node, in which conf is an object containing some tuning parameters for the MapReduce framework and HDFS.

- The runJob (conf) function and conf are comparable to the MapReduce (Spec, &Results) function and Spec in the first implementation of MapReduce by Google.

- Figure 5.2 depicts the data flow of running a MapReduce job in Hadoop.

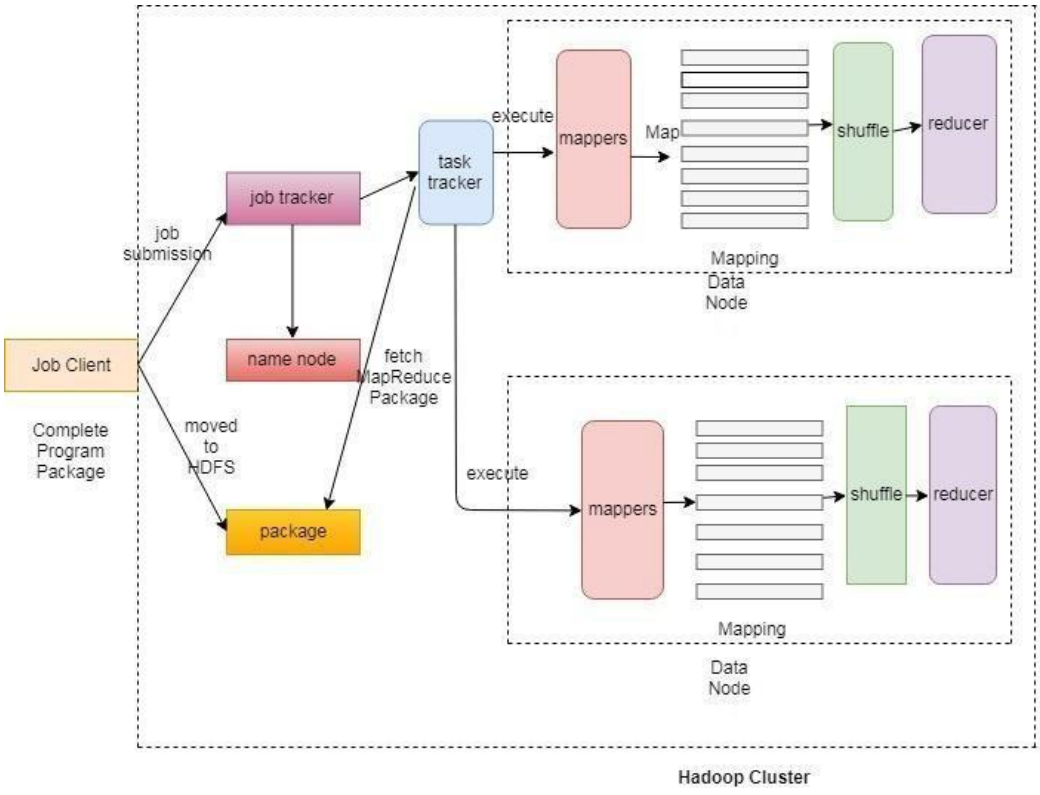


Figure 5.2 Data flow in Hadoop

- Job Submission Each job is submitted from a user node to the JobTracker node that might be situated in a different node within the cluster through the following procedure:

    - A user node asks for a new job ID from the JobTracker and computes input file splits.

- The user node copies some resources, such as the job's JAR file, configuration file, and computed input splits, to the JobTracker's file system.
- The user node submits the job to the JobTracker by calling the submitJob() function.
- Task assignment The JobTracker creates one map task for each computed input split by the user node and assigns the map tasks to the execution slots of the TaskTrackers.
    - The JobTracker considers the localization of the data when assigning the map tasks to the TaskTrackers.
    - The JobTracker also creates reduce tasks and assigns them to the TaskTrackers.
    - The number of reduce tasks is predetermined by the user, and there is no locality consideration in assigning them.

- Task execution The control flow to execute a task (either map or reduce) starts inside the TaskTracker by copying the job JAR file to its file system.
- Instructions inside the job JAR file are executed after launching a Java Virtual Machine (JVM) to run its map or reduce task.
- Task running check A task running check is performed by receiving periodic heartbeat messages to the JobTracker from the TaskTrackers.
- Each heartbeat notifies the JobTracker that the sending TaskTracker is alive, and whether the sending TaskTracker is ready to run a new task.

5.3 Virtual Box

- Oracle VM VirtualBox is a cross platform virtualization application.

- For one thing, it installs on the existing Intel or AMD-based computers, whether they are running Windows, Mac OS X, Linux, or Oracle Solaris operating systems (OSes).

- Secondly, it extends the capabilities of existing computer so that it can run multiple OSes, inside multiple virtual machines, at the same time.

- As an example, the end user can run Windows and Linux on your Mac, run Windows Server 2016 on your Linux server, run Linux on your Windows PC, and so on, all alongside the existing applications.

- The user can install and run as many virtual machines.

- The only practical limits are disk space and memory.

- Oracle VM VirtualBox is deceptively simple yet also very powerful.

- It can run everywhere from small embedded systems or desktop class machines all the way up to datacenter deployments and even Cloud environments.

- Virtual Box is created by Innotek and it was acquired by Sun Microsystems. In 2010, Virtual Box was acquired by Oracle.

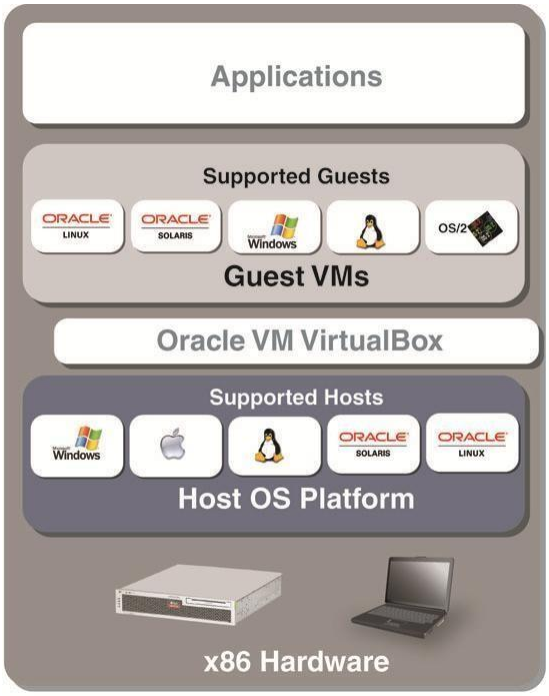


Figure 5.3 architecture of Virtual Box

- Virtual Box supported in Windows, macOS. Linux, Solaris and Open Solaris.

- Figure 5.3 depicts the architecture of Virtual Box

- The user can independently configure each VM and run it under a choice of software-based virtualization or hardware assisted virtualization if the underlying host hardware supports this.

- The host OS and guest OSs and applications can communicate with each other through a number of mechanisms including a common clipboard and a virtualized network facility.

- Guest VMs can also directly communicate with each other if configured to do so.

- The software based virtualization was dropped starting with VirtualBox 6.1. In earlier versions the absence of hardware assisted virtualization, VirtualBox adopts a standard software-based virtualization approach.

- This mode supports 32 bit guest OSs which run in rings 0 and 3 of the Intel ring architecture.

    - The system reconfigures the guest OS code, which would normally run in ring 0, to execute in ring 1 on the host hardware.
    - Because this code contains many privileged instructions which cannot run natively in ring 1, VirtualBox employs a Code Scanning and Analysis Manager (CSAM) to scan the ring 0 code recursively before its first execution to identify problematic instructions and then calls the Patch Manager (PATM) to perform in-situ patching.
    - This replaces the instruction with a jump to a VM-safe equivalent compiled code fragment in hypervisor memory.

    - The guest user mode code, running in ring 3, generally runs directly on the host hardware in ring 3.

- In both cases, VirtualBox uses CSAM and PATM to inspect and patch the offending instructions whenever a fault occurs.

- VirtualBox also contains a dynamic recompiler, based on QEMU to recompile any real mode or protected mode code entirely.

- Hardware assisted virtualization is starting with version 6.1, VirtualBox only supports.

- VirtualBox supports both Intel VT-X and AMD-V hardware assisted virtualization.

- Making use of these facilities, VirtualBox can run each guest VM in its own separate address-space.

- The guest OS ring 0 code runs on the host at ring 0 in VMX non-root mode rather than in ring 1.

- Until then, VirtualBox specifically supported some guests (including 64 bit guests, SMP guests and certain proprietary OSs) only on hosts with hardware-assisted virtualization

- The system emulates hard disks in one of three disk image formats:

    - VDI: This format is the VirtualBox-specific VirtualBox Disk Image and stores data in files bearing a ".vdi" .
    - VMDK: This open format is used by VMware products and stores data in one or more files bearing ".vmdk" filename extensions. A single virtual hard disk may span several files.
    - VHD: This format is used by Windows Virtual PC and Hyper-V and it is the native virtual disk format of the Microsoft Windows operating system. Data in this format are stored in a single file bearing the ".vhd" filename extension.

- A VirtualBox virtual machine can, therefore, use disks previously created in VMware or Microsoft Virtual PC, as well as its own native format.

- VirtualBox can also connect to iSCSI targets and to raw partitions on the host, using either as virtual hard disks.

- VirtualBox has supported Open Virtualization Format (OVF).

- By default, VirtualBox provides graphics support through a custom virtual graphics-card

- For an Ethernet network adapter, VirtualBox virtualizes these Network Interface Cards.

    - AMD PCnet PCI II
    - AMD PCnet-Fast III
    - Intel Pro/1000 MT Desktop
    - Intel Pro/1000 MT Server
    - Intel Pro/1000 T Server
    - Paravirtualized network adapter

- For a sound card, VirtualBox virtualizes Intel HD Audio.

- A USB controller is emulated so that any USB devices attached to the host can be seen in the guest.

- Oracle VM VirtualBox was designed to be modular and flexible.

- When the Oracle VM VirtualBox graphical user interface (GUI) is opened and a VM is started, at least the following three processes are running:

    - VBoxSVC is Oracle VM VirtualBox service process which always runs in the background. This process is started automatically by the first Oracle VM VirtualBox client process and exits a short time after the last client exits.

- The first Oracle VM VirtualBox service can be the GUI, VBoxManage, VBoxHeadless, the web service amongst others.

- The service is responsible for bookkeeping, maintaining the state of all VMs, and for providing communication between Oracle VM VirtualBox components.

- Oracle VM VirtualBox comes with comprehensive support for third-party developers.

- The Main API of Oracle VM VirtualBox exposes the entire feature set of the virtualization engine.

- The Main API is made available to C++ clients through COM on Windows hosts or XPCOM on other hosts. Bridges also exist for SOAP, Java and Python.

5.4 Google App Engine

- Google has the world's largest search engine facilities.

- The company has extensive experience in massive data processing that has led to new insights into data-center design and novel programming models that scale to incredible sizes.

- Google platform is based on its search engine expertise.

- Google has hundreds of data centers and has installed more than 460,000 servers worldwide.

- For example, 200 Google data centers are used at one time for a number of cloud applications.

- Data items are stored in text, images, and video and are replicated to tolerate faults or failures.

- Google's App Engine (GAE) which offers a PaaS platform supporting various cloud and web applications.

- Google has pioneered cloud development by leveraging the large number of data centers it operates.

- For example, Google pioneered cloud services in Gmail, Google Docs, and Google Earth, among other applications.

- These applications can support a large number of users simultaneously with HA.

- Notable technology achievements include the Google File System (GFS), MapReduce, BigTable, and Chubby.

- In 2008, Google announced the GAE web application platform which is becoming a common platform for many small cloud service providers.

- This platform specializes in supporting scalable (elastic) web applications.

- GAE enables users to run their applications on a large number of data centers associated with Google's search engine operations.

5.4.1 GAE Architecture

- Figure 5.4 shows the major building blocks of the Google cloud platform which has been used to deliver the cloud services highlighted earlier.

- GFS is used for storing large amounts of data.

- MapReduce is for use in application program development.

- Chubby is used for distributed application lock services.
- BigTable offers a storage service for accessing structured data.
- Users can interact with Google applications via the web interface provided by each application.
- Third-party application providers can use GAE to build cloud applications for providing services.
- The applications all run in data centers under tight management by Google engineers. Inside each data center, there are thousands of servers forming different clusters

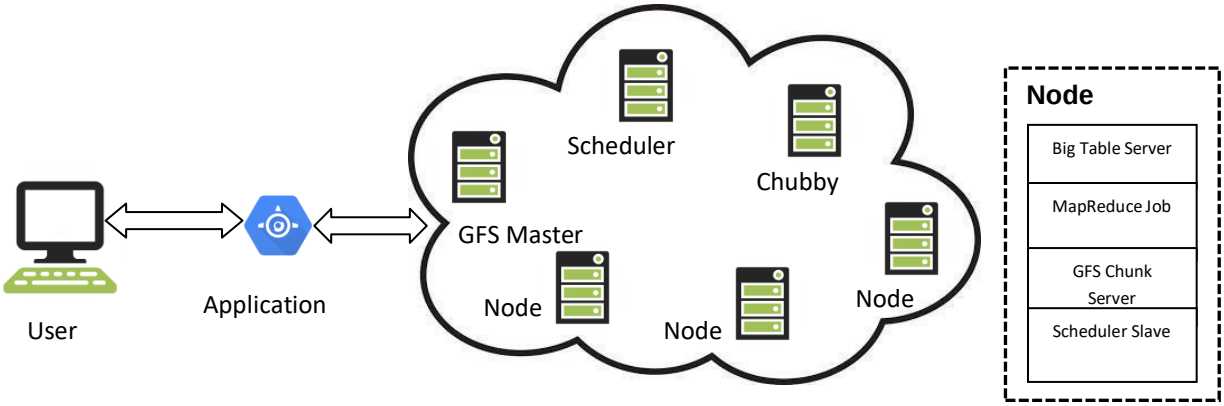


Figure 5.4 Google cloud platform

- Google is one of the larger cloud application providers, although it is fundamental service program is private and outside people cannot use the Google infrastructure to build their own service.
- The building blocks of Google's cloud computing application include the Google File System for storing large amounts of data, the MapReduce programming framework for application developers, Chubby for distributed application lock services, and BigTable as a storage service for accessing structural or semistructural data.

- With these building blocks, Google has built many cloud applications.

- Figure 5.4 shows the overall architecture of the Google cloud infrastructure.

- A typical cluster configuration can run the Google File System, MapReduce jobs and BigTable servers for structure data.

- Extra services such as Chubby for distributed locks can also run in the clusters.

- GAE runs the user program on Google's infrastructure. As it is a platform running third-party programs, application developers now do not need to worry about the maintenance of servers.

- GAE can be thought of as the combination of several software components.

- The frontend is an application framework which is similar to other web application frameworks such as ASP, J2EE and JSP.

- At the time of this writing, GAE supports Python and Java programming environments. The applications can run similar to web application containers.

- The frontend can be used as the dynamic web serving infrastructure which can provide the full support of common technologies.

5.4.2 Functional Modules of GAE

- The GAE platform comprises the following five major components.

- The GAE is not an infrastructure platform, but rather an application development platform for users.

    - The datastore offers object-oriented, distributed, structured data storage services based on BigTable techniques. The datastore secures data management operations.
    - The application runtime environment offers a platform for scalable web programming and execution. It supports two development languages: Python and Java.
    - The software development kit (SDK) is used for local application development. The SDK allows users to execute test runs of local applications and upload application code.
    - The administration console is used for easy management of user application development cycles, instead of for physical resource management.
    - The GAE web service infrastructure provides special interfaces to guarantee flexible use and management of storage and network resources by GAE.

- Google offers essentially free GAE services to all Gmail account owners.

- The user can register for a GAE account or use your Gmail account name to sign up for the service.

- The service is free within a quota.

- If the user exceeds the quota, the page instructs how to pay for the service. Then the user can download the SDK and read the Python or Java guide to get started.

- Note that GAE only accepts Python, Ruby and Java programming languages.

- The platform does not provide any IaaS services, unlike Amazon, which offers IaaS and PaaS.

- This model allows the user to deploy user-built applications on top of the cloud infrastructure that are built using the programming languages and software tools supported by the provider (e.g., Java, Python).

- Azure does this similarly for .NET. The user does not manage the underlying cloud infrastructure.

- The cloud provider facilitates support of application development, testing, and operation support on a well-defined service platform.

5.4.3 GAE Applications

- Best-known GAE applications include the Google Search Engine, Google Docs, Google Earth and Gmail.

- These applications can support large numbers of users simultaneously.

- Users can interact with Google applications via the web interface provided by each application.

- Third party application providers can use GAE to build cloud applications for providing services.

- The applications are all run in the Google data centers.

- Inside each data center, there might be thousands of server nodes to form different clusters.

- Each cluster can run multipurpose servers.

- GAE supports many web applications.

- One is a storage service to store application specific data in the Google infrastructure.

- The data can be persistently stored in the backend storage server while still providing the facility for queries, sorting and even transactions similar to traditional database systems.

- GAE also provides Google specific services, such as the Gmail account service. This can eliminate the tedious work of building customized user management components in web applications.

5.5 Programming Environment for Google App Engine

- Several web resources (e.g., http://code.google.com/appengine/) and specific books and articles discuss how to program GAE.

- Figure 5.5 summarizes some key features of GAE programming model for two supported languages: Java and Python.

- A client environment that includes an Eclipse plug-in for Java allows you to debug your GAE on your local machine.

- Also, the GWT Google Web Toolkit is available for Java web application developers. Developers can use this, or any other language using a JVM based interpreter or compiler, such as JavaScript or Ruby.

- Python is often used with frameworks such as Django and CherryPy, but Google also supplies a built in webapp Python environment.

- There are several powerful constructs for storing and accessing data.

- The data store is a NOSQL data management system for entities that can be, at most, 1 MB in size and are labeled by a set of schema-less properties.

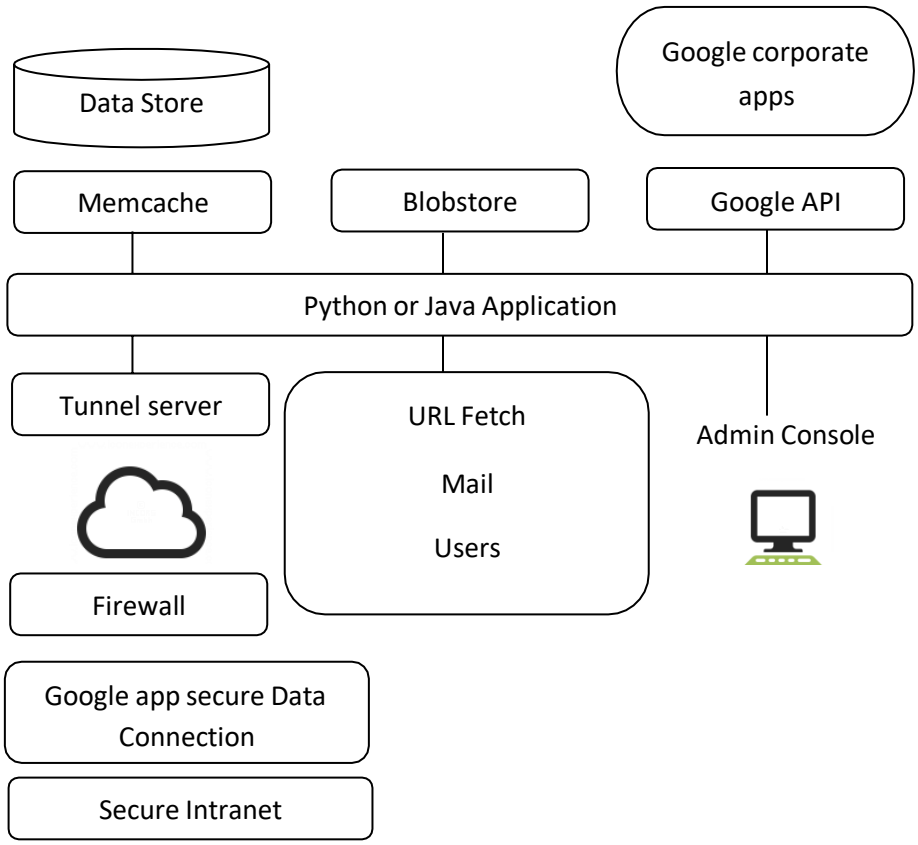


Figure 5.5 Programming Environment of Google AppEngine

- Queries can retrieve entities of a given kind filtered and sorted by the values of the properties.

- Java offers Java Data Object (JDO) and Java Persistence API (JPA) interfaces implemented by the open source Data Nucleus Access platform, while Python has a SQL-like query language called GQL.

- The data store is strongly consistent and uses optimistic concurrency control.

- An update of an entity occurs in a transaction that is retried a fixed number of times if other processes are trying to update the same entity simultaneously.

- The user application can execute multiple data store operations in a single transaction which either all succeed or all fail together.

- The data store implements transactions across its distributed network using entity groups.

- A transaction manipulates entities within a single group.

- Entities of the same group are stored together for efficient execution of transactions.

- The user GAE application can assign entities to groups when the entities are created.

- The performance of the data store can be enhanced by in-memory caching using the memcache, which can also be used independently of the data store.

- Recently, Google added the blobstore which is suitable for large files as its size limit is 2 GB.

- There are several mechanisms for incorporating external resources.

- The Google SDC Secure Data Connection can tunnel through the Internet and link your intranet to an external GAE application.

- The URL Fetch operation provides the ability for applications to fetch resources and communicate with other hosts over the Internet using HTTP and HTTPS requests.

- There is a specialized mail mechanism to send e-mail from your GAE application.

- Applications can access resources on the Internet, such as web services or other data, using GAE's URL fetch service.

- The URL fetch service retrieves web resources using the same high-speed Google infrastructure that retrieves web pages for many other Google products.

- There are dozens of Google "corporate" facilities including maps, sites, groups, calendar, docs, and YouTube, among others.

- These support the Google Data API which can be used inside GAE.

- An application can use Google Accounts for user authentication. Google Accounts handles user account creation and sign-in, and a user that already has a Google account (such as a Gmail account) can use that account with your app.

- GAE provides the ability to manipulate image data using a dedicated Images service which can resize, rotate, flip, crop and enhance images. An application can perform tasks outside of responding to web requests.

- A GAE application is configured to consume resources up to certain limits or quotas. With quotas, GAE ensures that your application would not exceed your budget and that other applications running on GAE would not impact the performance of your app.

- In particular, GAE use is free up to certain quotas.

- GFS was built primarily as the fundamental storage service for Google's search engine.

- As the size of the web data that was crawled and saved was quite substantial, Google needed a distributed file system to redundantly store massive amounts of data on cheap and unreliable computers.

- In addition, GFS was designed for Google applications and Google applications were built for GFS.

- In traditional file system design, such a philosophy is not attractive, as there should be a clear interface between applications and the file system such as a POSIX interface.

- GFS typically will hold a large number of huge files, each 100 MB or larger, with files that are multiple GB in size quite common. Thus, Google has chosen its file data block size to be 64 MB instead of the 4 KB in typical traditional file systems.

- The I/O pattern in the Google application is also special.

- Files are typically written once, and the write operations are often the appending data blocks to the end of files.

- Multiple appending operations might be concurrent.

- BigTable was designed to provide a service for storing and retrieving structured and semi structured data.

- BigTable applications include storage of web pages, per-user data, and geographic locations.

- The scale of such data is incredibly large. There will be billions of URLs, and each URL can have many versions, with an average page size of about 20 KB per version.

- The user scale is also huge.

- There are hundreds of millions of users and there will be thousands of queries per second.

- The same scale occurs in the geographic data, which might consume more than 100 TB of disk space.

- It is not possible to solve such a large scale of structured or semi structured data using a commercial database system.

- This is one reason to rebuild the data management system and the resultant system can be applied across many projects for a low incremental cost.

- The other motivation for rebuilding the data management system is performance.

- Low level storage optimizations help increase performance significantly which is much harder to do when running on top of a traditional database layer.

- The design and implementation of the BigTable system has the following goals.

    - The applications want asynchronous processes to be continuously updating different pieces of data and want access to the most current data at all times.
    - The database needs to support very high read/write rates and the scale might be millions of operations per second.
    - The application may need to examine data changes over time.

- Thus, BigTable can be viewed as a distributed multilevel map. It provides a fault tolerant and persistent database as in a storage service.

- The BigTable system is scalable, which means the system has thousands of servers, terabytes of in-memory data, peta bytes of disk based data, millions of reads/writes per second and efficient scans.

- BigTable is a self managing system (i.e., servers can be added/removed dynamically and it features automatic load balancing).

- Chubby, Google's Distributed Lock Service Chubby is intended to provide a coarse-grained locking service.

- It can store small files inside Chubby storage which provides a simple namespace as a file system tree.

- The files stored in Chubby are quite small compared to the huge files in GFS.

5.6 OpenStack

- The OpenStack project is an open source cloud computing platform for all types of clouds, which aims to be simple to implement, massively scalable and feature rich.

- Developers and cloud computing technologists from around the world create the OpenStack project.

- OpenStack provides an Infrastructure as a Service (IaaS) solution through a set of interrelated services.

- Each service offers an application programming interface (API) that facilitates this integration.

- Depending on their needs, administrator can install some or all services.

- OpenStack began in 2010 as a joint project of Rackspace Hosting and NASA.

- As of 2012, it is managed by the OpenStack Foundation, a non-profit corporate entity established in September 2013 to promote OpenStack software and its community.

- Now, More than 500 companies have joined the project

- The OpenStack system consists of several key services that are separately installed.

- These services work together depending on your cloud needs and include the Compute, Identity, Networking, Image, Block Storage, Object Storage, Telemetry, Orchestration, and Database services.

- The administrator can install any of these projects separately and configure them standalone or as connected entities.

- Figure 5.6 shows the relationships among the OpenStack services:

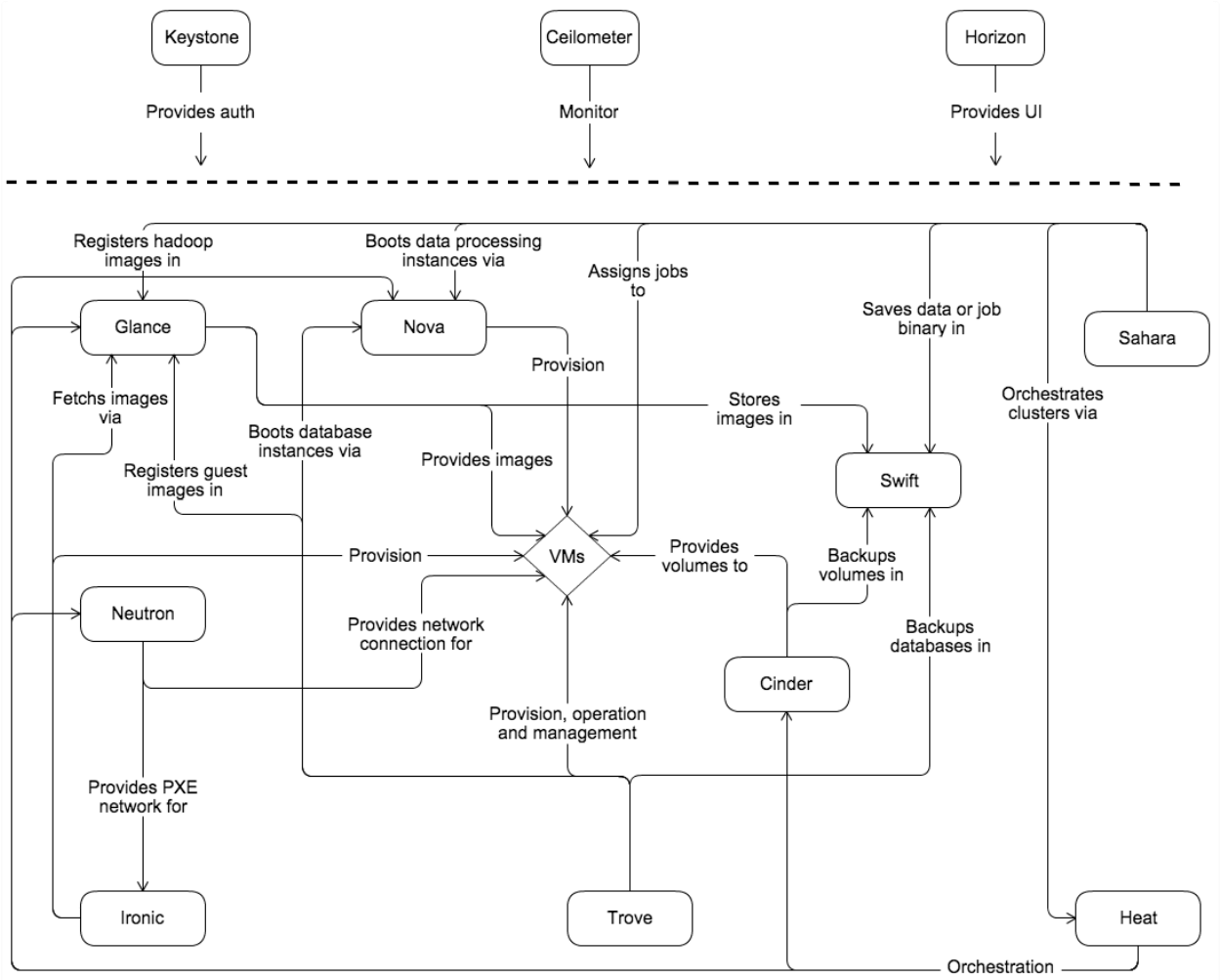


Figure 5.6 Relationship between OpenStack services

- To design, deploy, and configure OpenStack, administrators must understand the logical architecture.

- OpenStack consists of several independent parts, named the OpenStack services. All services authenticate through a common Identity service.

- Individual services interact with each other through public APIs, except where privileged administrator commands are necessary.

- Internally, OpenStack services are composed of several processes.

- All services have at least one API process, which listens for API requests, preprocesses them and passes them on to other parts of the service.

- With the exception of the Identity service, the actual work is done by distinct processes.

- For communication between the processes of one service, an AMQP message broker is used.

- The service's state is stored in a database.

- When deploying and configuring the OpenStack cloud, administrator can choose among several message broker and database solutions, such as RabbitMQ, MySQL, MariaDB, and SQLite.

- Users can access OpenStack via the web-based user interface implemented by the Horizon Dashboard, via command-line clients and by issuing API requests through tools like browser plug-ins or curl.

- For applications, several SDKs are available. Ultimately, all these access methods issue REST API calls to the various OpenStack services.

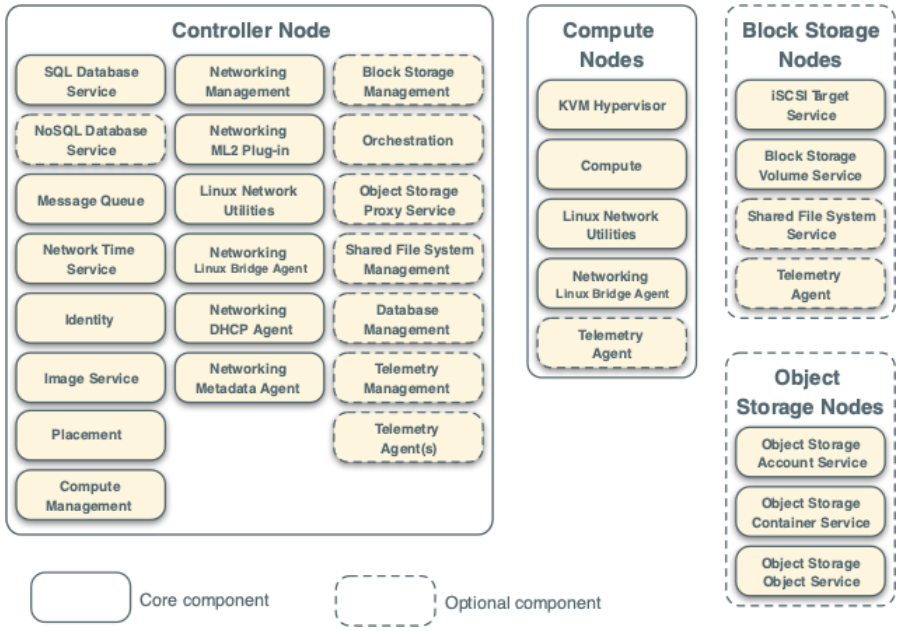


Figure 5.7 Example OpenStack architecture

- The controller node runs the Identity service, Image service, Placement service, management portions of Compute, management portion of Networking, various Networking agents, and the Dashboard.

- It also includes supporting services such as an SQL database, message queue, and NTP.

    - Optionally, the controller node runs portions of the Block Storage, Object Storage, Orchestration, and Telemetry services.

- The controller node requires a minimum of two network interfaces.

- The compute node runs the hypervisor portion of Compute that operates instances. By default, Compute uses the KVM hypervisor.

- The compute node also runs a Networking service agent that connects instances to virtual networks and provides firewalling services to instances via security groups.

- Administrator can deploy more than one compute node. Each node requires a minimum of two network interfaces.

- The optional Block Storage node contains the disks that the Block Storage and Shared File System services provision for instances.

- For simplicity, service traffic between compute nodes and this node uses the management network.

- Production environments should implement a separate storage network to increase performance and security.

- Administrator can deploy more than one block storage node. Each node requires a minimum of one network interface.

- The optional Object Storage node contains the disks that the Object Storage service uses for storing accounts, containers, and objects.

- For simplicity, service traffic between compute nodes and this node uses the management network.

- Production environments should implement a separate storage network to increase performance and security.

- This service requires two nodes. Each node requires a minimum of one network interface. Administrator can deploy more than two object storage nodes.

- The provider networks option deploys the OpenStack Networking service in the simplest way possible with primarily layer 2 (bridging/switching) services and VLAN segmentation of networks.

- Essentially, it bridges virtual networks to physical networks and relies on physical network infrastructure for layer-3 (routing) services.

- Additionally, a DHCP service provides IP address information to instances.

5.7 Federation in the Cloud

- One challenge in creating and managing a globally decentralized cloud computing environment is maintaining consistent connectivity between untrusted components while remaining fault tolerant.

- A key opportunity for the emerging cloud industry will be in defining a federated cloud ecosystem by connecting multiple cloud computing providers using a common standard.

- A notable research project being conducted by Microsoft called the Geneva Framework. This framework focuses on issues involved in cloud federation.

- Geneva has been described as claims based access platform and is said to help simplify access to applications and other systems.

- The concept allows for multiple providers to interact seamlessly with others and it enables developers to incorporate various authentication models that will work with any corporate identity system, including Active Directory,

- LDAPv3 based directories, application specific databases, and new user centric identity models such as LiveID, OpenID, and InfoCard systems.

- It also supports Microsoft's CardSpace and Novell's Digital Me.

- Federation in cloud is implemented by the use of Internet Engineering Task Force (IETF) standard Extensible Messaging and Presence Protocol (XMPP) and inter domain federation using the Jabber Extensible Communications Platform (Jabber XCP).

- Because this protocol is currently used by a wide range of existing services offered by providers as diverse as Google Talk, Live Journal, Earthlink, Facebook, ooVoo, Meebo, Twitter, the U.S. Marines Corps, the Defense Information Systems Agency (DISA), the U.S. Joint Forces Command (USJFCOM), and the National Weather Service.

- Session Initiation Protocol (SIP), which is the foundation of popular enterprise messaging systems such as IBM's Lotus Sametime and Microsoft's Live Communications Server (LCS) and Office Communications Server (OCS).

- Jabber XCP is a highly scalable, extensible, available, and device-agnostic presence solution built on XMPP and supports multiple protocols such as Session Initiation Protocol for Instant Messaging and Presence Leveraging Extensions (SIMPLE) and Instant Messaging and Presence Service (IMPS).

- Jabber XCP is a highly programmable platform, which makes it ideal for adding presence and messaging to existing applications or services and for building next-generation, presence based solutions.

- Over the last few years there has been a controversy brewing in web services architectures.

- Cloud services are being talked up as a fundamental shift in web architecture that promises to move us from interconnected silos to a collaborative network of services whose sum is greater than its parts.

- The problem is that the protocols powering current cloud services, SOAP (Simple Object Access Protocol) and a few other assorted HTTP based protocols, are all one-way information exchanges.

- Therefore cloud services are not real time, would not scale, and often cannot clear the firewall.

- Many believe that those barriers can be overcome by XMPP (also called Jabber) as the protocol that will fuel the Software as a Service (SaaS) models of tomorrow.

- Google, Apple, AOL, IBM, Live journal and Jive have all incorporated this protocol into their cloud based solutions in the last few years.

- Since the beginning of the Internet era, if the user wanted to synchronize services between two servers, the most common solution was to have the client "ping" the host at regular intervals, which is known as polling.

- Polling is how most of us check our email.

- XMPP's profile has been steadily gaining since its inception as the protocol behind the open source instant messenger (IM) server jabberd in 1998.

- XMPP's advantages include:

    - It is decentralized, meaning anyone may set up an XMPP server.
    - It is based on open standards.
    - It is mature multiple implementations of clients and servers exist.

- Robust security is supported via Simple Authentication and Security Layer (SASL) and Transport Layer Security (TLS).

- It is flexible and designed to be extended.

- XMPP is a good fit for cloud computing because it allows for easy two way communication

- XMPP eliminates the need for polling and focus on rich publish subscribe functionality

- It in it is XML-based and easily extensible, perfect for both new IM features and custom cloud services

- It is efficient and has been proven to scale to millions of concurrent users on a single service (such as Google's GTalk). And also it has a built-in worldwide federation model.

- Of course, XMPP is not the only pub-sub enabler getting a lot of interest from web application developers.

- An Amazon EC2-backed server can run Jetty and Cometd from Dojo.

- Unlike XMPP, Comet is based on HTTP and in conjunction with the Bayeux Protocol, uses JSON to exchange data.

- Given the current market penetration and extensive use of XMPP and XCP for federation in the cloud and that it is the dominant open protocol in that space.

- The ability to exchange data used for presence, messages, voice, video, files, notifications, etc., with people, devices and applications gain more power when they can be shared across organizations and with other service providers.

- Federation differs from peering, which requires a prior agreement between parties before a server-to-server (S2S) link can be established.

- In the past, peering was more common among traditional telecommunications providers (because of the high cost of transferring voice traffic).

- In the brave new Internet world, federation has become a de facto standard for most email systems because they are federated dynamically through Domain Name System (DNS) settings and server configurations.

5.8 Four Levels of Federation

- Federation is the ability for two XMPP servers in different domains to exchange XML stanzas.

- According to the XEP-0238: XMPP Protocol Flows for Inter-Domain Federation, there are at least four basic types of federation:

- Permissive federation

    - Permissive federation occurs when a server accepts a connection from a peer network server without verifying its identity using DNS lookups or certificate checking.
    - The lack of verification or authentication may lead to domain spoofing (the unauthorized use of a third-party domain name in an email message in order to pretend to be someone else), which opens the door to widespread spam and other abuses. With the release of the open source jabberd 1.2 server in October 2000, which included support for the Server Dialback protocol (fully supported in Jabber XCP), permissive federation met its demise on the XMPP network.

- Verified federation

    - This type of federation occurs when a server accepts a connection from a peer after the identity of the peer has been verified.
    - It uses information obtained via DNS and by means of domain-specific keys exchanged beforehand.
    - The connection is not encrypted, and the use of identity verification effectively prevents domain spoofing.
    - To make this work, federation requires proper DNS setup and that is still subject to DNS poisoning attacks.

- Verified federation has been the default service policy on the open XMPP since the release of the open-source jabberd 1.2 server.

- Encrypted federation

  - In this mode, a server accepts a connection from a peer if and only if the peer supports Transport Layer Security (TLS) as defined for XMPP in Request for Comments (RFC) 3920.
  - The peer must present a digital certificate.
  - The certificate may be self signed, but this prevents using mutual authentication.
  - If this is the case, both parties proceed to weakly verify identity using Server Dialback.
  - XEP-0220 defines the Server Dialback protocol, which is used between XMPP servers to provide identity verification.
  - Server Dialback uses the DNS as the basis for verifying identity
  - The basic approach is that when a receiving server receives a server-to-server connection request from an originating server, it does not accept the request until it has verified a key with an authoritative server for the domain asserted by the originating server.
  - Although Server Dialback does not provide strong authentication or trusted federation, and although it is subject to DNS poisoning attacks, it has effectively prevented most instances of address spoofing on the XMPP network since its release in 2000.
  - This results in an encrypted connection with weak identity verification.

- Trusted federation

  - In this federation, a server accepts a connection from a peer only under the stipulation that the peer supports TLS and the peer can present a digital certificate issued by a root certification authority (CA) that is trusted by the authenticating server.
  - The list of trusted root CAs may be determined by one or more factors, such as the operating system, XMPP server software or local service policy.

- In trusted federation, the use of digital certificates results not only in a channel encryption but also in strong authentication.
- The use of trusted domain certificates effectively prevents DNS poisoning attacks but makes federation more difficult, since such certificates have traditionally not been easy to obtain.

5.9 Federated Services and Applications

- S2S federation is a good start toward building a real-time communications cloud.
- Clouds typically consist of all the users, devices, services, and applications connected to the network.
- In order to fully leverage the capabilities of this cloud structure, a participant needs the ability to find other entities of interest.
- Such entities might be end users, multiuser chat rooms, real-time content feeds, user directories, data relays, messaging gateways, etc.
- Finding these entities is a process called discovery.
- XMPP uses service discovery (as defined in XEP-0030) to find the aforementioned entities.
- The discovery protocol enables any network participant to query another entity regarding its identity, capabilities and associated entities.
- When a participant connects to the network, it queries the authoritative server for its particular domain about the entities associated with that authoritative server.

- In response to a service discovery query, the authoritative server informs the inquirer about services hosted there and may also detail services that are available but hosted elsewhere.

- XMPP includes a method for maintaining personal lists of other entities, known as roster technology, which enables end users to keep track of various types of entities.

- Usually, these lists are comprised of other entities the users are interested in or interact with regularly.

- Most XMPP deployments include custom directories so that internal users of those services can easily find what they are looking for.

5.10 Future of Federation

- The implementation of federated communications is a precursor to building a seamless cloud that can interact with people, devices, information feeds, documents, application interfaces and other entities.

- The power of a federated, presence enabled communications infrastructure is that it enables software developers and service providers to build and deploy such applications without asking permission from a large, centralized communications operator.

- The process of server-to-server federation for the purpose of inter domain communication has played a large role in the success of XMPP, which relies on a small set of simple but powerful mechanisms for domain checking and security to generate verified, encrypted, and trusted connections between any two deployed servers.

- These mechanisms have provided a stable, secure foundation for growth of the XMPP network and similar real time technologies.

TWO MARK QUESTIONS

1. What is Hadoop?

   - Hadoop is an open source implementation of MapReduce coded and released in Java (rather than C) by Apache.
   - The Hadoop implementation of MapReduce uses the Hadoop Distributed File System (HDFS) as its underlying layer rather than GFS.

2. List the fundamental layers of Hadoop core.

   - The Hadoop core is divided into two fundamental layers:
     - MapReduce engine
     - HDFS

3. Describe about HDFS.

   - HDFS is a Hadoop distributed file system inspired by GFS that organizes files and stores their data on a distributed computing system.
   - HDFS has a master/slave architecture containing a single NameNode as the master and a number of DataNodes as workers (slaves).
   - To store a file in this architecture, HDFS splits the file into fixed-size blocks (e.g., 64 MB) and stores them on workers (DataNodes).
   - The mapping of blocks to DataNodes is determined by the NameNode.

4. Is HDFS provides fault tolerant?

   - One of the main aspects of HDFS is its fault tolerance characteristic. Since Hadoop is designed to be deployed on low-cost hardware by default, a hardware failure in this system is considered to be common rather than an exception.

5. List the issues to fulfill reliability requirements of the file system by hadoop.

- Block replication
- Replica placement
- Heartbeat and Block report messages

6. What is the purpose of heartbeat messages?

   - Heartbeat is a periodic message sent to the NameNode by each DataNode in a cluster.

7. List the advantages of HDFS.

   - The list of blocks per file will shrink as the size of individual blocks increases, and by keeping large amounts of data sequentially within a block, HDFS provides fast streaming reads of data.

8. Define MapReduce.

   - The topmost layer of Hadoop is the MapReduce engine that manages the data flow and control flow of MapReduce jobs over distributed computing systems.
   - Similar to HDFS, the MapReduce engine also has a master/slave architecture consisting of a single JobTracker as the master and a number of TaskTrackers as the slaves (workers).
   - The JobTracker manages the MapReduce job over a cluster and is responsible for monitoring jobs and assigning tasks to TaskTrackers.
   - The TaskTracker manages the execution of the map and/or reduce tasks on a single computation node in the cluster.

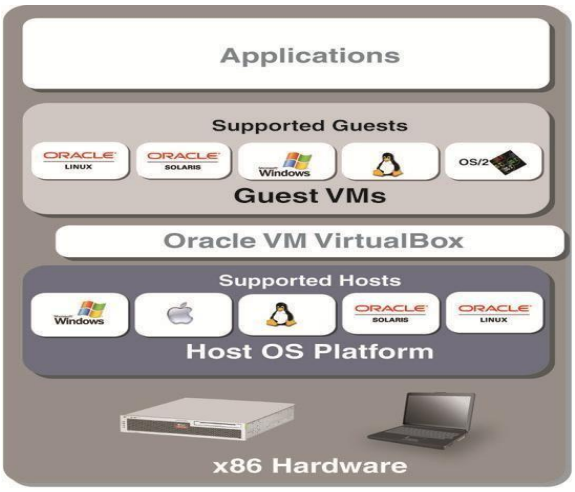
9. List the components contribute in running a job in Hadoop system.

   - a user node
   - a JobTracker
   - TaskTrackers

10. What is the use of VirtualBox?

- Oracle VM VirtualBox is a cross-platform virtualization application.
- For one thing, it installs on the existing Intel or AMD-based computers, whether they are running Windows, Mac OS X, Linux, or Oracle Solaris operating systems (OSes).
- Secondly, it extends the capabilities of existing computer so that it can run multiple OSes, inside multiple virtual machines, at the same time.

11. Illustrate the architecture of VirtualBox.



12. List the three disk image formats used in VirtualBox:

- VDI: This format is the VirtualBox-specific VirtualBox Disk Image and stores data in files bearing a ".vdi".
- VMDK: This open format is used by VMware products and stores data in one or more files bearing ".vmdk" filename extensions.
- VHD: This format is used by Windows Virtual PC and Hyper-V, and is the native virtual disk format of the Microsoft Windows operating system.

13. Describe about GAE.

- Google's App Engine (GAE) which offers a PaaS platform supporting various cloud and web applications.

- This platform specializes in supporting scalable (elastic) web applications.
- GAE enables users to run their applications on a large number of data centers associated with Google's search engine operations.

14. Mention the components maintained in a node of Google cloud platform.

- GFS is used for storing large amounts of data.
- MapReduce is for use in application program development.
- Chubby is used for distributed application lock services.
- BigTable offers a storage service for accessing structured data.

15. List the functional modules of GAE.

- Datastore
- Application runtime environment
- Software development kit (SDK)
- Administration console
- GAE web service infrastructure

16. List the applications of GAE.

- Well-known GAE applications include the Google Search Engine, Google Docs, Google Earth, and Gmail.
- These applications can support large numbers of users simultaneously.
- Users can interact with Google applications via the web interface provided by each application.
- Third-party application providers can use GAE to build cloud applications for providing services.

17. Mention the goals for design and implementation of the BigTable system.

- The applications want asynchronous processes to be continuously updating different pieces of data and want access to the most current data at all times.
- The database needs to support very high read/write rates and the scale might be millions of operations per second.
- The application may need to examine data changes over time.

18. Describe about Openstack.

- The OpenStack project is an open source cloud computing platform for all types of clouds, which aims to be simple to implement, massively scalable, and feature rich.
- Developers and cloud computing technologists from around the world create the OpenStack project.
- OpenStack provides an Infrastructure-as-a-Service (IaaS) solution through a set of interrelated services.

19. List the key services of OpenStack.

- The OpenStack system consists of several key services that are separately installed.
- Compute, Identity, Networking, Image, Block Storage, Object Storage, Telemetry, Orchestration and Database services.

20. What is the need of federated cloud ecosystem?

- One challenge in creating and managing a globally decentralized cloud computing environment is maintaining consistent connectivity between untrusted components while remaining fault-tolerant.
- A key opportunity for the emerging cloud industry will be in defining a federated cloud ecosystem by connecting multiple cloud computing providers using a common standard.
- A notable research project being conducted by Microsoft, called the Geneva Framework, focuses on issues involved in cloud federation.

21. List the advantages of Extensible Messaging and Presence Protocol.

- XMPP's is decentralized, meaning anyone may set up an XMPP server. It is based on open standards. It is mature multiple implementations of clients and servers exist.

22. List the levels of Federation.

- Permissive federation
- Verified federation
- Encrypted federation
- Trusted federation

23. What is S2S federation?

- S2S federation is a good start toward building a real-time communications cloud. Clouds typically consist of all the users, devices, services, and applications connected to the network.

24. What is the future of federation?

- The power of a federated, presence enabled communications infrastructure is that it enables software developers and service providers to build and deploy such applications without asking permission from a large, centralized communications operator.
- These mechanisms have provided a stable, secure foundation for growth of the XMPP network and similar real time technologies.

MODEL QUESTION PAPER - I

B.E./B.Tech. DEGREE EXAMINATION

Seventh Semester

Computer Science and Engineering

CS8791 – Cloud Computing

(Regulation 2017)

Time: Three hours Maximum: 100 marks

Answer ALL questions

PART A – (10 X 2 = 20 marks)

1. Define Cloud.
2. List the components of cloud model.
3. Mention the four characteristics to identify the service.
4. Differentiate between Full virtualization and Paravirtualization.
5. What are advantages of cloud storage?
6. What is Hardware as a Service?
7. What is the purpose of runtime support service named cluster monitoring?
8. Compare over provisioning and under provisioning?
9. Illustrate the architecture of VirtualBox.
10. List the merits of XMPP.

PART B – (5 X 16 = 80 marks)

11.(a) Explain about evolution of cloud computing.

Or

(b) (i) Explain about the elements of parallel and distributed computing. (8)

(ii) Explain about elasticity nature of cloud computing and on-demand provisioning. (8)

12.(a) (i) Explain about Service Oriented Architecture. (8)

(ii) Explain about Publish-Subscribe model. (8)

Or

(b) Explain about various implementation levels of virtualization.

13.(a) (i) Explain about layered architectural design of cloud computing. (8)

(ii) Explain about cloud deployment models. (8)

Or

(b) Explain about major architectural design challenges in cloud. (16)

14.(a) (i) Explain about inter cloud resource management with neat diagram. (8)

(ii) Explain about resource provisioning methods. (8)

Or

(b) (i) Explain about Identity Access Management. (8)

(ii) Explain about Virtual Machine Security. (8)

15.(a) Explain about HDFS and MapReduce in Hadoop framework. (16)

Or

(b) (i) Explain about Programming environment for Google AppEngine (8)

(ii) Explain about the levels of federation. (8)

MODEL QUESTION PAPER - II

B.E./B.Tech. DEGREE EXAMINATION

Seventh Semester

Computer Science and Engineering

CS8791 – Cloud Computing

(Regulation 2017)

Time: Three hours Maximum: 100 marks

Answer ALL questions

PART A – (10 X 2 = 20 marks)

1. Differentiate between Parallel and Distributed computing.
2. List the various models for message based communication.
3. Define service oriented architecture.
4. Illustrate ring based security with neat diagram.
5. Compare Public cloud and Private cloud.
6. What are the design requirements considers by Amazon to build S3?
7. What is Event-driven provisioning?
8. Mention the purpose of Security Governance.
9. What is purpose of Task tracker and Job tracker in Hadoop?
10. What is the need for federated cloud ecosystem?

PART B – (5 X 16 = 80 marks)

11.(a) Explain about the principles of Parallel and Distributed Computing. (16)

Or

(b) Explain about characteristics of cloud computing. (16)

12.(a) (i) Explain about RESTful Systems. (8)

(ii) Explain about Web service technologies stack. (8)

Or

(b) (i) Explain about CPU, Memory and I/O device virtualization . (8)

(ii) Explain about virtualization support and disaster recovery strategies.(8)

13.(a) Explain about NIST reference architecture with neat diagram. (16)

Or

(b) (i) Explain about cloud service model. (8)

(ii) Explain about Storage-as-a-Service. (8)

14.(a) (i) Explain about global exchange of cloud resources (8).

(ii) Explain about runtime support services in inter cloud management. (8)

Or

(b) Explain about cloud security and its challenges. Elaborate some standards specific to cloud security. (16)

15.(a) Explain about functional modules and programming environment of Google App Engine. (16)

Or

(b) Explain about OpenStack architecture with neat diagram. (16)

"Books are my favorite friends and I Consider my home library, with many thousand books, to be my greatest wealth. Every new book, based on some new idea inspires me and gives me a new thought to ponder."

— Dr. A.P.J. Abdul Kalam

MADHA
Expertise | Empathy | Excellence
ENGINEERING COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COMMON FOR: DEPARTMENT OF INFORMATION TECHNOLOGY

MG8591 – PRINCIPLES OF MANAGEMENT

R – 2017

LECTURE NOTES

PRINCIPLES OF MANAGEMENT (MG8591)

UNIT I

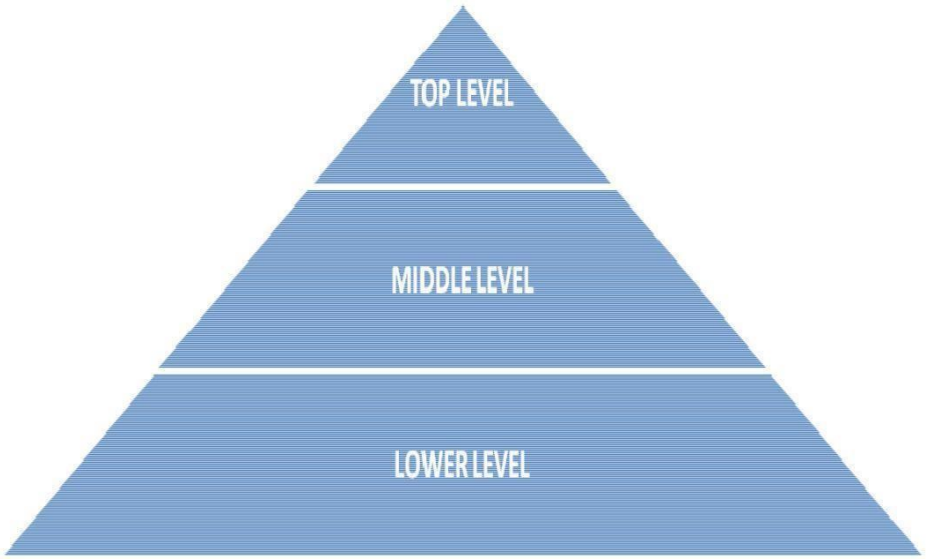
OVERVIEW OF MANAGEMENT

DEFINITION

According to Harold Koontz, "Management is an art of getting things done through and with the people in formally organized groups. It is an art of creating an environment in which people can perform and individuals and can co-operate towards attainment of group goals".

LEVELS OF MANAGEMENT

The three levels of management are as follows



1. **The Top Management**

   It consists of board of directors, chief executive or managing director. The top management is the ultimate source of authority and it manages goals and policies for an

enterprise. It devotes more time on planning and coordinating functions.The role of the top management can be summarized as follows –

a. Top management lays down the objectives and broad policies of the enterprise.
b. It issues necessary instructions for preparation of department budgets, procedures, schedules etc.
c. It prepares strategic plans & policies for the enterprise.
d. It appoints the executive for middle level i.e. departmental managers.
e. It controls & coordinates the activities of all the departments.
f. It is also responsible for maintaining a contact with the outside world.
g. It provides guidance and direction.
h. The top management is also responsible towards the shareholders for the performance of the enterprise.

2. Middle Level Management

The branch managers and departmental managers constitute middle level. They are responsible to the top management for the functioning of their department. They devote more time to organizational and directional functions. In small organization, there is only one layer of middle level of management but in big enterprises, there may be senior and junior middle level management. Their role can be emphasized as –

a. They execute the plans of the organization in accordance with the policies and directives of the top management.
b. They make plans for the sub-units of the organization.
c. They participate in employment & training of lower level management.
d. They interpret and explain policies from top level management to lower level.
e. They are responsible for coordinating the activities within the division or department.
f. It also sends important reports and other important data to top level management.
g. They evaluate performance of junior managers.
h. They are also responsible for inspiring lower level managers towards better performance.

3. Lower Level Management

Lower level is also known as supervisory / operative level of management. It consists of supervisors, foreman, section officers, superintendent etc. According to R.C. Davis,

"Supervisory management refers to those executives whose work has to be largely with personal oversight and direction of operative employees". In other words, they are concerned with direction and controlling function of management. Their activities include

a. Assigning of jobs and tasks to various workers.
b. They guide and instruct workers for day to day activities.
c. They are responsible for the quality as well as quantity of production.
d. They are also entrusted with the responsibility of maintaining good relation in the organization.
e. They communicate workers problems, suggestions, and recommendatory appeals etc to the higher level and higher level goals and objectives to the workers.
f. They help to solve the grievances of the workers.
g. They supervise & guide the sub-ordinates.
h. They are responsible for providing training to the workers.
i. They arrange necessary materials, machines, tools etc for getting the things done.
j. They prepare periodical reports about the performance of the workers.
k. They ensure discipline in the enterprise.
l. They motivate workers.
m. They are the image builders of the enterprise because they are in direct contact with the workers.

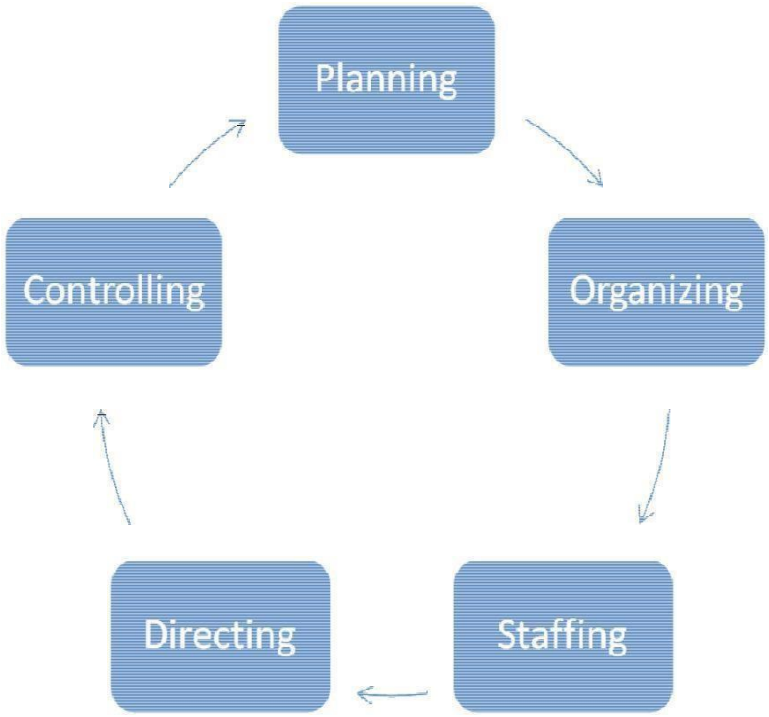
FUNCTIONS OF MANAGEMENT

Management has been described as a social process involving responsibility for economical and effective planning & regulation of operation of an enterprise in the fulfillment of given purposes. It is a dynamic process consisting of various elements and activities. These activities are different from operative functions like marketing, finance, purchase etc. Rather these activities are common to each and every manger irrespective of his level or status.

Different experts have classified functions of management. According to George & Jerry, "There are four fundamental functions of management i.e. planning, organizing, actuating and controlling". According to Henry Fayol, "To manage is to forecast and plan, to organize, to command, & to control". Whereas Luther Gullick has given a keyword '**POSDCORB**' where P

stands for Planning, O for Organizing, S for Staffing, D for Directing, Co for Co-ordination, R for reporting & B for Budgeting. But the most widely accepted are functions of management given by KOONTZ and O'DONNEL i.e. **Planning**, **Organizing**, **Staffing**, **Directing** and **Controlling**.

For theoretical purposes, it may be convenient to separate the function of management but practically these functions are overlapping in nature i.e. they are highly inseparable. Each function blends into the other & each affects the performance of others.



1. **Planning**

   It is the basic function of management. It deals with chalking out a future course of action & deciding in advance the most appropriate course of actions for achievement of pre-determined goals. According to KOONTZ, "Planning is deciding in advance – what to do, when to do & how to do. It bridges the gap from where we are & where we want to

be". A plan is a future course of actions. It is an exercise in problem solving & decision making. Planning is determination of courses of action to achieve desired goals. Thus, planning is a systematic thinking about ways & means for accomplishment of pre-determined goals. Planning is necessary to ensure proper utilization of human & non-human resources. It is all pervasive, it is an intellectual activity and it also helps in avoiding confusion, uncertainties, risks, wastages etc.

2. Organizing

It is the process of bringing together physical, financial and human resources and developing productive relationship amongst them for achievement of organizational goals. According to Henry Fayol, "To organize a business is to provide it with everything useful or its functioning i.e. raw material, tools, capital and personnel's". To organize a business involves determining & providing human and non-human resources to the organizational structure. Organizing as a process involves:

- Identification of activities.
- Classification of grouping of activities.
- Assignment of duties.
- Delegation of authority and creation of responsibility.
- Coordinating authority and responsibility relationships.

3. Staffing

It is the function of manning the organization structure and keeping it manned. Staffing has assumed greater importance in the recent years due to advancement of technology, increase in size of business, complexity of human behavior etc. The main purpose o staffing is to put right man on right job i.e. square pegs in square holes and round pegs in round holes. According to Kootz & O'Donell, "Managerial function of staffing involves manning the organization structure through proper and effective selection, appraisal & development of personnel to fill the roles designed un the structure". Staffing involves:

- Manpower Planning (estimating man power in terms of searching, choose the person and giving the right place).
- Recruitment, selection & placement.
- Training & development.
- Remuneration.
- Performance appraisal.

- Promotions & transfer.

4. Directing

It is that part of managerial function which actuates the organizational methods to work efficiently for achievement of organizational purposes. It is considered life-spark of the enterprise which sets it in motion the action of people because planning, organizing and staffing are the mere preparations for doing the work. Direction is that inert-personnel aspect of management which deals directly with influencing, guiding, supervising, motivating sub-ordinate for the achievement of organizational goals. Direction has following elements:

- Supervision
- Motivation
- Leadership
- Communication

(i) Supervision- implies overseeing the work of subordinates by their superiors. It is the act of watching & directing work & workers.

(ii) Motivation- means inspiring, stimulating or encouraging the sub-ordinates with zeal to work. Positive, negative, monetary, non-monetary incentives may be used for this purpose.

(iii) Leadership- may be defined as a process by which manager guides and influences the work of subordinates in desired direction.

(iv) Communications- is the process of passing information, experience, opinion etc from one person to another. It is a bridge of understanding.

5. Controlling

It implies measurement of accomplishment against the standards and correction of deviation if any to ensure achievement of organizational goals. The purpose of controlling is to ensure that everything occurs in conformities with the standards. An efficient system of control helps to predict deviations before they actually occur. According to Theo Haimann, "Controlling is the process of checking whether or not proper progress is being made towards the objectives and goals and acting if necessary, to correct any deviation". According to Koontz & O'Donell "Controlling is the measurement & correction of performance activities of subordinates in order to make

sure that the enterprise objectives and plans desired to obtain them as being accomplished". Therefore controlling has following steps:

(i) Establishment of standard performance.

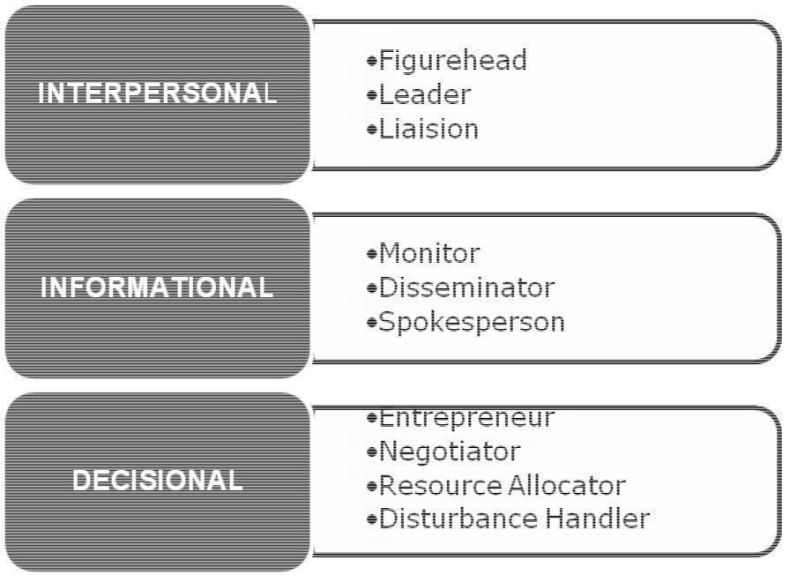
(ii) Measurement of actual performance.

(iii) Comparison of actual performance with the standards and finding out deviation if any.

(iv) Corrective action.

ROLES OF MANAGER

Henry Mintzberg identified ten different roles, separated into three categories. The categories he defined are as follows



a) Interpersonal Roles

The ones that, like the name suggests, involve people and other ceremonial duties. It can be further classified as follows

- Leader – Responsible for staffing, training, and associated duties.
- Figurehead – The symbolic head of the organization.
- Liaison – Maintains the communication between all contacts and informers that compose the organizational network.

b) Informational Roles

Related to collecting, receiving, and disseminating information.

- Monitor – Personally seek and receive information, to be able to understand the organization.
- Disseminator – Transmits all import information received from outsiders to the members of the organization.
- Spokesperson – On the contrary to the above role, here the manager transmits the organization's plans, policies and actions to outsiders.

c) Decisional Roles

Roles that revolve around making choices.

- Entrepreneur – Seeks opportunities. Basically they search for change, respond to it, and exploit it.
- Negotiator – Represents the organization at major negotiations.
- Resource Allocator – Makes or approves all significant decisions related to the allocation of resources.
- Disturbance Handler – Responsible for corrective action when the organization faces disturbances.

EVOLUTION OF MANAGEMENT THOUGHT

The practice of management is as old as human civilization. The ancient civilizations of Egypt (the great pyramids), Greece (leadership and war tactics of Alexander the great) and Rome displayed the marvelous results of good management practices.

The origin of management as a discipline was developed in the late 19^{th} century. Over time, management thinkers have sought ways to organize and classify the voluminous information about management that has been collected and disseminated. These attempts at classification have resulted in the identification of management approaches. The approaches of management are theoretical frameworks for the study of management. Each of the approaches of management are based on somewhat different assumptions about human beings and the organizations for which they work.

The different approaches of management are

a) Classical approach,

b) Behavioral approach,

c) Quantitative approach,

d) Systems approach,

e) Contingency approach.

The formal study of management is largely a twentieth-century phenomenon, and to some degree the relatively large number of management approaches reflects a lack of consensus among management scholars about basic questions of theory and practice.

a) THE CLASSICAL APPROACH:

The classical approach is the oldest formal approach of management thought. Its roots pre-date the twentieth century. The classical approach of thought generally concerns ways to manage work and organizations more efficiently. Three areas of study that can be grouped under the classical approach are scientific management, administrative management, and bureaucratic management.

(i) Scientific Management.

Frederick Winslow Taylor is known as the father of scientific management. Scientific management (also called Taylorism or the Taylor system) is a theory of management that analyzes and synthesizes workflows, with the objective of improving labor productivity. In other words, Traditional rules of thumb are replaced by precise procedures developed after careful study of an individual at work.

(ii) Administrative Management.

Administrative management focuses on the management process and principles of management. In contrast to scientific management, which deals largely with jobs and work at the individual level of analysis, administrative management provides a more general theory of management. Henri Fayol is the major contributor to this approach of management thought. (iii) Bureaucratic Management.

Bureaucratic management focuses on the ideal form of organization. Max Weber was the major contributor to bureaucratic management. Based on observation, Weber concluded that many early organizations were inefficiently managed, with decisions based on personal relationships and loyalty. He proposed that a form of organization, called a bureaucracy, characterized by division of labor, hierarchy, formalized rules, impersonality, and the selection and promotion of employees based on ability, would lead to more efficient management. Weber also contended that managers' authority in an organization should be based not on tradition or charisma but on the position held by managers in the organizational hierarchy.

b) THE BEHAVIORAL APPROACH:

The behavioral approach of management thought developed, in part, because of perceived weaknesses in the assumptions of the classical approach. The classical approach emphasized efficiency, process, and principles. Some felt that this emphasis disregarded important aspects of organizational life, particularly as it related to human behavior. Thus, the behavioral approach focused on trying to understand the factors that affect human behavior at work.

(i) Human Relations.

The Hawthorne Experiments began in 1924 and continued through the early 1930s. A variety of researchers participated in the studies, including Elton Mayo. One of the major conclusions of the Hawthorne studies was that workers' attitudes are associated with productivity. Another was that the workplace is a social system and informal group influence could exert a powerful effect on individual behavior. A third was that the style of supervision is an important factor in increasing workers' job satisfaction. (ii) Behavioral Science.

Behavioral science and the study of organizational behavior emerged in the 1950s and 1960s. The behavioral science approach was a natural progression of the human relations movement. It focused on applying conceptual and analytical tools to the problem of understanding and predicting behavior in the workplace.

The behavioral science approach has contributed to the study of management through its focus on personality, attitudes, values, motivation, group behavior, leadership, communication, and conflict, among other issues.

c) THE QUANTITATIVE APPROACH:

The quantitative approach focuses on improving decision making via the application of quantitative techniques. Its roots can be traced back to scientific management. (i) Management Science (Operations Research)

Management science (also called operations research) uses mathematical and statistical approaches to solve management problems. It developed during World War II as strategists tried to apply scientific knowledge and methods to the complex problems of war. Industry began to apply management science after the war. The advent of the computer made many management science tools and concepts more practical for industry

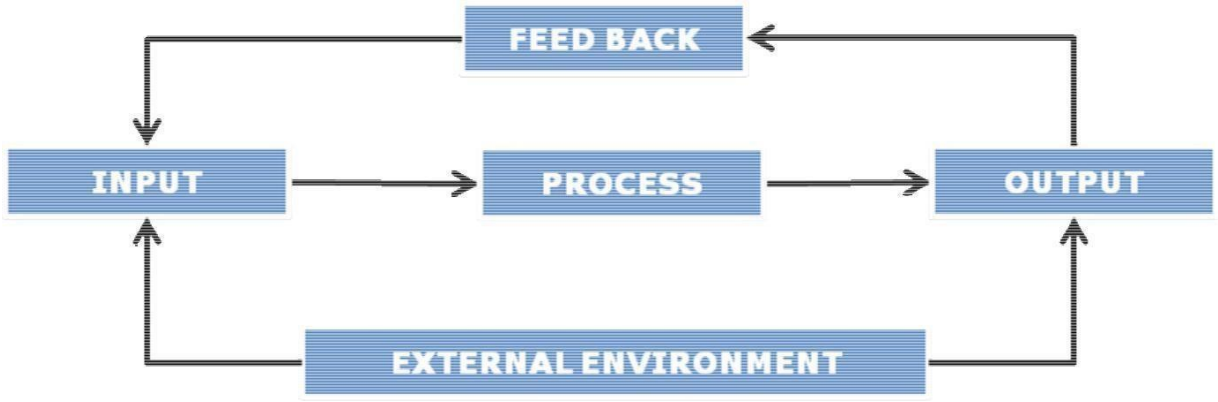
(ii) Production And Operations Management.

This approach focuses on the operation and control of the production process that transforms resources into finished goods and services. It has its roots in scientific management but became an identifiable area of management study after World War II. It uses many of the tools of management science.

Operations management emphasizes productivity and quality of both manufacturing and service organizations. W. Edwards Deming exerted a tremendous influence in shaping modern ideas about improving productivity and quality. Major areas of study within operations management include capacity planning, facilities location, facilities layout, materials requirement planning, scheduling, purchasing and inventory control, quality control, computer integrated manufacturing, just-in-time inventory systems, and flexible manufacturing systems.

d) SYSTEMS APPROACH:

The simplified block diagram of the systems approach is given below.



The systems approach focuses on understanding the organization as an open system that transforms inputs into outputs. The systems approach began to have a strong impact on management thought in the 1960s as a way of thinking about managing techniques that would allow managers to relate different specialties and parts of the company to one another, as well as to external environmental factors. The systems approach focuses on the organization as a whole, its interaction with the environment, and its need to achieve equilibrium

e) CONTINGENCY APPROACH:

The contingency approach focuses on applying management principles and processes as dictated by the unique characteristics of each situation. It emphasizes that there is no one best way to manage and that it depends on various situational factors, such as the external environment, technology, organizational characteristics, characteristics of the manager, and characteristics of the subordinates. Contingency theorists often implicitly or explicitly criticize the classical approach for its emphasis on the universality of management principles; however, most classical writers recognized the need to consider aspects of the situation when applying management principles.

MANAGEMENT APPROACHS	Beginning Dates	Emphasis
CLASSICAL APPROACH		
Scientific Management	1880s	Traditional rules of thumb are replaced by precise procedures developed after careful study of an individual at work.
Administrative Management	1940s	Gives idea about the primary functions of management and The 14 Principles of Administration
Bureaucratic Management	1920s	Replaces traditional leadership and charismatic leadership with legal leadership
BEHAVIORAL APPROACH		
Human Relations	1930s	workers' attitudes are associated with productivity
Behavioral Science	1950s	Gives idea to understand human behavior in the organization.

QUANTITATIVE APPROACH		
Management Science (Operation research)	1940s	Uses mathematical and statistical approaches to solve management problems.
Production and Operations Management	1940s	This approach focuses on the operation and control of the production process that transforms resources into finished goods and services
RECENT DEVELOPEMENTS		
SYSTEMS APPROACH	1950s	Considers the organization as a system that transforms inputs into outputs while in constan interaction with its' environment.
CONTINGENCY APPROACH	1960s	Applies management principles and processes as dictated by the unique characteristics of each situation.

CONTRIBUTION OF FAYOL AND TAYLOR

F.W. Taylor and Henry Fayol are generally regarded as the founders of scientific management and administrative management and both provided the bases for science and art of management.

Taylor's Scientific Management

Frederick Winslow Taylor well-known as the founder of scientific management was the first to recognize and emphasis the need for adopting a scientific approach to the task of managing an enterprise. He tried to diagnose the causes of low efficiency in industry and came to the conclusion that much of waste and inefficiency is due to the lack of order and system in the methods of management. He found that the management was usually ignorant of the amount of work that could be done by a worker in a day as also the best method of doing the job. As a result, it remained largely at the mercy of the workers who deliberately shirked work. He

therefore, suggested that those responsible for management should adopt a scientific approach in their work, and make use of "scientific method" for achieving higher efficiency. The scientific method consists essentially of

(a) Observation

(b) Measurement

(c) Experimentation and

(d) Inference.

He advocated a thorough planning of the job by the management and emphasized the necessity of perfect understanding and co-operation between the management and the workers both for the enlargement of profits and the use of scientific investigation and knowledge in industrial work. He summed up his approach in these words:

- Science, not rule of thumb
- Harmony, not discord
- Co-operation, not individualism
- Maximum output, in place of restricted output
- The development of each man to his greatest efficiency and prosperity.

Elements of Scientific Management: The techniques which Taylor regarded as its essential elements or features may be classified as under:

1. Scientific Task and Rate-setting, work improvement, etc.
2. Planning the Task.
3. Vocational Selection and Training
4. Standardization (of working conditions, material equipment etc.)
5. Specialization
6. Mental Revolution.

1. **Scientific Task and Rate-Setting (work study):** Work study may be defined as the systematic, objective and critical examination of all the factors governing the operational efficiency of any specified activity in order to effect improvement.

Work study includes.

(a) **Methods Study:** The management should try to ensure that the plant is laid out in the best manner and is equipped with the best tools and machinery. The possibilities of eliminating or combining certain operations may be studied.

(b) **Motion Study:** It is a study of the movement, of an operator (or even of a machine) in performing an operation with the purpose of eliminating useless motions.

(c) **Time Study (work measurement):** The basic purpose of time study is to determine the proper time for performing the operation. Such study may be conducted after the motion study. Both time study and motion study help in determining the best method of doing a job and the standard time allowed for it.

(d) **Fatigue Study:** If, a standard task is set without providing for measures to eliminate fatigue, it may either be beyond the workers or the workers may over strain themselves to attain it. It is necessary, therefore, to regulate the working hours and provide for rest pauses at scientifically determined intervals.

(e) **Rate-setting:** Taylor recommended the differential piece wage system, under which workers performing the standard task within prescribed time are paid a much higher rate per unit than inefficient workers who are not able to come up to the standard set.

2. **Planning the Task:** Having set the task which an average worker must strive to perform to get wages at the higher piece-rate, necessary steps have to be taken to plan the production thoroughly so that there is no bottlenecks and the work goes on systematically.

3. **Selection and Training:** Scientific Management requires a radical change in the methods and procedures of selecting workers. It is therefore necessary to entrust the task of selection to a central personnel department. The procedure of selection will also have to be systematised. Proper attention has also to be devoted to the training of the workers in the correct methods of work.

4. **Standardization:** Standardization may be introduced in respect of the following.

(a) **Tools and equipment:** By standardization is meant the process of bringing about uniformity. The management must select and store standard tools and implements which will be nearly the best or the best of their kind.

(b) **Speed:** There is usually an optimum speed for every machine. If it is exceeded, it is likely to result in damage to machinery.

(c) **Conditions of Work:** To attain standard performance, the maintenance of standard conditions of ventilation, heating, cooling, humidity, floor space, safety etc., is very essential.

(d) **Materials:** The efficiency of a worker depends on the quality of materials and the method of handling materials.

5. **Specialization:** Scientific management will not be complete without the introduction of specialization. Under this plan, the two functions of 'planning' and 'doing' are separated in the organization of the plant. The `functional foremen' are specialists who join their heads to give

thought to the planning of the performance of operations in the workshop. Taylor suggested eight functional foremen under his scheme of functional foremanship.

(a) **The Route Clerk:** To lay down the sequence of operations and instruct the workers concerned about it.

(b) **The Instruction Card Clerk:** To prepare detailed instructions regarding different aspects of work.

(c) **The Time and Cost Clerk:** To send all information relating to their pay to the workers and to secure proper returns of work from them.

(d) **The Shop Disciplinarian**: To deal with cases of breach of discipline and absenteeism.

(e) **The Gang Boss:** To assemble and set up tools and machines and to teach the workers to make all their personal motions in the quickest and best way.

(f) **The Speed Boss:** To ensure that machines are run at their best speeds and proper tools are used by the workers.

(g) **The Repair Boss:** To ensure that each worker keeps his machine in good order and maintains cleanliness around him and his machines.

(h) **The Inspector:** To show to the worker how to do the work.

6. **Mental Revolution:** At present, industry is divided into two groups – management and labour. The major problem between these two groups is the division of surplus. The management wants the maximum possible share of the surplus as profit; the workers want, as large share in the form of wages. Taylor has in mind the enormous gain that arises from higher productivity. Such gains can be shared both by the management and workers in the form of increased profits and increased wages.

Henry Fayol's 14 Principles of Management:

The principles of management are given below:

1. **Division of work:** Division of work or specialization alone can give maximum productivity and efficiency. Both technical and managerial activities can be performed in the best manner only through division of labour and specialization.

2. **Authority and Responsibility:** The right to give order is called authority. The obligation to accomplish is called responsibility. Authority and Responsibility are the two sides of the management coin. They exist together. They are complementary and mutually interdependent.

3. **Discipline:** The objectives, rules and regulations, the policies and procedures must be honoured by each member of an organization. There must be clear and fair agreement on the rules and objectives, on the policies and procedures. There must be penalties (punishment) for non-obedience or indiscipline. No organization can work smoothly without discipline - preferably voluntary discipline.

4. **Unity of Command:** In order to avoid any possible confusion and conflict, each member of an organization must received orders and instructions only from one superior (boss).

5. **Unity of Direction:** All members of an organization must work together to accomplish common objectives.

6. **Emphasis on Subordination of Personal Interest to General or Common Interest:** This is also called principle of co-operation. Each shall work for all and all for each. General or common interest must be supreme in any joint enterprise.

7. **Remuneration:** Fair pay with non-financial rewards can act as the best incentive or motivator for good performance. Exploitation of employees in any manner must be eliminated. Sound scheme of remuneration includes adequate financial and nonfinancial incentives.

8. **Centralization:** There must be a good balance between centralization and decentralization of authority and power. Extreme centralization and decentralization must be avoided.

9. **Scalar Chain:** The unity of command brings about a chain or hierarchy of command linking all members of the organization from the top to the bottom. Scalar denotes steps.

10. **Order:** Fayol suggested that there is a place for everything. Order or system alone can create a sound organization and efficient management.

11. **Equity:** An organization consists of a group of people involved in joint effort. Hence, equity (i.e., justice) must be there. Without equity, we cannot have sustained and adequate joint collaboration.

12. **Stability of Tenure:** A person needs time to adjust himself with the new work and demonstrate efficiency in due course. Hence, employees and managers must have job security. Security of income and employment is a pre-requisite of sound organization and management.

13. **Esprit of Co-operation:** Esprit de corps is the foundation of a sound organization. Union is strength. But unity demands co-operation. Pride, loyalty and sense of belonging are responsible for good performance.

14. **Initiative:** Creative thinking and capacity to take initiative can give us sound managerial planning and execution of predetermined plans.

ORGANIZATION AND ENVIRONMENTAL FACTORS

An organization is a group of people intentionally organized to accomplish a common or set of goals.

Types of Business Organizations

When organizing a new business, one of the most important decisions to be made is choosing the structure of a business.

a) Sole Proprietorships

The vast majority of small business starts out as sole proprietorships . . . very dangerous. These firms are owned by one person, usually the individual who has day-to-day responsibility for running the business. Sole proprietors own all the assets of the business and the profits generated by it. They also assume "complete personal" responsibility for all of its liabilities or debts. In the eyes of the law, you are one in the same with the business.

Merits:

- Easiest and least expensive form of ownership to organize.
- Sole proprietors are in complete control, within the law, to make all decisions.
- Sole proprietors receive all income generated by the business to keep or reinvest.
- Profits from the business flow-through directly to the owner's personal tax return.
- The business is easy to dissolve, if desired.

Demerits:

- Unlimited liability and are legally responsible for all debts against the business.
- Their business and personal assets are 100% at risk.
- Has almost been ability to raise investment funds.
- Are limited to using funds from personal savings or consumer loans.
- Have a hard time attracting high-caliber employees, or those that are motivated by the opportunity to own a part of the business.
- Employee benefits such as owner's medical insurance premiums are not directly deductible from business income (partially deductible as an adjustment to income).

b) Partnerships

In a Partnership, two or more people share ownership of a single business. Like proprietorships, the law does not distinguish between the business and its owners. The Partners should have a

legal agreement that sets forth how decisions will be made, profits will be shared, disputes will be resolved, how future partners will be admitted to the partnership, how partners can be bought out, or what steps will be taken to dissolve the partnership when needed. Yes, its hard to think about a "break-up" when the business is just getting started, but many partnerships split up at crisis times and unless there is a defined process, there will be even greater problems. They also must decide up front how much time and capital each will contribute, etc.

Merits:

- Partnerships are relatively easy to establish; however time should be invested in developing the partnership agreement.
- With more than one owner, the ability to raise funds may be increased.
- The profits from the business flow directly through to the partners' personal taxes.
- Prospective employees may be attracted to the business if given the incentive to become a partner.

Demerits:

- Partners are jointly and individually liable for the actions of the other partners.
- Profits must be shared with others.
- Since decisions are shared, disagreements can occur.
- Some employee benefits are not deductible from business income on tax returns.
- The partnerships have a limited life; it may end upon a partner withdrawal or death.

c) Corporations

A corporation, chartered by the state in which it is headquartered, is considered by law to be a unique "entity", separate and apart from those who own it. A corporation can be taxed; it can be sued; it can enter into contractual agreements. The owners of a corporation are its shareholders. The shareholders elect a board of directors to oversee the major policies and decisions. The corporation has a life of its own and does not dissolve when ownership changes.

Merits:

- Shareholders have limited liability for the corporation's debts or judgments against the corporations.
- Generally, shareholders can only be held accountable for their investment in stock of the company. (Note however, that officers can be held personally liable for their actions, such as the failure to withhold and pay employment taxes.)
- Corporations can raise additional funds through the sale of stock.

- A corporation may deduct the cost of benefits it provides to officers and employees.
- Can elect S corporation status if certain requirements are met. This election enables company to be taxed similar to a partnership.

Demerits:

- The process of incorporation requires more time and money than other forms of organization.
- Corporations are monitored by federal, state and some local agencies, and as a result may have more paperwork to comply with regulations.
- Incorporating may result in higher overall taxes. Dividends paid to shareholders are not deductible form business income, thus this income can be taxed twice.

d) Joint Stock Company:

Limited financial resources & heavy burden of risk involved in both of the previous forms of organization has led to the formation of joint stock companies these have limited dilutives.
The capital is raised by selling shares of different values. Persons who purchase the shares are called shareholder. The managing body known as; Board of Directors; is responsible for policy making important financial & technical decisions. There are two main types of joint stock Companies.

(i) Private limited company.
(ii) Public limited company

(i) Private limited company: This type company can be formed by two or more persons. Te maximum number of member ship is limited to 50. In this transfer of shares is limited to members only. The government also does not interfere in the working of the company.
(ii) Public Limited Company: Its is one whose membership is open to general public. The minimum number required to form such company is seven, but there is no upper limit. Such company's can advertise to offer its share to genera public through a prospectus. These public limited companies are subjected to greater control & supervision of control.

Merits:

- The liability being limited the shareholder bear no Rick& therefore more as make persons are encouraged to invest capital.
- Because of large numbers of investors, the risk of loss is divided.
- Joint stock companies are not affected by the death or the retirement of the shareholders.

Disadvantages:

- It is difficult to preserve secrecy in these companies.

- It requires a large number of legal formalities to be observed.
- Lack of personal interest.

e) Public Corporations:

A public corporation is wholly owned by the Government centre to state. It is established usually by a Special Act of the parliament. Special statute also prescribes its management pattern power duties & jurisdictions. Though the total capital is provided by the Government, they have separate entity & enjoy independence in matters related to appointments, promotions etc.

Merits:

- These are expected to provide better working conditions to the employees & supported to be better managed.
- Quick decisions can be possible, because of absence of bureaucratic control.
- More Hexibility as compared to departmental organization.
- Since the management is in the hands of experienced & capable directors & managers, these ate managed more efficiently than that of government departments.

Demerits:

- Any alteration in the power & Constitution of Corporation requires an amendment in the particular Act, which is difficult & time consuming.
- Public Corporations possess monopoly & in the absence of competition, these are not interested in adopting new techniques & in making improvement in their working.

f) Government Companies:

A state enterprise can also be organized in the form of a Joint stock company; A government company is any company in which of the share capital is held by the central government or partly by central government & party by one to more state governments. It is managed b the elected board of directors which may include private individuals. These are accountable for its working to the concerned ministry or department & its annual report is required to be placed ever year on the table of the parliament or state legislatures along with the comments of the government to concerned department.

Merits:

- It is easy to form.
- The directors of a government company are free to take decisions & are not bound by certain rigid rules & regulations.

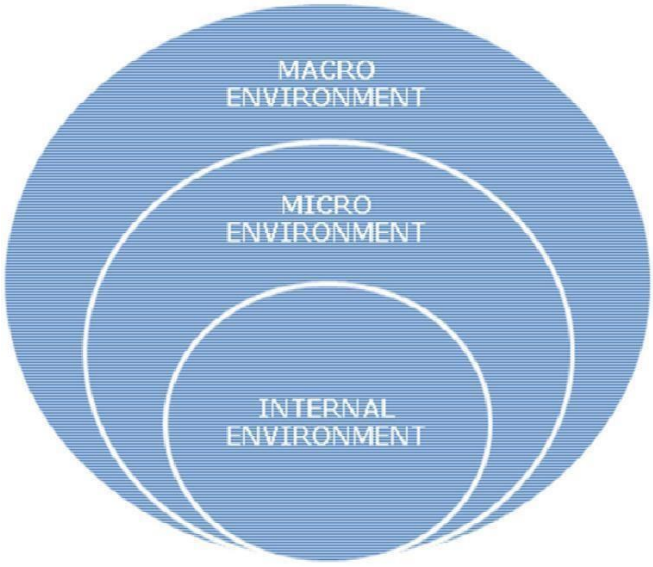
Demerits:

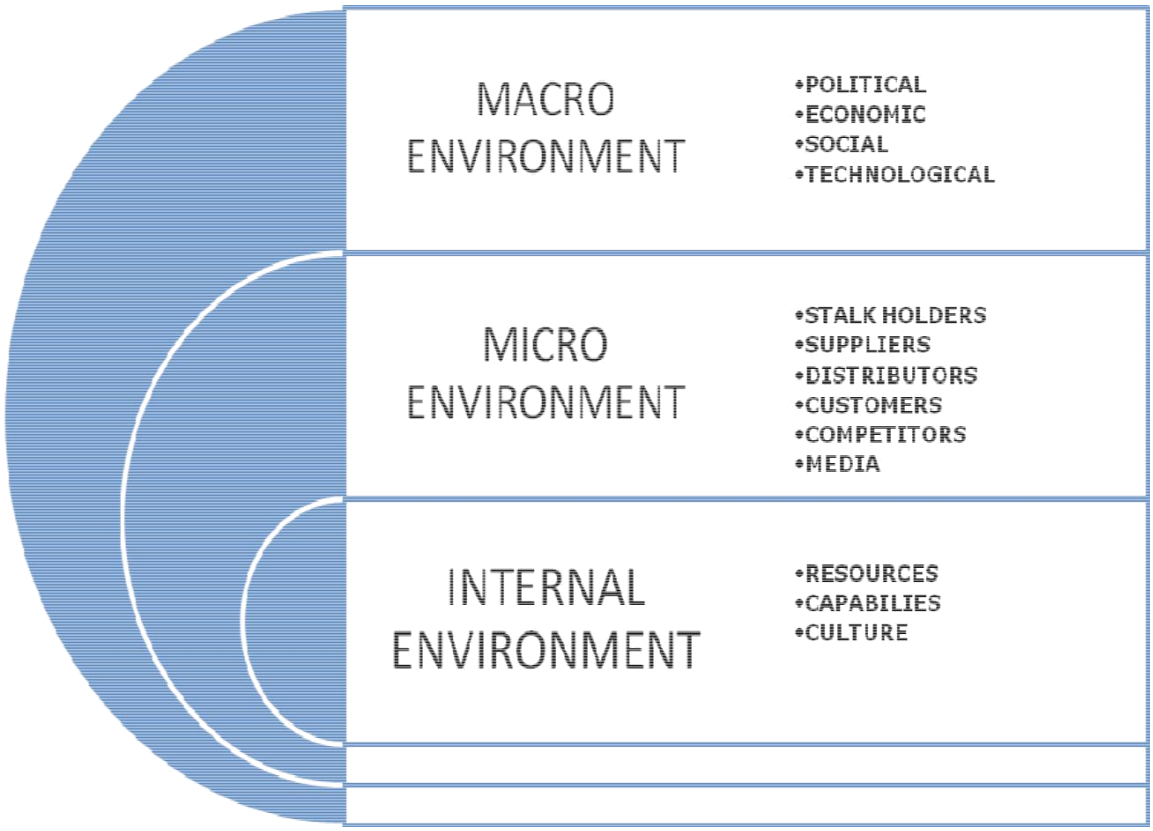
- Misuse of excessive freedom cannot be ruled out.

- The directors are appointed by the government so they spend more time in pleasing their political masters & top government officials, which results in inefficient management.

CLASSIFICATION OF ENVIRONMENTAL FACTORS

On the basis of the extent of intimacy with the firm, the environmental factors may be classified into different types namely internal and external.





1) INTERNAL ENVIRONMENTAL FACTORS

The internal environment is the environment that has a direct impact on the business. The internal factors are generally controllable because the company has control over these factors. It can alter or modify these factors. The internal environmental factors are resources, capabilities and culture.

i) Resources:

A good starting point to identify company resources is to look at tangible, intangible and human resources.

Tangible resources are the easiest to identify and evaluate: financial resources and physical assets are identifies and valued in the firm's financial statements.

Intangible resources are largely invisible, but over time become more important to the firm than tangible assets because they can be a main source for a competitive advantage. Such intangible recourses include reputational assets (brands, image, etc.) and technological assets (proprietary technology and know-how).

Human resources or human capital are the productive services human beings offer the firm in terms of their skills, knowledge, reasoning, and decision-making abilities.

ii) Capabilities:

Resources are not productive on their own. The most productive tasks require that resources collaborate closely together within teams. The term organizational capabilities are used to refer to a firm's capacity for undertaking a particular productive activity. Our interest is not in capabilities per se, but in capabilities relative to other firms. To identify the firm's capabilities we will use the functional classification approach. A functional classification identifies organizational capabilities in relation to each of the principal functional areas.

iii) Culture:

It is the specific collection of values and norms that are shared by people and groups in an organization and that helps in achieving the organizational goals.

2) EXTERNAL ENVIRONMENT FACTORS

It refers to the environment that has an indirect influence on the business. The factors are uncontrollable by the business. The two types of external environment are micro environment and macro environment.

a) MICRO ENVIRONMENTAL FACTORS

These are external factors close to the company that have a direct impact on the organizations process. These factors include:

i) Shareholders

Any person or company that owns at least one share (a percentage of ownership) in a company is known as shareholder. A shareholder may also be referred to as a "stockholder". As organization requires greater inward investment for growth they face increasing pressure to move from private ownership to public. However this movement unleashes the forces of shareholder pressure on the strategy of organizations.

ii) Suppliers

An individual or an organization involved in the process of making a product or service available for use or consumption by a consumer or business user is known as supplier. Increase in raw material prices will have a knock on affect on the marketing mix strategy of an organization. Prices may be forced up as a result. A closer supplier relationship is one way of ensuring competitive and quality products for an organization.

iii) Distributors

Entity that buys non-competing products or product-lines, warehouses them, and resells them to retailers or direct to the end users or customers is known as distributor. Most distributors provide strong manpower and cash support to the supplier or manufacturer's promotional efforts. They usually also provide a range of services (such as product information, estimates, technical support, after-sales services, credit) to their customers. Often getting products to the end customers can be a major issue for firms. The distributors used will determine the final price of the product and how it is presented to the end customer. When selling via retailers, for example, the retailer has control over where the products are displayed, how they are priced and how much they are promoted in-store. You can also gain a competitive advantage by using changing distribution channels.

iv) Customers

A person, company, or other entity which buys goods and services produced by another person, company, or other entity is known as customer. Organizations survive on the basis of meeting the needs, wants and providing benefits for their customers. Failure to do so will result in a failed business strategy.

v) Competitors

A company in the same industry or a similar industry which offers a similar product or service is known as competitor. The presence of one or more competitors can reduce the prices of goods and services as the companies attempt to gain a larger market share. Competition also requires companies to become more efficient in order to reduce costs. Fast-food restaurants McDonald's and Burger King are competitors, as are Coca-Cola and Pepsi, and Wal-Mart and Target.

vi) Media

Positive or adverse media attention on an organisations product or service can in some cases make or break an organisation.. Consumer programmes with a wider and more direct audience can also have a very powerful and positive impact, hforcing organisations to change their tactics.

b) MACRO ENVIRONMENTAL FACTORS

An organization's macro environment consists of nonspecific aspects in the organization's surroundings that have the potential to affect the organization's strategies. When compared to a firm's task environment, the impact of macro environmental variables is less direct and the organization has a more limited impact on these elements of the environment.

The macro environment consists of forces that originate outside of an organization and generally cannot be altered by actions of the organization. In other words, a firm may be influenced by changes within this element of its environment, but cannot itself influence the environment. The curved lines in Figure 1 indicate the indirect influence of the environment on the organization.

Macro environment includes political, economic, social and technological factors. A firm considers these as part of its environmental scanning to better understand the threats and opportunities created by the variables and how strategic plans need to be adjusted so the firm can obtain and retain competitive advantage.

i) Political Factors

Political factors include government regulations and legal issues and define both formal and informal rules under which the firm must operate. Some examples include:

- tax policy
- employment laws
- environmental regulations
- trade restrictions and tariffs
- political stability

ii) Economic Factors

Economic factors affect the purchasing power of potential customers and the firm's cost of capital. The following are examples of factors in the macroeconomy:

- economic growth
- interest rates
- exchange rates
- inflation rate

iii) Social Factors

Social factors include the demographic and cultural aspects of the external macro environment. These factors affect customer needs and the size of potential markets. Some social factors include:

- health consciousness
- population growth rate
- age distribution
- career attitudes
- emphasis on safety

iv) Technological Factors

Technological factors can lower barriers to entry, reduce minimum efficient production levels, and influence outsourcing decisions. Some technological factors include:

- R&D activity
- automation
- technology incentives
- rate of technological change

TRENDS AND CHALLENGES OF MANAGEMENT IN GLOBAL SCENARIO

The management functions are planning and decision making, organizing. leading, and controlling — are just as relevant to international managers as to domestic managers. International managers need to have a clear view of where they want their firm to be in the future; they have to organize to implement their plans: they have to motivate those who work lot them; and they have to develop appropriate control mechanisms.

a) Planning and Decision Making in a Global Scenario

To effectively plan and make decisions in a global economy, managers must have a broad-based understanding of both environmental issues and competitive issues. They need to understand local market conditions and technological factor that will affect their operations. At the corporate level, executives need a great deal of information to function effectively. Which markets are growing? Which markets are shrinking? Which are our domestic and foreign competitors doing in each market? They must also make a variety of strategic decisions about their organizations. For example, if a firm wishes to enter market in France, should it buy a local

firm there, build a plant, or seek a strategic alliance? Critical issues include understanding environmental circumstances, the role of goals and planning in a global organization, and how decision making affects the global organization.

b) Organizing in a Global Scenario

Managers in international businesses must also attend to a variety of organizing issues. For example, General Electric has operations scattered around the globe.The firm has made the decision to give local managers a great deal of responsibility for how they run their business. In contrast, many Japanese firms give managers of their foreign operations relatively little responsibility. As a result, those managers must frequently travel back to Japan to present problems or get decisions approved. Managers in an international business must address the basic issues of organization structure and design, managing change, and dealing with human resources.

c) Leading in a Global Scenario

We noted earlier some of the cultural factors that affect international organizations. Individual managers must be prepared to deal with these and other factors as they interact people from different cultural backgrounds .Supervising a group of five managers, each of whom is from a different state in the United States, is likely to be much simpler than supervising a group of five managers, each of whom is from a different culture. Managers must understand how cultural factors affect individuals. How motivational processes vary across cultures, how the role of leadership changes in different cultures, how communication varies across cultures, and how interpersonal and group processes depend on cultural background.

d) Controlling in a Global Scenario

Finally, managers in international organizations must also be concerned with control. Distances, time zone differences, and cultural factors also play a role in control. For example, in some cultures, close supervision is seen as being appropriate, whereas in other cultures, it is not Like-wise, executives in the United States and Japan may find it difficult to communicate vital information to one another because of the time zone differences. Basic control issues for the international manager revolve around operations management productivity, quality, technology and information systems.

UNIT II

PLANNING

DEFINITION

According to Koontz O'Donnel - "Planning is an intellectual process, the conscious determination of courses of action, the basing of decisions on purpose, acts and considered estimates".

NATURE AND PURPOSE OF PLANNING

Nature of Planning

1. **Planning is goal-oriented:** Every plan must contribute in some positive way towards the accomplishment of group objectives. Planning has no meaning without being related to goals.
2. **Primacy of Planning:** Planning is the first of the managerial functions. It precedes all other management functions.
3. **Pervasiveness of Planning:** Planning is found at all levels of management. Top management looks after strategic planning. Middle management is in charge of administrative planning. Lower management has to concentrate on operational planning.
4. **Efficiency, Economy and Accuracy:** Efficiency of plan is measured by its contribution to the objectives as economically as possible. Planning also focuses on accurate forecasts.
5. **Co-ordination:** Planning co-ordinates the what, who, how, where and why of planning. Without co-ordination of all activities, we cannot have united efforts.
6. **Limiting Factors:** A planner must recognize the limiting factors (money, manpower etc) and formulate plans in the light of these critical factors.
7. **Flexibility:** The process of planning should be adaptable to changing environmental conditions.
8. **Planning is an intellectual process:** The quality of planning will vary according to the quality of the mind of the manager.

Purpose of Planning

As a managerial function planning is important due to the following reasons:-

1. **To manage by objectives:** All the activities of an organization are designed to achieve certain specified objectives. However, planning makes the objectives more concrete by focusing attention on them.

2. **To offset uncertainty and change:** Future is always full of uncertainties and changes. Planning foresees the future and makes the necessary provisions for it.

3. **To secure economy in operation:** Planning involves, the selection of most profitable course of action that would lead to the best result at the minimum costs.

4. **To help in co-ordination:** Co-ordination is, indeed, the essence of management, the planning is the base of it. Without planning it is not possible to co-ordinate the different activities of an organization.

5. **To make control effective:** The controlling function of management relates to the comparison of the planned performance with the actual performance. In the absence of plans, a management will have no standards for controlling other's performance.

6. **To increase organizational effectiveness:** Mere efficiency in the organization is not important; it should also lead to productivity and effectiveness. Planning enables the manager to measure the organizational effectiveness in the context of the stated objectives and take further actions in this direction.

Features of Planning

- It is primary function of management.
- It is an intellectual process
- Focuses on determining the objectives
- Involves choice and decision making
- It is a continuous process
- It is a pervasive function

Classification of Planning

On the basis of content

- Strategic Planning
  - It is the process of deciding on Long-term objectives of the organization.
  - It encompasses all the functional areas of business

- Tactical Planning

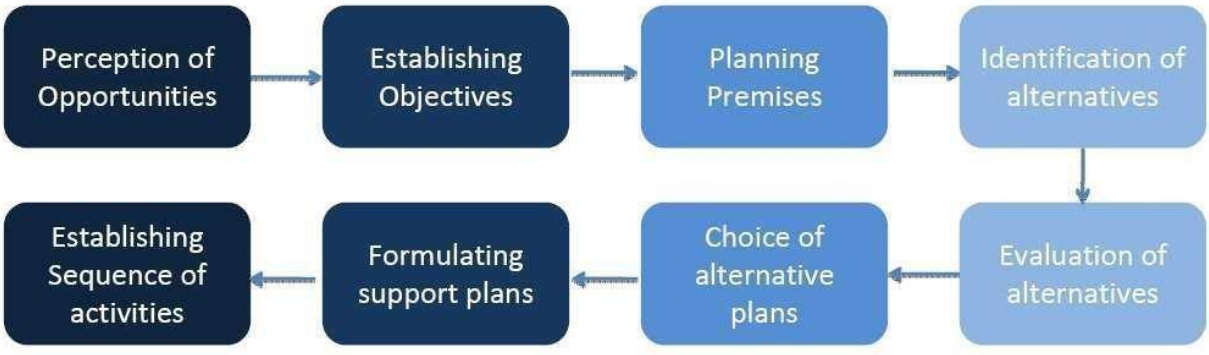
- It involves conversion of detailed and specific plans into detailed and specific action plans.
- It is the blue print for current action and it supports the strategic plans.

On the basis of time period

- Long term planning
  - Time frame beyond five years.
  - It specifies what the organization wants to become in long run.
  - It involves great deal of uncertainty.
- Intermediate term planning
  - Time frame between two and five years.
  - It is designed to implement long term plans.
- Short term planning
  - Time frame of one year or less.
  - It provide basis for day to day operations.

PLANNING PROCESS

The various steps involved in planning are given below



Planning Process

a) Perception of Opportunities:

Although preceding actual planning and therefore not strictly a part of the planning process, awareness of an opportunity is the real starting point for planning. It includes a preliminary look at possible future opportunities and the ability to see them clearly and completely, knowledge of where we stand in the light of our strengths and weaknesses, an understanding of why we wish to solve uncertainties, and a vision of what we expect to gain. Setting realistic objectives depends on this awareness. Planning requires realistic diagnosis of the opportunity situation.

b) Establishing Objectives:

The first step in planning itself is to establish objectives for the entire enterprise and then for each subordinate unit. Objectives specifying the results expected indicate the end points of what is to be done, where the primary emphasis is to be placed, and what is to be accomplished by the network of strategies, policies, procedures, rules, budgets and programs.

Enterprise objectives should give direction to the nature of all major plans which, by reflecting these objectives, define the objectives of major departments. Major department objectives, in turn, control the objectives of subordinate departments, and so on down the line. The objectives of lesser departments will be better framed, however, if subdivision managers understand the overall enterprise objectives and the implied derivative goals and if they are given an opportunity to contribute their ideas to them and to the setting of their own goals.

c) Considering the Planning Premises:

Another logical step in planning is to establish, obtain agreement to utilize and disseminate critical planning premises. These are forecast data of a factual nature, applicable basic policies, and existing company plans. Premises, then, are planning assumptions – in other words, the expected environment of plans in operation. This step leads to one of the major principles of planning.

The more individuals charged with planning understand and agree to utilize consistent planning premises, the more coordinated enterprise planning will be.

Planning premises include far more than the usual basic forecasts of population, prices, costs, production, markets, and similar matters.

Because the future environment of plans is so complex, it would not be profitable or realistic to make assumptions about every detail of the future environment of a plan.

Since agreement to utilize a given set of premises is important to coordinate planning, it becomes a major responsibility of managers, starting with those at the top, to make sure that subordinate managers understand the premises upon which they are expected to plan. It is not unusual for chief executives in well- managed companies to force top managers with differing views, through group deliberation, to arrive at a set of major premises that all can accept.

d) Identification of alternatives:

Once the organizational objectives have been clearly stated and the planning premises have been developed, the manager should list as many available alternatives as possible for reaching those objectives.

The focus of this step is to search for and examine alternative courses of action, especially those not immediately apparent. There is seldom a plan for which reasonable alternatives do not exist, and quite often an alternative that is not obvious proves to be the best.

The more common problem is not finding alternatives, but reducing the number of alternatives so that the most promising may be analyzed. Even with mathematical techniques and the computer, there is a limit to the number of alternatives that may be examined. It is therefore usually necessary for the planner to reduce by preliminary examination the number of alternatives to those promising the most fruitful possibilities or by mathematically eliminating, through the process of approximation, the least promising ones.

e) Evaluation of alternatives

Having sought out alternative courses and examined their strong and weak points, the following step is to evaluate them by weighing the various factors in the light of premises and goals. One course may appear to be the most profitable but require a large cash outlay and a slow payback; another may be less profitable but involve less risk; still another may better suit the company in long–range objectives.

If the only objective were to examine profits in a certain business immediately, if the future were not uncertain, if cash position and capital availability were not worrisome, and if most factors could be reduced to definite data, this evaluation should be relatively easy. But typical planning is replete with uncertainties, problems of capital shortages, and intangible factors, and so evaluation is usually very difficult, even with relatively simple problems. A company may wish to enter a new product line primarily for purposes of prestige; the forecast of expected results may show a clear financial loss, but the question is still open as to whether the loss is worth the gain.

f) Choice of alternative plans

An evaluation of alternatives must include an evaluation of the premises on which the alternatives are based. A manager usually finds that some premises are unreasonable and can therefore be excluded from further consideration. This elimination process helps the manager determine which alternative would best accomplish organizational objectives.

g) Formulating of Supporting Plans

After decisions are made and plans are set, the final step to give them meaning is to numberize them by converting them to budgets. The overall budgets of an enterprise represent the sum total of income and expenses with resultant profit or surplus and budgets of major balance–sheet items such as cash and capital expenditures. Each department or program of a business or other enterprise can have its own budgets, usually of expenses and capital expenditures, which tie into the overall budget.

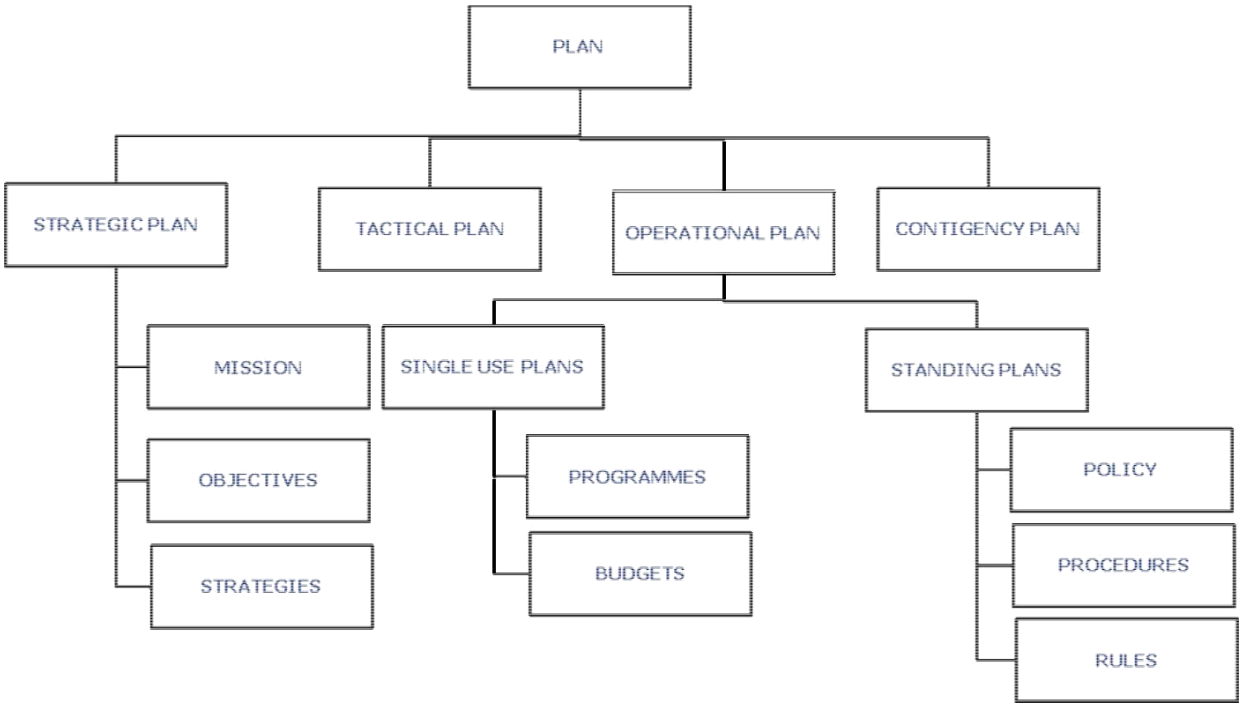
If this process is done well, budgets become a means of adding together the various plans and also important standards against which planning progress can be measured.

h) Establishing sequence of activities

Once plans that furnish the organization with both long-range and short-range direction have been developed, they must be implemented. Obviously, the organization can not directly benefit from planning process until this step is performed.

TYPES OF PLANS / COMPONENTS OF PLANNING

In the process of planning, several plans are prepared which are known as components of planning.



Plans can be broadly classified as

a) Strategic plans

b) Tactical plans

c) Operational plans

Operational plans lead to the achievement of tactical plans, which in turn lead to the attainment of strategic plans. In addition to these three types of plans, managers should also develop a contingency plan in case their original plans fail.

a) Strategic plans:

A strategic plan is an outline of steps designed with the goals of the entire organization as a whole in mind, rather than with the goals of specific divisions or departments. It is further classified as

i) Mission:

. The mission is a statement that reflects the basic purpose and focus of the organization which normally remain unchanged. The mission of the company is the answer of the question : why does the organization exists?

Properly crafted mission statements serve as filters to separate what is important from what is not, clearly state which markets will be served and how, and communicate a sense of intended direction to the entire organization.

Mission of Ford: "we are a global, diverse family with a proud inheritance, providing exceptional products and services".

ii) Objectives or goals:

Both goal and objective can be defined as statements that reflect the end towards which the organization is aiming to achieve. However, there are significant differences between the two. A goal is an abstract and general umbrella statement, under which specific objectives can be clustered. Objectives are statements that describe—in precise, measurable, and obtainable terms which reflect the desired organization's outcomes.

iii) Strategies:

Strategy is the determination of the basic long term objectives of an organization and the adoption of action and collection of action and allocation of resources necessary to achieve these goals.

Strategic planning begins with an organization's mission. Strategic plans look ahead over the next two, three, five, or even more years to move the organization from where it currently is to where it wants to be. Requiring multilevel involvement, these plans demand harmony among all levels of management within the organization. Top-level management develops the directional objectives for the entire organization, while lower levels of management develop compatible objectives and plans to achieve them. Top management's strategic plan for the entire organization becomes the framework and sets dimensions for the lower level planning.

b) Tactical plans:

A tactical plan is concerned with what the lower level units within each division must do, how they must do it, and who is in charge at each level. Tactics are the means needed to activate a strategy and make it work.

Tactical plans are concerned with shorter time frames and narrower scopes than are strategic plans. These plans usually span one year or less because they are considered short-term goals. Long-term goals, on the other hand, can take several years or more to accomplish. Normally, it

is the middle manager's responsibility to take the broad strategic plan and identify specific tactical actions.

c) Operational plans

The specific results expected from departments, work groups, and individuals are the operational goals. These goals are precise and measurable. "Process 150 sales applications each week" or "Publish 20 books this quarter" are examples of operational goals.

An operational plan is one that a manager uses to accomplish his or her job responsibilities. Supervisors, team leaders, and facilitators develop operational plans to support tactical plans (see the next section). Operational plans can be a single-use plan or a standing plan.

i) Single-use plans apply to activities that do not recur or repeat. A one-time occurrence, such as a special sales program, is a single-use plan because it deals with the who, what, where, how, and how much of an activity.

- **Programme**: Programme consists of an ordered list of events to be followed to execute a project.
- **Budget**: A budget predicts sources and amounts of income and how much they are used for a specific project.

ii) Standing plans are usually made once and retain their value over a period of years while undergoing periodic revisions and updates. The following are examples of ongoing plans:

- **Policy**: A policy provides a broad guideline for managers to follow when dealing with important areas of decision making. Policies are general statements that explain how a manager should attempt to handle routine management responsibilities. Typical human resources policies, for example, address such matters as employee hiring, terminations, performance appraisals, pay increases, and discipline.
- **Procedure**: A procedure is a set of step-by-step directions that explains how activities or tasks are to be carried out. Most organizations have procedures for purchasing supplies and equipment, for example. This procedure usually begins with a supervisor completing a purchasing requisition. The requisition is then sent to the next level of management for approval. The approved requisition is forwarded to the purchasing department. Depending on the amount of the request, the purchasing department may place an order, or they may need to secure quotations and/or bids for several vendors before placing the order. By defining the steps to be taken and

the order in which they are to be done, procedures provide a standardized way of responding to a repetitive problem.

- **Rule**: A rule is an explicit statement that tells an employee what he or she can and cannot do. Rules are "do" and "don't" statements put into place to promote the safety of employees and the uniform treatment and behavior of employees. For example, rules about tardiness and absenteeism permit supervisors to make discipline decisions rapidly and with a high degree of fairness.

d) Contingency plans

Intelligent and successful management depends upon a constant pursuit of adaptation, flexibility, and mastery of changing conditions. Strong management requires a "keeping all options open" approach at all times — that's where contingency planning comes in.

Contingency planning involves identifying alternative courses of action that can be implemented if and when the original plan proves inadequate because of changing circumstances.

Keep in mind that events beyond a manager's control may cause even the most carefully prepared alternative future scenarios to go awry. Unexpected problems and events frequently occur. When they do, managers may need to change their plans. Anticipating change during the planning process is best in case things don't go as expected. Management can then develop alternatives to the existing plan and ready them for use when and if circumstances make these alternatives appropriate.

OBJECTIVES

Objectives may be defined as the goals which an organisation tries to achieve. Objectives are described as the end- points of planning. According to Koontz and O'Donnell, "an objective is a term commonly used to indicate the end point of a management programme." Objectives constitute the purpose of the enterprise and without them no intelligent planning can take place.

Objectives are, therefore, the ends towards which the activities of the enterprise are aimed. They are present not only the end-point of planning but also the end towards which organizing, directing and controlling are aimed. Objectives provide direction to various activities. They also serve as the benchmark of measuring the efficiency and effectiveness of the enterprise. Objectives make every human activity purposeful. Planning has no meaning if it is not related to certain objectives.

Features of Objectives

- The objectives must be predetermined.
- A clearly defined objective provides the clear direction for managerial effort.
- Objectives must be realistic.
- Objectives must be measurable.
- Objectives must have social sanction.
- All objectives are interconnected and mutually supportive.
- Objectives may be short-range, medium-range and long-range.
- Objectives may be constructed into a hierarchy.

Advantages of Objectives

- Clear definition of objectives encourages unified planning.
- Objectives provide motivation to people in the organization.
- When the work is goal-oriented, unproductive tasks can be avoided.
- Objectives provide standards which aid in the control of human efforts in an organization.
- Objectives serve to identify the organization and to link it to the groups upon which its existence depends.
- Objectives act as a sound basis for developing administrative controls.
- Objectives contribute to the management process: they influence the purpose of the organization, policies, personnel, leadership as well as managerial control.

Process of Setting Objectives

Objectives are the keystone of management planning. It is the most important task of management. Objectives are required to be set in every area which directly and vitally effects the survival and prosperity of the business. In the setting of objectives, the following points should be borne in mind.

- Objectives are required to be set by management in every area which directly and vitally affects the survival and prosperity of the business.
- The objectives to be set in various areas have to be identified.
- While setting the objectives, the past performance must be reviewed, since past performance indicates what the organization will be able to accomplish in future.

- The objectives should be set in realistic terms i.e., the objectives to be set should be reasonable and capable of attainment.
- Objectives must be consistent with one and other.
- Objectives must be set in clear-cut terms.
- For the successful accomplishment of the objectives, there should be effective communication.

MANAGEMENT BY OBJECTIVES (MBO)

MBO was first popularized by Peter Drucker in 1954 in his book 'The practice of Management'. It is a process of agreeing within an organization so that management and employees buy into the objectives and understand what they are. It has a precise and written description objectives ahead, timelines for their motoring and achievement.

The employees and manager agree to what the employee will attempt to achieve in a period ahead and the employee will accept and buy into the objectives.

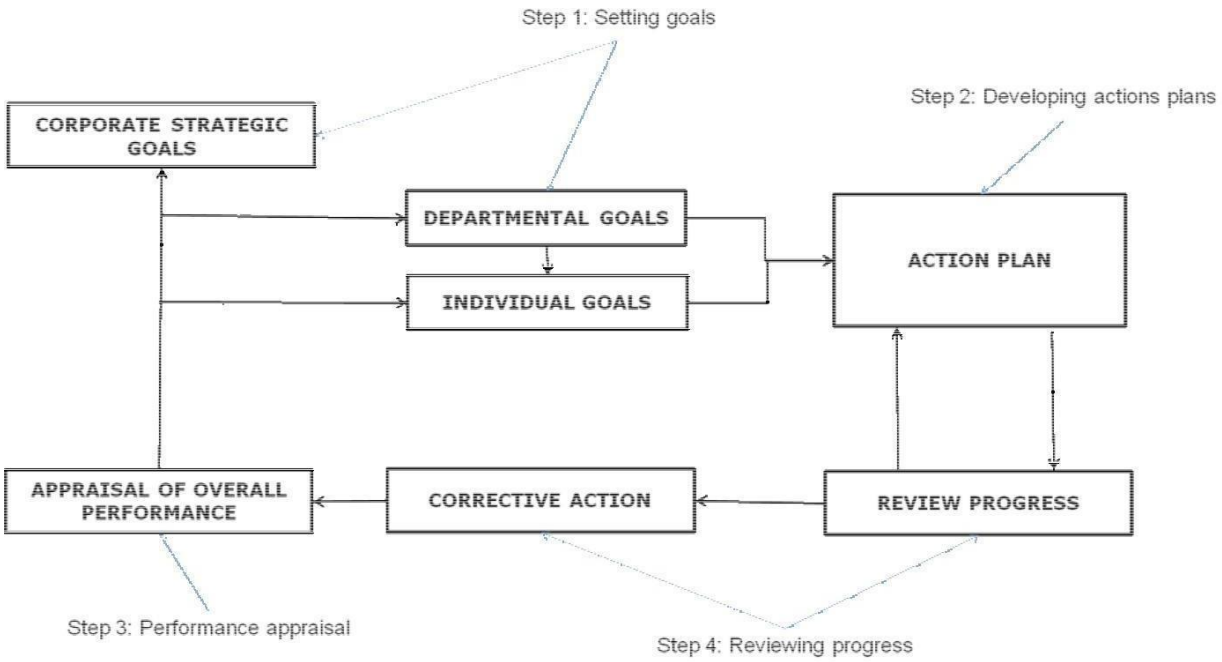
Definition

"MBO is a process whereby the superior and the mangers of an organization jointly identify its common goals, define each individual's major area of responsibility in terms of results expected of him, and use these measures as guides for operating the unit and assessing the contribution of each of its members."

Features of MBO

1. MBO is concerned with goal setting and planning for individual managers and their units.
2. The essence of MBO is a process of joint goal setting between a supervisor and a subordinate.
3. Managers work with their subordinates to establish the performance goals that are consistent with their higher organizational objectives.
4. MBO focuses attention on appropriate goals and plans.
5. MBO facilitates control through the periodic development and subsequent evaluation of individual goals and plans.

Steps in MBO:



The typical MBO process consists of:

1) Establishing a clear and precisely defined statement of objectives for the employee
2) Developing an action plan indicating how these objectives are to be achieved
3) Reviewing the performance of the employees
4) Appraising performance based on objective achievement

1) Setting objectives:

For Management by Objectives (MBO) to be effective, individual managers must understand the specific objectives of their job and how those objectives fit in with the overall company objectives set by the board of directors.

The managers of the various units or sub-units, or sections of an organization should know not only the objectives of their unit but should also actively participate in setting these objectives and make responsibility for them.

Management by Objective (MBO) systems, objectives are written down for each level of the organization, and individuals are given specific aims and targets.

Managers need to identify and set objectives both for themselves, their units, and their organizations.

2) Developing action plans

Actions plans specify the actions needed to address each of the top organizational issues and to reach each of the associated goals, who will complete each action and according to what timeline. An overall, top-level action plan that depicts how each strategic goal will be reached is developed by the top level management. The format of the action plan depends on the objective of the organization.

3) Reviewing Progress:

Performance is measured in terms of results. Job performance is the net effect of an employee's effort as modified by abilities, role perceptions and results produced. Effort refers to the amount of energy an employee uses in performing a job. Abilities are personal characteristics used in performing a job and usually do not fluctuate widely over short periods of time. Role perception refers to the direction in which employees believe they should channel their efforts on their jobs, and they are defined by the activities and behaviors they believe are necessary.

4) Performance appraisal:

Performance appraisals communicate to employees how they are performing their jobs, and they establish a plan for improvement. Performance appraisals are extremely important to both employee and employer, as they are often used to provide predictive information related to possible promotion. Appraisals can also provide input for determining both individual and organizational training and development needs. Performance appraisals encourage performance improvement. Feedback on behavior, attitude, skill or knowledge clarifies for employees the job expectations their managers hold for them. In order to be effective, performance appraisals must be supported by documentation and management commitment.

Advantages

- Motivation – Involving employees in the whole process of goal setting and increasing employee empowerment. This increases employee job satisfaction and commitment.
- Better communication and Coordination – Frequent reviews and interactions between superiors and subordinates helps to maintain harmonious relationships within the organization and also to solve many problems.

- Clarity of goals
- Subordinates have a higher commitment to objectives they set themselves than those imposed on them by another person.
- Managers can ensure that objectives of the subordinates are linked to the organization's objectives.

Limitations

There are several limitations to the assumptive base underlying the impact of managing by objectives, including:

- It over-emphasizes the setting of goals over the working of a plan as a driver of outcomes.
- It underemphasizes the importance of the environment or context in which the goals are set. That context includes everything from the availability and quality of resources, to relative buy-in by leadership and stake-holders.
- Companies evaluated their employees by comparing them with the "ideal" employee. Trait appraisal only looks at what employees should be, not at what they should do.

When this approach is not properly set, agreed and managed by organizations, self-centered employees might be prone to distort results, falsely representing achievement of targets that were set in a short-term, narrow fashion. In this case, managing by objectives would be counterproductive.

STRATEGIES

The term 'Strategy' has been adapted from war and is being increasingly used in business to reflect broad overall objectives and policies of an enterprise. Literally speaking, the term 'Strategy' stands for the war-art of the military general, compelling the enemy to fight as per out chosen terms and conditions.

According to Koontz and O' Donnell, "Strategies must often denote a general programme of action and deployment of emphasis and resources to attain comprehensive objectives". Strategies are plans made in the light of the plans of the competitors because a modern business institution operates in a competitive environment. They are a useful framework for guiding enterprise thinking and action. A perfect strategy can be built only on perfect knowledge of the plans of others in the industry. This may be done by the management of a firm putting itself in the place of a rival firm and trying to estimate their plans.

Characteristics of Strategy

- It is the right combination of different factors.
- It relates the business organization to the environment.
- It is an action to meet a particular challenge, to solve particular problems or to attain desired objectives.
- Strategy is a means to an end and not an end in itself.
- It is formulated at the top management level.
- It involves assumption of certain calculated risks.

Strategic Planning Process / Strategic Formulation Process

1. **Input to the Organization**: Various Inputs (People, Capital, Management and Technical skills, others) including goals input of claimants (Employees, Consumers, Suppliers, Stockholders, Government, Community and others)need to be elaborated.
2. **Industry Analysis**: Formulation of strategy requires the evaluation of the attractiveness of an industry by analyzing the external environment. The focus should be on the kind of compaction within an industry, the possibility of new firms entering the market, the availability of substitute products or services, the bargaining positions of the suppliers, and buyers or customers.
3. **Enterprise Profile**: Enterprise profile is usually the starting point for determining where the company is and where it should go. Top managers determine the basic purpose of the enterprise and clarify the firm's geographic orientation.
4. **Orientation, Values, and Vision of Executives**: The enterprise profile is shaped by people, especially executives, and their orientation and values are important for formulation the strategy. They set the organizational climate, and they determine the direction of the firm though their vision. Consequently, their values, their preferences, and their attitudes toward risk have to be carefully examined because they have an impact on the strategy.
5. **Mission (Purpose), Major Objectives, and Strategic Intent**: Mission or Purpose is the answer to the question: What is our business? The major Objectives are the end points towards which the activates of the enterprise are directed. Strategic intent is the commitment (obsession) to win in the competitive environment, not only at the top-level but also throughout the organization.
6. **Present and Future External Environment**: The present and future external environment must be assessed in terms of threats and opportunities.

7. **Internal Environment**: Internal Environment should be audited and evaluated with respect to its resources and its weaknesses, and strengths in research and development, production, operation, procurement, marketing and products and services. Other internal factors include, human resources and financial resources as well as the company image, the organization structure and climate, the planning and control system, and relations with customers.
8. **Development of Alternative Strategies**: Strategic alternatives are developed on the basis of an analysis of the external and internal environment. Strategies may be specialize or concentrate. Alternatively, a firm may diversify, extending the operation into new and profitable markets. Other examples of possible strategies are joint ventures, and strategic alliances which may be an appropriate strategy for some firms.
9. **Evaluation and Choice of Strategies**: Strategic choices must be considered in the light of the risk involved in a particular decision. Some profitable opportunities may not be pursued because a failure in a risky venture could result in bankruptcy of the firm. Another critical element in choosing a strategy is timing. Even the best product may fail if it is introduced to the market at an inappropriate time.
10. **Medium/Short Range Planning, Implementation through Reengineering the Organization Structure, Leadership and Control**: Implementation of the Strategy often requires reengineering the organization, staffing the organization structure and providing leadership. Controls must also be installed monitoring performance against plans.
11. **Consistency Testing and Contingency Planning**: The last key aspect of the strategic planning process is the testing for consistency and preparing for contingency plans.

TYPES OF STRATEGIES

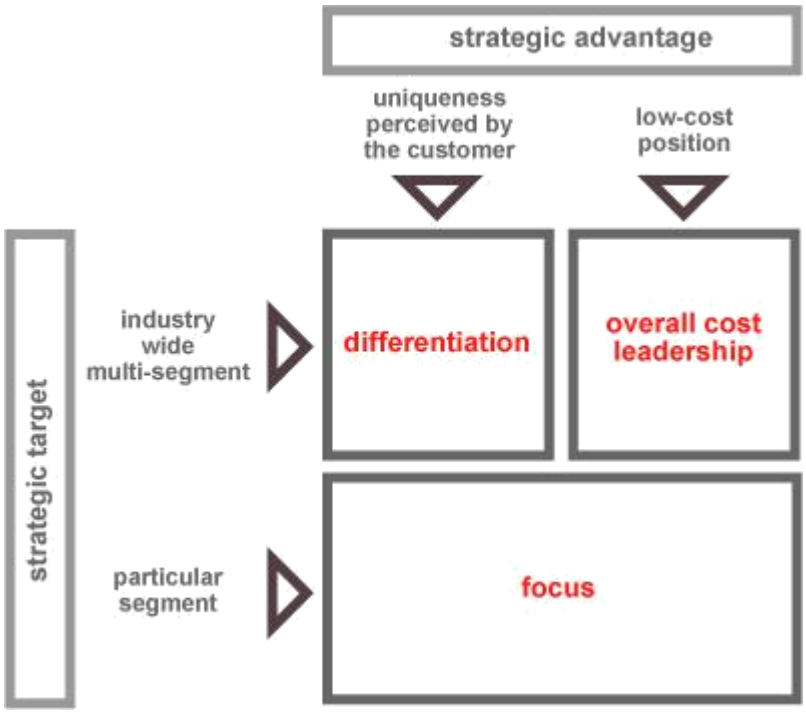
According to Michel Porter, the strategies can be classified into three types. They are

a) Cost leadership strategy

b) Differentiation strategy

c) Focus strategy

The following table illustrates Porter's generic strategies:



a) Cost Leadership Strategy

This generic strategy calls for being the low cost producer in an industry for a given level of quality. The firm sells its products either at average industry prices to earn a profit higher than that of rivals, or below the average industry prices to gain market share. In the event of a price war, the firm can maintain some profitability while the competition suffers losses. Even without a price war, as the industry matures and prices decline, the firms that can produce more cheaply will remain profitable for a longer period of time. The cost leadership strategy usually targets a broad market.

Some of the ways that firms acquire cost advantages are by improving process efficiencies, gaining unique access to a large source of lower cost materials, making optimal outsourcing and vertical integration decisions, or avoiding some costs altogether. If competing firms are unable to lower their costs by a similar amount, the firm may be able to sustain a competitive advantage based on cost leadership.

Firms that succeed in cost leadership often have the following internal strengths:

- Access to the capital required to make a significant investment in production assets; this investment represents a barrier to entry that many firms may not overcome.
- Skill in designing products for efficient manufacturing, for example, having a small component count to shorten the assembly process.
- High level of expertise in manufacturing process engineering.
- Efficient distribution channels.

Each generic strategy has its risks, including the low-cost strategy. For example, other firms may be able to lower their costs as well. As technology improves, the competition may be able to leapfrog the production capabilities, thus eliminating the competitive advantage. Additionally, several firms following a focus strategy and targeting various narrow markets may be able to achieve an even lower cost within their segments and as a group gain significant market share.

b) Differentiation Strategy

A differentiation strategy calls for the development of a product or service that offers unique attributes that are valued by customers and that customers perceive to be better than or different from the products of the competition. The value added by the uniqueness of the product may allow the firm to charge a premium price for it. The firm hopes that the higher price will more than cover the extra costs incurred in offering the unique product. Because of the product's unique attributes, if suppliers increase their prices the firm may be able to pass along the costs to its customers who cannot find substitute products easily.

Firms that succeed in a differentiation strategy often have the following internal strengths:

- Access to leading scientific research.
- Highly skilled and creative product development team.
- Strong sales team with the ability to successfully communicate the perceived strengths of the product.
- Corporate reputation for quality and innovation.

The risks associated with a differentiation strategy include imitation by competitors and changes in customer tastes. Additionally, various firms pursuing focus strategies may be able to achieve even greater differentiation in their market segments.

c) Focus Strategy

The focus strategy concentrates on a narrow segment and within that segment attempts to achieve either a cost advantage or differentiation. The premise is that the needs of the group can be better serviced by focusing entirely on it. A firm using a focus strategy often enjoys a high degree of customer loyalty, and this entrenched loyalty discourages other firms from competing directly.

Because of their narrow market focus, firms pursuing a focus strategy have lower volumes and therefore less bargaining power with their suppliers. However, firms pursuing a differentiation-focused strategy may be able to pass higher costs on to customers since close substitute products do not exist.

Firms that succeed in a focus strategy are able to tailor a broad range of product development strengths to a relatively narrow market segment that they know very well.

Some risks of focus strategies include imitation and changes in the target segments. Furthermore, it may be fairly easy for a broad-market cost leader to adapt its product in order to compete directly. Finally, other focusers may be able to carve out sub-segments that they can serve even better.

A Combination of Generic Strategies

These generic strategies are not necessarily compatible with one another. If a firm attempts to achieve an advantage on all fronts, in this attempt it may achieve no advantage at all. For example, if a firm differentiates itself by supplying very high quality products, it risks undermining that quality if it seeks to become a cost leader. Even if the quality did not suffer, the firm would risk projecting a confusing image. For this reason, Michael Porter argued that to be successful over the long-term, a firm must select only one of these three generic strategies. Otherwise, with more than one single generic strategy the firm will be "stuck in the middle" and will not achieve a competitive advantage.

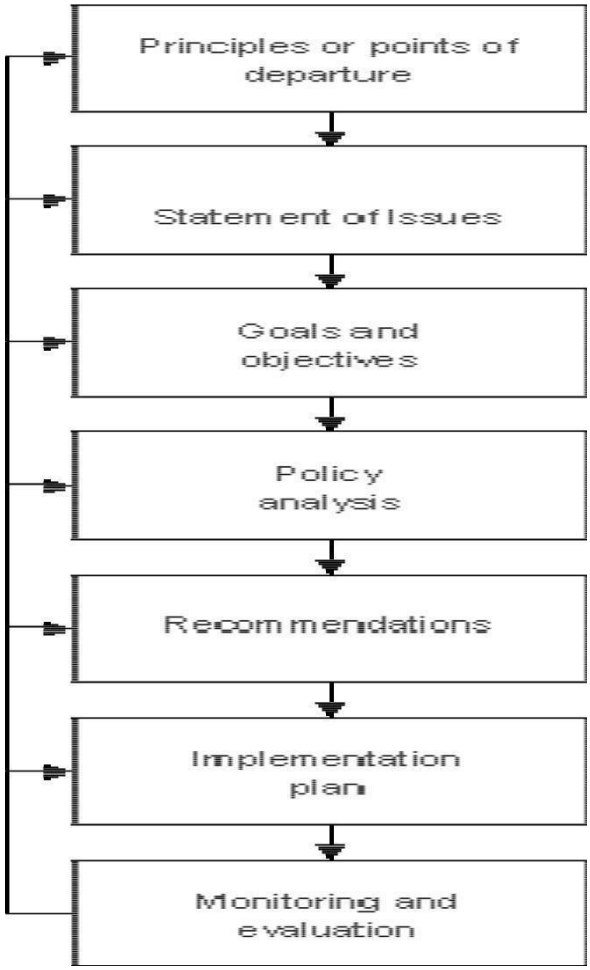
Porter argued that firms that are able to succeed at multiple strategies often do so by creating separate business units for each strategy. By separating the strategies into different units having different policies and even different cultures, a corporation is less likely to become "stuck in the middle."

However, there exists a viewpoint that a single generic strategy is not always best because within the same product customers often seek multi-dimensional satisfactions such as a combination of quality, style, convenience, and price. There have been cases in which high quality producers faithfully followed a single strategy and then suffered greatly when another

firm entered the market with a lower-quality product that better met the overall needs of the customers.

POLICIES

Policies are general statements or understandings that guide managers' thinking in decision making. They usually do not require action but are intended to guide managers in their commitment to the decision they ultimately make.



The first step in the process of policy formulation, as shown in the diagram below, is to capture the values or principles that will guide the rest of the process and form the basis on which to produce a statement of issues. The statement of issues involves identifying the opportunities and constraints affecting the local housing market, and is to be produced by

thoroughly analyzing the housing market. The kit provides the user with access to a housing data base to facilitate this analysis.
The statement of issues will provide the basis for the formulation of a set of housing goals and objectives, designed to address the problems identified and to exploit the opportunities which present themselves.

The next step is to identify and analyze the various policy options which can be applied to achieve the set of goals and objectives. The options available to each local government will depend on local circumstances as much as the broader context and each local authority will have to develop its own unique approach to addressing the housing needs of its residents.
An implementation program for realizing the policy recommendations must then be prepared, addressing budgetary and programming requirements, and allocating roles and responsibilities. Finally, the implementation of the housing strategy needs to be systematically monitored and evaluated against the stated goals and objectives, and the various components of the strategy modified or strengthened, as required.
At each step of the way, each component of the strategy needs to be discussed and debated, and a public consultation process engaged in. The extent of consultation and the participants involved will vary with each step.

Essentials of Policy Formulation

The essentials of policy formation may be listed as below:

- A policy should be definite, positive and clear. It should be understood by everyone in the organization.
- A policy should be translatable into the practices.
- A policy should be flexible and at the same time have a high degree of permanency.
- A policy should be formulated to cover all reasonable anticipatable conditions.
- A policy should be founded upon facts and sound judgment.
- A policy should conform to economic principles, statutes and regulations.
- A policy should be a general statement of the established rule.

Importance of Policies

Policies are useful for the following reasons:

- They provide guides to thinking and action and provide support to the subordinates.
- They delimit the area within which a decision is to be made.

- They save time and effort by pre-deciding problems and
- They permit delegation of authority to mangers at the lower levels.

DECISION MAKING

The word decision has been derived from the Latin word "decidere" which means "cutting off". Thus, decision involves cutting off of alternatives between those that are desirable and those that are not desirable.

In the words of George R. Terry, "Decision-making is the selection based on some criteria from two or more possible alternatives".

Characteristics of Decision Making

- Decision making implies that there are various alternatives and the most desirable alternative is chosen to solve the problem or to arrive at expected results.
- The decision-maker has freedom to choose an alternative.
- Decision-making may not be completely rational but may be judgemental and emotional.
- Decision-making is goal-oriented.
- Decision-making is a mental or intellectual process because the final decision is made by the decision-maker.
- A decision may be expressed in words or may be implied from behaviour.
- Choosing from among the alternative courses of operation implies uncertainty about the final result of each possible course of operation.
- Decision making is rational. It is taken only after a thorough analysis and reasoning and weighing the consequences of the various alternatives.

TYPES OF DECISIONS

a) Programmed and Non-Programmed Decisions: Herbert Simon has grouped organizational decisions into two categories based on the procedure followed. They are:

i) Programmed decisions: Programmed decisions are routine and repetitive and are made within the framework of organizational policies and rules. These policies and rules are established well in advance to solve recurring problems in the organization. Programmed decisions have short-run impact. They are, generally, taken at the lower level of management.

ii) Non-Programmed Decisions: Non-programmed decisions are decisions taken to meet non-repetitive problems. Non-programmed decisions are relevant for solving unique/ unusual problems in which various alternatives cannot be decided in advance. A common feature of non-programmed decisions is that they are novel and non-recurring and therefore, readymade solutions are not available. Since these decisions are of high importance and have long-term consequences, they are made by top level management.

b) Strategic and Tactical Decisions: Organizational decisions may also be classified as strategic or tactical.

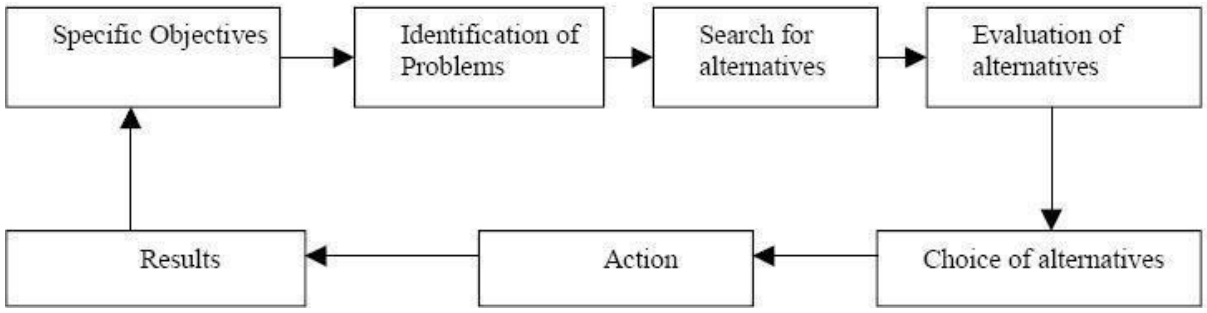
i) Strategic Decisions: Basic decisions or strategic decisions are decisions which are of crucial importance. Strategic decisions a major choice of actions concerning allocation of resources and contribution to the achievement of organizational objectives. Decisions like plant location, product diversification, entering into new markets, selection of channels of distribution, capital expenditure etc are examples of basic or strategic decisions.

ii) Tactical Decisions: Routine decisions or tactical decisions are decisions which are routine and repetitive. They are derived out of strategic decisions. The various features of a tactical decision are as follows:

- Tactical decision relates to day-to-day operation of the organization and has to be taken very frequently.
- Tactical decision is mostly a programmed one. Therefore, the decision can be made within the context of these variables.
- The outcome of tactical decision is of short-term nature and affects a narrow part of the organization.
- The authority for making tactical decisions can be delegated to lower level managers because: first, the impact of tactical decision is narrow and of short-term nature and Second, by delegating authority for such decisions to lower-level managers, higher level managers are free to devote more time on strategic decisions.

DECISION MAKING PROCESS

The decision making process is presented in the figure below:



1. Specific Objective: The need for decision making arises in order to achieve certain specific objectives. The starting point in any analysis of decision making involves the determination of whether a decision needs to be made.

2. Problem Identification: A problem is a felt need, a question which needs a solution. In the words of Joseph L Massie "A good decision is dependent upon the recognition of the right problem". The objective of problem identification is that if the problem is precisely and specifically identifies, it will provide a clue in finding a possible solution. A problem can be identified clearly, if managers go through diagnosis and analysis of the problem.

Diagnosis: Diagnosis is the process of identifying a problem from its signs and symptoms. A symptom is a condition or set of conditions that indicates the existence of a problem. Diagnosing the real problem implies knowing the gap between what is and what ought to be, identifying the reasons for the gap and understanding the problem in relation to higher objectives of the organization.

Analysis: Diagnosis gives rise to analysis. Analysis of a problem requires:

- Who would make decision?
- What information would be needed?
- From where the information is available?

3. Search for Alternatives: A problem can be solved in several ways; however, all the ways cannot be equally satisfying. Therefore, the decision maker must try to find out the various alternatives available in order to get the most satisfactory result of a decision. A decision maker can use several sources for identifying alternatives:

- His own past experiences
- Practices followed by others and
- Using creative techniques.

4. Evaluation of Alternatives: After the various alternatives are identified, the next step is to evaluate them and select the one that will meet the choice criteria. /the decision maker must check proposed alternatives against limits, and if an alternative does not meet them, he can discard it. Having narrowed down the alternatives which require serious consideration, the decision maker will go for evaluating how each alternative may contribute towards the objective supposed to be achieved by implementing the decision.

5. Choice of Alternative: The evaluation of various alternatives presents a clear picture as to how each one of them contribute to the objectives under question. A comparison is made among the likely outcomes of various alternatives and the best one is chosen.

6. Action: Once the alternative is selected, it is put into action. The actual process of decision making ends with the choice of an alternative through which the objectives can be achieved.

7. Results: When the decision is put into action, it brings certain results. These results must correspond with objectives, the starting point of decision process, if good decision has been made and implemented properly. Thus, results provide indication whether decision making and its implementation is proper.

Characteristics of Effective Decisions

An effective decision is one which should contain three aspects. These aspects are given below:

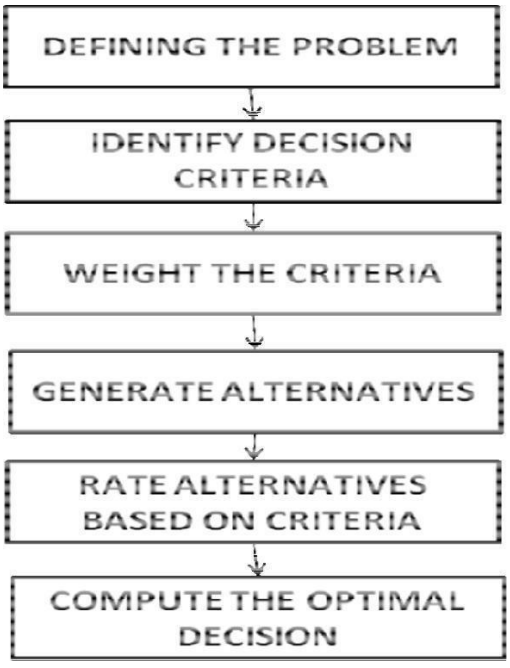
- **Action Orientation:** Decisions are action-oriented and are directed towards relevant and controllable aspects of the environment. Decisions should ultimately find their utility in implementation.
- **Goal Direction:** Decision making should be goal-directed to enable the organization to meet its objectives.
- **Effective in Implementation:** Decision making should take into account all the possible factors not only in terms of external context but also in internal context so that a decision can be implemented properly.

RATIONAL DECISION MAKING MODEL

The Rational Decision Making Model is a model which emerges from Organizational Behavior. The process is one that is logical and follows the orderly path from problem identification through solution. It provides a structured and sequenced approach to decision making. Using such an approach can help to ensure discipline and consistency is built into your decision making process.

The Six-Step Rational Decision-Making Model

1. Define the problem.
2. Identify decision criteria
3. Weight the criteria
4. Generate alternatives
5. Rate each alternative on each criterion
6. Compute the optimal decision



1) Defining the problem

This is the initial step of the rational decision making process. First the problem is identied and then defined to get a clear view of the situation.

2) Identify decision criteria

Once a decision maker has defined the problem, he or she needs to identify the decision criteria that will be important in solving the problem. In this step, the decision maker is determining what's relevant in making the decision.

This step brings the decision maker's interests, values, and personal preferences into the process.

Identifying criteria is important because what one person thinks is relevant, another may not.

Also keep in mind that any factors not identified in this step are considered as irrelevant to the decision maker.

3) Weight the criteria

The decision-maker weights the previously identified criteria in order to give them correct priority in the decision.

4) Generate alternatives

The decision maker generates possible alternatives that could succeed in resolving the problem. No attempt is made in this step to appraise these alternatives, only to list them.

5) Rate each alternative on each criterion

The decision maker must critically analyze and evaluate each one. The strengths and weakness of each alternative become evident as they compared with the criteria and weights established in second and third steps.

6) Compute the optimal decision

Evaluating each alternative against the weighted criteria and selecting the alternative with the highest total score.

DECISION MAKING UNDER VARIOUS CONDITIONS

The conditions for making decisions can be divided into three types. Namely a) Certainty, b) Uncertainty and c) Risk

Virtually all decisions are made in an environment to at least some uncertainty However; the degree will vary from relative certainty to great uncertainty. There are certain risks involved in making decisions.

a) Certainty:

In a situation involving certainty, people are reasonably sure about what will happen when they make a decision. The information is available and is considered to be reliable, and the cause and effect relationships are known.

b) Uncertainty

In a situation of uncertainty, on the other hand, people have only a meager database, they do not know whether or not the data are reliable, and they are very unsure about whether or not the situation may change.

Moreover, they cannot evaluate the interactions of the different variables. For example, a corporation that decides to expand its Operation to an unfamiliar country may know little about the country, culture, laws, economic environment, and politics. The political situation may be volatile that even experts cannot predict a possible change in government.

c) Risk

In a situation with risks, factual information may exist, but it may be incomplete. 1o improve decision making One may estimate the objective probability of an outcome by using, for example, mathematical models On the other hand, subjective probability, based on judgment and experience may be used

All intelligent decision makers dealing with uncertainty like to know the degree and nature of the risk they are taking in choosing a course of action. One of the deficiencies in using the traditional approaches of operations research for problem solving is that many of the data used in model are merely estimates and others are based on probabilities. The ordinary practice is to have staff specialists conic up with best estimates.

Virtually every decision is based on the interaction of a number of important variables, many of which has e an element of uncertainty but, perhaps, a fairly high degree of probability. Thus, the wisdom of launching a new product might depend on a number of critical variables: the cost of introducing the product, the cost of producing it, the capital investment that will he required, the price that can be set for the product, the size of the potential market, and the share of the total market that it will represent.

UNIT III

ORGANIZING

DEFINITION

According to Koontz and O'Donnell, "Organization involves the grouping of activities necessary to accomplish goals and plans, the assignment of these activities to appropriate departments and the provision of authority, delegation and co-ordination."

Organization involves division of work among people whose efforts must be co-ordinated to achieve specific objectives and to implement pre-determined strategies.

NATURE OR CHARACTERISTICS OF ORGANIZING

From the study of the various definitions given by different management experts we get the following information about the characteristics or nature of organization,

(1) Division of Work: Division of work is the basis of an organization. In other words, there can be no organization without division of work. Under division of work the entire work of business is divided into many departments .The work of every department is further sub-divided into sub-works. In this way each individual has to do the saran work repeatedly which gradually makes that person an expert.

(2) Coordination: Under organizing different persons are assigned different works but the aim of all these persons happens to be the some - the attainment of the objectives of the enterprise. Organization ensures that the work of all the persons depends on each other's work even though it happens to be different. The work of one person starts from where the work of another person ends. The non-completion of the work of one person affects the work of everybody. Therefore, everybody completes his work in time and does not hinder the work of others. It is thus, clear that it is in the nature of an organization to establish coordination among different works, departments and posts in the enterprise.

(3) Plurality of Persons: Organization is a group of many persons who assemble to fulfill a common purpose. A single individual cannot create an organization.

(4) Common Objectives: There are various parts of an organization with different functions to perform but all move in the direction of achieving a general objective.

(5) Well-defined Authority and Responsibility: Under organization a chain is established between different posts right from the top to the bottom. It is clearly specified as to what will be

the authority and responsibility of every post. In other words, every individual working in the organization is given some authority for the efficient work performance and it is also decided simultaneously as to what will be the responsibility of that individual in case of unsatisfactory work performance.

(6) Organization is a Structure of Relationship: Relationship between persons working on different posts in the organization is decided. In other words, it is decided as to who will be the superior and who will be the subordinate. Leaving the top level post and the lowest level post everybody is somebody's superior and somebody's subordinate. The person working on the top level post has no superior and the person working on the lowest level post has no subordinate.

(7) Organization is a Machine of Management: Organization is considered to be a machine of management because the efficiency of all the functions depends on an effective organization. In the absence of organization no function can be performed in a planned manner. It is appropriate to call organization a machine of management from another point of view. It is that machine in which no part can afford tube ill-fitting or non-functional. In other words, if the division of work is not done properly or posts are not created correctly the whole system of management collapses.

(8) Organization is a Universal Process: Organization is needed both in business and non-business organizations. Not only this, organization will be needed where two or mom than two people work jointly. Therefore, organization has the quality of universality. (9) Organization is a Dynamic Process: Organization is related to people and the knowledge and experience of the people undergo a change. The impact of this change affects the various functions of the organizations. Thus, organization is not a process that can be decided for all times to come but it undergoes changes according to the needs. The example in this case can be the creation or abolition of a new post according to the need.

IMPORTANCE OR ADVANTAGES OF ORGANIZING

Organization is an instrument that defines relations among different people which helps them to understand as in who happens to be their superior and who is their subordinate. This information helps in fixing responsibility and developing coordination. In such circumstances the objectives of the organization can be easily achieved. That is why, it is said that Organization Is a mechanism of management. In addition to that it helps in the other functions of management like planning, staffing, leading, controlling, etc. The importance of organization or its merits becomes clear from the following facts,

(1) Increase In Managerial Efficiency: A good and balanced organization helps the managers to increase their efficiency. Managers, through the medium of organization, make a proper distribution of the whole work among different people according to their ability.

(2) Proper Utilization of Resources: Through the medium of organization optimum utilization of all the available human and material resources of an enterprise becomes possible. Work is allotted to every individual according to his ability and capacity and conditions ant created to enable him to utilize his ability to the maximum extent. For example, if an employee possesses the knowledge of modem machinery but the modem machinery is not available in the organization, in that case, efforts are made to make available the modem machinery.

(3) Sound Communication Possible: Communication is essential for taking the right decision at the right time. However, the establishment of a good communication system is possible only through an organization. In an organization the time of communication is decided so that all the useful information reaches the officers concerned which. in turn, helps the decision-making.

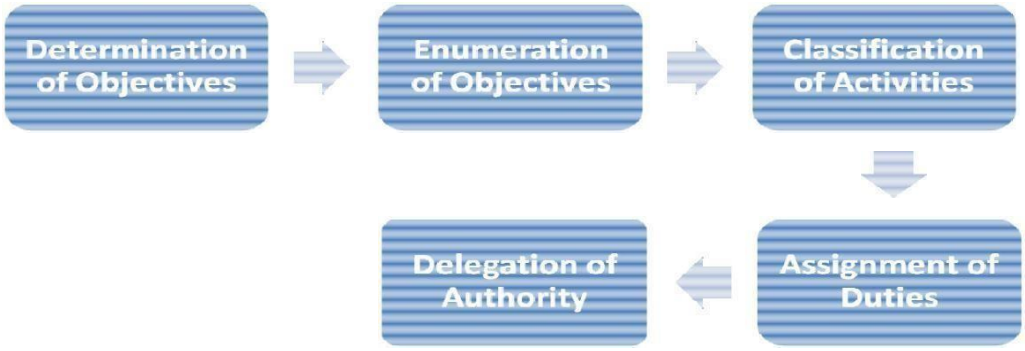
(4) Facilitates Coordination: In order to attain successfully the objectives of the organization, coordination among various activities in the organization is essential. Organization is the only medium which makes coordination possible. Under organization the division of work is made in such a manner as to make all the activities complementary to each other increasing their inter-dependence. Inter-dependence gives rise to the establishment of relations which, in turn, increases coordination.

(5) Increase in Specialization: Under organization the whole work is divided into different parts. Competent persons are appointed to handle all the sub-works and by handling a particular work repeatedly they become specialists. This enables them to have maximum work performance in the minimum time while the organization gets the benefit of specialization.

(6) Helpful in Expansion: A good organization helps the enterprise in facing competition. When an enterprise starts making available good quality product at cheap rates, it increases the demand for its products. In order to meet the increasing demand for its products an organization has to expand its business. On the other hand, a good organization has an element of flexibility which far from impeding the expansion work encourages it.

ORGANIZING PROCESS

Organization is the process of establishing relationship among the members of the enterprise. The relationships are created in terms of authority and responsibility. To organize is to harmonize, coordinate or arrange in a logical and orderly manner. Each member in the organization is assigned a specific responsibility or duty to perform and is granted the corresponding authority to perform his duty. The managerial function of organizing consists in making a rational division of work into groups of activities and tying together the positions representing grouping of activities so as to achieve a rational, well coordinated and orderly structure for the accomplishment of work. According to Louis A Allen, "Organizing involves identification and grouping the activities to be performed and dividing them among the individuals and creating authority and responsibility relationships among them for the accomplishment of organizational objectives." The various steps involved in this process are:



a) Determination of Objectives:

It is the first step in building up an organization. Organization is always related to certain objectives. Therefore, it is essential for the management to identify the objectives before starting any activity. Organization structure is built on the basis of the objectives of the enterprise. That means, the structure of the organization can be determined by the management only after knowing the objectives to be accomplished through the organization. This step helps the management not only in framing the organization structure but also in achieving the enterprise objectives with minimum cost and efforts. Determination of objectives will consist in deciding as to why the proposed organization is to be set up and, therefore, what will be the nature of the work to be accomplished through the organization.

b) Enumeration of Objectives:

If the members of the group are to pool their efforts effectively, there must be proper division of the major activities. The first step in organizing group effort is the division of the total job into essential activities. Each job should be properly classified and grouped. This will enable the people to know what is expected of them as members of the group and will help in avoiding duplication of efforts. For example, the work of an industrial concern may be divided into the following major functions – production, financing, personnel, sales, purchase, etc.

c) Classification of Activities:

The next step will be to classify activities according to similarities and common purposes and functions and taking the human and material resources into account. Then, closely related and similar activities are grouped into divisions and departments and the departmental activities are further divided into sections.

d) Assignment of Duties:

Here, specific job assignments are made to different subordinates for ensuring a certainty of work performance. Each individual should be given a specific job to do according to his ability and made responsible for that. He should also be given the adequate authority to do the job assigned to him. In the words of Kimball and Kimball - "Organization embraces the duties of designating the departments and the personnel that are to carry on the work, defining their functions and specifying the relations that are to exist between department and individuals."

e) Delegation of Authority:

Since so many individuals work in the same organization, it is the responsibility of management to lay down structure of relationship in the organization. Authority without responsibility is a dangerous thing and similarly responsibility without authority is an empty vessel. Everybody should clearly know to whom he is accountable; corresponding to the responsibility authority is delegated to the subordinates for enabling them to show work performance. This will help in the smooth working of the enterprise by facilitating delegation of responsibility and authority.

ORGANIZATION STRUCTURE

An organization structure is a framework that allots a particular space for a particular department or an individual and shows its relationship to the other. An organization structure shows the authority and responsibility relationships between the various positions in the organization by showing who reports to whom. It is an established pattern of relationship among the components of the organization.

March and Simon have stated that-"Organization structure consists simply of those aspects of pattern of behavior in the organization that are relatively stable and change only slowly." The structure of an organization is generally shown on an organization chart. It shows the authority and responsibility relationships between various positions in the organization while designing the organization structure, due attention should be given to the principles of sound organization.

Significance of Organization Structure

- Properly designed organization can help improve teamwork and productivity by providing a framework within which the people can work together most effectively.
- Organization structure determines the location of decision-making in the organization.
- Sound organization structure stimulates creative thinking and initiative among organizational members by providing well defined patterns of authority.
- A sound organization structure facilitates growth of enterprise by increasing its capacity to handle increased level of authority.
- Organization structure provides the pattern of communication and coordination.
- The organization structure helps a member to know what his role is and how it relates to other roles.

PRINCIPLES OF ORGANIZATION STRUCTURE

Modern organizational structures have evolved from several organizational theories, which have identified certain principles as basic to any organization structure.

a) Line and Staff Relationships:

Line authority refers to the scalar chain, or to the superior-subordinate linkages, that extend throughout the hierarchy (Koontz, O'Donnell and Weihrich). Line employees are responsible for achieving the basic or strategic objectives of the organization, while staff plays a supporting role to line employees and provides services. The relationship between line and staff

is crucial in organizational structure, design and efficiency. It is also an important aid to information processing and coordination.

b) Departmentalization:

Departmentalization is a process of horizontal clustering of different types of functions and activities on any one level of the hierarchy. Departmentalization is conventionally based on purpose, product, process, function, personal things and place.

c) Span of Control:

This refers to the number of specialized activities or individuals supervised by one person. Deciding the span of control is important for coordinating different types of activities effectively.

d) De-centralization and Centralization:

De-centralization refers to decision making at lower levels in the hierarchy of authority. In contrast, decision making in a centralized type of organizational structure is at higher levels. The degree of centralization and de-centralization depends on the number of levels of hierarchy, degree of coordination, specialization and span of control.

Every organizational structure contains both centralization and de-centralization, but to varying degrees. The extent of this can be determined by identifying how much of the decision making is concentrated at the top and how much is delegated to lower levels. Modern organizational structures show a strong tendency towards de-centralization.

FORMAL AND INFORMAL ORGANIZATION

The formal organization refers to the structure of jobs and positions with clearly defined functions and relationships as prescribed by the top management. This type of organization is built by the management to realize objectives of an enterprise and is bound by rules, systems and procedures. Everybody is assigned a certain responsibility for the performance of the given task and given the required amount of authority for carrying it out. Informal organization, which does not appear on the organization chart, supplements the formal organization in achieving organizational goals effectively and efficiently. The working of informal groups and leaders is not as simple as it may appear to be. Therefore, it is obligatory for every manager to study thoroughly the working pattern of informal relationships in the organization and to use them for achieving organizational objectives.

FORMAL ORGANIZATION

Chester I Bernard defines formal organization as -"a system of consciously coordinated activities or forces of two or more persons. It refers to the structure of well-defined jobs, each bearing a definite measure of authority, responsibility and accountability." The essence of formal organization is conscious common purpose and comes into being when persons–

(i) Are able to communicate with each other

(ii) Are willing to act and

(iii) Share a purpose.

- Division of labor
- Scalar and functional processes
- Structure and
- Span of control

Thus, a formal organization is one resulting from planning where the pattern of structure has already been determined by the top management.

Characteristic Features of formal organization

- Formal organization structure is laid down by the top management to achieve organizational goals.
- Formal organization prescribes the relationships amongst the people working in the organization.
- The organization structures is consciously designed to enable the people of the organization to work together for accomplishing the common objectives of the enterprise
- Organization structure concentrates on the jobs to be performed and not the individuals who are to perform jobs.
- In a formal organization, individuals are fitted into jobs and positions and work as per the managerial decisions. Thus, the formal relations in the organization arise from the pattern of responsibilities that are created by the management.
- A formal organization is bound by rules, regulations and procedures.
- In a formal organization, the position, authority, responsibility and accountability of each level are clearly defined.
- Organization structure is based on division of labor and specialization to achieve efficiency in operations.

- A formal organization is deliberately impersonal. The organization does not take into consideration the sentiments of organizational members.
- The authority and responsibility relationships created by the organization structure are to be honored by everyone.
- In a formal organization, coordination proceeds according to the prescribed pattern.

Advantages of formal organization

- The formal organization structure concentrates on the jobs to be performed. It, therefore, makes everybody responsible for a given task.
- A formal organization is bound by rules, regulations and procedures. It thus ensures law and order in the organization.
- The organization structure enables the people of the organization to work together for accomplishing the common objectives of the enterprise

Disadvantages or criticisms of formal organization

- The formal organization does not take into consideration the sentiments of organizational members.
- The formal organization does not consider the goals of the individuals. It is designed to achieve the goals of the organization only.
- The formal organization is bound by rigid rules, regulations and procedures. This makes the achievement of goals difficult.

INFORMAL ORGANIZATION

Informal organization refers to the relationship between people in the organization based on personal attitudes, emotions, prejudices, likes, dislikes etc. an informal organization is an organization which is not established by any formal authority, but arises from the personal and social relations of the people. These relations are not developed according to procedures and regulations laid down in the formal organization structure; generally large formal groups give rise to small informal or social groups. These groups may be based on same taste, language, culture or some other factor. These groups are not pre-planned, but they develop automatically within the organization according to its environment.

Characteristics features of informal organization

- Informal organization is not established by any formal authority. It is unplanned and arises spontaneously.
- Informal organizations reflect human relationships. It arises from the personal and social relations amongst the people working in the organization.
- Formation of informal organizations is a natural process. It is not based on rules, regulations and procedures.
- The inter-relations amongst the people in an informal organization cannot be shown in an organization chart.
- In the case of informal organization, the people cut across formal channels of communications and communicate amongst themselves.
- The membership of informal organizations is voluntary. It arises spontaneously and not by deliberate or conscious efforts.
- Membership of informal groups can be overlapping as a person may be member of a number of informal groups.
- Informal organizations are based on common taste, problem, language, religion, culture, etc. it is influenced by the personal attitudes, emotions, whims, likes and dislikes etc. of the people in the organization.

Benefits of Informal organization

- It blends with the formal organization to make it more effective.
- Many things which cannot be achieved through formal organization can be achieved through informal organization.
- The presence of informal organization in an enterprise makes the managers plan and act more carefully.
- Informal organization acts as a means by which the workers achieve a sense of security and belonging. It provides social satisfaction to group members.
- An informal organization has a powerful influence on productivity and job satisfaction.
- The informal leader lightens the burden of the formal manager and tries to fill in the gaps in the manager's ability.
- Informal organization helps the group members to attain specific personal objectives.
- Informal organization is the best means of employee communication. It is very fast.

- Informal organization gives psychological satisfaction to the members. It acts as a safety valve for the emotional problems and frustrations of the workers of the organization because they get a platform to express their feelings.
- It serves as an agency for social control of human behavior.

DIFFERENCES BETWEEN FORMAL AND INFORMAL ORGANIZATION

Formal Organization	**Informal Organization**
1. Formal organization is established with the explicit aim of achieving well-defined goals.	1. Informal organization springs on its own. Its goals are ill defined and intangible.
2. Formal organization is bound together by authority relationships among members. A hierarchical structure is created, constituting top management, middle management and supervisory management.	2. Informal organization is characterized by a generalized sort of power relationships. Power in informal organization has bases other than rational legal right.
3. Formal organization recognizes certain tasks which are to be carried out to achieve its goals.	3. Informal organization does not have any well-defined tasks.
4. The roles and relationships of people in formal organization are impersonally defined	4. In informa organization the relationships among people are interpersonal.
5. In formal organization, much emphasis is placed on efficiency, discipline, conformity, consistency and control.	5. Informal organization is characterized by relative freedom, spontaneity, by relative freedom, spontaneity, homeliness and warmth.
6. In formal organization, the social and psychological needs and interests of members of the organization get little attention.	6. In informa organization the sociopsychological needs, interests and aspirations of members get priority.
7. The communication system in formal organization follows certain pre-determined patterns and paths.	7. In informal organization, The communication pattern is haphazard, intricate and natural.
8. Formal organization is relatively slow to respond and adapt to changing situations and realities.	8. Informal organization is dynamic and very vigilant. It is sensitive to its surroundings.

LINE AND STAFF AUTHORITY

In an organization, the line authority flows from top to bottom and the staff authority is exercised by the specialists over the line managers who advise them on important matters. These specialists stand ready with their specialty to serve line mangers as and when their services are called for, to collect information and to give help which will enable the line officials to carry out their activities better. The staff officers do not have any power of command in the organization as they are employed to provide expert advice to the line officers. The 'line' maintains discipline and stability; the 'staff' provides expert information. The line gets out the production, the staffs carries on the research, planning, scheduling, establishing of standards and recording of performance. The authority by which the staff performs these functions is delegated by the line and the performance must be acceptable to the line before action is taken. The following figure depicts the line and staff authority:

Types of Staff

The staff position established as a measure of support for the line managers may take the following forms:

1. Personal Staff: Here the staff official is attached as a personal assistant or adviser to the line manager. For example, Assistant to managing director.
2. Specialized Staff: Such staff acts as the fountainhead of expertise in specialized areas like R & D, personnel, accounting etc.
3. General Staff: This category of staff consists of a set of experts in different areas who are meant to advise and assist the top management on matters called for expertise. For example, Financial advisor, technical advisor etc.

Features of line and staff organization

- Under this system, there are line officers who have authority and command over the subordinates and are accountable for the tasks entrusted to them. The staff officers are specialists who offer expert advice to the line officers to perform their tasks efficiently.
- Under this system, the staff officers prepare the plans and give advice to the line officers and the line officers execute the plan with the help of workers.
- The line and staff organization is based on the principle of specialization.

Advantages

- It brings expert knowledge to bear upon management and operating problems. Thus, the line managers get the benefit of specialized knowledge of staff specialists at various levels.
- The expert advice and guidance given by the staff officers to the line officers benefit the entire organization.
- As the staff officers look after the detailed analysis of each important managerial activity, it relieves the line managers of the botheration of concentrating on specialized functions.
- Staff specialists help the line managers in taking better decisions by providing expert advice. Therefore, there will be sound managerial decisions under this system.
- It makes possible the principle of undivided responsibility and authority, and at the same time permits staff specialization. Thus, the organization takes advantage of functional organization while maintaining the unity of command.
- It is based upon planned specialization.
- Line and staff organization has greater flexibility, in the sense that new specialized activities can be added to the line activities without disturbing the line procedure.

Disadvantages

- Unless the duties and responsibilities of the staff members are clearly indicated by charts and manuals, there may be considerable confusion throughout the organization as to the functions and positions of staff members with relation to the line supervisors.
- There is generally a conflict between the line and staff executives. The line managers feel that staff specialists do not always give right type of advice, and staff officials generally complain that their advice is not properly attended to.
- Line managers sometimes may resent the activities of staff members, feeling that prestige and influence of line managers suffer from the presence of the specialists.
- The staff experts may be ineffective because they do not get the authority to implement their recommendations.
- This type of organization requires the appointment of large number of staff officers or experts in addition to the line officers. As a result, this system becomes quite expensive.
- Although expert information and advice are available, they reach the workers through the officers and thus run the risk of misunderstanding and misinterpretation.

- Since staff managers are not accountable for the results, they may not be performing their duties well.
- Line mangers deal with problems in a more practical manner. But staff officials who are specialists in their fields tend to be more theoretical. This may hamper coordination in the organization.

DEPARTMENTATION BY DIFFERENT STRATEGIES

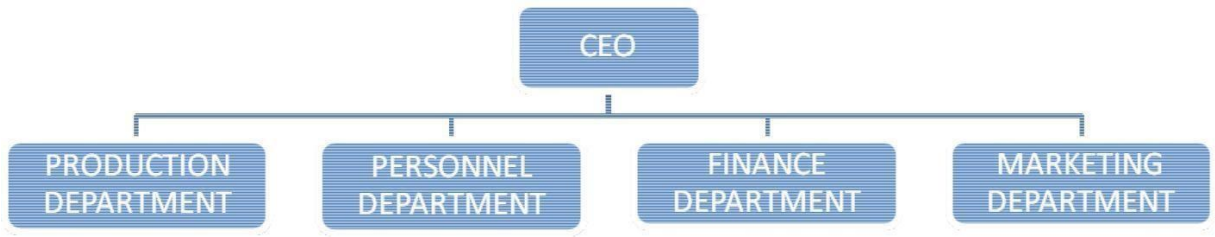
DEPARTMENTATION refers to the process of grouping activities into departments. Departmentation is the process of grouping of work activities into departments, divisions, and other homogenous units.

Key Factors in Departmentation

- It should facilitate control.
- It should ensure proper coordination.
- It should take into consideration the benefits of specialization.
- It should not result in excess cost.
- It should give due consideration to Human Aspects.

Departmentation takes place in various patterns like departmentation by functions, products, customers, geographic location, process, and its combinations.

a) FUNCTIONAL DEPARTMENTATION



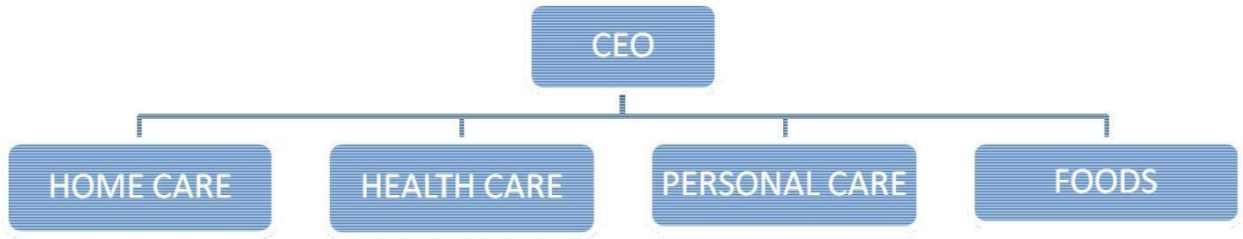
Functional departmentation is the process of grouping activities by functions performed. Activities can be grouped according to function (work being done) to pursue economies of scale by placing employees with shared skills and knowledge into departments for example human resources, finance, production, and marketing. Functional departmentation can be used in all types of organizations.

Advantages:

- Advantage of specialization
- Easy control over functions
- Pinpointing training needs of manager
- It is very simple process of grouping activities.

- Lack of responsibility for the end result
- Overspecialization or lack of general management
- It leads to increase conflicts and coordination problems among departments.

b) PRODUCT DEPARTMENTATION



Product departmentation is the process of grouping activities by product line. Tasks can also be grouped according to a specific product or service, thus placing all activities related to the product or the service under one manager. Each major product area in the corporation is under the authority of a senior manager who is specialist in, and is responsible for, everything related to the product line. Dabur India Limited is the India's largest Ayurvedic medicine manufacturer is an example of company that uses product departmentation. Its structure is based on its varied product lines which include Home care, Health care, Personal care and Foods.

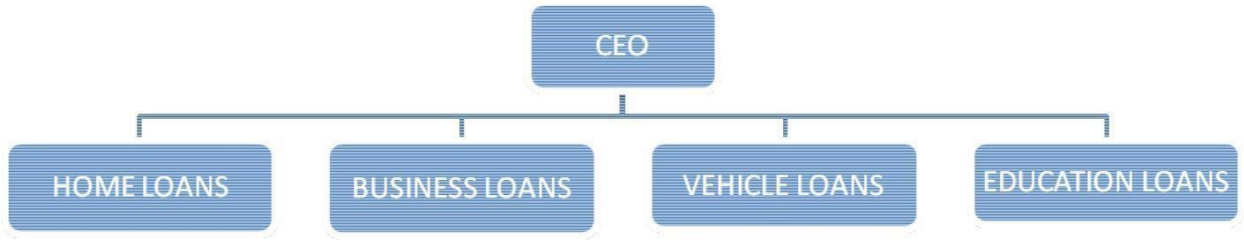
Advantages

- It ensures better customer service
- Unprofitable products may be easily determined
- It assists in development of all around managerial talent
- Makes control effective
- It is flexible and new product line can be added easily.

Disadvantages

- It is expensive as duplication of service functions occurs in various product divisions
- Customers and dealers have to deal with different persons for complaint and information of different products.

c) CUSTOMER DEPARTMENTATION



Customer departmentation is the process of grouping activities on the basis of common customers or types of customers. Jobs may be grouped according to the type of customer served by the organization. The assumption is that customers in each department have a common set of problems and needs that can best be met by specialists. UCO is the one of the largest commercial banks of India is an example of company that uses customer departmentation. Its structure is based on various services which includes Home loans, Business loans, Vehicle loans and Educational loans.

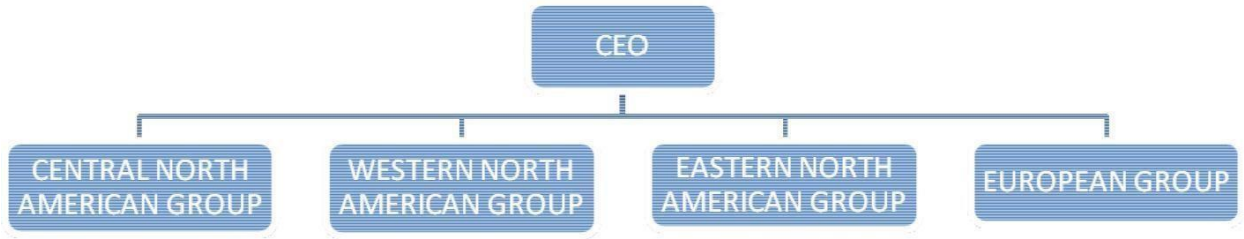
Advantages

- It focused on customers who are ultimate suppliers of money
- Better service to customer having different needs and tastes
- Development in general managerial skills

Disadvantages

- Sales being the exclusive field of its application, co-ordination may appear difficult between sales function and other enterprise functions.
- Specialized sales staff may become idle with the downward movement of sales to any specified group of customers.

d) GEOGRAPHIC DEPARTMENTATION



Geographic departmentation is the process of grouping activities on the basis of territory. If an organization's customers are geographically dispersed, it can group jobs based on geography. For example, the organization structure of Coca-Cola Ltd has reflected the company's operation in various geographic areas such as Central North American group, Western North American group, Eastern North American group and European group

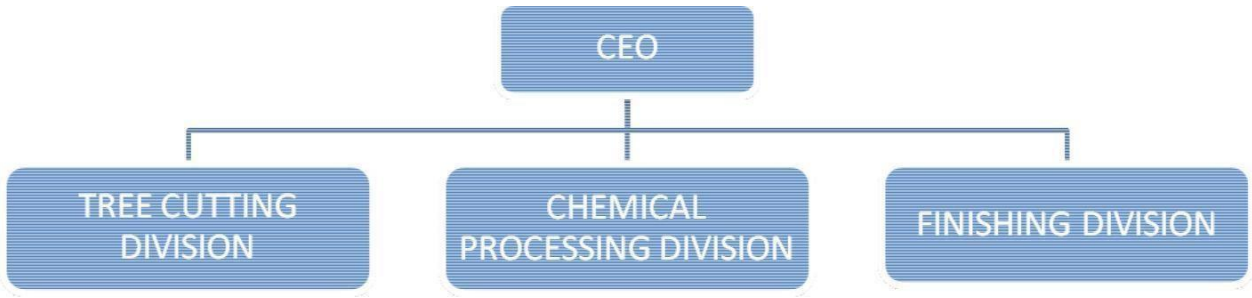
Advantages

- Help to cater to the needs of local people more satisfactorily.
- It facilitates effective control
- Assists in development of all-round managerial skills

Disadvantages

- Communication problem between head office and regional office due to lack of means of communication at some location
- Coordination between various divisions may become difficult.
- Distance between policy framers and executors
- It leads to duplication of activities which may cost higher.

e) PROCESS DEPARTMENTATION



Geographic departmentation is the process of grouping activities on the basis of product or service or customer flow. Because each process requires different skills, process departmentation allows homogenous activities to be categorized. For example, Bowater Thunder Bay, a Canadian company that harvests trees and processes wood into newsprint and pulp. Bowater has three divisions namely tree cutting, chemical processing, and finishing (which makes newsprint).

Departmentation by process: -

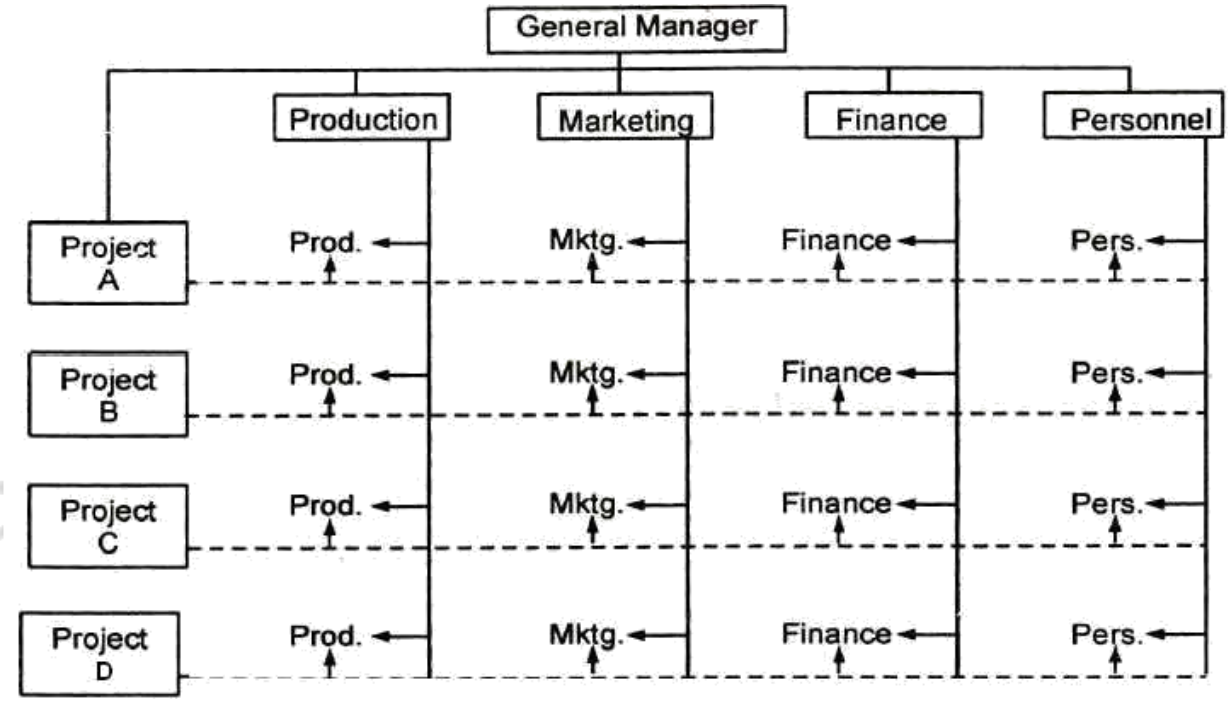
Advantages

- Oriented towards end result.
- Professional identification is maintained.
- Pinpoints product-profit responsibility.

Disadvantage

- Conflict in organization authority exists.
- Possibility of disunity of command.
- Requires managers effective in human relation

f) MARTIX DEPARTMENTATION



In actual practice, no single pattern of grouping activities is applied in the organization structure with all its levels. Different bases are used in different segments of the enterprise. Composite or hybrid method forms the common basis for classifying activities rather than one particular method,. One of the mixed forms of organization is referred to as matrix or grid organization's According to the situations, the patterns of Organizing varies from case to case. The form of structure must reflect the tasks, goals and technology if the originations the type of people employed and the environmental conditions that it faces. It is not unusual to see firms that utilize the function and project organization combination. The same is true for process and project as well as other combinations. For instance, a large hospital could have an accounting department, surgery department, marketing department, and a satellite center project team that make up its organizational structure.

Advantages

- Efficiently manage large, complex tasks
- Effectively carry out large, complex tasks

Disadvantages

- Requires high levels of coordination
- Conflict between bosses
- Requires high levels of management skills

SPAN OF CONTROL

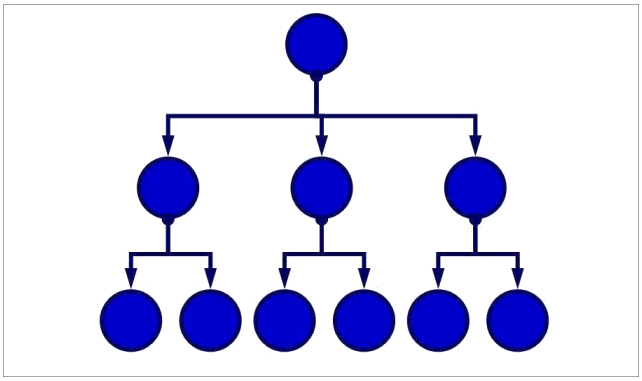
Span of Control means the number of subordinates that can be managed efficiently and effectively by a superior in an organization. It suggests how the relations are designed between a superior and a subordinate in an organization.

Factors Affecting Span of control:

a) Capacity of Superior:
   Different ability and capacity of leadership, communication affect management of subordinates.
b) Capacity of Subordinates:
   Efficient and trained subordinates affects the degree of span of management.
c) Nature of Work:
   Different types of work require different patterns of management.
d) Degree of Centralization or Decentralization:
   Degree of centralization or decentralization affects the span of management by affecting the degree of involvement of the superior in decision making.
e) Degree of Planning:
   Plans which can provide rules, procedures in doing the work higher would be the degree of span of management.
f) Communication Techniques:
   Pattern of communication, its means, and media affect the time requirement in managing subordinates and consequently span of management.
g) Use of Staff Assistance:
   Use of Staff assistance in reducing the work load of managers enables them to manage more number of subordinates.
h) Supervision of others:
   If subordinate receives supervision form several other personnel besides his direct supervisor. In such a case, the work load of direct superior is reduced and he can supervise more number of persons.

Span of control is of two types:

1. Narrow span of control: Narrow Span of control means a single manager or supervisor oversees few subordinates. This gives rise to a tall organizational structure.



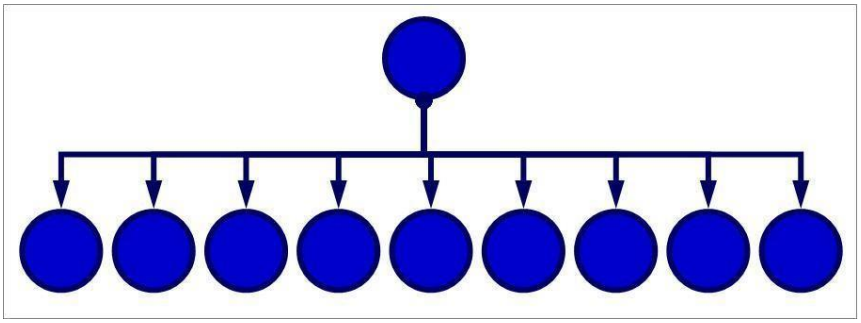
Advantages:

- Close supervision
- Close control of subordinates
- Fast communication

Disadvantages:

- Too much control
- Many levels of management
- High costs
- Excessive distance between lowest level and highest level

2. Wide span of control: Wide span of control means a single manager or supervisor oversees a large number of subordinates. This gives rise to a flat organizational structure.



Advantages:

- More Delegation of Authority
- Development of Managers
- Clear policies

Disadvantages:

- Overloaded supervisors

- Danger of superiors loss of control
- Requirement of highly trained managerial personnel
- Block in decision making

CENTRALIZATION AND DECENTRALIZATION

CENTRALIZATION:

It is the process of transferring and assigning decision-making authority to higher levels of an organizational hierarchy. The span of control of top managers is relatively broad, and there are relatively many tiers in the organization.

Characteristics

- Philosophy / emphasis on: top-down control, leadership, vision, strategy.
- Decision-making: strong, authoritarian, visionary, charismatic.
- Organizational change: shaped by top, vision of leader.
- Execution: decisive, fast, coordinated. Able to respond quickly to major issues and changes.
- Uniformity. Low risk of dissent or conflicts between parts of the organization.

Advantages of Centralization

- Provide Power and prestige for manager
- Promote uniformity of policies, practices and decisions
- Minimal extensive controlling procedures and practices
- Minimize duplication of function

Disadvantages of Centralization

- Neglected functions for mid. Level, and less motivated beside personnel.
- Nursing supervisor functions as a link officer between nursing director and first-line management.

DECENTRALIZATION:

It is the process of transferring and assigning decision-making authority to lower levels of an organizational hierarchy. The span of control of top managers is relatively small, and there are relatively few tears in the organization, because there is more autonomy in the lower ranks.

Characteristics

- Philosophy / emphasis on: bottom-up, political, cultural and learning dynamics.
- Decision-making: democratic, participative, detailed.
- Organizational change: emerging from interactions, organizational dynamics.
- Execution: evolutionary, emergent. Flexible to adapt to minor issues and changes.
- Participation, accountability. Low risk of not-invented-here behavior.

Three Forms of decentralization

- **De-concentration**. The weakest form of decentralization. Decision making authority is redistributed to lower or regional levels of the same central organization.
- **Delegation**. A more extensive form of decentralization. Through delegation the responsibility for decision-making are transferred to semi-autonomous organizations not wholly controlled by the central organization, but ultimately accountable to it.
- **Devolution**. A third type of decentralization is devolution. The authority for decision-making is transferred completely to autonomous organizational units.

Advantages of Decentralization

- Raise morale and promote interpersonal relationships
- Relieve from the daily administration
- Bring decision-making close to action
- Develop Second-line managers
- Promote employee's enthusiasm and coordination
- Facilitate actions by lower-level managers

Disadvantages of Decentralization

- Top-level administration may feel it would decrease their status
- Managers may not permit full and maximum utilization of highly qualified personnel
- Increased costs. It requires more managers and large staff
- It may lead to overlapping and duplication of effort

Centralization and Decentralization are two opposite ways to transfer decision-making power and to change the organizational structure of organizations accordingly.

There must be a good balance between centralization and decentralization of authority and power. Extreme centralization and decentralization must be avoided.

DELEGATION OF AUTHORITY

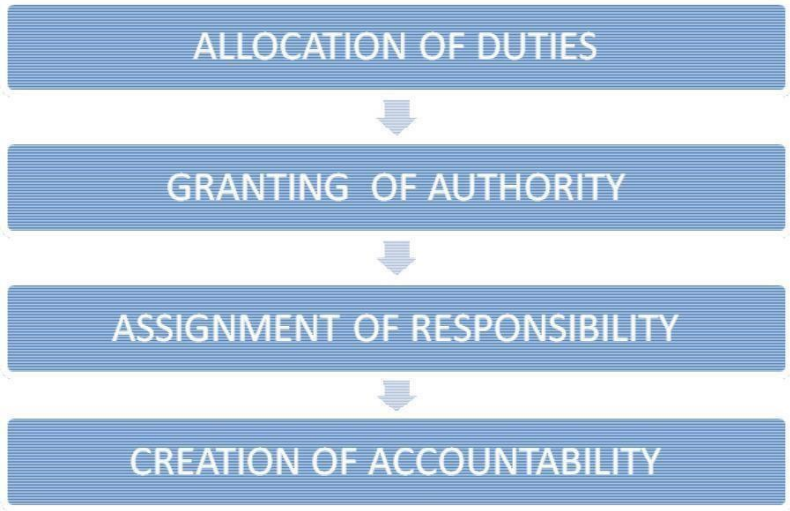
A manager alone cannot perform all the tasks assigned to him. In order to meet the targets, the manager should delegate authority. Delegation of Authority means division of authority and powers downwards to the subordinate. Delegation is about entrusting someone else to do parts of your job. Delegation of authority can be defined as subdivision and sub-allocation of powers to the subordinates in order to achieve effective results. Elements of Delegation

1. **Authority** - in context of a business organization, authority can be defined as the power and right of a person to use and allocate the resources efficiently, to take decisions and to give orders so as to achieve the organizational objectives. Authority must be well- defined. All people who have the authority should know what is the scope of their authority is and they shouldn't misutilize it. Authority is the right to give commands, orders and get the things done. The top level management has greatest authority. Authority always flows from top to bottom. It explains how a superior gets work done from his subordinate by clearly explaining what is expected of him and how he should go about it. Authority should be accompanied with an equal amount of responsibility. Delegating the authority to someone else doesn't imply escaping from accountability. Accountability still rest with the person having the utmost authority.
2. **Responsibility** - is the duty of the person to complete the task assigned to him. A person who is given the responsibility should ensure that he accomplishes the tasks assigned to him. If the tasks for which he was held responsible are not completed, then he should not give explanations or excuses. Responsibility without adequate authority leads to discontent and dissatisfaction among the person. Responsibility flows from bottom to top. The middle level and lower level management holds more responsibility. The person held responsible for a job is answerable for it. If he performs the tasks assigned as expected, he is bound for praises. While if he doesn't accomplish tasks assigned as expected, then also he is answerable for that.
3. **Accountability** - means giving explanations for any variance in the actual performance from the expectations set. Accountability cannot be delegated. For example, if 'A' is given a task with sufficient authority, and 'A' delegates this task to B and asks him to ensure that task is done well, responsibility rest with 'B', but accountability still rest with 'A'. The top level

management is most accountable. Being accountable means being innovative as the person will think beyond his scope of job. Accountability ,in short, means being answerable for the end result. Accountability can't be escaped. It arises from responsibility.

DELEGATION PROCESS

The steps involved in delegation are given below



1. **Allocation of duties –** The delegator first tries to define the task and duties to the subordinate. He also has to define the result expected from the subordinates. Clarity of duty as well as result expected has to be the first step in delegation.
2. **Granting of authority –** Subdivision of authority takes place when a superior divides and shares his authority with the subordinate. It is for this reason; every subordinate should be given enough independence to carry the task given to him by his superiors. The managers at all levels delegate authority and power which is attached to their job positions. The subdivision of powers is very important to get effective results.
3. **Assigning of Responsibility and Accountability –** The delegation process does not end once powers are granted to the subordinates. They at the same time have to be obligatory towards the duties assigned to them. Responsibility is said to be the factor or obligation of an individual to carry out his duties in best of his ability as per the directions of superior. Therefore, it is that which gives effectiveness to authority. At the same time, responsibility is absolute and cannot be shifted.

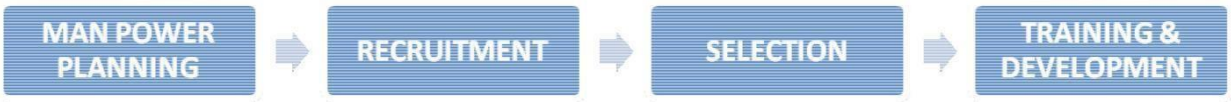
4. **Creation of accountability –** Accountability, on the others hand, is the obligation of the individual to carry out his duties as per the standards of performance. Therefore, it is said that authority is delegated, responsibility is created and accountability is imposed. Accountability arises out of responsibility and responsibility arises out of authority. Therefore, it becomes important that with every authority position an equal and opposite responsibility should be attached.

Therefore every manager, i.e., the delegator has to follow a system to finish up the delegation process. Equally important is the delegatee's role which means his responsibility and accountability is attached with the authority over to here.

STAFFING

Staffing involves filling the positions needed in the organization structure by appointing competent and qualified persons for the job.

The staffing process encompasses man power planning, recruitment, selection, and training.



a) Manpower requirements:

Manpower Planning which is also called as Human Resource Planning consists of putting right number of people, right kind of people at the right place, right time, doing the right things for which they are suited for the achievement of goals of the organization. The primary function of man power planning is to analyze and evaluate the human resources available in the organization, and to determine how to obtain the kinds of personnel needed to staff positions ranging from assembly line workers to chief executives.

b) Recruitment:

Recruitment is the process of finding and attempting to attract job candidates who are capable of effectively filling job vacancies.

Job descriptions and job specifications are important in the recruiting process because they specify the nature of the job and the qualifications required of job candidates.

c) Selection:

Selecting a suitable candidate can be the biggest challenge for any organization. The success of an organization largely depends on its staff. Selection of the right candidate builds the foundation of any organization's success and helps in reducing turnovers.

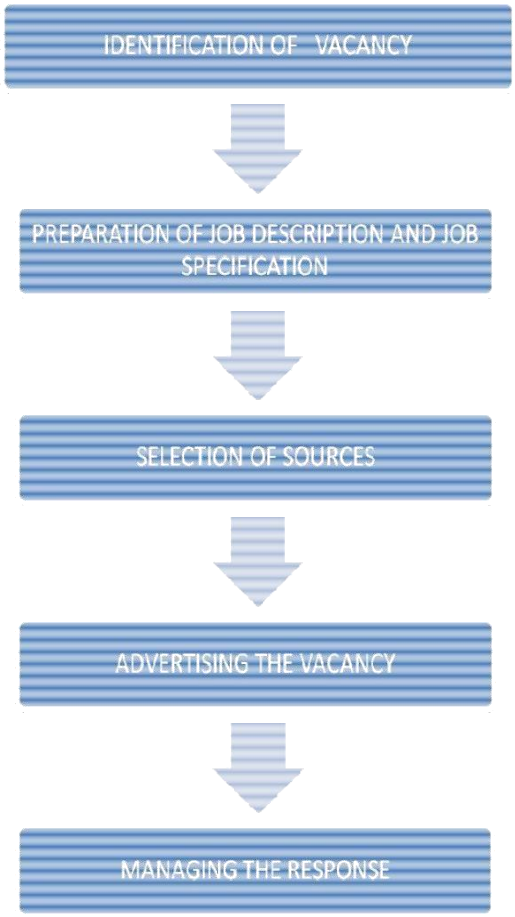
d) Training and Development:

Training and Development is a planned effort to facilitate employee learning of job-related behaviors in order to improve employee performance. Experts sometimes distinguish between the terms "training" and "development"; "training" denotes efforts to increase employee skills on present jobs, while "development" refers to efforts oriented toward improvements relevant to future jobs. In practice, though, the distinction is often blurred (mainly because upgrading skills in present jobs usually improves performance in future jobs).

RECRUITMENT PROCESS

Recruitment is the process of finding and attempting to attract job candidates who are capable of effectively filling job vacancies. The recruitment process consists of the following steps

- Identification of vacancy
- Preparation of job description and job specification
- Selection of sources
- Advertising the vacancy
- Managing the response



a) Identification of vacancy:

The recruitment process begins with the human resource department receiving requisitions for recruitment from any department of the company. These contain:

- Posts to be filled
- Number of persons
- Duties to be performed
- Qualifications required

b) Preparation of job description and job specification:

A job description is a list of the general tasks, or functions, and responsibilities of a position. It may often include to whom the position reports, specifications such as the qualifications or skills needed by the person in the job, or a salary range. A job specification describes the

knowledge, skills, education, experience, and abilities you believe are essential to performing a particular job.

c) Selection of sources:

Every organization has the option of choosing the candidates for its recruitment processes from two kinds of sources: internal and external sources. The sources within the organization itself (like transfer of employees from one department to other, promotions) to fill a position are known as the internal sources of recruitment. Recruitment candidates from all the other sources (like outsourcing agencies etc.) are known as the external sources of the recruitment.

d) Advertising the vacancy:

After choosing the appropriate sources, the vacancy is communicated to the candidates by means of a suitable media such as television, radio, newspaper, internet, direct mail etc.

e) Managing the response:

After receiving an adequate number of responses from job seekers, the sieving process of the resumes begins. This is a very essential step of the recruitment selection process, because selecting the correct resumes that match the job profile, is very important. Naturally, it has to be done rather competently by a person who understands all the responsibilities associated with the designation in its entirety. Candidates with the given skill set are then chosen and further called for interview. Also, the applications of candidates that do not match the present nature of the position but may be considered for future requirements are filed separately and preserved.

The recruitment process is immediately followed by the selection process.

JOB ANALYSIS

Job Analysis is the process of describing and recording aspects of jobs and specifying the skills and other requirements necessary to perform the job. The outputs of job analysis are

a) Job description
b) Job specification

Job Description

A job description (JD) is a written statement of what the job holder does, how it is done, under what conditions it is done and why it is done. It describes what the job is all about, throwing light on job content, environment and conditions of employment. It is descriptive in nature and defines the purpose and scope of a job. The main purpose of writing a job description is to differentiate the job from other jobs and state its outer limits.

Contents

A job description usually covers the following information:

- ♣ Job title: Tells about the job title, code number and the department where it is done.
- ♣ Job summary: A brief write-up about what the job is all about.
- ♣ Job activities: A description of the tasks done, facilities used, extent of supervisory help, etc.
- ♣ Working conditions: The physical environment of job in terms of heat, light, noise and other hazards.
- ♣ Social environment: Size of work group and interpersonal interactions required to do the job.

Job Specification

Job specification summarizes the human characteristics needed for satisfactory job completion. It tries to describe the key qualifications someone needs to perform the job successfully. It spells out the important attributes of a person in terms of education, experience, skills, knowledge and abilities (SKAs) to perform a particular job. The job specification is a logical outgrowth of a job description. For each job description, it is desirable to have a job specification. This helps the organization to find what kinds of persons are needed to take up specific jobs.

Contents

A job specification usually covers the following information:

- Education
- Experience
- Skill, Knowledge, Abilities
- Work Orientation Factors
- Age

SELECTION PROCESS

Selecting a suitable candidate can be the biggest challenge for any organisation. The success of an organization largely depends on its staff. Selection of the right candidate builds the foundation of any organization's success and helps in reducing turnovers.

Though there is no fool proof selection procedure that will ensure low turnover and high profits, the following steps generally make up the selection process-

a) Initial Screening

This is generally the starting point of any employee selection process. Initial Screening eliminates unqualified applicants and helps save time. Applications received from various sources are scrutinized and irrelevant ones are discarded.

b) Preliminary Interview

It is used to eliminate those candidates who do not meet the minimum eligibility criteria laid down by the organization. The skills, academic and family background, competencies and interests of the candidate are examined during preliminary interview. Preliminary interviews are less formalized and planned than the final interviews. The candidates are given a brief up about the company and the job profile; and it is also examined how much the candidate knows about the company. Preliminary interviews are also called screening interviews.

c) Filling Application Form

An candidate who passes the preliminary interview and is found to be eligible for the job is asked to fill in a formal application form. Such a form is designed in a way that it records the personal as well professional details of the candidates such as age, qualifications, reason for leaving previous job, experience, etc.

d) Personal Interview

Most employers believe that the personal interview is very important. It helps them in obtaining more information about the prospective employee. It also helps them in interacting with the candidate and judging his communication abilities, his ease of handling pressure etc. In some Companies, the selection process comprises only of the Interview.

e) References check

Most application forms include a section that requires prospective candidates to put down names of a few references. References can be classified into - former employer, former customers, business references, reputable persons. Such references are contacted to get a feedback on the person in question including his behaviour, skills, conduct etc.

f) Background Verification

A background check is a review of a person's commercial, criminal and (occasionally) financial records. Employers often perform background checks on employers or candidates for employment to confirm information given in a job application, verify a person's identity, or ensure that the individual does not have a history of criminal activity, etc., that could be an issue upon employment.

g) Final Interview

Final interview is a process in which a potential employee is evaluated by an employer for prospective employment in their organization. During this process, the employer hopes to

determine whether or not the applicant is suitable for the job. Different types of tests are conducted to evaluate the capabilities of an applicant, his behaviour, special qualities etc. Separate tests are conducted for various types of jobs.

h) Physical Examination

If all goes well, then at this stage, a physical examination is conducted to make sure that the candidate has sound health and does not suffer from any serious ailment.

i) Job Offer

A candidate who clears all the steps is finally considered right for a particular job and is presented with the job offer. An applicant can be dropped at any given stage if considered unfit for the job.

Employee Induction / Orientation

Orientation or induction is the process of introducing new employees to an organization, to their specific jobs & departments, and in some instances, to their community.

Purposes of Orientation

Orientation isn't a nicety! It is used for the following purposes:

1. To Reduce Startup-Costs:

Proper orientation can help the employee get "up to speed" much more quickly, thereby reducing the costs associated with learning the job.

2. To Reduce Anxiety:

Any employee, when put into a new, strange situation, will experience anxiety that can impede his or her ability to learn to do the job. Proper orientation helps to reduce anxiety that results from entering into an unknown situation, and helps provide guidelines for behaviour and conduct, so the employee doesn't have to experience the stress of guessing.

3. To Reduce Employee Turnover:

Employee turnover increases as employees feel they are not valued, or are put in positions where they can't possibly do their jobs. Orientation shows that the organization values the employee, and helps provide tools necessary for succeeding in the job.

4. To Save Time for Supervisor & Co-Workers:

Simply put, the better the initial orientation, the less likely supervisors and co-workers will have to spend time teaching the employee.

5. To Develop Realistic Job Expectations, Positive Attitudes and Job Satisfaction:

It is important that employees learn early on what is expected of them, and what to expect from others, in addition to learning about the values and attitudes of the organization. While people can learn from experience, they will make many mistakes that are unnecessary and potentially damaging.

An orientation program principally conveys 3 types of information, namely:

a) General information about the daily work routine to be followed

b) A review of the organization's history, founders, objectives, operations & products or services, as well as how the employee's job contributes to the organization's needs.

c) A detailed presentation of the organization's policies, work rules & employee benefits.

Two Kinds of Orientation

There are two related kinds of orientation. The first we will call Overview Orientation, and deals with the basic information an employee will need to understand the broader system he or she works in.

Overview Orientation includes helping employees understand:

- Management in general
- Department and the branch
- Important policies
- General procedures (non-job specific)
- Information about compensation
- Accident prevention measures
- Employee and union issues (rights, responsibilities)
- Physical facilities

Often, Overview Orientation can be conducted by the personnel department with a little help from the branch manager or immediate supervisor, since much of the content is generic in nature.

The second kind of orientation is called Job-Specific Orientation, and is the process that is used to help employees understand:

- Function of the organization,
- Responsibilities,

- Expectations,
- Duties
- Policies, procedures, rules and regulations
- Layout of workplace
- Introduction to co-workers and other people in the broader organization.

Job specific orientation is best conducted by the immediate supervisor, and/or manager, since much of the content will be specific to the individual. Often the orientation process will be ongoing, with supervisors and co-workers supplying coaching.

CARRER DEVELOPMENT

Career development not only improves job performance but also brings about the growth of the personality. Individuals not only mature regarding their potential capacities but also become better individuals.

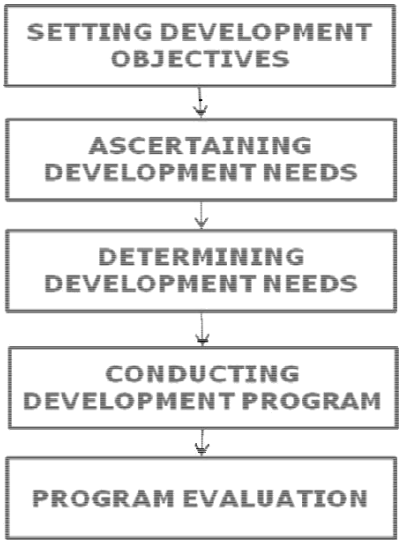
Purpose of development

Management development attempts to improve managerial performance by imparting

- Knowledge
- Changing attitudes
- Increasing skills

The major objective of development is managerial effectiveness through a planned and a deliberate process of learning. This provides for a planned growth of managers to meet the future organizational needs.

Development Process:



The development process consists of the following steps **1. Setting Development Objectives:**

It develops a framework from which executive need can be determined.

2. Ascertaining Development Needs:

It aims at organizational planning & forecast the present and future growth.

3. Determining Development Needs: This consists of

- Appraisal of present management talent
- Management Manpower Inventory

The above two processes will determine the skill deficiencies that are relative to the future needs of the organization.

4. Conducting Development Programs:

It is carried out on the basis of needs of different individuals, differences in their attitudes and behavior, also their physical, intellectual and emotional qualities. Thus a comprehensive and well conceived program is prepared depending on the organizational needs and the time & cost involved.

5. Program Evaluation:

It is an attempt to assess the value of training in order to achieve organizational objectives.

TRAINING

Training is a process of learning a sequence of programmed behaviour. It improves the employee's performance on the current job and prepares them for an intended job. Purpose of Training:

1) To improve Productivity: Training leads to increased operational productivity and increased company profit.

2) To improve Quality: Better trained workers are less likely to make operational mistakes.

3) To improve Organizational Climate: Training leads to improved production and product quality which enhances financial incentives. This in turn increases the overall morale of the organization.

4) To increase Health and Safety: Proper training prevents industrial accidents.

5) Personal Growth: Training gives employees a wider awareness, an enlarged skill base and that leads to enhanced personal growth.

Steps in Training Process:

1) Identifying Training needs: A training program is designed to assist in providing solutions for specific operational problems or to improve performance of a trainee.

- Organizational determination and Analysis: Allocation of resources that relate to organizational goal.
- Operational Analysis: Determination of a specific employee behaviour required for a particular task.

- Man Analysis: Knowledge, attitude and skill one must possess for attainment of organizational objectives

2) Getting ready for the job: The trainer has to be prepared for the job. And also who needs to be trained - the newcomer or the existing employee or the supervisory staff.

Preparation of the learner:

- Putting the learner at ease
- Stating the importance and ingredients of the job
- Creating interest
- Placing the learner as close to his normal working position
- Familiarizing him with the equipment, materials and trade terms

3) Presentation of Operation and Knowledge: The trainer should clearly tell, show, illustrate and question in order to convey the new knowledge and operations. The trainee should be encouraged to ask questions in order to indicate that he really knows and understands the job.

4) Performance Try out: The trainee is asked to go through the job several times. This gradually builds up his skill, speed and confidence.

5) Follow-up: This evaluates the effectiveness of the entire training effort

TRAINING METHODS

Training methods can be broadly classified as on-the-job training and off-the-job taining

a) On-the-job training

On the job training occurs when workers pick up skills whilst working along side experienced workers at their place of work. For example this could be the actual assembly line or offices where the employee works. New workers may simply **"shadow"** or observe fellow employees to begin with and are often given instruction manuals or interactive training programmes to work through.

b) Off-the-job training

This occurs when workers are **taken away from their place of work** to be trained. This may take place at training agency or local college, although many larger firms also have their own training centres. Training can take the form of lectures or self-study and can be used to develop more general skills and knowledge that can be used in a variety of situations. The various types of off-the-job training are

(i) Instructor presentation: The trainer orally presents new information to the trainees, usually through lecture. Instructor presentation may include classroom lecture, seminar, workshop, and the like.

(ii) Group discussion: The trainer leads the group of trainees in discussing a topic.
(iii) Demonstration: The trainer shows the correct steps for completing a task, or shows an example of a correctly completed task.
(iv) Assigned reading: The trainer gives the trainees reading assignments that provide new information.
(v) Exercise: The trainer assigns problems to be solved either on paper or in real situations related to the topic of the training activity.
(vi) Case study: The trainer gives the trainees information about a situation and directs them to come to a decision or solve a problem concerning the situation.
(vii) Role play: Trainees act out a real-life situation in an instructional setting.
(viii) Field visit and study tour: Trainees are given the opportunity to observe and interact with the problem being solved or skill being learned.

CAREER STAGES

What people want from their careers also varies according to the stage of one's career. What may have been important in an early stage may not be important in a later one. Four distinct career stages have been identified: trial, establishment/advancement, mid-career, and late career. Each stage represents different career needs and interests of the individual

a) Trial stage: The trial stage begins with an individual's exploration of career-related matters and ends usually at about age 25 with a commitment on the part of the individual to a particular occupation. Until the decision is made to settle down, the individual may try a number of jobs and a number of organizations. Unfortunately for many organizations, this trial and exploration stage results in high level of turnover among new employees. Employees in this stage need opportunities for self-exploration and a variety of job activities or assignments.

b) Establishment Stage: The establishment/advancement stage tends to occur between ages 25 and 44. In this stage, the individual has made his or her career choice and is concerned with achievement, performance, and advancement. This stage is marked by high employee productivity and career growth, as the individual is motivated to succeed in the organization and in his or her chosen occupation. Opportunities for job challenge and use of special competencies are desired in this stage. The employee strives for creativity and innovation

through new job assignments. Employees also need a certain degree of autonomy in this stage so that they can experience feelings of individual achievement and personal success.

c) Mid Career Crisis Sub Stage: The period occurring between the mid-thirties and mid-forties during which people often make a major reassessment of their progress relative to their original career ambitions and goals.

d) Maintenance stage: The mid-career stage, which occurs roughly between the ages 45 and 64, has also been referred to as the maintenance stage. This stage is typified by a continuation of established patterns of work behavior. The person is no longer trying to establish a place for himself or herself in the organization, but seeks to maintain his or her position. This stage is viewed as a mid-career plateau in which little new ground is broken. The individual in this stage may need some technical updating in his or her field. The employee should be encouraged to develop new job skills in order to avoid early stagnation and decline.

e) Late-career stage: In this stage the career lessens in importance and the employee plans for retirement and seeks to develop a sense of identity outside the work environment.

PERFORMANCE APPRAISAL

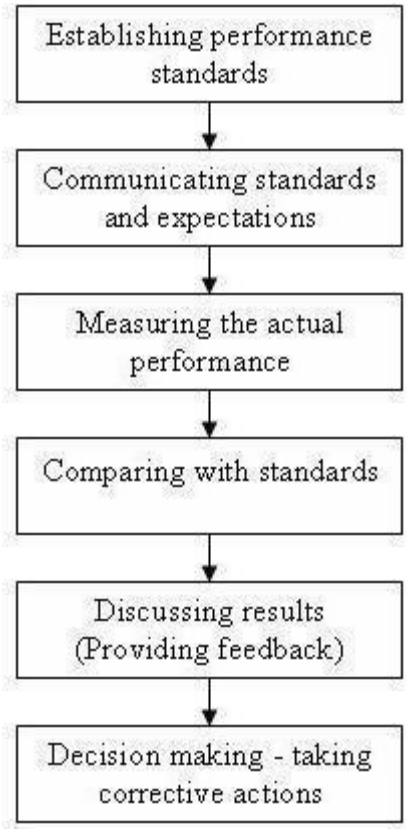
Performance appraisal is the process of obtaining, analyzing and recording information about the relative worth of an employee. The focus of the performance appraisal is measuring and improving the actual performance of the employee and also the future potential of the employee. Its aim is to measure what an employee does.

Objectives of Performance appraisal:

- To review the performance of the employees over a given period of time.
- To judge the gap between the actual and the desired performance.
- To help the management in exercising organizational control.
- Helps to strengthen the relationship and communication between superior – subordinates and management – employees.
- To diagnose the strengths and weaknesses of the individuals so as to identify the training and development needs of the future.
- To provide feedback to the employees regarding their past performance.
- Provide information to assist in the other personal decisions in the organization.

- Provide clarity of the expectations and responsibilities of the functions to be performed by the employees.
- To judge the effectiveness of the other human resource functions of the organization such as recruitment, selection, training and development.
- To reduce the grievances of the employees.

Process of performance appraisal:



a) Establishing performance standards:

The first step in the **process of performance appraisal** is the setting up of the standards which will be used to as the base to compare the actual performance of the employees. This step requires setting the criteria to judge the performance of the employees as successful or unsuccessful and the degrees of their contribution to the organizational goals and objectives. The standards set should be clear, easily understandable and in measurable terms.

In case the performance of the employee cannot be measured, great care should be taken to describe the standards.

b) Communicating the standards:

After establishing the standards, it is the responsibility of the management to communicate the standards to all the employees of the organization.

The employees should be informed and the standards should be clearly explained to the. This will help them to understand their roles and to know what exactly is expected from them. The standards should also be communicated to the appraisers or the evaluators and if required, the standards can also be modified at this stage itself according to the relevant feedback from the employees or the evaluators.

c) Measuring the actual performance:

The most difficult part of the Performance appraisal process is measuring the actual performance of the employees that is the work done by the employees during the specified period of time. It is a continuous process which involves monitoring the performance throughout the year. This stage requires the careful selection of the appropriate techniques of measurement, taking care that personal bias does not affect the outcome of the process and providing assistance rather than interfering in an employees work.

d) Comparing the actual with the desired performance:

The actual performance is compared with the desired or the standard performance. The comparison tells the deviations in the performance of the employees from the standards set. The result can show the actual performance being more than the desired performance or, the actual performance being less than the desired performance depicting a negative deviation in the organizational performance. It includes recalling, evaluating and analysis of data related to the employees' performance.

e) Discussing results:

The **result of the appraisal** is communicated and discussed with the employees on one-to-one basis. The focus of this discussion is on communication and listening. The results, the problems and the possible solutions are discussed with the aim of problem solving and reaching consensus. The feedback should be given with a positive attitude as this can have an

effect on the employees' future performance. The purpose of the meeting should be to solve the problems faced and motivate the employees to perform better.

f) Decision making:

The last step of the process is to take decisions which can be taken either to improve the performance of the employees, take the required corrective actions, or the related HR decisions like rewards, promotions, demotions, transfers etc.

UNIT IV

DIRECTING

DEFINITION

"Activating deals with the steps a manager takes to get sub-ordinates and others to carry out plans" - Newman and Warren.

Directing concerns the total manner in which a manager influences the actions of subordinates. It is the final action of a manager in getting others to act after all preparations have been completed.

Characteristics

- Elements of Management
- Continuing Function
- Pervasive Function
- Creative Function
- Linking function
- Management of Human Factor

Scope of Directing

- Initiates action
- Ensures coordination
- Improves efficiency
- Facilitates change
- Assists stability and growth

Elements of Directing

The three elements of directing are

- Motivation
- Leadership
- Communication

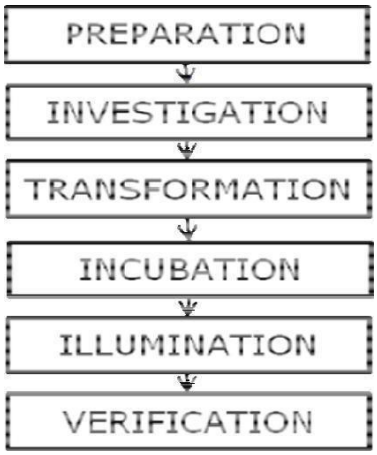
CREATIVITY AND INNOVATION

Often used interchangeably, they should to be considered separate and distinct. Creativity can be described as problem identification and idea generation and innovation is considered as idea selection, development and commercialization.

Creativity is creation of new ideas and Innovation is implementation of the new ideas.

There cannot be innovation without creativity. There can be creativity without innovation but it has no value.

Steps involved in creativity



a) Preparation: This is the first stage at which the base for creativity and innovation is defined; the mind is prepared for subsequent use in creative thinking. During preparation the individual is encouraged to appreciate the fact that every opportunity provides situations that can educate and experiences from which to learn.

The creativity aspect is kindled through a quest to become more knowledgeable. This can be done through reading about various topics and/or subjects and engaging in discussions with others. Taking part in brainstorming sessions in various forums like professional and trade association seminars, and taking time to study other countries and cultures to identify viable opportunities is also part of preparation. Of importance is the need to cultivate a personal ability to listen and learn from others.

b) Investigation: This stage of enhancing entrepreneurial creativity and innovation involves the business owner taking time to study the problem at hand and what its various components are.

c) Transformation: The information thus accumulated and acquired should then be subjected to convergent and divergent thinking which will serve to highlight the inherent similarities and differences. Convergent thinking will help identify aspects that are similar and connected while divergent thinking will highlight the differences. This twin manner of thinking is of particular importance in realizing creativity and innovation for the following reasons:

- ✶ One will be able to skim the details and see what the bigger picture is the situation/problem's components can be reordered and in doing so new patterns can be identified.
- ✶ It will help visualize a number of approaches that can be used to simultaneously tackle the problem and the opportunity.
- ✶ One's decision-making abilities will be bettered such that the urge to make snap decisions will be resisted.

d) Incubation: At this stage in the quest for creativity and innovation it is imperative that the subconscious reflect on the accumulated information, i.e. through incubation, and this can be improved or augmented when the entrepreneur:

- ✶ Engages in an activity completely unrelated to the problem/opportunity under scrutiny.
- ✶ Takes time to daydream i.e. letting the mind roam beyond any restrictions self-imposed or otherwise.
- ✶ Relax and play
- ✶ Study the problem/opportunity in a wholly different environment

e) Illumination: This happens during the incubation stage and will often be spontaneous. The realizations from the past stages combine at this instance to form a breakthrough.

f) Verification: This is where the entrepreneur attempts to ascertain whether the creativity of thought and the action of innovation are truly effective as anticipated. It may involve activities like simulation, piloting, prototype building, test marketing, and various experiments. While the tendency to ignore this stage and plunge headlong with the breakthrough may be tempting, the transformation stage should ensure that the new idea is put to the test.

MOTIVATION AND SATISFACTION

MOTIVATION

"Motivation" is a Latin word, meaning "to move". Human motives are internalized goals within individuals. Motivation may be defined as those forces that cause people to behave in certain ways. Motivation encompasses all those pressures and influences that trigger, channel, and sustain human behavior. Most successful managers have learned to understand the concept of human motivation and are able to use that understanding to achieve higher standards of subordinate work performance.

According to Koontz and O'Donnell, "Motivation is a class of drives, needs, wishes and similar forces".

NATURE AND CHARACTERISTICS OF MOTIVATION

Psychologists generally agree that all behavior is motivated, and that people have reasons for doing the things they do or for behaving in the manner that they do. Motivating is the work a manager performs to inspire, encourage and impel people to take required action.

The characteristics of motivation are given below:-

- **Motivation is an Internal Feeling**

Motivation is a psychological phenomenon which generates in the mind of an individual the feeling that he lacks certain things and needs those things. Motivation is a force within an individual that drives him to behave in a certain way.

- **Motivation is Related to Needs**

Needs are deficiencies which are created whenever there is a physiological or psychological imbalance. In order to motivate a person, we have to understand his needs that call for satisfaction.

- **Motivation Produces Goal-Directed Behaviour**

Goals are anything which will alleviate a need and reduce a drive. An individual's behavior is directed towards a goal.

✶ **Motivation can be either Positive or Negative**

Positive or incentive motivation is generally based on reward. According to Flippo - "positive motivation is a process of attempting to influence others to do your will through the possibility of gain or reward".

Negative or fear motivation is based on force and fear. Fear causes persons to act in a certain way because they are afraid of the consequences if they don't.

IMPORTANCE OF MOTIVATION

A manager's primary task is to motivate others to perform the tasks of the organization. Therefore, the manager must find the keys to get subordinates to come to work regularly and on time, to work hard, and to make positive contributions towards the effective and efficient achievement of organizational objectives. Motivation is an effective instrument in the hands of a manager for inspiring the work force and creating confidence in it. By motivating the work force, management creates "will to work" which is necessary for the achievement of organizational goals. The various benefits of motivation are:-

- Motivation is one of the important elements in the directing process. By motivating the workers, a manager directs or guides the workers' actions in the desired direction for accomplishing the goals of the organization.
- Workers will tend to be as efficient as possible by improving upon their skills and knowledge so that they are able to contribute to the progress of the organization thereby increasing productivity.
- For performing any tasks, two things are necessary. They are: (a) ability to work and (b) willingness to work. Without willingness to work, ability to work is of no use. The willingness to work can be created only by motivation.
- Organizational effectiveness becomes, to some degree, a question of management's ability to motivate its employees, to direct at least a reasonable effort towards the goals of the organization.
- Motivation contributes to good industrial relations in the organization. When the workers are motivated, contented and disciplined, the frictions between the workers and the management will be reduced.

- Motivation is the best remedy for resistance to changes. When changes are introduced in an organization, generally, there will be resistance from the workers. But if the workers of an

organization are motivated, they will accept, introduce and implement the changes whole heartily and help to keep the organization on the right track of progress.

- Motivation facilitates the maximum utilization of all factors of production, human, physical and financial resources and thereby contributes to higher production.
- Motivation promotes a sense of belonging among the workers. The workers feel that the enterprise belongs to them and the interest of the enterprise is their interests.
- Many organizations are now beginning to pay increasing attention to developing their employees as future resources upon which they can draw as they grow and develop.

SATISFACTION

Employee satisfaction (Job satisfaction) is the terminology used to describe whether employees are happy and contented and fulfilling their desires and needs at work. Many measures purport that employee satisfaction is a factor in employee motivation, employee goal achievement, and positive employee morale in the workplace.

Employee satisfaction, while generally a positive in your organization, can also be a downer if mediocre employees stay because they are satisfied with your work environment.

Factors contributing to employee satisfaction include treating employees with respect, providing regular employee recognition, empowering employees, offering above industry-average benefits and compensation, providing employee perks and company activities, and positive management within a success framework of goals, measurements, and expectations.

Employee satisfaction is often measured by anonymous employee satisfaction surveys administered periodically that gauge employee satisfaction in areas such as:

- management,
- understanding of mission and vision,
- empowerment,
- teamwork,
- communication, and
- Coworker interaction.

The facets of employee satisfaction measured vary from company to company.

A second method used to measure employee satisfaction is meeting with small groups of employees and asking the same questions verbally. Depending on the culture of the company, either method can contribute knowledge about employee satisfaction to managers and employees.

JOB DESIGN

It is the process of Work arrangement (or rearrangement) aimed at reducing or overcoming job dissatisfaction and employee alienation arising from repetitive and mechanistic tasks. Through job design, organizations try to raise productivity levels by offering non-monetary rewards such as greater satisfaction from a sense of personal achievement in meeting the increased challenge and responsibility of one's work.

Approaches to job design include:

✶ **Job Enlargement:** Job enlargement changes the jobs to include more and/or different tasks. Job enlargement should add interest to the work but may or may not give employees more responsibility.

✶ **Job Rotation:** Job rotation moves employees from one task to another. It distributes the group tasks among a number of employees.

✶ **Job Enrichment:** Job enrichment allows employees to assume more responsibility, accountability, and independence when learning new tasks or to allow for greater participation and new opportunities.

TYPES OF MOTIVATION TECHNIQUES

If a manager wants to get work done by his employees, he may either hold out a promise of a reward (positive motivation) or he/she may install fear (negative motivation). Both these types are widely used by managements.

a) Positive Motivation:

This type of motivation is generally based on reward. A positive motivation involves the possibility of increased motive satisfaction. According to Flippo - "Positive motivation is a process of attempting to influence others to do your will through the possibility of gain or reward". Incentive motivation is the "pull" mechanism. The receipt of awards, due recognition and praise for work-well done definitely lead to good team spirit, co-operation and a feeling of happiness.

- Positive motivation include:-
- Praise and credit for work done
- Wages and Salaries

- Appreciation
- A sincere interest in subordinates as individuals
- Delegation of authority and responsibility

b) Negative Motivation:

This type of motivation is based on force and fear. Fear causes persons to act in a certain way because they fear the consequences. Negative motivation involves the possibility of decreased motive satisfaction. It is a "push" mechanism. The imposition of punishment frequently results in frustration among those punished, leading to the development of maladaptive behaviour. It also creates a hostile state of mind and an unfavourable attitude to the job. However, there is no management which has not used the negative motivation at some time or the other.

MOTIVATION THEORIES

Some of the motivation theories are discussed below

a) McGregor's Theory X and Theory Y:

McGregor states that people inside the organization can be managed in two ways. The first is basically negative, which falls under the category X and the other is basically positive, which falls under the category Y. After viewing the way in which the manager dealt with employees, McGregor concluded that a manager's view of the nature of human beings is based on a certain grouping of assumptions and that he or she tends to mold his or her behavior towards subordinates according to these assumptions.

Under the assumptions of theory X :

- Employees inherently do not like work and whenever possible, will attempt to avoid it.
- Because employees dislike work, they have to be forced, coerced or threatened with punishment to achieve goals.
- Employees avoid responsibilities and do not work fill formal directions are issued.
- Most workers place a greater importance on security over all other factors and display little ambition.

- Physical and mental effort at work is as natural as rest or play.

- People do exercise self-control and self-direction and if they are committed to those goals.
- Average human beings are willing to take responsibility and exercise imagination, ingenuity and creativity in solving the problems of the organization.
- That the way the things are organized, the average human being's brainpower is only partly used.

On analysis of the assumptions it can be detected that theory X assumes that lower-order needs dominate individuals and theory Y assumes that higher-order needs dominate individuals. An organization that is run on Theory X lines tends to be authoritarian in nature, the word "authoritarian" suggests such ideas as the "power to enforce obedience" and the "right to command." In contrast Theory Y organizations can be described as "participative", where the aims of the organization and of the individuals in it are integrated; individuals can achieve their own goals best by directing their efforts towards the success of the organization.

b) Abraham Maslow's "Need Hierarchy Theory":

One of the most widely mentioned theories of motivation is the hierarchy of needs theory put forth by psychologist Abraham Maslow. Maslow saw human needs in the form of a hierarchy, ascending from the lowest to the highest, and he concluded that when one set of needs is satisfied, this kind of need ceases to be a motivator. As per his theory these needs are:

(i) Physiological needs:

These are important needs for sustaining the human life. Food, water, warmth, shelter, sleep, medicine and education are the basic physiological needs which fall in the primary list of need satisfaction. Maslow was of an opinion that until these needs were satisfied to a degree to maintain life, no other motivating factors can work.

(ii) Security or Safety needs:

These are the needs to be free of physical danger and of the fear of losing a job, property, food or shelter. It also includes protection against any emotional harm.

(iii) Social needs:

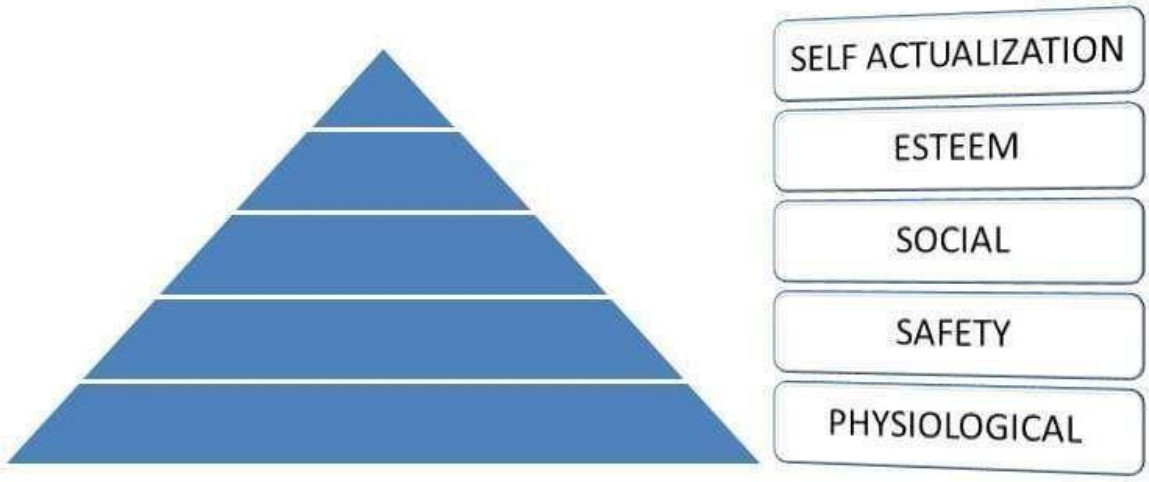
Since people are social beings, they need to belong and be accepted by others. People try to satisfy their need for affection, acceptance and friendship.

(iv) Esteem needs:

According to Maslow, once people begin to satisfy their need to belong, they tend to want to be held in esteem both by themselves and by others. This kind of need produces such satisfaction as power, prestige status and self-confidence. It includes both internal esteem factors like self-respect, autonomy and achievements and external esteem factors such as states, recognition and attention.

(v) Need for self-actualization:

Maslow regards this as the highest need in his hierarchy. It is the drive to become what one is capable of becoming; it includes growth, achieving one's potential and self-fulfillment. It is to maximize one's potential and to accomplish something.



All of the needs are structured into a hierarchy and only once a lower level of need has been fully met, would a worker be motivated by the opportunity of having the next need up in the hierarchy satisfied. For example a person who is dying of hunger will be motivated to achieve a basic wage in order to buy food before worrying about having a secure job contract or the respect of others.

A business should therefore offer different incentives to workers in order to help them fulfill each need in turn and progress up the hierarchy. Managers should also recognize that workers are not all motivated in the same way and do not all move up the hierarchy at the same pace. They may therefore have to offer a slightly different set of incentives from worker to worker.

c) Frederick Herzberg's motivation-hygiene theory:

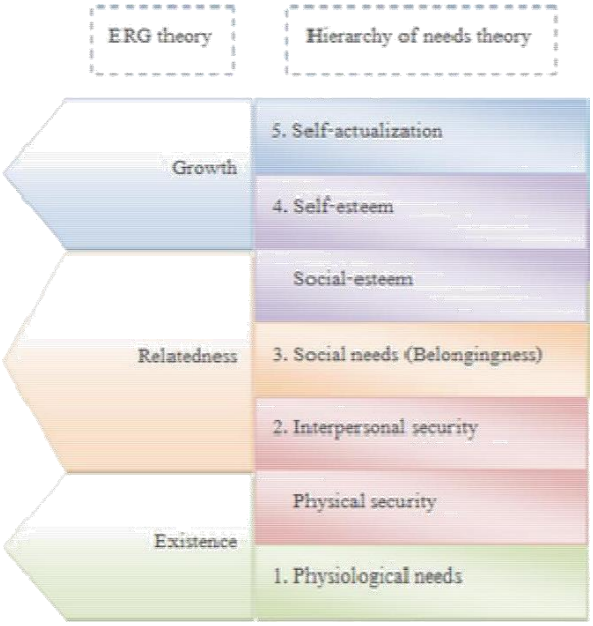
Frederick has tried to modify Maslow's need Hierarchy theory. His theory is also known as two-factor theory or Hygiene theory. He stated that there are certain satisfiers and dissatisfiers for employees at work. Intrinsic factors are related to job satisfaction, while extrinsic factors are associated with dissatisfaction. He devised his theory on the question: "What do people want from their jobs?" He asked people to describe in detail, such situations when they felt exceptionally good or exceptionally bad. From the responses that he received, he concluded that opposite of satisfaction is not dissatisfaction. Removing dissatisfying characteristics from a job does not necessarily make the job satisfying. He states that presence of certain factors in the organization is natural and the presence of the same does not lead to motivation. However, their non-presence leads to de-motivation. In similar manner there are certain factors, the absence of which causes no dissatisfaction, but their presence has motivational impact.
Examples of Hygiene factors are:
Security, status, relationship with subordinates, personal life, salary, work conditions, relationship with supervisor and company policy and administration. Examples of Motivational factors are:
Growth prospectus job advancement, responsibility, challenges, recognition and achievements.

d) Victor Vroom's Expectancy theory:

The most widely accepted explanations of motivation have been propounded by Victor Vroom. His theory is commonly known as expectancy theory. The theory argues that the strength of a tendency to act in a specific way depends on the strength of an expectation that the act will be followed by a given outcome and on the attractiveness of that outcome to the individual to make this simple, expectancy theory says that an employee can be motivated to perform better when there is a belief that the better performance will lead to good performance appraisal and that this shall result into realization of personal goal in form of some reward.
Therefore an employee is:
Motivation = Valence x Expectancy.
The theory focuses on three things:

- Efforts and performance relationship
- Performance and reward relationship
- Rewards and personal goal relationship

e) Clayton Alderfer's ERG Theory:



Alderfer has tried to rebuild the hierarchy of needs of Maslow into another model named ERG i.e. Existence – Relatedness – Growth. According to him there are 3 groups of core needs as mentioned above. The existence group is concerned mainly with providing basic material existence. The second group is the individuals need to maintain interpersonal relationship with other members in the group. The final group is the intrinsic desire to grow and develop personally. The major conclusions of this theory are :

- In an individual, more than one need may be operative at the same time.
- If a higher need goes unsatisfied than the desire to satisfy a lower need intensifies.
- It also contains the frustration-regression dimension.

f) McClelland's Theory of Needs:

David McClelland has developed a theory on three types of motivating needs :

(i) Need for Power
(ii) Need for Affiliation
(iii) Need for Achievement

Basically people for high need for power are inclined towards influence and control. They like to be at the center and are good orators. They are demanding in nature, forceful in manners and ambitious in life. They can be motivated to perform if they are given key positions or power positions.

In the second category are the people who are social in nature. They try to affiliate themselves with individuals and groups. They are driven by love and faith. They like to build a friendly environment around themselves. Social recognition and affiliation with others provides them motivation.

People in the third area are driven by the challenge of success and the fear of failure. Their need for achievement is moderate and they set for themselves moderately difficult tasks. They are analytical in nature and take calculated risks. Such people are motivated to perform when they see at least some chances of success.

McClelland observed that with the advancement in hierarchy the need for power and achievement increased rather than Affiliation. He also observed that people who were at the top, later ceased to be motivated by this drives.

g) Stacey Adams' Equity Theory:

As per the equity theory of J. Stacey Adams, people are motivated by their beliefs about the reward structure as being fair or unfair, relative to the inputs. People have a tendency to use subjective judgment to balance the outcomes and inputs in the relationship for comparisons between different individuals. Accordingly:

$$\frac{\text{Out comes by a person}}{\text{Inputs by a person}} = \frac{\text{Out comes by another person}}{\text{Input by another person}}$$

If people feel that they are not equally rewarded they either reduce the quantity or quality of work or migrate to some other organization. However, if people perceive that they are rewarded higher, they may be motivated to work harder.

h) Skinner's Reinforcement Theory:

B.F. Skinner, who propounded the reinforcement theory, holds that by designing the environment properly, individuals can be motivated. Instead of considering internal factors like impressions, feelings, attitudes and other cognitive behavior, individuals are directed by what happens in the environment external to them. Skinner states that work environment should be made suitable to the individuals and that punishment actually leads to frustration and de-motivation. Hence, the only way to motivate is to keep on making positive changes in the external environment of the organization.

LEADERSHIP

Definition

Leadership is defined as influence, the art or process of influencing people so that they will strive willingly and enthusiastically toward the achievement of group goals.

- Leaders act to help a group attain objectives through the maximum application of its capabilities.
- Leaders must instill values – whether it be concern for quality, honesty and calculated risk taking or for employees and customers.

Importance of Leadership

1. Aid to authority
2. Motive power to group efforts
3. Basis for co operation
4. Integration of Formal and Informal Organization.

LEADERSHIP STYLES

The leadership style we will discuss here are:

a) Autocratic style
b) Democratic Style
c) Laissez Faire Style

a) Autocratic style

Manager retains as much power and decision-making authority as possible. The manager does not consult employees, nor are they allowed to give any input. Employees are expected to obey orders without receiving any explanations. The motivation environment is produced by creating a structured set of rewards and punishments.

Autocratic leadership is a classical leadership style with the following characteristics:

- Manager seeks to make as many decisions as possible
- Manager seeks to have the most authority and control in decision making
- Manager seeks to retain responsibility rather than utilize complete delegation

- Consultation with other colleagues in minimal and decision making becomes a solitary process
- Managers are less concerned with investing their own leadership development, and prefer to simply work on the task at hand.

Advantages

Reduced stress due to increased control

A more productive group 'while the leader is watching'

Improved logistics of operations

Faster decision making

Disadvantages

Short-termistic approach to management.

Manager perceived as having poor leadership skills Increased workload for the manager

People dislike being ordered around

Teams become dependent upon their leader

b) Democratic Style

Democratic Leadership is the leadership style that promotes the sharing of responsibility, the exercise of delegation and continual consultation.

The style has the following characteristics:

- Manager seeks consultation on all major issues and decisions.
- Manager effectively delegate tasks to subordinates and give them full control and responsibility for those tasks.
- Manager welcomes feedback on the results of intiatives and the work environment.
- Manager encourages others to become leaders and be involved in leadership development.

Advantages

Positive work environment

Successful initiatives

Creative thinking

Reduction of friction and office politics

Reduced employee turnover

Disadvantages

Takes long time to take decisions

Danger of pseudo participation

Like the other styles, the democratic style is not always appropriate. It is most successful when used with highly skilled or experienced employees or when implementing operational changes or resolving individual or group problems.

c) Laissez-Faire Style

This French phrase means "leave it be" and is used to describe a leader who leaves his/her colleagues to get on with their work. The style is largely a "hands off" view that tends to minimize the amount of direction and face time required.

Advantages

- No work for the leader
- Frustration may force others into leadership roles
- Allows the visionary worker the opportunity to do what they want, free from interference
- Empowers the group

Disadvantages

It makes employees feel insecure at the unavailability of a manager.

The manager cannot provide regular feedback to let employees know how well they are doing.

Managers are unable to thank employees for their good work.

The manager doesn't understand his or her responsibilities and is hoping the employees can cover for him or her.

LEADERSHIP THEORIES

The various leadership theories are

a) Great Man Theory:

Assumptions

- Leaders are born and not made.
- Great leaders will arise when there is a great need.

Description

Early research on leadership was based on the study of people who were already great leaders. These people were often from the aristocracy, as few from lower classes had the opportunity to lead. This contributed to the notion that leadership had something to do with breeding.

The idea of the Great Man also strayed into the mythic domain, with notions that in times of need, a Great Man would arise, almost by magic. This was easy to verify, by pointing to people such as Eisenhower and Churchill, let alone those further back along the timeline, even to Jesus, Moses, Mohammed and the Buddah.

Discussion

Gender issues were not on the table when the 'Great Man' theory was proposed. Most leaders were male and the thought of a Great Woman was generally in areas other than leadership. Most researchers were also male, and concerns about androcentric bias were a long way from being realized.

b) Trait Theory:

Assumptions

- People are born with inherited traits.
- Some traits are particularly suited to leadership.
- People who make good leaders have the right (or sufficient) combination of traits.

Description

Early research on leadership was based on the psychological focus of the day, which was of people having inherited characteristics or traits. Attention was thus put on discovering these

traits, often by studying successful leaders, but with the underlying assumption that if other people could also be found with these traits, then they, too, could also become great leaders.

McCall and Lombardo (1983) researched both success and failure identified four primary traits by which leaders could succeed or 'derail':
Emotional stability and composure: Calm, confident and predictable, particularly when under stress.
Admitting error: Owning up to mistakes, rather than putting energy into covering up.
Good interpersonal skills: able to communicate and persuade others without resort to negative or coercive tactics.
Intellectual breadth: Able to understand a wide range of areas, rather than having a narrow (and narrow-minded) area of expertise.

c) Behavioral Theory:

Assumptions

- Leaders can be made, rather than are born.
- Successful leadership is based in definable, learnable behavior.

Description

Behavioral theories of leadership do not seek inborn traits or capabilities. Rather, they look at what leaders actually do.
If success can be defined in terms of describable actions, then it should be relatively easy for other people to act in the same way. This is easier to teach and learn then to adopt the more ephemeral 'traits' or 'capabilities'.

d) Participative

Leadership: Assumptions

- Involvement in decision-making improves the understanding of the issues involved by those who must carry out the decisions.
- People are more committed to actions where they have involved in the relevant decision-making.
- People are less competitive and more collaborative when they are working on joint goals.

- When people make decisions together, the social commitment to one another is greater and thus increases their commitment to the decision.
- Several people deciding together make better decisions than one person alone.

Description

A Participative Leader, rather than taking autocratic decisions, seeks to involve other people in the process, possibly including subordinates, peers, superiors and other stakeholders. Often, however, as it is within the managers' whim to give or deny control to his or her subordinates, most participative activity is within the immediate team. The question of how much influence others are given thus may vary on the manager's preferences and beliefs, and a whole spectrum of participation is possible

e) Situational Leadership:

Assumptions

- The best action of the leader depends on a range of situational factors.

Description

When a decision is needed, an effective leader does not just fall into a single preferred style. In practice, as they say, things are not that simple.

Factors that affect situational decisions include motivation and capability of followers. This, in turn, is affected by factors within the particular situation. The relationship between followers and the leader may be another factor that affects leader behavior as much as it does follower behavior.

The leaders' perception of the follower and the situation will affect what they do rather than the truth of the situation. The leader's perception of themselves and other factors such as stress and mood will also modify the leaders' behavior.

f) Contingency

Theory: Assumptions

- The leader's ability to lead is contingent upon various situational factors, including the leader's preferred style, the capabilities and behaviors of followers and also various other situational factors.

Description

Contingency theories are a class of behavioral theory that contend that there is no one best way of leading and that a leadership style that is effective in some situations may not be successful in others.

An effect of this is that leaders who are very effective at one place and time may become unsuccessful either when transplanted to another situation or when the factors around them change.

Contingency theory is similar to situational theory in that there is an assumption of no simple one right way. The main difference is that situational theory tends to focus more on the behaviors that the leader should adopt, given situational factors (often about follower behavior), whereas contingency theory takes a broader view that includes contingent factors about leader capability and other variables within the situation.

g) Transactional

Leadership: Assumptions

- People are motivated by reward and punishment.
- Social systems work best with a clear chain of command.
- When people have agreed to do a job, a part of the deal is that they cede all authority to their manager.
- The prime purpose of a subordinate is to do what their manager tells them to do.

Description

The transactional leader works through creating clear structures whereby it is clear what is required of their subordinates, and the rewards that they get for following orders. Punishments are not always mentioned, but they are also well-understood and formal systems of discipline are usually in place.

The early stage of Transactional Leadership is in negotiating the contract whereby the subordinate is given a salary and other benefits, and the company (and by implication the subordinate's manager) gets authority over the subordinate.

When the Transactional Leader allocates work to a subordinate, they are considered to be fully responsible for it, whether or not they have the resources or capability to carry it out. When things go wrong, then the subordinate is considered to be personally at fault, and is punished for their failure (just as they are rewarded for succeeding).

h) Transformational Leadership:

Assumptions

- People will follow a person who inspires them.
- A person with vision and passion can achieve great things.
- The way to get things done is by injecting enthusiasm and energy.

Description

Working for a Transformational Leader can be a wonderful and uplifting experience. They put passion and energy into everything. They care about you and want you to succeed.

Transformational Leaders are often charismatic, but are not as narcissistic as pure Charismatic Leaders, who succeed through a belief in themselves rather than a belief in others.

One of the traps of Transformational Leadership is that passion and confidence can easily be mistaken for truth and reality.

Transformational Leaders, by definition, seek to transform. When the organization does not need transforming and people are happy as they are, then such a leader will be frustrated. Like wartime leaders, however, given the right situation they come into their own and can be personally responsible for saving entire companies.

COMMUNICATION

Communication is the exchange of messages between people for the purpose of achieving common meanings. Unless common meanings are shared, managers find it extremely difficult to influence others. Whenever group of people interact, communication takes place. Communication is the exchange of information using a shared set of symbols. It is the process that links group members and enables them to coordinate their activities. Therefore, when managers foster effective communication, they strengthen the

connections between employees and build cooperation. Communication also functions to build and reinforce interdependence between various parts of the organization. As a linking mechanism among the different organizational subsystems, communication is a central feature of the structure of groups and organizations. It helps to coordinate tasks and activities within and between organizations.

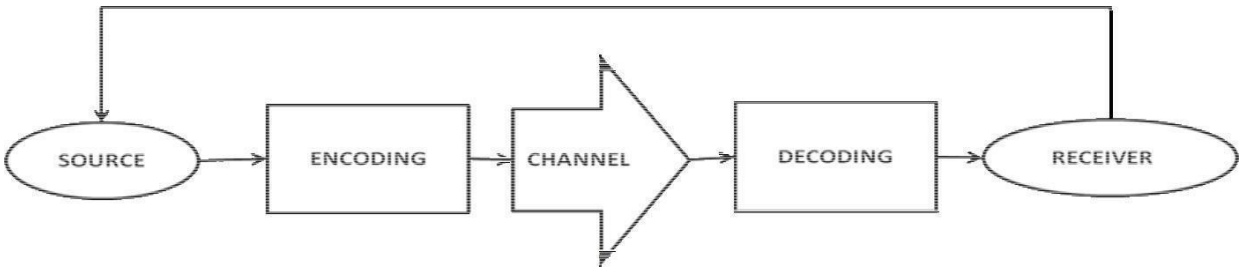
DEFINITION

According to Koontz and O'Donnell, "Communication, is an intercourse by words, letters symbols or messages, and is a way that the organization members shares meaning and understanding with another".

THE COMMUNICATION PROCESS

Communication is important in building and sustaining human relationships at work. Communication can be thought of as a process or flow. Before communication can take place, a purpose, expressed as a message to be conveyed is needed. It passes between the sender and the receiver. The result is transference of meaning from one person to another.

The figure below depicts the communication process. This model is made up of seven parts: (1) Source, (2) Encoding, (3) Message, (4) Channel, (5) Decoding, (6) Receiver, and (7) Feedback.



a) Source:

The source initiates a message. This is the origin of the communication and can be an individual, group or inanimate object. The effectiveness of a communication depends to a considerable degree on the characteristics of the source. The person who initiates the communication process is known as sender, source or communicator. In an organization, the sender will be a person who has a need or desire to send a message to others. The sender has some information which he wants to communicate to some other person to achieve some purpose. By initiating the message, the sender attempts to achieve understanding and change in the behaviour of the receiver.

b) Encoding:

Once the source has decided what message to communicate, the content of the message must be put in a form the receiver can understand. As the background for encoding information, the sender uses his or her own frame of reference. It includes the individual's view of the organization or situation as a function of personal education, interpersonal relationships, attitudes, knowledge and experience. Three conditions are necessary for successful encoding the message.

- **Skill:** Successful communicating depends on the skill you posses. Without the requisite skills, the message of the communicator will not reach the requisite skills; the message of the communicator will not reach the receiver in the desired form. One's total communicative success includes speaking, reading, listening and reasoning skills.
- **Attitudes:** Our attitudes influence our behaviour. We hold predisposed ideas on a number of topics and our communications are affected by these attitudes.
- **Knowledge:** We cannot communicate what we don't know. The amount of knowledge the source holds about his or her subject will affect the message he or she seeks to transfer.

c) The Message:

The message is the actual physical product from the source encoding. The message contains the thoughts and feelings that the communicator intends to evoke in the receiver. The message has two primary components:-

- **The Content:** The thought or conceptual component of the message is contained in the words, ideas, symbols and concepts chosen to relay the message.
- **The Affect:** The feeling or emotional component of the message is contained in the intensity, force, demeanour (conduct or behaviour), and sometimes the gestures of the communicator.

d) The Channel:

The actual means by which the message is transmitted to the receiver (Visual, auditory, written or some combination of these three) is called the channel. The channel is the medium through which the message travels. The channel is the observable carrier of the message. Communication in which the sender's voice is used as the channel is called oral communication. When the channel involves written language, the sender is using written communication. The sender's choice of a channel conveys additional information beyond that contained in the

message itself. For example, documenting an employee's poor performance in writing conveys that the manager has taken the problem seriously.

f) Decoding:

Decoding means interpreting what the message means. The extent to which the decoding by the receiver depends heavily on the individual characteristics of the sender and receiver. The greater the similarity in the background or status factors of the communicators, the greater the probability that a message will be perceived accurately. Most messages can be decoded in more than one way. Receiving and decoding a message are a type of perception. The decoding process is therefore subject to the perception biases.

g) The Receiver:

The receiver is the object to whom the message is directed. Receiving the message means one or more of the receiver's senses register the message - for example, hearing the sound of a supplier's voice over the telephone or seeing the boss give a thumbs-up signal. Like the sender, the receiver is subject to many influences that can affect the understanding of the message. Most important, the receiver will perceive a communication in a manner that is consistent with previous experiences. Communications that are not consistent with expectations is likely to be rejected.

h) Feedback:

The final link in the communication process is a feedback loop. Feedback, in effect, is communication travelling in the opposite direction. If the sender pays attention to the feedback and interprets it accurately, the feedback can help the sender learn whether the original communication was decoded accurately. Without feedback, one-way communication occurs between managers and their employees. Faced with differences in their power, lack of time, and a desire to save face by not passing on negative information, employees may be discouraged from providing the necessary feedback to their managers.

Guidelines for effective Communication

(i) Senders of message must clarify in their minds what they want to communicate. Purpose of the message and making a plan to achieve the intended end must be clarified.

(ii) Encoding and decoding be done with symbols that are familiar to the sender and the receiver of the message.

(iii) For the planning of the communication, other people should be consulted and encouraged to participate.

(iv) It is important to consider the needs of the receivers of the information. Whenever appropriate, one should communicate something that is of value to them, in the short run as well as in the more distant future.

(v) In communication, tone of voice, the choice of language and the congruency between what is said and how it is said influence the reactions of the receiver of the message.

(vi) Communication is complete only when the message is understood by the receiver. And one never knows whether communication is understood unless the sender gets a feedback.

(vii) The function of communication is more than transmitting the information. It also deals with emotions that are very important in interpersonal relationships between superiors, subordinates and colleagues in an organization.

(viii) Effective communicating is the responsibility not only of the sender but also of the receiver of the information.

BARRIERS TO EFFECTIVE COMMUNICATION

Barriers to communication are factors that block or significantly distort successful communication. Effective managerial communication skills helps overcome some, but not all, barriers to communication in organizations. The more prominent barriers to effective communication which every manager should be aware of is given below:

a) Filtering:

Filtering refers to a sender manipulating information so it will be seen more favourably by the receiver. The major determinant of filtering is the number of levels in an organization's structure. The more vertical levels in the organization's hierarchy, the more opportunities for filtering. Sometimes the information is filtered by the sender himself. If the sender is hiding some meaning and disclosing in such a fashion as appealing to the receiver, then he is "filtering" the message deliberately. A manager in the process of altering communication in his favour is attempting to filter the information.

b) Selective Perception:

Selective perception means seeing what one wants to see. The receiver, in the communication process, generally resorts to selective perception i.e., he selectively perceives the message based on the organizational requirements, the needs and characteristics, background of the employees etc. Perceptual distortion is one of the distressing barriers to the effective communication. People interpret what they see and call it a reality. In our regular activities, we tend to see those things that please us and to reject or ignore unpleasant things. Selective perception allows us to keep out dissonance (the existence of conflicting elements in our perceptual set) at a tolerable level. If we encounter something that does not fit out current image of reality, we structure the situation to minimize our dissonance. Thus, we manage to overlook many stimuli from the environment that do not fit into out current perception of the world. This process has significant implications for managerial activities. For example, the employment interviewer who expects a female job applicant to put her family ahead of her career is likely to see that in female applicants, regardless of whether the applicants feel that way or not.

c) Emotions:

How the receiver feels at the time of receipt of information influences effectively how he interprets the information. For example, if the receiver feels that the communicator is in a jovial mood, he interprets that the information being sent by the communicator to be good and interesting. Extreme emotions and jubilation or depression are quite likely to hinder the effectiveness of communication. A person's ability to encode a message can become impaired when the person is feeling strong emotions. For example, when you are angry, it is harder to consider the other person's viewpoint and to choose words carefully. The angrier you are, the harder this task becomes. Extreme emotions – such as jubilation or depression - are most likely to hinder effective communication. In such instances, we are most prone to disregard our rational and objective thinking processes and substitute emotional judgments.

d) Language:

Communicated message must be understandable to the receiver. Words mean different things to different people. Language reflects not only the personality of the individual but also the culture of society in which the individual is living. In organizations, people from different regions, different backgrounds, and speak different languages. People will have different academic backgrounds, different intellectual facilities, and hence the jargon they use varies. Often, communication gap arises because the language the sender is using may be

incomprehensible, vague and indigestible. Language is a central element in communication. It may pose a barrier if its use obscures meaning and distorts intent. Words mean different things to different people. Age, education and cultural background are three of the more obvious variables that influence the language a person uses and the definitions he or she gives to words. Therefore, use simple, direct, declarative language.

Speak in brief sentences and use terms or words you have heard from you audience. As much as possible, speak in the language of the listener. Do not use jargon or technical language except with those who clearly understand it.

e) Stereotyping:

Stereotyping is the application of selective perception. When we have preconceived ideas about other people and refuse to discriminate between individual behaviours, we are applying selective perception to our relationship with other people. Stereotyping is a barrier to communications because those who stereotype others use selective perception in their communication and tend to hear only those things that confirm their stereotyped images. Consequently, stereotypes become more deeply ingrained as we find more "evidence" to confirm our original opinion. Stereotyping has a convenience function in our interpersonal relations. Since people are all different, ideally we should react and interact with each person differently. To do this, however, requires considerable psychological effort. It is much easier to categorize (stereotype) people so that we can interact with them as members of a particular category. Since the number of categories is small, we end up treating many people the same even though they are quite different. Our communications, then, may be directed at an individual as a member of a category at the sacrifice of the more effective communication on a personal level.

f) Status Difference:

The organizational hierarchy pose another barrier to communication within organization, especially when the communication is between employee and manager. This is so because the employee is dependent on the manager as the primary link to the organization and hence more likely to distort upward communication than either horizontal or downward communication. Effective supervisory skills make the supervisor more approachable and help reduce the risk of problems related to status differences. In addition, when employees feel secure, they are more likely to be straightforward in upward communication.

g) Use of Conflicting Signals:

A sender is using conflicting signals when he or she sends inconsistent messages. A vertical message might conflict with a nonverbal one. For example, if a manager says to his employees, "If you have a problem, just come to me. My door is always open", but he looks annoyed whenever an employee knocks on his door". Then we say the manager is sending conflicting messages. When signals conflict, the receivers of the message have to decide which, if any, to believe.

h) Reluctance to Communicate:

For a variety of reasons, managers are sometimes reluctant to transmit messages. The reasons could be:-

- They may doubt their ability to do so.
- They may dislike or be weary of writing or talking to others.
- They may hesitate to deliver bad news because they do not want to face a negative reaction.

When someone gives in to these feelings, they become a barrier to effective communications.

i) Projection:

Projection has two meanings.

(a) Projecting one's own motives into others behavior. For example, managers who are motivated by money may assume their subordinates are also motivated by it. If the subordinate's prime motive is something other than money, serious problems may arise.

(b) The use of defense mechanism to avoid placing blame on oneself. As a defense mechanism, the projection phenomenon operates to protect the ego from unpleasant communications. Frequently, individuals who have a particular fault will see the same fault in others, making their own fault seem not so serious.

j) The "Halo Effect":

The term "halo effect" refers to the process of forming opinions based on one element from a group of elements and generalizing that perception to all other elements. For example, in an organization, a good attendance record may cause positive judgments about productivity, attitude, or quality of work. In performance evaluation system, the halo effect refers to the practice of singling out one trait of an employee (either good or bad) and using this as a basis for judgments of the total employee.

CHANNELS OF COMMUNICATION

a) Formal Communication

Formal communication follows the route formally laid down in the organization structure. There are three directions in which communications flow: downward, upward and laterally (horizontal).

i) Downward Communication

Downward communication involves a message travelling to one or more receivers at the lower level in the hierarchy. The message frequently involves directions or performance feedback. The downward flow of communication generally corresponds to the formal organizational communications system, which is usually synonymous with the chain of command or line of authority. This system has received a great deal of attention from both managers and behavioral scientists since it is crucial to organizational functioning.

ii) Upward Communication

In upward communication, the message is directed toward a higher level in the hierarchy. It is often takes the form of progress reports or information about successes and failures of the individuals or work groups reporting to the receiver of the message. Sometimes employees also send suggestions or complaints upward through the organization's hierarchy.

The upward flow of communication involves two distinct manager-subordinate activities in addition to feedback:

- The participation by employees in formal organizational decisions.
- Employee appeal is a result against formal organization decisions. The employee appeal is a result of the industrial democracy concept that provides for two-way communication in areas of disagreement.

iii) Horizontal Communication

When takes place among members of the same work group, among members of work groups at the same level, among managers at the same level or among any horizontally equivalent personnel, we describe it as lateral communications. In lateral communication, the sender and receiver(s) are at the same level in the hierarchy. Formal communications that travel laterally involve employees engaged in carrying out the same or related tasks.

The messages might concern advice, problem solving, or coordination of activities.

b) Informal Communication or Grapevine

Informal communication, generally associated with interpersonal communication, was primarily seen as a potential hindrance to effective organizational performance. This is no longer the case. Informal communication has become more important to ensuring the effective conduct of work in modern organizations.

Probably the most common term used for the informal communication in the workplace is "grapevine" and this communication that is sent through the organizational grapevine is often considered gossip or rumor. While grapevine communication can spread information quickly and can easily cross established organizational boundaries, the information it carries can be changed through the deletion or exaggeration crucial details thus causing the information inaccurate – even if it's based on truth.

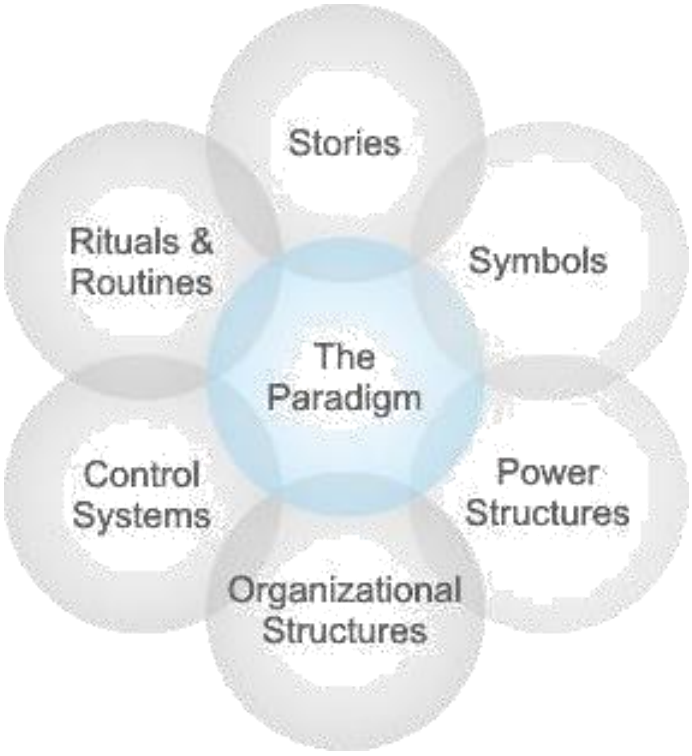
The use of the organizational grapevine as an informal communication channel often results when employees feel threatened, vulnerable, or when the organization is experiencing change and when communication from management is restricted and not forthcoming.

ORGANIZATIONAL CULTURE

Organizational culture is an idea in the field of organizational studies and management which describes the psychology, attitudes, experiences, beliefs and values (personal and cultural values) of an organization. It has been defined as "the specific collection of values and norms that are shared by people and groups in an organization and that control the way they interact with each other and with stakeholders outside the organization."

ELEMENTS OF ORGANIZATIONAL CULTURE

Johnson and Scholes described a cultural web, identifying a number of elements that can be used to describe or influence Organizational Culture:



The six elements are:

a) Stories: The past events and people talked about inside and outside the company. Who and what the company chooses to immortalize says a great deal about what it values, and perceives as great behavior.
b) Rituals and Routines: The daily behavior and actions of people that signal acceptable behavior. This determines what is expected to happen in given situations, and what is valued by management.
c) Symbols: The visual representations of the company including logos, how plush the offices are, and the formal or informal dress codes.
d) Organizational Structure: This includes both the structure defined by the organization chart, and the unwritten lines of power and influence that indicate whose contributions are most valued.
e) Control Systems: The ways that the organization is controlled. These include financial systems, quality systems, and rewards (including the way they are measured and distributed within the organization.)

f) Power Structures: The pockets of real power in the company. This may involve one or two key senior executives, a whole group of executives, or even a department. The key is that these people have the greatest amount of influence on decisions, operations, and strategic direction.

TYPES OF ORGANIZATIONAL CULTURE

Deal and Kennedy argue organizational culture is based on based on two elements:

1. **Feedback Speed**: How quickly are feedback and rewards provided (through which the people are told they are doing a good or a bad job).
2. **Degree of Risk**: The level of risk taking (degree of uncertainty).

The combination of these two elements results in **four types of corporate cultures**:

a) Tough-Guy Culture or Macho Culture (Fast feedback and reward, high risk):

- Stress results from the high risk and the high potential decrease or increase of the reward.
- Focus on now, individualism prevails over teamwork.
- Typical examples: advertising, brokerage, sports.

The most important aspect of this kind of culture is big rewards and quick feedback. This kind of culture is mostly associated with quick financial activities like brokerage and currency trading. It can also be related with activities, like a sports team or branding of an athlete, and also the police team. This kind of culture is considered to carry along, a high amount of stress, and people working within the organization are expected to possess a strong mentality, for survival in the organization.

b) Work Hard/Play Hard (Fast feedback and reward, low risk):

- Stress results from quantity of work rather than uncertainty.
- Focus on high-speed action, high levels of energy.
- Typical examples: sales, restaurants, software companies.

This type of organization does not involve much risk, as the organizations already consist of a firm base along with a strong client relationship. This kind of culture is mostly opted by large

organizations which have strong customer service. The organization with this kind of culture is equipped with specialized jargons and is qualified with multiple team meetings.

c) Bet Your Company Culture (Slow feedback and reward, high risk):

- Stress results from high risk and delay before knowing if actions have paid off.
- Focus on long-term, preparation and planning.
- Typical examples: pharmaceutical companies, aircraft manufacturers, oil prospecting companies.

In this kind of culture, the company makes big and important decisions over high stakes endeavors. It takes time to see the consequence of these decisions. Companies that postulate experimental projects and researches as their core business, adopt this kind of culture. This kind of culture can be adopted by a company designing experimental military weapons for example.

d) Process Culture (Slow feedback and reward, low risk):

- Stress is generally low, but may come from internal politics and stupidity of the system.
- Focus on details and process excellence.
- Typical examples: bureaucracies, banks, insurance companies, public services.

This type of culture does not include the process of feedback. In this kind of culture, the organization is extremely cautious about the adherence to laws and prefer to abide by them. This culture provides consistency to the organization and is good for public services.

One of the most difficult tasks to undertake in an organization, is to change its work culture. An organizational culture change requires an organization to make amendments to its policies, its workplace ethics and its management system. It needs to start right from its base functions which includes support functions, operations and the production floor, which finally affects the overall output of the organization. It requires a complete overhaul of the entire system, and not many organizations prefer it as the process is a long and tedious one, which requires patience and endurance. However, when an organization succeeds in making a change on such a massive level, the results are almost always positive and fruitful. The different types of organizational cultures mentioned above must have surely helped you to understand them. You

can also adopt one of them for your own organization, however, persistence and patience is ultimately of the essence.

MANAGING CULTURAL DIVERSITY

Experts indicate that business owners and managers who hope to create and manage an effective, harmonious multicultural work force should remember the importance of the following:

- Setting a good example—This basic tool can be particularly valuable for small business owners who hope to establish a healthy environment for people of different cultural backgrounds, since they are generally able to wield significant control over the business's basic outlook and atmosphere.
- Communicate in writing—Company policies that explicitly forbid prejudice and discriminatory behavior should be included in employee manuals, mission statements, and other written communications. Jorgensen referred to this and other similar practices as "internal broadcasting of the diversity message in order to create a common language for all members of the organization."
- Training programs—Training programs designed to engender appreciation and knowledge of the characteristics and benefits of multicultural work forces have become ubiquitous in recent years. "Two types of training are most popular: awareness and skill-building," wrote Cox. "The former introduces the topic of managing diversity and generally includes information on work force demographics, the meaning of diversity, and exercises to get participants thinking about relevant issues and raising their own self-awareness. The skill-building training provides more specific information on cultural norms of different groups and how they may affect work behavior." New employee orientation programs are also ideal for introducing workers to the company's expectations regarding treatment of fellow workers, whatever their cultural or ethnic background.
- Recognize individual differences—Writing in The Complete MBA Companion, contributor Rob Goffee stated that "there are various dimensions around which differences in human relationships may be understood. These include such factors as orientation towards authority; acceptance of power inequalities; desire for orderliness and structure; the need to belong to a wider social group and so on. Around these dimensions researchers have demonstrated systematic differences between national, ethnic, and religious groups." Yet Goffee also cautioned business owners, managers, and executives to recognize that differences between individuals can not always be traced back to easily understood

differences in cultural background: "Do not assume differences are always 'cultural.' There are several sources of difference. Some relate to factors such as personality, aptitude, or competence. It is a mistake to assume that all perceived differences are cultural in origin. Too many managers tend to fall back on the easy 'explanation' that individual behavior or performance can be attributed to the fact that someone is 'Italian' or 'a Catholic' or 'a woman.' Such conclusions are more likely to reflect intellectually lazy rather than culturally sensitive managers."

- Actively seek input from minority groups—Soliciting the opinions and involvement of minority groups on important work committees, etc., is beneficial not only because of the contributions that they can make, but also because such overtures confirm that they are valued by the company. Serving on relevant committees and task forces can increase their feelings of belonging to the organization. Conversely, relegating minority members to superfluous committees or projects can trigger a downward spiral in relations between different cultural groups.
- Revamp reward systems—An organization's performance appraisal and reward systems should reinforce the importance of effective diversity management, according to Cox. This includes assuring that minorities are provided with adequate opportunities for career development.
- Make room for social events—Company sponsored social events—picnics, softball games, volleyball leagues, bowling leagues, Christmas parties, etc.—can be tremendously useful in getting members of different ethnic and cultural backgrounds together and providing them with opportunities to learn about one another.
- Flexible work environment—Cox indicated that flexible work environments—which he characterized as a positive development for all workers—could have particularly "beneficial to people from nontraditional cultural backgrounds because their approaches to problems are more likely to be different from past norms."
- Don't assume similar values and opinions—Goffee noted that "in the absence of reliable information there is a well-documented tendency for individuals to assume that others are 'like them.' In any setting this is likely to be an inappropriate assumption; for those who manage diverse work forces this tendency towards 'cultural assimilation' can prove particularly damaging."
- Continuous monitoring—Experts recommend that business owners and managers establish and maintain systems that can continually monitor the organization's policies and practices to ensure that it continues to be a good environment for all employees. This, wrote

Jorgensen, should include "research into employees' needs through periodic attitude surveys."

"Increased diversity presents challenges to business leaders who must maximize the opportunities that it presents while minimizing its costs," summarized Cox. "The multicultural organization is characterized by pluralism, full integration of minority-culture members both formally and informally, an absence of prejudice and discrimination, and low levels of inter-group conflict…. The organization that achieves these conditions will create an environment in which all members can contribute to their maximum potential, and in which the 'value in diversity ' can be fully realized."

UNIT V

CONTROLLING

DEFINITION

Control is the process through which managers assure that actual activities conform to planned activities.

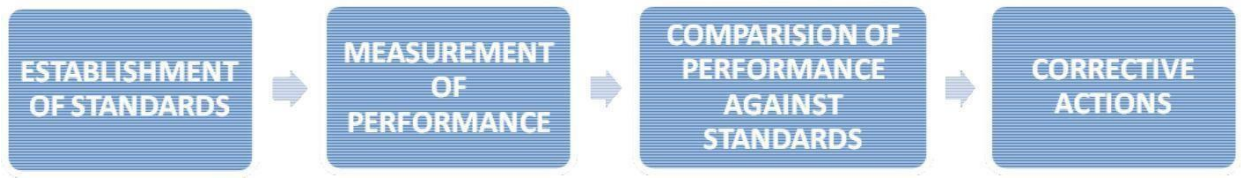
In the words of Koontz and O'Donnell - "Managerial control implies measurement of accomplishment against the standard and the correction of deviations to assure attainment of objectives according to plans."

Nature & Purpose of Control

- Control is an essential function of management
- Control is an ongoing process
- Control is forward – working because pas cannot be controlled
- Control involves measurement
- The essence of control is action
- Control is an integrated system

CONTROL PROCESS

The basic control process involves mainly these steps as shown in Figure



a) The Establishment of Standards:

Because plans are the yardsticks against which controls must be revised, it follows logically that the first step in the control process would be to accomplish plans. Plans can be considered as the criterion or the standards against which we compare the actual performance in order to figure out the deviations.

Examples for the standards

- Profitability standards: In general, these standards indicate how much the company would like to make as profit over a given time period- that is, its return on investment.
- Market position standards: These standards indicate the share of total sales in a particular market that the company would like to have relative to its competitors.
- Productivity standards: How much that various segments of the organization should produce is the focus of these standards.
- Product leadership standards: These indicate what must be done to attain such a position.
- Employee attitude standards: These standards indicate what types of attitudes the company managers should strive to indicate in the company's employees.
- Social responsibility standards: Such as making contribution to the society.
- Standards reflecting the relative balance between short and long range goals.

b) Measurement of Performance:

The measurement of performance against standards should be on a forward looking basis so that deviations may be detected in advance by appropriate actions. The degree of difficulty in measuring various types of organizational performance, of course, is determined primarily by the activity being measured. For example, it is far more difficult to measure the performance of highway maintenance worker than to measure the performance of a student enrolled in a college level management course.

c) Comparing Measured Performance to Stated Standards:

When managers have taken a measure of organizational performance, their next step in controlling is to compare this measure against some standard. A standard is the level of activity established to serve as a model for evaluating organizational performance. The performance evaluated can be for the organization as a whole or for some individuals working within the organization. In essence, standards are the yardsticks that determine whether organizational performance is adequate or inadequate.

d) Taking Corrective Actions:

After actual performance has been measured compared with established performance standards, the next step in the controlling process is to take corrective action, if necessary. Corrective action is managerial activity aimed at bringing organizational performance up to the level of performance standards. In other words, corrective action focuses on correcting organizational mistakes that hinder organizational performance. Before taking any corrective action, however, managers should make sure that the standards they are using were properly established and that their measurements of organizational performance are valid and reliable.

At first glance, it seems a fairly simple proposition that managers should take corrective action to eliminate problems - the factors within an organization that are barriers to organizational goal attainment. In practice, however, it is often difficult to pinpoint the problem causing some undesirable organizational effect.

BARRIERS FOR CONTROLLING

There are many barriers, among the most important of them:

- Control activities can create an undesirable overemphasis on short-term production as opposed to long- term production.
- Control activities can increase employees' frustration with their jobs and thereby reduce morale. This reaction tends to occur primarily where management exerts too much control.
- Control activities can encourage the falsification of reports.
- Control activities can cause the perspectives of organization members to be too narrow for the good of the organization.
- Control activities can be perceived as the goals of the control process rather than the means by which corrective action is taken.

REQUIREMENTS FOR EFFECTIVE CONTROL

The requirements for effective control are

a) Control should be tailored to plans and positions

This means that, all control techniques and systems should reflect the plans they are designed to follow. This is because every plan and every kind and phase of an operation has its unique characteristics.

b) Control must be tailored to individual managers and their responsibilities

This means that controls must be tailored to the personality of individual managers. This because control systems and information are intended to help individual managers carry out their function of control. If they are not of a type that a manager can or will understand, they will not be useful.

c) Control should point up exceptions as critical points

This is because by concentration on exceptions from planned performance, controls based on the time honored exception principle allow managers to detect those places where their attention is required and should be given. However, it is not enough to look at exceptions, because some deviations from standards have little meaning and others have a great deal of significance.

d) Control should be objective

This is because when controls are subjective, a manager's personality may influence judgments of performance inaccuracy. Objective standards can be quantitative such as costs or man hours per unit or date of job completion. They can also be qualitative in the case of training programs that have specific characteristics or are designed to accomplish a specific kind of upgrading of the quality of personnel.

e) Control should be flexible

This means that controls should remain workable in the case of changed plans, unforeseen circumstances, or outsight failures.Much flexibility in control can be provided by having alternative plans for various probable situations.

f) Control should be economical

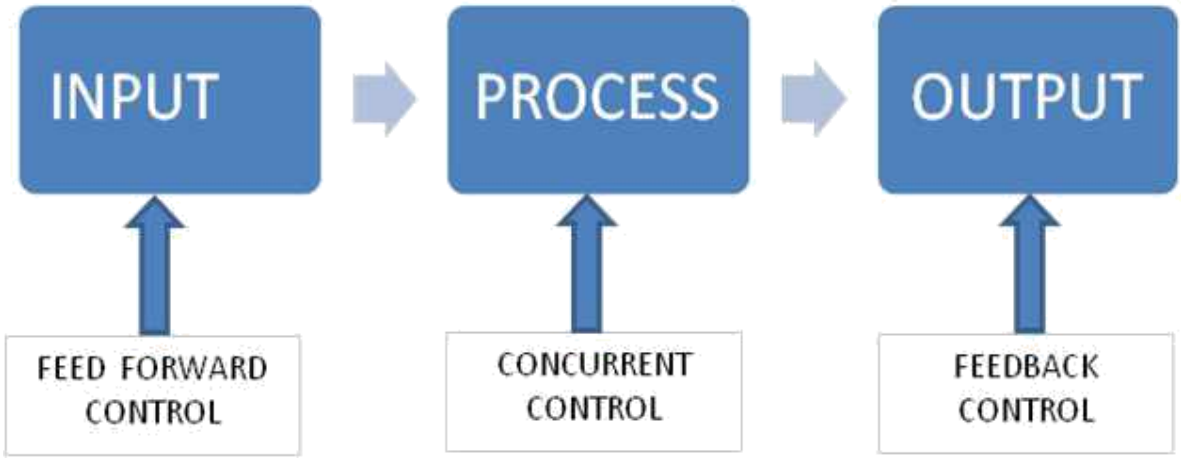
This means that control must worth their cost. Although this requirement is simple, its practice is often complex. This is because a manager may find it difficult to know what a particular system is worth, or to know what it costs.

g) Control should lead to corrective actions

This is because a control system will be of little benefit if it does not lead to corrective action, control is justified only if the indicated or experienced deviations from plans are corrected through appropriate planning, organizing, directing, and leading.

TYPES OF CONTROL SYSTEMS

The control systems can be classified into three types namely feed forward, concurrent and feedback control systems.



a) Feed forward controls: They are preventive controls that try to anticipate problems and take corrective action before they occur. Example – a team leader checks the quality, completeness and reliability of their tools prior to going to the site.

b) Concurrent controls: They (sometimes called screening controls) occur while an activity is taking place. Example – the team leader checks the quality or performance of his members while performing.

c) Feedback controls: They measure activities that have already been completed. Thus corrections can take place after performance is over. Example – feedback from facilities engineers regarding the completed job.

BUDGETARY CONTROL

Definition: Budgetary Control is defined as "the establishment of budgets, relating the responsibilities of executives to the requirements of a policy, and the continuous comparison of actual with budgeted results either to secure by individual action the objective of that policy or to provide a base for its revision.

Salient features:

a. Objectives: Determining the objectives to be achieved, over the budget period, and the policy(ies) that might be adopted for the achievement of these ends.

b. Activities: Determining the variety of activities that should be undertaken for achievement of

the objectives.

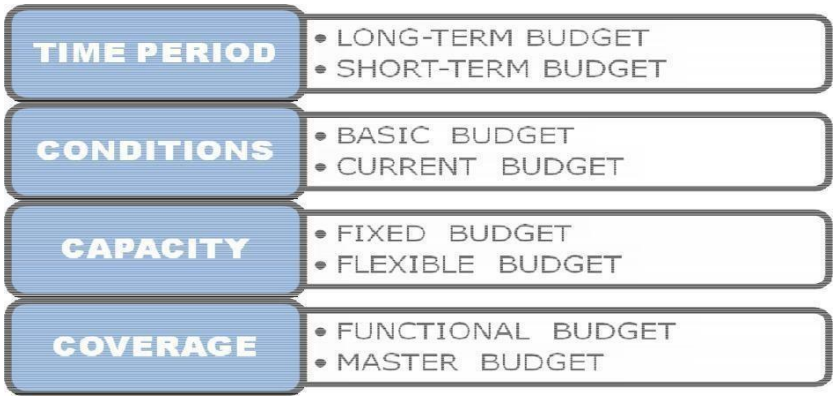
c. Plans: Drawing up a plan or a scheme of operation in respect of each class of activity, in physical a well as monetary terms for the full budget period and its parts.

d. Performance Evaluation: Laying out a system of comparison of actual performance by each person section or department with the relevant budget and determination of causes for the discrepancies, if any.

e. Control Action: Ensuring that when the plans are not achieved, corrective actions are taken; and when corrective actions are not possible, ensuring that the plans are revised and objective achieved

CLASSIFICATION OF BUDGETS

Budgets may be classified on the following bases –



a) BASED ON TIME PERIOD:

(i) Long Term Budget

Budgets which are prepared for periods longer than a year are called LongTerm Budgets. Such Budgets are helpful in business forecasting and forward planning. Eg: Capital Expenditure Budget and R&D Budget.

(ii) Short Term Budget

Budgets which are prepared for periods less than a year are known as ShortTerm Budgets. Such Budgets are prepared in cases where a specific action has to be immediately taken to bring any variation under control.

Eg: Cash Budget.

b) BASED ON CONDITION:

(i) Basic Budget

A Budget, which remains unaltered over a long period of time, is called Basic Budget.

(ii) Current Budget

A Budget, which is established for use over a short period of time and is related to the current conditions, is called Current Budget.

c) BASED ON CAPACITY:

(i) Fixed Budget

It is a Budget designed to remain unchanged irrespective of the level of activity actually attained. It operates on one level of activity and less than one set of conditions. It assumes that there will be no change in the prevailing conditions, which is unrealistic.

(ii) Flexible Budget

It is a Budget, which by recognizing the difference between fixed, semi variable and variable costs is designed to change in relation to level of activity attained. It consists of various budgets for different levels of activity

d) BASED ON COVERAGE:

(i) Functional Budget

Budgets, which relate to the individual functions in an organization, are known as Functional Budgets, e.g. purchase Budget, Sales Budget, Production Budget, plant Utilization Budget and Cash Budget.

(ii) Master Budget

It is a consolidated summary of the various functional budgets. It serves as the basis upon which budgeted Profit & Loss Account and forecasted Balance Sheet are built up.

BUDGETARY CONTROL TECHNIQUES

The various types of budgets are as follows

i) Revenue and Expense Budgets:

The most common budgets spell out plans for revenues and operating expenses in rupee terms. The most basic of revenue budget is the sales budget which is a formal and

detailed expression of the sales forecast. The revenue from sales of products or services furnishes the principal income to pay operating expenses and yield profits. Expense budgets may deal with individual items of expense, such as travel, data processing, entertainment, advertising, telephone, and insurance.

ii) Time, Space, Material, and Product Budgets:

Many budgets are better expressed in quantities rather than in monetary terms. e.g. direct-labor-hours, machine-hours, units of materials, square feet allocated, and units produced. The Rupee cost would not accurately measure the resources used or the results intended.

iii) Capital Expenditure Budgets:

Capital expenditure budgets outline specifically capital expenditures for plant, machinery, equipment, inventories, and other items. These budgets require care because they give definite form to plans for spending the funds of an enterprise. Since a business takes a long time to recover its investment in plant and equipment, (Payback period or gestation period) capital expenditure budgets should usually be tied in with fairly long-range planning.

iv) Cash Budgets:

The cash budget is simply a forecast of cash receipts and disbursements against which actual cash "experience" is measured. The availability of cash to meet obligations as they fall due is the first requirement of existence, and handsome business profits do little good when tied up in inventory, machinery, or other noncash assets.

v) Variable Budget:

The variable budget is based on an analysis of expense items to determine how individual costs should vary with volume of output.

Some costs do not vary with volume, particularly in so short a period as 1 month, 6 months, or a year. Among these are depreciation, property taxes and insurance, maintenance of plant and equipment, and costs of keeping a minimum staff of supervisory and other key personnel. Costs that vary with volume of output range from those that are completely variable to those that are only slightly variable.

The task of variable budgeting involves selecting some unit of measure that reflects volume; inspecting the various categories of costs (usually by reference to the chart of

accounts); and, by statistical studies, methods of engineering analyses, and other means, determining how these costs should vary with volume of output.

vi) Zero Based Budget:

The idea behind this technique is to divide enterprise programs into "packages" composed of goals, activities, and needed resources and then to calculate costs for each package from the ground up. By starting the budget of each package from base zero, budgeters calculate costs afresh for each budget period; thus they avoid the common tendency in budgeting of looking only at changes from a previous period.

Advantages

There are a number of advantages of budgetary control:

- Compels management to think about the future, which is probably the most important feature of a budgetary planning and control system. Forces management to look ahead, to set out detailed plans for achieving the targets for each department, operation and (ideally) each manager, to anticipate and give the organization purpose and direction.
- Promotes coordination and communication.
- Clearly defines areas of responsibility. Requires managers of budget centre's to be made responsible for the achievement of budget targets for the operations under their personal control.
- Provides a basis for performance appraisal (variance analysis). A budget is basically a yardstick against which actual performance is measured and assessed. Control is provided by comparisons of actual results against budget plan. Departures from budget can then be investigated and the reasons for the differences can be divided into controllable and non-controllable factors.
- Enables remedial action to be taken as variances emerge.
- Motivates employees by participating in the setting of budgets.
- Improves the allocation of scarce resources.
- Economises management time by using the management by exception principle.

Problems in budgeting

- Whilst budgets may be an essential part of any marketing activity they do have a number of disadvantages, particularly in perception terms.

- Budgets can be seen as pressure devices imposed by management, thus resulting in:
    a) bad labour relations
    b) inaccurate record-keeping.
- Departmental conflict arises due to:
    a) disputes over resource allocation
    b) departments blaming each other if targets are not attained.
- It is difficult to reconcile personal/individual and corporate goals.
- Waste may arise as managers adopt the view, "we had better spend it or we will lose it". This is often coupled with "empire building" in order to enhance the prestige of a department.
- Responsibility versus controlling, i.e. some costs are under the influence of more than one person, e.g. power costs.
- Managers may overestimate costs so that they will not be blamed in the future should they overspend.

NON-BUDGETARY CONTROL TECHNIQUES

There are, of course, many traditional control devices not connected with budgets, although some may be related to, and used with, budgetary controls. Among the most important of these are: statistical data, special reports and analysis, analysis of break- even points, the operational audit, and the personal observation.

i) Statistical data:

Statistical analyses of innumerable aspects of a business operation and the clear presentation of statistical data, whether of a historical or forecast nature are, of course, important to control. Some managers can readily interpret tabular statistical data, but most managers prefer presentation of the data on charts.

ii) Break- even point analysis:

An interesting control device is the break even chart. This chart depicts the relationship of sales and expenses in such a way as to show at what volume revenues exactly cover expenses.

iii) Operational audit:

Another effective tool of managerial control is the internal audit or, as it is now coming to be called, the operational audit. Operational auditing, in its broadest sense, is the regular and independent appraisal, by a staff of internal auditors, of the accounting, financial, and other operations of a business.

iv) Personal observation:

In any preoccupation with the devices of managerial control, one should never overlook the importance of control through personal observation.

v) PERT:

The Program (or Project) Evaluation and Review Technique, commonly abbreviated PERT, is a is a method to analyze the involved tasks in completing a given project, especially the time needed to complete each task, and identifying the minimum time needed to complete the total project.

vi) GANTT CHART:

A Gantt chart is a type of bar chart that illustrates a project schedule. Gantt charts illustrate the start and finish dates of the terminal elements and summary elements of a project. Terminal elements and summary elements comprise the work breakdown structure of the project. Some Gantt charts also show the dependency (i.e., precedence network) relationships between activities.

PRODUCTIVITY

Productivity refers to the ratio between the output from production processes to its input. Productivity may be conceived of as a measure of the technical or engineering efficiency of production. As such quantitative measures of input, and sometimes output, are emphasized.

Typical Productivity Calculations

Measures of size and resources may be combined in many different ways. The three common approaches to defining productivity based on the model of Figure 2 are referred to as physical, functional, and economic productivity. Regardless of the approach selected, adjustments may be needed for the factors of diseconomy of scale, reuse, requirements churn, and quality at delivery.

a) Physical Productivity

This is a ratio of the amount of product to the resources consumed (usually effort). Product may be measured in lines of code, classes, screens, or any other unit of product. Typically, effort is measured in terms of staff hours, days, or months. The physical size also may be used to estimate software performance factors (e.g., memory utilization as a function of lines of code).

b) Functional Productivity

This is a ratio of the amount of the functionality delivered to the resources consumed (usually effort). Functionality may be measured in terms of use cases, requirements, features, or function points (as appropriate to the nature of the software and the development method). Typically, effort is measured in terms of staff hours, days, or months. Traditional measures of Function Points work best with information processing systems. The effort involved in embedded and scientific software is likely to be underestimated with these measures, although several variations of Function Points have been developed that attempt to deal with this issue.

c) Economic Productivity

This is a ratio of the value of the product produced to the cost of the resources used to produce it. Economic productivity helps to evaluate the economic efficiency of an organization. Economic productivity usually is not used to predict project cost because the outcome can be affected by many factors outside the control of the project, such as sales volume, inflation, interest rates, and substitutions in resources or materials, as well as all the other factors that affect physical and functional measures of productivity. However, understanding economic productivity is essential to making good decisions about outsourcing and subcontracting. The basic calculation of economic productivity is as follows:

Economic Productivity = Value/Cost

PROBLEMS IN MEASUREMENT OF PRODUCTIVITY OF KNOWLEDGE WORKERS

Productivity implies measurement, which in turn, is an essential step in the control process. Although there is a general agreement about the need for improving productivity, there

is little consensus about the fundamental causes of the problem and what to do about them. The blame has been assigned to various factors. Some people place it on the greater proportion of less skilled workers with respect to the total labor force, but others disagree. There are those who see cutback in research and the emphasis on immediate results as the main culprit. Another reason given for the productivity dilemma is the growing affluence of people, which makes them less ambitious. Still others cite the breakdown in family structure, the workers' attitudes, and government policies and regulations. Another problem is that the measurement of skills work is relatively easy, but it becomes more difficult for knowledge work. The difference between the two kinds is the relative use of knowledge and skills.

COST CONTROL

Cost control is the measure taken by management to assure that the cost objectives set down in the planning stage are attained and to assure that all segments of the organization function in a manner consistent with its policies.

Steps involved in designing process of cost control system:

- **Establishing norms:** To exercise cost control it is essential to establish norms, targets or parameters which may serve as yardsticks to achieve the ultimate objective. These standards, norms or targets may be set on the basis of research, study or past actual.
- **Appraisal:** The actual results are compared with the set norms to ascertain the degree of utilization of men, machines and materials. The deviations are analyzed so as to arrive at the causes which are controllable and uncontrollable.
- **Corrective measures:** The variances are reviewed and remedial measures or revision of targets, norms, standards etc., as required are taken.

Advantages of cost control

- Better utilization of resources
- To prepare for meeting a future competitive position.
- Reasonable price for the customers
- Firm standing in domestic and export markets.
- Improved methods of production and use of latest manufacturing techniques which have the effect of rising productivity and minimizing cost.

- By a continuous search for improvement creates proper climate for the increase efficiency.
- Improves the image of company for long-term benefits.
- Improve the rate of return on investment.

PURCHASE CONTROL

Purchase control is an element of material control. Material procurement is known as the purchase function. The functional responsibility of purchasing is that of the purchase manager or the purchaser. Purchasing is an important function of materials management because in purchase of materials, a substantial portion of the company's finance is committed which affects cash flow position of the company. Success of a business is to a large extent influenced by the efficiency of its purchase organization. The advantages derived from a good and adequate system of the purchase control are as follows:

a) Continuous availability of materials: It ensures the continuous flow of materials. so production work may not be held up for want of materials. A manufacturer can complete schedule of production in time.

b) Purchasing of right quantity: Purchase of right quantity of materials avoids locking up of working capital. It minimizes risk of surplus and obsolete stores. It means there should not be possibility of overstocking and understocking.

c) Purchasing of right quality: Purchase of materials of proper quality and specification avoids waste of materials and loss in production. Effective purchase control prevents wastes and losses of materials right from the purchase till their consumptions. It enables the management to reduce cost of production.

d) Economy in purchasing: The purchasing of materials is a highly specialized function. By purchasing materials at reasonable prices, the efficient purchaser is able to make a valuable contribution to the success of a business.

e) Works as information centre: It serves as a function centre on the materials knowledge relating to prices, sources of supply, specifications, mode of delivery, etc. By providing continuous information to the management it is possible to prepare planning for production.

f) Development of business relationship: Purchasing of materials from the best market and from reliable suppliers develops business relationships. The result is that there may be smooth supply of materials in time and so it avoid disputes and financial losses.

g) Finding of alternative source of supply: If a particular supplier fails to supply the materials in time, it is possible to develop alternate sources of supply. the effect of this is that the production work is not disturbed.

h) Fixing responsibilities: Effective purchase control fix the responsibilities of operating units and individuals connected with the purchase, storage and handling of materials.

In short, the basic objective of the effective purchase control is to ensure continuity of supply of requisite quantity of material, to avoid held up of production and loss in production and at the same time reduces the ultimate cost of the finished products.

MAINTENANCE CONTROL

Maintenance department has to excercise effective cost control, to carry out the maintenance functions in a pre-specified budget, which is possible only through the following measures:

First line supervisors must be apprised of the cost information of the various materials so that the objective of the management can be met without extra expenditure on maintenance functions

A monthly review of the budget provisions and expenditures actually incurred in respect of each center/shop will provide guidlines to the departmental head to exercise better cost control.

The total expenditure to be incurred can be uniformly spread over the year for better budgetary control. however, the same may not be true in all cases particularly where overhauling of equipment has to be carried out due to unforseen breakdowns. some budgetary provisions must be set aside, to meet out unforeseen exigencies.

The controllable elements of cost such as manpower cost and material cost can be discussed with the concerned personnel, which may help in reducing the total cost of maintenance. Emphasis should be given to reduce the overhead expenditures, as other expenditures cannot be compromised.

It is observed through studies that the manpower cost is normally fixed, but the same way increase due to overtime cost. however, the material cost, which is the prime factor in maintenance cost, can be reduced by timely inspections designed, to detect failures. If the

inspection is carried out as per schedule, the total failure of parts may be avoided, which otherwise would increase the maintenance cost. the proper handling of the equipment by the operators also reduces the frequency of repair and material requirements. Operators, who check their equipment regularly and use it within the operating limits, can help avoid many unwanted repairs. In the same way a good record of equipment failures/ maintenance would indicate the nature of failures, which can then be corrected even permanently.

QUALITY CONTROL

Quality control refers to the technical process that gathers, examines, analyze & report the progress of the project & conformance with the performance requirements

The steps involved in quality control process are

1) Determine what parameter is to be controlled.
2) Establish its criticality and whether you need to control before, during or after results are produced.
3) Establish a specification for the parameter to be controlled which provides limits of acceptability and units of measure.
4) Produce plans for control which specify the means by which the characteristics will be achieved and variation detected and removed.
5) Organize resources to implement the plans for quality control.
6) Install a sensor at an appropriate point in the process to sense variance from specification.
7) Collect and transmit data to a place for analysis.
8) Verify the results and diagnose the cause of variance.
9) Propose remedies and decide on the action needed to restore the status quo.
10) Take the agreed action and check that the variance has been corrected.

Advantages and disadvantages

- Advantages include better products and services ultimately establishing a good reputation for a company and higher revenue from having more satisfied customers.
- Disadvantages include needing more man power/operations to maintain quality control and adding more time to the initial process.

PLANNING OPERATIONS

An **operational planning** is a subset of strategic work plan. It describes short-term ways of achieving milestones and explains how, or what portion of, a strategic plan will be put into operation during a given operational period, in the case of commercial application, a fiscal year or another given budgetary term. An operational plan is the basis for, and justification of an annual operating budget request. Therefore, a five-year strategic plan would need five operational plans funded by five operating budgets.

Operational plans should establish the activities and budgets for each part of the organization for the next 1 – 3 years. They link the strategic plan with the activities the organization will deliver and the resources required to deliver them.

An operational plan draws directly from agency and program strategic plans to describe agency and program missions and goals, program objectives, and program activities. Like a strategic plan, an operational plan addresses four questions:

- Where are we now?
- Where do we want to be?
- How do we get there?
- How do we measure our progress?

The OP is both the first and the last step in preparing an operating budget request. As the first step, the OP provides a plan for resource allocation; as the last step, the OP may be modified to reflect policy decisions or financial changes made during the budget development process.

Operational plans should be prepared by the people who will be involved in implementation.

There is often a need for significant cross-departmental dialogue as plans created by one part of the organization inevitably have implications for other parts. Operational plans should contain:

- clear objectives
- activities to be delivered
- quality standards
- desired outcomes
- staffing and resource requirements
- implementation timetables
- a process for monitoring progress.

MADHA

Expertise | Empathy | Excellence

ENGINEERING COLLEGE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

COMMON FOR: DEPARTMENT OF INFORMATION TECHNOLOGY

CS8792 – CRYPTOGRAPHY AND NETWORK SECURITY

R – 2017

LECTURE NOTES

DIFFIE HELLMAN KEY EXCHANGE

Example :

$q = 11$, $\alpha = 2$, $X_A = 6$, $X_B = 8$.

1. Choosen the prime number $q = 11$.
   Choose α and it is a primitive root of q. $\alpha = 2$.

2. Choose any random value key for user A and User B.

   $X_A = 6$ $\qquad$ $X_B = 8$.

3. Find $Y_A = \alpha^{X_A} \bmod q$ $\qquad$ $Y_B = \alpha^{X_B} \bmod q$

   $Y_A = 2^6 \bmod 11$ $\qquad$ $Y_B = 2^8 \bmod 11$

   $= 64 \bmod 11$ $\qquad$ $= 256 \bmod 11$

   $\boxed{Y_A = 9}$ $\qquad$ $\boxed{Y_B = 3}$

Y_A is tranferred to User B.
Y_B is transferred to User A.

Shared Secret Key K of User A

$$K = (Y_B)^{X_A} \bmod q$$

$$K = 3^6 \bmod 11$$

$$\equiv 729 \bmod 11$$

$$= 3.$$

Shared Secret Key K of User B.

$$K = (Y_A)^{X_B} \bmod q$$

$$= 9^8 \bmod 11$$

$$= 43046721 \bmod 11 = 2.$$

To find $43046721 \bmod 11$

$43046721 / 11 = 3913338$

$3913338 \times 11 = 43046718$

$43046721 - 43046718 = \boxed{3}$

RSA Algorithm

Example: 2. $M = 8$, $e = 17$

1. Choose 2 prime numbers.

$p = 7$, $q = 11$, e.

2. $n = p \times q$

$= 7 \times 11$

$n = 77$

3. $\varphi(n) = (p-1)(q-1)$

$= (7-1)(11-1)$

$= 6 \times 10$

$\varphi(n) = 60$

4. choose e such that i) $1 < e < \varphi(n)$.

& $\gcd(e, \varphi(n)) = 1$. $1 < 17 < \varphi(n)$.

$\gcd(e, 60) = 1$

$\gcd(17, 60) = 1$.

5. $d = \dfrac{1 + k \cdot \varphi(n)}{e}$

$$d = \frac{1 + \phi(n)}{e}$$

$$d = \frac{1 + 60}{17} = \frac{61}{17} = 3.588$$

Substitute $k = 0$ to 16.

K=0 $d = \frac{1 + k \cdot \phi(n)}{e} = \frac{1 + 0.60}{17} = 1/17$.

K=1 $d = \frac{1+60}{17} = 3.58$

K=2 $d = \frac{1 + 2 \times 60}{17} = \frac{121}{17} = 7.11$

K=3 $d = \frac{1 + 3 * 60}{17} = \frac{181}{17} = 10.6$

K=4 $d = \frac{1 + 4 * 60}{17} = \frac{240}{17} = 14.1$

K=5 $d = \frac{1 + 5 * 60}{17} = \frac{301}{17} = 17.7$

- - - - - -

K=15 $d = \frac{1 + 15 * 60}{17} = \frac{901}{17} = 53$ whole no.

$d = 53$

Public Key $K_U = \{e, n\}$

$K_U = \{17, 77\}$

Private Key $K_R = \{d, n\}$

$K_R = \{53, 77\}$.

Encryption:

$C = M^e \bmod n$

$C = 8^{17} \bmod 77$

$= 57$.

How to find: $8^{17} \bmod 77$.

$8^{17} \bmod 77 = 8^4 \cdot (8 \cdot 8 \cdot 8^4 \cdot 8^4 \cdot 8^1) \bmod 77$

$= [8^4 \bmod 77 \times 8^4 \bmod 77 \times 8^4 \bmod 77 \times 8^4 \bmod 77 \times 8^1 \bmod 77] \bmod 77$

$= [15 \times 15 \times 15 \times 15 \times 8] \bmod 77$

$= 405000 \bmod 77$

$8^{17} \bmod 77 = 57.$

Note:

$8^4 \bmod 77 = 4096 \bmod 77$

$= 15.$

$4096 \bmod 77 = ?$

1) $4096 / 77 = 53$

2) $53 \times 77 = 4081$

3) $4096 - 4081 = 15$

Decryption:

$M = C^d \bmod n$

$= 57^{53} \bmod 77.$

$= (57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^5 \cdot 57^3) \bmod 77.$

$$= (57^5 \bmod 77 \times 54^5 \bmod 77 \times 57^5 \bmod 77 \times$$
$$57^5 \bmod 77 \times 57^5 \bmod 77 \times 57^5 \bmod 77 \times$$
$$57^5 \bmod 77 \times 57^5 \bmod 77 \times 57^5 \bmod 77$$
$$\times 57^3 \bmod 77) \bmod 77$$

$$57^5 \bmod 77 = 60169205 \bmod 77$$
$$= 43.$$

$57^3 \bmod 77 = 185193 \bmod 77$

$= 8$

$M = (43 \times 43 \times 43 \times 43 \times 43 \times 43 \times 43 \times 43 \times 43 \times 43 \times 43 \times 8) \bmod 77$

$= (43^5 \cdot 43^5 \cdot 8) \bmod 77$

$= (43^5 \bmod 77 \times 43^5 \bmod 77 \times 8 \bmod 77) \times \bmod 77$

$43^5 \bmod 77 = 147008443 \bmod 77$

$= 43$

$= (43 \times 43 \times 8) \bmod 77$

$= 14792 \bmod 77$

$M = 8.$

EULER'S THEOREM

- called as Euler totient function.

Theorem:

It states that if x and n are co-prime +ve integers, then $x^{\phi(n)} \equiv 1 \bmod n$.

where $\phi(n)$ is Euler Totient function.

$\Downarrow$

$x^{\phi(n)} \bmod n = 1 \bmod n$

$\phi(n) = n - 1.$

$\phi(a * b) = \phi(a) * \phi(b)$

Example:

$x = 11, \; n = 10$. Both are co-prime.

∴ we can represent them as

$x^{\phi(n)} \equiv 1 \bmod n.$

$11^{\phi(10)} \equiv 1 \bmod 10.$

$\phi(10) = \phi(2 \times 5) = \phi(2) * \phi(5)$

$= 2 - 1 * 5 - 1$

$= 1 * 4$

$\phi(10) = 4$

Substitute $\varphi(10) = 4$

$$11^{\varphi(10)} \equiv 1 \bmod 10$$

$$11^{4} \equiv 1 \bmod 10.$$

$$14641 \equiv 1 \bmod 10 \quad \text{or}$$

$$14641 \bmod 10 = 1.$$

Note: $x^{\varphi(n)\cdot a} \equiv 1 \bmod n.$

ie) $11^{4\cdot 2} \equiv 1 \bmod n$

$$11^{8} \equiv 1 \bmod 10$$

$$214358881 \equiv 1 \bmod 10$$

$214358881 \bmod 10 = 1$. true.

ie) Any multiple of $\varphi(n)$ will give the same result.

n	$\varphi(n)$	nos. coprime to n.	n	$\varphi(n)$	nos coprime to n
1	1	1	7	6	1, 2, 3, 4, 5, 6
2	1	1	8	4	1, 3, 5, 7
3	2	1, 2	9	6	1, 2, 4, 5, 7, 8
4	2	1, 3	10	4	1, 3, 7, 9
5	4	1, 2, 3, 4			
6	2	1, 5			

Note: Two integers a, b are said to be relatively prime, mutually prime or co-prime, if the only +ve integer / factor that divides both of them is 1.

* If n is prime, $\phi(n) = n-1$.
* If n is not prime, $\phi(a*b) = \phi(a) * \phi(b)$.
  - a and b should be co-prime.

* Eg: $n = 11$.

$\phi(n) = n-1$

$\phi(11) = 11-1 = 10$.

$\boxed{\phi(11) = 10}$

* Eg: $n = 35$

$\phi(a*b) = \phi(a) * a(b)$

$\phi(35) = \phi(5*7)$

$= \phi(5) * \phi(7)$

$= 4 * 6$

$\boxed{\phi(35) = 24}$

* Euler totient function ϕn is defined as the no. of +ve integers less than n that are co-prime to n.

(1) $\phi(6) = \{1, 5\}$. $n = 6$.

No. of +ve integers less than n } 1, 2, 3, 4, 5.
6 6 6
× × ×

These integers should not have common divisor with n.

$\phi(6) = \{1, 5\}$.

No. of elements in these set is the totient function.

(2) $\phi(8) = ?$ $n = 8$.

No. of +ve ints less than $n(8) = 1, \underline{2}, 3, \underline{4}, 5, \underline{6}, 7$

Nos. have common divisor with 8 = 2, 4, 6

Except these numbers = $\{1, 3, 5, 7\}$.

$\phi(8) = \{1, 3, 5, 7\}$.

$\phi(8) = 4$.

CHINESE REMAINDER THEOREM

Given :

$x \equiv a_1 \pmod{m_1}$

$x \equiv a_2 \pmod{m_2}$

$x \equiv a_3 \pmod{m_3}$

$x \equiv a_4 \pmod{m_4}$ ..

* These are called as simultaneous eqns.
* CRT is used to solve simultaneous equations.

* Given $x \equiv a_i \pmod{m_i}$
* Check whether m_1, m_2, m_3, m_4 .. are relatively prime or coprime. For coprime $\gcd(m_1, m_2) = 1$
* Calculate $M = m_1 \times m_2 \times m_3 \ldots m_i$
* Calculate $M_i = \frac{M}{m_i}$, $M_1 = \frac{M}{m_1}$, $M_2 = \frac{M}{m_2}$, $M_3 = \frac{M}{m_3}$ ...
* Calculate M_i^{-1}.

$$M_i \cdot M_i^{-1} \equiv 1 \pmod{m_i}$$

* Calculate $x = \sum \left[a_i \times M_i \times M_i^{-1} \right] \bmod M$.

NOTE:

If m_i is prime, $M_i^{-1} = M_i^{P-2} \bmod P$

By fermat's little theorem,

$$a^{-1} = a^{P-2} \pmod P$$

CRT states that there always exists x an 'x' that satisfies the given congruence.

$x \equiv a_1 \bmod m_1$

$x \equiv a_2 \bmod m_2$ ..

and m_1, m_2.. must be co-prime to one another.

Example:

$x \equiv 1 \bmod 5$ * find x?

$x \equiv 1 \bmod 7$

$x \equiv 3 \bmod 11$.

Solution:

Given:

$a_1 = 1 \quad a_2 = 1 \quad a_3 = 3$

$m_1 = 5 \quad m_2 = 7 \quad m_3 = 11$.

i) Check m_1, m_2 & m_3 are co-prime to each other.

$\gcd(5, 7) = 1$

$\gcd(5, 11) = 1$

$\gcd(7, 11) = 1$.

So m_1, m_2, m_2 are co-prime to each other.

ii) Calculate $M = m_1 \times m_2 \times m_3$

$M = 5 \times 7 \times 11$

$\boxed{M = 385}$

iii) Calculate M_1, M_2, M_3.

$M_1 = \frac{M}{m_1}$, $M_2 = \frac{M}{m_2}$, $M_3 = \frac{M}{m_3}$

$$M_1 = \frac{385}{5} = 77$$

$$M_2 = \frac{385}{7} = 55$$

$$M_3 = \frac{385}{11} = 35$$

$M_1 = 77$
$M_2 = 55$
$M_3 = 35$

$M = 385$
$m_1 = 5$
$m_2 = 7$
$m_3 = 11$

iv) Calculate M_i^{-1}.

Since m_1, m_2 and m_3 are prime numbers, we can find multiplicative inverse using Fermat theorem.

$$a^{-1} = a^{p-2} \bmod p.$$

To find M_1^{-1}:

$M_1 \cdot M_1^{-1} \equiv 1 \bmod m_i$

$M_1 = 77 \quad M_1^{-1} = ? \quad m_1 = 5$

$a = 77 \quad a^{-1} = ? \quad p = 5$

$$a^{-1} = a^{p-2} \bmod p$$

$$M_1^{-1} = M_1^{m_1-2} \bmod m_1$$

$$M_1^{-1} = 77^{5-2} \bmod 5$$

$$M_1^{-1} = 77^3 \bmod 5 = 456533 \bmod 5$$

$$M_1^{-1} = 3$$

$M_1 \cdot M_1^{-1} \equiv 1 \bmod m_1$

$x \times \frac{1}{x} = 1$

(17)

Find M_2^{-1}:

$M_2 = 55$, $M_2^{-1} = ?$ $m_2 = 7$

$M_2^{-1} = M_2^{m_2-2} \bmod 7$

$M_2^{-1} = 55^{7-2} \bmod 7$

$= 55^5 \bmod 7$

$= 503,284,375 \bmod 7$

$M_2^{-1} = 6.$

Find M_3^{-1}:

$M_3 = 35$, $M_3^{-1} = ?$ $m_3 = 11$.

$M_3^{-1} = M_3^{m_3-2} \bmod m_3$

$= 35^{11-2} \bmod 11$

$= 35^9 \bmod 11 = (35^4 \times 35^5) \bmod 11$

4774715 136420

$= (35^4 \bmod 11 \times 35^5 \bmod 11) \bmod 11$

$= (1500625 \bmod 11 \times 52521875 \bmod 11) \bmod 11$

$= (5 \times 10) \bmod 11$

$= 50 \bmod 11$

$M_3^{-1} = 6$

$M_1^{-1} = 3$

$M_2^{-1} = 6$

$M_3^{-1} = 6$

v) Calculate x.

$$x = (a_1 M_1 M_1^{-1} + a_2 M_2 M_2^{-1} + a_3 M_3 M_3^{-1}) \bmod M$$

$a_1 = 1$	$a_2 = 1$	$a_3 = 3$	$M = 385$
$M_1 = 77$	$M_2 = 55$	$M_3 = 35$	
$M_1^{-1} = 3$	$M_2^{-1} = 6$	$M_3^{-1} = 6$	

$x = (1 \times 77 \times 3 + 1 \times 55 \times 6 + 3 \times 35 \times 6) \bmod 385$

$= (231 + 330 + 630) \bmod 385$

$= 1191 \bmod 385$

$x = 36$

(19)

We will verify the answer.

$x \equiv 1 \bmod 5$	$36 \equiv 1 \bmod 5$	$36 \bmod 5 = 1$
$x \equiv 1 \bmod 7$	$36 \equiv 1 \bmod 7$	$36 \bmod 7 = 1$
$x \equiv 3 \bmod 11$	$36 \equiv 3 \bmod 11$	$36 \bmod 11 = 3$

UNIT III
PUBLIC KEY CRYPTOGRAPHY

MATHEMATICS OF ASYMMETRIC KEY CRYPTOGRAPHY: Primes – Primality Testing – Factorization – Euler's totient function, Fermat's and Euler's Theorem – Chinese Remainder Theorem – Exponentiation and logarithm – ASYMMETRIC KEY CIPHERS: RSA cryptosystem – Key distribution – Key management – Diffie Hellman key exchange – ElGamal cryptosystem – Elliptic curve arithmetic-Elliptic curve cryptography.

MATHEMATICS OF ASYMMETRIC KEY CRYPTOGRAPHY

5.1. PRIMES

- An integer $p > 1$ is a prime number if and only if its only divisors are ± 1 and $\pm p$. **Prime numbers** play a critical role in number theory.. In particular, note the number of primes in each range of 100 numbers.

  Any integer $a > 1$ can be factored in a unique way as

  $$a = p_1^{a_1} \times p_2^{a_2} \times \cdots \times p_t^{a_t} \qquad (8.1)$$

- where $p_1 < p_2 < \ldots . < p_t$ are prime numbers and where each ai is a positive integer. This is known as the fundamental theorem of arithmetic; a proof can be found in any text on number theory.

  $$91 = 7 \times 13$$
  $$3600 = 2^4 \times 3^2 \times 5^2$$
  $$11011 = 7 \times 11^2 \times 13$$

- It is useful for what follows to express this another way. If P is the set of all prime numbers, then any positive integer a can be written uniquely in the following form:

  $$a = \prod_{p \in P} p^{a_p} \quad \text{where each } a_p \geq 0$$

- The right-hand side is the product over all possible prime numbers p; for any particular value of a, most of the exponents ap will be 0.
- The value of any given positive integer can be specified by simply listing all the nonzero exponents in the foregoing formulation.

  The integer 12 is represented by $\{a_2 = 2, \ a_3 = 1\}$.
  The integer 18 is represented by $\{a_2 = 1, \ a_3 = 2\}$.
  The integer 91 is represented by $\{a_7 = 1, a_{13} = 1\}$.

- Multiplication of two numbers is equivalent to adding the corresponding exponents. Given $a = \prod_{p \in P} p^{a_p}, \ b = \prod_{p \in P} p^{b_p}$.
- Define k = ab. We know that the integer k can be expressed as the product of powers of primes: $k = \prod_{p \in P} p^{k_p}$.

It follows that $k_p = a_p + b_p$ for all $p \in P$.

$$k = 12 \times 18 = (2^2 \times 3) \times (2 \times 3^2) = 216$$
$$k_2 = 2 + 1 = 3; \ k_3 = 1 + 2 = 3$$
$$216 = 2^3 \times 3^3 = 8 \times 27$$

- What does it mean, in terms of the prime factors of *a* and*b*, to say that a divides b? Any integer of the form p^n can be divided only by an integer that is of a lesser or equal power of the same prime number, $p^j$with $j \leq n$. Thus, we can say the following.

  Given

  $$a = \prod_{p \in P} p^{a_p}, b = \prod_{p \in P} p^{b_p}$$

  If a|b, then $a_p \leq b_p$ for all p.

  $a = 12; b = 36; 12|36$
  $12 = 2^2 \times 3; 36 = 2^2 \times 3^2$
  $a_2 = 2 = b_2$
  $a_3 = 1 \leq 2 = b_3$
  Thus, the inequality $a_p \leq b_p$ is satisfied for all prime numbers.

It is easy to determine the greatest common divisor of two positive integers if we express each integer as the product of primes.

$$300 = 2^2 \times 3^1 \times 5^2$$
$$18 = 2^1 \times 3^2$$
$$\gcd(18, 300) = 2^1 \times 3^1 \times 5^0 = 6$$

The following relationship always holds:

If $k = \gcd(a, b)$, then $k_p = \min(a_p, b_p)$ for all p.

Determining the prime factors of a large number is no easy task, so the preceding relationship does not directly lead to a practical method of calculating the greatest common divisor.

5.2. PRIMALITY TESTING

Contents
• **Testing for primality** ✓ **Miller-Rabin Algorithm** • **Two Properties of Prime Numbers** • **Details of the Algorithm** • **A Deterministic Primality Algorithm** • **Distribution of Primes**

Testing for primality:

Miller-Rabin Algorithm

- The algorithm due to Miller and Rabin [MILL75, RABI80] is typically used to test a large number for primality. Before explaining the algorithm, we need some background.
- First, any positive odd integer n ≥ 3 can be expressed as
  $n - 1 = 2^k q$ with $k > 0, q$ odd

Two Properties of Prime Numbers

The first property is stated as follows:

- If p is prime and a is a positive integer less than p, then a^2 mod p = 1 if and only if either a mod p = 1 or a mod p = −1 mod p = p −1. By the rules of modular arithmetic (a mod p) (a mod p) = a^2 mod p.

The second property is stated as follows:

1. a^q is congruent to *1 modulo p*. That is, a^q mod p = 1, or equivalently, $a^q \equiv 1 (\text{mod } p)$.
2. One of the numbers a^q, $a^{2qa2k-l}q$ is congruent to −1 modulo p.

Details of the Algorithm

- The procedure TEST takes a candidate integer n as input and returns the result composite if n is definitely not a prime, and the result inconclusive if n may or may not be a prime.

```
TEST (n)
1. Find integers k, q, with k > 0, q odd, so that
   (n − 1 = 2^k q);
2. Select a random integer a, 1 < a < n − 1;
3. if a^q mod n = 1 then return("inconclusive");
4. for j = 0 to k − 1 do
5. if a^(2^j q) mod n = n − 1 then return("inconclusive");
6. return("composite");
```

A Deterministic Primality Algorithm

- All of the algorithms in use, including the most popular (Miller-Rabin), produced a probabilistic result.
- AKS developed a relatively simple deterministic algorithm that efficiently determines whether a given large number is a prime. The algorithm, known as the AKS algorithm, does not appear to be as efficient as the Miller- Rabin algorithm.

Distribution of Primes

- A result from number theory, known as the prime number theorem, states that the primes near n are spaced on the average one every ln(n) integers.
- Thus, on average, one would have to test on the order of ln(n) integers before a prime is found. Because all even integers can be immediately rejected, the correct figure is 0.5 ln(n).

5.3. FACTORIZATION

5.4. EULER'S TOTIENT FUNCTION

- ✓ Euler's totient function Φ(n) defined as the number of positive integers less than n and Relatively prime to n. by conversion Φ(1)=1.

5.5. FERMAT'S AND EULER'S THEOREM

Contents
• **Fermat's Theorem**   • Proof: • **Euler's Totient Function**

- **Euler's Theorem**
  - Proof:

Two theorems that play important roles in public-key cryptography are Fermat's theorem and Euler's theorem.

Fermat's Theorem

- Fermat's theorem states the following: If p is prime and a is a positive integer not divisible by p, then

$$a^{p-1} \equiv 1 \pmod{p} \qquad (8.2)$$

Proof:

- Consider the set of positive integers less than $p = \{1, 2, \ldots, p - 1\}$ and multiply each element by " a modulo p", to get the set $X = \{a \bmod p, 2a \bmod p, \ldots, (p - 1)a \bmod p\}$.
- None of the elements of X is equal to zero because p does not divide a.
- Multiplying the numbers in both sets (p and X) and taking the result mod p yields

$$a \times 2a \times \cdots \times (p-1)a \equiv [(1 \times 2 \times \cdots \times (p-1)] \pmod{p}$$
$$a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$$

- We can cancel the (p - 1)! term because it is relatively prime to p .

$$a^{p-1} \equiv 1 \pmod{p}$$

- Hence proved.
- Example

$$a = 7, p = 19$$
$$7^2 = 49 \equiv 11 \pmod{19}$$
$$7^4 \equiv 121 \equiv 7 \pmod{19}$$
$$7^8 \equiv 49 \equiv 11 \pmod{19}$$
$$7^{16} \equiv 121 \equiv 7 \pmod{19}$$
$$a^{p-1} = 7^{18} = 7^{16} \times 7^2 \equiv 7 \times 11 \equiv 1 \pmod{19}$$

- An alternative form of Fermat's theorem is also useful: If p is prime and a is a positive integer, then

$$a^p \equiv a \pmod{p} \qquad (8.3)$$

Euler's Theorem

Euler's theorem states that for every a and n that are relatively prime:

$$a^{\phi(n)} \equiv 1 \pmod{n} \qquad (8.4)$$

Proof:

- Equation (8.4) is true if n is prime, because in that case, f(n) = (n - 1) and Fermat's theorem holds.
- However, it also holds for any integer n. Recall that f(n) is the number of positive integers less than n that are relatively prime to n. Consider the set of such integers, labeled as

$$R = \{x_1, x_2, \ldots, x_{\phi(n)}\}$$

- That is, each element x_i of R is a unique positive integer less than n with gcd(x_i, n) = 1.Now multiply each element by a, modulo n:

$$S = \{(ax_1 \bmod n), (ax_2 \bmod n), \ldots, (ax_{\phi(n)} \bmod n)\}$$

The set *S* is a permutation of *R*, by the following line of reasoning:

1. Because *a* is relatively prime to *n* and x_i is relatively prime to *n*, ax_i must also be relatively prime to *n*. Thus, all the members of S are integers that are less than *n* and that are relatively primeto *n*.
2. There are no duplicates in *S*. Refer to Equation (4.5). If ax_i mod n = ax_j mod *n*, then $x_i = x_j$.

Therefore,

$$\prod_{i=1}^{\phi(n)} (ax_i \bmod n) = \prod_{i=1}^{\phi(n)} x_i$$

$$\prod_{i=1}^{\phi(n)} ax_i \equiv \prod_{i=1}^{\phi(n)} x_i \pmod n$$

$$a^{\phi(n)} \times \left[\prod_{i=1}^{\phi(n)} x_i\right] \equiv \prod_{i=1}^{\phi(n)} x_i \pmod n$$

$$a^{\phi(n)} \equiv 1 \pmod n$$

- which completes the proof.

5.5. CHINESE REMAINDER THEOREM

Let $m_1,\ldots,m_k$ be integers that are pairwise relatively prime integers. Define *M* to be the product of all the m_i's. Let $a_1, \ldots, a_k$ be integers. Then the set of congruences.

$x \equiv a1 (mod\ m1)$
$x \equiv a2 (mod\ m2)$
.
.
.
$x \equiv ak (mod\ mk)$
has a unique solution modulo M.

Proof:

- Put M= $m_1, m_2, \ldots\ldots m_r$ and for each k= 1,2,……r.

Let ,

$$M_k = \frac{M}{m_k}$$

- Then gcd (M_K, m_k) = 1 for all k.
- Let, y_k be an inverse of M_K modulo m_k for each k.
- Then by definition of inverse we have

$$M_K y_k = 1 (mod\ m_k)$$

- Let , $x = a_1 M_1 y_1 + a_2 M_2 y_2 + \ldots\ldots + a_k M_k y_k$.

- Then x is a simultaneous solution to all of the congruence.
- Since the modulo $m_1, m_2, \ldots\ldots m_r$ are pairwise relatively prime, any two simultaneous solution to the system must be congruent modulo M.
- Thus , the solution is a unique congruence class modulo M, and the value of x computed above is in that class.

5.6. EXPONENTIATION AND LOGARITHM

Contents
• **Introduction** • **The Powers of an Integer, Modulo *n*** • **Logarithms for Modular Arithmetic** • **Calculation of Discrete Logarithms**

Introduction

- Discrete logarithms are fundamental to a number of public-key algorithms, including Diffie-Hellman key exchange and the digital signature algorithm (DSA).

The Powers of an Integer, Modulo *n*

- Recall from Euler's theorem [Equation (8.4)] that, for every a and n that are relatively prime,

$$a^{\phi(n)} \equiv 1 \pmod{n}$$

- where $\varphi(n)$, Euler's totient function, is the number of positive integers less than n and relatively prime to *n*. Now consider the more general expression:

$$a^m \equiv 1 \pmod{n} \qquad \textbf{(8.10)}$$

- If a and n are relatively prime, then there is at least one integer *m* that satisfies Equation (8.10), namely, $M = \varphi(n)$. The least positive exponent *m* for which Equation (8.10) holds is referred to in several ways:

  • The order of *a (mod n)*
  • The exponent to which a belongs *(mod n)*
  • The length of the period generated by *a*

- Table 8.3 shows all the powers of *a*, modulo 19 for all positive a < 19. The length of the sequence for each base value is indicated by shading. Note the following:

  1. All sequences end in 1. This is consistent with the reasoning of the preceding few paragraphs.
  2. The length of a sequence divides $\varphi(19) = 18$. That is, an integral number of sequences occur in each row of the table.
  3. Some of the sequences are of length 18. In this case, it is said that the base integer a generates (via powers) the set of nonzero integers modulo 19. Each such integer is called a primitive root of the modulus 19.

Table 8.3 Powers of Integers, Modulo 19

a	a^2	a^3	a^4	a^5	a^6	a^7	a^8	a^9	a^{10}	a^{11}	a^{12}	a^{13}	a^{14}	a^{15}	a^{16}	a^{17}	a^{18}
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	4	8	16	13	7	14	9	18	17	15	11	3	6	12	5	10	1
3	9	8	5	15	7	2	6	18	16	10	11	14	4	12	17	13	1
4	16	7	9	17	11	6	5	1	4	16	7	9	17	11	6	5	1
5	6	11	17	9	7	16	4	1	5	6	11	17	9	7	16	4	1

Logarithms for Modular Arithmetic:

- With ordinary positive real numbers, the logarithm function is the inverse of exponentiation. An analogous function exists for modular arithmetic.
- The logarithm of a number is defined to be the power to which some positive base (except 1) must be raised in order to equal the number. That is, for base x and for a value y,

  $$y = x^{\log_x(y)}$$

  The properties of logarithms include

  $$\log_x(1) = 0$$
  $$\log_x(x) = 1$$

  $$\log_x(yz) = log_x(y) + \log_x(z) \qquad \textbf{(8.11)}$$
  $$\log_x(y^r) = r \times \log_x(y) \qquad \textbf{(8.12)}$$

- Consider a primitive root a for some prime number p (the argument can be developed for nonprimes as well). Then we know that the powers of a from 1 through (p - 1) produce each integer from 1 through (p - 1) exactly once. We also know that any integer b satisfies

  $$b \equiv r \pmod p \text{ for some } r, \text{ where } 0 \leq r \leq (p-1)$$

- By the definition of modular arithmetic. It follows that for any integer b and a primitive root a of prime number p, we can find a unique exponent i such that

  $$b \equiv a^i \pmod p \quad \text{where } 0 \leq i \leq (p-1)$$

- This exponent i is referred to as the discrete logarithm of the number b for the base a (mod p). We denote this value as dloga,p(b).

  Note the following:

  $$\text{dlog}_{a,p}(1) = 0 \text{ because } a^0 \bmod p = 1 \bmod p = 1 \qquad \textbf{(8.13)}$$
  $$\text{dlog}_{a,p}(a) = 1 \text{ because } a^1 \bmod p = a \qquad \textbf{(8.14)}$$

Calculation of Discrete Logarithms

Consider the equation

$$y = g^x \bmod p$$

- Given g, x, and p, it is a straightforward matter to calculate y. At the worst, we must perform x repeated multiplications, and algorithms exist for achieving greater efficiency.

(a) Discrete logarithms to the base 2, modulo 19

a	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$log_{2,19}(a)$	18	1	13	2	16	14	6	3	8	17	12	15	5	7	11	4	10	9

(b) Discrete logarithms to the base 3, modulo 19

a	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$log_{3,19}(a)$	18	7	1	14	4	8	6	3	2	11	12	15	17	13	5	10	16	9

ASYMMETRIC KEY CIPHERS

5.7. RSA CRYPTOSYSTEM

Contents

- introduction
- Description of the Algorithm
- Computational Aspects
- The Security of RSA

Introduction:

- It was developed by **Rivest, Shamir and Adleman**. This algorithm makes use of an expression with exponentials.
- Plaintext is encrypted in blocks, with each block having a binary value less than some number n.
- The RSA scheme is a cipher in which the plaintext and cipher text are integers between 0 and n - 1 for some n. A typical size for n is 1024 bits, or 309 decimal digits. That is, n is less than 21024.

Description of the Algorithm

- That is, the block size must be less than or equal to $\log_2(n)$; in practice, the block size is k-bits, where $2^k < n < 2^{k+1}$.Encryption and decryption are of the following form, for some Plaintext block M and Cipher text block C:

$$\mathbf{C = M^e \bmod n}$$

$$M = C^d \bmod n$$

- Both the sender and receiver know the value of n. the sender knows the value of e and only the receiver knows the value of d. thus, this is a public key encryption algorithm with a public key of KU = {e, n} and a private key of KR = {d, n}.
- Let us focus on the first requirement. We need to find the relationship of the form:

$$\mathbf{M^{ed} = M \bmod n}$$

- A corollary to Euler's theorem fits the bill: Given two prime numbers p and q Integers,

n and m, such that n=pq and 0<m<n, and arbitrary integer k, the following relationship holds

$$m^{k\Phi(n)+1} = m^{k(p-1)(q-1)+1} = m \bmod n$$

- where Φ(n) – Euler totient function, which is the number of positive integers less than n and relatively prime to n.we can achieve the desired relationship, if

$$d = e^{-1} \bmod \Phi(n)$$

- That is, e and d are multiplicative inverses mod Φ(n). According to the rule of modular arithmetic, this is true only if d (and therefore e) is relatively prime to Φ(n). Equivalently, gcd (Φ(n), d) = 1.

The steps involved in RSA algorithm for generating the key are

- Select two prime numbers, p = 17 and q = 11.
- Calculate n = p*q = 17*11 = 187
- Calculate Φ (n) = (p-1) (q-1) = 16*10 = 160.
- Select e such that e is relatively prime to Φ (n) = 160 and less than Φ(n);
- we choose e =7.
- Determine d such that ed ≡ 1 mod Φ (n) and d<160. The correct value is d = 23, because 23*7 = 161 = 1 mod 160.

RSA algorithm is summarized below.

Figure 9.5. The RSA Algorithm

Key Generation	
Select p, q	p and q both prime, $p \neq q$
Calculate $n = p \times q$	
Calculate $\phi(n) = (p-1)(q-1)$	
Select integer e	$\gcd(\phi(n), e) = 1; 1 < e < \phi(n)$
Calculate d	$d \equiv e^{-1} \pmod{\phi(n)}$
Public key	$PU = \{e, n\}$
Private key	$PR = \{d, n\}$

Encryption	
Plaintext:	$M < n$
Ciphertext:	$C = M^e \bmod n$

Decryption	
Ciphertext:	C
Plaintext:	$M = C^d \bmod n$

5.8. KEY DISTRIBUTION AND KEY MANAGEMENT

Distribution of Public Keys

Several techniques have been proposed for the distribution of public keys. **Virtually all these proposals can be grouped into the following general schemes:**

- Public announcement

- Publicly available directory
- Public-key authority
- Public-key certificates

Public Announcement of Public Keys

- On the face of it, the point of public-key encryption is that the public key is public. Thus, if there is some broadly accepted public-key algorithm, such as RSA, any participant can send his or her public key to any other participant or broadcast the key to the community at large (Figure 10.1).

- **For example**, because of the growing popularity of **PGP (pretty good privacy),** which makes use of RSA, many PGP users have adopted the practice of appending their public key to messages that they send to public forums, such as USENET newsgroups and Internet mailing lists.

Figure 10.1. Uncontrolled Public-Key Distribution

- Although this approach is convenient, it has a major weakness. Anyone can forge such a public announcement. That is, some user could pretend to be user A and send a public key to another participant **or broadcast such a public key.**

- Until such time as user A discovers the forgery and alerts other participants, the forger is able to read all encrypted messages intended for A and can use the forged keys for authentication.

Publicly Available Directory:

- A greater degree of security can be achieved by maintaining a publicly available dynamic directory of public keys.
- Maintenance and distribution of the public directory would have to be the responsibility of some trusted entity or organization (Figure 10.2). Such a scheme would include the following elements:

    1. The authority maintains a directory with a {name, public key} entry for each Participant.

    2. Each participant registers a public key with the directory authority. Registration would

have to be in person or by some form of secure authenticated communication.

3. A participant may replace the existing key with a new one at any time, either because of the desire to replace a public key that has already been used for a large amount of data, or because the corresponding private key has been compromised in some way.

4. Participants could also access the directory electronically. For this purpose, secure, authenticated communication from the authority to the participant is mandatory.

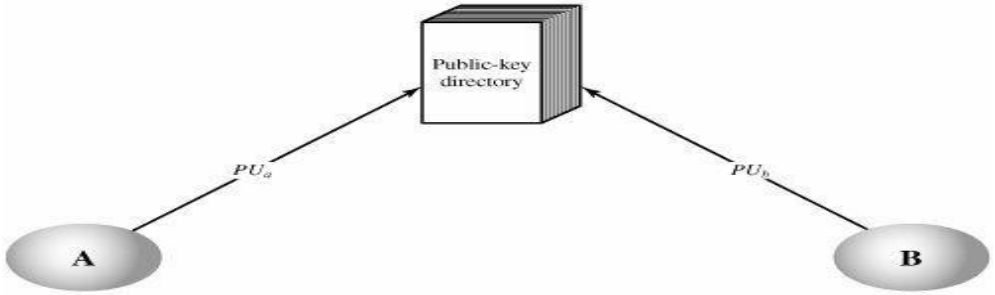


Figure 10.2. Public-Key Publication

- This scheme is clearly more secure than individual public announcements but still has vulnerabilities.
- If an adversary succeeds in obtaining or computing the private key of the directory authority, the adversary could authoritatively pass out counterfeit public keys and subsequently impersonate any participant and eavesdrop on messages sent to any participant.

- Another way to achieve the same end is for the adversary to tamper with the records kept by the authority.

Public-Key Authority:

- Stronger security for public-key distribution can be achieved by providing tighter control over the
  distribution of public keys from the directory.
- A typical scenario is illustrated in Figure 10.3, As before, the scenario assumes that a central authority maintains a dynamic directory of public keys of all participants.
- In addition, each participant reliably knows a public key for the authority, with only the authority knowing the corresponding private key.

The following steps:

1. A sends a timestamped message to the public-key authority containing a request for the current public key of B.

2. The authority responds with a message that is encrypted using the authority's private key, PR_{auth}

Thus, A is able to decrypt the message using the authority's public key. Therefore, A is assured that the message originated with the authority. The message includes the following:

- B's public key, $\boldsymbol{P_{Ub}}$ which A can use to encrypt messages destined for B
- The original request, to enable A to match this response with the corresponding earlier request and to verify that the original request was not altered before reception by the authority
- The original timestamp, so A can determine that this is not an old message from the authority containing a key other than B's current public key

3. A stores B's public key and also uses it to encrypt a message to B containing an identifier of A (*IDA*) and a nonce (*N*1), which is used to identify this transaction uniquely.

4. B retrieves A's public key from the authority in the same manner as A retrieved B's public key.

5. At this point, public keys have been securely delivered to A and B, and they may begin their protected exchange. However, two additional steps are desirable:

6. B sends a message to A encrypted with PU_a and containing A's nonce (*N*1) as well as a new nonce

generated by B (*N*2) Because only B could have decrypted message (3), the presence of *N*1 in message (6) assures A that the correspondent is B.

7. A returns *N*2, encrypted using B's public key, to assure B that its correspondent is A.

- ❖ Thus, a total of seven messages are required. However, the initial four messages need be used only infrequently because both A and B can save the other's public key for future use, a technique known as caching.

- ❖ Periodically, a user should request fresh copies of the public keys of its correspondents to ensure currency.

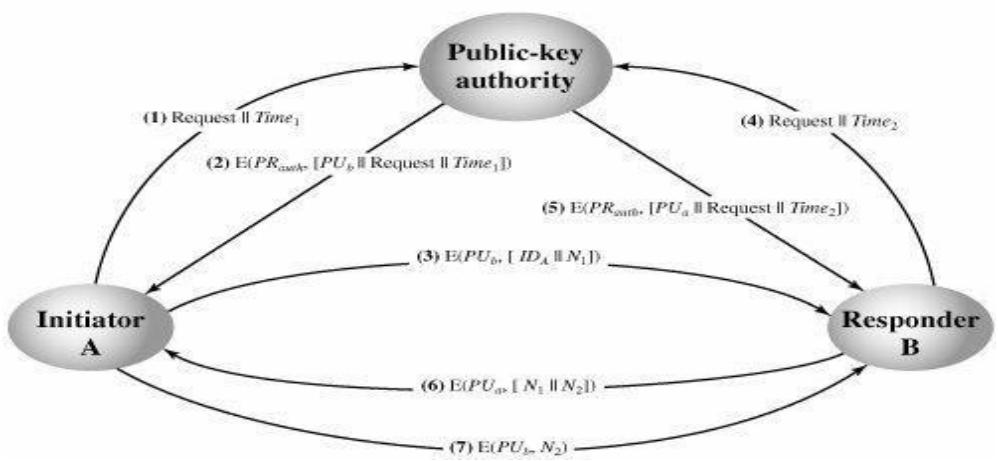


Figure 10.3. Public-Key Distribution Scenario

Public-Key Certificates

The scenario of Figure 10.3 is attractive, yet it has **some drawbacks**.

- The public-key authority could be somewhat of a bottleneck in the system, for a user must appeal to the authority for a public key for every other user that it wishes to contact.

- As before, the directory of names and public keys maintained by the authority is vulnerable to tampering.
- **An alternative approach,** first suggested by Kohnfelder [KOHN78], is to use **certificates** that can be used by participants to exchange keys without contacting a public-key authority, in a way that is as reliable as if the keys were obtained directly from a public-key authority.

- In essence, a certificate consists of a public key plus an identifier of the key owner, with the whole block signed by a trusted third party. Typically, the third party is a certificate authority, **such as a government agency or a financial institution,** that is trusted by the user community.

- A user can present his or her public key to the authority in a secure manner, and obtain a certificate. The user can then publish the certificate.

- Anyone needed this user's public key can obtain the certificate and verify that it is valid by way of the attached trusted signature. A participant can also convey its key information to another by transmitting its certificate. Other participants can verify that the certificate was created by the authority.

We can place the following requirements on this scheme:

1. Any participant can read a certificate to determine the name and public key of the certificate's

   Owner.

2. Any participant can verify that the certificate originated from the certificate authority and is not Counterfeit.
3. Only the certificate authority can create and update certificates.

Following additional requirement:

4. Any participant can verify the currency of the certificate.

- A certificate scheme is illustrated in Figure 10.4. Each participant applies to the certificate authority, supplying a public key and requesting a certificate.

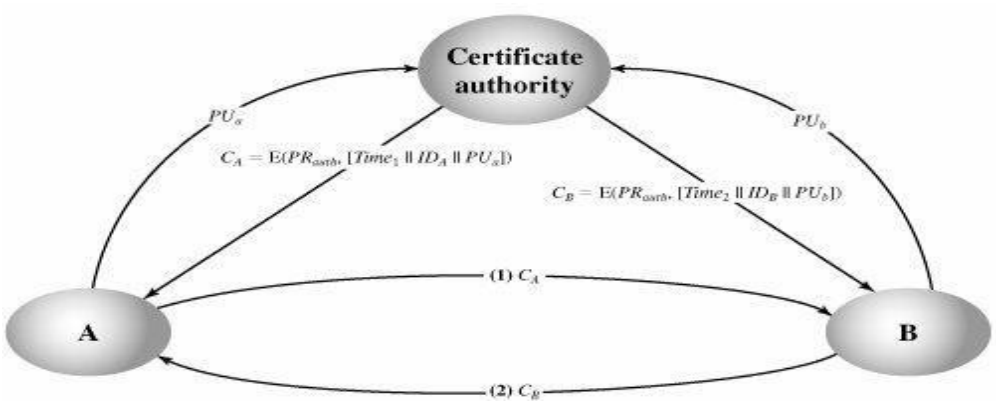


Figure 10.4. Exchange of Public-Key Certificates

- Application must be in person or by some form of secure authenticated communication. For participant A, the authority provides a certificate of the form

$$C_A = E(PR_{auth}, [T||ID_A||PU_a])$$

- Where PR_{auth} is the private key used by the authority and T is a timestamp. A may then pass this certificate on to any other participant, who reads and verifies the certificate as follows:

$$D(PU_{auth}, C_A) = D(PU_{auth}, E(PR_{auth}, [T||ID_A||PU_a])) = (T||ID_A||PU_a)$$

- The recipient uses the authority's public key, PU_{auth} to decrypt the certificate. Because the certificate is readable only using the authority's public key, this verifies that the certificate came from the certificate authority.
- The elements ID_A and PU_a provide the recipient with the name and public key of the certificate's holder. The timestamp T validates the currency of the certificate. The timestamp counters the following scenario. A's private key is learned by an adversary.
- A generates a new private/public key pair and applies to the certificate authority for a new certificate. Meanwhile, the adversary replays the old certificate to B.
- If B then encrypts messages using the compromised old public key, the adversary can read those messages. In this context, the compromise of a private key is comparable to the loss of a credit card.

- ❖ The owner cancels the credit card number but is at risk until all possible communicants are aware that the old credit card is obsolete. Thus, the timestamp serves as something like an expiration date. If a certificate is sufficiently old, it is assumed to be expired.

- ❖ One scheme has become universally accepted for formatting public-key certificates: the X.509 standard. X.509 certificates are used in most network security applications, including IP security, **secure sockets layer (SSL), secure electronic transactions (SET), and S/MIME,**

5.9. KEY MANAGEMENT

Contents
• Introduction • Distribution of Public Keys • Distribution of Secret Keys Using Public-Key Cryptography

Introduction:

One of the major roles of public-key encryption has been to address the problem of key distribution. There are actually two distinct aspects to the use of public-key cryptography in this regard:

- The distribution of public keys
- The use of public-key encryption to distribute secret keys.

Distribution of Secret Keys Using Public-Key Cryptography

- ❖ Once public keys have been distributed or have become accessible, secure communication that thwarts However, few users will wish to make exclusive use of public-key encryption for communication because of the relatively slow data rates that can be achieved.

- ❖ Accordingly, public-key encryption provides for the distribution of secret keys to be used for conventional encryption.

Simple Secret Key Distribution:

An extremely simple scheme was put forward by Merkle [MERK79], as illustrated in Figure 10.5. If A
wishes to communicate with B, the following procedure is employed:

1. A generates a public/private key pair $\{PU_a, PR_a\}$ and transmits a message to B consisting of PU_a
and an identifier of A, ID_A.
2. B generates a secret key, K_s, and transmits it to A, encrypted with A's public key.
3. A computes $D(PR_a, E(PU_a, K_s))$ to recover the secret key. Because only A can decrypt the

message, only A and B will know the identity of K_s.
4. A discards PU_a and PR_a and B discards PU_a.

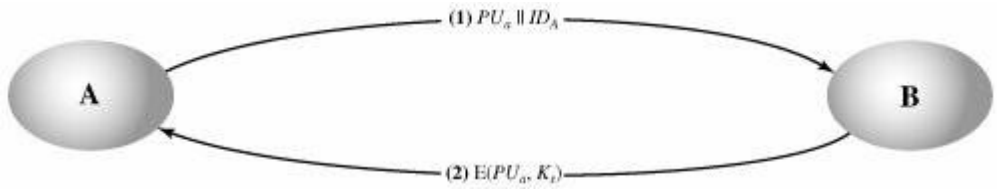


Figure 10.5. Simple Use of Public-Key Encryption to Establish a Session Key

- A and B can now securely communicate using conventional encryption and the session key *Ks*. At the completion of the exchange, both A and B discard *Ks*. Despite its simplicity, this is an attractive protocol.
- No keys exist before the start of the communication and none exist after the completion of communication. Thus, the risk of compromise of the keys is minimal. At the same time, the communication is secure from eavesdropping.
- The protocol depicted in Figure 10.5 is insecure against an adversary who can intercept messages and then either relay the intercepted message or substitute another message (see Figure 1.4c). Such an attack is known as a **man-in-the-middle attack .**

In this case, if an adversary, E, has control of the intervening communication channel, then E can compromise the communication in the following fashion without being detected:

1. A generates a public/private key pair $\{PU_a, PR_a\}$ and transmits a message intended for B
   consisting of PU_a and an identifier of A, ID_A.
2. E intercepts the message, creates its own public/private key pair $\{PU_e, PR_e\}$ and transmits $PUe \| ID_A$ to B.
3. B generates a secret key, *Ks*, and transmits $E(PU_e, K_s)$.
4. E intercepts the message, and learns *Ks* by computing $D(PR_e, E(PU_e, K_s))$.
5. E transmits $E(PU_a, K_s)$ to A.

- The result is that both A and B know K_s and are unaware that K_s has also been revealed to E. A and B can now exchange messages using K_s E no longer actively interferes with the communications channel but simply eavesdrops.
- Knowing K_s E can decrypt all messages, and both A and B are unaware of the problem. Thus, this simple protocol is only useful in an environment where the only threat is eavesdropping.

Secret Key Distribution with Confidentiality and Authentication:

Figure 10.6, based on an approach, provides protection against both active and passive attacks. We begin at a point when it is assumed that A and B have exchanged public keys by one of the schemes described earlier in this section.

Then the following steps occur:

1. A uses B's public key to encrypt a message to B containing an identifier of A (ID_A) and a
   Nonce (N_1), which is used to identify this transaction uniquely.
2. B sends a message to A encrypted with PU_a and containing A's nonce (N_1) as well as a new
   Nonce generated by B (N_2) Because only B could have decrypted message (1), the presence of N_1 in message (2) assures A that the correspondent is B.
3. A returns N_2 encrypted using B's public key, to assure B that its correspondent is A.
4. A selects a secret key K_s and sends $\boldsymbol{M = E(PU_b, E(PR_a, K_s))}$ to B. Encryption of this message
   With B's public key ensures that only B can read it; encryption with A's private key ensures
   That only A could have sent it.
5. B computes D(PU_a, D(PR_b, M)) to recover the secret key.

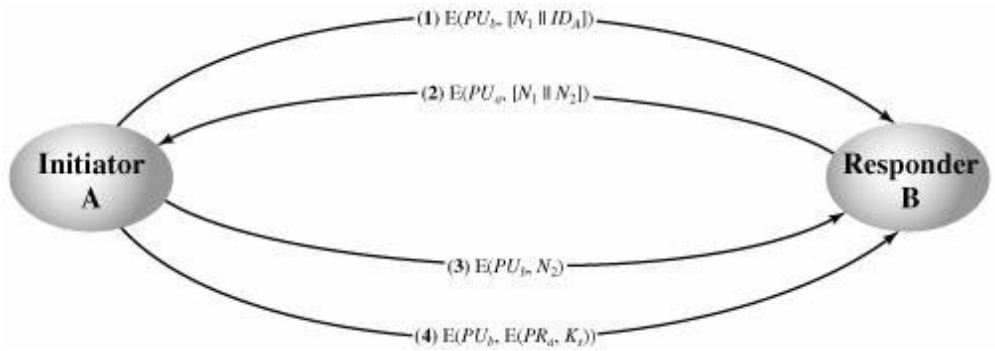


Figure 10.6. Public-Key Distribution of Secret Keys

Notice that the first three steps of this scheme are the same as the last three steps of Figure 10.3. The
result is that this scheme ensures both confidentiality and authentication in the exchange of a secret key.

A Hybrid Scheme

- Yet another way to use public-key encryption to distribute secret keys is a hybrid approach in use on IBM mainframes scheme retains the use of a **key distribution center (KDC**) that shares a secret master key with each user and distributes secret session keys encrypted with the master key.
- A public key scheme is used to distribute the master keys. The following rationale is provided for using this **three-level approach:**

● **Performance:** There are many applications, especially transaction-oriented applications, in which the session keys change frequently.Distribution of session keys by public-key encryption could degrade overall system performance because of the relatively high computational load of public-key encryption and decryption.

With a three-level hierarchy, public-key encryption is used only occasionally to update the master key between a user and the KDC.

● **Backward compatibility:** The hybrid scheme is easily overlaid on an existing KDC scheme, with
minimal disruption or software changes.

The addition of a public-key layer provides a secure, efficient means of distributing master keys. This is
an advantage in a configuration in which a single KDC serves a widely distributed set of users.

5.10. DIFFIE HELLMAN KEY EXCHANGE

Contents

- Introduction
- The Algorithm
- Key Exchange Protocols
- Man-in-the-Middle Attack

Introduction

- The Diffie-Hellman algorithm depends for its effectiveness on the difficulty of computing discrete logarithms.
- That is, if a is a primitive root of the prime number p, then the numbers a mod p, a2 mod p,c, ap-1 mod p are distinct and consist of the integers from 1 through p - 1 in some permutation. For any integer b and a primitive root a of prime number p, we can find a unique exponent i such that **b K ai (mod p)** where 0 … i … (p - 1).
- The exponent i is referred to as the discrete logarithm of b for the base a, mod p.

The Algorithm:

- Figure 10.1 summarizes the **Diffie-Hellman key exchange algorithm.** For this scheme, there are two publicly known numbers: a prime number q and an integer a that is a primitive root of q. Suppose the users A and B wish to create a shared key.

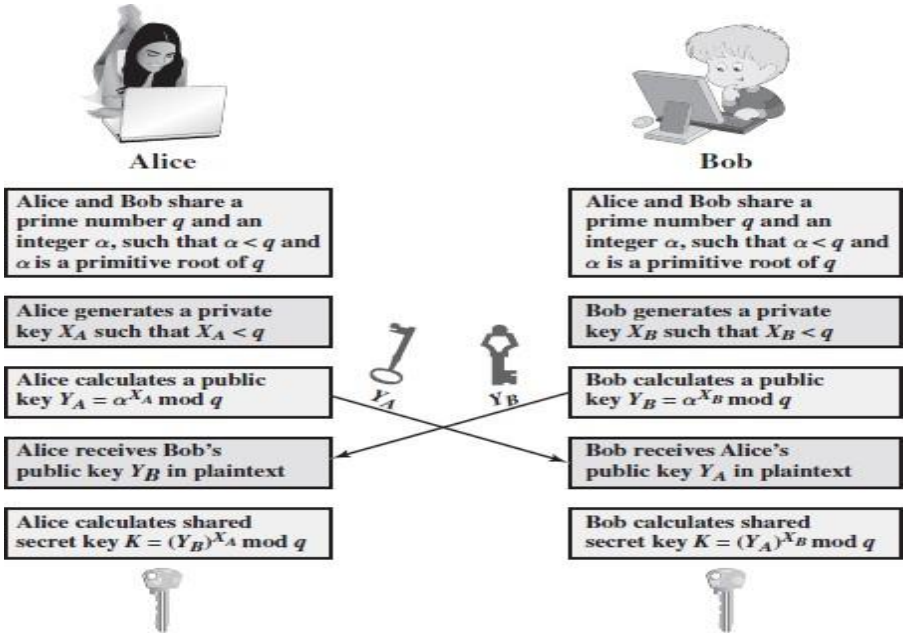


Figure 10.1 The Diffie-Hellman Key Exchange

- User A selects a random integer XA 6 q and computes **YA = aXA mod q.** Similarly, user B independently selects a random integer XB 6 q and computes YB = aXB mod q. Each side keeps the X value private and makes the Y value available publicly to the other side.

- Thus, XA is A's private key and YA is A's corresponding public key, and similarly for B. User A computes the key as K = (YB)XA mod q and user B computes the key as K = (YA)XB mod q. These two calculations produce identical results:

$$
\begin{aligned}
K &= (Y_B)^{X_A} \bmod q \\
&= (\alpha^{X_B} \bmod q)^{X_A} \bmod q \\
&= (\alpha^{X_B})^{X_A} \bmod q && \text{by the rules of modular arithmetic} \\
&= \alpha^{X_B X_A} \bmod q \\
&= (\alpha^{X_A})^{X_B} \bmod q \\
&= (\alpha^{X_A} \bmod q)^{X_B} \bmod q \\
&= (Y_A)^{X_B} \bmod q
\end{aligned}
$$

- The result is that the two sides have exchanged a secret value. Typically, this secret value is used as shared symmetric secret key. Now consider an adversary who can observe the key exchange and wishes to determine the secret key K.

- Because XA and XB are private, an adversary only has the following ingredients to work with: q, a, YA, and YB. Thus, the adversary is forced to take a discrete logarithm to determine the key.
- The adversary can then calculate the key K in the same manner as user B calculates it. That is, the adversary can calculate K as

$$K = (Y_A)^{X_B} \bmod q$$

- The security of the Diffie-Hellman key exchange lies in the fact that, while it is relatively easy to calculate exponentials modulo a prime, it is very difficult to calculate discrete logarithms. For large primes, the latter task is considered infeasible.

Here is an example. Key exchange is based on the use of the prime number $q = 353$ and a primitive root of 353, in this case $\alpha = 3$. A and B select private keys $X_A = 97$ and $X_B = 233$, respectively. Each computes its public key:

A computes $Y_A = 3^{97} \bmod 353 = 40$.
B computes $Y_B = 3^{233} \bmod 353 = 248$.

After they exchange public keys, each can compute the common secret key:

A computes $K = (Y_B)^{X_A} \bmod 353 = 248^{97} \bmod 353 = 160$.
B computes $K = (Y_A)^{X_B} \bmod 353 = 40^{233} \bmod 353 = 160$.

We assume an attacker would have available the following information:

$$q = 353; \alpha = 3; Y_A = 40; Y_B = 248$$

Key Exchange Protocols

Figure 10.1 shows a simple protocol that makes use of the Diffie-Hellman calculation. Suppose that user A wishes to set up a connection with user B and use a secret key to encrypt messages on that connection.

- User A can generate a one-time private key XA, calculate YA, and send that to user B. User B responds by generating a private value XB, calculating YB, and sending YB to user A. Both users can now calculate the key.
- The necessary public values q and α would need to be known ahead of time. Alternatively, user A could pick values for q and a and include those in the first message.

Man-in-the-Middle Attack:

- The protocol depicted in Figure 10.1 is insecure against a man-in-the-middle attack. Suppose Alice and Bob wish to exchange keys, and Darth is the adversary. The Attack proceeds as follows (Figure 10.2).

1. Darth prepares for the attack by generating two random private keys X_{D1} and X_{D2} and then computing the corresponding public keys Y_{D1} and Y_{D2}.
2. Alice transmits Y_A to Bob.
3. Darth intercepts Y_A and transmits Y_{D1} to Bob. Darth also calculates $K2 = (Y_A)^{X_{D2}} \bmod q$.
4. Bob receives Y_{D1} and calculates $K1 = (Y_{D1})^{X_B} \bmod q$.
5. Bob transmits Y_B to Alice.
6. Darth intercepts Y_B and transmits Y_{D2} to Alice. Darth calculates $K1 = (Y_B)^{X_{D1}} \bmod q$.
7. Alice receives Y_{D2} and calculates $K2 = (Y_{D2})^{X_A} \bmod q$.

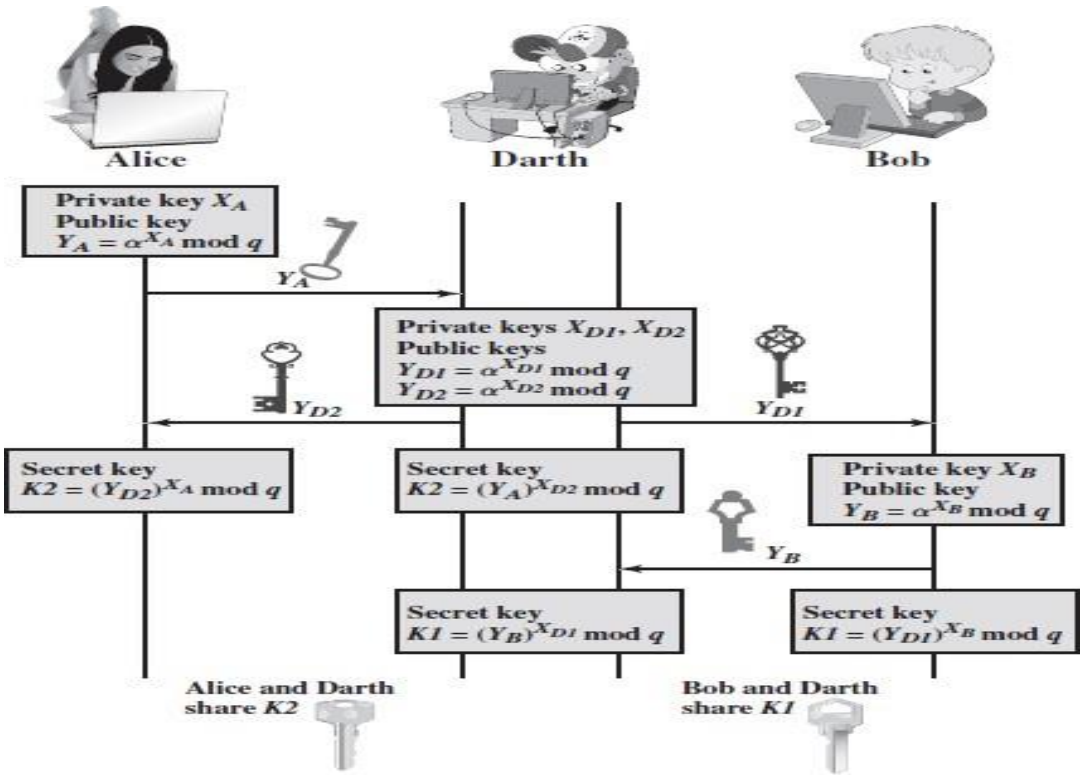


Figure 10.2 Man-in-the-Middle Attack

- At this point, Bob and Alice think that they share a secret key, but instead Bob and Darth share secret key K1 and Alice and Darth share secret key K2. All future communication between Bob and Alice is compromised in the following way.

  1. Alice sends an encrypted message M: E(K2, M).
  2. Darth intercepts the encrypted message and decrypts it to recover M.
  3. Darth sends Bob E(K1, M) or E(K1, M′), where M′ is any message. In the first case, Darth simply wants to eavesdrop on the communication without altering it. In the second case, Darth wants to modify the message going to Bob.

5.11. ELGAMAL CRYPTOSYSTEM

In 1984, T. Elgamal announced a public-key scheme based on discrete logarithms, closely related to the Diffie-Hellman technique. The ElGamal2 cryptosystem is used in some form in a number of standards including the digital signature standard (DSS), and the S/MIME e-mail standard.
As with Diffie-Hellman, the global elements of ElGamal are a prime number and , which is a primitive root of .

User A generates a private/public key pair as follows:

1. Generate a random integer X_A, such that $1 < X_A < q - 1$.
2. Compute $Y^A = \alpha^{X_A} \bmod q$.
3. A's private key is X_A; A's pubic key is $\{q, \alpha, Y_A\}$.

Any user B that has access to A's public key can encrypt a message as follows:

1. Represent the message as an integer M in the range $0 \leq M \leq q - 1$. Longer messages are sent as a sequence of blocks, with each block being an integer less than q.
2. Choose a random integer k such that $1 \leq k \leq q - 1$.
3. Compute a one-time key $K = (Y_A)^k \bmod q$.
4. Encrypt M as the pair of integers (C_1, C_2) where

$$C_1 = \alpha^k \bmod q; \; C_2 = KM \bmod q$$

User A recovers the plaintext as follows:

1. Recover the key by computing $K = (C_1)^{X_A} \bmod q$.
2. Compute $M = (C_2 K^{-1}) \bmod q$.

These steps are summarized in following figure.
It corresponds to following scenario :
Alice generates a public/private key pair; Bob encrypts using Alice's public key; and Alice decrypts using her private key.

Let us demonstrate why the ElGamal scheme works. First, we show how is recovered by the decryption process:

$K = (Y_A)^k \bmod q$	K is defined during the encryption process
$K = (\alpha^{X_A} \bmod q)^k \bmod q$	substitute using $Y_A = \alpha^{X_A} \bmod q$
$K = \alpha^{kX_A} \bmod q$	by the rules of modular arithmetic
$K = (C_1)^{X_A} \bmod q$	substitute using $C_1 = \alpha^k \bmod q$

Next, using K, we recover the plaintext as

$$C_2 = KM \bmod q$$

$$(C_2 K^{-1}) \bmod q = KMK^{-1} \bmod q = M \bmod q = M$$

We can restate the ElGamal process as follows, using Figure 10.3.

1. Bob generates a random integer k.
2. Bob generates a one-time key K using Alice's public-key components Y_A, q, and k.
3. Bob encrypts k using the public-key component α, yielding C_1. C_1 provides sufficient information for Alice to recover K.
4. Bob encrypts the plaintext message M using K.
5. Alice recovers K from C_1 using her private key.
6. Alice uses K^{-1} to recover the plaintext message from C_2.

The, K functions as a one-time key, used to encrypt and decrypt the message.*K C*1

Elgamal Cryptosystem

Global Public Elements	
q	prime number
α	$\alpha < q$ and α a primitive root of q

Key Generation by Alice	
Select private X_A	$X_A < q - 1$
Calculate Y_A	$Y_A = \alpha^{XA} \bmod q$
Public key	$PU = \{q, \alpha, Y_A\}$
Private key	X_A

Encryption by Bob with Alice's Public Key	
Plaintext:	$M < q$
Select random integer k	$k < q$
Calculate K	$K = (Y_A)^k \bmod q$
Calculate C_1	$C_1 = \alpha^k \bmod q$
Calculate C_2	$C_2 = KM \bmod q$
Ciphertext:	(C_1, C_2)

Decryption by Alice with Alice's Private Key	
Ciphertext:	(C_1, C_2)
Calculate K	$K = (C_1)^{XA} \bmod q$
Plaintext:	$M = (C_2 K^{-1}) \bmod q$

5.12. ELLIPTIC CURVE ARITHMETIC

Contents

- Elliptic Curves over Real Numbers
- Elliptic Curves over Zp
- Elliptic Curves over GF(2m)

- Most of the products and standards that use public-key cryptography for encryption and digital signatures use RSA. As we have seen, the key length for secure RSA use has increased over recent years, and this has put a heavier processing load on applications using RSA.
- This burden has ramifications, especially for electronic commerce sites that conduct large numbers of secure transactions. A competing system challenges RSA: **elliptic curve cryptography (ECC).** ECC is showing up in standardization efforts, including the IEEE P1363 Standard for Public-Key Cryptography.

- The principal attraction of ECC, compared to RSA, is that it appears to offer equal security for a far smaller key size, thereby reducing processing overhead.

- On the other hand, although the theory of ECC has been around for some time, it is only recently that products have begun to appear and that there has been sustained cryptanalytic interest in probing for weaknesses. Accordingly, the confidence level in ECC is not yet as high as that in RSA.
- ECC is fundamentally more difficult to explain than either RSA or Diffie- Hellman, and a full mathematical description is beyond the scope of this book. This section and the next give some background on elliptic curves and ECC.

- We begin with a brief review of the concept of abelian group. Next, we examine the concept of elliptic curves defined over the real numbers. This is followed by a look at elliptic curves defined over finite fields. Finally, we are able to examine elliptic curve ciphers.

- Abelian Groups an abelian group G, sometimes denoted by {G, .}, is a set of elements with a binary operation, denoted by . , that associates to each ordered pair (a, b) of elements in G an element (a . b) in G, such that the following axioms are obeyed:3

(A1) Closure:	If a and b belong to G, then $a \cdot b$ is also in G.
(A2) Associative:	$a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all a, b, c in G.
(A3) Identity element:	There is an element e in G such that $a \cdot \mathrm{e} = \mathrm{e} \cdot a = a$ for all a in G.
(A4) Inverse element:	For each a in G there is an element a' in G such that $a \cdot a' = a' \cdot a = \mathrm{e}$.
(A5) Commutative:	$a \cdot b = b \cdot a$ for all a, b in G.

- A number of public-key ciphers are based on the use of an abelian group.

- **For example, Diffie-Hellman key exchang**e involves multiplying pairs of nonzero integers modulo a prime number q. Keys are generated by exponentiation

the group, with exponentiation defined as repeated multiplication. For example, $a^k \bmod q = \underbrace{(a \times a \times \ldots \times a)}_{k \text{ times}} \bmod q$. To attack Diffie-Hellman, the attacker must determine k given a and a^k; this is the discrete logarithm problem.

For elliptic curve cryptography, an operation over elliptic curves, called addition, is used. Multiplication is defined by repeated addition. For example,

$$a \times k = \underbrace{(a + a + \ldots + a)}_{k \text{ times}}$$

where the addition is performed over an elliptic curve. Cryptanalysis involves determining k given a and $(a \times k)$.

- An elliptic curve is defined by an equation in two variables with coefficients. For cryptography, the variables and coefficients are restricted to elements in a finite field, which results in the definition of a finite abelian group.
- Before looking at this, we first look at elliptic curves in which the variables and coefficients are real numbers. This case is perhaps easier to visualize.

Elliptic Curves over Real Numbers:1

- Elliptic curves are not ellipses. They are so named because they are described by cubic equations, similar to those used for calculating the circumference of an ellipse. In general, cubic equations for elliptic curves take the following form, known as a **Weierstrass equation:**

$$y^2 + axy + by = x^3 + cx^2 + dx + e$$

- where a, b, c, d, e are real numbers and x and y take on values in the real numbers.4 For our purpose, it is sufficient to limit ourselves to equations of the form

$$y^2 = x^3 + ax + b$$

- Such equations are said to be cubic, or of degree 3, because the highest

- Exponent they contain is a 3. Also included in the definition of an elliptic curve is a single element denoted O and called the point at infinity or the zero point, which we discuss subsequently. To plot such a curve, we need to compute

$$y = \sqrt{x^3 + ax + b}$$

- For given values of a and b, the plot consists of positive and negative values of y for each value of x. Thus, each curve is symmetric about y = 0. Figure 10.4 shows two examples of elliptic curves. As you can see, the formula sometimes produces weirdlooking curves.

- Now, consider the set of points E(a, b) consisting of all of the points (x, y) that satisfy Equation (10.1) together with the element O. Using a different value of the pair (a, b) results in a different set E(a, b).
- Using this terminology, the two curves in Figure 10.4 depict the sets E(-1, 0) and E(1, 1), respectively.

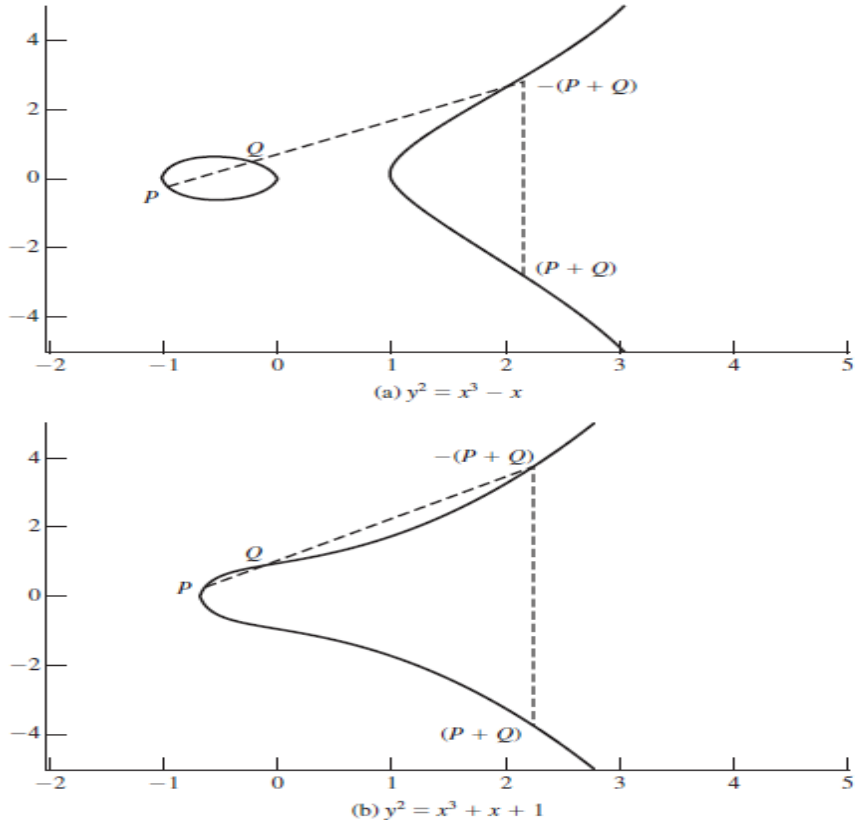


Figure 10.4 Example of Elliptic Curves

Elliptic Curves over Zp

- Elliptic curve cryptography makes use of elliptic curves in which the variables and coefficients are all restricted to elements of a finite field.
- Two families of elliptic curves are used in cryptographic applications: prime curves over Zp and binary
- curves over GF(2m). For a prime curve over Zp, we use a cubic equation in which the variables and coefficients all take on values in the set of integers from 0 through p - 1 and in which calculations are performed modulo p.

- For a **binary curve defined** over GF(2m), the variables and coefficients all take on values in GF(2m) and in calculations are performed over GF(2m).
- There is no obvious geometric interpretation of elliptic curve arithmetic over finite fields. The algebraic interpretation used for elliptic curve arithmetic over real numbers does readily carry over, and this is the approach we take.
- For elliptic curves over Zp, as with real numbers, we limit ourselves to equations of the form of Equation (10.1), but in this case with coefficients and variables limited to Zp:

$$y^2 \bmod p = (x^3 + ax + b) \bmod p$$

For example, Equation (10.5) is satisfied for $a = 1, b = 1, x = 9, y = 7, p = 23$:

$$7^2 \bmod 23 = (9^3 + 9 + 1) \bmod 23$$
$$49 \bmod 23 = 739 \bmod 23$$
$$3 = 3$$

- Now consider the set Ep(a, b) consisting of all pairs of integers (x, y) that satisfy Equation (10.5), together with a point at infinity O. The coefficients a and b and the variables x and y are all elements of Zp.

- **For example,** let p = 23 and consider the elliptic curve y2 = x3 + x + 1. In this case, a = b = 1. Note that this equation is the same as that of Figure 10.4b. The figure shows a continuous curve with all of the real points that satisfy the equation. For the set E23(1, 1), we are only interested in the nonnegative integers in the quadrant from (0, 0) through (p - 1, p - 1) that satisfy the equation mod p.

- Table 10.1 lists the points (other than O) that are part of E23(1, 1). Figure 10.5 plots the points of E23(1, 1); note that the points, with one exception, are symmetric about y = 11.5.

Table 10.1 Points (other than *O*) on the Elliptic Curve E_{23} (1,1)

(0, 1)	(6, 4)	(12, 19)
(0, 22)	(6, 19)	(13, 7)
(1, 7)	(7, 11)	(13, 16)
(1, 16)	(7, 12)	(17, 3)
(3, 10)	(9, 7)	(17, 20)
(3, 13)	(9, 16)	(18, 3)
(4, 0)	(11, 3)	(18, 20)
(5, 4)	(11, 20)	(19, 5)
(5, 19)	(12, 4)	(19, 18)

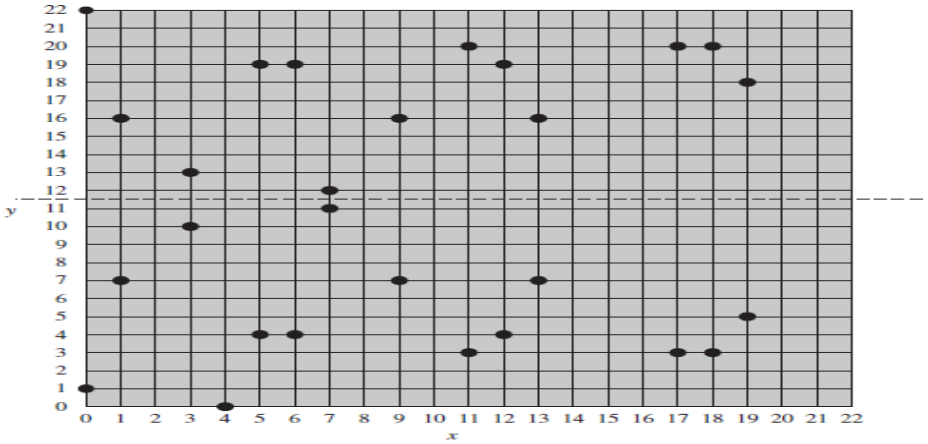


Figure 10.5 The Elliptic Curve $E_{23}(1, 1)$

- It can be shown that a finite abelian group can be defined based on the set Ep(a, b) provided that (x3 + ax + b) mod p has no repeated factors. This is equivalent to the condition

$$(4a^3 + 27b^2) \bmod p \neq 0 \bmod p$$

The rules for addition over $E_p(a, b)$, correspond to the algebraic technique described for elliptic curves defined over real numbers. For all points $P, Q \in E_p(a, b)$:

1. $P + O = P$.
2. If $P = (x_P, y_P)$, then $P + (x_P, -y_P) = O$. The point $(x_P, -y_P)$ is the negative of P, denoted as $-P$. For example, in $E_{23}(1, 1)$, for $P = (13, 7)$, we have $-P = (13, -7)$. But $-7 \bmod 23 = 16$. Therefore, $-P = (13, 16)$, which is also in $E_{23}(1, 1)$.
3. If $P = (x_p, y_p)$ and $Q = (x_Q, y_Q)$ with $P \neq -Q$, then $R = P + Q = (x_R, y_R)$ is determined by the following rules:

$$x_R = (\lambda^2 - x_P - x_Q) \bmod p$$
$$y_R = (\lambda(x_P - x_R) - y_P) \bmod p$$

where

$$\lambda = \begin{cases} \left(\frac{y_Q - y_P}{x_Q - x_P}\right) \bmod p & \text{if } P \neq Q \\ \left(\frac{3x_P^2 + a}{2y_P}\right) \bmod p & \text{if } P = Q \end{cases}$$

4. Multiplication is defined as repeated addition; for example, $4P$ $P + P + P + P$.

For example, let $P = (3, 10)$ and $Q = (9, 7)$ in $E_{23}(1, 1)$. Then

$$\lambda = \left(\frac{7 - 10}{9 - 3}\right) \bmod 23 = \left(\frac{-3}{6}\right) \bmod 23 = \left(\frac{-1}{2}\right) \bmod 23 = 11$$

$$x_R = (11^2 - 3 - 9) \bmod 23 = 109 \bmod 23 = 17$$

$$y_R = (11(3 - 17) - 10) \bmod 23 = -164 \bmod 23 = 20$$

So $P + Q = (17, 20)$. To find $2P$,

$$\lambda = \left(\frac{3(3^2) + 1}{2 \times 10}\right) \bmod 23 = \left(\frac{5}{20}\right) \bmod 23 = \left(\frac{1}{4}\right) \bmod 23 = 6$$

- The last step in the preceding equation involves taking the multiplicative inverse of 4 in Z23. This can be done using the extended Euclidean algorithm defined in Section 4.4. To confirm, note that (6 * 4) mod 23 = 24 mod 23 = 1.

$$x_R = (6^2 - 3 - 3) \bmod 23 = 30 \bmod 23 = 7$$
$$y_R = (6(3 - 7) - 10) \bmod 23 = (-34) \bmod 23 = 12$$

and $2P = (7, 12)$.

- For determining the security of various elliptic curve ciphers, it is of some interest to know the number of points in a finite abelian group defined over an elliptic curve. In the case of the finite group EP(a, b), the number of points N is bounded by
  - $p + 1 - 2\sqrt{p} \leq N \leq p + 1 + 2\sqrt{p}$
- Note that the number of points in Ep(a, b) is approximately equal to the number of elements in Zp, namely p elements.

Elliptic Curves over GF(2m)

- Recall from Chapter 4 that a finite field GF(2m) consists of 2m elements, together with addition and multiplication operations that can be defined over polynomials.

- For elliptic curves over GF(2m), we use a cubic equation in which the variables and coefficients all take on values in GF(2m) for some number m and in which calculations are performed using the rules of arithmetic in GF(2m).

- It turns out that the form of cubic equation appropriate for cryptographic applications for elliptic curves is somewhat different for GF(2m) than for Zp. The form is
  - $y^2 + xy = x^3 + ax^2 + b$

Table 10.2 Points (other than O) on the Elliptic Curve $E_{2^4}(g^4, 1)$

$(0, 1)$	(g^5, g^3)	(g^9, g^{13})
$(1, g^6)$	(g^5, g^{11})	(g^{10}, g)
$(1, g^{13})$	(g^6, g^8)	(g^{10}, g^8)
(g^3, g^8)	(g^6, g^{14})	$(g^{12}, 0)$
(g^3, g^{13})	(g^9, g^{10})	(g^{12}, g^{12})

- Where it is understood that the variables x and y and the coefficients a and b are elements of GF(2m) and that calculations are performed in GF(2m).

- Now consider the set E2m(a, b) consisting of all pairs of integers (x, y) that satisfy Equation (10.7), together with a point at infinity O.

- For example, let us use the finite field GF(24) with the irreducible polynomial f(x) = x4 + x + 1. This yields a generator g that satisfies f(g) = 0 with a value of g4 = g + 1, or in binary, g = 0010. We can develop the powers of g as follows.

$g^0 = 0001$	$g^4 = 0011$	$g^8 = 0101$	$g^{12} = 1111$
$g^1 = 0010$	$g^5 = 0110$	$g^9 = 1010$	$g^{13} = 1101$
$g^2 = 0100$	$g^6 = 1100$	$g^{10} = 0111$	$g^{14} = 1001$
$g^3 = 1000$	$g^7 = 1011$	$g^{11} = 1110$	$g^{15} = 0001$

For example, $g^5 = (g^4)(g) = (g + 1)(g) = g^2 + g = 0110$.

Now consider the elliptic curve $y^2 + xy = x^3 + g^4x^2 + 1$. In this case, $a = g^4$ and $b = g^0 = 1$. One point that satisfies this equation is (g^5, g^3):

$$(g^3)^2 + (g^5)(g^3) = (g^5)^3 + (g^4)(g^5)^2 + 1$$
$$g^6 + g^8 = g^{15} + g^{14} + 1$$
$$1100 + 0101 = 0001 + 1001 + 0001$$
$$1001 = 1001$$

Table 10.2 lists the points (other than O) that are part of $E_{2^4}(g^4, 1)$. Figure 10.6 plots the points of $E_{2^4}(g^4, 1)$.

It can be shown that a finite abelian group can be defined based on the set $E_{2^m}(a, b)$, provided that $b \neq 0$. The rules for addition can be stated as follows. For all points $P, Q \in E_{2^m}(a, b)$:

1. $P + O = P$.
2. If $P = (x_P, y_P)$, then $P + (x_P, x_P + y_P) = O$. The point $(x_P, x_P + y_P)$ is the negative of P, which is denoted as $-P$.
3. If $P = (x_P, y_P)$ and $Q = (x_Q, y_Q)$ with $P \neq -Q$ and $P \neq Q$, then $R = P + Q = (x_R, y_R)$ is determined by the following rules:

$$x_R = \lambda^2 + \lambda + x_P + x_Q + a$$
$$y_R = \lambda(x_P + x_R) + x_R + y_P$$

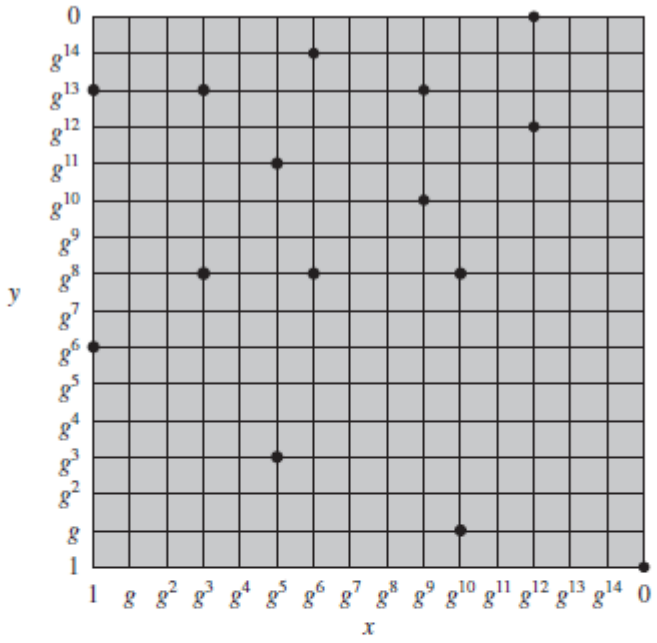


Figure 10.6 The Elliptic Curve $E_{2^4}(g^4, 1)$

where

$$\lambda = \frac{y_Q + y_P}{x_Q + x_P}$$

4. If $P = (x_P, y_P)$ then $R = 2P = (x_R, y_R)$ is determined by the following rules:

$$x_R = \lambda^2 + \lambda + a$$
$$y_R = x_P^2 + (\lambda + 1)x_R$$

where

$$\lambda = x_P + \frac{y_P}{x_P}$$

5.13. ELLIPTIC CURVE CRYPTOGRAPHY

Contents
• Analog of Diffie-Hellman Key Exchange • Elliptic Curve Encryption/Decryption • Security of Elliptic Curve Cryptography

- The addition operation in ECC is the counterpart of modular multiplication in RSA, and multiple addition is the counterpart of modular exponentiation.

- To form a cryptographic system using elliptic curves, we need to find a "hard problem" corresponding to factoring the product of two primes or taking the discrete logarithm.

- Consider the equation $Q = kP$ where $Q, P \in EP(a, b)$ and k 6 p. It is relatively easy to calculate Q given k and P, but it is hard to determine k given Q and P. This is called the discrete logarithm problem for elliptic curves.

- This is the group defined by the equation $y2 \bmod 23 = (x3 + 9x + 17) \bmod 23$. What is the discrete logarithm k of $Q = (4, 5)$ to the base $P = (16, 5)$? The brute-force method is to compute multiples of P until

Q is found. Thus,

$$P = (16, 5); 2P = (20, 20); 3P = (14, 14); 4P = (19, 20); 5P = (13, 10);$$
$$6P = (7, 3); 7P = (8, 7); 8P = (12, 17); 9P = (4, 5)$$

Because $9P = (4, 5) = Q$, the discrete logarithm $Q = (4, 5)$ to the base $P = (16, 5)$ is $k = 9$. In a real application, k would be so large as to make the brute-force approach infeasible.

In the remainder of this section, we show two approaches to ECC that give the flavor of this technique.

Analog of Diffie-Hellman Key Exchange

- Key exchange using elliptic curves can be done in the following manner. First pick a large integer q, which is either a prime number p or an integer of the form 2m, and elliptic curve parameters a and b for Equation (10.5) or Equation (10.7). This defines the elliptic group of points Eq(a, b).

- Next, pick a base point **G = (x1, y1) in Ep(a, b)** whose order is a very large value n. The order n of a point G on an elliptic curve is the smallest positive integer n such that nG = 0 and G are parameters of the cryptosystem known to all participants.

- A key exchange between users A and B can be accomplished as follows (Figure 10.7).

1. A selects an integer n_A less than n. This is A's private key. A then generates a public key $P_A = n_A \times G$; the public key is a point in $E_q(a, b)$.
2. B similarly selects a private key n_B and computes a public key P_B.
3. A generates the secret key $k = n_A \times P_B$. B generates the secret key $k = n_B \times P_A$.

The two calculations in step 3 produce the same result because

$$n_A \times P_B = n_A \times (n_B \times G) = n_B \times (n_A \times G) = n_B \times P_A$$

Elliptic Curve Encryption/Decryption

- Several approaches to encryption/decryption using elliptic curves have been analyzed in the literature. In this subsection, we look at perhaps the simplest. The first task in this system is to encode the plaintext message m to be sent as an (x, y) point Pm

- It is the point Pm that will be encrypted as a ciphertext and subsequently decrypted. Note that we cannot simply encode the message as the x or y coordinate of a point, because not all such coordinates are in Eq(a, b);

Global Public Elements

$E_q(a, b)$	elliptic curve with parameters a, b, and q, where q is a prime or an integer of the form 2^m
G	point on elliptic curve whose order is large value n

User A Key Generation

Select private n_A	$n_A < n$
Calculate public P_A	$P_A = n_A \times G$

User B Key Generation

Select private n_B	$n_B < n$
Calculate public P_B	$P_B = n_B \times G$

Calculation of Secret Key by User A

$K = n_A \times P_B$

Calculation of Secret Key by User B

$K = n_B \times P_A$

Figure 10.7 ECC Diffie-Hellman Key Exchange

Security of Elliptic Curve Cryptography

- The security of ECC depends on how difficult it is to determine k given kP and P. This is referred to as the elliptic curve logarithm problem. The fastest known technique for taking the elliptic curve logarithm is known as the Pollard rho method..

UNIT V
SECURITY PRACTICE AND SYSTEM SECURITY

Electronic Mail security – PGP, S/MIME – IP security – Web Security – SYSTEM SECURITY: Intruders – Malicious software – viruses – Firewalls.

5.1. ELECTRONIC MAIL SECURITY

- In virtually all distributed environments, electronic mail is the most heavily used network- based application. Users expect to be able to, and do, send e-mail to others who are connected directly or indirectly to the Internet, regardless of host operating system or communications suite.
- With the explosively growing reliance on e-mail, there grows a demand for authentication and confidentiality services. Two schemes stand out as approaches that enjoy widespread use: Pretty Good Privacy (PGP) and S/MIME. Both are examined in this chapter and Domain Keys Identified Mail.

Contents
• **Pretty Good Privacy** o Notation o Operational Description • **S/MIME** o RFC 5322 o Multipurpose Internet Mail Extensions o S/MIME Functionality o S/MIME Messages o S/MIME Certificate Processing o Enhanced Security Services

5.1.1. PGP

Contents
• **Pretty Good Privacy** o Notation o Operational Description

Pretty Good Privacy(PGP)

- PGP is a remarkable phenomenon. Largely the effort of a single person, Phil Zimmermann, PGP provides a confidentiality and authentication service that can be used for electronic mail and file storage applications.

- **In essence, Zimmermann has done the following**:

1. Selected the best available cryptographic algorithms as building blocks.
2. Integrated these algorithms into a general-purpose application that is independent of operating system and processor and that is based on a small set of easy-to-use commands.
3. Made the package and its documentation, including the source code, freely available via the Internet, bulletin boards, and commercial networks such as AOL (America On Line).
4.Entered into an agreement with a company (Viacrypt, now Network Associates) to provide a fully compatible, low-cost commercial version of PGP.

- **Characteristics of PGP or PGP has grown explosively and is now widely used. A number of reasons can be cited for this growth.**
    1. It is available free worldwide in versions that run on a variety of platforms, including Windows, UNIX, Macintosh, and many more.
    .
    2. It is based on algorithms that have survived extensive public review and are considered extremely secure. Specifically, the package includes RSA, DSS, and Diffie-Hellman for public-key encryption; CAST-128, IDEA, and 3DES for symmetric encryption; and SHA-1 for hash coding.
    3. It has a wide range of applicability
    4. It was not developed by, nor is it controlled by, any governmental or standards organization.
    .
    5. PGP is now on an Internet standards track (RFC 3156; *MIME Security with OpenPGP*).
    6. The algorithms used are extremely secure

Notation

- Most of the notation used in this chapter has been used before, but a few terms are new. It is perhaps best to summarize those at the beginning. The following symbols are used.
    Ks = session key used in symmetric encryption scheme
    PRa = private key of user A, used in public-key encryption scheme
    PUa = public key of user A, used in public-key encryption scheme
    EP = public-key encryption
    DP = public-key decryption
    EC = symmetric encryption
    DC = symmetric decryption
    H = hash function
    } = concatenation
    Z = compression using ZIP algorithm
    R64 = conversion to radix 64 ASCII format1

- The PGP documentation often uses the term ***secret key*** to refer to a key paired with a public key in a public-key encryption scheme.
- As was mentioned earlier, this practice risks confusion with a secret key used for symmetric encryption. Hence, we use the term ***private key*** instead.

Operational Description in PGP

- The actual operation of PGP, as opposed to the management of keys, consists of four services**:**
    - **Authentication,**
    - **Confidentiality,**
    - **Confidentiality and Authentication,**
    - **E-mail**
    - **Compatibility**
- (Table 19.1). We examine each of these in turn.

Authentication

- Figure 19.1a illustrates the digital signature service provided by PGP. This is the digital signature scheme discussed in Chapter 13 and illustrated in Figure 13.2. The sequence is as follows.
    1. The sender creates a message.
    2. SHA-1 is used to generate a 160-bit hash code of the message.
    3. The hash code is encrypted with RSA using the sender's private key, and the result is prepended to the message.
    4. The receiver uses RSA with the sender's public key to decrypt and recover the hash code.
    5. The receiver generates a new hash code for the message and compares it with the decrypted hash code. If the two match, the message is accepted as authentic

Table 19.1 Summary of PGP Services

Function	Algorithms Used	Description
Digital signature	DSS/SHA or RSA/SHA	A hash code of a message is created using SHA-1. This message digest is encrypted using DSS or RSA with the sender's private key and included with the message.
Message encryption	CAST or IDEA or Three-key Triple DES with Diffie-Hellman or RSA	A message is encrypted using CAST-128 or IDEA or 3DES with a one-time session key generated by the sender. The session key is encrypted using Diffie-Hellman or RSA with the recipient's public key and included with the message.
Compression	ZIP	A message may be compressed for storage or transmission using ZIP.
E-mail compatibility	Radix-64 conversion	To provide transparency for e-mail applications, an encrypted message may be converted to an ASCII string using radix-64 conversion.

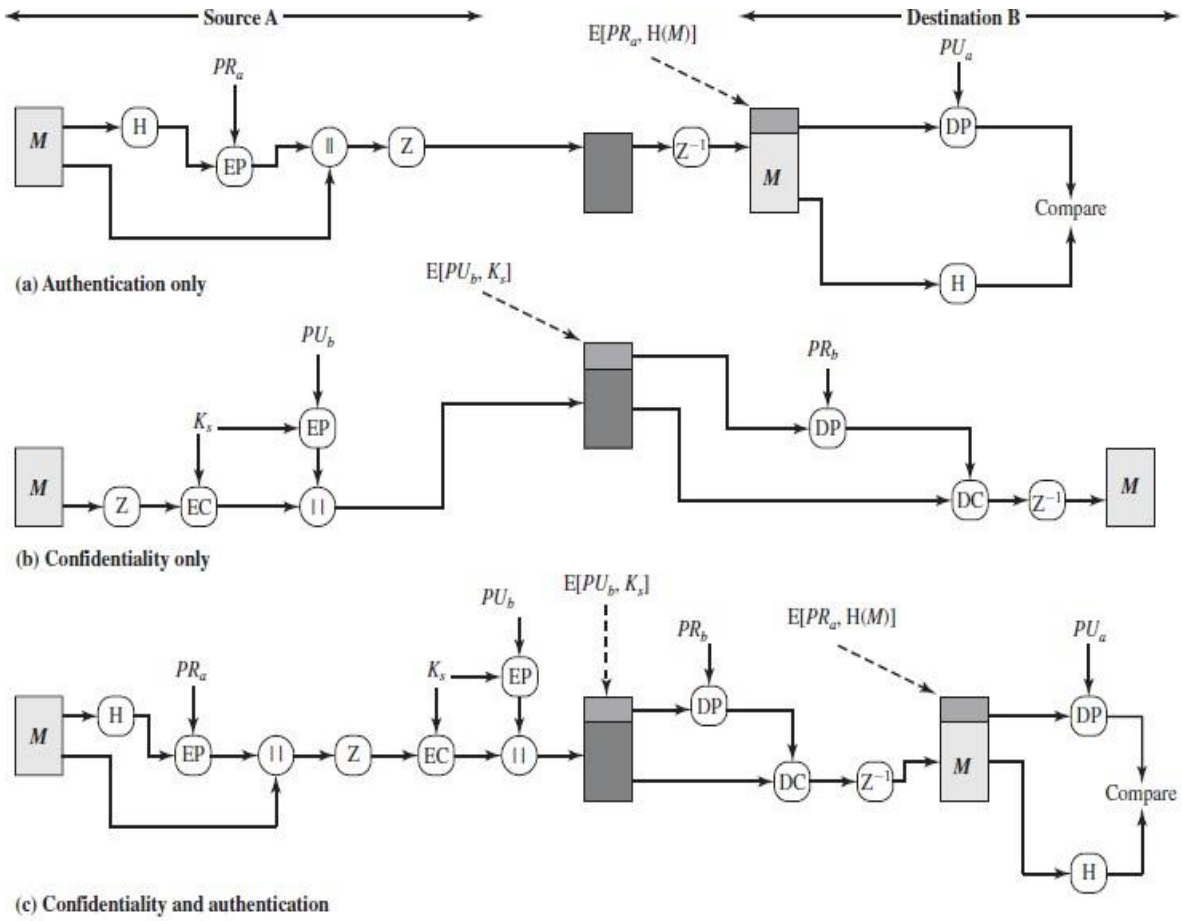


Figure 19.1 PGP Cryptographic Functions

- The combination of **SHA-1 and RSA** provides an effective digital signature scheme. Because of the strength of RSA, the recipient is assured that only the possessor of the matching private key can generate the signature.
- Because of the strength of SHA-1, the recipient is assured that no one else could generate a new message that matches the hash code and, hence, the signature of the original message. As an alternative, signatures can be generated using **DSS/SHA-1**.
- Although signatures normally are found attached to the message or file that they sign, this is not always the case: Detached signatures are supported.
- A detached signature may be stored and transmitted separately from the message it signs. This is useful in several contexts. A user may wish to maintain a separate signature log of all messages sent or received. A detached signature of an executable program can detect subsequent virus infection.
- Finally, detached signatures can be used when more than one party must sign a document, such as a legal contract. Each person's signature is independent and therefore is applied only to the document. Otherwise, signatures would have to be nested, with the second signer signing both the document and the first signature, and so on.

Confidentiality

- Another basic service provided by PGP is confidentiality, which is provided by encrypting messages to be transmitted or to be stored locally as files.
- In both cases, the symmetric encryption algorithm CAST-128 may be used.

- Alternatively, IDEA or 3DES may be used. The 64-bit cipher feedback (CFB) mode is used.
- As always, one must address the problem of key distribution. In PGP, each symmetric key is used only once.
- That is, a new key is generated as a random 128-bit number for each message. Thus, although this is referred to in the documentation as a session key, it is in reality a one-time key. Because it is to be used only once, the session key is bound to the message and transmitted with it.
- To protect the key, it is encrypted with the receiver's public key. Figure 19.1b illustrates the sequence,

Which can be described as follows?

1. The sender generates a message and a random 128-bit number to be used as a session key for this message only.
2. The message is encrypted using CAST-128 (or IDEA or 3DES) with the session key.
3. The session key is encrypted with RSA using the recipient's public key and is prepended to the message.
4. The receiver uses RSA with its private key to decrypt and recover the session key.
5. The session key is used to decrypt the message.

Confidentiality and Authentication

- As Figure 19.1c illustrates, both services may be used for the same message. First, a signature is generated for the plaintextmessage and prepended to the message.
- Then the plaintext message plus signatureis encrypted using CAST-128 (or IDEA or 3DES), and the session key is encrypted using RSA (or ElGamal).
- This sequence is preferable to the opposite: encrypting the message and then generating a signature for the encrypted message.
- It is generally more convenient to store a signature with a plaintext version of a message. Furthermore, for purposes of third-party verification, if the signature is performed first, a third party need not be concerned with the symmetric key when verifying the signature.

Compression

- As a default, PGP compresses the message after applying the signature but before encryption. This has the benefit of saving space both for e-mail transmission and for file storage.
- The placement of the compression algorithm, indicated by Z for compression and Z-1 for decompression in Figure 19.1, is critical.

  1. The signature is generated before compression for two reasons:

  a. It is preferable to sign an uncompressed message so that one can store only the uncompressed message together with the signature for future verification. If one signed a compressed document, then it would be necessary either to store a compressed version of the message for later verification or to recompress the message when verification is required.

  b. Even if one were willing to generate dynamically a recompressed message for verification, PGP's compression algorithm presents a difficulty. The algorithm is not deterministic; various implementations of the algorithm achieve different tradeoffs in

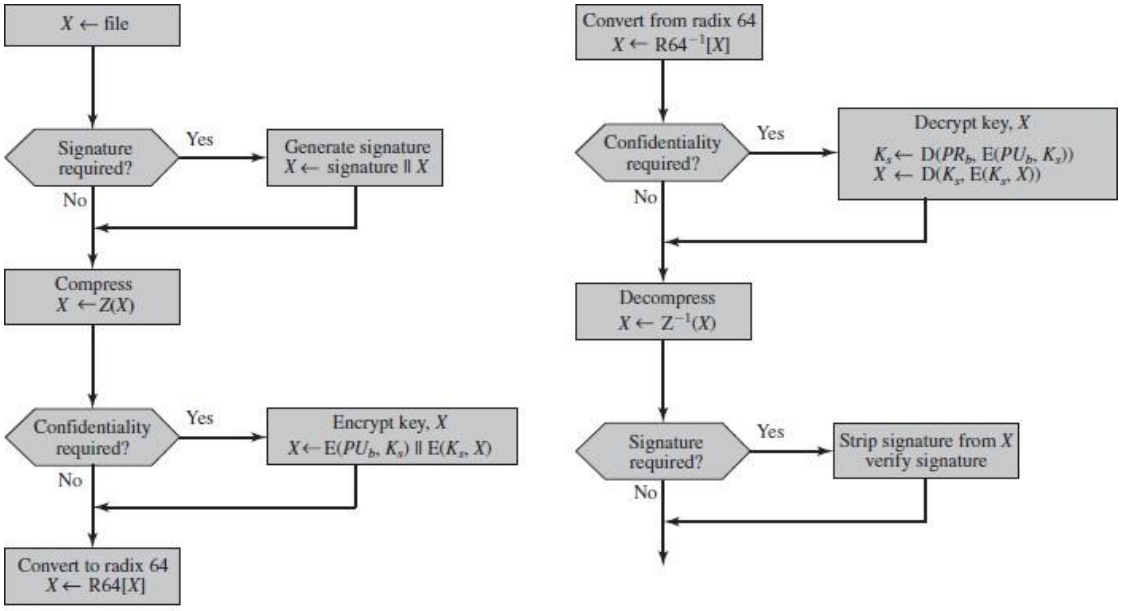
running speed versus compression ratio and, as a result, produce different compressed forms. However, these different compression algorithms are interoperable because any version of the algorithm can correctly decompress the output of any other version. Applying the hash function and signature after compression would constrain all PGP implementations to the same version of the compression algorithm.

2. Message encryption is applied after compression to strengthen cryptographic security. Because the compressed message has less redundancy than the original plaintext, cryptanalysis is more difficult.

.

E-mail Compatibility

- When PGP is used, at least part of the block to be transmitted is encrypted. If only the signature service is used, then the message digest is encrypted (with the sender's private key). If the confidentiality service is used, the message plus signature (if present) are encrypted (with a one-time symmetric key).
- Thus, part or the entire resulting block consists of a stream of arbitrary 8-bit octets. However, many electronic mail systems only permit the use of blocks consisting of ASCII text. To accommodate this restriction, PGP provides the service of converting the raw 8-bit binary stream to a stream of printable ASCII characters.
- The scheme used for this purpose is radix-64 conversion. Each group of three octets of binary data is mapped into four ASCII characters. This format also appends a CRC to detect transmission errors.
- The use of radix 64 expands a message by 33%. Fortunately, the session key and signature portions of the message are relatively compact, and the plaintext message has been compressed. In fact, the compression should be more than enough to compensate for the radix-64 expansion.



(a) Generic transmission diagram (from A)

(b) Generic reception diagram (to B)

Figure 19.2 Transmission and Reception of PGP Messages

5.1.2. S/MIME

Contents

S/MIME

- Secure/Multipurpose Internet Mail Extension (S/MIME) is a security enhancement to the MIME Internet e-mail format standard based on technology from RSA Data Security.
- Although both PGP and S/MIME are on an IETF standards track, it appears likely that S/MIME will emerge as the industry standard for commercial and organizational use, while PGP will remain the choice for personal e-mail security for many users. S/MIME is defined in a number of documents—most importantly RFCs 3370, 3850, 3851, and 3852.

RFC 5322

- RFC 5322 defines a format for text messages that are sent using electronic mail. It has been the standard for Internet-based text mail messages and remains in common use.
- In the RFC 5322 context, messages are viewed as having an envelope and contents. The envelope contains whatever information is needed to accomplish transmission and delivery. The contents compose the object to be delivered to the recipient.
- The RFC 5322 standard applies only to the contents. However, the content standard includes a set of header fields that may be used by the mail system to create the envelope, and the standard is intended to facilitate the acquisition of such information by programs.
- The overall structure of a message that conforms to RFC 5322 is very simple. A message consists of some number of header lines **(*the header*)** followed by unrestricted text **(*the body*).**
- The header is separated from the body by a blank line. Put differently, a message is ASCII text, and all lines up to the first blank line are assumed to be header lines used by the user agent part of the mail system.
- A header line usually consists of a keyword, followed by a colon, followed by the keyword's arguments; the format allows a long line to be broken up into several lines. The most frequently used keywords **are *From*, *To*, *Subject*, and *Date*.**
- **Here is an example message:**

```
Date: October 8, 2009 2:15:49 PM EDT
From: "William Stallings" <ws@shore.net>
Subject: The Syntax in RFC 5322
To: Smith@Other-host.com
Cc: Jones@Yet-Another-Host.com

Hello. This section begins the actual
message body, which is delimited from the
message heading by a blank line.
```

- Another field that is commonly found in RFC 5322 headers is *Message-ID*. This field contains a unique identifier associated with this message.

Multipurpose Internet Mail Extensions

- Multipurpose Internet Mail Extension (MIME) is an extension to the RFC 5322 framework that is intended to address some of the problems and limitations of the use of Simple Mail Transfer Protocol (SMTP), defined in RFC 821, or some other mail transfer protocol and RFC 5322 for electronic mail. [PARZ06]
- **Lists the following limitations of the SMTP/5322 scheme**.

  1. SMTP cannot transmit executable files or other binary objects. A number of schemes are in use for converting binary files into a text form that can be used by SMTP mail systems, including the popular UNIX UUencode/ UUdecode scheme. However, none of these is a standard or even a *de facto* standard.
  2. SMTP cannot transmit text data that includes national language characters, because these are represented by 8-bit codes with values of 128 decimal or higher, and SMTP is limited to 7-bit ASCII.
  3. SMTP servers may reject mail message over a certain size.
  4. SMTP gateways that translate between ASCII and the character code EBCDIC do not use a consistent set of mappings, resulting in translation problems.
  5. SMTP gateways to X.400 electronic mail networks cannot handle nontextual
  6. Some SMTP implementations do not adhere completely to the SMTP **standards defined in RFC 821. Common problems include:**
     - Deletion, addition, or reordering of carriage return and linefeed
     - Truncating or wrapping lines longer than 76 characters
     - Removal of trailing white space (tab and space characters)
     - Padding of lines in a message to the same length
     - Conversion of tab characters into multiple
- **Overview the MIME specification includes the following elements.**

  1. Five new message header fields are defined, which may be included in an RFC 5322 header. These fields provide information about the body of the message.
  2. A number of content formats are defined, thus standardizing representations that support multimedia electronic mail.
  3. Transfer encodings are defined that enable the conversion of any content format into a form that is protected from alteration by the mail system.

- **The five header fields defined in MIME are**

- **MIME-Version:** Must have the parameter value 1.0. This field indicates that the message conforms to RFCs 2045 and 2046.
- **Content-Type:** Describes the data contained in the body with sufficient detail that the receiving user agent can pick an appropriate agent or mechanism to represent the data to the user or otherwise deal with the data in an appropriate manner.
- **Content-Transfer-Encoding:** Indicates the type of transformation that has been used to represent the body of the message in a way that is acceptable for mail transport.
- **Content-ID:** Used to identify MIME entities uniquely in multiple contexts.
- **Content-Description:** A text description of the object with the body; this is useful when the object is not readable (e.g., audio data).

Mail Message Header

```
MIME-Version: 1.0
Content-type: multipart/mixed; boundary="simple boundary"
This is the preamble. It is to be ignored, though it is a
handy place for mail composers to include an explanatory
note to non-MIME conformant readers.
–simple boundary
This is implicitly typed plain ASCII text. It does NOT
end with a linebreak.
–simple boundary
Content-type: text/plain; charset=us-ascii
This is explicitly typed plain ASCII text. It DOES end
with a linebreak.
–simple boundary–
This is the epilogue. It is also to be ignored.
```

- **There are four subtypes of the multipart type,** all of which have the same overall syntax.
- The **multipart/mixed subtype** is used when there are multiple independent body parts that need to be bundled in a particular order.
- For the **multipart/ parallel subtype**, the order of the parts is not significant. If the recipient's system is appropriate, the multiple parts can be presented in parallel.
- For example, a picture

Table 19.2 MIME Content Types

Type	Subtype	Description
Text	Plain	Unformatted text; may be ASCII or ISO 8859.
	Enriched	Provides greater format flexibility.
Multipart	Mixed	The different parts are independent but are to be transmitted together. They should be presented to the receiver in the order that they appear in the mail message.
	Parallel	Differs from Mixed only in that no order is defined for delivering the parts to the receiver.
	Alternative	The different parts are alternative versions of the same information. They are ordered in increasing faithfulness to the original, and the recipient's mail system should display the "best" version to the user.
	Digest	Similar to Mixed, but the default type/subtype of each part is message/rfc822.
Message	rfc822	The body is itself an encapsulated message that conforms to RFC 822.
	Partial	Used to allow fragmentation of large mail items, in a way that is transparent to the recipient.
	External-body	Contains a pointer to an object that exists elsewhere.
Image	jpeg	The image is in JPEG format, JFIF encoding.
	gif	The image is in GIF format.
Video	mpeg	MPEG format.
Audio	Basic	Single-channel 8-bit ISDN mu-law encoding at a sample rate of 8 kHz.
Application	PostScript	Adobe Postscript format.
	octet-stream	General binary data consisting of 8-bit bytes.

- For the **multipart/alternative subtype**, the various parts are different representations of the same information. The following is an example:

```
From: Nathaniel Borenstein <nsb@bellcore.com>
To: Ned Freed <ned@innosoft.com>
Subject: Formatted text mail
MIME-Version: 1.0
Content-Type: multipart/alternative;
boundary = boundary42

   –boundary42
Content-Type: text/plain; charset = us-ascii
   ...plain text version of message goes here....
   –boundary42
Content-Type: text/enriched
   ...RFC 1896 text/enriched version of same message
goes here...
   –boundary42–
```

- The **multipart/digest subtype** is used when each of the body parts is interpreted as an RFC 5322 message with headers

MIME Transfer Encodings

.

- The MIME standard defines two methods of encoding data. The Content- Transfer-Encoding field can actually take on six values, as listed in Table 19.3. However, three of these values (7bit, 8bit, and binary) indicate that no encoding has been done but provide some information about the nature of the data.

Table 19.3 MIME Transfer Encodings

7bit	The data are all represented by short lines of ASCII characters.
8bit	The lines are short, but there may be non-ASCII characters (octets with the high-order bit set).
binary	Not only may non-ASCII characters be present, but the lines are not necessarily short enough for SMTP transport.
quoted-printable	Encodes the data in such a way that if the data being encoded are mostly ASCII text, the encoded form of the data remains largely recognizable by humans.
base64	Encodes data by mapping 6-bit blocks of input to 8-bit blocks of output, all of which are printable ASCII characters.
x-token	A named nonstandard encoding.

Mail Message Format

MIME-Version: 1.0
From: Nathaniel Borenstein <nsb@bellcore.com>
To: Ned Freed <ned@innosoft.com>
Subject: A multipart example
Content-Type: multipart/mixed;
Content-type: text/plain; charset=US-ASCII

Content-Type: multipart/parallel; boundary=unique-boundary-2
Content-Type: audio/basic
Content-Transfer-Encoding: base64
Content-type: text/enriched
Content-Type: message/rfc822
From: (mailbox in US-ASCII)
To: (address in US-ASCII)
Subject: (subject in US-ASCII)
Content-Type: Text/plain; charset=ISO-8859-1
Content-Transfer-Encoding: Quoted-printable

Figure 19.3 Example MIME Message Structure

S/MIME Functionality

- In terms of general functionality, S/MIME is very similar to PGP. Both offer the ability to sign and/or encrypt messages.

S/MIME provides the following functions.

- **Enveloped data:** This consists of encrypted content of any type and encryptedcontent encryption keys for one or more recipients.

- **Signed data:** A digital signature is formed by taking the message digest of the content to be signed and then encrypting that with the private key of the signer. The content plus signature are then encoded using base64 encoding. A signed data message can only be viewed by a recipient with S/MIME capability.
- **Clear-signed data:** As with signed data, a digital signature of the content is formed. However, in this case, only the digital signature is encoded using base64. As a result, recipients without S/MIME capability can view the message content, although they cannot verify the signature.
- **Signed and enveloped data:** Signed-only and encrypted-only entities may be nested, so that encrypted data may be signed and signed data or clear-signed data may be encrypted.

Cryptographic Algorithms

- Table 19.5 summarizes the cryptographic algorithms used in S/MIME. S/MIME uses the following terminology taken from RFC 2119 **(*Key Words for use in RFCs to Indicate Requirement Levels*)** to specify the requirement level:

  - **MUST:** The definition is an absolute requirement of the specification. An implementation must include this feature or function to be in conformance with the specification.
  - **SHOULD:** There may exist valid reasons in particular circumstances to ignore this feature or function, but it is recommended that an implementation include the feature or function.

Table 19.5 Cryptographic Algorithms Used in S/MIME

Function	Requirement
Create a message digest to be used in forming a digital signature.	MUST support SHA-1. Receiver SHOULD support MD5 for backward compatibility.
Encrypt message digest to form a digital signature.	Sending and receiving agents MUST support DSS. Sending agents SHOULD support RSA encryption. Receiving agents SHOULD support verification of RSA signatures with key sizes 512 bits to 1024 bits.
Encrypt session key for transmission with a message.	Sending and receiving agents SHOULD support Diffie-Hellman. Sending and receiving agents MUST support RSA encryption with key sizes 512 bits to 1024 bits.
Encrypt message for transmission with a one-time session key.	Sending and receiving agents MUST support encryption with tripleDES. Sending agents SHOULD support encryption with AES. Sending agents SHOULD support encryption with RC2/40.
Create a message authentication code.	Receiving agents MUST support HMAC with SHA-1. Sending agents SHOULD support HMAC with SHA-1.

S/MIME Messages

- The general procedures for S/MIME message preparation

1. Securing a MIME Entity

- S/MIME secures a MIME entity with a signature, encryption, or both.
- A MIME entity may be an entire message (except for the RFC 5322 headers), or if the MIME content type is multipart, then a MIME entity is one or more of the subparts of the message. The MIME entity is prepared according to the normal rules for MIME message preparation. Then the MIME entity plus some security-related data, such as algorithm identifiers and certificates, are processed by S/MIME to produce what is known as a PKCS object.
- A PKCS object is then treated as message content and wrapped in MIME (provided with appropriate MIME headers).
- The message to be sent is converted to canonical form. In particular, for a given type and subtype, the appropriate canonical form is used for the message content. For a multipart message, the appropriate canonical form is used for each subpart.

2. Enveloped Data

- The steps for preparing an envelopedData MIME entity are
  1. Generate a pseudorandom session key for a particular symmetric encryption algorithm (RC2/40 or triple DES).
  2. For each recipient, encrypt the session key with the recipient's public RSA key.
  3. For each recipient, prepare a block known as RecipientInfo that contains an identifier of the recipient's public-key certificate,2 an identifier of the algorithm used to encrypt the session key, and the encrypted session key.
  4. Encrypt the message content with the session key.

A sample message (excluding the RFC 5322 headers) is

Content-Type: application/pkcs7-mime; smime-type=envelopeddata;
name=smime.p7m
Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename=smime.p7m
rfvbnj756tbBghyHhHUujhJhjH77n8HHGT9HG4VQpfyF467GhIGfHfYT6
7n8HHGghyHhHUujhJh4VQpfyF467GhIGfHfYGTrfvbnjT6jH7756tbB9H
f8HHGTrfvhJhjH776tbB9HG4VQbnj7567GhIGfHfYT6ghyHhHUujpfyF4
0GhIGfHfQbnj756YT64V

- To recover the encrypted message, the recipient first strips off the base64 encoding. Then the recipient's private key is used to recover the session key. Finally, the message content is decrypted with the session key.

3. SignedData

- . The steps for preparing a signedData MIME entity are
  1. Select a message digest algorithm (SHA or MD5).
  2. Compute the message digest (hash function) of the content to be signed.
  3. Encrypt the message digest with the signer's private key.
  4. Prepare a block known as SignerInfo that contains the signer's public-key certificate, an identifier of the message digest algorithm, an identifier of the algorithm used to encrypt the message digest, and the encrypted message digest.

A sample message as follows

Content-Type: application/pkcs7-mime; smime-type=
signed-data; name=smime.p7m

Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename=smime.p7m
567GhIGfHfYT6ghyHhHUujpfyF4f8HHGTrfvhJhjH776tbB9HG4VQbnj7
77n8HHGT9HG4VQpfyF467GhIGfHfYT6rfvbnj756tbBghyHhHUujhJhjH
HUujhJh4VQpfyF467GhIGfHfYGTrfvbnjT6jH7756tbB9H7n8HHGghyHh
6YT64V0GhIGfHfQbnj75

- The recipient independently computes the message digest and compares it to the decrypted message digest to verify the signature.

4. Clear Signing

- Clear signing is achieved using the multipart content type with a signed subtype. As was mentioned, this signing process does not involve transforming the message to be signed, so that the message is sent "in the clear." Thus, recipients with MIME capability but not S/MIME capability are able to read the incoming message.

- A multipart/signed message has two parts.
    - The first part can be any **MIME type** but must be prepared so that it will not be altered during transfer from source to destination. This means that if the first part is not 7bit, then it needs to be encoded using base64 or quoted-printable.
    - This second part has **a MIME content type** of application and a subtype of pkcs7-signature. Here is a sample message:

Content-Type: multipart/signed;
protocol="application/pkcs7-signature";
micalg=sha1; boundary=boundary42
—boundary42
Content-Type: text/plain
This is a clear-signed message.
—boundary42
Content-Type: application/pkcs7-signature; name=smime.p7s
Content-Transfer-Encoding: base64
Content-Disposition: attachment; filename=smime.p7s
ghyHhHUujhJhjH77n8HHGTrfvbnj756tbB9HG4VQpfyF467GhIGfHf
YT6
4VQpfyF467GhIGfHfYT6jH77n8HHGghyHhHUujhJh756tbB9HGTrf
vbnj
n8HHGTrfvhJhjH776tbB9HG4VQbnj7567GhIGfHfYT6ghyHhHUujp
fyF4
7GhIGfHfYT64VQbnj756
—boundary42—

4. Registration Request

- **The certification request includes**

1. Certification Request Info block
2. Identifier of the public key encryption algorithm
3. Signature of the certification RequestInfo block

S/MIME Certificate Processing

- S/MIME uses public-key certificates that conform to version 3 of X.509 . The key-management scheme used by S/MIME is in some ways a hybrid between a strict X.509 certification hierarchy and PGP's web of trust.
- As with the PGP model, S/MIME managers and/or users must configure each client with a list of trusted keys and with certificate revocation lists.

1. User Agent Role

- An S/MIME user has several key-management functions to perform.
  - **Key generation:**. A user agent SHOULD generate RSA key pairs with a length in the range of 768 to 1024 bits and MUST NOT generate a length of less than 512 bits.
  - **Registration:** A user's public key must be registered with a certification authority in order to receive an X.509 public-key certificate.
  - **Certificate storage and retrieval:** A user requires access to a local list of certificates in order to verify incoming signatures and to encrypt outgoing messages.

2. VeriSign Certificates

- VeriSign provides a CA service that is intended to be compatible with S/MIME and a variety of other applications. VeriSign issues X.509 certificates with the product name VeriSign Digital ID.
- Each Digital ID contains the following :
  - Owner's public key
  - Owner's name or alias
  - Expiration date of the Digital ID
  - Serial number of the Digital ID
  - Name of the certification authority that issued the Digital ID
  - Digital signature of the certification authority that issued the Digital ID

- Digital IDs can also contain other user-supplied information, including
  - Address
  - E-mail address
  - Basic registration information (country, zip code, age, and gender)

Enhanced Security Services

The details of these may change, and additional services may be added**. The three services are**

- **Signed receipts:** A signed receipt may be requested in a SignedData object. Returning a signed receipt provides proof of delivery to the originator of a message and allows the originator to demonstrate to a third party that the recipient received the message
- **Security labels:** A security label may be included in the authenticated attributes of a SignedData object. A security label is a set of security information regarding the sensitivity of the content that is protected by S/MIME encapsulation. The labels may be used for access control, by indicating which users are permitted access to an object.
- **Secure mailing lists:** When a user sends a message to multiple recipients, a certain amount of per-recipient processing is required, including the use of each recipient's public key. The user can be relieved of this work by employing the services of an S/MIME Mail List Agent (MLA). An MLA can take a single incoming message, perform the recipient-specific encryption for each recipient, and forward the message..

5.2. IP SECURITY

Contents
• **IP Security Overview** o Applications of IPsec o Benefits of IPsec o Routing Applications • **IP Security Architecture** o IPsec Documents o IPsec Services o Security Associations (SA) • **Transport and Tunnel Modes**

IP Security Overview

- To provide security, the IAB included authentication and encryption as necessary security features in the next-generation IP, which has been issued as IPv6. Fortunately, these security capabilities were designed to be usable both with the current IPv4 and the future IPv6. This means that vendors can begin offering these features now, and many vendors now do have some IPsec capability in their products.
- The IPsec specification now exists as a set of Internet standards.

Applications of IPSec

- IPSec provides the capability to secure communications across a LAN, across private and public WANs, and across the Internet. Examples of its use include the following:
  - **Secure branch office connectivity over the Internet:** A company can build a secure virtual private network over the Internet or over a public WAN. This enables a business to rely heavily on the Internet and reduce its need for private networks, saving costs and network management overhead.
  - **Secure remote access over the Internet:** An end user whose system is equipped with IP security protocols can make a local call to an Internet service provider (ISP) and gain secure access to a company network. This reduces the cost of toll charges for traveling employees and telecommuters.
  - **Establishing extranet and intranet connectivity with partners:** IPSec can be used to secure communication with other organizations, ensuring authentication and confidentiality and providing a key exchange mechanism.

- **Enhancing electronic commerce security:** Even though some Web and electronic commerce applications have built-in security protocols, the use of IPSec enhances that security.

- The principal feature of IPSec that enables it to support these varied applications is that it can encrypt and/or authenticate *all* traffic at the IP level.
- Thus, all distributed applications, including remote logon, client/server, e-mail, file transfer, Web access, and so on, can be secured.
- Figure 20.1 is a typical scenario of IPsec usage. An organization maintains LANs at dispersed locations. Nonsecure IP traffic is conducted on each LAN. For traffic offsite, through some sort of private or public WAN, IPsec protocols are used.
- These protocols operate in networking devices, such as a router or firewall, that connect each LAN to the outside world. The IPsec networking device will typically encrypt and compress all traffic going into the WAN and decrypt and decompress traffic coming from the WAN; these operations are transparent to workstations and servers on the LAN.
- Secure transmission is also possible with individual users who dial into the WAN. Such user workstations must implement the IPsec protocols to provide security.

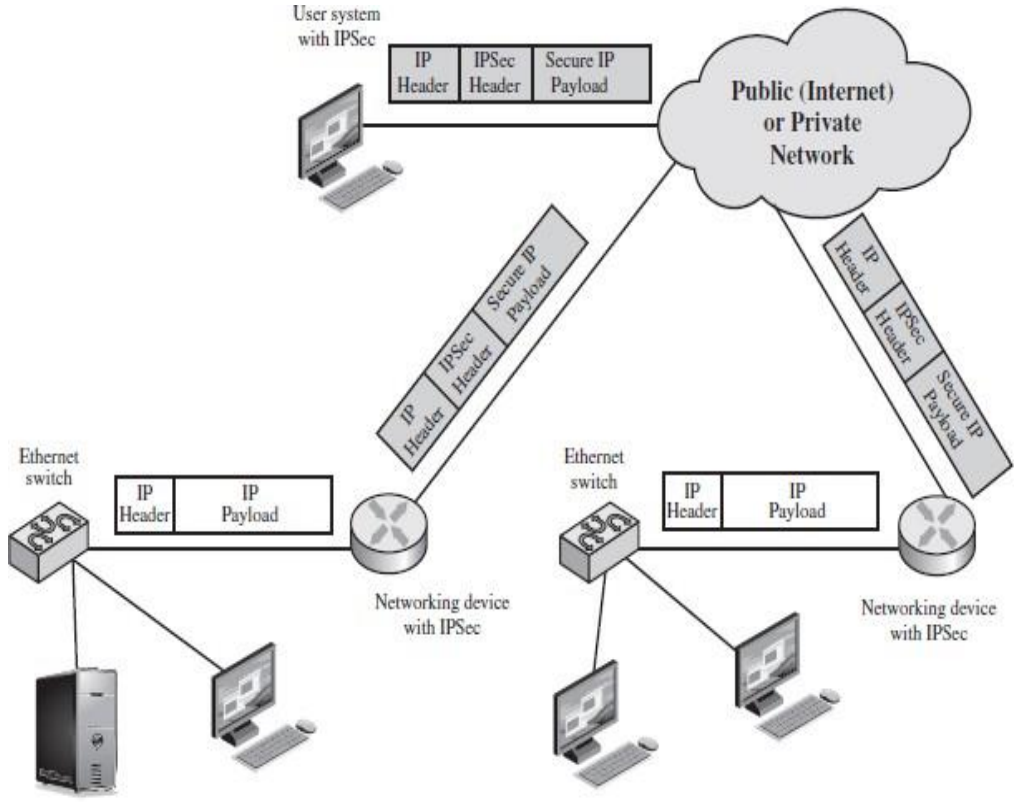


Figure 20.1 An IP Security Scenario

Benefits of IPSec

- [MARK97] lists the following benefits of IPSec: When IPSec is implemented in a firewall or router, it provides strong security that can be applied to all traffic crossing

the perimeter. Traffic within a company or workgroup does not incur the overhead of security-related processing.

- IPSec in a firewall is resistant to bypass if all traffic from the outside must use IP, and the firewall is the only means of entrance from the Internet into the organization.
- IPSec is below the transport layer (TCP, UDP) and so is transparent to applications.
- There is no need to change software on a user or server system when IPSec is implemented in the firewall or router.
- Even if IPSec is implemented in end systems, upper-layer software, including applications, is not affected.
- IPSec can be transparent to end users. There is no need to train users on security mechanisms, issue keying material on a per-user basis, or revoke keying material when users leave the organization.
- IPSec can provide security for individual users if needed. This is useful for offsite workers and for setting up a secure virtual sub network within an organization for sensitive applications.

Routing Applications

- A router advertisement (a new router advertises its presence) comes from an authorized router.
- neighbor advertisement (a router seeks to establish or maintain a neighbor
- relationship with a router in another routing domain) comes from an authorized router.
- A redirect message comes from the router to which the initial IP packet was sent.
- A routing update is not forged.

IP Security Architecture

- The IPSec specification has become quite complex. To get a feel for the overall architecture, we begin with a look at the **documents that define IPSec**. Then we discuss **IPSec services** and introduce the concept of **security association(SA).**

1. IPsec Documents

- The IPSec specification consists of numerous documents. The most important of these, issued in November of 1998, are RFCs 2401, 2402, 2406, and 2408:
  - **RFC 2401: An overview of a security architecture**
  - **RFC 2402: Description of a packet authentication extension to IPv4 and IPv6**
  - **RFC 2406: Description of a packet encryption extension to IPv4 and IPv6**
  - **RFC 2408: Specification of key management capabilities**
- Support for these features is mandatory for IPv6 and optional for IPv4. In both cases, the security features are implemented as extension headers that follow the main IP header.
- The extension header for authentication is known as the Authentication header; that for encryption is known as the Encapsulating Security Payload (ESP) header.

- IPsec encompasses three functional areas: authentication, confidentiality, and key anagement. The totality of the IPsec specification is scattered across dozens of RFCs and draft IETF documents, making this the most complex and difficult to grasp of all IETF specifications.

- The best way to grasp the scope of IPsec is to consult the latest version of the IPsec document roadmap, which as of this writing is RFC 6071 **[*IP Security (IPsec) and Internet Key Exchange (IKE) Document Roadmap*,**
- **The documents can be categorized into the following groups**.

    - **Architecture:** Covers the general concepts, security requirements, definitions, and mechanisms defining IPsec technology. The current specification is RFC 4301, ***Security Architecture for the Internet Protocol***.

    - **Authentication Header (AH):** AH is an extension header to provide message authentication. The current specification is RFC 4302, ***IP Authentication***

    - ***Heade****r*. Because message authentication is provided by ESP, the use of AH is deprecated. It is included in IPsecv3 for backward compatibility but should not be used in new applications.

    - **Encapsulating Security Payload (ESP):** ESP consists of an encapsulating header and trailer used to provide encryption or combined encryption/authentication. The current specification is RFC 4303, *IP **Encapsulating Security Payload (ESP)*.

    - **Internet Key Exchange (IKE):** This is a collection of documents describing the key management schemes for use with IPsec. The main specification is RFC 5996, ***Internet Key Exchange (IKEv2) Protocol***, but there are a number of related RFCs.

    - **Cryptographic algorithms:** This category encompasses a large set of documents that define and describe cryptographic algorithms for encryption, message authentication, pseudorandom functions (PRFs), and cryptographic key exchange.

    - **Other:** There are a variety of other IPsec-related RFCs, including those dealing with security policy and management information base (MIB) content.

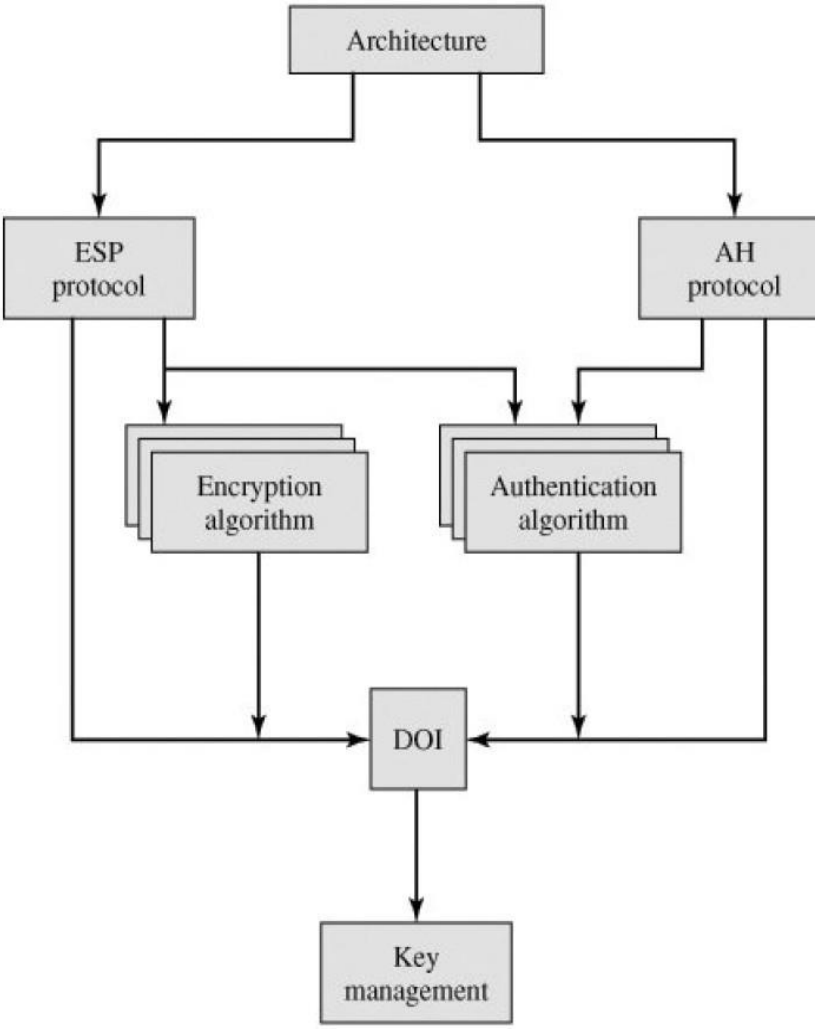


Figure 16.2. IPSec Document Overview

2. IPsec Services

- IPsec provides security services at the IP layer by enabling a system to select required ecurity protocols, determine the algorithm(s) to use for the service(s), and put in place any cryptographic keys required to provide the requested services.
- Two protocols are used to provide security: an authentication protocol designated by the header of the protocol, **Authentication Header (AH);** and a combined encryption/ authentication protocol designated by the format of the packet for that protocol, **Encapsulating Security Payload (ESP).**
- RFC 4301 lists the following services:
    - Access control
    - Connectionless integrity
    - Data origin authentication
    - Rejection of replayed packets (a form of partial sequence integrity)
    - Confidentiality (encryption)
    - Limited traffic flow confidentiality

	AH	ESP (encryption only)	ESP (encryption plus authentication)
Access control	✔	✔	✔
Connectionless integrity	✔		✔
Data origin authentication	✔		✔
Rejection of replayed packets	✔	✔	✔
Confidentiality		✔	✔
Limited traffic flow confidentiality		✔	✔

3. Security Associations (SA)

- A key concept that appears in both the authentication and confidentiality mechanisms for IP is the security association (SA). An association is a one-way relationship between a sender and a receiver that affords security services to the traffic carried on it.
- If a peer relationship is needed, for two-way secure exchange, then two security associations are required.
- A security association is uniquely identified by three parameters:

Security Parameters Index (SPI): A bit string assigned to this SA and having local significance only. The SPI is carried in AH and ESP headers to enable the receiving system to select the SA under which a received packet will be processed.

IP Destination Address: Currently, only unicast addresses are allowed; this is the address of the destination endpoint of the SA, which may be an end user system or a network system such as a firewall or router.

Security Protocol Identifier: This indicates whether the association is an AH or ESP security Association

SA Parameters

- In each IPSec implementation, there is a nominal Security Association Database that defines the parameters associated with each SA. A security association is normally defined by the following parameters:
  - [**Sequence Number Counter:** A 32-bit value used to generate the Sequence Number field in AH or ESP headers,
  - **Sequence Counter Overflow:** A flag indicating whether overflow of the Sequence Number Counter should generate an auditable event and prevent further transmission of packets on this SA (required for all implementations).
  - **ESP Information:** Encryption and authentication algorithm, keys, initialization values, key lifetimes, and related parameters being used with ESP (required for ESP implementations).
  - **Lifetime of This Security Association:** A time interval or byte count after which an SA must be replaced with a new SA (and new SPI) or terminated, plus an indication of which of these actions should occur (required for all implementations).
  - **IPSec Protocol Mode:** Tunnel, transport, or wildcard (required for all implementations). These modes are discussed later in this section.
  - **Path MTU:** Any observed path maximum transmission unit (maximum size of a packet that can
  - be transmitted without fragmentation) and aging variables (required for all implementations). **Anti-Replay Window:** Used to determine whether an inbound AH or ESP packet is a replay,
  - described in Section 16.3 (required for all implementations).

- **AH Information:** Authentication algorithm, keys, key lifetimes, and related parameters being used with AH (required for AH implementations).

SA Selectors

- IPSec provides the user with considerable flexibility in the way in which IPSec services are applied to IP traffic.
- The means by which IP traffic is related to specific SAs (or no SA in the case of traffic allowed to bypass IPSec) is the nominal Security Policy Database (SPD).
- In its simplest form, an SPD contains entries, each of which defines a subset of IP traffic and points to an SA for that traffic.
- In more complex environments, there may be multiple entries that potentially relate to a single SA or multiple SAs associated with a single SPD entry. The reader is referred to the relevant IPSec documents for a full discussion.
- Each SPD entry is defined by a set of IP and upper-layer protocol field values, called ***selectors*.**
- In effect, these selectors are used to filter outgoing traffic in order to map it into a particular SA. Outbound processing obeys the following general sequence for each IP packet:

Transport and Tunnel Modes

- Both **AH and ESP support** two modes of use: **transport and tunnel mode**. The operation of these two modes is best understood in the context of a description of ESP. Here we provide a brief overview.

1. Transport Mode

- Transport mode provides protection primarily for upper-layer protocols. That is, transport mode protection extends to the payload of an IP packet.1 Examples include a TCP or UDP segment or an ICMP packet,
- The transport mode is used for end - to- end communication between two hosts (e.g., a client and a server, or two workstations).
- When a host runs AH or ESP over IPv4, the payload is the data that normally follow the IP header. For IPv6, the payload is the data that normally follow both the IP header and any IPv6 extensions headers that are present, with the possible exception of the destination options header, which may be included in the protection.
- ESP in transport mode encrypts and optionally authenticates the IP payload but not the IP header. AH in transport mode authenticates the IP payload and selected portions of the IP header.

2. Tunnel Mode

- Tunnel mode provides protection to the entire IP packet. To achieve this, after the AH or ESP fields are added to the IP packet, the entire packet plus security fields is treated as the payload of new outer IP packet with a new outer IP header.
- The entire original, inner, packet travels through a tunnel from one point of an IP network to another; no routers along the way are able to examine the inner IP header. Because the original packet is encapsulated, the new, larger packet may have totally different source and destination addresses, adding to the security.

- Tunnel mode is used when one or both ends of a security association (SA) are a security gateway, such as a firewall or router that implements IPsec.
- In tunnel mode, a number of hosts on networks behind firewalls may engage in secure communications without implementing IPsec. The unprotected packets generated by such hosts are tunneled through external networks by tunnel mode SAs set up by the IPsec software in the firewall or secure router at the boundary of the local network.
- ESP in tunnel mode encrypts and optionally authenticates the entire inner IP packet, including the inner IP header.
- AH in tunnel mode authenticates the entire inner IP packet and selected portions of the outer IP header.

Table 20.1 Tunnel Mode and Transport Mode Functionality

	Transport Mode SA	Tunnel Mode SA
AH	Authenticates IP payload and selected portions of IP header and IPv6 extension headers.	Authenticates entire inner IP packet (inner header plus IP payload) plus selected portions of outer IP header and outer IPv6 extension headers.
ESP	Encrypts IP payload and any IPv6 extension headers following the ESP header.	Encrypts entire inner IP packet.
ESP with Authentication	Encrypts IP payload and any IPv6 extension headers following the ESP header. Authenticates IP payload but not IP header.	Encrypts entire inner IP packet. Authenticates inner IP packet.

-

Authentication header (AH)

Contents
○ **Authentication header (AH)** ○ **Anti-Replay Service** ○ **Integrity Check Value** ○ **Transport and Tunnel Modes**

Authentication header (AH)

- .The Authentication Header provides support for data integrity and authentication of IP packets. The data integrity feature ensures that undetected modification to a packet's content in transit is not possible.

- The authentication feature enables an end system or network device to authenticate the user or application and filter traffic accordingly; it also prevents the address spoofing attacks
- Authentication is based on the use of a message authentication code (MAC), protocol hence the two parties must share a secret key.
- The Authentication Header consists of the following fields (Figure 16.3):
  - **Next Header (8 bits):** Identifies the type of header immediately following this header.
  - **Payload Length (8 bits):** Length of Authentication Header in 32-bit words, minus 2. For

- example, the default length of the authentication data field is 96 bits, or three 32-bit words. With a three-word fixed header, there are a total of six words in the header, and the Payload Length field has a value of 4.
- **Reserved (16 bits):** For future use.
- **Security Parameters Index (32 bits):** Identifies a security association.
- **Sequence Number (32 bits):** A monotonically increasing counter value,
- **Authentication Data (variable):** A variable-length field (must be an integral number of 32-bit words) that contains the Integrity Check Value (ICV), or MAC, for this packet,

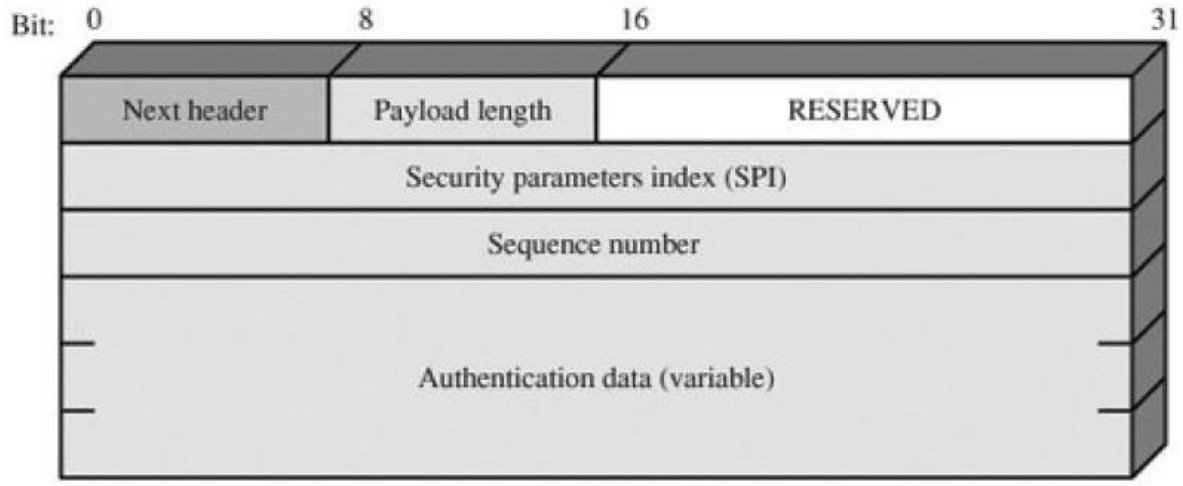


Figure - IPSec Authentication Header

Anti-Replay Service

- A replay attack is one in which an attacker obtains a copy of an authenticated packet and later transmits

- It to the intended destination. The receipt of duplicate, authenticated IP packets may disrupt service in some way or may have some other undesired consequence. The Sequence Number field is designed to thwart such attacks. First, we discuss sequence number generation by the sender, and then we look at how it is processed by the recipient.
- When a new SA is established, the **sender** initializes a sequence number counter to 0. Each time that a packet is sent on this SA, the sender increments the counter and places the value in the Sequence Number field. Thus, the first value to be used is 1. If anti-replay is enabled (the default), the sender must not allow the sequence number to cycle past 232 1 back to zero. Otherwise, there would be multiple valid packets with the same sequence number. If the limit of 232 1 is reached, the sender should terminate this SA and negotiate a new SA with a new key. Because IP is a connectionless, unreliable service, the protocol does not guarantee that packets will be delivered in order and does not guarantee that all packets will be delivered.
- Therefore, the IPSec authentication document dictates that the **receiver** should implement a window of size *W*, with a default of $W = 64$. The right edge of the window represents the highest sequence number, *N*, so far received for a valid packet. For any packet with a sequence number in the range from N $W + 1$ to N that has been correctly

received (i.e., properly authenticated), the corresponding slot in the window is marked (Figure 16.4). Inbound processing proceeds as follows when a packet is received:

1.If the received packet falls within the window and is new, the MAC is checked. If the packet is authenticated, the corresponding slot in the window is marked.

2. If the received packet is to the right of the window and is new, the MAC is checked. If the packet is authenticated, the window is advanced so that this sequence number is the right edge of the window, and the corresponding slot in the window is marked.

3. If the received packet is to the left of the window, or if authentication fails, the packet is discarded; this is an auditable event.

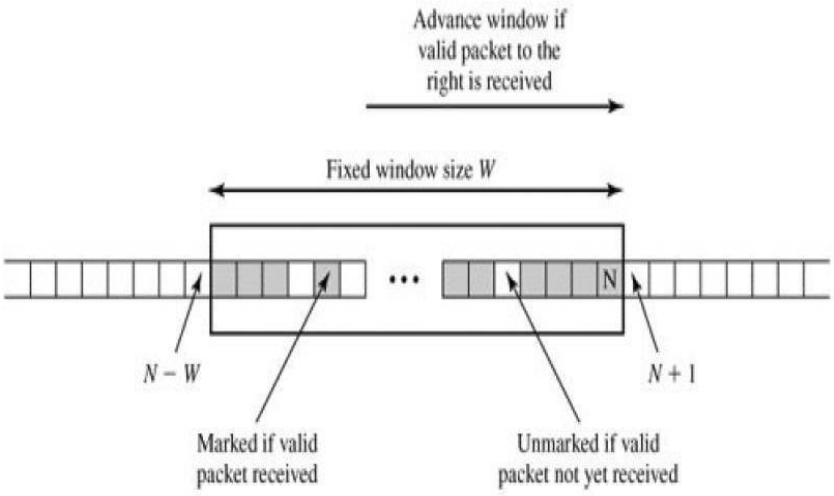


Figure 16.4. Antireplay Mechanism

Integrity Check Value

- The Authentication Data field holds a value referred to as the Integrity Check Value. The ICV is a message authentication code or a truncated version of a code produced by a MAC algorithm. The current specification dictates that a compliant implementation must support
    - HMAC-MD5-96
    - HMAC-SHA-1-96
- Both of these use the HMAC algorithm, the first with the MD5 hash code and the second with the SHA-1 hash code
- In both cases, the full HMAC value is calculated but then truncated by using the first 96 bits, which is the default length for the Authentication Data field.

Transport and Tunnel Modes

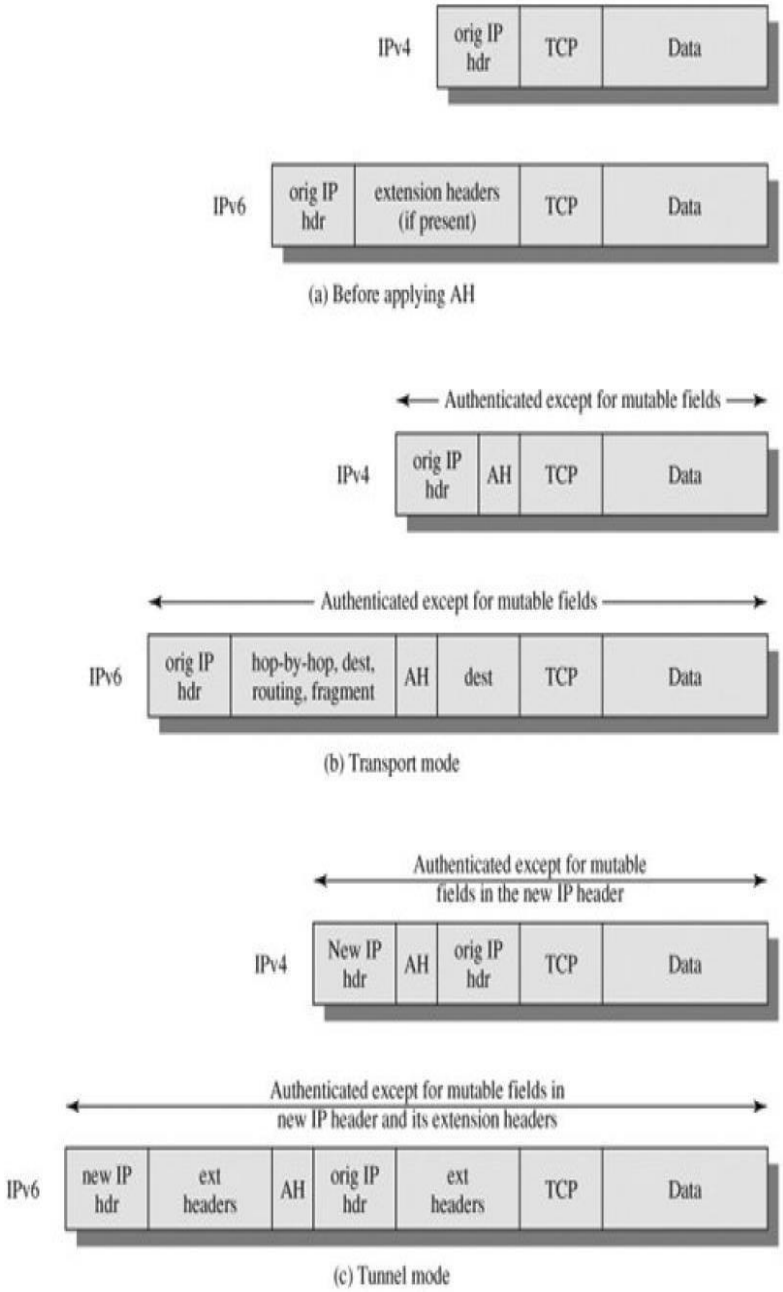
- Figure 16.5 shows two ways in which the IPSec authentication service can be used. In one case, authentication is provided directly between a server and client workstations; the workstation can be either on the same network as the server or on an external network.

- As long as the workstation and the server share a protected secret key, the authentication process is secure. This case uses a transport mode SA.
- In the other case, a remote workstation authenticates itself to the corporate firewall, either for access to the entire internal network or because the requested server does not support the authentication feature. This case uses a tunnel mode SA.
- In this subsection, we look at the scope of authentication provided by AH and the authentication header location for the two modes. The considerations are somewhat different for IPv4 and IPv6. Figure 16.6a shows typical IPv4 and IPv6 packets.
- In this case, the IP payload is a TCP segment; it could also be a data unit for any other protocol that uses IP, such as UDP or ICMP.

.

- For **transport mode AH** using IPv4, the AH is inserted after the original IP header and before the IP payload (e.g., a TCP segment); this is shown in the upper part of Figure 16.6b. Authentication covers the entire packet, excluding mutable fields in the IPv4 header that are set to zero for MAC calculation

Figure 16.6. Scope of AH Authentication



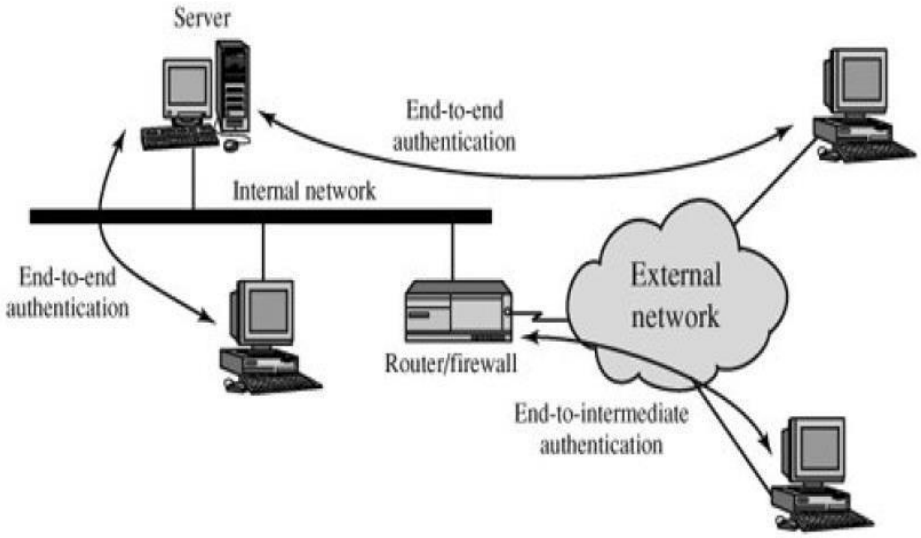


Figure 16.5. End-to-End versus End-to-Intermediate Authentication

- In the context of IPv6, AH is viewed as an end-to-end payload; that is, it is not examined or processedherefore, the AH appears after the IPv6 base header and the hop-by-hop, routing, and fragment extension headers. The destination options extension header could appear before or after the AH header, depending on the semantics desired. Again, authentication covers the entire packet, excluding mutable fields that are set to zero for MAC calculation.
- For **tunnel mode AH**, the entire original IP packet is authenticated, and the AH is inserted between the original IP header and a new outer IP header (Figure 16.6c). The inner IP header carries the ultimate source and destination addresses, while an outer IP header may contain different IP addresses (e.g., addresses of firewalls or other security gateways). With tunnel mode, the entire inner IP packet, including the entire inner IP header is protected by AH.
- The outer IP header (and in the case of IPv6, the outer IP extension headers) is protected except for mutable and unpredictable fields.

Encapsulating Security Payload (ESP)

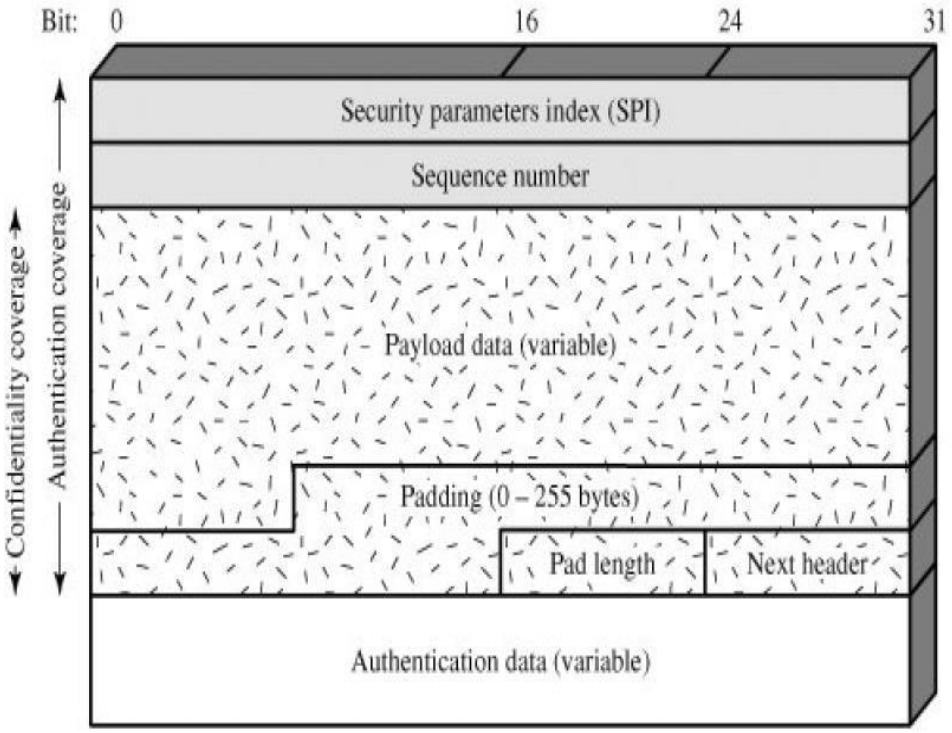
Contents
Encapsulating Security Payload • ESP Format • Encryption and Authentication Algorithms • Padding • Anti-Replay Service • Transport and Tunnel Modes

Encapsulating Security Payload

- The Encapsulating Security Payload provides confidentiality services, including confidentiality of message contents and limited traffic flow confidentiality. As an optional feature, ESP can also provide an authentication service.

ESP Format

- Figure 16.7 shows the format of an ESP packet. It contains the following fields:
  - **Security Parameters Index (32 bits):** Identifies a security association. **Sequence Number (32 bits):** A monotonically increasing counter value; this provides an antireplay function, as discussed for AH.
  - **Payload Data (variable):** This is a transport-level segment (transport mode) or IP packet (tunnel mode) that is protected by encryption.
  - **Padding (0255 bytes):** The purpose of this field is discussed later.
  - **Pad Length (8 bits):** Indicates the number of pad bytes immediately preceding this field.
  - **Next Header (8 bits):** Identifies the type of data contained in the payload data field by identifying the first header in that payload (for example, an extension header in IPv6, or an upper-layer protocol such as TCP).



  - **Authentication Data (variable):** A variable-length field (must be an integral number of 32-bit words) that contains the Integrity Check Value computed over the ESP packet minus the Authentication Data field.

Encryption and Authentication Algorithms

- The Payload Data, Padding, Pad Length, and Next Header fields are encrypted by the ESP service. If the algorithm used to encrypt the payload requires cryptographic

synchronization data, such as an initialization vector (IV), then these data may be carried explicitly at the beginning of the Payload Datafield.

- If included, an IV is usually not encrypted, although it is often referred to as being part of the ciphertext. The current specification dictates that a compliant implementation must support DES in cipher block chaining (CBC) mode).
- A number of other algorithms have been assigned identifiers in the DOI document and could therefore easily be used for encryption; these include
    - Three-key triple DES
    - RC5
    - IDEA
    - Three-key triple IDEA
    - CAST
    - Blowfish
- As with AH, ESP supports the use of a MAC with a default length of 96 bits. Also as with AH, the current specification dictates that a compliant implementation must support HMAC-MD5-96 and HMAC-SHA-1-96.

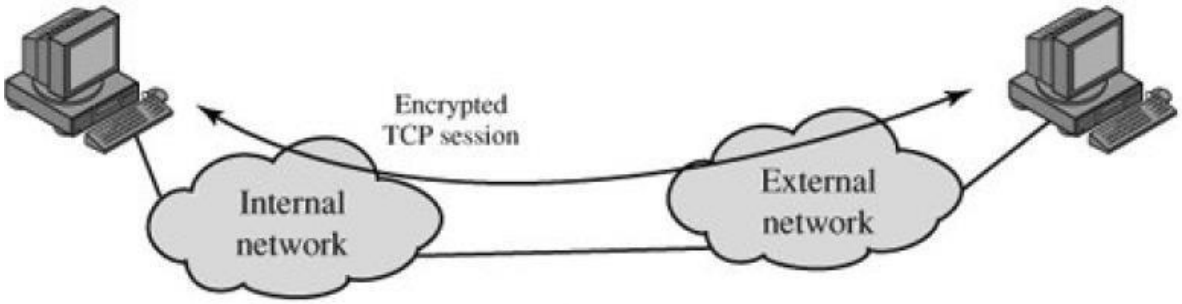
Padding

The Padding field serves several purposes:

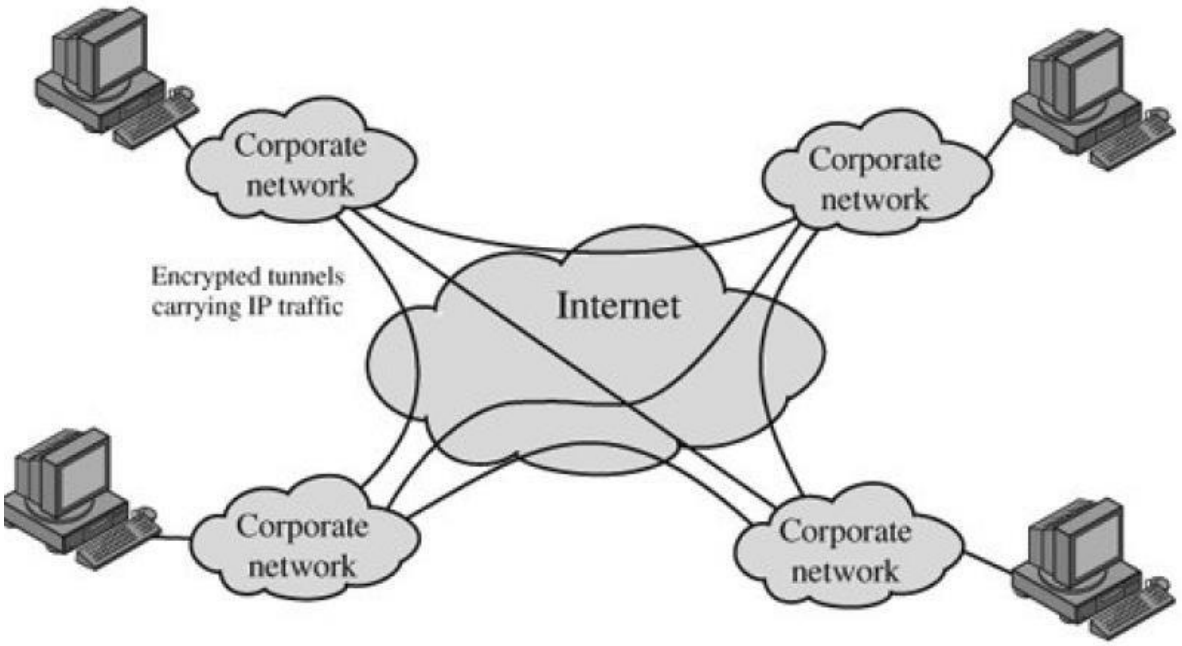
- If an encryption algorithm requires the plaintext to be a multiple of some number of bytes (e.g., the multiple of a single block for a block cipher), the Padding field is used to expand the plaintext (**consisting of the Payload Data, Padding, Pad Length, and Next Header fields**) to the required length.
- The ESP format requires that the Pad Length and Next Header fields be right aligned within a 32- bit word. Equivalently, the ciphertext must be an integer multiple of 32 bits. The Padding field is used to assure this alignment.
- Additional padding may be added to provide partial traffic flow confidentiality by concealing the actual length of the payload.

Transport and Tunnel Modes

- Figure 16.8 shows two ways in which the IPSec ESP service can be used. In the upper part of the figure, encryption (and optionally authentication) is provided directly between two hosts. Figure 16.8b shows how tunnel mode operation can be used to set up a *virtual private network*.
- In this example, an organization has four private networks interconnected across the Internet. Hosts on the internal networks use the Internet for transport of data but do not interact with other Internet-based hosts.
- By terminating the tunnels at the security gateway to each internal network, the configuration allows the hosts to avoid implementing the security capability. The former technique is support by a transport mode SA, while the latter technique uses a tunnel mode SA.



(a) Transport-level security



(b) A virtual private network via tunnel mode

Transport Mode ESP

- Transport mode ESP is used to encrypt and optionally authenticate the data carried by IP (e.g., a TCP segment), as shown in Figure 16.9a. For this mode using IPv4, the ESP header is inserted into the IP packet immediately prior to the transport-layer header (e.g., TCP, UDP, ICMP) and an ESP trailer (Padding, Pad Length, and Next Header fields) is placed after the IP packet; if authentication is selected, the ESP Authentication Data field is added after the ESP trailer.
- The entire transport-level segment plus the ESP trailer are encrypted. Authentication covers all of the ciphertext plus the ESP header.

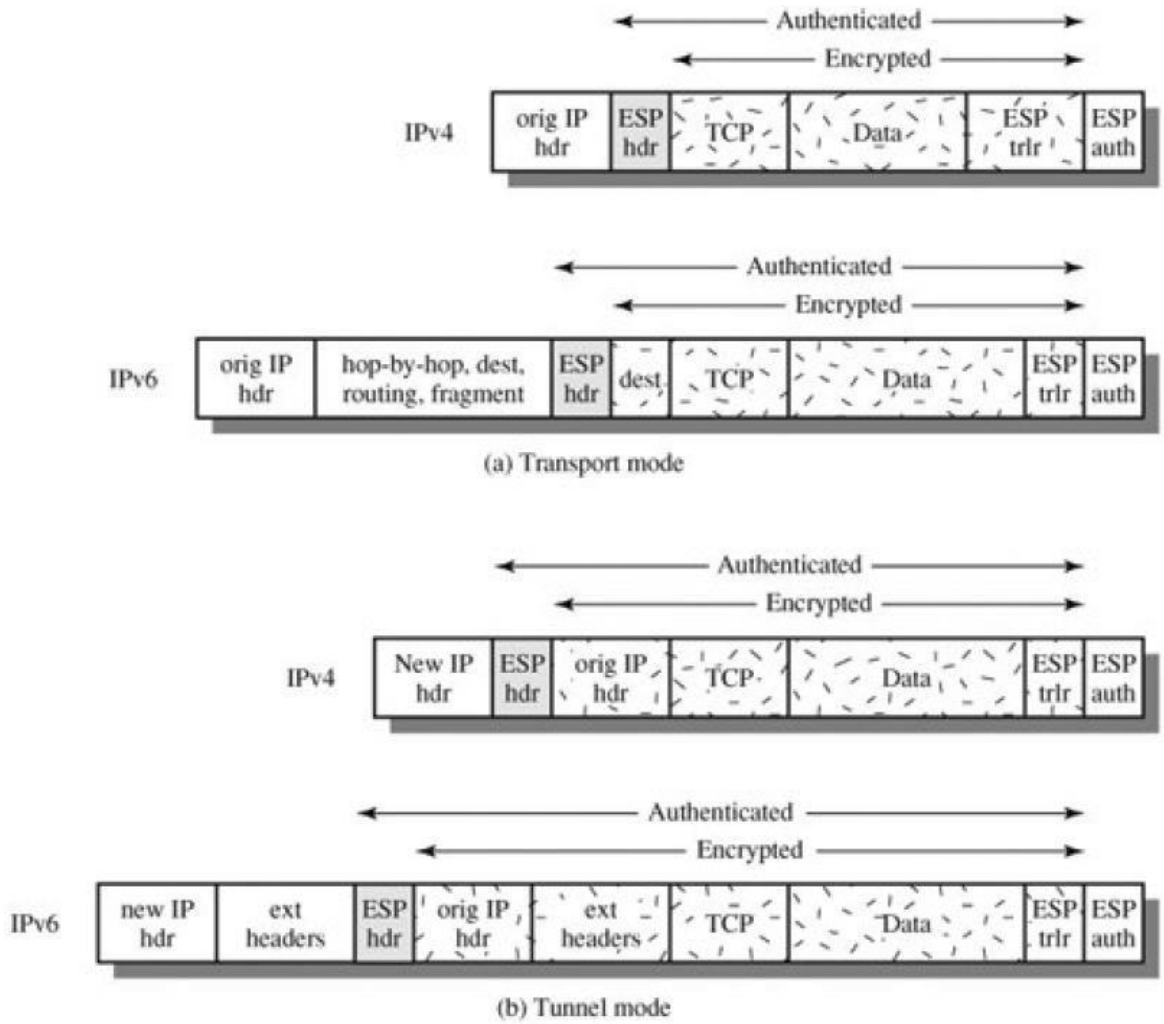


Figure 16.9. Scope of ESP Encryption and Authentication

- In the context of IPv6, ESP is viewed as an end-to-end payload; that is, it is not examined or processed by intermediate routers. Therefore, the ESP header appears after the IPv6 base header and the hop-byhop, routing, and fragment extension headers.
- The destination options extension header could appear before or after the ESP header, depending on the semantics desired. For IPv6, encryption covers the entire transport-level segment plus the ESP trailer plus the destination options extension header if it occurs after the ESP header. Again, authentication covers the ciphertext plus the ESP header.

Tunnel Mode ESP

- Tunnel mode ESP is used to encrypt an entire IP packet (Figure 16.9b). For this mode, the ESP header is prefixed to the packet and then the packet plus the ESP trailer is encrypted. This method can be used to counter traffic analysis. Because the IP header contains the destination address and possibly source routing directives and hopby- hop option information, it is not possible simply to transmit the encrypted IP packet prefixed by the ESP header. Intermediate routers would be unable to process such a packet.
- Therefore, it is necessary to encapsulate the entire block (ESP header plus ciphertext plus Authentication Data, if present) with a new IP header that will contain sufficient information for routing but not for traffic analysis.Whereas the transport mode is suitable for protecting connections between hosts that support the ESP feature, the tunnel mode is useful in a configuration that includes a firewall or other sort of security

gateway that protects a trusted network from external networks. In this latter case, encryption occurs only between an external host and the security gateway or between two security gateways.

- This relieves hosts on the internal network of the processing burden of encryption and simplifies the key distribution task by reducing the number of needed keys. Further, it thwarts traffic analysis based on ultimate destination.

VARIOUS ASPECTS OF IPV6

Contents
IPv6 (Internet Protocol Version 6) o Advantages of IPv6: o IPv6 Addresses o CIDR Notation: o IPv6 Packet Format:

IPv6 (Internet Protocol Version 6)

- IPv6 is the next generation Internet Protocol designed as a successor to the IP version 4.
- IPv6 was designed to enable high-performance, scalable Internet.
- This was achieved by overcoming many of the weaknesses of IPv4 protocol and by adding several new features.

Advantages of IPv6:

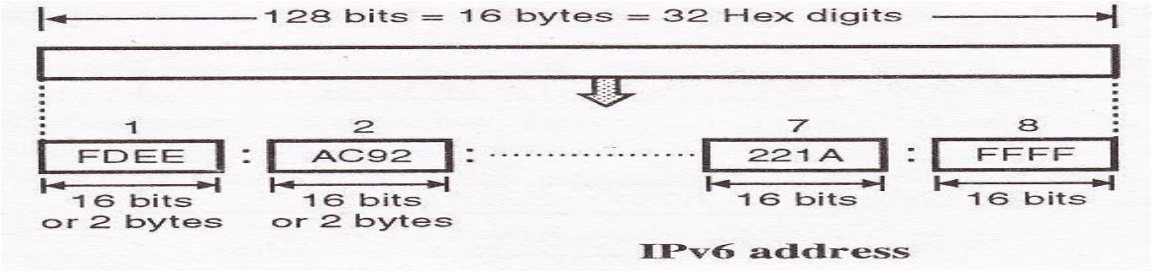
1. *Larger address space*
   - IPv6 has 128-bit address space, which is 4 times wider in bits in compared to IPv4's 32-bit address space.
   - So there is a huge increase in the address space.
2. *Better header forma*
   - IPv6 uses a better header format. In its header format the options are separated from the base header.
   - The options are inserted when needed, between the base header and upper layer data.
   - The helps in speeding up the routing process.
3. *New option*
   - New options have been added in IPv6 to increase the functionality.
4. *Possibility of extension*
   - IPv6 has been designed in such a way that there is a possibility of extension of protocol if required.
5. *More security*

- IPv6 includes security in the basic specification.
- It includes encryption of packets (ESP: Encapsulated Security Payload) and authentication of the sender of packets (AH: Authentication Header).

6. *Support to resource allocation*
   - To implement better support for real time traffic (such as video conference), IPv6 includes flow label in the specification.
   - With flow label mechanism, routers can recognize to which end-to-end flow the packets belongs.
7. *Plug and play*
   - IPv6 includes plug and play in the standard specification.
   - It therefore must be easier for novice users to connect their machines to the network, it will be done automatically.
8. *Clearer specification and optimization*
   - IPv6 follows good practices of IPv4, and rejects minor flaws/obsolete items of IPv4.

IPv6 Addresses

An IPv6 addresses consists of 16 bytes (octets) i.e. it is 128 bits long as shown in the below figure.

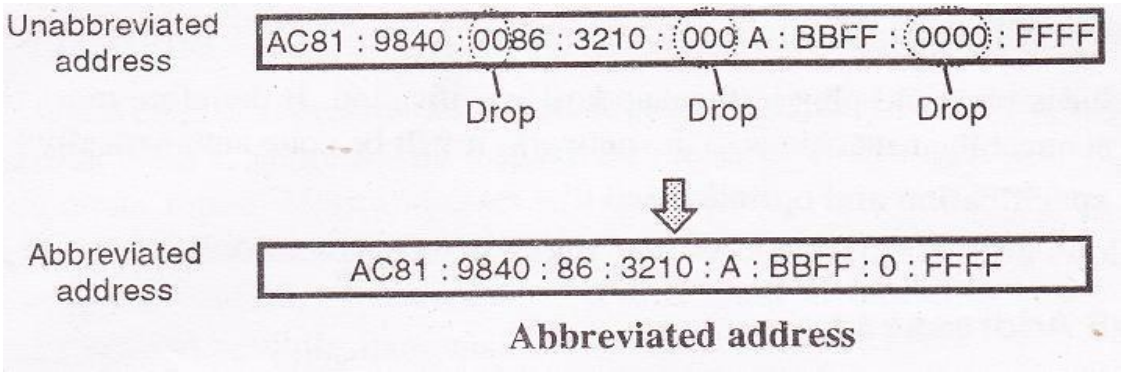


Hexadecimal colon notation:

- IPv6 uses a special notation called hexadecimal colon notation. In this, the 128 bits are divided into 8 sections; each one is 2 bytes long.
- 2 bytes correspond to 16 bits. So in hexadecimal notation will require four hexadecimal digits.
- Hence the IPv6 address consists of 32 hex digits and every group of 4 digits is separated by a colon as shown in the above figure.
- IPv6 uses 128-bit addresses. Only about 15% of the address space is initially allocated, the remaining 85% being reserved for future use.
- This remainder may be used in the future for expanding the address spaces of existing address types or for totally new uses.

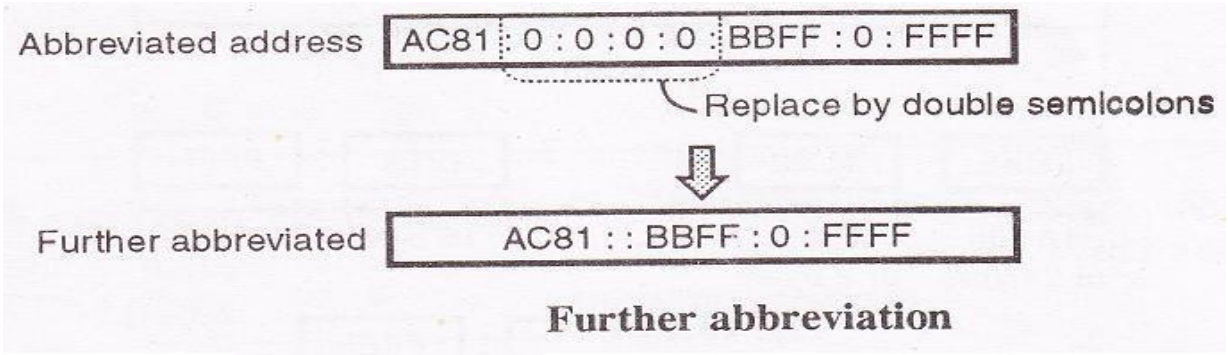
Abbreviation:

- The IPv6 address, even in hexadecimal format is very long. But in this address there are many of the zero digits in it.
- In such a case, we can abbreviate the address. The leading zeros of a section (four digits between two colons) can be omitted.
- Note that oOnly the leading zeros can be dropped but the trailing zeros can not drop. This is illustrated in the below figure.



Further abbreviation:

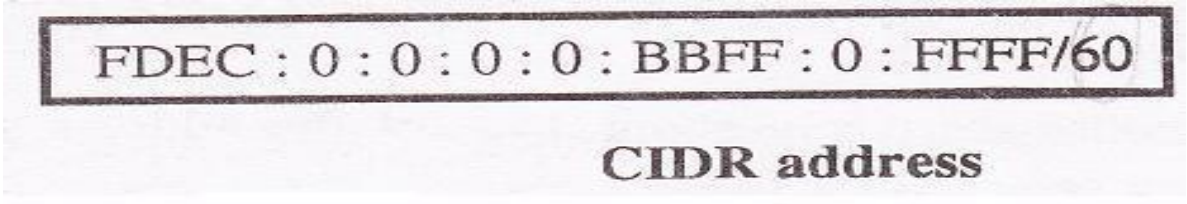
- Further abbreviations are possible if there is consecutive section consisting of only zeros.
- We can remove the zeros completely and replace them with double semicolon as shown in the below figure.



- It is important to note abbreviation is allowed only once per address. Also note that if there are two runs of zero sections, then only one of them can be abbreviated.

CIDR Notation:

- IPv6 protocol allows classless addressing and CIDR notation.
- The below figure shows how to define a prefix of 60 bits using CIDR.



Categories of Address:

IPv6 defines three different types of addresses.

Unicast

- A unicast address defines a single computer.
- A packet sent to a unicast address is delivered to that specific computer.

Anycast

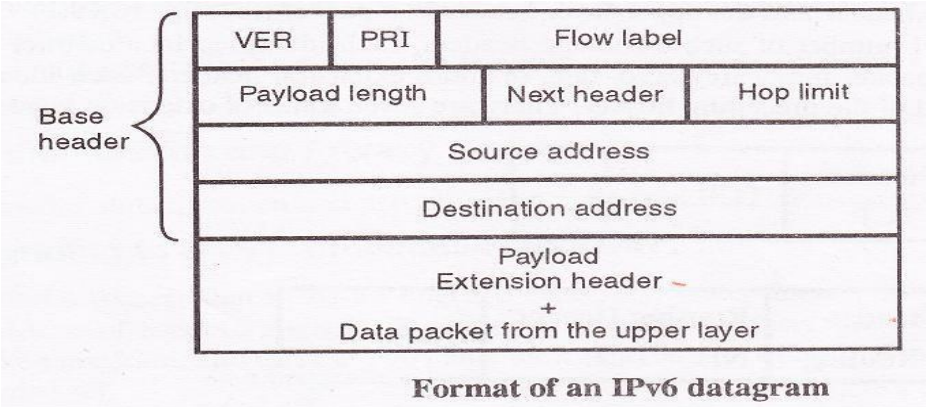
- This is a type of address defines a group of computers with addresses which have the same prefix.
- A packet sent to an anycast address must be delivered to exactly one of the members of the group which is the closest or the most easily accessible.

Multicast Addresses:

- A multicast address defines a group of computers which may or may not share the same prefix and may or may not be connected to the same physical network.
- A packet sent to a multicast address must be delivered to each member of the set.
- There are no broadcast addresses in IPv6, because multicast addresses can perform the same function. The type of address is determined by the leading bits.
- Multicast addresses all start with FF (1111 1111) and all other addresses are unicast addresses.
- Anycast addresses are assigned from the unicast address space and they do not differ syntactically from unicast addresses.
- Anycast addressing is a rather new concept and there is little experience with the widespread use of anycast addresses.
- Therefore, some restrictions apply to anycast addressing in IPv6 until more experience is gained.
- An anycast address may not be used as the Source Address of an IPv6 packet and anycast addresses may not be assigned to hosts but to routers only.

IPv6 Packet Format:

- The IPv6 packet is shown in the below figure. Each packet consists of a base header which is mandatory followed by the payload.
- The payload is made up of two parts
  - Optional extension headers and
  - Data from an upper layer



Format of an IPv6 datagram

- The base header is 40 byte length whereas the extension header and the data from upper layer contain upto 65,535 bytes of information.

Base header

In the base header have eight fields.

These fields are as follows:

1) **Version (VER):** It is a 4 bit field which defines the version of IP such as IPv4 or IPv6. For IPv6 the value of this field is 6.
2) **Priority:** It is a 4 bit field which defines the priority of the packet which is important in connection with the traffic congestion.
3) **Flow label:** It is a 24 bit (3 byte) field which is designed for providing special handling for a particular flow of data.
4) **Payload length:** This is a 2 byte length field which is used to define the total length of the IP datagram excluding the base header.
5) **Next header:** It is an 8 bit field which defines the header which follows the base header in the datagram.
6) **Hop limit:** This is an 8 bit field which has the same purpose as TTL in IPv4.
7) **Source address:** It is a 16 byte (128) Internet address which identifies the original source of datagram.
8) **Destination address:** This is a 16 byte (128) internet address which identifies the final destination of datagram. But this field will contain the address of the next router if source routing is being used.

KEY MANAGEMENT OF IPSEC OR INTERNET KEY EXCHANGE PROTOCOL

Contents
Internet Key Exchange • Key Determination Protocol • Header and Payload Formats

Internet Key Exchange

- The key management portion of IPsec involves the determination and distribution of secret keys. A typical requirement is four keys for communication between two applications: transmit and receive pairs for both integrity and confidentiality.
- The IPsec Architecture document mandates **support for two types of key management:**
  - **Manual:** A system administrator manually configures each system with its own keys and with the keys of other communicating systems. This is practical for small, relatively static environments.
  - **Automated:** An automated system enables the on-demand creation of keys for SAs and facilitates the use of keys in a large distributed system with an evolving configuration.
  - The default automated key management protocol for IPsec is referred to as ISAKMP/Oakley and **consists of the following elements:**

- **Oakley Key Determination Protocol:**
  - ✓ Oakley is a key exchange protocol based on the Diffie-Hellman algorithm but providing added security. Oakley is generic in that it does not dictate specific formats.
- **Internet Security Association and Key Management Protocol (ISAKMP):**
  - ✓ ISAKMP provides a framework for Internet key management and provides the specific protocol support, including formats, for negotiation of security attributes.
  - ✓ ISAKMP by itself does not dictate a specific key exchange algorithm; rather, ISAKMP consists of a set of message types that enable the use of a variety of key exchange algorithms. Oakley is the specific key exchange algorithm mandated for use with the initial version of ISAKMP.
  - ✓ In IKEv2, the terms Oakley and ISAKMP are no longer used, and there are significant differences from the use of Oakley and ISAKMP in IKEv1. Nevertheless, the basic functionality is the same. In this section, we describe the IKEv2 specification.

Key Determination Protocol

- IKE key determination is a refinement of the Diffie-Hellman key exchange algorithm. Recall that Diffie-Hellman involves the following interaction between users A and B.
- There is prior agreement on two global parameters: *q*, a large prime number; and a, a primitive root of *q*. A selects a random integer *XA* as its private key and transmits to B its public key *_A* = a*XA* mod *q*. Similarly, B selects a random integer *XB* as its private key and transmits to A its public key *_B* = a*XB* mod *q*. Each side can now compute the secret session key:

$$K = (\mathrm{Y}_B)^{X_A} \bmod q = (\mathrm{Y}_A)^{X_B} \bmod q = \alpha^{X_A X_B} \bmod q$$

The Diffie-Hellman algorithm has two attractive features:

- Secret keys are created only when needed. There is no need to store secret keys for a long period of time, exposing them to increased vulnerability.
- The exchange requires no pre-existing infrastructure other than an agreement on the global parameters. However, there are a number of weaknesses to Diffie-Hellman, as pointed out in [HUIT98]. • It does not provide any information about the identities of the parties.
- It is subject to a man-in-the-middle attack, in which a third party C impersonates B while communicating with A and impersonates A while communicating with B. Both A and B end up negotiating a key with C, which can then listen to and pass on traffic.
  The man-in-the-middle attack proceeds as
  1. B sends his public key *YB* in a message addressed to A (see Figure 10.2).
  2. The enemy (E) intercepts this message. E saves B's public key and sends a message to A that has B's User ID but E's public key *YE*. This message is sent in such a way that it appears as though it was sent from B's host system. A receives E's message and stores E's public key with B's User ID. Similarly, E sends a message to B with E's public key, purporting to come

3. from A.
4. B computes a secret key *K*1 based on B's private key and *YE*. A computes a secret key *K*2 based on A's private key and *YE*. E computes *K*1 using E's secret key *XE* and *YB* and computers *K*2 using *XE* and *YA*.
5. From now on, E is able to relay messages from A to B and from B to A, appropriately changing their encipherment en route in such a way that neither A nor B will know that they share their communication with E.

- It is computationally intensive. As a result, it is vulnerable to a clogging attack, in which an opponent requests a high number of keys. The victim spends considerable computing resources doing useless modular exponentiation rather than real work.
- IKE key determination is designed to retain the advantages of Diffie-Hellman, while countering its weaknesses.

Features of IKE key determination

- **The IKE key determination algorithm is characterized by five important features**:
  - ✓ .It employs a mechanism known as cookies to thwart clogging attacks.
  - ✓ It enables the two parties to negotiate a *group*; this, in essence, specifies the global parameters of the Diffie-Hellman key exchange.
  - ✓ It uses nonces to ensure against replay attacks.
  - ✓ It enables the exchange of Diffie-Hellman public key values.. It authenticates the Diffie-Hellman exchange to thwart man-in-the-middle attacks.

- **IKE mandates that cookie generation satisfy three basic requirements**:
  - ✓ The cookie must depend on the specific parties. This prevents an attacker from obtaining a cookie using a real IP address and UDP port and then using it to swamp the victim with requests from randomly chosen IP addresses or ports.
  - ✓ It must not be possible for anyone other than the issuing entity to generate cookies that will be accepted by that entity. This implies that the issuing entity will use local secret information in the generation and subsequent verification of a cookie. It must not be possible to deduce this secret information from any particular cookie. The point of this requirement is that the issuing entity need not save copies of its cookies, which are then more vulnerable to discovery, but can verify an incoming cookie acknowledgment when it needs to.
  - ✓ The cookie generation and verification methods must be fast to thwart attacks intended to sabotage processor resources.

- IKE key determination employs **nonces** to ensure against replay attacks. Each nonce is a locally generated pseudorandom number. Nonces appear in responses and are encrypted during certain portions of the exchange to secure their use.

- Three different **authentication** methods can be used with IKE key determination:

  - **Digital signatures:** The exchange is authenticated by signing a mutually obtainable hash; each party encrypts the hash with its private key. The hash is generated over important parameters, such as user IDs and nonces.
  - **Public-key encryption:** The exchange is authenticated by encrypting parameters such as IDs and nonces with the sender's private key.
  - **Symmetric-key encryption:** A key derived by some out-of-band mechanism can be used to authenticate the exchange by symmetric encryption of exchange parameters.

Header and Payload Formats

- IKE defines procedures and packet formats to establish, negotiate, modify, and delete security associations. As part of SA establishment, IKE defines payloads for exchanging key generation and authentication data.
- These payload formats provide a consistent framework independent of the specific key exchange protocol, encryption algorithm, and authentication mechanism.

1. IKE Header Format

- An IKE message consists of an IKE header followed by one or more payloads
- .All of this is carried in a transport protocol. The specification dictates that implementations must support the use of UDP for the transport protocol.

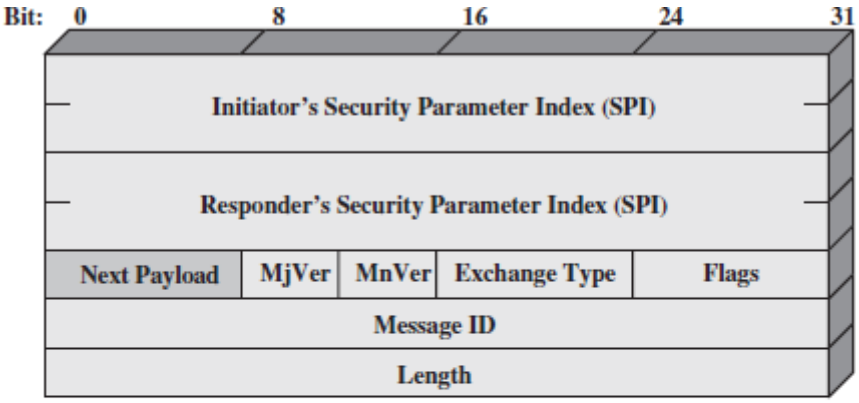


Figure 20.12 IKE Formats

- Figure 20.12a shows the header format for an IKE message. It consists of the following fields.
  - **Initiator SPI (64 bits):** A value chosen by the initiator to identify a unique IKE security association (SA).
  - **Responder SPI (64 bits):** A value chosen by the responder to identify a unique IKE SA.
  - **Next Payload (8 bits):** Indicates the type of the first payload in the message; payloads are discussed in the next subsection.
  - **Major Version (4 bits):** Indicates major version of IKE in use.
  - **Minor Version (4 bits):** Indicates minor version in use.
  - **Exchange Type (8 bits):** Indicates the type of exchange
  - **Flags (8 bits):** Indicates specific options set for this IKE exchange. Three bits are defined so far. The initiator bit indicates whether this packet is sent by the SA initiator. The version bit indicates whether the transmitter is capable of using a higher major version number than the one currently indicated. The response bit indicates whether this is a response to a message containing the same message ID.
  - **Message ID (32 bits):** Used to control retransmission of lost packets and matching of requests and responses.

- **Length (32 bits):** Length of total message (header plus all payloads) in octets.

2. IKE Payload Types

- All IKE payloads begin with the same generic payload header shown in Figure 20.12b. The Next Payload field has a value of 0 if this is the last payload in the message; otherwise its value is the type of the next payload. The Payload Length field indicates the length in octets of this payload, including the generic payload header.
- The critical bit is 0 if the sender wants the recipient to skip this payload if it does not understand the payload type code in the Next Payload field of the previous payload. It is set to 1 if the sender wants the recipient to reject this entire message if it does not understand the payload type.

- These elements are formatted as substructures within the payload as follows.

  - **Proposal:** This substructure includes a proposal number, a protocol ID (AH, ESP, or IKE), an indicator of the number of transforms, and then a transform substructure. If more than one protocol is to be included in a proposal, then there is a subsequent proposal substructure with the same proposal number.
  - **Transform:** Different protocols support different transform types. The transforms are used primarily to define cryptographic algorithms to be used with a particular protocol.
  - **Attribute:** Each transform may include attributes that modify or complete the specification of the transform. An example is key length.

Table 20.3 IKE Payload Types

Type	Parameters
Security Association	Proposals
Key Exchange	DH Group #, Key Exchange Data
Identification	ID Type, ID Data
Certificate	Cert Encoding, Certificate Data
Certificate Request	Cert Encoding, Certification Authority
Authentication	Auth Method, Authentication Data
Nonce	Nonce Data
Notify	Protocol-ID, SPI Size, Notify Message Type, SPI, Notification Data
Delete	Protocol-ID, SPI Size, # of SPIs, SPI (one or more)
Vendor ID	Vendor ID
Traffic Selector	Number of TSs, Traffic Selectors
Encrypted	IV, Encrypted IKE payloads, Padding, Pad Length, ICV
Configuration	CFG Type, Configuration Attributes
Extensible Authentication Protocol	EAP Message

  -

- The **Key Exchange payload** can be used for a variety of key exchange techniques, including Oakley, Diffie-Hellman, and the RSA-based key exchange used by PGP. The Key Exchange data field contains the data required to generate a session key and is dependent on the key exchange algorithm used.

- The **Identification payload** is used to determine the identity of communicating peers and may be used for determining authenticity of information. Typically the ID Data field will contain an IPv4 or IPv6 address.
- The **Certificate payload** transfers a public-key certificate. The Certificate Encoding field indicates the type of certificate or certificate-related information, which may include the following:
    - PKCS #7 wrapped X.509 certificate
    - PGP certificate
    - DNS signed key
    - X.509 certificate—signature
    - X.509 certificate—key exchange
    - Kerberos tokens
    - Certificate Revocation List (CRL)
    - Authority Revocation List (ARL)
    - SPKI certificate
- At any point in an IKE exchange, the sender may include a **Certificate Request** payload to request the certificate of the other communicating entity. The payload may list more than one certificate type that is acceptable and more than one certificate authority that is acceptable.
- The **Authentication** payload contains data used for message authentication purposes. The authentication method types so far defined are RSA digital signature, shared-key message integrity code, and DSS digital signature.
- The **Nonce** payload contains random data used to guarantee liveness during an exchange and to protect against replay attacks.
- The **Notify** payload contains either error or status information associated with this SA or this SA negotiation. The following table lists the IKE notify messages.

Error Messages	Status Messages
Unsupported Critical	Initial Contact
Payload	Set Window Size
Invalid IKE SPI	Additional TS Possible
Invalid Major Version	IPCOMP Supported
Invalid Syntax	NAT Detection Source IP
Invalid Payload Type	NAT Detection Destination IP
Invalid Message ID	Cookie
Invalid SPI	Use Transport Mode

- The **Delete** payload indicates one or more SAs that the sender has deleted from its database and that therefore are no longer valid.
- The **Vendor ID** payload contains a vendor-defined constant. The constant is used by vendors to identify and recognize remote instances of their implementations. This mechanism allows a vendor to experiment with new features while maintaining backward compatibility.
- The **Traffic Selector** payload allows peers to identify packet flows for processing by IPsec services.

- The **Encrypted** payload contains other payloads in encrypted form. The encrypted payload format is similar to that of ESP. It may include an IV if the encryption algorithm requires it and an ICV if authentication is selected.
- The **Configuration** payload is used to exchange configuration information between IKE peers.
- The **Extensible Authentication Protocol (EAP)** payload allows IKE SAs to be authenticated using EAP,

5.3. WEB SECURITY

Contents
Web Security
Web Security Considerations
Web Security Threats
Web Traffic Security Approaches
Secure Socket Layer and Transport Layer Security
SSL Architecture
SSL Record Protocol
Change Cipher Spec Protocol
Alert Protocol
Handshake Protocol
Cryptographic Computations
Transport Layer Security
Secure Electronic Transaction
SET Overview
Dual Signature
Payment Processing

Web Security Considerations

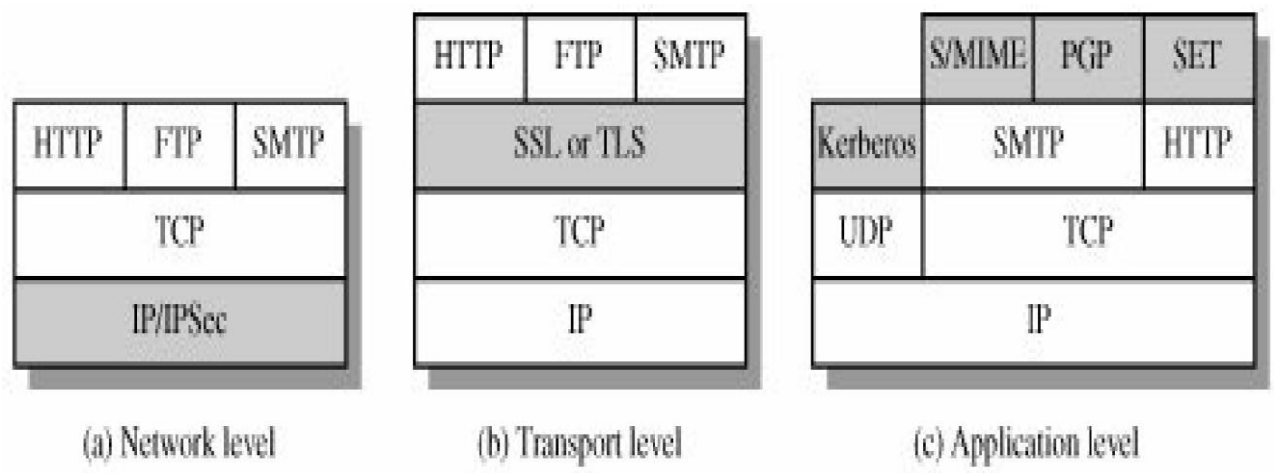
- The World Wide Web is fundamentally a client/server application running over the Internet and TCP/IP intranets.
- As such, the security tools and approaches discussed so far in this book are relevant to the issue of Web security.
- But, as pointed out in [GARF97], the Web presents new challenges not generally appreciated in the context of computer and network security:
- The Internet is two way. Unlike traditional publishing environments, even electronic publishing systems involving tele text, voice response, or fax-back, the Web is vulnerable to attacks on the Web servers over the Internet.
- The Web is increasingly serving as a highly visible outlet for corporate and product information and as the platform for business transactions.
- Reputations can be damaged and money can be lost if the Web servers are subverted.

Web Security Threats

- Table 17.1 provides a summary of the types of security threats faced in using the Web. One way to group these threats is in terms of passive and active attacks.
- Passive attacks include eavesdropping on network traffic between browser and server and gaining access to information on a Web site that is supposed to be restricted.
- Active attacks include impersonating another user, altering messages in transit between client and server, and altering information on a Web site.

Web Traffic Security Approaches

- A number of approaches to providing Web security are possible. The various approaches that have been considered are similar in the services they provide and, to some extent, in the mechanisms that they use, but they differ with respect to their scope of applicability and their relative location within the TCP/IP protocol stack.
- The advantage of using IPSec is that it is transparent to end users and applications and provides a general-purpose solution. Further, IPSec includes a filtering capability so that only selected traffic need incur the overhead of IPSec processing.



	Threats	Consequences	Countermeasures
Integrity	•Modification of user data •Trojan horse browser •Modification of memory •Modification of message traffic in transit	•Loss of information •Compromise of machine •Vulnerabilty to all other threats	Cryptographic checksums
Confidentiality	•Eavesdropping on the Net •Theft of info from server •Theft of data from client •Info about network configuration •Info about which client talks to server	•Loss of information •Loss of privacy	Encryption, web proxies
Denial of Service	•Killing of user threads •Flooding machine with bogus requests •Filling up disk or memory •Isolating machine by DNS attacks	•Disruptive •Annoying •Prevent user from getting work done	Difficult to prevent
Authentication	•Impersonation of legitimate users •Data forgery	•Misrepresentation of user •Belief that false information is valid	Cryptographic techniques

Table 17.1. A Comparison of Threats on the Web

SECURE SOCKET LAYER SECURITY(SSL)

Contents

Secure Socket Layer

- SSL protocol is an internet protocol for secure exchange of information between a web browser and web server
- SSL is Designed to make use of TCP to provide a reliable end to end secure service
- SSL provides security services between TCP and applications that use TCP. The SSL protocol is an internet protocol for secure exchange of information between a web browser and web server
- Subsequently, when a consensus was reached to submit the protocol for Internet standardization, the TLS working group was formed within IETF to develop a common standard. This first published version of TLS can be viewed as essentially an SSLv3.1 and is very close to and backward compatible with SSLv3.
- The bulk of this section is devoted to a discussion of SSLv3. At the end of the section, the principal differences between SSLv3 and TLS are described.

SSL Architecture

- SSL is designed to make use of TCP to provide a reliable end-to-end secure service. SSL is not a single protocol but rather two layers of protocols, as illustrated in Figure 17.2.
- The SSL Record Protocol provides basic security services to various higher-layer protocols. In particular, the Hypertext Transfer Protocol (HTTP), which provides the transfer service for Web client/server interaction, can operate on top of SSL. Three higher-layer protocols are defined as part of SSL:
- The Handshake Protocol, The Change Cipher Spec Protocol, and the Alert Protocol.

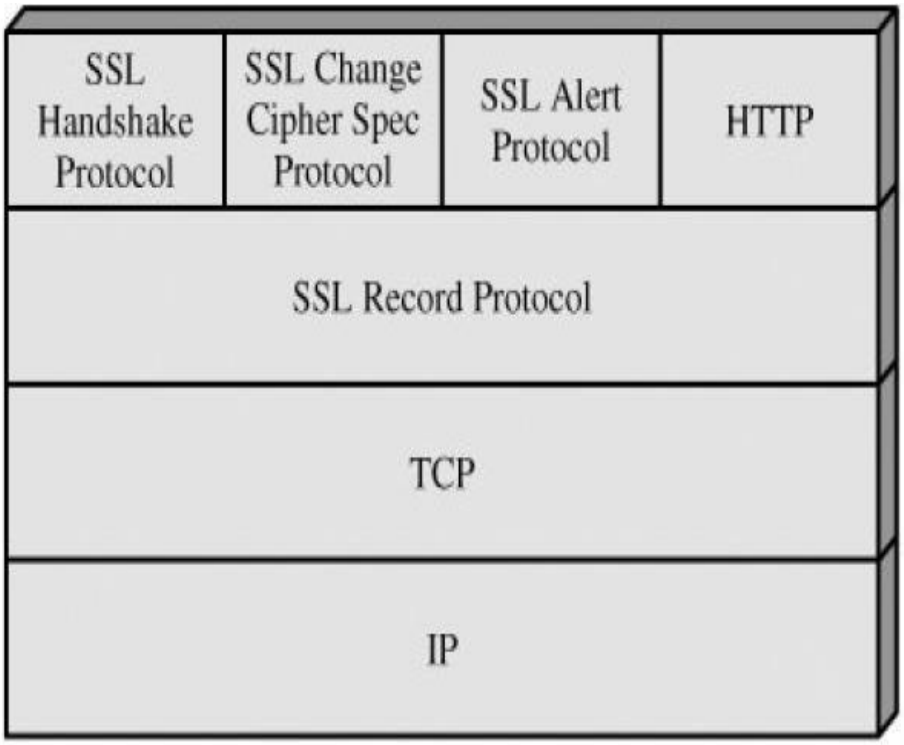


Figure 17.2. SSL Protocol Stack

- Two important SSL concepts are the SSL session and the SSL connection, which are defined in the specification as follows:
    - **Connection:** A connection is a transport (in the OSI layering model definition) that provides a suitable type of service. For SSL, such connections are peer-to-peer relationships. The connections are transient. Every connection is associated with one session.
    - **Session:** An SSL session is an association between a client and a server. Sessions are created by the Handshake Protocol. Sessions define a set of cryptographic security parameters, which can be shared among multiple connections.
- Sessions are used to avoid the expensive negotiation of new security parameters for each connection. Between any pair of parties (applications such as HTTP on client and server), there may be multiple secure connections.
- In theory, there may also be multiple simultaneous sessions between parties, but this feature is not used in practice.
- There are actually a number of states associated with each session.
- Once a session is established, there is a current operating state for both read and write (i.e., receive and send). In addition, during the Handshake Protocol, pending read and write states are created.

- Upon successful conclusion of the Handshake Protocol, the pending states become the current states. A session state is defined by the following parameters (definitions taken from the SSL specification):
    - **Session identifier:** An arbitrary byte sequence chosen by the server to identify an active or resumable session state.
    - **Peer certificate:** An X509.v3 certificate of the peer. This element of the state may be null.
    - **Compression method:** The algorithm used to compress data prior to encryption.
    - **Cipher spec:** Specifies the bulk data encryption algorithm (such as null, AES, etc.) and a hash algorithm (such as MD5 or SHA-1) used for MAC calculation. It also defines cryptographic attributes such as the hash size.
    - **Master secret:** 48-byte secret shared between the client and server.
    - **Is resumable:** A flag indicating whether the session can be used to initiate new connections.
- A connection state is defined by the following parameters:
    - **Server and client random:** Byte sequences that are chosen by the server and client for each connection.
    - **Server write MAC secret:** The secret key used in MAC operations on data sent by the server.
    - **Client write MAC secret:** The secret key used in MAC operations on data sent by the client.
    - **Server write key:** The conventional encryption key for data encrypted by the server and decrypted by the client.
    - **Client write key:** The conventional encryption key for data encrypted by the client and decrypted by the server.
    - **Initialization vectors:** When a block cipher in CBC mode is used, an initialization vector (IV) is maintained for each key. This field is first initialized by the SSL Handshake Protocol. Thereafter the final ciphertext block from each record is preserved for use as the IV with the following record.
    - **Sequence numbers:** Each party maintains separate sequence numbers for transmitted and received messages for each connection. When a party sends or receives a change cipher spec message, the appropriate sequence number is set to zero. Sequence numbers may not exceed 264

- **Confidentiality:** The Handshake Protocol defines a shared secret key that is used for conventional encryption of SSL payloads.
- **Message Integrity:** The Handshake Protocol also defines a shared secret key that is used to form a message authentication code (MAC).

- **SSL components:**
  - SSL Record Protocol
  - SSL Handshake Protocol
  - SSL Alert Protocol
  - SSL Change Cipher Spec Protocol

SSL Record Protocol

- The SSL Record Protocol provides two services for SSL connections:
  - **Confidentiality:** The Handshake Protocol defines a shared secret key that is used for conventional encryption of SSL payloads.
  - **Message Integrity:** The Handshake Protocol also defines a shared secret key that is used to form a message authentication code (MAC).

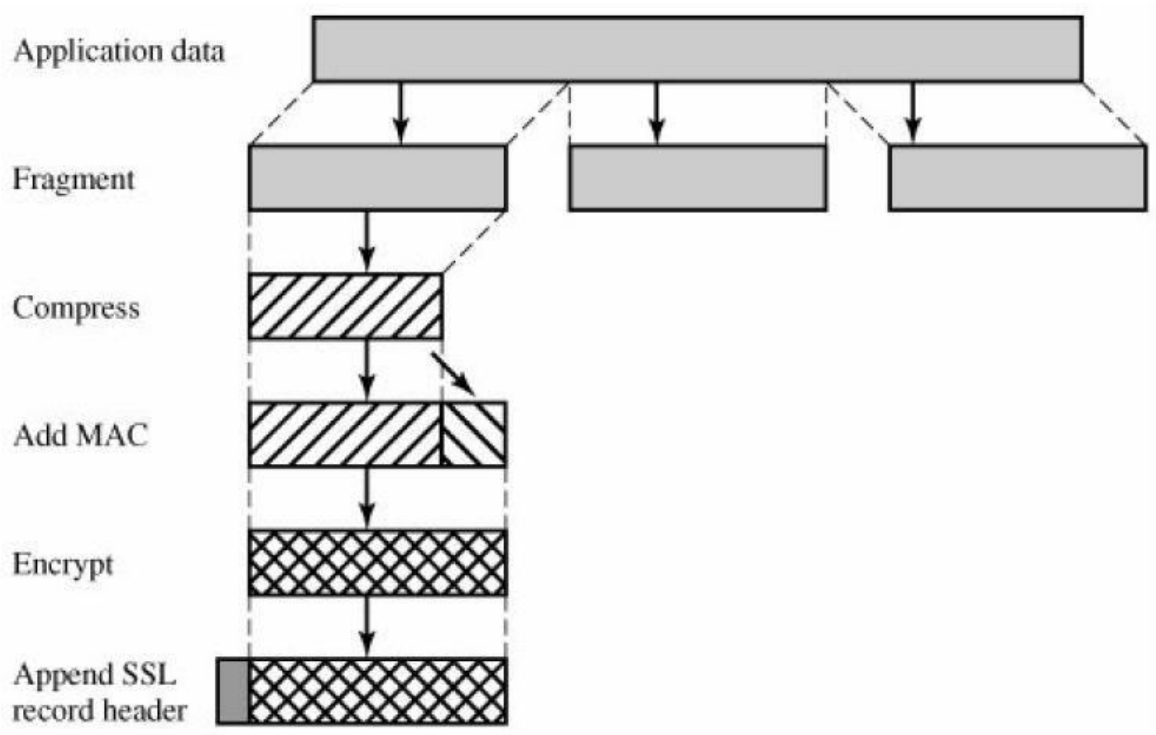


Figure 17.3. SSL Record Protocol Operation

- Figure 17.3 indicates the overall operation of the SSL Record Protocol. The Record Protocol takes an application message to be transmitted, fragments the data into manageable blocks, optionally compresses the data, applies a MAC, encrypts, adds a header, and transmits the resulting unit in a TCP segment.

- Received data are decrypted, verified, decompressed, and reassembled and then delivered to higher-level users.
- The first step is **fragmentation**. Each upper-layer message is fragmented into blocks of 214 bytes (16384 bytes) or less.
- Next, **compression** is optionally applied. Compression must be lossless and may not increase the content length by more than 1024 bytes. In SSLv3 (as well as the current version of TLS), no compression algorithm is specified, so the default compression algorithm is null.

- The next step in processing is to compute a **message authentication code** over the compressed data. For this purpose, a shared secret key is used. The calculation is defined as

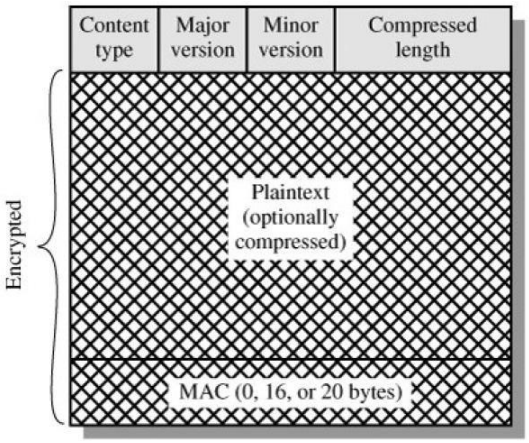
hash(MAC_write_secret || pad_2 ||
hash(MAC_write_secret || pad_1 || seq_num ||
SSLCompressed.type ||
SSLCompressed.length || SSLCompressed.fragment))

Where

|| = concatenation
MAC_write_secret = shared secret key
hash = cryptographic hash algorithm; either MD5 or SHA-1
pad_1 = the byte 0x36 (0011 0110) repeated 48 times (384 bits) for MD5 and 40 times (320 bits) for SHA-1
pad_2 = the byte 0x5C (0101 1100) repeated 48 times for MD5 and 40 times for
SHA-1
seq_num = the sequence number for this message
SSLCompressed.type = the higher-level protocol used to process this fragment
SSLCompressed.length = the length of the compressed fragment
SSLCompressed.fragment = the compressed fragment (if compression is not used, the plaintext Fragment

- Next, the compressed message plus the MAC are **encrypted** using symmetric encryption. Encryption may not increase the content length by more than 1024 bytes, so that the total length may not exceed 214 + 2048.
- The final step of SSL Record Protocol processing is to prepend a header, consisting of the following fields:
  - **Content Type (8 bits):** The higher layer protocol used to process the enclosed fragment.
  - **Major Version (8 bits):** Indicates major version of SSL in use. For SSLv3, the value is 3.
  - **Minor Version (8 bits):** Indicates minor version in use. For SSLv3, the value is 0.
  - **Compressed Length (16 bits):** The length in bytes of the plaintext fragment (or compressed
  - fragment if compression is used). The maximum value is 214 + 2048.

Figure 17.4 illustrates the SSL record format.



Change Cipher Spec Protocol

- The Change Cipher Spec Protocol is one of the three SSL-specific protocols that use the SSL Record Protocol, and it is the simplest. This protocol consists of a single message (Figure 17.5a), which consists of a single byte with the value 1.
- The sole purpose of this message is to cause the pending state to be copied into the current state, which updates the cipher suite to be used on this connection.

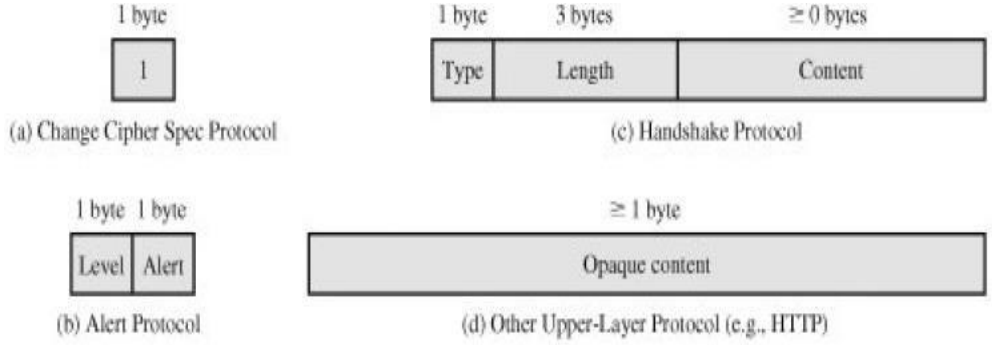


Figure 17.5. SSL Record Protocol Payload

Alert Protocol

- The Alert Protocol is used to convey SSL-related alerts to the peer entity. As with other applications that use SSL, alert messages are compressed and encrypted, as specified by the current state.
- Each message in this protocol consists of two bytes (Figure 17.5b). The first byte takes the value warning(1) or fatal(2) to convey the severity of the message.
- If the level is fatal, SSL immediately terminates the connection. Other connections on the same session may continue, but no new connections on this session may be established. The second byte contains a code that indicates the specific alert.
- First, we list those alerts that are always fatal (definitions from the SSL specification):
    - **Unexpected message:** An inappropriate message was received.

- **bad_record_mac:** An incorrect MAC was received.
- **Decompression failure:** The decompression function received improper input (e.g., unable to decompress or decompress to greater than maximum allowable length).
- **handshake failure:** Sender was unable to negotiate an acceptable set of security parameters given the options available.
- **illegal parameter:** A field in a handshake message was out of range or inconsistent with other fields.

- **The remainder of the alerts is the following:**
  - **close notify:** Notifies the recipient that the sender will not send any more messages on this connection. Each party is required to send a close_notify alert before closing the write side of a connection.
  - **no certificate:** May be sent in response to a certificate request if no appropriate certificate is available.
  - **bad certificate:** A received certificate was corrupt (e.g., contained a signature that did not verify).
  - **unsupported certificate:** The type of the received certificate is not supported.
  - **certificate revoked:** A certificate has been revoked by its signer.
  - **certificate expired:** A certificate has expired.
  - **certificate unknown:** Some other unspecified issue arose in processing the certificate, rendering it unacceptable.

Handshake Protocol

- The most complex part of SSL is the Handshake Protocol. This protocol allows the server and client to authenticate each other and to negotiate an encryption and MAC algorithm and cryptographic keys to be used to protect data sent in an SSL record.
- The Handshake Protocol is used before any application data is transmitted.
- Each message has three fields:
  - **Type (1 byte):** Indicates one of 10 messages. Table 17.2 lists the defined message types.
  - **Length (3 bytes):** The length of the message in bytes.
  - **Content (0 bytes):** The parameters associated with this message; these are listed in Table17.2.

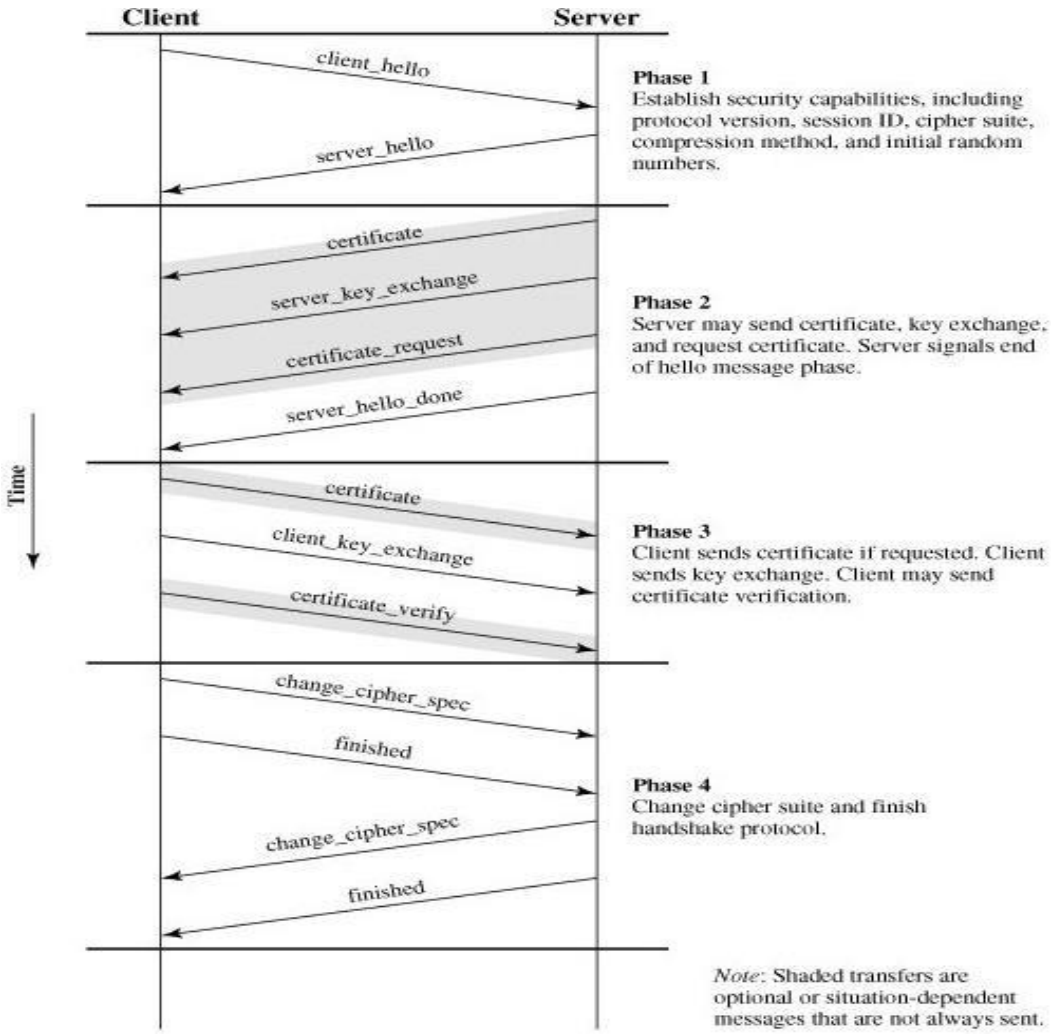


Figure 17.6. Handshake Protocol Action

Cryptographic Computations

- Two further items are of interest: the creation of a shared master secret by means of the key exchange, and the generation of cryptographic parameters from the master secret.

Master Secret Creation

- The shared master secret is a one-time 48-byte value (384 bits) generated for this session by means of secure key exchange. The creation is in two stages. First, a pre_master_secret is exchanged.
- Second, the master_secret is calculated by both parties. For pre_master_secret exchange, there are two possibilities:
    - **RSA:** A 48-byte pre_master_secret is generated by the client, encrypted with the server's public RSA key, and sent to the server. The server decrypts the ciphertext using its private key to recover the pre_master_secret.
    - **Diffie-Hellman:** Both client and server generate a Diffie-Hellman public key. After these are exchanged, each side performs the Diffie-Hellman calculation to create the shared pre_master_secret.

      Both sides now compute the master_secret as follows:
      master_secret = MD5(pre_master_secret || SHA('A' ||

pre_master_secret ||ClientHello.random ||
ServerHello.random)) ||
MD5(pre_master_secret || SHA('BB' ||
pre_master_secret || ClientHello.random ||
ServerHello.random)) ||
MD5(pre_master_secret || SHA('CCC' ||
pre_master_secret || ClientHello.random ||
ServerHello.random))

- where ClientHello.random and ServerHello.random are the two nonce values exchanged in the initial hello messages

Transport Layer Security(TLS)

- TLS is an IETF standardization initiative whose goal is to produce an Internet standard version of SSL.
- TLS is defined as a Proposed Internet Standard in RFC 2246. RFC 2246 is very similar to SSLv3. In this section, we highlight the differences.

Version Number

- The TLS Record Format is the same as that of the SSL Record Format (Figure 17.4), and the fields in the header have the same meanings. The one difference is in version values. For the current version of TLS, the Major Version is 3 and the Minor Version is 1.

Message Authentication Code

- There are two differences between the SSLv3 and TLS MAC schemes: the actual algorithm and the scope of the MAC calculation. TLS makes use of the HMAC algorithm defined in RFC 2104.
- HMAC is defined as follows:

$$\mathbf{HMAC}_K(M) = \mathbf{H}[(K^+ \oplus \mathbf{opad}) \| \mathbf{H}[(K^+ \oplus \mathbf{ipad}) \| M]]$$

Where

H = embedded hash function (for TLS, either MD5 or SHA-1)
M = message input to HMAC
K+ = secret key padded with zeros on the left so that the result is equal to the block length of the hash code(for MD5 and SHA-1, block length = 512 bits)
ipad = 00110110 (36 in hexadecimal) repeated 64 times (512 bits)
opad = 01011100 (5C in hexadecimal) repeated 64 times (512 bits)

- SSLv3 uses the same algorithm, except that the padding bytes are concatenated with the secret key rather than being XORed with the secret key padded to the block length. The level of security should be about the same in both cases.
- For TLS, the MAC calculation encompasses the fields indicated in the following expression:

HMAC_hash(MAC_write_secret, seq_num || TLSCompressed.type ||
TLSCompressed.version || TLSCompressed.length ||

TLSCompressed.fragment)

- The MAC calculation covers all of the fields covered by the SSLv3 calculation, plus the field
- TLSCompressed.version, which is the version of the protocol being employed.

Pseudorandom Function

- TLS makes use of a pseudorandom function referred to as PRF to expand secrets into blocks of data for purposes of key generation or validation.
- The objective is to make use of a relatively small shared secret value but to generate longer blocks of data in a way that is secure from the kinds of attacks made on hash functions and MACs.The PRF is based on the data expansion function (Figure 16.7) given as

P_hash(secret, seed)= HMAC_hash(secret,A(1) || seed) ||
HMAC_hash(secret, A(2) || seed) ||
HMAC_hash(secret, A(3) || seed) || . . .

where

A() is defined as

A(0) = seed

A(i) = HMAC_hash(secret,A(i – 1))

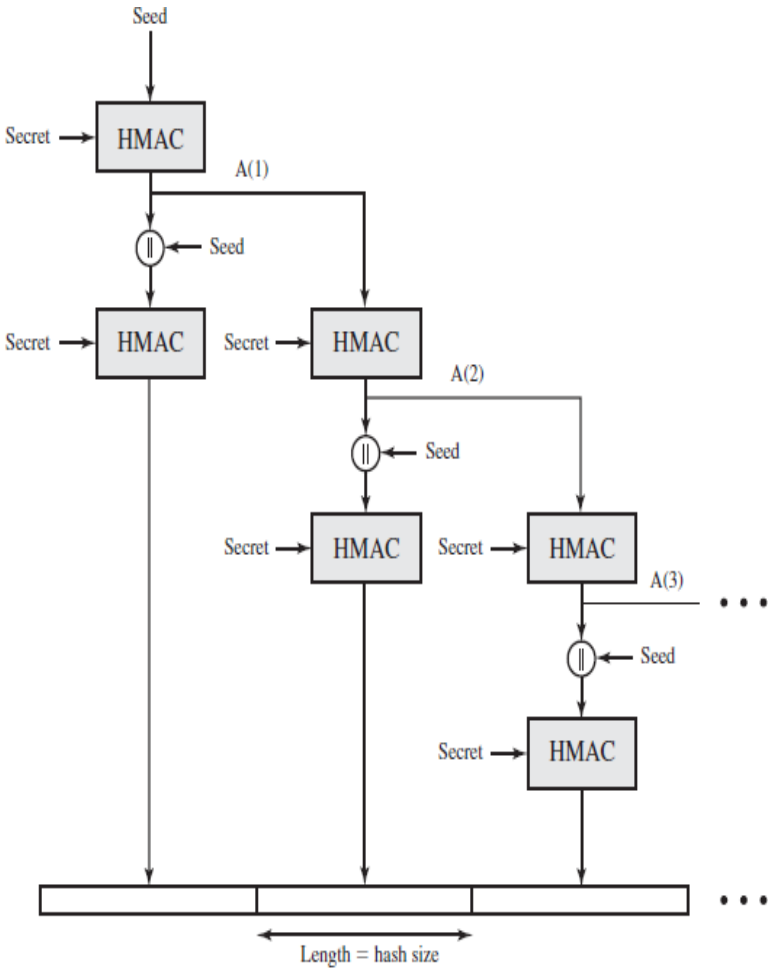


Figure 16.7 TLS Function `P_hash(secret, seed)`

Alert Codes

- TLS supports all of the alert codes defined in SSLv3 with the exception of no_certificate.A number of additional codes are defined in TLS; of these, the following are always fatal.
    - **record_overflow:** A TLS record was received with a payload (ciphertext) whose length exceeds bytes, or the ciphertext decrypted to a length of greater than bytes.
    - **unknown_ca:** A valid certificate chain or partial chain was received, but the certificate was not accepted because the CA certificate could not be located or could not be matched with a known, trusted CA.
    - **access_denied:** A valid certificate was received, but when access control was applied, the sender decided not to proceed with the negotiation.
    - **decode_error:** A message could not be decoded, because either a field was out of its specified range or the length of the message was incorrect.
    - **protocol_version:** The protocol version the client attempted to negotiate is recognized but not supported.
    - **insufficient_security:** Returned instead of handshake failure when a negotiation has failed specifically because the server requires ciphers more secure than those supported by the client.
    - **unsupported_extension:** Sent by clients that receive an extended server hello containing an extension not in the corresponding client hello.

- **internal_error:** An internal error unrelated to the peer or the correctness of the protocol makes it impossible to continue.
- **decrypt_error:** A handshake cryptographic operation failed, including being unable to verify a signature, decrypt a key exchange, or validate a finished message.

- **The remaining alerts include the following**.
  - **user_canceled:** This handshake is being canceled for some reason unrelated to a protocol failure.
  - n**o_renegotiation:** Sent by a client in response to a hello request or by the server in response to a client hello after initial handshaking. Either of these messages would normally result in renegotiation, but this alert indicates that the sender is not able to renegotiate.This message is always a warning.

Cipher Suites

- There are several small differences between the cipher suites available under SSLv3 and under TLS:
  - K**ey Exchange:** TLS supports all of the key exchange techniques of SSLv3 with the exception of Fortezza.
  - **Symmetric Encryption Algorithms:** TLS includes all of the symmetric encryption algorithms found in SSLv3, with the exception of Fortezza

5.4. SYSTEM SECURITY

5.4.1. INTRUDERS

Contents
• **Intruder** ✓ **Masquerader:** ✓ **Misfeasor:** ✓ **Clandestine user:** • **Intruder Behavior Patterns** ✓ **Hackers** ✓ **Criminals** ✓ **Insider Attacks** • **Intrusion Techniques** ✓ **One-way function:** ✓ **Access control:**

Intruder

One of the two most publicized threats to security is the intruder, often referred to as a hacker or cracker.

- **The identified three classes of intruders:**

  • **Masquerader:** An individual who is not authorized to use the computer and who penetrates a system's access controls to exploit a legitimate user's account

• **Misfeasor:** A legitimate user who accesses data, programs, or resources for which such access is not authorized, or who is authorized for such access but misuses his or her privileges

• **Clandestine user:** An individual who seizes supervisory control of the system and uses this control to evade auditing and access controls or to suppress audit collection.

- The masquerader is likely to be an outsider; the misfeasor generally is an insider; and the clandestine user can be either an outsider or an insider.

The following are examples of intrusion:

• Performing a remote root compromise of an e-mail server

• Defacing a Web server

• Guessing and cracking passwords

• Copying a database containing credit card numbers

• Viewing sensitive data, including payroll records and medical information, without authorization

• Running a packet sniffer on a workstation to capture usernames and passwords

• Using a permission error on an anonymous FTP server to distribute pirated software and music files

• Dialing into an unsecured modem and gaining internal network access

• Posing as an executive, calling the help desk, resetting the executive's e-mail password, and learning the new password

• Using an unattended, logged-in workstation without permission.

Intruder Behavior Patterns:

The three broad examples of intruder behavior patterns are

- **Hackers** Traditionally, those who hack into computers do so for the thrill of it or for status. The hacking community is a strong meritocracy in which status is determined by level of competence.

(a) Hacker

1. Select the target using IP lookup tools such as NSLookup, Dig, and others.
2. Map network for accessible services using tools such as NMAP.
3. Identify potentially vulnerable services (in this case, pcAnywhere).
4. Brute force (guess) pcAnywhere password.
5. Install remote administration tool called DameWare.
6. Wait for administrator to log on and capture his password.
7. Use that password to access remainder of network.

- **Criminals** Organized groups of hackers have become a widespread and common threat to Internet-based systems. These groups can be in the employ of a corporation or government but often are loosely affiliated gangs of hackers.

(b) Criminal Enterprise

1. Act quickly and precisely to make their activities harder to detect.
2. Exploit perimeter through vulnerable ports.
3. Use Trojan horses (hidden software) to leave back doors for reentry.
4. Use sniffers to capture passwords.
5. Do not stick around until noticed.
6. Make few or no mistakes.

- **Insider Attacks** Insider attacks are among the most difficult to detect and prevent. Employees already have access and knowledge about the structure and content of corporate databases. Insider attacks can be motivated by revenge or simply a feeling of entitlement.

(c) Internal Threat

1. Create network accounts for themselves and their friends.
2. Access accounts and applications they wouldn't normally use for their daily jobs.
3. E-mail former and prospective employers.
4. Conduct furtive instant-messaging chats.
5. Visit Web sites that cater to disgruntled employees, such as f'dcompany.com.
6. Perform large downloads and file copying.
7. Access the network during off hours.

Intrusion Techniques:

The password file can be protected in one of two ways:

• **One-way function:** The system stores only the value of a function based on the user's password. When the user presents a password, the system transforms that password and compares it with the stored value.
• **Access control:** Access to the password file is limited to one or a very few accounts.

The techniques for learning passwords:

1. Try default passwords used with standard accounts that are shipped with the system. Many administrators do not bother to change these defaults.
2. Exhaustively try all short passwords (those of one to three characters).
3. Try words in the system's online dictionary or a list of likely passwords. Examples of the latter are readily available on hacker bulletin boards.
4. Collect information about users, such as their full names, the names of their spouse and children, pictures in their office, and books in their office that are related to hobbies.
5. Try users' phone numbers, Social Security numbers, and room numbers.
6. Try all legitimate license plate numbers for this state.
7. Use a Trojan horse to bypass restrictions on access.
8. Tap the line between a remote user and the host system.

Contents

Intrusion detection system:

1. If an intrusion is detected quickly enough, the intruder can be identified and ejected from the system before any damage is done or any data are compromised.
2. An effective intrusion detection system can serve as a deterrent, so acting to prevent intrusions.
3. Intrusion detection enables the collection of information about intrusion techniques that can be used to strengthen the intrusion prevention facility.

The following approaches to intrusion detection:

1. **Statistical anomaly detection:** Involves the collection of data relating to the behavior of legitimate users over a period of time.

   a. ***Threshold detection:*** This approach involves defining thresholds, independent of user, for the frequency of occurrence of various events.
   b. ***Profile based:*** A profile of the activity of each user is developed and used to detect changes in the behavior of individual accounts.

2. **Rule-based detection:** Involves an attempt to define a set of rules that can be used to decide that a given behavior is that of an intruder.

   a. ***Anomaly detection:*** Rules are developed to detect deviation from previous usage patterns.
   b. ***Penetration identification:*** An expert system approach that searches for suspicious behavior.

Audit Records

- A fundamental tool for intrusion detection is the audit record. Some record of ongoing activity by users must be maintained as input to an intrusion detection system. Basically, two plans are used:

  • **Native audit records:** Virtually all multiuser operating systems include accounting software that collects information on user activity.

  • **Detection-specific audit records:** A collection facility can be implemented that generates audit records containing only that information required by the intrusion detection system.

Each audit record contains the following fields:

• **Subject:** Initiators of actions.
• **Action:** Operation performed by the subject on or with an object.
• **Object:** Receptors of actions.
• **Exception-Condition:** Denotes which, if any, exception condition is raised on return.
• **Resource-Usage:** A list of quantitative elements in which each element gives the amount used of some resource.
• **Time-Stamp:** Unique time-and-date stamp identifying when the action took place.

Statistical Anomaly Detection:

Statistical anomaly detection techniques fall into two broad categories:

- ***Threshold detection systems:*** Threshold detection involves counting the number of occurrences of a specific event type over an interval of time.
- ***Profile-based systems:*** Profile-based anomaly detection focuses on characterizing the past behavior of individual users or related groups of users and then detecting significant deviations.

Examples of metrics that are useful for profile-based intrusion detection are the following:

• **Counter:** A nonnegative integer that may be incremented but not decremented until it is reset by management action.
• **Gauge:** A nonnegative integer that may be incremented or decremented. Typically, a gauge is used to measure the current value of some entity.
• **Interval timer:** The length of time between two related events.
• **Resource utilization:** Quantity of resources consumed during a specified period.

Given these general metrics, various tests can be performed to determine whether current activity fits within acceptable limits. the following approaches that may be taken:

- **Mean and standard deviation**- of a parameter over some historical period. This gives a reflection of the average behavior and its variability.

- **Multivariate** - A **multivariate** model is based on correlations between two or more variables.
  - **Markov process-** A **Markov process** model is used to establish transition probabilities among various states.
  - **Time series-** A **time series** model focuses on time intervals, looking for sequences of events that happen too rapidly or too slowly.
  - **Operational-** Finally, an **operational model** is based on a judgment of what is considered abnormal, rather than an automated analysis of past audit records.

Rule-Based Intrusion Detection:

- Rule-based techniques detect intrusion by observing events in the system and applying a set of rules that lead to a decision regarding whether a given pattern of activity is or is not suspicious.
- **Rule-based anomaly detection** is similar in terms of its approach and strengths to statistical anomaly detection.
- With the rule-based approach, historical audit records are analyzed to identify usage patterns and to generate automatically rules that describe those patterns. Rules may represent past behavior patterns of users, programs, privileges, time slots, terminals, and so on.
- **Rule-based penetration identification** takes a very different approach to intrusion detection. The key feature of such systems is the use of rules for identifying known penetrations or penetrations that would exploit known weaknesses.
- Rules can also be defined that identify suspicious behavior, even when the behavior is within the bounds of established patterns of usage.

Distributed Intrusion Detection

- Until recently, work on intrusion detection systems focused on single-system standalone facilities.
- Porras points out the following major issues in the design of a distributed intrusion detection system :
  - A distributed intrusion detection system may need to deal with different audit record formats.
  - One or more nodes in the network will serve as collection and analysis points for the data from the systems on the network.
  - Either a centralized or decentralized architecture can be used. With a centralized architecture, there is a single central point of collection and analysis of all audit data.

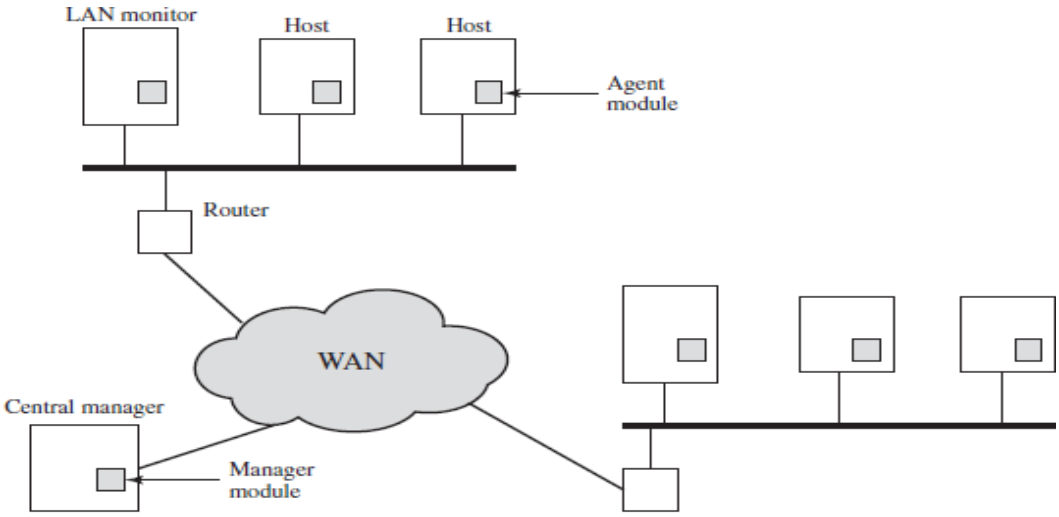


Figure 20.2 Architecture for Distributed Intrusion Detection

A good example of a distributed intrusion detection system, which consists of three main components:

• **Host agent module:** An audit collection module operating as a background process on a monitored system. Its purpose is to collect data on security related events on the host and transmit these to the central manager.
• **LAN monitor agent module:** Operates in the same fashion as a host agent module except that it analyzes LAN traffic and reports the results to the central manager.
• **Central manager module:** Receives reports from LAN monitor and host agents and processes and correlates these reports to detect intrusion.

- Figure 20.3 shows the general approach that is taken. The agent captures each audit record produced by the native audit collection system. A filter is applied that retains only those records that are of security interest.
- At the lowest level, the agent scans for notable events that are of interest independent of any past events.
- At the next higher level, the agent looks for sequences of events, such as known attack patterns (signatures).
- Finally, the agent looks for anomalous behavior of an individual user based on a historical profile of that user, such as number of programs executed, number of files accessed, and the like.

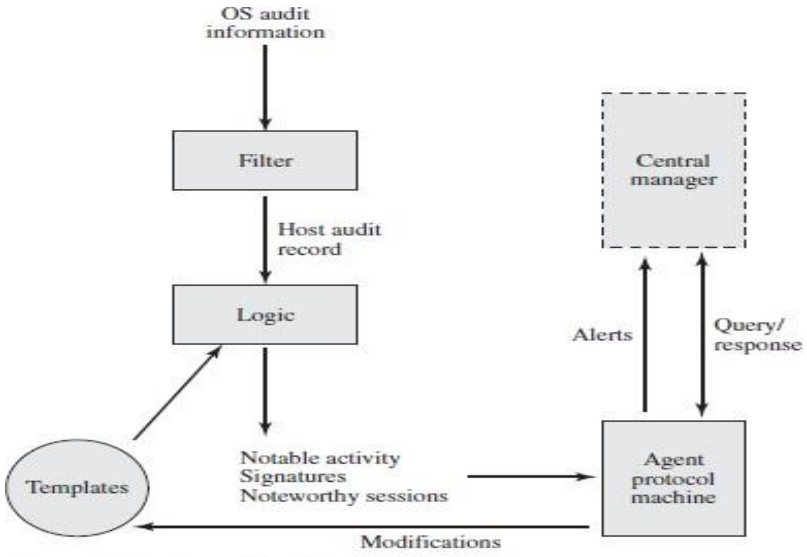


Figure 20.3 Agent Architecture

Honeypots

- A relatively recent innovation in intrusion detection technology is the honeypot. Honeypots are decoy systems that are designed to *lure a potential attacker* away from critical systems.

- Honeypots are designed to
  - divert an attacker from accessing critical systems
  - collect information about the attacker's activity
  - encourage the attacker to stay on the system long enough for administrators to respond

- Initial efforts involved a single honeypot computer with IP addresses designed to attract hackers. More recent research has focused on building entire honeypot networks that emulate an enterprise, possibly with actual or simulated traffic and data. Once hackers are within the network, administrators can observe their behavior in detail and figure out defenses.

Intrusion Detection Exchange Format:

- To facilitate the development of distributed intrusion detection systems that can function across a wide range of platforms and environments, standards are needed to support interoperability. Such standards are the focus of the IETF Intrusion Detection Working Group.

- The purpose of the working group is to define data formats and exchange procedures for sharing information of interest to intrusion detection and response systems and to management systems that may need to interact with them.

  The outputs of this working group include:

  1. A requirements document, which describes the high-level functional requirements

  2. A common intrusion language specification, which describes data formats that satisfy the requirements.

  3. A framework document, which identifies existing protocols best used for communication

5.4.2. MALICIOUS SOFTWARE

Malicious software can be divided into two categories:

- Those that need a host program, and those that are independent.
- The former are essentially fragments of programs that cannot exist independently of some actual application program, utility, or system program.
- Viruses, logic bombs, and backdoors are examples. The latter are self-contained programs that can be scheduled and run by the operating system. Worms and zombie programs are examples.

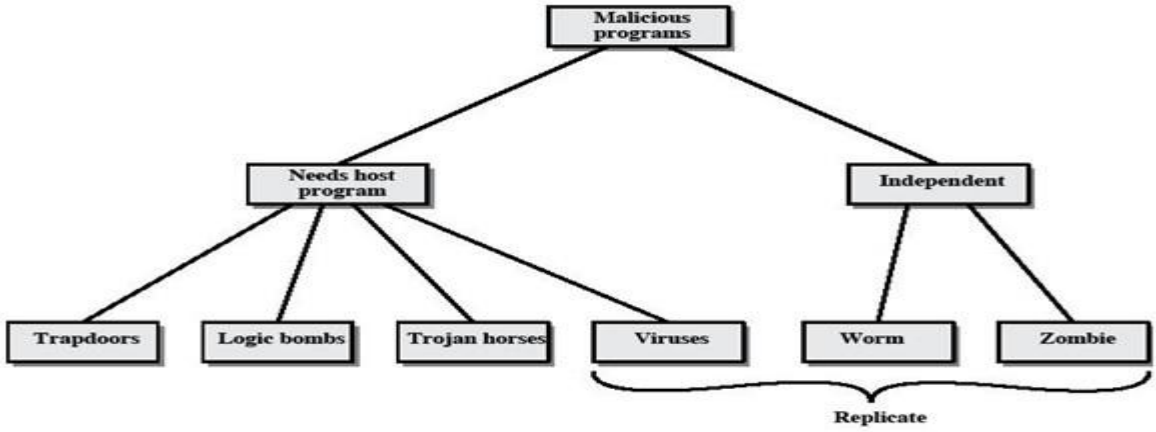


Figure 19.1 Taxonomy of Malicious Programs

5.4.3. VIRUSES

Contents
• **The Nature of Viruses** ✓ **Infection mechanism:** ✓ **Trigger:** ✓ **Payload:** • **Virus Structure** • **Viruses Classification** ✓ **classification by target** ✓ **classification by concealment strategy** • **Virus Kits** • **Macro Viruses** • **E-Mail Viruses**

- Perhaps the most sophisticated types of threats to computer systems are presented by programs that exploit vulnerabilities in computing systems.

Backdoor:

- A backdoor, also known as a **trapdoor,** is a secret entry point into a program that allows someone that is aware of the backdoor to gain access without going through the usual security access procedures.
- Programmers have used backdoors legitimately for many years to debug and test programs.
- The backdoor is code that recognizes some special sequence of input or is triggered by being run from a certain user ID or by an unlikely sequence of events.
- Backdoors become threats when unscrupulous programmers use them to gain unauthorized access.
- The backdoor was the basic idea for the vulnerability portrayed in the movie *War Games*.

Logic Bomb

- One of the oldest types of program threat, predating viruses and worms, is the logic bomb.
- The logic bomb is code embedded in some legitimate program that is set to "explode" when certain conditions are met.
- Examples of conditions that can be used as triggers for a logic bomb are the presence or absence of certain files, a particular day of the week or date, or a particular user running the application.
- Once triggered, a bomb may alter or delete data or entire files, cause a machine halt, or do some other damage.

Trojan Horses

- A Trojan horse is a useful, or apparently useful, program or command procedure containing hidden code that, when invoked, performs some unwanted or harmful function.
- Trojan horse programs can be used to accomplish functions indirectly that an unauthorized user could not accomplish directly.
- For example, to gain access to the files of another user on a shared system, a user could create a Trojan horse program that, when executed,changed the invoking user's file permissions so that the files are readable by any user.
- The author could then induce users to run the program by placing it in a common directory and naming it such that it appears to be a useful utility.
- The code creates a backdoor in the login program that permits the author to log on to the system using a special password.
- This Trojan horse can never be discovered by reading the source code of the login program.

Zombie

- A zombie is a program that secretly takes over another Internet-attached computer and then uses that computer to launch attacks that are difficult to trace to the zombie's creator.
- Zombies are used in denial-of-service attacks, typically against targeted Web sites.

- The zombie is planted on hundreds of computers belonging to unsuspecting third parties, and then used to overwhelm the target Web site by launching an overwhelming onslaught of Internet traffic.

The Nature of Viruses

- A computer virus is a piece of software that can **"infect**" other programs by modifying them; the modification includes injecting the original program with a routine to make copies of the virus program, which can then go on to infect other programs.
- A virus can do anything that other programs do. The difference is that a virus attaches itself to another program and executes secretly when the host program is run.

A computer virus has three parts:

- **Infection mechanism:** The means by which a virus spreads, enabling it to replicate. The mechanism is also referred to as the **infection vector**.
- **Trigger:** The event or condition that determines when the payload is activated or delivered.
- **Payload:** What the virus does, besides spreading. During its lifetime.

A typical virus goes through the following four phases:

- ✓ **Dormant phase:** The virus will eventually be activated by some event, such as a date, the presence of another program or file, or the capacity of the disk exceeding some limit.
- ✓ **Propagation phase:** The virus places a copy of itself into other programs or into certain system areas on the disk.
- ✓ **Triggering phase:** As with the dormant phase, the triggering phase can be caused by a variety of system events, including a count of the number of times that this copy of the virus has made copies of itself.
- ✓ **Execution phase:** The function may be harmless, such as a message on the screen, or damaging, such as the destruction of programs and data files.

Virus Structure A virus can be prepended or postpended to an executable program, or it can be embedded in some other fashion. In this case, the virus code,V, is prepended to infected programs, and it is assumed that the entry point to the program, when invoked, is the first line of the program.

```
     program V :=

{goto main;
     1234567;

     subroutine infect-executable :=
          {loop:
          file := get-random-executable-file;
          if (first-line-of-file = 1234567)
             then goto loop
             else prepend V to file; }

     subroutine do-damage :=
          {whatever damage is to be done}

     subroutine trigger-pulled :=
          {return true if some condition holds}

main:    main-program :=
          {infect-executable;
          if trigger-pulled then do-damage;
          goto next;}
next:

}
```

Figure 21.1 A Simple Virus

When this program is invoked, control passes to its virus, which performs the following steps:

1. For each uninfected file P2 that is found, the virus first compresses that file to produce ,which is shorter than the original program by the size of the virus.
2. A copy of the virus is prepended to the compressed program.
3. The compressed version of the original infected program is uncompressed.
4. The uncompressed original program is executed.

Viruses Classification:

Viruses are classified along two orthogonal axes:

- ✓ The type of target the virus tries to infect and
- ✓ The method the virus uses to conceal itself from detection by users and antivirus software.

A virus **classification by target** includes the following categories:

• **Boot sector infector:** Infects a master boot record or boot record and spreads when a system is booted from the disk containing the virus.

• **File infector:** Infects files that the operating system or shell consider to be executable.

• **Macro virus:** Infects files with macro code that is interpreted by an application.

- A virus classification **by concealment strategy** includes the following categories:

• **Encrypted virus:** A typical approach is as follows. A portion of the virus creates a random encryption key and encrypts the remainder of the virus. The key is stored with the virus.

• **Stealth virus:** A form of virus explicitly designed to hide itself from detection by antivirus software. Thus, the entire virus, not just a payload is hidden.

• **Polymorphic virus:** A virus that mutates with every infection, making detection by the "signature" of the virus impossible.

• **Metamorphic virus:** As with a polymorphic virus, a metamorphic virus mutates with every infection. Metamorphic viruses may change their behavior as well as their appearance.

Virus Kits

- Another weapon in the virus writers' armory is the virus-creation toolkit. Such a toolkit enables a relative novice to quickly create a number of different viruses.
- Although viruses created with toolkits tend to be less sophisticated than viruses designed from scratch, the sheer number of new viruses that can be generated using a toolkit creates a problem for antivirus schemes.
-

Macro Viruses

- In the mid-1990s, macro viruses became by far the most prevalent type of virus. Macro viruses are particularly threatening for a number of reasons:
  1. A macro virus is platform independent.
  2. Macro viruses infect documents, not executable portions of code.
  3. Macro viruses are easily spread. A very common method is by electronic mail.
  4. Because macro viruses infect user documents rather than system programs, traditional file system access controls are of limited use in preventing their spread.

E-Mail Viruses:

- A more recent development in malicious software is the e-mail virus. The first rapidly spreading e-mail viruses, such as Melissa, made use of a Microsoft Word macro embedded in an attachment. If the recipient opens the e-mail attachment, the Word macro is activated. Then
  1. The e-mail virus sends itself to everyone on the mailing list in the user's e-mail package.
  2. The virus does local damage on the user's system.

WORM COUNTERMEASURES OR DIGITAL IMMUNE SYSTEM

Contents
• **Countermeasures**

- **Antivirus Approaches**
  - ✓ **Detection:**
  - ✓ **Identification:**
  - ✓ **Removal:**
- **First generation: simple scanners**
- **Second generation: heuristic scanners**
- **Third generation: activity traps**
- **Fourth generation: full-featured protection**
- **Advanced Antivirus Techniques**
  - **Generic Decryption:**
    - ➢ **CPU emulator:**
    - ➢ **Virus signature scanner:**
    - ➢ **Emulation control module:**
  - **Digital Immune System:**
    - ➢ **Integrated mail systems:**
    - ➢ **Mobile-program systems:**
  - **Behavior-Blocking Software:**

Countermeasures:
Antivirus Approaches:

- The ideal solution to the threat of viruses is prevention:
  - Do not allow a virus to get into the system in the first place, or
  - block the ability of a virus to modify any files containing executable code or macros.

- The next best approach is to be able to do the following:

  • **Detection:** Once the infection has occurred, determine that it has occurred and locate the virus.
  • **Identification:** Once detection has been achieved, identify the specific virus that has infected a program.
  • **Removal:** Once the specific virus has been identified, remove all traces of the virus from the infected program and restore it to its original state. Remove the virus from all infected systems so that the virus cannot spread further.

- The four generations of antivirus software:
  - **• First generation: simple scanners**
  - **• Second generation: heuristic scanners**
  - **• Third generation: activity traps**
  - **• Fourth generation: full-featured protection**

- A **first-generation** scanner requires a virus signature to identify a virus. The virus may contain "wildcards" but has essentially the same structure and bit pattern in all copies.
- A **second-generation** scanner does not rely on a specific signature. Rather, the scanner uses heuristic rules to search for probable virus infection.
- **Third-generation** programs are memory-resident programs that identify a virus by its actions rather than its structure in an infected program.

- **Fourth-generation** products are packages consisting of a variety of antivirus techniques used in conjunction.

Advanced Antivirus Techniques

- More sophisticated antivirus approaches and products continue to appear. In this subsection, we highlight some of the most important.

Generic Decryption:

- Generic decryption (GD) technology enables the antivirus program to easily detect even the most complex polymorphic viruses while maintaining fast scanning speeds. In order to detect such a structure, executable files are run through a GD scanner,

Which contains the following elements?

• **CPU emulator:** A software-based virtual computer. Instructions in an executable file are interpreted by the emulator rather than executed on the underlying processor.

• **Virus signature scanner:** A module that scans the target code looking for known virus signatures.

• **Emulation control module:** Controls the execution of the target code.

Digital Immune System:

- The motivation for this development has been the rising threat of Internet-based virus propagation.
- **The two major trends in Internet technology are**

  • **Integrated mail systems:** Systems such as Lotus Notes and Microsoft Outlook make it very simple to send anything to anyone and to work with objects that are received.

  • **Mobile-program systems:** Capabilities such as Java and ActiveX allow programs to move on their own from one system to another.
- Figure 21.4 illustrates the typical steps in digital immune system operation:

1. A monitoring program on each PC uses a variety of heuristics based on system behavior.

2. The administrative machine encrypts the sample and sends it to a central virus analysis machine.

3. This machine creates an environment in which the infected program can be safely run for analysis.

4. The resulting prescription is sent back to the administrative machine.

5. The administrative machine forwards the prescription to the infected client.

6. The prescription is also forwarded to other clients in the organization.

7. Subscribers around the world receive regular antivirus updates that protect them from the new virus.

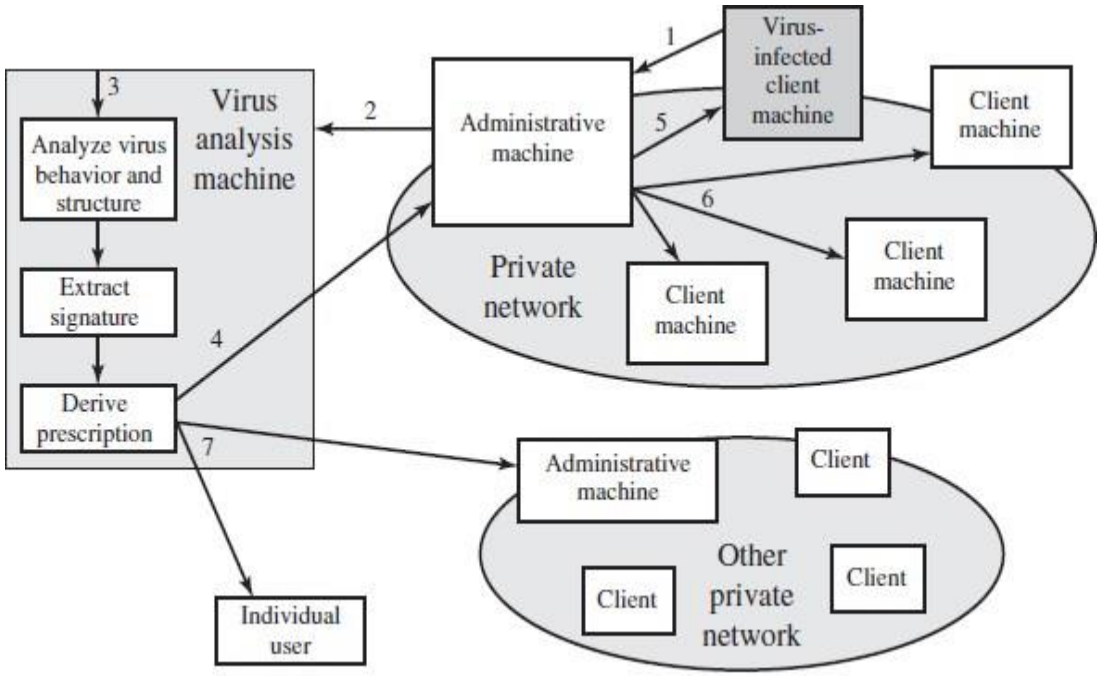


Figure 21.4 Digital Immune System

- ***Behavior-Blocking Software:***

  - *T*he behavior blocking software blocks potentially malicious actions before they have a chance to affect the system.
  - Monitored behaviors can include
    - Attempts to open, view, delete, and/or modify files;
    - Attempts to format disk drives and other unrecoverable disk operations;
    - Modifications to the logic of executable files or macros;
    - Modification of critical system settings, such as start-up settings;
    - Scripting of e-mail and instant messaging clients to send executable content; and
    - Initiation of network communications.

  - Figure 21.5 illustrates the operation of a behavior blocker. Behavior-blocking software runs on server and desktop computers and is instructed through policies set by the network administrator to let benign actions take place but to intercede when unauthorized or suspicious actions occur.

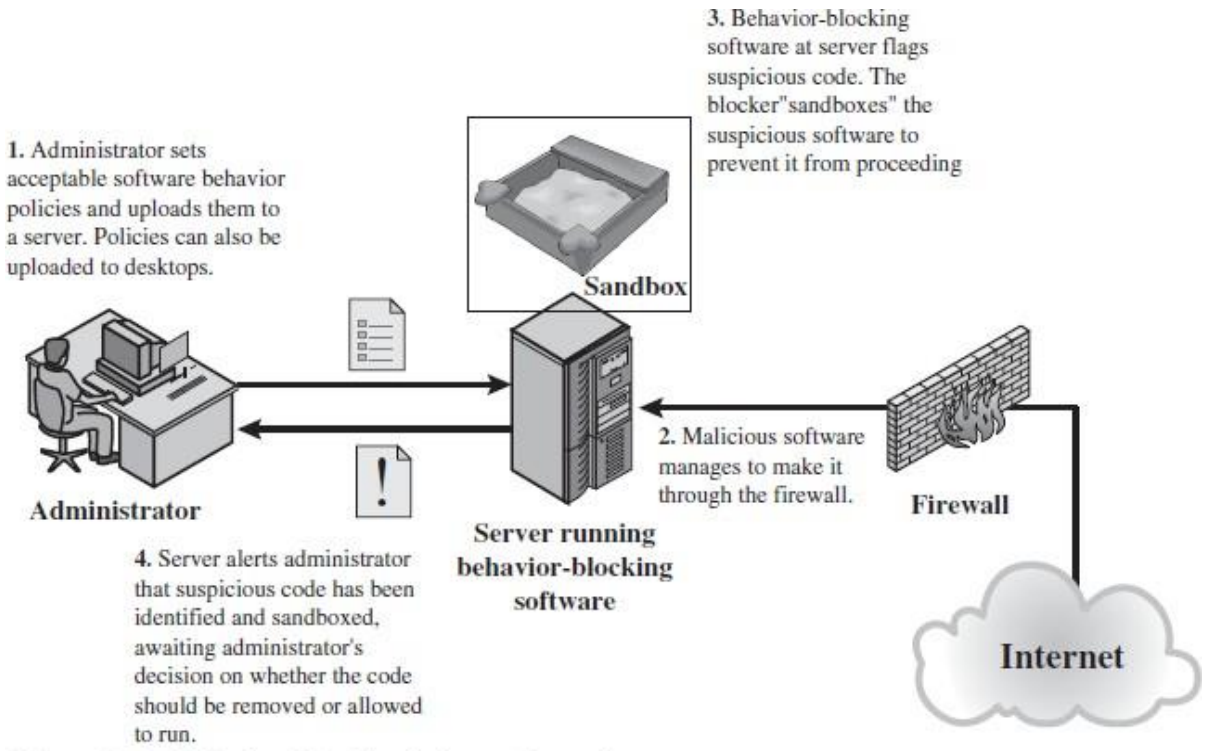


Figure 21.5 Behavior-Blocking Software Operation

5.4.4. FIREWALLS

Internet Firewalls for Trusted Systems

- A firewall is a device or group of devices that controls access between networks.
- A firewall generally consists of filters and gateway(s), varying from firewall to firewall.
- It is a security gateway that controls access between the public Internet and an intranet (a private internal network) and is a secure computer system placed between a trusted network and an untrusted internet.
- Firewalls act as an intermediate server in handling SMTP and HTTP connections in either direction.
- Firewalls can be classified into three main categories: packet filters, circuit-level gateways and application-level gateways

Contents
• Role of Firewalls • Firewall-Related Terminology • Types of Firewalls • Firewall Designs

Screened Host Firewall (Single-homed Bastion Host)

Screened Host Firewall (Dual-homed Bastion Host)

Screened Subnet Firewall

Role of Firewalls:

- The firewall itself must be immune to penetration.
- Firewalls create checkpoints (or choke points) between an internal private network and an untrusted Internet .
- The firewall may filter on the basis of IP source and destination addresses and TCP port number.
- The firewall also enforces logging, and provides alarm capacities as well.
- Firewalls may block TELNET or RLOGIN connections from the Internet to the intranet.
- They also block SMTP and FTP connections to the Internet from internal systems not authorised to send e-mail or to move files.
- The firewall provides protection from various kinds of IP spoofing and routing attacks.

Firewall-Related Terminology:

To design and configure a firewall, some familiarity with the basic terminology is required.
.

- Bastion Host
- Proxy Server
- SOCKS
- Choke Point
- De-militarised Zone (DMZ)
- Logging and Alarms
- VPN

Bastion Host

- A bastion host is a publicly accessible device for the network's security, which has a direct connection to a public network such as the Internet.

- The bastion host serves as a platform for any one of the three types of firewalls: packet filter, circuit-level gateway or application-level gateway.
- Bastion hosts must check all incoming and outgoing traffic and enforce the rules specified in the security policy. They must be prepared for attacks from external and possibly internal sources. They should be built with the least amount of hardware and software in order for a potential hacker to have less opportunity to overcome the firewall.

- **The bastion host's role falls into the following three common types:**

    - ***Single-homed bastion host***: This is a device with only one network interface, normally used for an **application-level gateway**. The external router is configured to send all incoming data to the bastion host, and all internal clients are configured to send all outgoing data to the host.
    - ***Dual-homed bastion host*:** This is a firewall device with at least two network interfaces. Dual-homed bastion hosts serve as **application-level gateways**, and as **packet filters and circuit-level gateways as well**. The advantage of using such hosts is that they create a complete break between the external network and the internal network.
    - ***Multihomed bastion host***: Single-purpose or internal bastion hosts can be classified as either single-homed or multihomed bastion hosts. The latter are used to allow the user to enforce strict security mechanisms.

Proxy Server

- When the security policy requires all inbound and outbound traffic to be sent through a proxy server, a new proxy server should be created for the new streaming application. On the new proxy server, it is necessary to implement strict security mechanisms such as authentication.

SOCKS

- The SOCKS protocol version 4 provides for unsecured firewall traversal for TCP-based client/server applications, including HTTP, TELNET and FTP.
- The new protocol extends the SOCKS version 4 model to include UDP, and allows the framework to include provision for generalized strong authentication schemes, and extends the addressing scheme to encompass domain name and IPv6 addresses.

Choke Point

- The most important aspect of firewall placement is to create choke points. A choke point is the point at which a public internet can access the internal network. The most comprehensive and extensive monitoring tools should be configured on the choke points.
- Proper implementation requires that all traffic be funnelled through these choke points

De-militarised Zone (DMZ)

- The DMZ is an expression that originates from the Korean War. It meant a strip of land forcibly kept clear of enemy soldiers. In terms of a firewall, the DMZ is a network that lies between an internal private network and the external public network.
- DMZ networks are sometimes called perimeter networks. A DMZ is used as an additional buffer to further separate the public network from the internal network.
- A gateway is a machine that provides relay services to compensate for the effects of a filter. The network inhabited by the gateway is often called the DMZ. A gateway in the DMZ is sometimes assisted by an internal gateway.

Logging and Alarms

- Logging is usually implemented at every device in the firewall, but these individual logs combine to become the entire record of user activity. Packet filters normally do not enable logging by default so as not to degrade performance. Packet filters as well as circuit-level gateways log only the most basic information.

VPN

- Some firewalls are now providing VPN services. VPNs are appropriate for any organization requiring secure external access to internal resources. All VPNs are tunneling protocols in the sense that their information packets or payloads are encapsulated or tunneled into the network packets.
- All data transmitted over a VPN is usually encrypted because an opponent with access to the Internet could eavesdrop on the data as it travels over the public network. The VPN encapsulates all the encrypted data within an IP packet. Authentication, message integrity and encryption are very important fundamentals for implementing a VPN.

.

Types of firewall

Contents
• **Types of Firewalls** ✓ **Packet Filtering Firewall** ✓ **Stateful Inspection Firewalls** ✓ **Application-Level Gateway** ✓ **Circuit-Level Gateway**

Types of Firewalls

- A firewall may act as a packet filter. It can operate as a positive filter, allowing to pass only packets that meet specific criteria, or as a negative filter, rejecting any packet that meets certain criteria.

Packet Filtering Firewall

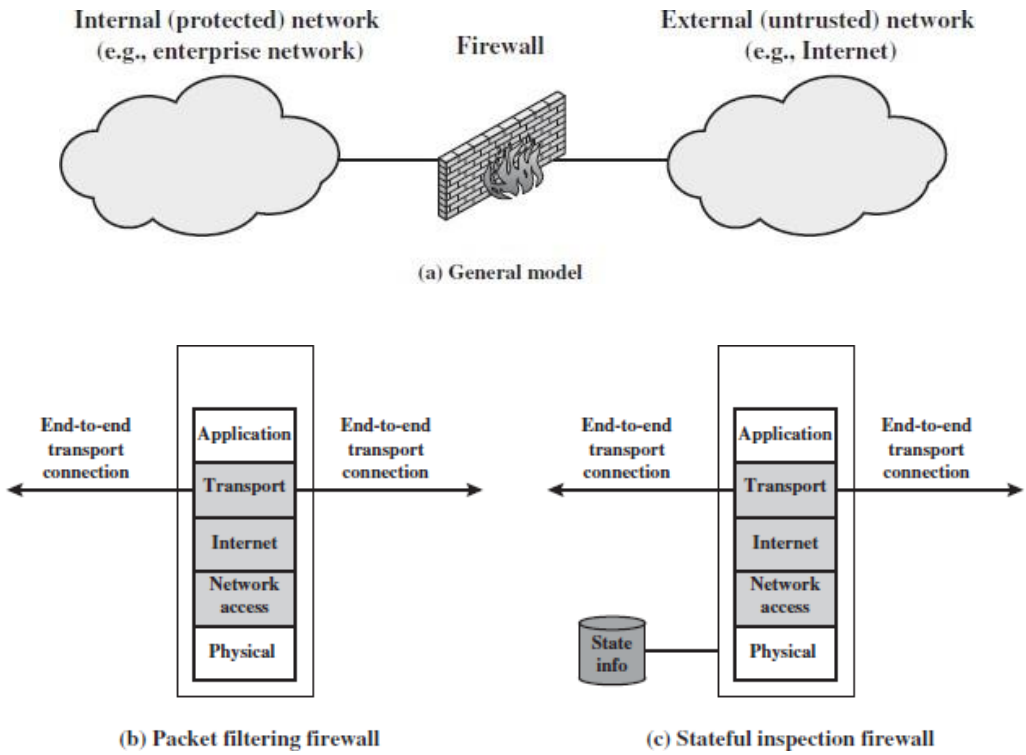
- A packet filtering firewall applies a set of rules to each incoming and outgoing IP packet and then forwards or discards the packet.
- The firewall is typically configured to filter packets going in both directions (from and to the internal network). Filtering rules are based on information contained in a network packet:
    - **Source IP address:** The IP address of the system that originated the IP packet

- **Destination IP address:** The IP address of the system the IP packet is trying to reach
- **Source and destination transport-level address:** The transport-level (e.g., TCP or UDP) port number, which defines applications such as SNMP or TELNET
- **IP protocol field:** Defines the transport protocol
- **Interface:** For a firewall with three or more ports, which interface of the firewall the packet came from or which interface of the firewall the packet is destined for

- The packet filter is typically set up as a list of rules based on matches to fields in the IP or TCP header. **Two default policies are possible:**

  • **Default = discard:** That which is not expressly permitted is prohibited.
  • **Default = forward:** That which is not expressly prohibited is permitted.

- The default discard policy is more conservative. Initially, everything is blocked, and services must be added on a case-by-case basis.
- The default forward policy increases ease of use for end users but provides reduced security; the security administrator must, in essence, react to each new security threat as it becomes known.



A. Inbound mail is allowed (port 25 is for SMTP incoming), but only to a gateway host. However, packets from a particular external host, SPIGOT, are blocked because that host has a history of sending massive files in e-mail messages.

B. This is an explicit statement of the default policy. All rulesets include this rule implicitly as the last rule.

- **One *advantage* of** a packet filtering firewall is its **simplicit**y. Also, packet filters typically are transparent to users and are very fast.

- The following are ***disadvantages*** of packet filter firewalls:

    - Because packet filter firewalls do not examine upper-layer data, they ***cannot prevent attacks*** that employ application-specific vulnerabilities or functions.
    - Because of the ***limited information*** available to the firewall, the logging functionality present in packet filter firewalls is limited.
    - Most packet filter firewalls ***do not support advanced user*** authentication schemes.
    - Packet filter firewalls are generally *vulnerable to attacks*, such as ***network layer address spoofing*.**
    - Finally, due to the small number of variables used in access control decisions, packet filter firewalls **are *susceptible to security breaches*** caused by improper configurations.

- Some of the attacks that can be made on packet filtering firewalls and the appropriate countermeasures are the following:
-
    - **IP address spoofing:** The intruder transmits packets from the outside with a source IP address field containing an address of an internal host.
    - **Source routing attacks:** The source station specifies the route that a packet should take as it crosses the Internet, in the hopes that this will bypass security measures that do not analyze the source routing information.
    - **Tiny fragment attacks:** The intruder uses the IP fragmentation option to create extremely small fragments and force the TCP header information into a separate packet fragment.

Stateful Inspection Firewalls

- A simple packet filtering firewall must permit inbound network traffic on all these high-numbered ports for TCP-based traffic to occur. This creates a vulnerability that can be exploited by unauthorized users.

- A stateful inspection packet firewall tightens up the rules for TCP traffic by creating a directory of outbound TCP connections. There is an entry for each currently established connection. The packet filter will now allow incoming traffic to high-numbered ports only for those packets that fit the profile of one of the entries in this directory.

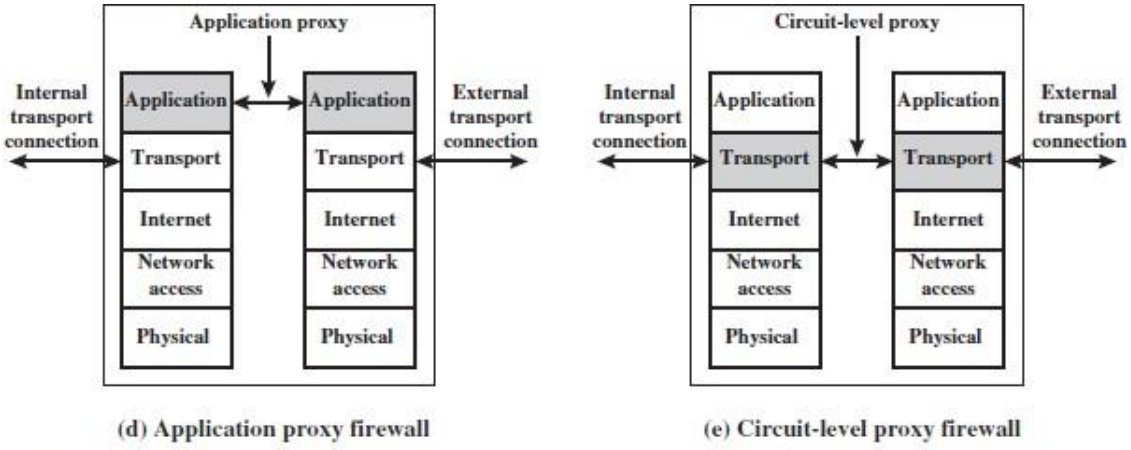
- A stateful packet inspection firewall reviews the same packet information as a packet filtering firewall, but also records information about TCP connections. Some stateful firewalls also keep track of TCP sequence numbers to prevent attacks that depend on the sequence number, such as session hijacking. Some even inspect limited amounts of application data for some well-known protocols like FTP, IM and SIPS commands, in order to identify and track related connections.

Application-Level Gateway:

- An application-level gateway, also called an **application proxy,** acts as a relay of application-level traffic. The user contacts the gateway using a TCP/IP application, such as Telnet or FTP, and the gateway asks the user for the name of the remote host to be accessed.

- When the user responds and provides a valid user ID and authentication information, the gateway contacts the application on the remote host and relays TCP segments containing the application data between the two endpoints. If the gateway does not implement the proxy code for a specific application, the service is not supported and cannot be forwarded across the firewall.

- A prime disadvantage of this type of gateway is the additional processing overhead on each connection. In effect, there are two spliced connections between the end users, with the gateway at the splice point, and the gateway must examine and forward all traffic in both directions.



(d) Application proxy firewall (e) Circuit-level proxy firewall

Figure 22.1 Types of Firewalls

Circuit-Level Gateway:

- A fourth type of firewall is the circuit-level gateway or **circuit-level proxy**. This can be a stand-alone system or it can be a specialized function performed by an application-level gateway for certain applications.

- As with an application gateway, a circuit-level gateway does not permit an end-to-end TCP connection; rather, the gateway sets up two TCP connections, one between itself and a TCP user on an inner host and one between itself and a TCP user on an outside host.

- A typical use of circuit-level gateways is a situation in which the system administrator trusts the internal users. The gateway can be configured to support application- level or proxy service on inbound connections and circuit-level functions for outbound connections. In this configuration, the gateway can incur the processing overhead of examining incoming application data for forbidden functions but does not incur that overhead on outgoing data.

- An example of a circuit-level gateway implementation is the SOCKS package; version 5 of SOCKS is specified in RFC 1928.The RFC defines SOCKS in the following fashion:

  SOCKS consists of the following components:

  - The SOCKS server, which often runs on a UNIX-based firewall. SOCKS is also implemented on Windows systems.
  - The SOCKS client library, which runs on internal hosts protected by the firewall.
  - SOCKS-ified versions of several standard client programs such as FTP and TELNET.

FIREWALL DESIGN OR FIREWALL CONFIGURATION

Contents

• **Firewall Designs**
  ○ Screened Host Firewall (Single-homed Bastion Host)
  ○ Screened Host Firewall (Dual-homed Bastion Host)
  ○ Screened Subnet Firewall

Firewall design

- A security administrator must decide on the location and on the number of firewalls needed. to implement a firewall strategy. The primary step in designing a secure firewall is obviously to prevent the firewall devices from being compromised by threats.
- To provide a certain level of security, the three basic firewall designs are considered: a single-homed bastion host, a dual-homed bastion host and a screened subnet firewall. The first two options are for creating a screened host firewall, and the third option contains an additional packet-filtering router to achieve another level of security.

Screened Host Firewall (Single-homed Bastion Host)

- The first type of firewall is a **screened host** which uses a single-homed bastion host plus a packet-filtering router, as shown in Figure 10.4.
- Single-homed bastion hosts can be configured as either circuit-level or application-level gateways.
- NAT is essentially needed for developing an address scheme internally. It is a critical component of any firewall strategy. It translates the internal IP addresses to IANA

registered addresses to access the Internet. Hence, using NAT allows network administrators to use any internal IP address scheme.

- The screened host firewall is designed such that all incoming and outgoing information is passed through the bastion host.
- The external screening router is configured to route all incoming traffic directly to the bastion host as indicated in Figure 10.4.

- The screening router is also configured to route outgoing traffic only if it originates from the bastion host.

- **A single-homed implementation** may allow a hacker to modify the router not to forward packets to the bastion host. This action would bypass the bastion host and allow the hacker directly into the network.
- But such a bypass usually does not happen because a network using a single-homed bastion host is normally configured to send packets only to the bastion host, and ot directly to the Internet.

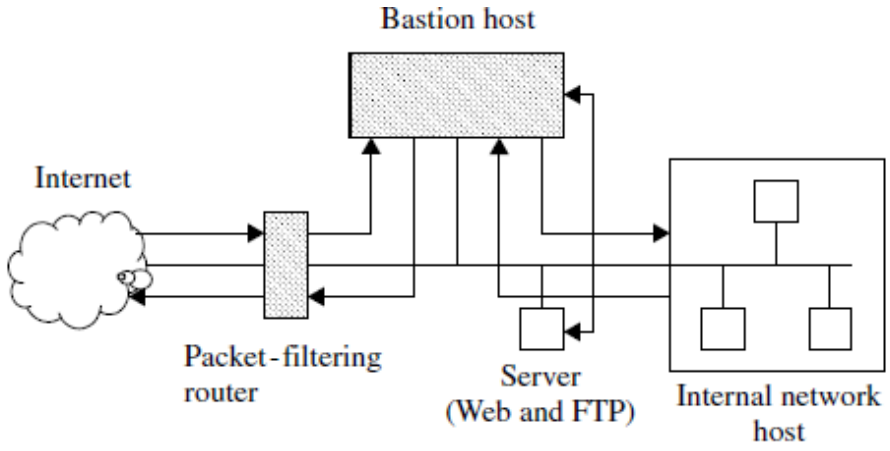


Figure 10.4 Screened host firewall system (single-homed bastion host).

Screened Host Firewall (Dual-homed Bastion Host)

- The configuration of the screened host firewall using a dual-homed bastion host adds significant security, compared with a single-homed bastion host. As shown in Figure 10.5, a dual-homed bastion host has two network interfaces.

- This firewall implementation is secure due to the fact that it creates a complete break between the internal network and the external Internet. As with the single-homed bastion, all external traffic is forwarded directly to the bastion host for processing.

- However, a hacker may try to subvert the bastion host and the router to bypass the firewall mechanisms. Even if a hacker could defeat either the screening router or the dual-homed bastion host, the hacker would still have to penetrate the other. Nevertheless, a dual-homed bastion host removes even this possibility. It is also possible to implement NAT for dual-homed bastion hosts.

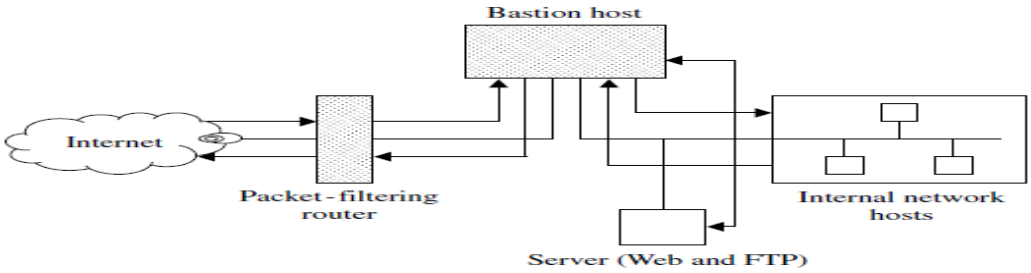


Figure 10.5 Screened host firewall system (dual-homed bastion host).

Screened Subnet Firewall

- The third implementation of a firewall is the screened subnet, which is also known as a DMZ. This firewall is the most secure one among the three implementations, simply because it uses a bastion host to support both circuit- and application-level gateways.
- As shown in Figure 10.6, all publicly accessible devices, including modem and server, are placed inside the DMZ.
- These DMZ then functions as a small isolated network positioned between the Internet and the internal network.
- The screened subnet firewall contains external and internal screening routers. Each is configured such that its traffic flows only to or from the bastion host. This arrangement prevents any traffic from directly traversing the DMZ subnetwork.
- The external screening router uses standard filtering to restrict external access to the bastion host, and rejects any traffic that does not come from the bastion host.
- The benefits of the screened subnet firewall are based on the following facts.
- First, a hacker must subvert three separate tri-homed interfaces when he or she wants to access the internal network. But it is almost infeasible.
- Second, the internal network is effectively invisible to the Internet because all inbound/outbound packets go directly through the DMZ.
- Third, internal users cannot access the Internet without going through the bastion host because the routing information is contained within the network.

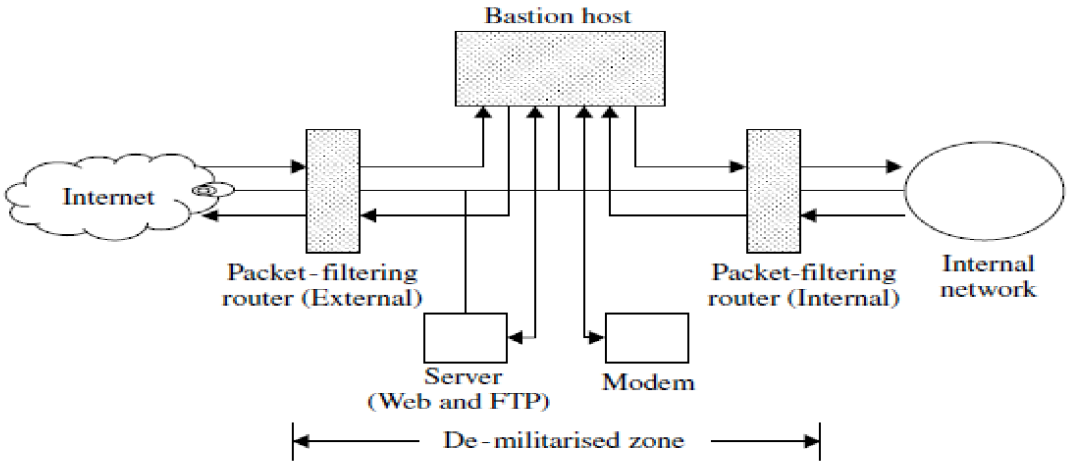


Figure 10.6 Screened subnet firewall system.

UNIT IV
MESSAGE AUTHENTICATION AND INTEGRITY

Authentication requirement – Authentication function – MAC – Hash function – Security of hash function and MAC – SHA –Digital signature and authentication protocols – DSS- Entity Authentication: Biometrics, Passwords, Challenge Response protocols- Authentication applications – Kerberos, X.509

4.1. AUTHENTICATION REQUIREMENT

- In the context of communications across a network, the following attacks can be identified.

  1. Disclosure: Release of message contents to any person or process not possessing the appropriate cryptographic key.

  2. Traffic analysis: Discovery of the pattern of traffic between parties. In a connection-oriented application, the frequency and duration of connections could be determined. In either a connection-oriented or connectionless environment, the number and length of messages between parties could be determined.

  3. Masquerade: Insertion of messages into the network from a fraudulent source. This includes the creation of messages by an opponent that are purported to come from an authorized entity.

  4. Content modification: Changes to the contents of a message, including insertion, deletion, transposition, and modification.

  5. Sequence modification: Any modification to a sequence of messages between parties, including insertion, deletion, and reordering.

  6. Timing modification: Delay or replay of messages. In a connection-oriented application, an entire session or sequence of messages could be a replay of some previous valid session, or individual messages in the sequence could be delayed or replayed.

  7. Source repudiation: Denial of transmission of message by source.

  8. Destination repudiation: Denial of receipt of message by destination.
- **In summary**, message authentication is a procedure to verify that received messages come from the alleged source and have not been altered. Message authentication may also verify sequencing and timeliness.
- A digital signature is an authentication technique that also includes measures to counter repudiation by the source.

4.2. AUTHENTICATION FUNCTION

- Any message authentication or digital signature mechanism has two levels of functionality.
- At the lower level, there must be some sort of function that produces an authenticator: a value to be used to authenticate a message. This lower-level function is then used as a primitive in a higher-level authentication protocol that enables a receiver to verify the authenticity of a message.
- These may be grouped into three classes
  - ***Hash function***
  - ***Message Encryption***

- ***Message Authentication Code***

.

• Hash function: A function that maps a message of any length into a fixedlength hash value, which serves as the authenticator

• Message encryption: The ciphertext of the entire message serves as its authenticator

• Message authentication code (MAC): A function of the message and a secret key that produces a fixed-length value that serves as the authenticator

4.3. MAC - MESSAGE AUTHENTICATION CODE

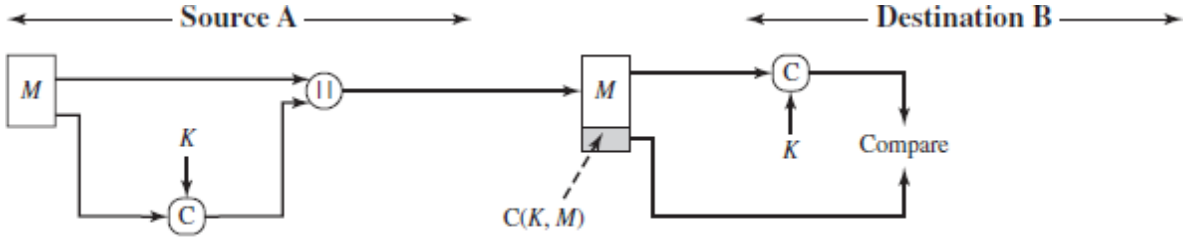
- An alternative authentication technique involves the use of a secret key to generate a small fixed-size block of data, known as a cryptographic checksum or MAC, that is appended to the message. This technique assumes that two communicating parties, say A and B, share a common secret key K.
- When A has a message to send to B, it calculates the MAC as a function of the message and the key:

$$\text{MAC} = C(K, M)$$

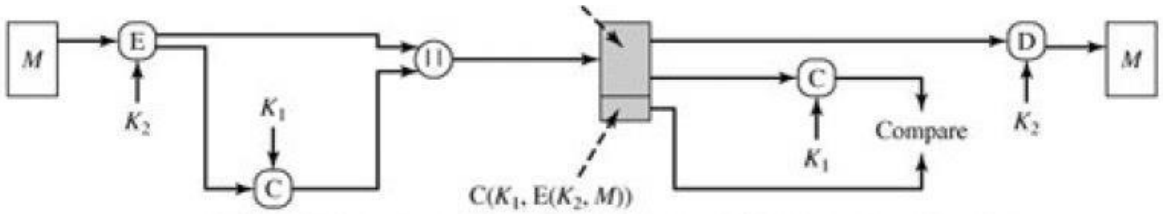
where

M = input message
C = MAC function
K = shared secret key
MAC = message authentication code

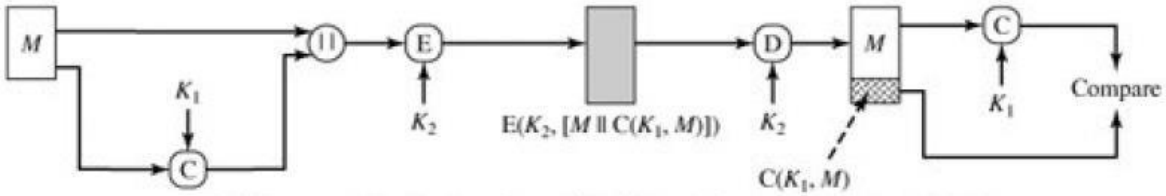
- If we assume that only the receiver and the sender know the identity of the secret key, and if the received MAC matches the calculated MAC, then
  1. The receiver is assured that the message has not been altered.
  2. The receiver is assured that the message is from the alleged sender.
  3. If the message includes a sequence number (such as is used with HDLC, X.25, and TCP), then the receiver can be assured of the proper sequence because an attacker cannot successfully alter the sequence number.



(a) Message authentication



(c) Message authentication and confidentiality; authentication tied to ciphertext



(b) Message authentication and confidentiality; authentication tied to plaintext

$E(K_2, M)$

- The process depicted in Figure a provides authentication but not confidentiality, because the message as a whole is transmitted in the clear.
- Confidentiality can be provided by performing message encryption either after (Figure .b) or before (Figure c) the MAC algorithm. In both these cases, two separate keys are needed, each of which is shared by the sender and the receiver.
- In the first case, the MAC is calculated with the message as input and is then concatenated to the message. The entire block is then encrypted. In the second case, the message is encrypted first.
- Then the MAC is calculated using the resulting ciphertext and is concatenated to the ciphertext to form the transmitted block. Typically, it is preferable to tie the authentication directly to the plaintext, so the method of Figure b is used.

Application of MAC

- Application in message is broadcast to a number of destinations.
- Authentication of a computer program in plain text is an attractive service

4.4. HASH FUNCTION

- A variation on the message authentication code is the one way hash function. As with MAC, a hash function accepts a variable size message M as input and produces affixed-size output, referred to as hash code h=**H(M).**

- Unlike a MAC, a hash code does not use a key but is a function only of the input message. The hash code is also referred to as a **message digest or hash value.**

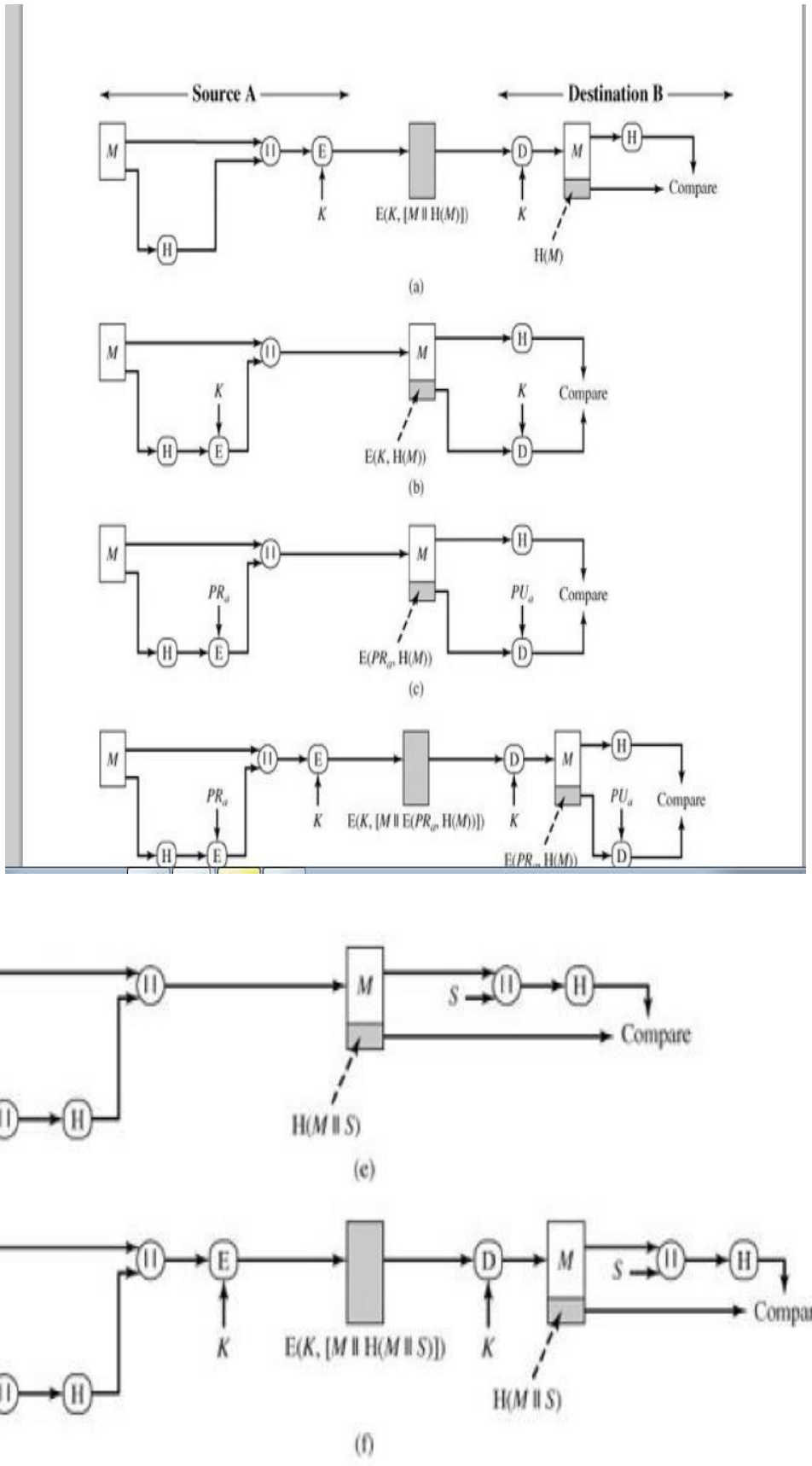


Fig .a. Encrypt message plus hash code

Sender:

- The sender creates a message using SHA to generate a 160 bit hashcode.
- The message and hashcode is concatenated and the result is encrypted using symmetric Encryption Algorithm.

Receiver:

- The receiver uses RSA (or) DSA algorithm to decrypt the message and recover hashcode.
- The receiver generates a new hashcode for the message and compare it with decrypted hashcode
- If two hashcodes are match ,the message is accepted,else it is rejected.

Fig .b. Encrypt hash code shared secret key

Sender:

- The sender creates a message using SHA to generate a 160 bit hashcode.
- Only the hash code is encrypted using Symmetric Encryption.
- The message and hash code encrypted and result is concatented

Receiver:

- The receiver uses RSA (or) DSA algorithm to decrypt the message and recover hashcode.
- The receiver generates a new hashcode for the message and compare it with decrypted hashcode
- If two hashcodes are match ,the message is accepted,else it is rejected.

Fig .c. Encrypt hash code sender's private key

Sender:

- The sender creates a message using SHA to generate a 160 bit hashcode.
- Only the hash code is encrypted using public key encryption and using the sender's private key.
- The message and hash code encrypted and result is concatented

Receiver:

- The receiver uses RSA (or) DSA algorithm to decrypt the message and recover hashcode.
- The receiver generates a new hashcode for the message and compare it with decrypted hashcode
- If two hashcodes are match ,the message is accepted,else it is rejected.

D) Sender:

- The sender creates a message using SHA to generate a 160 bit hashcode.
- Only the hash code is encrypted using public key encryption and using the sender's private key.
- The message and hash code encrypted and result is concatented

Receiver:

- The receiver uses RSA (or) DSA algorithm to decrypt the message and

recover hashcode.

- The receiver generates a new hashcode for the message and compare it with decrypted hashcode which uses the public key Encryption Algorithm of public key of sender.
- If two hashcodes are match ,the message is accepted,else it is rejected.

E) Sender:

- The sender creates a message M using SHA to generate a 160 bit hashcode.\
- This technique uses a hash fuction,but no encryption for message authentication
- This technique assumes that the two communicating parties share a common secret value 'S'.
- The source computes the hash value over the concatenation of M and S and appends the resulting hashvalue to M.

Receiver:

- The receiver uses RSA (or) DSA algorithm to decrypt the message and recover hashcode.
- The receiver generates a new hashcode for the message and compare it with decrypted hashcode which uses the public key Encryption Algorithm of public key of sender.
- The Message concatenated with hash value and 'S' is compared with receiver 's hash value.
- If two hashcodes are match ,the message is accepted,else it is rejected.

f) Confidentiality can be added to the previous approach by encrypting the entire message plus the hash code.

Requirements for a Hash Function

1. H can be applied to a block of data of any size.

2.H produces a fixed-length output.

3. H(x) is relatively easy to compute for any given x, making both hardware and Software implementations practical.

4. For any* given value h, it is computationally infeasible to find x such that H(x) =h. This is sometimes referred to in the literature as the one-way property.

5. For any given block x, it is computationally infeasible to find y x such that H(y) = H(x). This is
sometimes referred to as weak hash function.

6. It is computationally infeasible to find any pair (x, y) such that H(x) = H(y). This is sometimes referred to as **strong collision resistance**.

- ✓ The first three properties are requirements for the practical application of a hash function to message authentication.
- ✓ The fourth property, the one-way property, states that it is easy to generate a code given a message but virtually impossible to generate a message given a code.
- ✓ The fifth property guarantees that an alternative message hashing to the same value as a given message cannot be found.
- ✓ This prevents forgery when an encrypted hash code is used (Figures b and c).
- ✓ The sixth property refers to how resistant the hash function is to a type of attack known as the birthday attack, which we examine shortly.

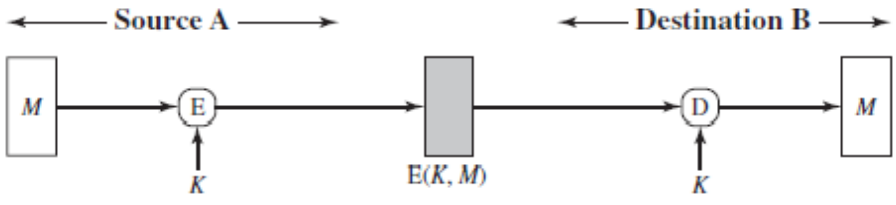
Message Encryption

- Message encryption by itself can provide a measure of authentication.
- The analysis differs for
  - **symmetric and**
  - **public-key encryption schemes**.

Symmetric Encryption

- Consider the straightforward use of symmetric encryption (Figure 12.1a). A message M transmitted from source A to destination B is encrypted using a secret key K shared by A and B. If no other party knows the key, then confidentiality is provided: No other party can recover the plaintext of the message.
- In addition, B is assured that the message was generated by A. Why? The message must have come from A, because A is the only other party that possesses K and therefore the only other party with the information necessary to construct ciphertext that can be decrypted with K.
- Furthermore, if M is recovered, B knows that none of the bits of M have been altered, because an opponent that does not know K would not know how to alter bits in the ciphertext to produce the desired changes in the plaintext.

Figure 12.1 Basic Uses of Message Encryption



(a) Symmetric encryption: confidentiality and authentication

- Thus, in general, we require that only a small subset of all possible bit patterns be considered legitimate plaintext. In that case, any spurious ciphertext is unlikely to produce legitimate plaintext.

- For example, suppose that only one bit pattern in 106 is legitimate plaintext. Then the probability that any randomly chosen bit pattern, treated as ciphertext, will produce a legitimate plaintext message is only 10-6. For a number of applications and encryption schemes, the desired conditions prevail as a matter of course.
- For example, suppose that we are transmitting Englishlanguage messages using a Caesar cipher with a shift of one (K = 1). A sends the following legitimate ciphertext:
    - `nbsftfbupbutboeepftfbupbutboemjuumfmbnctfbujwz`

  B decrypts to produce the following plaintext:

  `mareseatoatsanddoeseatoatsandlittlelambseativy`

- A simple frequency analysis confirms that this message has the profile of ordinary English. On the other hand, if an opponent generates the following random sequence of letters:

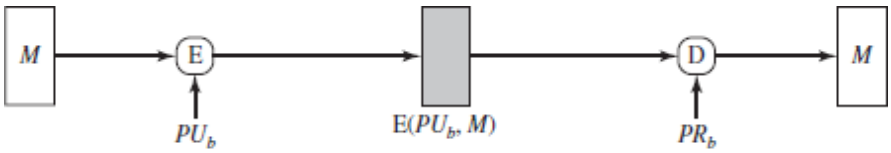
  `zuvrsoevgqxlzwigamdvnmhpmccxiuureosfbcebtqxsxq`

  this decrypts to

  `ytuqrndufpwkyvhfzlcumlgolbbwhttqdnreabdaspwrwp`

Public-Key Encryption

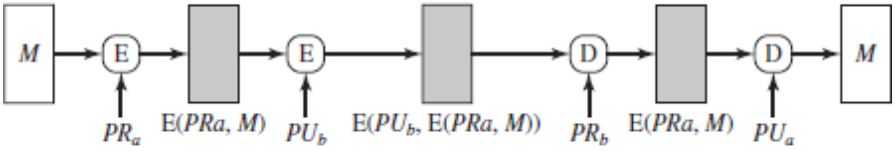
- The straightforward use of public-key encryption (Figure 12.1b) provides confidentiality but not authentication. The source (A) uses the public key PU_b of the destination (B) to encrypt M. Because only B has the corresponding private key PR_b, only B can decrypt the message. This scheme provides no authentication, because any opponent could also use B's public key to encrypt a message and claim to be A.



(b) Public-key encryption: confidentiality

- To provide authentication, A uses its private key to encrypt the message, and B uses A's public key to decrypt (Figure 12.1c). This provides authentication using the same type of reasoning as in the symmetric encryption case: The message must have come from A because A is the only party that possesses PR_a and therefore the only party with the information necessary to construct ciphertext that can be decrypted with PU_a.
- Again, the same reasoning as before applies: There must be some internal structure to the plaintext so that the receiver can distinguish between well-formed plaintext and random bits.

(c) Public-key encryption: authentication and signature



(d) Public-key encryption: confidentiality, authentication, and signature

4.5. SECURITY OF HASH FUNCTION AND MAC

Contents

• **Security of hash function and MAC**
• **Brute-Force Attacks**
  ○ ***Hash functions***
  ○ ***Message Authentication Code***.
• **Cryptanalysis**

Security of hash function and MAC

- We can group attacks on MACs into two categories: brute-force attacks and cryptanalysis.

Brute-Force Attacks

- A brute-force attack on a MAC is a more difficult undertaking than a brute-force attack on a hash function because it requires known message-tag pairs. Let us see why this is so. To attack a hash code, we can proceed in the following way.

Hash functions

- The strength of a hash function against brute-force attacks depends solely on the length of the hash code produced by the algorithm. Recall from our discussion of hash **functions that there are three desirable properties**:
  - **One-way**: For any given code *h*, it is computationally infeasible to find *x* such that H(*x*) = *h*.
  - **Weak collision resistance**: For any given block *x*, it is computationally infeasible to find *y x* with H(*y*) = H(*x*).
  - **Strong collision resistance**: It is computationally infeasible to find any pair (*x*, *y*) such that H (*x*) = H(*y*).
- For a hash code of length *n*, the level of effort required, as we have seen is proportional to the following

One way	2^n
Weak collision resistance	2^n
Strong collision resistance	$2^n/2$

Message Authentication Codes

A brute-force attack on a MAC is a more difficult undertaking because it requires known message-MAC pairs. Let us see why this is so.

To attack a hash code, we can proceed **in the following way**.

- Given a fixed message x with n-bit hash code $h = H(x)$, a brute-force method of finding a collision is to pick a random bit string y and check if $H(y) = H(x)$.
- The attacker can do this repeatedly off line. Whether an off-line attack can be used on a MAC algorithm depends on the relative size of the key and the MAC.
- To proceed, we need to state the desired security property of a MAC algorithm, which can be expressed as follows:

Computation resistance:

- Given one or more text-MAC pairs [xi, $C(K, xi)$], it is computationally infeasible to compute any text-MAC pair [x, $C(K, x)$] for any new input $x \neq xi$.
- The attacker would like to come up with the valid MAC code for a given message x.
- There are two lines of attack possible: Attack the key space and attack the MAC value
- If an attacker can determine the MAC key, then it is possible to generate a valid MAC value for any input x.
- Suppose the key size is k bits and that the attacker has one known text-MAC pair. Then the attacker can compute the n-bit MAC on the known text for all possible keys. At least one key is guaranteed to produce the correct MAC, namely, the valid key that was initially used to produce the known text-MAC pair. This phase of the attack takes a level of effort proportional to 2^k (that is, one operation for each of the 2^k possible key values).
- It can be shown that the level of effort drops off rapidly with each additional text-MAC pair and that the overall level of effort is roughly 2^k
- To summarize, the level of effort for brute-force attack on a MAC algorithm can be expressed as $\min(2^k, 2^n)$

Cryptanalysis

- The way to measure the resistance of a MAC algorithm to cryptanalysis is to compare its strength to the effort required for a bruteforce attack. That is, an ideal MAC algorithm will require a cryptanalytic effort greater than or equal to the brute-force effort.
- There is much more variety in the structure of MACs than in hash functions, so it is difficult to generalize about the cryptanalysis of MACs. Furthermore, far less work has been done on developing such attacks.

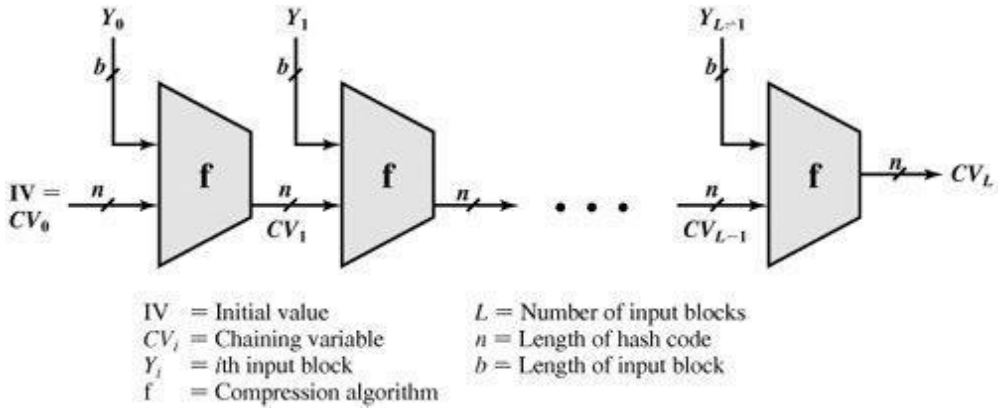


Figure 11.9. General Structure of Secure Hash Code

- The hash algorithm involves repeated use of a **compression function**, f, that takes two inputs (an *n*-bit input from the previous step, called the ***chaining variable***, and a *b*-bit block) and produces an *n*-bit output.
- At the start of hashing, the chaining variable has an initial value that is specified as part of the algorithm. The final value of the chaining variable is the hash value. Often, *b* > *n*; hence the term
- ***Compression***. The hash function can be summarized as follows:

$$CV_o = IV = \text{initial } n\text{-bit value}$$

$$CV_i = f(CV_{i1}, Y_{i1})\ 1 \leq i \leq L$$

$$H(M) = CV_L$$

- Where the input to the hash function is a message *M* consisting of the blocks *Y*o, *Y*1,..., *YL*1.

4.6. SHA

Contents
• **Secure Hash Algorithm(SHA)** ○ **SHA-512 Logic** ○ **SHA-512 Round Function**

Secure Hash Algorithm (SHA)

- SHA was developed by the Nation-bits foral Institute of Standards and Technology (NIST) and published as a federal information processing standard (FIPS 180) in 1993. When weaknesses were discovered in SHA, now known as **SHA-0**, a revised version was issued as FIPS 180-1 in 1995 and is referred to as **SHA-1**.

Table 11.3 Comparison of SHA Parameters

	SHA-1	SHA-224	SHA-256	SHA-384	SHA-512
Message Digest Size	160	224	256	384	512
Message Size	$< 2^{64}$	$< 2^{64}$	$< 2^{64}$	$< 2^{128}$	$< 2^{128}$
Block Size	512	512	512	1024	1024
Word Size	32	32	32	64	64
Number of Steps	80	64	64	80	80

- The actual standards document is entitled "Secure Hash Standard." SHA is based on the hash function MD4, and its design closely models MD4. SHA-1 produces a hash value of 160 bits.
- NIST produced a revised version of the standard, FIPS 180-2, that defined three new versions of SHA, with hash value lengths of 256, 384, and 512 bits, known as SHA-256, SHA-384, and SHA-512, respectively.
- Collectively, these hash algorithms are known as **SHA-2**. These new versions have the same underlying structure and use the same types of modular arithmetic and logical binary operations as SHA-1. A revised document was issued as FIP PUB 180-3 in 2008, which added a 224-bit version (Table 11.3).

SHA-512 Logic

- The algorithm takes as input a message with a maximum length of less than 2128 bits and produces as output a 512-bit message digest. The input is processed in 1024-bit blocks.
- Figure 11.9 depicts the overall processing of a message to produce a digest. This follows the general structure depicted in Figure 11.8. The processing consists of the following steps.

  - **Step 1 Append padding bits.** The message is padded so that its length is congruent to 896 modulo 1024 [length K 896(mod 1024)]. Padding is always added, even if the message is already of the desired length. Thus, the number of padding bits is in the range of 1 to 1024. The padding consists of a single 1 bit followed by the necessary number of 0 bits.
  - **Step 2 Append length.** A block of 128 bits is appended to the message. This block is treated as an unsigned 128-bit integer (most significant byte first) and contains the length of the original message (before the padding).

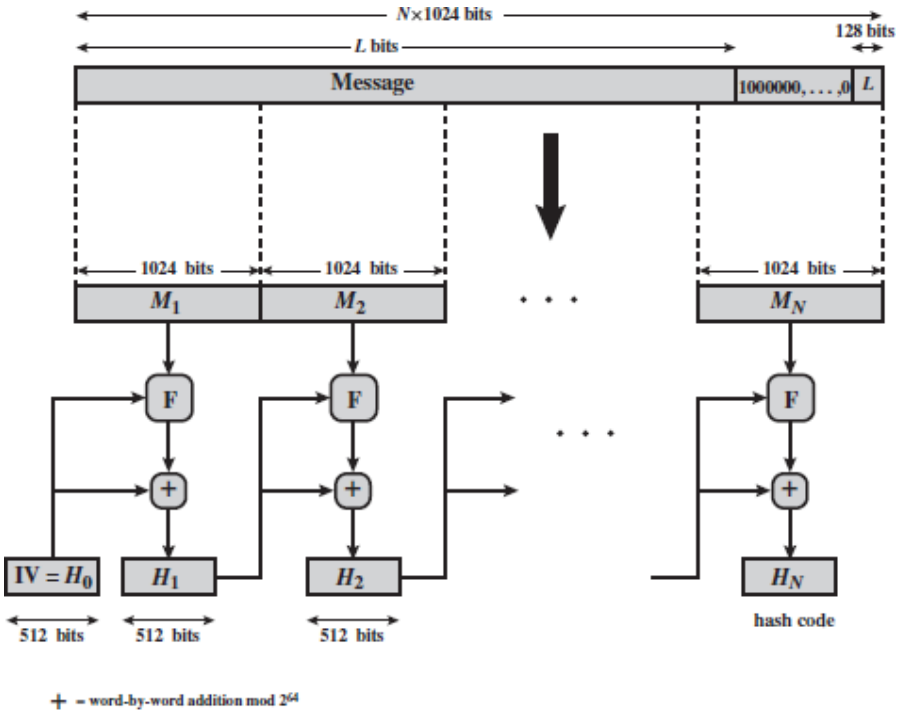


Figure 11.9 Message Digest Generation Using SHA-512

- The outcome of the first two steps yields a message that is an integer multiple of 1024 bits in length. In Figure 11.9, the expanded message is represented as the sequence of 1024-bit blocks *M*1, *M*2, c, *MN*, so that the total length of the expanded message is *N* * 1024 bits.

❖ **Step 3 Initialize hash buffer.** A 512-bit buffer is used to hold intermediate and final results of the hash function. The buffer can be represented as eight 64-bit registers (a, b, c, d, e, f, g, h). These registers are initialized to the following 64-bit integers (hexadecimal values):

```
a = 6A09E667F3BCC908    e = 510E527FADE682D1
b = BB67AE8584CAA73B    f = 9B05688C2B3E6C1F
c = 3C6EF372FE94F82B    g = 1F83D9ABFB41BD6B
d = A54FF53A5F1D36F1    h = 5BE0CD19137E2179
```

- These values are stored in **big-endian** format, which is the most significant byte of a word in the low-address (leftmost) byte position. These words were obtained by taking the first sixty-four bits of the fractional parts of the square roots of the first eight prime numbers.

❖ **Step 4 Process message in 1024-bit (128-word) blocks.** The heart of the algorithm is a module that consists of 80 rounds; this module is labeled F in Figure 11.9. The logic is illustrated in Figure 11.10.

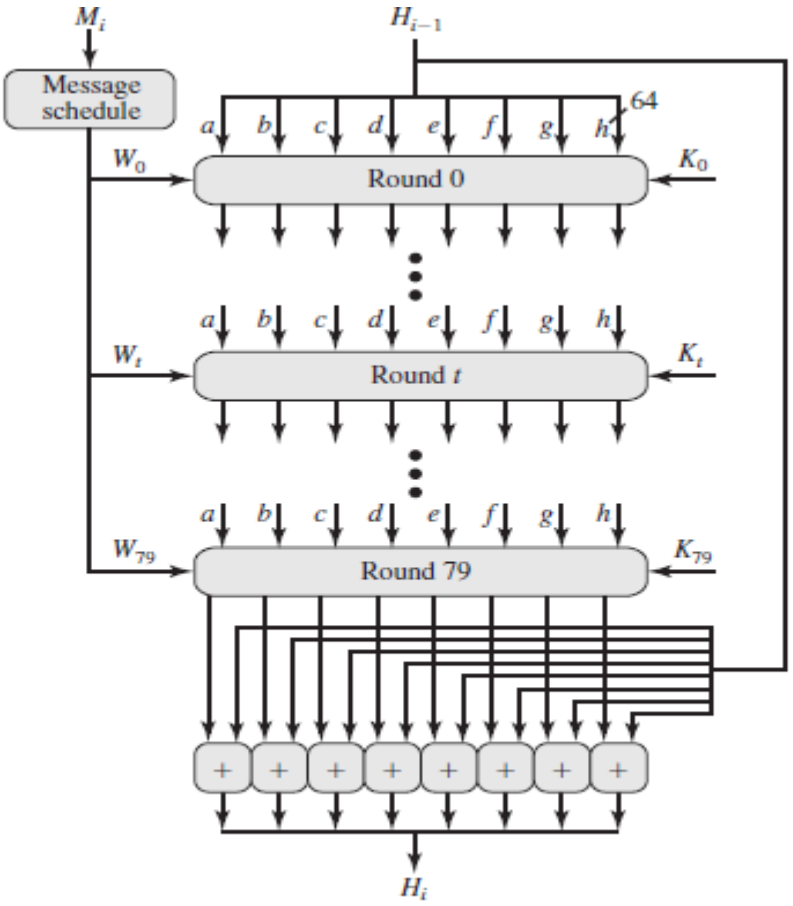


Figure 11.10 SHA-512 Processing of a Single 1024-Bit Block

- Each round takes as input the 512-bit buffer value, abcdefgh, and updates the contents of the buffer. At input to the first round, the buffer has the value of the intermediate hash value, *Hi*-1.
- Each round *t* makes use of a 64-bit value *Wt*, derived from the current 1024-bit block being processed (*Mi*). These values are derived using a message schedule described subsequently.Each round also makes use of an additive constant *Kt*, where 0 … *t* … 79 indicates one of the 80 rounds.

❖ **Step 5 Output.** After all *N* 1024-bit blocks have been processed, the output from the *N*th stage is the 512-bit message digest. We can summarize the behavior of SHA-512 as follows:

$$H_0 = \text{IV}$$
$$H_i = \text{SUM}_{64}(H_{i-1}, \text{abcdefgh}_i)$$
$$MD = H_N$$

where

IV ----------→ initial value of the abcdefgh buffer, defined in step 3

abcdefgh_i --→ the output of the last round of processing of the *i*th message block

N ----------→ the number of blocks in the message (including padding and length fields)

SUM64-----→addition modulo 264

MD ---------→final message digest value

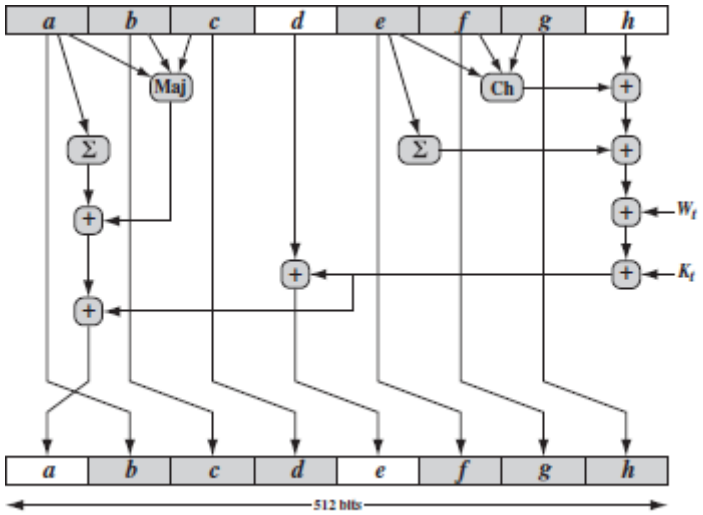


Figure 11.11 Elementary SHA-512 Operation (single round)

SHA-512 Round Function

Let us look in more detail at the logic in each of the 80 steps of the processing of one 512-bit block (Figure 11.11). Each round is defined by the following set of equations:

$$T_1 = h + \text{Ch}(e, f, g) + \left(\sum_1^{512} e\right) + W_t + K_t$$
$$T_2 = \left(\sum_0^{512} a\right) + \text{Maj}(a, b, c)$$
$$h = g$$
$$g = f$$
$$f = e$$
$$e = d + T_1$$
$$d = c$$
$$c = b$$
$$b = a$$
$$a = T_1 + T_2$$

Where

t = step number; $0 \le t \le 79$

$\text{Ch}(e, f, g)$ = $(e \text{ AND } f) \oplus (\text{NOT } e \text{ AND } g)$
the conditional function: If e then f else g

$\text{Maj}(a, b, c)$ = $(a \text{ AND } b) \oplus (a \text{ AND } c) \oplus (b \text{ AND } c)$
the function is true only of the majority (two or three) of the arguments are true

$\left(\sum_0^{512} a\right)$ = $\text{ROTR}^{28}(a) \oplus \text{ROTR}^{34}(a) \oplus \text{ROTR}^{39}(a)$

$\left(\sum_1^{512} e\right)$ = $\text{ROTR}^{14}(e) \oplus \text{ROTR}^{18}(e) \oplus \text{ROTR}^{41}(e)$

$\text{ROTR}^n(x)$ = circular right shift (rotation) of the 64-bit argument x by n bits

W_t = a 64-bit word derived from the current 1024-bit input block

K_t = a 64-bit additive constant

$+$ = addition modulo 2^{64}

Two observations can be made about the round function.

1. Six of the eight words of the output of the round function involve simply permutation (b, c, d, f, g, h) by means of rotation. This is indicated by shading in Figure 11.11.

2. Only two of the output words (*a*, *e*) are generated by substitution. Word *e* is a function of input variables (*d*, *e*, *f* , *g*, *h*), as well as the round word *Wt* and the constant *Kt*. Word *a* is a function of all of the input variables except *d*, as well as the round word *Wt* and the constant *Kt*. It remains to indicate how the 64-bit word values *Wt* are derived from the 1024-bit message. The remaining values are defined as

$$W_t = \sigma_1^{512}(W_{t-2}) + W_{t-7} + \sigma_0^{512}(W_{t-15}) + W_{t-16}$$

where

$$\sigma_0^{512}(x) = \text{ROTR}^1(x) \oplus \text{ROTR}^8(x) \oplus \text{SHR}^7(x)$$
$$\sigma_1^{512}(x) = \text{ROTR}^{19}(x) \oplus \text{ROTR}^{61}(x) \oplus \text{SHR}^6(x)$$
$$\text{ROTR}^n(x) = \text{circular right shift (rotation) of the 64-bit argument } x \text{ by } n \text{ bits}$$
$$\text{SHR}^n(x) = \text{left shift of the 64-bit argument } x \text{ by } n \text{ bits with padding by zeros on the right}$$
$$+ = \text{addition modulo } 2^{64}$$

- Thus, in the first 16 steps of processing, the value of *Wt* is equal to the corresponding word in the message block. For the remaining 64 steps, the value of *Wt* consists of the circular left shift by one bit of the XOR of four of the preceding values of *Wt*, with two of those values subjected to shift and rotate operations.
- This introduces a great deal of redundancy and interdependence into the message blocks that are compressed, which complicates the task of finding a different message block that maps to the same compression function output

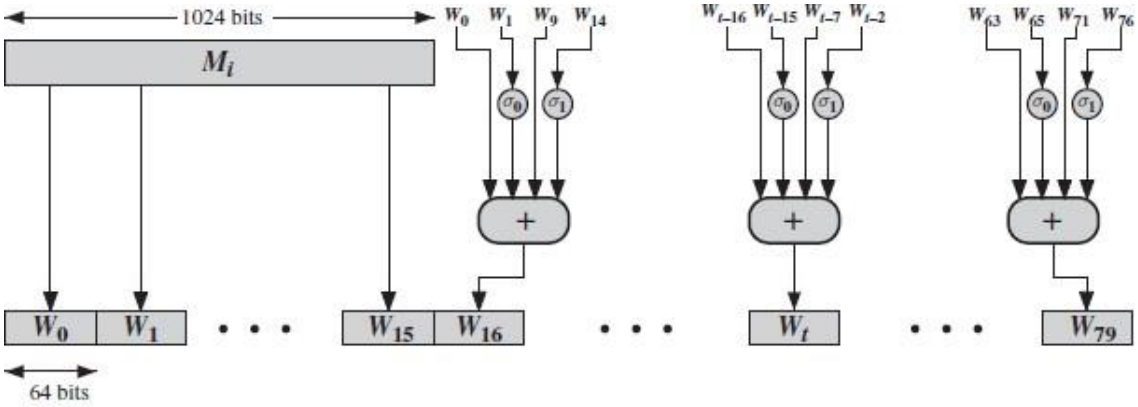


Figure 11.12 Creation of 80-word Input Sequence for SHA-512 Processing of Single Block

4.7. DIGITAL SIGNATURE

Contents
• **Digital Signatures** ○ Properties ○ Attacks and Forgeries ○ Digital Signature Requirements ○ Direct Digital Signature

Digital Signatures

- A digital signature is an authentication mechanism that enables the creator of a message to attach a code that acts as a signature. The signature is formed by taking the hash of the message and encrypting the message with the creators private key

Properties

- Message authentication protects two parties who exchange messages from any third party. However, it does not protect the two parties against each other. Several forms of dispute between the two are possible.
- For example, suppose that John sends an authenticated message to Mary, using one of the schemes of Figure 12.1. Consider the following disputes that could arise.

  1. Mary may forge a different message and claim that it came from John. Mary would simply have to create a message and append an authentication code using the key that John and Mary share.

  2. John can deny sending the message. Because it is possible for Mary to forge a message, there is no way to prove that John did in fact send the message.

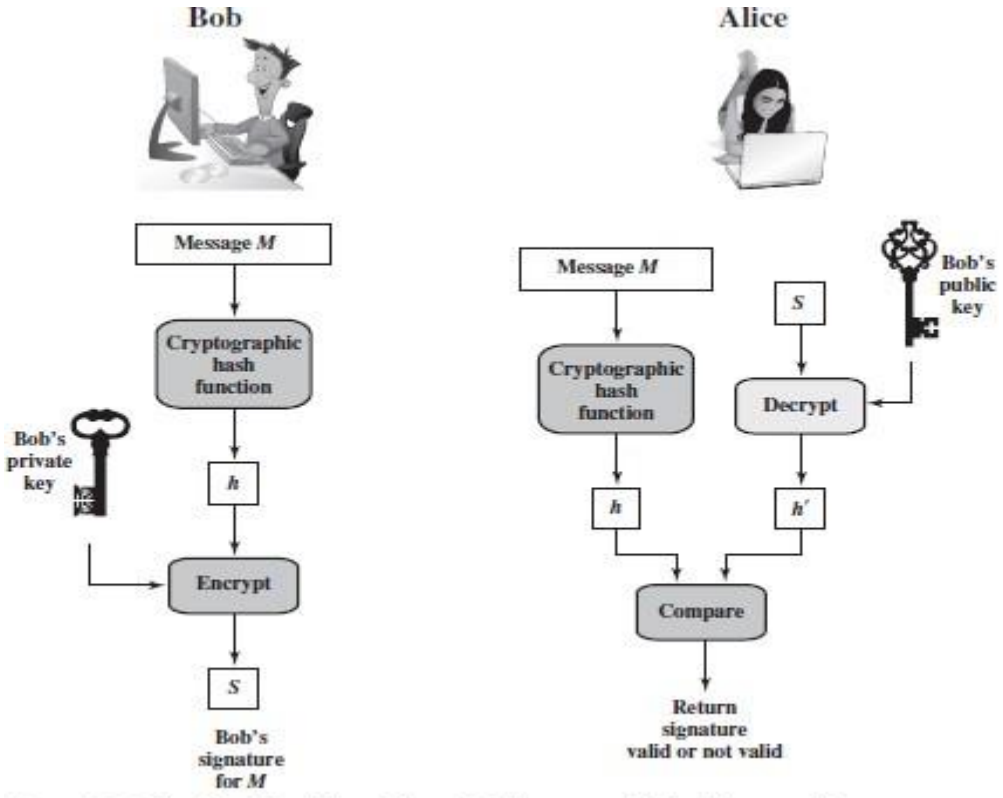


Figure 13.2 Simplified Depiction of Essential Elements of Digital Signature Process

- Both scenarios are of legitimate concern. Here is an example of the first scenario: An electronic funds transfer takes place, and the receiver increases the amount of funds transferred and claims that the larger amount had arrived from the sender.
- An example of the second scenario is that an electronic mail message contains instructions to a stockbroker for a transaction that subsequently turns out badly. The sender pretends that the message was never sent.
- In situations where there is not complete trust between sender and receiver, something more than authentication is needed. The most attractive solution to this problem is the digital signature.
- The digital signature must have the following properties:
  - It must verify the author and the date and time of the signature.
  - It must authenticate the contents at the time of the signature.
  - It must be verifiable by third parties, to resolve disputes.

Thus, the digital signature function includes the authentication function.

Attacks and Forgeries

- [GOLD88] lists the following types of attacks, in order of increasing severity. Here A denotes the user whose signature method is being attacked, and C denotes the attacker.
  • **Key-only attack:** C only knows A's public key.
  • **Known message attack:** C is given access to a set of messages and their signatures.
  • **Generic chosen message attack:** C chooses a list of messages before attempting to breaks A's signature scheme, independent of A's public key. C then obtains from A valid signatures for the chosen messages.
  • **Directed chosen message attack:** Similar to the generic attack, except that the list of messages to be signed is chosen after C knows A's public key but before any signatures are seen.
  • **Adaptive chosen message attack:** C is allowed to use A as an "oracle." This means that C may request from A signatures of messages that depend on previously obtained message-signature pairs.

- [GOLD88] then defines success at breaking a signature scheme as an outcome in which C can do any of the following with a non-negligible probability:
  • **Total break:** C determines A's private key.
  • **Universal forgery:** C finds an efficient signing algorithm that provides an equivalent way of constructing signatures on arbitrary messages.
  • **Selective forgery:** C forges a signature for a particular message chosen by C.
  • **Existential forgery:** C forges a signature for at least one message. C has no control over the message. Consequently, this forgery may only be a minor nuisance to A.

Digital Signature Requirements

- On the basis of the properties and attacks just discussed, we can formulate the following requirements for a digital signature.
  • The signature must be a bit pattern that depends on the message being signed.
  • The signature must use some information unique to the sender to prevent both forgery and denial.
  • It must be relatively easy to produce the digital signature.
  • It must be relatively easy to recognize and verify the digital signature.
  • It must be computationally infeasible to forge a digital signature, either by constructing a new message for an existing digital signature or by constructing a fraudulent digital signature for a given message.
  • It must be practical to retain a copy of the digital signature in storage.
- **Two general schemes for digital signatures**
  - Direct Digital Signature
  - Arbitrated Digital Signature

Direct Digital Signature

- The term **direct digital signature** refers to a digital signature scheme that involves only the communicating parties (source, destination). It is assumed that the destination knows the public key of the source.
- Confidentiality can be provided by encrypting the entire message plus signature with a shared secret key (symmetric encryption). Note that it is important to perform the signature function first and then an outer confidentiality function.
- In case of dispute, some third party must view the message and its signature. If the signature is calculated on an encrypted message, then the third party also needs access to the decryption key to read the original message. However, if the signature is the inner

operation, then the recipient can store the plaintext message and its signature for later use in dispute resolution.

- The validity of the scheme just described depends on the security of the sender's private key. If a sender later wishes to deny sending a particular message, the sender can claim that the private key was lost or stolen and that someone else forged his or her signature.
- Administrative controls relating to the security of private keys can be employed to thwart or at least weaken this ploy, but the threat is still there, at least to some degree. One example is to require every signed message to include a **timestamp** (date and time) and to require prompt reporting of compromised keys to a central authority.
- Another threat is that some private key might actually be stolen from X at time T. The opponent can then send a message signed with X's signature and stamped with a time before or equal to T.

Arbitrated Digital Signature

- The problems associated with direct digital signatures can be addressed by using an arbiter.
- As with direct signature schemes, there is a variety of arbitrated signature schemes.

In general terms, they all operate as follows.

- Every signed message from a sender X to a receiver Y goes first to an arbiter A, who subjects the message and its signature to a number of tests to check its origin and content.
- The message is then dated and sent to Y with an indication that it has been verified to the satisfaction of the arbiter. The presence of A solves the problem faced by direct signature schemes: that X might disown the message.
- The arbiter plays a sensitive and crucial role in this sort of scheme, and all parties must have a great deal of trust that the arbitration mechanism is working properly Table 13.1, based on scenarios described in [AKL83] and [MITC92], gives several examples of arbitrated digital signatures.
- In the first, symmetric encryption is used. It is assumed that the sender X and the arbiter A share a secret key *Kxa* and that A and Y share secret key *Kay*. X constructs a message *M* and computes its hash value H(*M*).
- Then X transmits the message plus a signature to A. The signature consists of an identifier *IDX* of X plus the hash value, all encrypted using *Kxa*. A decrypts the signature and checks the hash value to validate the message.
- Then A transmits a message to Y, encrypted with *Kay*. The message includes *IDX*, the original message from X, the signature, and a timestamp. Y can decrypt this to recover the message and the signature. The timestamp informs Y that this message is timely and not a replay. Y can store *M* and the signature. In case of dispute, Y, who claims to have received *M* from X, sends the following message to A:

- The arbiter uses *Kay* to recover *IDX*, *M*, and the signature, and then uses *Kxa* to decrypt the signature and verify the hash code. In this scheme, Y cannot directly check X's signature; the signature is there solely to settle disputes. Y considers the message from X authentic because it comes through A.

In this scenario, both sides must have a high degree of trust in A:

  - X must trust A not to reveal *Kxa* and not to generate false signatures of the form E(*Kxa*, [*IDX*||H (*M*)]).
  - Y must trust A to send E(*Kay*, [*IDX*||*M*||E(*Kxa*, [*IDX*||H(*M*)])||*T*]) only if the hash value is correct and the signature was generated by X.

- Both sides must trust A to resolve disputes fairly.
- If the arbiter does live up to this trust, then X is assured that no one can forge his signature and Y is assured that X cannot disavow his signature.

4.8. AUTHENTICATION PROTOCOLS

Contents
• **Authentication Protocols** o Mutual Authentication o One-Way Authentication

Authentication Protocols

- Authentication Protocols are used to convince parties of each others identity and to exchange session keys.they may be **(mutual authentication and one-way authentication)**

Mutual Authentication

- **Symmetric Encryption Approaches**
- **Public-Key Encryption Approaches**

- An important application area is that of mutual authentication protocols. Such protocols enable communicating parties to satisfy themselves mutually about each other's identity and to exchange session keys.
- Central to the problem of authenticated key exchange are two issues: **confidentiality and timeliness.**
- To prevent masquerade and to prevent compromise of session keys, essential identification and session key information must be communicated in encrypted form.
- This requires the prior existence of secret or public keys that can be used for this purpose. The second issue, timeliness, is important because of the threat of message replays. Such replays, at worst, could allow an opponent to compromise a session key or successfully impersonate another party.
- At minimum, a successful replay can disrupt operations by presenting parties with messages that appear genuine but are not.

Lists the following examples of replay attacks:

- **Simple replay:** The opponent simply copies a message and replays it later.
- **Repetition that can be logged:** An opponent can replay a timestamped message within the valid time window.
- **Repetition that cannot be detected:** This situation could arise because the original message could have been suppressed and thus did not arrive at its destination; only the replay message arrives.
- **Backward replay without modification:** This is a replay back to the message sender. This attack is possible if symmetric encryption is used and the sender cannot easily recognize the difference between messages sent and messages received on the basis of content.

- **The following two general approaches is used:**

  - **Timestamps:** Party A accepts a message as fresh only if the message contains a timestamp that, in A's judgment, is close enough to A's knowledge of current time. This approach requires that clocks among the various participants be synchronized.

- **Challenge/response:** Party A, expecting a fresh message from B, first sends B a nonce (challenge) and requires that the subsequent message (response) received from B contain the correct nonce value. It can be argued.

Symmetric Encryption Approaches

- A two-level hierarchy of symmetric encryption keys can be used to provide confidentiality for communication in a distributed environment. In general, this strategy involves the use of a trusted key distribution center (KDC).
- Each party in the network shares a secret key, known as a master key, with the KDC. The KDC is responsible for generating keys to be used for a short time over a connection between two parties, known as session keys, and for distributing those keys using the master keys to protect the distribution.

1. A $\longrightarrow$ KDC: $ID_A||ID_B||N_1$
2. KDC $\longrightarrow$ A: $E(K_a, [K_s||ID_B||N_1||E(K_b, [K_s||ID_A])])$
3. A $\longrightarrow$ B: $E(K_b, [K_s||ID_A])$
4. A $\longrightarrow$ A: $E(K_s, N_2)$
5. A $\longrightarrow$ B: $E(K_s, f(N_2))$

- In step 1.Secret keys *Ka* and *Kb* are shared between A and the KDC and B and the KDC, respectively. The purpose of the protocol is to distribute securely a session key *Ks* to A and B. A securely acquires a new session key in step 2.
- The message in step 3 can be decrypted, and hence understood, only by B. Step 4 reflects B's knowledge of *Ks*, and step 5 assures B of A's knowledge of *Ks* and assures B that this is a fresh message because of the use of the nonce *N*2 that the purpose of steps 4 and 5 is to prevent a certain type of replay attack.

Suppress-replay attacks

- One way to counter suppress-replay attacks is to enforce the requirement that parties regularly check their clocks against the KDC's clock. The other alternative, which avoids the need for clock synchronization, is to rely on handshaking protocols using nonces.
- This latter alternative is not vulnerable to a suppress-replay attack because the nonces the recipient will choose in the future are unpredictable to the sender. The Needham/Schroeder protocol relies on nonces only but, as we have seen, has other vulnerabilities.

The protocol is as follows:

1. A ⟶ B: $ID_A||N_a$

2. B ⟶ KDC: $ID_B||N_b||E(K_b, [ID_A||N_a||T_b])$

3. KDC ⟶ A: $E(K_a, [ID_B||N_a||K_s||T_b])||E(K_b,[ID_A||K_s||T_b])||N_b$

4. A ⟶ B: $E(K_b, [ID_A||K_s||T_b])||E(K_s, N_b)$

Let us follow this exchange step by step.

1. A initiates the authentication exchange by generating a nonce, *Na*, and sending that plus its identifier to B in plaintext. This nonce will be returned to A in an encrypted message that includes the session key, assuring A of its timeliness.

2. B alerts the KDC that a session key is needed. Its message to the KDC includes its identifier and a nonce, *Nb* This nonce will be returned to B in an encrypted message that includes the session key, assuring B of its timeliness. B's message to the KDC also includes a block encrypted with the secret key shared by B and the KDC. This block is used to instruct the KDC to issue credentials to A; the block specifies the intended recipient of the credentials, a suggested expiration time for the credentials, and the nonce received from A.

3. The KDC passes on to A B's nonce and a block encrypted with the secret key that B shares with the KDC. The block serves as a "ticket" that can be used by A for subsequent authentications, as will be seen. The KDC also sends to A block encrypted with the secret key shared by A and the KDC. This block verifies that B has received A's initial message (*IDB*) and that this is a timely message and not a replay (*Na*) and it provides A with a session key (*Ks*) and the time limit on its use (*Tb*).

4. A transmits the ticket to B, together with the B's nonce, the latter encrypted with the session key. The ticket provides B with the secret key that is used to decrypt E(*Ks, Nb*) to recover the nonce. The fact that B's nonce is encrypted with the session key authenticates that the message came from A and is not a replay.

Public-Key Encryption Approaches

- This protocol assumes that each of the two parties is in possession of the current public key of the other.

A protocol using timestamps is provided in

1. A ⟶ AS: $ID_A||ID_B$

2. AS ⟶ A: $E(PR_{as}, [ID_A||PU_a||T])||E(PR_{as}, [ID_B||PU_b||T])$

3. A ⟶ B: $E(PR_{as}, [ID_A||PU_a||T])||E(PR_{as}, [ID_B||PU_b||T])$
$||E(PU_b, E(PR_a, [K_s||T]))$

- In this case, the central system is referred to as an authentication server (AS), because it is not actually responsible for secret key distribution. Rather, the AS provides public-key certificates. The session key is chosen and encrypted by A; hence, there is no risk of exposure by the AS.

- The timestamps protect against replays of compromised keys. This protocol is compact but, as before, requires synchronization of clocks. Another approach, proposed by Woo and Lam [WOO92a], makes use of nonces.

The protocol consists of the following steps:

1. A ⟶ KDC: $ID_A||ID_B$
2. KDC ⟶ A: $E(PR_{auth}, [ID_B||PU_b])$
3. A ⟶ B: $E(PU_b, [N_a||ID_A])$
4. B ⟶ KDC: $ID_A||ID_B||E(PU_{auth}, N_a)$
5. KDC ⟶ B: $E(PR_{auth}, [ID_A||PU_a])||E(PU_b, E(PR_{auth}, [N_a||K_s||ID_B]))$
6. B ⟶ A: $E(PU_a, E(PR_{auth}, [(N_a||K_s||ID_B)||N_b]))$
7. A ⟶ B: $E(K_s, N_b)$

- In step 1, A informs the KDC of its intention to establish a secure connection with B. The KDC returns to A. a copy of B's public-key certificate (step 2). Using B's public key, A informs B of its desire to communicate and sends a nonce *Na* (step 3). In step 4, B asks the KDC for A's public-key certificate and requests a session key; B includes A's nonce so that the KDC can stamp the session key with that nonce.
- The nonce is protected using the KDC's public key. In step 5, the KDC returns to B a copy of A's public-key certificate, plus the information {*Na*, *Ks*, *IDB*}. This information basically says that *Ks* is a secret key generated by the KDC on behalf of B and tied to *Na*; the binding of *Ks* and *Na* will assure A that *Ks* is fresh. This triple is encrypted, using the KDC's private key, to allow B to verify that the triple is in fact from the KDC.
- It is also encrypted using B's public key, so that no other entity may use the triple in an attempt to establish a fraudulent connection with A. In step 6, the triple {*Na*, *Ks*, *IDB*}, still encrypted with the KDC's private key, is relayed to A, together with a nonce *Nb* generated by B. All the foregoing are encrypted using A's public key.
- A retrieves the session key *Ks* and uses it to encrypt *Nb* and return it to B. This last message assures B of A's knowledge of the session key. This seems to be a secure protocol that takes into account the various attacks.
- However, the authors themselves spotted a flaw and submitted a revised version of the algorithm in [WOO92b]:

1. A ⟶ KDC: $ID_A||ID_B$

2. KDC ⟶ A: $E(PR_{auth}, [ID_B||PU_b])$

3. A ⟶ B: $E(PU_b, [N_a||ID_A])$

4. B ⟶ KDC: $ID_A||ID_B||E(PU_{auth}, N_a)$

5. KDC ⟶ B: $E(PR_{auth}, [ID_A||PU_a])||E(PU_b, E(PR_{auth}, [N_a||K_s||ID_A||ID_B])$

6. B ⟶ A: $E(PU_a, E(PR_{auth}, [(N_a||K_s||ID_A||ID_B)||N_b]))$

7. A ⟶ B: $E(K_s, N_b)$

- The identifier of A, *IDA*, is added to the set of items encrypted with the KDC's private key in steps 5 and 6. This binds the session key *Ks* to the identities of the two parties that will be engaged in the session.
- This inclusion of *IDA* accounts for the fact that the nonce value *Na* is considered unique only among all nonces generated by A, not among all nonces generated by all parties. Thus, it is the pair {*IDA*, *Na*} that uniquely identifies the connection request of A.

One-Way Authentication

- One application for which encryption is growing in popularity is electronic mail (e-mail). The very nature of electronic mail, and its chief benefit, is that it is not necessary for the sender and receiver to be online at the same time. Instead, the e-mail message is forwarded to the receiver's electronic mailbox, where it is buffered until the receiver is available to read it.
- The "envelope" or header of the e-mail message must be in the clear, so that the message can be handled by the store-and-forward e-mail protocol, such as the Simple Mail Transfer Protocol (SMTP) or X.400. However, it is often desirable that the mail-handling protocol not require access to the plaintext form of the message, because that would require trusting the mail-handling mechanism. Accordingly, the e-mail message should be encrypted such that the mail-handling system is not in possession of the decryption key.
- A second requirement is that of authentication. Typically, the recipient wants some assurance that the message is from the alleged sender.

Symmetric Encryption Approach

- Using symmetric encryption, the decentralized key distribution scenario illustrated in Figure 7.11 is impractical. This scheme requires the sender to issue a request to the intended recipient, await a response that includes a session key, and only then send the message.
- With some refinement, the KDC strategy illustrated in Figure 7.9 is a candidate for encrypted electronic mail. Because we wish to avoid requiring that the recipient (B) be on line at the same time as the sender (A), steps 4 and 5 must be eliminated. For a message with content *M*, the sequence is as follows:

1. A ⟶ KDC: $ID_A||ID_B||N_1$

2. KDC ⟶ A: $E(K_a, [K_s||ID_B||N_1||E(K_b, [K_s||ID_A])])$

3. A ⟶ B: $E(K_b, [K_s||ID_A])||E(K_s, M)$

- This approach guarantees that only the intended recipient of a message will be able to read it. It also provides a level of authentication that the sender is A. As specified, the protocol does not protect against replays. Some measure of defense could be provided by including a timestamp with the message.
- However, because of the potential delays in the e-mail process, such timestamps may have limited usefulness.

Public-Key Encryption Approaches

- We have already presented public-key encryption approaches that are suited to electronic mail, including the straightforward encryption of the entire message for confidentiality, authentication or both
- These approaches require that either the sender know the recipient's public key (confidentiality) or the recipient know the sender's public key (authentication) or both (confidentiality plus authentication). In addition, the public-key algorithm must be applied once or twice to what may be a long message.
- If confidentiality is the primary concern, then the following may be more efficient:

$$A \longrightarrow B: E(PU_b, K_s)||E(K_s, M)$$

- In this case, the message is encrypted with a one-time secret key. A also encrypts this one-time key with B's public key. Only B will be able to use the corresponding private key to recover the one-time key and then use that key to decrypt the message. This scheme is more efficient than simply encrypting the entire message with B's public key.
- If authentication is the primary concern, then a digital signature may suffice, as was illustrated in

$$A \longrightarrow B: M||E(PR_a, H(M))$$

- This method guarantees that A cannot later deny having sent the message. However, this technique is open to another kind of fraud. Bob composes a message to his boss Alice that contains an idea that will save the company money. He appends his digital signature and sends it into the e-mail system.
- Eventually, the message will get delivered to Alice's mailbox. But suppose that Max has heard of Bob's idea and gains access to the mail queue before delivery. He finds Bob's message, strips off his signature, appends his, and requeues the message to be delivered to Alice. Max gets credit for Bob's idea.
- To counter such a scheme, both the message and signature can be encrypted with the recipient's public key:

$$A \longrightarrow B: E(PU_b, [M||E(PR_a, H(M))])$$

4.9. DIGITAL SIGNATURE STANDARD

Contents
• **Digital Signatures Standard** ○ The DSS Approach ○ The Digital Signature Algorithm

Digital Signatures Standard

- The DSS makes use of the Secure Hash Algorithm (SHA) presents a new digital signature technique, the **Digital Signature Algorithm (DSA)**.

The DSS Approach

- The DSS uses an algorithm that is designed to provide only the digital signature function. Unlike RSA, it cannot be used for encryption or key exchange. Nevertheless, it is a public-key technique.
- Figure 13.3 contrasts the DSS approach for generating digital signatures to that used with RSA. In the RSA approach, the message to be signed is input to a hash function that produces a secure hash code of fixed length. This hash code is then encrypted using the sender's private key to form the signature.
- The DSS approach also makes use of a hash function. The hash code is provided as input to a signature function along with a random number generated for this particular signature.
- The signature function also depends on the sender's private key and a set of parameters known to a group of communicating principals. We can consider this set to constitute a global public key .1 The result is a signature consisting of two components, labeled *s* and *r*.

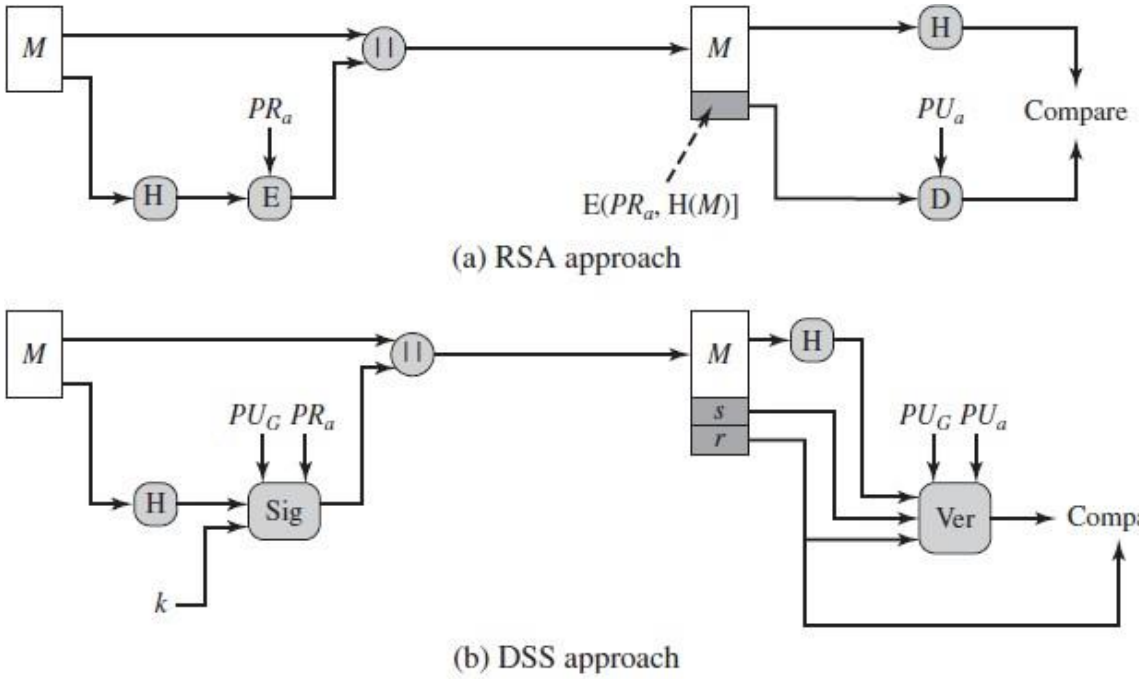


Figure 13.3 Two Approaches to Digital Signatures

- At the receiving end, the hash code of the incoming message is generated. This plus the signature is input to a verification function. The verification function also depends on the global public key as well as the sender's public key , which is paired with the sender's private key.
- The output of the verification function is a value that is equal to the signature component if the signature is valid. The signature function is such that only the sender, with knowledge of the private key, could have produced the valid signature. We turn now to the details of the algorithm.

The Digital Signature Algorithm

- Figure 13.4 summarizes the algorithm. There are three parameters that are public and can be common to a group of users. A 160-bit prime number is chosen. Next, a prime number is selected with a length between 512 and 1024 bits such that divides (*p* - 1).
- Finally, *g* is chosen to be of the form *h*(*p*-1)/*q*mod *p*, where *h* is an*p q q r* (*PUa*) 2In number-theoretic terms, g is of order q modp; see Chapter 8. integer between 1 and with the restriction that must be greater than 1.2
- Thus, the global public-key components of DSA have the same for as in the Schnorr signature scheme. With these numbers in hand, each user selects a private key and generates a public key. The private key must be a number from 1 to and should be chosen randomly or pseudorandomly.
- The public key is calculated from the private key as .The calculation of given is relatively straightforward. However, given the public key, it is believed to be computationally infeasible to determine , which is the discrete logarithm of *y* to the base *g*, mod p.

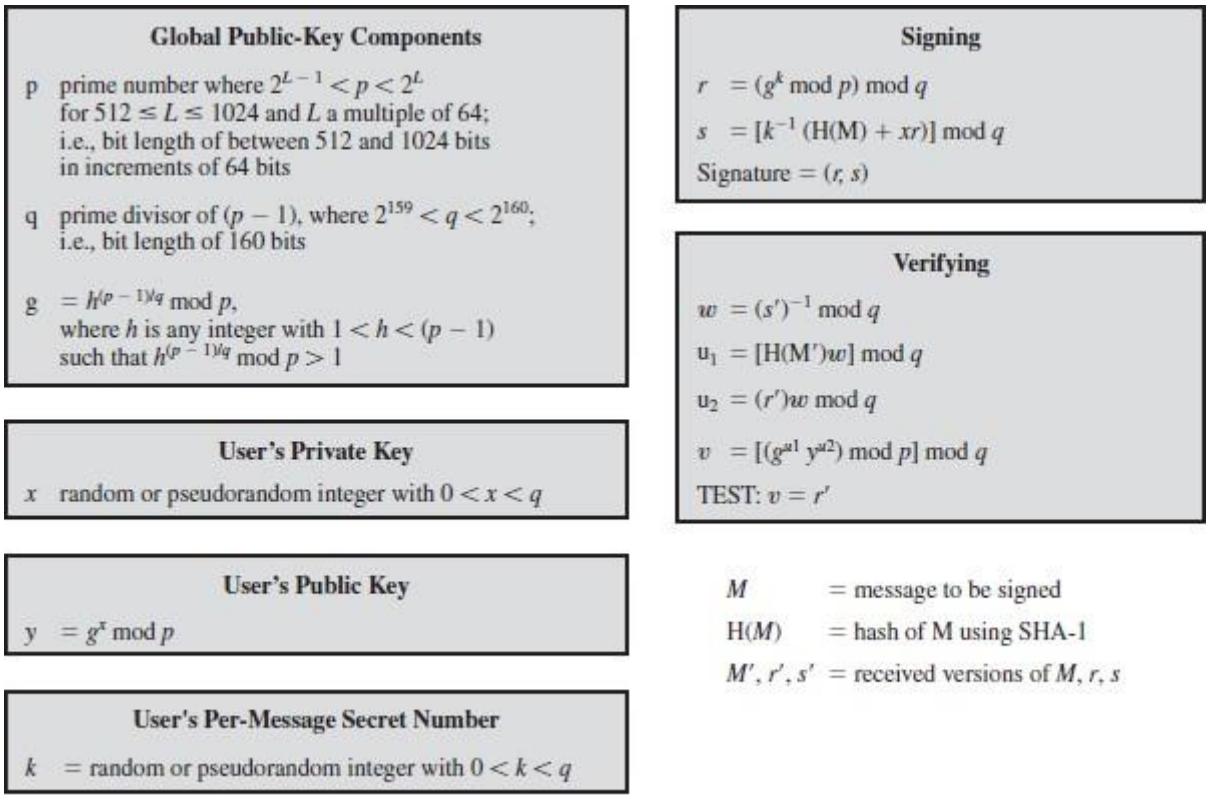
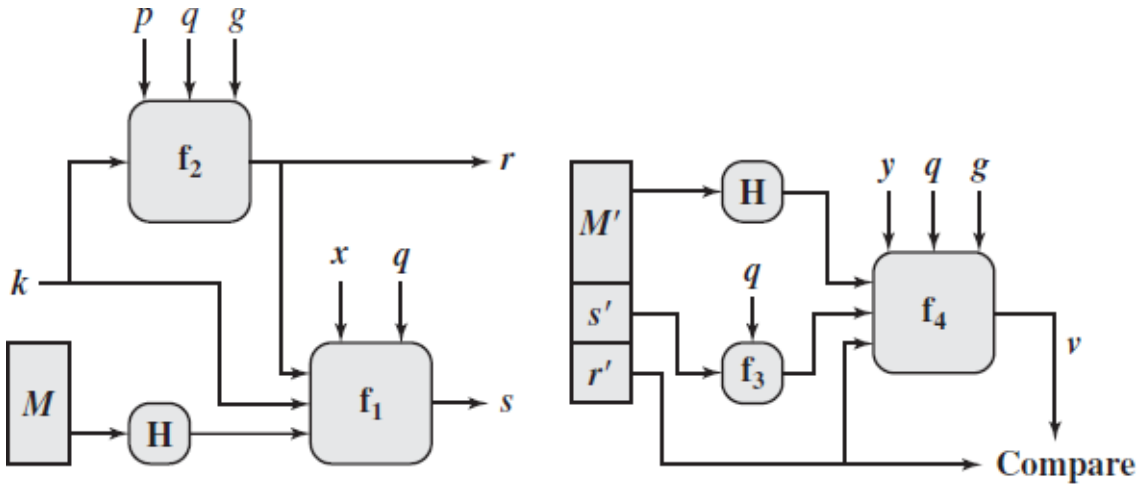


Figure 13.4 The Digital Signature Algorithm (DSA)

- To create a signature, a user calculates two quantities, and , that are functions of the public key components , the user's private key , the hash code of the message , and an additional integer that should be generated randomly or pseudorandomly and be unique for each signing.
- At the receiving end, verification is performed using the formulas shown in Figure 13.4.The receiver generates a quantity that is a function of the public key components, the sender's public key, and the hash code of the incoming message. If this quantity matches the component of the signature, then the signature is validated. Figure 13.5 depicts the functions of signing and verifying.

- The structure of the algorithm, as revealed in Figure 13.5, is quite interesting. Note that the test at the end is on the value, which does not depend on the message at all. Instead, is a function of and the three global public-key components.
- The multiplicative inverse of is passed to a function that also has as inputs the message hash code and the user's private key. The structure of this function is such that the receiver can recover using the incoming message and signature, the public key of the user, and the global public key. It is certainly not obvious from Figure 13.4 or Figure 13.5 that such a scheme would work. Because this value does not depend on the message to be signed, it can be computed ahead of time.
- Indeed, a user could precalculate a number of values of to be used to sign documents as needed. The only other somewhat demanding task is the determination of a multiplicative inverse, Again, a number of these values can be precalculated.



$s = f_1(H(M), k, x, r, q) = (k^{-1} (H(M) + xr)) \bmod q$

$r = f_2(k, p, q, g) = (g^k \bmod p) \bmod q$

(a) Signing

$w = f_3(s', q) = (s')^{-1} \bmod q$

$v = f_4(y, q, g, H(M'), w, r')$

$= ((g^{(H(M')w) \bmod q} y^{r'w \bmod q}) \bmod p) \bmod q$

(b) Verifying

Figure 13.5 DSS Signing and Verifying

4.10. ENTITY AUTHENTICATION

4.10.1. BIOMETRICS

4.10.2. PASSWORDS

4.11. CHALLENGE RESPONSE PROTOCOLS

4.12. AUTHENTICATION APPLICATIONS

- **Kerberos**
  - Motivation
  - Kerberos Version 4
  - Kerberos Version 5
- **X.509 Authentication Service**
  - Certificates
  - Authentication Procedures
  - X.509 Version 3

4.12.1. KERBEROS

Contents

Kerberos

- Kerberos4 is an authentication service developed as part of Project Athena at MIT. The problem that Kerberos addresses is this:
- Assume an open distributed environment in which users at workstations wish to access services on servers distributed throughout the network.

In particular, the following three threats exist:

1. A user may gain access to a particular workstation and pretend to be another user operating from that workstation.
2. A user may alter the network address of a workstation so that the requests sent from the altered workstation appear to come from the impersonated workstation.
3. A user may eavesdrop on exchanges and use a replay attack to gain entrance to a server or to disrupt operations.

- In any of these cases, an unauthorized user may be able to gain access to services and data that he or she is not authorized to access.

- Rather than building in elaborate authentication protocols at each server, Kerberos provides a centralized authentication server whose function is to authenticate users to servers and servers to users.

Motivation

1. Rely on each individual client workstation to assure the identity of its user or users and rely on each server to enforce a security policy based on user identification (ID).

2. Require that client systems authenticate themselves to servers, but trust the client system concerning the identity of its user.
3. Require the user to prove his or her identity for each service invoked. Also require that servers prove their identity to clients.

Kerberos Version 4

- Version 4 of Kerberos makes use of DES, in a rather elaborate protocol, to provide the authentication service.
- Viewing the protocol as a whole, it is difficult to see the need for the many elements contained therein.

- Therefore, we adopt a strategy used by Bill Bryant of Project Athena [BRYA88] and build up to the full protocol by looking first at several hypothetical dialogues.
- Each successive dialogue adds additional complexity to counter security vulnerabilities revealed in the preceding dialogue.

A Simple Authentication Dialogue

In an unprotected network environment, any client can apply to any server for service. The obvious security risk is that of impersonation.

- An opponent can pretend to be another client and obtain unauthorized privileges on server machines.
- To counter this threat, servers must be able to confirm the identities of clients who request service.
- Each server can be required to undertake this task for each client/server interaction, but in an open environment, this places a substantial burden on each server.
- An alternative is to use an authentication server (AS) that knows the passwords of all users and stores these in a centralized database.
- In addition, the AS shares a unique secret key with each server. These keys have been distributed physically or in some other secure manner.
- Consider the following hypothetical dialogue:

(1) C ⟶ AS: $ID_C||P_C||ID_V$

(2) AS ⟶ C: $Ticket$

(3) C ⟶ V: $ID_C||Ticket$

$Ticket = E(K_V, [ID_C||AD_C||ID_V])$

Where

C = client
AS = authentication server
V =server
ID_C = identifier of user on C
ID_V = identifier of V
P_C = password of user on C
AD_C = network address of C
Kv = secret encryption key shared by AS and V

A More Secure Authentication Dialogue

- Although the foregoing scenario solves some of the problems of authentication in an open network environment, problems remain.

Once per user logon session:

(1) C ⟶ AS: $ID_C||ID_{tgs}$

(2) AS ⟶ C: $E(K_C, Ticket_{tgs})$

Once per type of service:

(3) C ⟶ TGS: $ID_C||ID_V||Ticket_{tgs}$

(4) TGS ⟶ C: $Ticket_V$

Once per service session:

(5) C ⟶ V: $ID_C||Ticket_V$

$Ticket_{tgs} = E(K_{tgs}, [ID_C||AD_C||ID_{tgs}||TS_1||Lifetime_1])$

$Ticket_V = E(K_V, [ID_C||AD_C||ID_V||TS_2||Lifetime_2])$

Let us look at the details of this scheme:

1.The client requests a ticket-granting ticket on behalf of the user by sending its user's ID and password to the AS, together with the TGS ID, indicating a request to use the TGS service.

2. The AS responds with a ticket that is encrypted with a key that is derived from the user's password. When this response arrives at the client, the client prompts the user for his or her password, generates the key, and attempts to decrypt the incoming message.

- If the correct password is supplied, the ticket is successfully recovered. Because only the correct user should know the password, only the correct user can recover the ticket.
- Thus, we have used the password to obtain credentials from Kerberos without having to transmit the password in plaintext.
- The ticket itself consists of the ID and network address of the user, and the ID of the TGS. This corresponds to the first scenario.

3. The client requests a service-granting ticket on behalf of the user. For this purpose, the client transmits a message to the TGS containing the user's ID, the ID of the desired service, and the ticket-granting ticket.

4. The TGS decrypts the incoming ticket and verifies the success of the decryption by the presence of its ID. It checks to make sure that the lifetime has not expired.

Then it compares the user ID and network address with the incoming information to authenticate the user.

- If the user is permitted access to the server V, the TGS issues a ticket to grant access to the requested.

- The service-granting ticket has the same structure as the ticket-granting ticket. Indeed, because the TGS is a server, we would expect that the same elements are needed to authenticate a client to the TGS and to authenticate a client to an application server.
- Again, the ticket contains a timestamp and lifetime. If the user wants access to the same service at a later time, the client can simply use the previously acquired service-granting ticket and need not bother the user for a password.
- Note that the ticket is encrypted with a secret key (*Kv*).

The Version 4 Authentication Dialogue:

Although the foregoing scenario enhances security compared to the first attempt, two additional problems remain.

- The heart of the first problem is the lifetime associated with the ticket-granting ticket.
- If this lifetime is very short (e.g., minutes), then the user will be repeatedly asked for a password. If the lifetime is long (e.g., hours), then an opponent has a greater opportunity for replay.
- An opponent could eavesdrop on the network and capture a copy of the ticket-granting ticket and then wait for the legitimate user to log out.
- Then the opponent could forge the legitimate user's network address and send the message of step (3) to the TGS.
- This would give the opponent unlimited access to the resources and files available to the legitimate user.

Summary of Kerberos Version 4 Message Exchanges

(a) Authentication Service Exchange: to obtain ticket-granting ticket
(1) C → AS: $ID_c \| ID_{tgs} \| TS_1$
(2) AS → C: $E_{K_c}[K_{c,tgs} \| ID_{tgs} \| TS_2 \| Lifetime_2 \| Ticket_{tgs}]$
$Ticket_{tgs} = E_{K_{tgs}}[K_{c,tgs} \| ID_C \| AD_C \| ID_{tgs} \| TS_2 \| Lifetime_2]$
(b) Ticket-Granting Service Exchange: to obtain service-granting ticket
(3) C → TGS: $ID_v \| Ticket_{tgs} \| Authenticator_c$
(4) TGS → C: $E_{K_{c,tgs}}[K_{c,v} \| ID_v \| TS_4 \| Ticket_v]$
$Ticket_{tgs} = E_{K_{tgs}}[K_{c,tgs} \| ID_C \| AD_C \| ID_{tgs} \| TS_2 \| Lifetime_2]$
$Ticket_v = E_{K_v}[K_{c,v} \| ID_C \| AD_C \| ID_v \| TS_4 \| Lifetime_4]$
$Authenticator_c = E_{K_{tgs}}[ID_C \| AD_C \| TS_3]$
(c) Client/Server Authentication Exchange: to obtain service
(5) C → V: $Ticket_v \| Authenticator_c$
(6) V → C: $E_{K_{c,v}}[TS_5 + 1]$ (for mutual authentication)
$Ticket_v = E_{K_v}[K_{c,v} \| ID_C \| AD_C \| ID_v \| TS_4 \| Lifetime_4]$
$Authenticator_c = E_{K_{c,v}}[ID_C \| AD_C \| TS_5]$

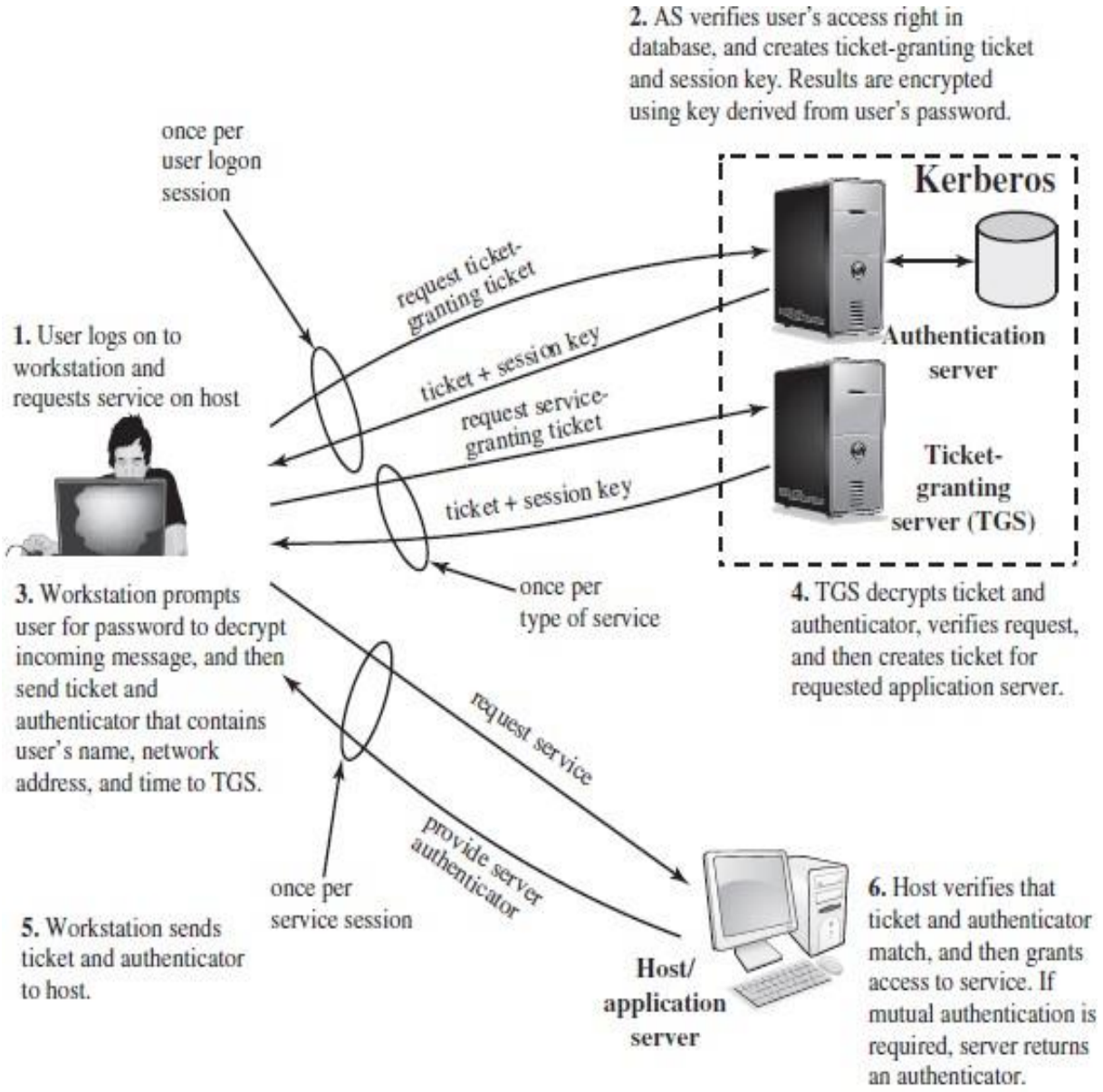
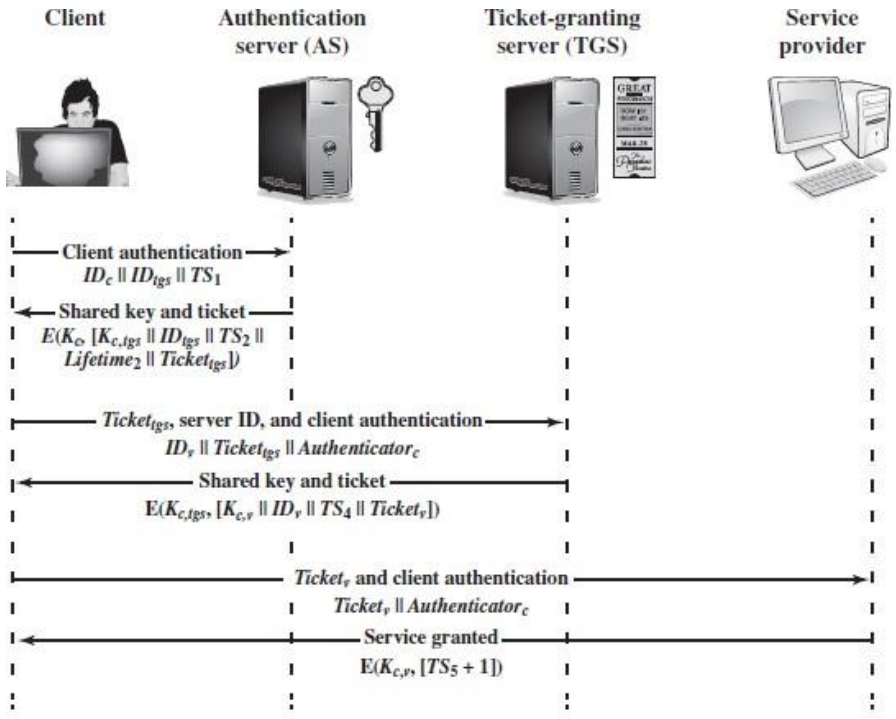


Figure 15.1 Overview of Kerberos

Figure 15.2 illustrates the Kerberos exchanges among the parties. Table 15.2 summarizes the justification for each of the elements in the Kerberos protocol.



ure 15.2 Kerberos Exchanges

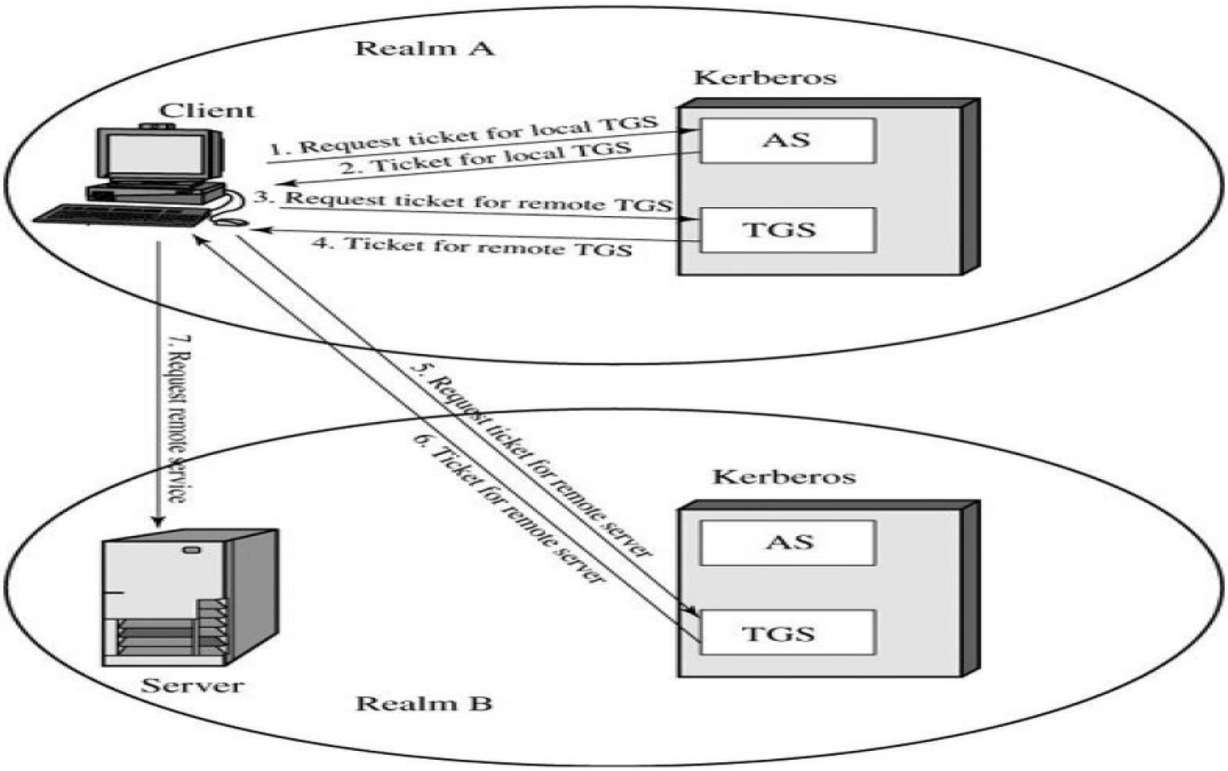
Kerberos Realms and Multiple Kerberi

- A full-service Kerberos environment consisting of a Kerberos server, a number of clients, and a number of application servers requires the following:
  1. The Kerberos server must have the user ID and hashed passwords of all participating
     users in its database. All users are registered with the Kerberos server.
  2. The Kerberos server must share a secret key with each server. All servers are registered with the Kerberos server. Such an environment is referred to as a **Kerberos realm**.
- The concept of **realm** can be explained as follows. A Kerberos realm is a set of managed nodes that share the same Kerberos database.

- Changing or accessing the contents of a Kerberos database requires the Kerberos master password. A related concept is that of a **Kerberos principal**, which is a service or user that is known to the Kerberos system.
- Each Kerberos principal is identified by its principal name. Principal names consist of three parts: a service or user name, an instance name, and a realm name.

  3. The Kerberos server in each interoperating realm shares a secret key with the server in the other realm. The two Kerberos servers are registered with each other.

- The scheme requires that the Kerberos server in one realm trust the Kerberos server in the other realm to authenticate its users..



(1) C ⟶ AS: $ID_c||ID_{tgs}||TS_1$

(2) AS ⟶ C: $E(K_c, [K_{c,tgs}||ID_{tgs}||TS_2||Lifetime_2||Ticket_{tgs}])$

(3) C ⟶ TGS: $ID_{tgsrem}||Ticket_{tgs}||Authenticator_c$

(4) TGS ⟶ C: $E(K_{c,tgs}, [K_{c,tgsrem}||ID_{tgsrem}||TS_4||Ticket_{tgsrem}])$

(5) C ⟶ TGS_{rem}: $ID_{vrem}||Ticket_{tgsrem}||Authenticator_c$

(6) TGS_{rem} ⟶ C: $E(K_{c,tgsrem}, [K_{c,vrem}||ID_{vrem}||TS_6||Ticket_{vrem}])$

(7) C ⟶ V_{rem}: $Ticket_{vrem}||Authenticator_c$

Kerberos Version 5

- Kerberos Version 5 is specified in RFC 1510 and provides a number of improvements over version 4

Differences between Versions 4 and 5

- Version 5 is intended to address the limitations of version 4 in two areas: environmental shortcomings and technical deficiencies
- Kerberos Version 4 was developed for use within the Project Athena environment and, accordingly, did not fully address the need to be of general purpose.

The Version 5 Authentication Dialogue

(a) Authentication Service Exchange: to obtain ticket-granting ticket
(1) C → AS: $\text{Options} \| ID_c \| Realm_c \| ID_{tgs} \| Times \| Nonce_1$
(2) AS → C: $Realm_c \| ID_C \| Ticket_{tgs} \| E_{K_c}\left[K_{c,tgs} \| Times \| Nonce_1 \| Realm_{tgs} \| ID_{tgs}\right]$
$Ticket_{tgs} = E_{K_{tgs}}\left[Flags \| K_{c,tgs} \| Realm_c \| ID_C \| AD_C \| Times\right]$
(b) Ticket-Granting Service Exchange: to obtain service-granting ticket
(3) C → TGS: $\text{Options} \| ID_v \| Times \| \| Nonce_2 \| Ticket_{tgs} \| Authenticator_c$
(4) TGS → C: $Realm_c \| ID_C \| Ticket_v \| E_{K_{c,tgs}}\left[K_{c,v} \| Times \| Nonce_2 \| Realm_v \| ID_V\right]$
$Ticket_{tgs} = E_{K_{tgs}}\left[Flags \| K_{c,tgs} \| Realm_c \| ID_C \| AD_C \| Times\right]$
$Ticket_v = E_{K_v}\left[Flags \| K_{c,v} \| Realm_c \| ID_C \| AD_C \| Times\right]$
$Authenticator_c = E_{K_{c,tgs}}\left[ID_C \| Realm_c \| TS_1\right]$
(c) Client/Server Authentication Exchange: to obtain service
(5) C → V: $\text{Options} \| \text{Ticket}_v \| \text{Authenticator}_c$
(6) V → C: $E_{K_{c,V}}[\ TS_2 \| \text{Subkey} \| \text{Seq\#}\ ]$
$Ticket_v = E_{K_v}\left[Flags \| K_{c,v} \| Realm_c \| ID_C \| AD_C \| Times\right]$
$Authenticator_c = E_{K_{c,V}}\left[ID_C \| Realm_c \| TS_2 \| Subkey \| Seq\#\right]$

The following new elements are added:

- **Realm:** Indicates realm of user
- **Options:** Used to request that certain flags be set in the returned ticket
- **Times:** Used by the client to request the following time settings in the ticket:
- **Nonce:** A random value to be repeated in message (2) to assure that the response is fresh and has not been replaced by an opponent.
-

The authenticator includes several new fields as follows:

Subkey: The client's choice for an encryption key to be used to protect this specific application session. If this field is omitted, the session key from the ticket (Kc,v) is used.

Sequence number: An optional field that specifies the starting sequence number to be used by the server for messages sent to the client during this session. Messages may be sequence numbered to detect replays.

Ticket Flags*: The* flags field included in tickets in version 5 supports expanded functionality compared to that available in version 4.

4.12.2. X.509

Contents

• **X.509 Authentication Service**
  o Certificates
  o Authentication Procedures
  o X.509 Version 3

X.509 Authentication services:

- X.509 defines a framework for the provision of authentication services by the X.500 directory to its users. Each certificate contains the public key of a user and is signed with the private key of a trusted certification authority.
- In addition, X.509 defines alternative authentication protocols based on the use of public-key certificates.
- X.509 is based on the use of public-key cryptography and digital signatures. The standard does not dictate the use of a specific algorithm but recommends RSA. The digital signature scheme is assumed to require the use of a hash function.

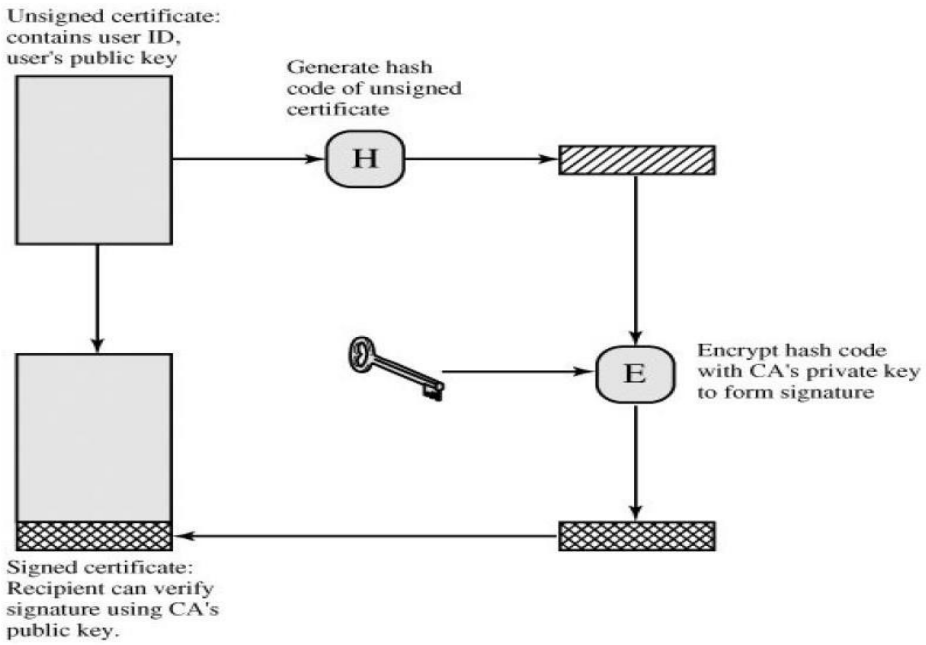


Figure 14.3. Public-Key Certificate Use

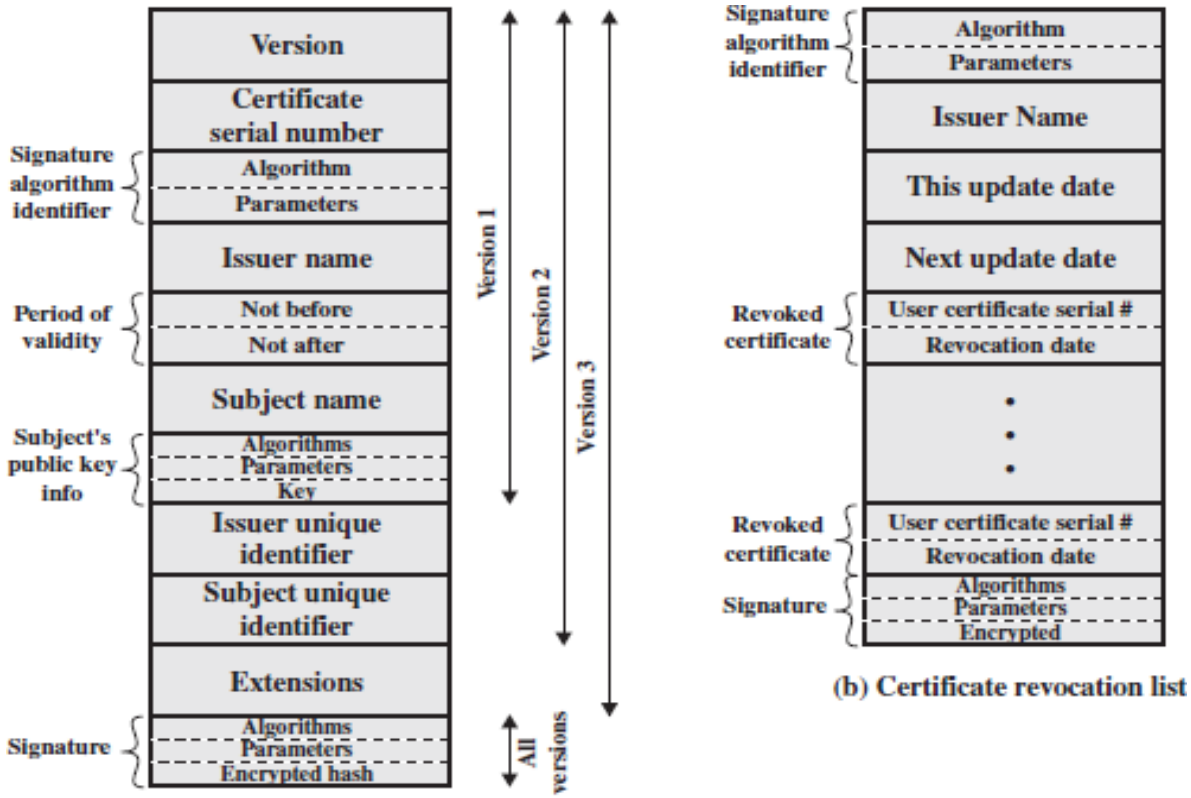
Certificates:

- The heart of the X.509 scheme is the public-key certificate associated with each user. These user certificates are assumed to be created by some trusted certification authority (CA) and placed in the directory by the CA or by the user.

- The directory server itself is not responsible for the creation of public keys or for the certification function; it merely provides an easily accessible location for users to obtain certificates.

- Figure 14.15a shows the general format of a certificate, which includes the following elements.

    - **Version:** Differentiates among successive versions of the certificate format; the default is version 1. If the Issuer Unique Identifier or Subject Unique Identifier are present, the value must be version 2. If one or more extensions are present, the version must be version 3.
    - **Serial number:** An integer value, unique within the issuing CA, that is unambiguously associated with this certificate.
    - **Signature algorithm identifier:** The algorithm used to sign the certificate, together with any associated parameters. Because this information is repeated in the Signature field at the end of the certificate, this field has little, if any, utility.
    - **Issuer name:** X.500 name of the CA that created and signed this certificate.
    - **Period of validity:** Consists of two dates: the first and last on which the certificate is valid.
    - **Subject name:** The name of the user to whom this certificate refers. That is, this certificate certifies the public key of the subject who holds the corresponding private key.
    - **Subject's public-key information:** The public key of the subject, plus an identifier of the algorithm for which this key is to be used, together with any associated parameters.
    - **Issuer unique identifier:** An optional bit string field used to identify uniquely the issuing CA in the event the X.500 name has been reused for different entities.
    - **Subject unique identifier:** An optional bit string field used to identify uniquely the subject in the event the X.500 name has been reused for different entities.
    - **Extensions:** A set of one or more extension fields. Extensions were added in version 3 and are discussed later in this section.
    - **Signature:** Covers all of the other fields of the certificate; it contains the hash code of the other fields, encrypted with the CA's private key. This field includes the signature algorithm identifier.



(a) X.509 certificate

(b) Certificate revocation list

- The unique identifier fields were added in version 2 to handle the possible reuse of subject and/or issuer names over time. These fields are rarely used. The standard uses the following notation to define a certificate:

$$CA \ll A \gg = CA \{V, SN, AI, CA, UCA, A, UA, Ap, T^A\}$$

Where

V = version of the certificate
SN = serial number of the certificate
AI = identifier of the algorithm used to sign the certificate
CA = name of certificate authority
UCA = optional unique identifier of the CA
A = name of user A
UA = optional unique identifier of the user A
Ap = public key of user A
T^A = period of validity of the certificate

Obtaining a User's Certificate

- User certificates generated by a CA have the following characteristics:
- Any user with access to the public key of the CA can verify the user public key that was certified.
- No party other than the certification authority can modify the certificate without this being detected.

- The CA signs the certificate with its private key. If the corresponding public key is known to a user, then that user can verify that a certificate signed by the CA is valid.
- Figure 14.16, taken from X.509, is an example of such a hierarchy. The connected circles indicate the hierarchical relationship among the CAs; the associated boxes indicate certificates maintained in the directory for each CA entry.

- The directory entry for each **CA includes two types of certificates:**
  • **Forward certificates:** Certificates of X generated by other CAs
  • **Reverse certificates:** Certificates generated by X that are the certificates of other CAs

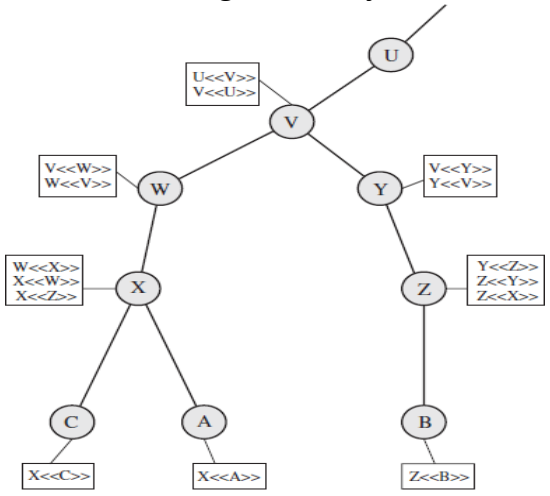


Figure 14.16 X.509 Hierarchy: A Hypothetical Example

- In this example, user A can acquire the following certificates from the directory to establish a certification path to B:

$$X \ll W \gg W \ll V \gg V \ll Y \gg Y \ll Z \gg Z \ll B \gg$$

- When A has obtained these certificates, it can unwrap the certification path in sequence to recover a trusted copy of B's public key. Using this public key, A can send encrypted messages to B. If A wishes to receive encrypted messages back from B, or to sign messages sent to B, then B will require A's public key, which can be obtained from the following certification path:

$$Z \ll Y \gg Y \ll V \gg V \ll W \gg W \ll X \gg X \ll A \gg$$

- B can obtain this set of certificates from the directory, or A can provide them as part of its initial message to B.

Revocation of Certificates

- Recall from Figure 14.4 that each certificate includes a period of validity, much like a credit card. Typically, a new certificate is issued just before the expiration of the old one.
- In addition, it may be desirable on occasion to revoke a certificate before it expires, for one of the following reasons:
  1. The user's private key is assumed to be compromised.
  2. The user is no longer certified by this CA.
  3. The CA's certificate is assumed to be compromised.

Authentication Procedures

- X.509 also includes three alternative authentication procedures that are intended for use across a variety of applications.
- All these procedures make use of public-key signatures. It is assumed that the two parties know each other's public key, either by obtaining each other's certificates from the directory or because the certificate is included in the initial message from each side.

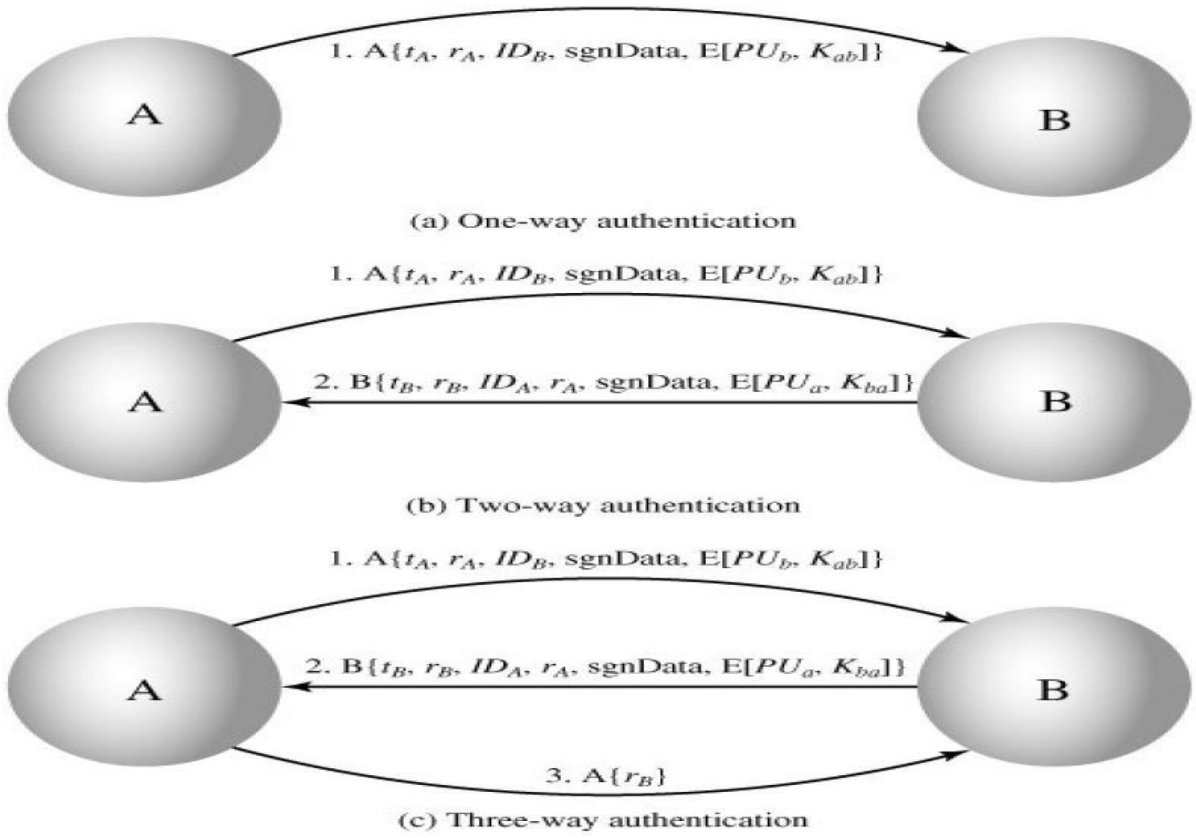
One-Way Authentication

- One way authentication involves a single transfer of information from one user (A) to another (B), and establishes the following:

  1. The identity of A and that the message was generated by A

  2. That the message was intended for B

  3. The integrity and originality (it has not been sent multiple times) of the message. At a minimum, the message includes a timestamp *tA*, a nonce *rA* and the identity of B and is signed with A's private key.



- **Figure 14.6. X.509 Strong Authentication Procedures**

- The timestamp consists of an optional generation time and an expiration time. This prevents delayed delivery of messages.
- The nonce can be used to detect replay attacks. The nonce value must be unique within the expiration time of the message.

- Thus, B can store the nonce until it expires and reject any new messages with the same nonce.

Two-Way Authentication

- In addition to the three elements just listed, two-way authentication establishes the following elements:

    4. The identity of B and that the reply message was generated by B

    5. That the message was intended for A

    6. The integrity and originality of the reply

- Two-way authentication thus permits both parties in a communication to verify the identity of the other.
- The reply message includes the nonce from A, to validate the reply. It also includes a timestamp and nonce generated by B. As before, the message may include signed additional information and a session key encrypted with A's public key.

Three-Way Authentication

- In three-way authentication, a final message from A to B is included, which contains a signed copy of the nonce *rB*.
- The intent of this design is that timestamps need not be checked: Because both nonces are echoed back by the other side, each side can check the returned nonce to detect replay attacks. This approach is needed when synchronized clocks are not available.

X.509 Version 3

- The X.509 version 2 format does not convey all of the information that recent design and implementation experience has shown to be needed.

Lists the following requirements not satisfied by version 2:

1. The Subject field is inadequate to convey the identity of a key owner to a public-key user. X.509 names may be relatively short and lacking in obvious identification details that may be needed by the user.

2. The Subject field is also inadequate for many applications, which typically recognize entities by an Internet e-mail address, URL, or some other Internet-related identification.

3. There is a need to indicate security policy information. This enables a security application or function, such as IPSec, to relate an X.509 certificate to a given policy.

4. There is a need to limit the damage that can result from a faulty or malicious CA by setting constraints on the applicability of a particular certificate.

4. It is important to be able to identify different keys used by the same owner at different times. This feature supports key life cycle management, in particular the ability to update key pairs for users and CAs on a regular basis or under exceptional circumstances.

Certificate Subject and Issuer Attributes

- These extensions support alternative names, in alternative formats, for a certificate subject or certificate issuer and can convey additional information about the certificate subject, to increase a certificate user's confidence that the certificate subject is a particular person or entity.
- For example, information such as postal address, position within a corporation, or picture image may be required.

The extension fields in this area include the following:

- **Subject alternative name:** Contains one or more alternative names, using any of a variety of forms. This field is important for supporting certain applications, such as electronic mail, EDI, and IPSec, which may employ their own name forms.
- **Issuer alternative name:** Contains one or more alternative names, using any of a variety of forms.
- **Subject directory attributes:** Conveys any desired X.500 directory attribute values for the subject of this certificate.
-

Certification Path Constraints

- These extensions allow constraint specifications to be included in certificates issued for CAs by other CAs.
- The constraints may restrict the types of certificates that can be issued by the subject CA or that may occur subsequently in a certification chain.

The extension fields in this area include the following:

Basic constraints: Indicates if the subject may act as a CA. If so, a certification path length constraint may be specified.

Name constraints: Indicates a name space within which all subject names in subsequent certificates in a certification path must be located.

Policy constraints: Specifies constraints that may require explicit certificate policy identification or inhibit policy mapping for the remainder of the certification path.

UNIT II
SYMMETRIC KEY CRYPTOGRAPHY

MATHEMATICS OF SYMMETRIC KEY CRYPTOGRAPHY: Algebraic structures – Modular arithmetic-Euclid‟s algorithm- Congruence and matrices – Groups, Rings, Fields-Finite fields- SYMMETRIC KEY CIPHERS: SDES – Block cipher Principles of DES – Strength of DES – Differential and linear cryptanalysis – Block cipher design principles – Block cipher mode of operation – Evaluation criteria for AES – Advanced Encryption Standard – RC4 – Key distribution.

MATHEMATICS OF SYMMETRIC KEY CRYPTOGRAPHY
2.2. MODULAR ARITHMETIC

Contents
• **The Modulus** • **Properties of Congruences** • **Modular Arithmetic Operations** • **Properties of Modular Arithmetic**

The Modulus

- If a is an integer and n is a positive integer, we define a mod n to be the remainder when a is divided by n. The integer n is called the modulus. Thus, for any integer a, we can rewrite Equation (4.1) as follows:

$$a = qn + r \qquad 0 \le r < n; q = \lfloor a/n \rfloor$$
$$a = \lfloor a/n \rfloor \times n + (a \bmod n)$$

$$11 \bmod 7 = 4; \qquad -11 \bmod 7 = 3$$

- Two integers a and b are said to be congruent modulo n, if (a mod n) = (b mod n). This is written as a K b (mod n).

$$73 \equiv 4 \pmod{23}; \qquad 21 \equiv -9 \pmod{10}$$

Note that if $a \equiv 0 \pmod{n}$, then $n \mid a$.

Properties of Congruences

Congruences have the following properties:

1. $a \equiv b \pmod{n}$ if $n \mid (a - b)$.
2. $a \equiv b \pmod{n}$ implies $b \equiv a \pmod{n}$.
3. $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ imply $a \equiv c \pmod{n}$.

To demonstrate the first point, if n | (a - b), then (a - b) = k*n* for some k.

- So when b is divided by n) = (b mod n).we can write a = b + k*n*. Therefore, (a mod n) = (remainder when b + k*n* is divided by n) = remainder

$$23 \equiv 8 \pmod{5} \quad \text{because} \quad 23 - 8 = 15 = 5 \times 3$$
$$-11 \equiv 5 \pmod{8} \quad \text{because} \quad -11 - 5 = -16 = 8 \times (-2)$$
$$81 \equiv 0 \pmod{27} \quad \text{because} \quad 81 - 0 = 81 = 27 \times 3$$

The remaining points are as easily proved.

Modular Arithmetic Operations

❖ Note that, by definition (Figure 4.1), the (mod n) operator maps all integers into the set of integers {0, 1, c, (n - 1)}. this technique is known as modular arithmetic.

Modular arithmetic exhibits the following properties:

1. $[(a \bmod n) + (b \bmod n)] \bmod n = (a + b) \bmod n$
2. $[(a \bmod n) - (b \bmod n)] \bmod n = (a - b) \bmod n$
3. $[(a \bmod n) \times (b \bmod n)] \bmod n = (a \times b) \bmod n$

Properties of Modular Arithmetic

Define the set Zn as the set of nonnegative integers less than n:

$$Z_n = \{0, 1, \ldots, (n - 1)\}$$

This is referred to as the set of residues, or residue classes (mod n). To be more precise, each integer in Zn represents a residue class. We can label the residue classes (mod n) as [0], [1], [2], c, [n - 1], where

The residue classes (mod 4) are

$[0] = \{\ldots, -16, -12, -8, -4, 0, 4, 8, 12, 16, \ldots\}$

$[1] = \{\ldots, -15, -11, -7, -3, 1, 5, 9, 13, 17, \ldots\}$

Table 4.3 Properties of Modular Arithmetic for Integers in Z_n

Property	Expression
Commutative Laws	$(w + x) \bmod n = (x + w) \bmod n$ $(w \times x) \bmod n = (x \times w) \bmod n$
Associative Laws	$[(w + x) + y] \bmod n = [w + (x + y)] \bmod n$ $[(w \times x) \times y] \bmod n = [w \times (x \times y)] \bmod n$
Distributive Law	$[w \times (x + y)] \bmod n = [(w \times x) + (w \times y)] \bmod n$
Identities	$(0 + w) \bmod n = w \bmod n$ $(1 \times w) \bmod n = w \bmod n$
Additive Inverse $(-w)$	For each $w \in Z_n$, there exists a z such that $w + z = 0 \bmod n$

2.4. EUCLID'S ALGORITHM

Contents
• **Introduction** • **Greatest Common Divisor** • **Finding the Greatest Common Divisor**

Introduction

❖ One of the basic techniques of number theory is the Euclidean algorithm, which is a simple procedure for determining the greatest common divisor of two positive integers. First, we need a simple definition: Two integers are **relatively prime** if their only common positive integer factor is 1.

Greatest Common Divisor:

- Recall that nonzero b is defined to be a divisor of a if $a = mb$ for some m, where a, b, and m are integers.
- We will use the notation gcd(a, b) to mean the **greatest common divisor** of a and b. The greatest common divisor of a and$b$ is the largest integer that divides both a and b.
- We also define gcd(0, 0) = 0.More formally, the positive integer c is said to be the greatest common divisor of a and b if
    1. c is a divisor of a and of b.
    2. Any divisor of a and b is a divisor of c.

An equivalent definition is the following:

$$\gcd(a, b) = \max[k, \text{such that } k|a \text{ and } k|b]$$

- Because we require that the greatest common divisor be positive, gcd(a, b) = gcd(a, −b) = gcd(−a, b) = gcd(−a,−b). In general, gcd(a, b) = gcd(|a| , |b|).

$$\gcd(60, 24) = \gcd(60, -24) = 12$$

Finding the Greatest Common Divisor

- Suppose we have integers a, b such that d = gcd(a, b). Because gcdgcd(|a| , |b|) = gcd(a, b), there is no harm in assuming a ≥ b > 0. Now dividing a by b and applying the division algorithm, we can state:

$$a = q_1 b + r_1 \qquad 0 \le r_1 < b \qquad \textbf{(4.2)}$$

- Let us now return to Equation (4.2) and assume that $r_1 \neq 0$. Because $b > r_1$, we can divide b by r_1 and apply the division algorithm to obtain:

$$b = q_2 r_1 + r_2 \qquad 0 \le r_2 < r_1$$

The result is the following system of equations:

$$\left.\begin{array}{ll} a = q_1 b + r_1 & 0 < r_1 < b \\ b = q_2 r_1 + r_2 & 0 < r_2 < r_1 \\ r_1 = q_3 r_2 + r_3 & 0 < r_3 < r_2 \\ \quad\vdots & \quad\vdots \\ r_{n-2} = q_n r_{n-1} + r_n & 0 < r_n < r_{n-1} \\ r_{n-1} = q_{n+1} r_n + 0 & \\ d = \gcd(a, b) = r_n & \end{array}\right\} \qquad \textbf{(4.3)}$$

Let us now look at an example with relatively large numbers to see the power of this algorithm:

To find d = gcd (a,b) = gcd (1160718174, 316258250)		
$a = q_1 b + r_1$	$1160718174 = 3 \times 316258250 + 211943424$	d = gcd(316258250, 211943424)
$b = q_2 r_1 + r_2$	$316258250 = 1 \times 211943424 + 104314826$	d = gcd(211943424, 104314826)
$r_1 = q_3 r_2 + r_3$	$211943424 = 2 \times 104314826 + 3313772$	d = gcd(104314826, 3313772)
$r_2 = q_4 r_3 + r_4$	$104314826 = 31 \times 3313772 + 1587894$	d = gcd(3313772, 1587894)
$r_3 = q_5 r_4 + r_5$	$3313772 = 2 \times 1587894 + 137984$	d = gcd(1587894, 137984)
$r_4 = q_6 r_5 + r_6$	$1587894 = 11 \times 137984 + 70070$	d = gcd(137984, 70070)
$r_5 = q_7 r_6 + r_7$	$137984 = 1 \times 70070 + 67914$	d = gcd(70070, 67914)
$r_6 = q_8 r_7 + r_8$	$70070 = 1 \times 67914 + 2156$	d = gcd(67914, 2156)
$r_7 = q_9 r_8 + r_9$	$67914 = 31 \times 2516 + 1078$	d = gcd(2156, 1078)
$r_8 = q_{10} r_9 + r_{10}$	$2156 = 2 \times 1078 + 0$	d = gcd(1078, 0) = 1078
Therefore, d = gcd(1160718174, 316258250) = 1078		

In this example, we begin by dividing 1160718174 by 316258250, which gives 3 with a remainder of 211943424. Next we take 316258250 and divide it by 211943424. The process continues until we get a remainder of 0, yielding a result of 1078.

2.5. CONGRUENCE AND MATRICES

2.6. GROUPS, RINGS, FIELDS

Contents

- **Groups**
  - A1- Closure
  - A2 - Associative
  - A3 - Identity
  - A4 - Inverse
  - A5 - Commutative
- **Rings**
  - M1- Closure under multiplication
  - M2 - Associativity of multiplication
  - M3 - Distributive law
  - M4 – Commutativity of multiplication
  - M5 – Multiplicative Identity
  - M6 – No zero divisors
- **Fields**
  - M7 – Multiplicative Inverse

Groups, rings, and fields are the fundamental elements of a branch of mathematicsknown as abstract algebra, or modern algebra.

Groups

- A **group** G, sometimes denoted by $\{G, \cdot\}$, is a set of elements with a binary operationdenoted by $\cdot$ that associates to each ordered pair (a, b) of elements in G an element$(a \cdot b)$ in G, such that the following axioms are obeyed:

(A1) Closure:	If a and b belong to G, then $a \cdot b$ is also in G.
(A2) Associative:	$a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for all a, b, c in G.
(A3) Identity element:	There is an element e in G such that $a \cdot e = e \cdot a = a$ for all a in G.
(A4) Inverse element:	For each a in G, there is an element a' in G such that $a \cdot a' = a' \cdot a = e$.

- If a group has a finite number of elements, it is referred to as a **finite group**, and the **order** of the group is equal to the number of elements in the group. Otherwise, the group is an **infinite group**.
- A group is said to be **abelian** if it satisfies the following additional condition:

(A5) Commutative:	$a \cdot b = b \cdot a$ for all a, b in G.

Rings

- A **ring** R, sometimes denoted by $\{R, +, *\}$, is a set of elements with two binary operations, called ***addition* and *multiplication***, such that for all a, b, c in R the following axioms are obeyed.

(A1–A5)	R is an abelian group with respect to addition; that is, R satisfies axioms A1 through A5. For the case of an additive group, we denote the identity element as 0 and the inverse of a as $-a$.
(M1) Closure under multiplication:	If a and b belong to R, then ab is also in R.
(M2) Associativity of multiplication:	$a(bc) = (ab)c$ for all a, b, c in R.
(M3) Distributive laws:	$a(b + c) = ab + ac$ for all a, b, c in R. $(a + b)c = ac + bc$ for all a, b, c in R.

- A ring is said to be **commutative** if it satisfies the following additional condition:

(M4) Commutativity of multiplication:	$ab = ba$ for all a, b in R.

- Next, we define an **integral domain**, which is a commutative ring that obeys the following axioms.

(M5) Multiplicative identity:	There is an element 1 in R such that $a1 = 1a = a$ for all a in R.
(M6) No zero divisors:	If a, b in R and $ab = 0$, then either $a = 0$ or $b = 0$.

Fields:

- A **field** F, sometimes denoted by $\{F, +, *\}$, is a set of elements with two binary operations, called *addition* and *multiplication*, such that for all a, b, c in F the following axioms are obeyed.

(A1–M6) F is an integral domain; that is, F satisfies axioms A1 through A5 and M1 through M6.

(M7) Multiplicative inverse: For each a in F, except 0, there is an element a^{-1} in F such that $aa^{-1} = (a^{-1})a = 1$.

2.7. FINITE FIELDS

SYMMETRIC KEY CIPHERS

2.8. SDES

Contents

Introduction

- Proposed by NIST in 1977.
- It is a block cipher and encrypts 64 bits data using 56 bit key.

DES Encryption:

- There are two inputs to the encryption function: the plaintext to be encrypted and the key. In this case, the plaintext must be **64 bits in length and key is 56** in length.
- Looking at the left-hand side of the figure, we can see that the processing of the plaintext proceeds in **three phases.**
- First, the 64-bit plaintext passes through an **initial permutation (IP)** that rearranges the bits to produce the **permuted input.**
- This is followed by a phase consisting of **sixteen rounds** of the same function, which involves both permutation and substitution functions.
- The output of the last (sixteenth) round consists of 64 bits that are a function of the input plaintext and the key.
- The left and right halves of the output are swapped to produce the **pre output.**
- Finally, the preoutput is passed through a permutation [IP -1] that is the inverse of the initial permutation function, to produce the 64-bit **ciphertext.** With the exception of the initial and final permutations, DES has the exact structure of a **Feistel Cipher.**

- The right-hand portion of Figure shows the way in which the **56-bit key is used.**
- Initially, the key is passed through a permutation function.
- Then, for each of the sixteen rounds, a **subkey (Ki)** is produced by the combination of a left circular shift and a permutation.
- The permutation function is the same for each round, but a different subkey is produced because of the repeated shifts of the key bits.

DES Decryption

- As with any **Feistel cipher**, decryption uses the same algorithm as encryption, except that the application of the subkeys is reversed. Additionally, the initial and final permutations are reversed.

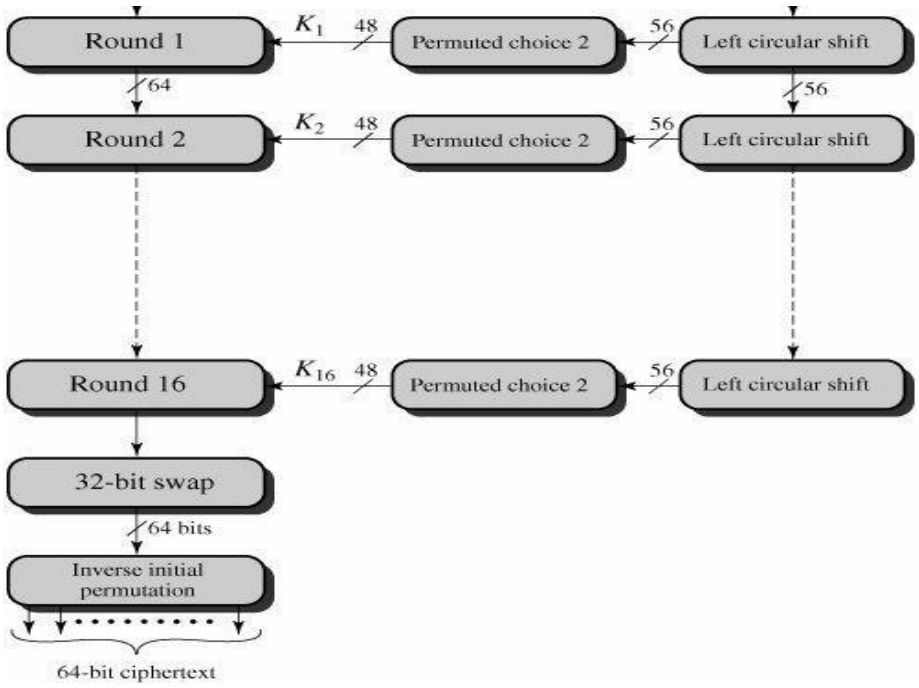


Fig .General Depiction of DES Encryption Algorithm

The Avalanche Effect

- A desirable property of any encryption algorithm is that a small change in either the plaintext or the key should produce a significant change in the cipher text.
- In particular, a change in one bit of the plaintext or one bit of the key should produce a change in many bits of the cipher text.
- This is referred to as the avalanche effect.

DES Round structure.

- ✓ Uses two 32 bit L & R halves.
- ✓ As in any classic Feistel cipher, the overall processing at each round can be summarized in the following formulas:

Li = Ri-1
Ri= Li 1 { F(Ri 1, Ki)

- ✓ The round key is 48 bits. The input is 32 bits.
- ✓ This input is first expanded to 48 bits by using a table that defines a permutation plus an expansion that involves duplication of 16 of the bits.
- ✓ The resulting 48 bits are XOR ed with. This 48-bit result passes through a substitution function that produces a 32-bit output, which is permuted as defined by table.
- ✓ The role of the S-boxes in the function F is illustrated in Figure.
- ✓ The substitution consists of a set of eight S-boxes, each of which accepts 6 bits as input and Produces 4 bits as output.

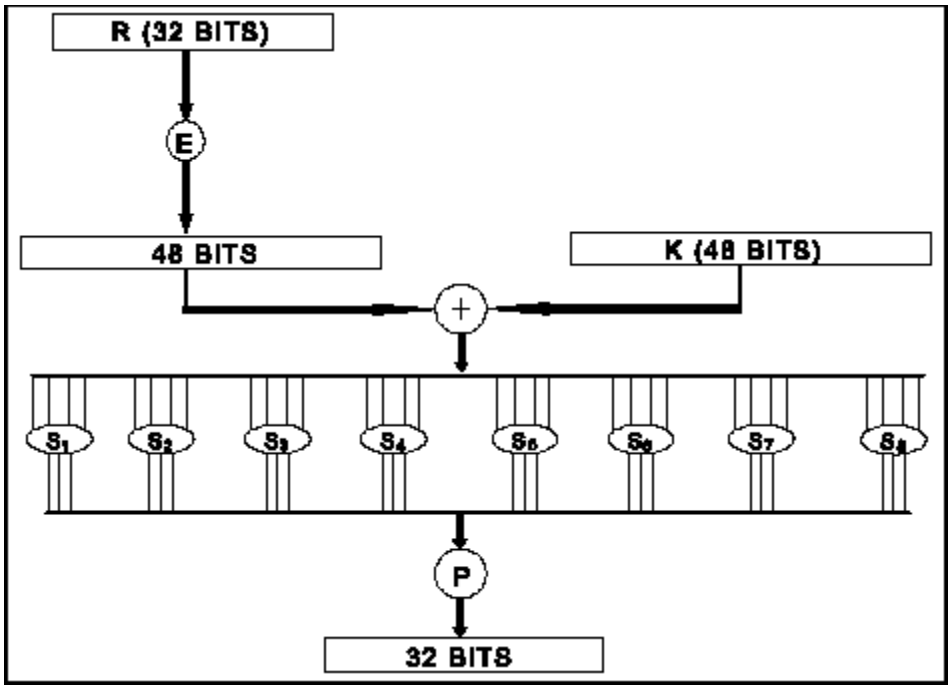


Fig: Calculation of F(R, K)

2.9. STRENGTH OF DES

- The Use of 56-Bit Keys
- The Nature of the DES Algorithm
- Timing Attacks

2.9. DIFFERENTIAL AND LINEAR CRYPTANALYSIS

2.10. BLOCK CIPHER DESIGN PRINCIPLES

– Block cipher Principles of DES

Contents
• Introduction • Number of rounds • Design of function • Key Schedule Algorithm

Introduction

- There are three critical aspects of block cipher design: the number of rounds, design of the function F, and key scheduling

Number of Rounds

- The greater the number of rounds, the more difficult it is to perform cryptanalysis, even for a relatively weak F.
- In general, the criterion should be that the number of rounds is chosen so that known cryptanalytic efforts require greater effort than a simple **brute-force key search attack.**
- The differential cryptanalysis attack requires 255.1 operations, whereas brute force requires 255.
- If DES had 15 or fewer rounds, differential cryptanalysis would require less effort than a brute-force key search.
- This criterion is attractive, because it makes it easy to judge the strength of an algorithm and to compare different algorithms.
- In the absence of a cryptanalytic breakthrough, the strength of any algorithm that satisfies the criterion can be judged solely on key length.

Design of Function F

- The heart of a **Feistel block cipher** is the function F, which provides the element of confusion in a Feistel cipher. Thus, it must be difficult to **"unscramble**" the substitution performed by F.
- One obvious criterion is that F be **nonlinear**. The more nonlinear F, the more difficult any type of cryptanalysis will be.
- The more difficult it is to approximate F by a set of linear equations, the more nonlinear F is. Several other criteria should be considered in designing F.
- We would like the algorithm to have good avalanche properties.
- A more stringent version of this is the **strict avalanche criterion (SAC)**, which states that any output bit *j* of an S-box should change with probability 1/2 when any single input bit *i* is inverted for all *i*, *j*.

- Although SAC is expressed in terms of S-boxes, a similar criterion could be applied to F as a whole. This is important when considering designs that do not include S-boxes.
- Another criterion proposed is the **bit independence criterion (BIC)**, which states that output bits *j* and *k* should change independently when any single input bit *i* is inverted for all *i*, *j*, and *k*. The SAC and BIC criteria appear to strengthen the effectiveness of the confusion function.

Key Schedule Algorithm

- With **any Feistel block cipher**, the key is used to generate one subkey for each round.
- In general, we would like to select subkeys to maximize the difficulty of deducing individual subkeys and the difficulty of working back to the main key.
- No general principles for this have yet been promulgated.
- At minimum, the key schedule should **guarantee key/ciphertext Strict Avalanche Criterion** and **Bit Independence Criterion.**

2.11. BLOCK CIPHER MODE OF OPERATION

Contents
• Electronic Code Book • Cipher Block Chaining Mode • Cipher Feedback Mode • Output Feedback Mode • Counter Mode

Electronic Code Book

- The simplest mode is the **Electronic codebook (ECB) mode**, in which plaintext is handled one block at a time and each block of plaintext is encrypted using the same key (Figure 6.3).
- The term codebook is used because, for a given key, there is a **unique ciphertext for every b-bit block of plaintext**.

Table 6.1 Block Cipher Modes of Operation

Mode	Description	Typical Application
Electronic Codebook (ECB)	Each block of plaintext bits is encoded independently using the same key.	• Secure transmission of single values (e.g., an encryption key)
Cipher Block Chaining (CBC)	The input to the encryption algorithm is the XOR of the next block of plaintext and the preceding block of ciphertext.	• General-purpose block-oriented transmission • Authentication
Cipher Feedback (CFB)	Input is processed s bits at a time. Preceding ciphertext is used as input to the encryption algorithm to produce pseudorandom output, which is XORed with plaintext to produce next unit of ciphertext.	• General-purpose stream-oriented transmission • Authentication
Output Feedback (OFB)	Similar to CFB, except that the input to the encryption algorithm is the preceding encryption output, and full blocks are used.	• Stream-oriented transmission over noisy channel (e.g., satellite communication)
Counter (CTR)	Each block of plaintext is XORed with an encrypted counter. The counter is incremented for each subsequent block.	• General-purpose block-oriented transmission • Useful for high-speed requirements

- For a message longer than b bits, the procedure is simply to break the message into b-bit blocks, padding the last block if necessary.
- **Decryption is performed one block at a time**, always using the same key.
- We can define ECB mode as follows.

ECB	$C_j = E(K, P_j) \qquad j = 1, \ldots, N$	$P_j = D(K, C_j) \qquad j = 1, \ldots, N$

- The ECB method is ideal for a **short amount of data**, such as an encryption key.
- For lengthy messages, the ECB mode may not be secure. If the message is highly structured, it may be possible for a cryptanalyst to exploit these regularities.

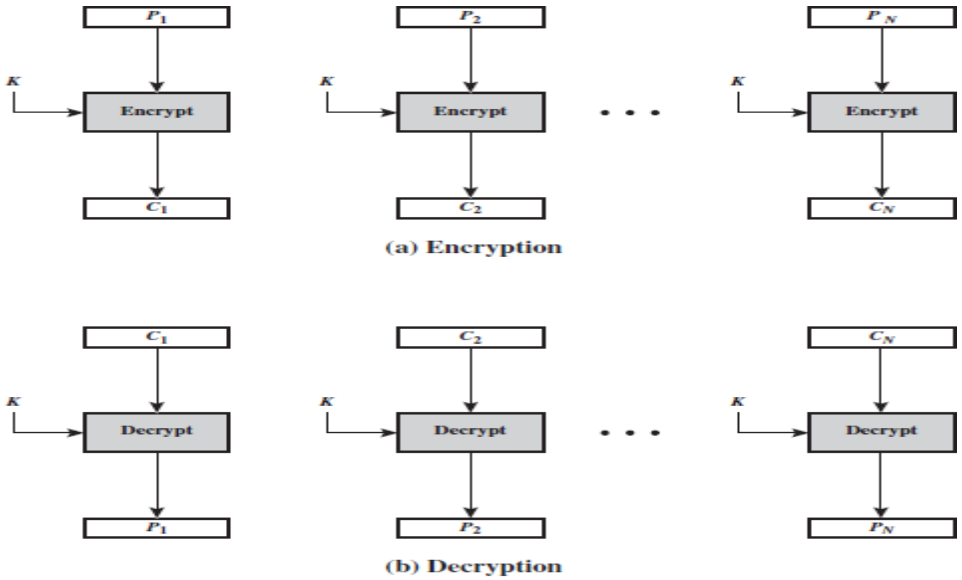


Figure 6.3 Electronic Codebook (ECB) Mode

Cipher Block Chaining Mode

- In this scheme, the input to the encryption algorithm is the XOR of the current plaintext block and the preceding ciphertext block; the same key is used for each block.
- In effect, we have chained together the processing of the sequence of plaintext blocks.
- The input to the encryption function for each plaintext block bears no fixed relationship to the plaintext block. Therefore, repeating patterns of b bits are not exposed.
- The result is XORed with the preceding ciphertext block to produce the plaintext block. To see that this works, we can write

$$C_j = \mathrm{E}(K, [C_{j-1} \oplus P_j])$$

Then

$$\mathrm{D}(K, C_j) = \mathrm{D}(K, E(K, [C_{j-1} \oplus P_j]))$$
$$\mathrm{D}(K, C_j) = C_{j-1} \oplus P_j$$
$$C_{j-1} \oplus \mathrm{D}(K, C_j) = C_{j-1} \oplus C_{j-1} \oplus P_j = P_j$$

- To produce the first block of ciphertext, an initialization vector (IV) is XORed with the first block of plaintext. On decryption, the IV is XORed with the output of the decryption algorithm to recover the first block of plaintext.
- We can define CBC mode as

CBC	$C_1 = \mathrm{E}(K, [P_1 \oplus \mathrm{IV}])$ $C_j = \mathrm{E}(K, [P_j \oplus C_{j-1}])\ j = 2, \ldots, N$	$P_1 = \mathrm{D}(K, C_1) \oplus \mathrm{IV}$ $P_j = \mathrm{D}(K, C_j) \oplus C_{j-1}\ j = 2, \ldots, N$

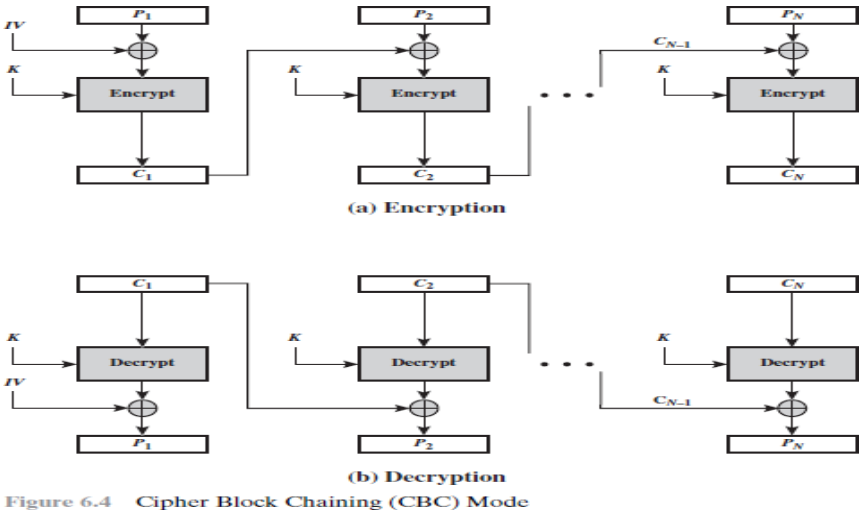


Figure 6.4 Cipher Block Chaining (CBC) Mode

Cipher Feedback Mode

- As with CBC, the units of plaintext are chained together, so that the cipher text of any plaintext unit is a function of all the preceding plaintext.
- In this case, rather than blocks of b bits, the plaintext is divided into segments of s bits.
- First, consider encryption. The input to the encryption function is a b-bit shift register that is initially set to some initialization vector (IV).
- **The leftmost (most significant) s** bits of the output of the encryption function are XORed with the first segment of plaintext P1 to produce the first unit of ciphertext C1, which is then transmitted.
- In addition, the contents of the shift register are shifted left by s bits, and C1 is placed in the rightmost (least significant) s bits of the shift register.
- This process continues until all plaintext units have been encrypted. For decryption, the same scheme is used, except that the received ciphertext unit is XORed with the output of the encryption function to produce the plaintext unit
- This is easily explained. Let MSBs(X) be defined as the **most significant s bits of X.** Then

$$C_1 = P_1 \oplus \text{MSB}_s[\text{E}(K, \text{IV})]$$

Therefore, by rearranging terms:

$$P_1 = C_1 \oplus \text{MSB}_s[\text{E}(K, \text{IV})]$$

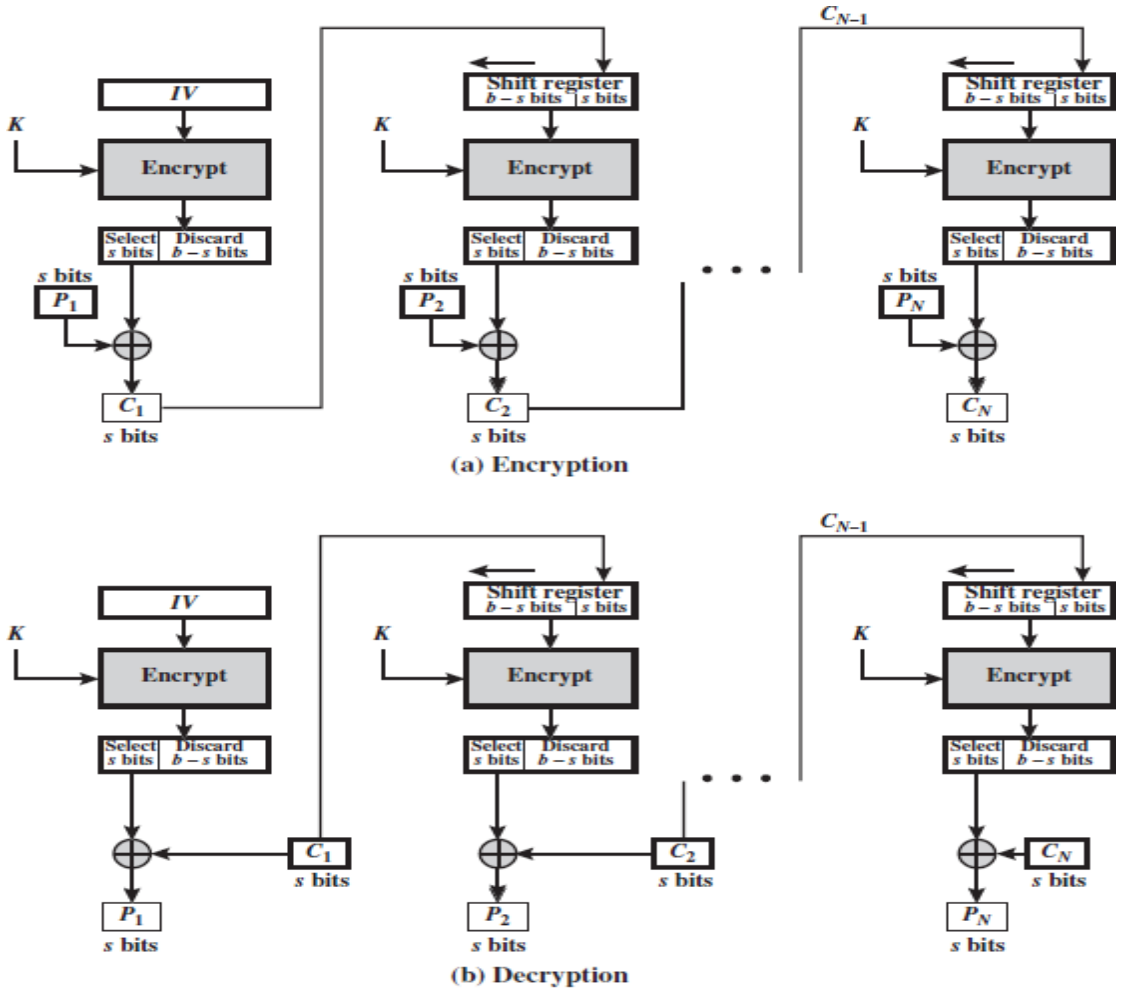


Figure 6.5 *s*-bit Cipher Feedback (CFB) Mode

We can define CFB mode as follows.

CFB	$I_1 = IV$ $I_j = \text{LSB}_{b-s}(I_{j-1}) \| C_{j-1}$ $j = 2, \ldots, N$ $O_j = \text{E}(K, I_j)$ $j = 1, \ldots, N$ $C_j = P_j \oplus \text{MSB}_s(O_j)$ $j = 1, \ldots, N$	$I_1 = IV$ $I_j = \text{LSB}_{b-s}(I_{j-1}) \| C_{j-1}$ $j = 2, \ldots, N$ $O_j = \text{E}(K, I_j)$ $j = 1, \ldots, N$ $P_j = C_j \oplus \text{MSB}_s(O_j)$ $j = 1, \ldots, N$

Output feedback (OFB) mode

- The output feedback (OFB) mode is similar in structure to that of CFB. For OFB, the output of the encryption function is feed back to become the input for encrypting the next block of plaintext (Figure 6.6).
- The other difference is that the OFB mode operates on full blocks of plaintext and ciphertext, whereas CFB operates on an s-bit subset. OFB encryption can be expressed as

$$C_j = P_j \oplus \mathrm{E}(K, O_{j-1})$$

where

$$O_{j-1} = \mathrm{E}(K, O_{j-2})$$

Some thought should convince you that we can rewrite the encryption expression as:

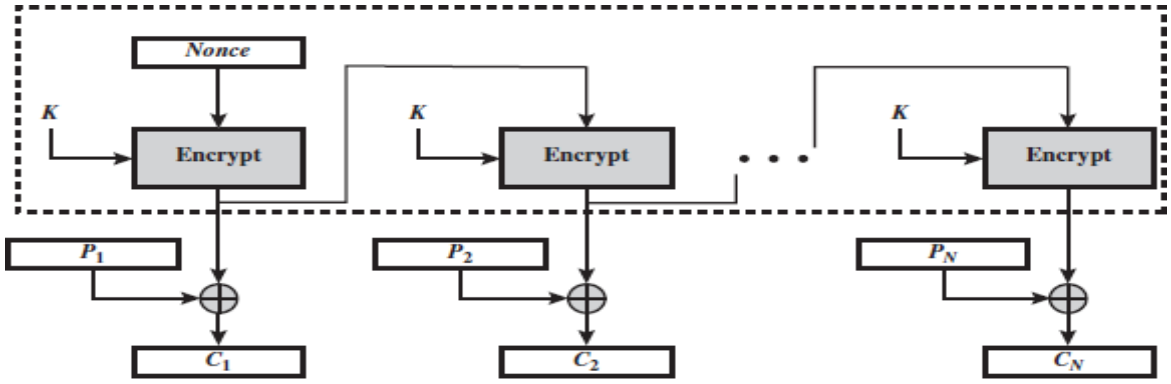
$$C_j = P_j \oplus \mathrm{E}(K, [C_{j-1} \oplus P_{j-1}])$$

By rearranging terms, we can demonstrate that decryption works.

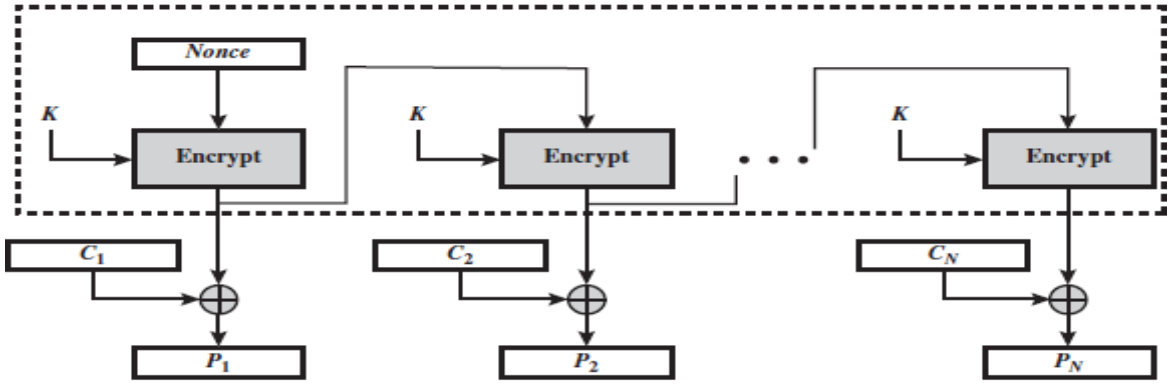
$$P_j = C_j \oplus \mathrm{E}(K, [C_{j-1} \oplus P_{j-1}])$$

We can define OFB mode as follows.

OFB	$I_1 = Nonce$ $I_j = O_{j-1}$ $\quad j = 2, \ldots, N$ $O_j = \mathrm{E}(K, I_j)$ $\quad j = 1, \ldots, N$ $C_j = P_j \oplus O_j$ $\quad j = 1, \ldots, N-1$ $C_N^* = P_N^* \oplus \mathrm{MSB}_u(O_N)$	$I_1 = Nonce$ $I_j = O_{j-1}$ $\quad j = 2, \ldots, N$ $O_j = \mathrm{E}(K, I_j)$ $\quad j = 1, \ldots, N$ $P_j = C_j \oplus O_j$ $\quad j = 1, \ldots, N-1$ $P_N^* = C_N^* \oplus \mathrm{MSB}_u(O_N)$



(a) Encryption



(b) Decryption

Figure 6.6 Output Feedback (OFB) Mode

Counter Mode

- Although interest in the **counter (CTR**) mode has increased recently with applications to ATM **(asynchronous transfer mode) network security and IP sec (IP security**), this mode was proposed early on (e.g., [DIFF79]).
- Figure 6.7 depicts the CTR mode. A counter equal to the plaintext block size is used.
- Typically, the counter is initialized to some value and then incremented by 1 for each subsequent block (modulo 2b, where b is the block size).
- For encryption, the counter is encrypted and then XORed with the plaintext block to produce the ciphertext block; there is no chaining.
- For decryption, the same sequence of counter values is used, with each encrypted counter XORed with a ciphertext block to recover the corresponding plaintext block. Thus, the initial counter value must be made available for decryption. Given a sequence of counters T1, T2, c, TN, we can define CTR mode as follows.

CTR	$C_j = P_j \oplus \mathrm{E}(K, T_j) \quad j = 1, \ldots, N-1$ $C_N^* = P_N^* \oplus \mathrm{MSB}_u[\mathrm{E}(K, T_N)]$	$P_j = C_j \oplus \mathrm{E}(K, T_j) \quad j = 1, \ldots, N-1$ $P_N^* = C_N^* \oplus \mathrm{MSB}_u[\mathrm{E}(K, T_N)]$

2.12. EVALUATION CRITERIA FOR AES

2.13. ADVANCED ENCRYPTION STANDARD

Finite Field Arithmetic

- In AES, all operations are performed on 8-bit bytes. In particular, the arithmetic operations of addition, multiplication, and division are performed over the finite field.
- In essence, a field is a set in which we can do **addition, subtraction, multiplication, and division** without leaving the set.
- Division is defined with the following rule: ***a*/*b* = *a*(*b*-1).**

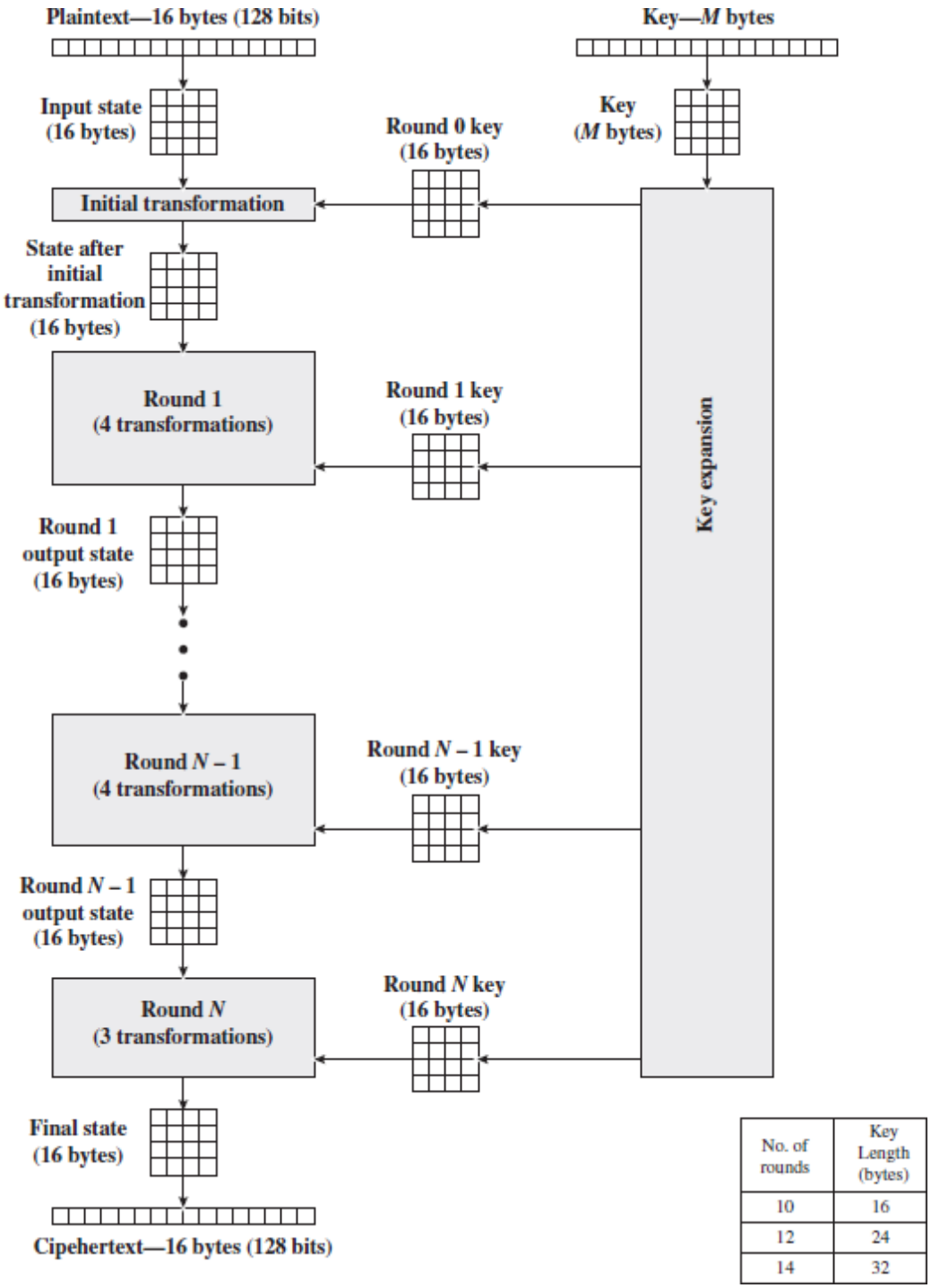
AES Structure

- General Structure
- Detailed Structure

General Structure

- Figure(5.1). shows the overall structure of the **AES encryption process**.
- The cipher takes a plaintext block size of **128 bits, or 16 bytes**.
- The key length can be 16, 24, or 32 bytes (128, 192, or 256 bits). The algorithm is referred to as AES-128, AES-192, or AES-256, depending on the key length.

- The input to the encryption and decryption algorithms is a single 128-bit block. In FIPS PUB 197, this block is depicted as a 4 * 4 square matrix of bytes. This block is copied into the **State** array, which is modified at each stage of **encryption or decryption**.
- After the final stage, **State** is copied to an output matrix.
- This key is then expanded into an array of key schedule words.
- Each word is four bytes, and the total key schedule is 44 words for the 128-bit key
- The **cipher** consists of *N* rounds, where the number of rounds depends on the key length: 10 rounds for a 16-byte key, 12 rounds for a 24-byte key, and 14 rounds for a 32-byte key.
- The first *N* - 1 rounds consist of four distinct transformation functions: **SubBytes, ShiftRows, MixColumns, and AddRoundKey**, which are described subsequently.
- The final round contains only three transformations, and there is a initial single transformation **(AddRoundKey)** before the first round, which can be considered Round 0. Each transformation takes one or more 4 * 4 matrices as input and produces a 4 * 4 matrix as output.



No. of rounds	Key Length (bytes)
10	16
12	24
14	32

Fig : AES Encryption Process

Table 5.1 AES Parameters

Key Size (words/bytes/bits)
Plaintext Block Size (words/bytes/bits)
Number of Rounds
Round Key Size (words/bytes/bits)
Expanded Key Size (words/bytes)

Detailed Structure

Figure (5.1) shows the AES cipher in more detail, indicating the sequence of transformations in each round and showing the corresponding decryption function.

We can make several comments about the overall AES structure.

1. One noteworthy feature of this structure is that it **is not a Feistel structure**
2. The key that is provided as input is expanded into an array of forty-four 32-bit words, **w**[*i*]. Four

   distinct words (128 bits) serve as a round key for each round; these are indicated in Figure 5.**3**

3. **Four different stages are used, one of permutation and three of substitution:**

• **Substitute bytes:** Uses an S-box to perform a byte-by-byte substitution of the block

• **ShiftRows:** A simple permutation

• **MixColumns:** A substitution that makes use of arithmetic over GF(28)

• **AddRoundKey:** A simple bitwise XOR of the current block with a portion of the expanded Key.

4. The structure is **quite simple**.

5. Only the **AddRoundKey** stage makes use of the key. For this reason, the cipher begins and ends with an AddRoundKey stage.

6. The AddRoundKey stage is, in effect, a form of Vernam cipher and by itself would not be formidable. The other three stages together provide confusion, diffusion, and nonlinearity, but by themselves would provide no security because they do not use the key.

7. Each stage is easily reversible.

8. Once it is established that all four stages are reversible, it is easy to verify that decryption does recover the plaintext.

9. The final round of both encryption and decryption consists of only three stages.Again, this is a consequence of the particular structure of AES and is required to make the cipher reversible

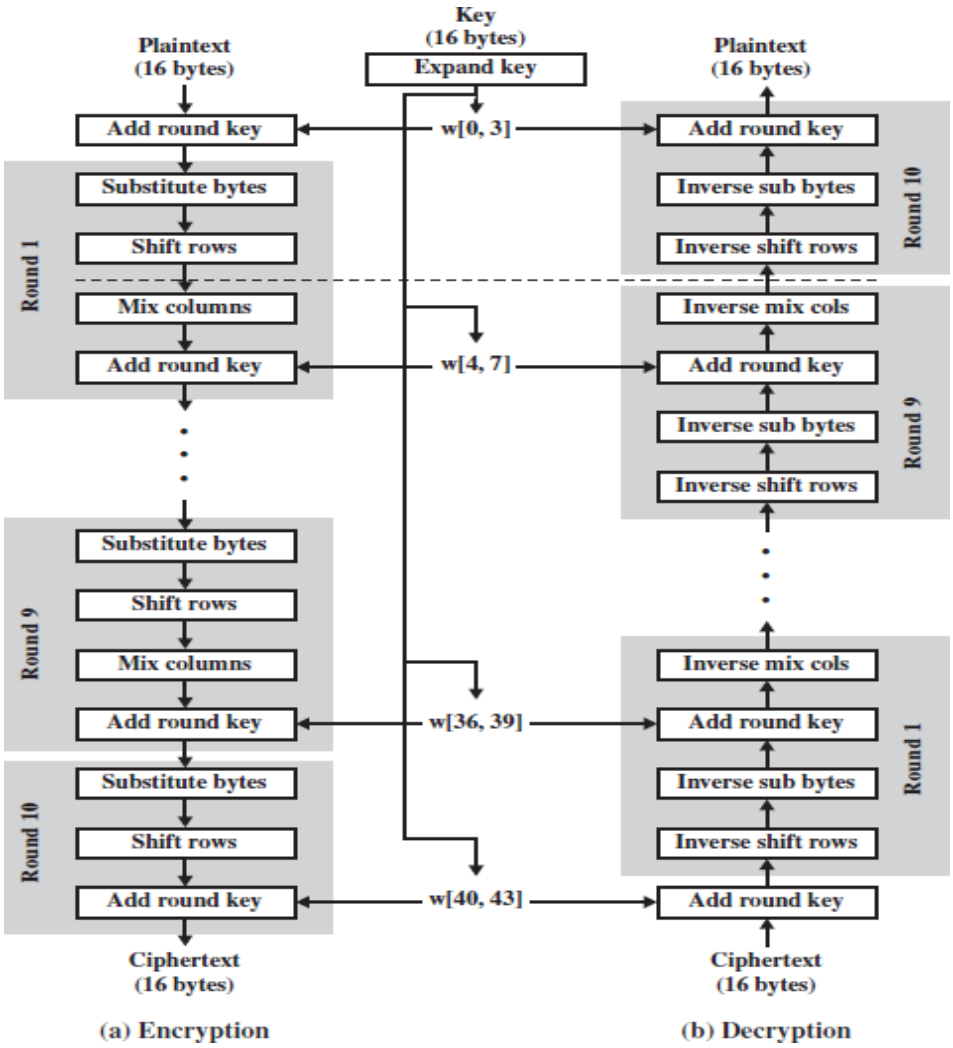


Fig (5.3): AES Encryption and Decryption

AES Transformation Functions

The four transformations used in AES. For each stage, we describe the forward (encryption) algorithm, the inverse (decryption) algorithm, and the rationale for the stage.

- Substitute Bytes Transformation
- Shift Rows Transformation
- Mix Columns Transformation
- AddRoundKey Transformation

Substitute Bytes Transformation

- The **forward substitute byte transformation**, called SubBytes, is a simple table lookup. AES defines a 16 * 16 matrix of byte values, called **an S-box**, that contains a permutation of all possible 256 8-bit values.
- Each individual byte of **State is mapped into a new byte in the following way:**

- ❖ The leftmost 4 bits of the byte are used as a row value and the rightmost 4 bits are used as a column value.
- ❖ These row and column values serve as indexes into the S-box to select a unique 8-bit output value.

Shift Rows Transformation

- ❖ The first row of State is not altered.
- ❖ For the second row, a 1-byte circular left shift is performed.
- ❖ For the third row, a 2-byte circular left shift is performed.
- ❖ For the fourth row, a 3-byte circular left shift is performed.
- ❖ The inverse shift row transformation, called InvShiftRows, performs the circular shifts in the opposite direction for each of the last three rows, with a 1-byte circular right shift for the second row, and so on.

MixColumns Transformation

- ❖ MixColumns, operates on each column individually.
- ❖ Each byte of a column is mapped into a new value that is a function of all four bytes in that column.

AddRoundKey Transformation

- ❖ AddRoundKey, the 128 bits of State are bitwise XORed with the 128 bits of the round key.

2.14. RC4

Contents
• **Characteristics** • **RC5 Parameters** • **Key Expansion** • **Encryption** • **Decryption** • **RC5 Modes**

RC5 is a symmetric encryption algorithm developed by Ron Rivest. RC5 was designed to have the **following characteristics:**

- **Suitable for hardware or software**
- **Fast**
- **Adaptable to processors of different word lengths**
- **Variable number of rounds:**
- **Variable-length key**

- **Simple** .
- **Low memory requirement**
- **High security**

❖ RC5 has been incorporated into RSA Data Security, Inc-'s major products, including BSAFE, JSAFE, and S/MAIL.

RC5 Parameters

RC5 is actually a family of encryption algorithms determined by three parameters, as follows:

Parameter	Definition	Allowable Values
w	Word size in bits. RC5 encrypts 2-word blocks	16, 32, 64
r	Number of rounds	0, 1, ..., 255
b	Number of 8-bit bytes (octets) in the secret key K	0, 1, ..., 255

Key Expansion

❖ RC5 performs a complex set of operations on the secret key to produce a total of *t* subkeys. Two subkeys are used in each round, and two subkeys are used on an additional operation that is not part of any round, so $t = 2r + 2$. Each subkey is one Word (w bits) in length.

❖ Figure 4-11 illustrates the technique used to generate subkeys; The subkeys are stored in a *t*-word array labeled S[0], S[1], ….,' S[t-1]. Using the parameters *r* and *w* as inputs, this array is initialized to a particular fixed pseudorandom bit pattern.

❖ Then the *b*-byte key, K[0…. *b* - 1], is converted into a *c*-word array L[0…. *c -1*]. On a little endian machine, this is accomplished by zeroing out the array *L* and copying the string *K* directly into the memory positions represented by L.

❖ If *b* is not an integer multiple of w, then a portion of L at the right end remains zero-Finally, a mixing operation is performed that applies the contents of *L* to the initialized value of S to produce a final value for the array S.

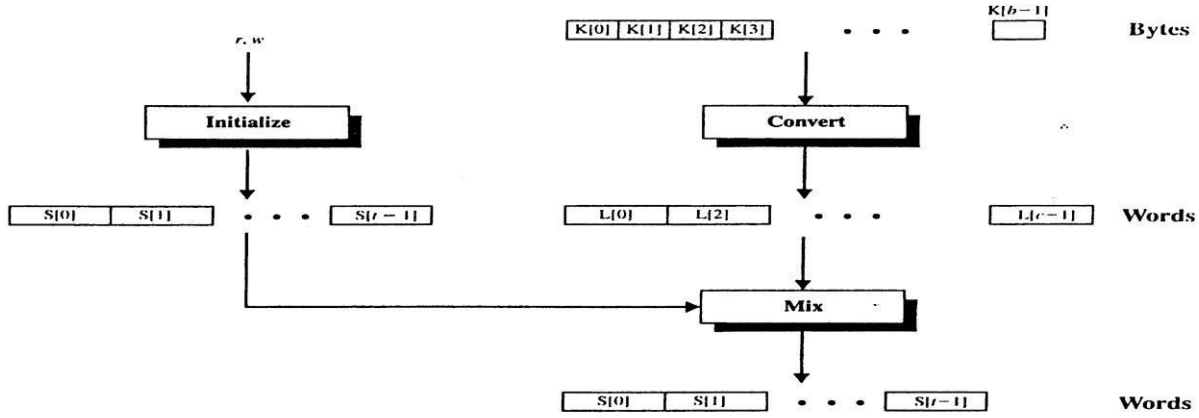


Figure 4.11 RC5 Key Expansion.

Let us look at this operations in detail. The initialize operation makes use of two word-length constants defined as follows,

$$P_w = \text{Odd}[(e-2)2^w]$$
$$Q_w = \text{Odd}[(\phi-1)2^w]$$

Where

$e = 2.718281828459 \ldots$ (base of natural logarithms)

$\phi = 1.618033988749 \ldots$ (golden ratio[3]) $= \left(\frac{1+\sqrt{5}}{2}\right)$

Encryption:

- RC5 uses three primitive operations (and their inverses):
  - **Addition:** Addition of words, denoted by +, is performed modulo 2^w. The inverse operation, denoted by -, is subtraction modulo 2^w.
  - **Bitwise exclusive-OR:** This operation is denoted by $\oplus$.
  - **Left circular rotation:** The cyclic rotation of word x left by y bits is denoted by $x <<< y$. The inverse is the right circular rotation of word x by y bits, denoted by $x >>> y$.

Figure 4-12a depicts the encryption operation. Note that this is not a classic Feistel structure. The plaintext is assumed to initially reside in the two w-bit registers A and B.

We use the variables LE_i and RE_i to refer to the left and right half of the data after round i has completed.

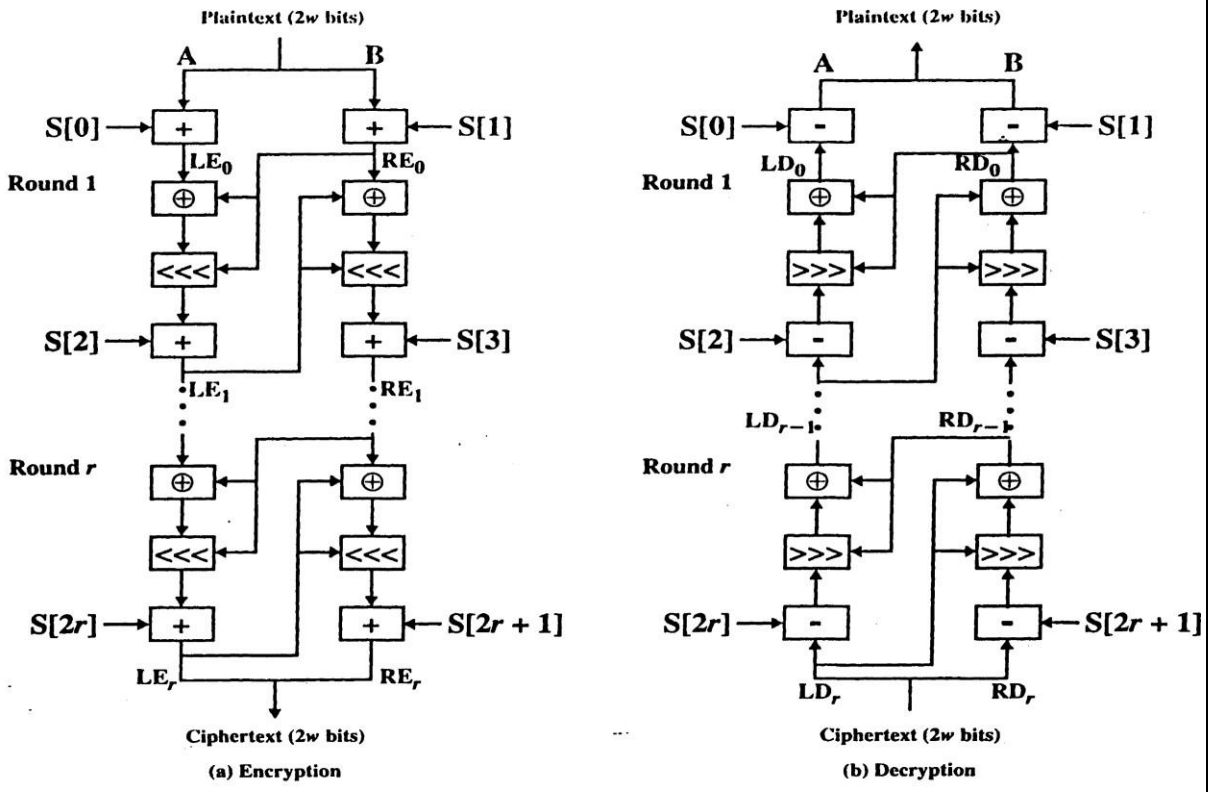


Figure 4.12 RC5 Encryption and Decryption.

Decryption

- Decryption, shown in Figure 4-12b, is easily derived from the encryption algorithm. In this case, the 2*w* bits of ciphertext are initially assigned to the two one-word variables LD_r, and RD_r.
- We use the variables LD_i and RD_i to refer to the left and right half of the data before round *i* has begun, where the rounds are numbered from *r* down to 1.

RC5 Modes:

To enhance the effectiveness of RC5 in interoperable implementations, RFC 2040 defines four different modes of operation:

- **RC5 block cipher:** This is the raw encryption algorithm that takes a fixed—size input block (2*w* bits) and produces a ciphertext block of the same length using a transformation that depends on a key.
- **RCS-CBC:** This is the cipher block chaining mode for RC5- CBC. CBC processes messages whose length is a multiple of the RC5 block size (multiples of 2w bits. CBC provides enhanced security compared to ECB because repeated blocks of plaintext produce different blocks of ciphertext.

- **RCS-CBC-Pad:** This is a CBC style of algorithm that handles plaintext of any length- The ciphertext will be longer than the plaintext by at most the size of a single RC5 block.
- **RCS-CTS:** This is the ciphertext stealing mode, which is also a CBC style of algorithm- This mode handles plaintext of any length and produces ciphertext of equal length.

The encryption sequence is as follows:

1. Encrypt the first $(N - 2)$ blocks using the traditional CBC technique.
2. Exclusive-OR P_{N-1} with the previous ciphertext block C_{N-2} to create Y_{N-1}.
3. Encrypt Y_{N-1} to create E_{N-1}.
4. Select the first L bytes of E_{N-1} to create C_N.
5. Pad P_N with zeros at the end and exclusive-OR with E_{N-1} to create Y_N.
6. Encrypt Y_N to create C_{N-1}.

2.15. KEY DISTRIBUTION

CS8792 CRYPTOGRAPHY AND NETWORK SECURITY

UNIT I
INTRODUCTION

Security trends – Legal, Ethical and Professional Aspects of Security, Need for Security at Multiple levels, Security Policies – Model of network security – Security attacks, services and mechanisms – OSI security architecture – Classical encryption techniques: substitution techniques, transposition techniques, steganography- Foundations of modern cryptography: perfect security – information theory – product cryptosystem – cryptanalysis.

1.1. SECURITY TRENDS

Internet Architecture Board (IAB) has issued report entitled "Security in the Internet Architecture" where they have identified key areas for security mechanisms. Among these were 1. need to secure network and 2. need to secure end to end transmission.

These concerns are fully justified. As confirmation, consider the trends reported by the Computer Emergency Response Team (CERT) Coordination Center (CERT/CC).

Figure 1.1 a shows the trend in Internet-related vulnerabilities reported to CERT over a 10-year period. These include security weaknesses in the operating systems of attached computers (e.g., Windows, Linux) as well as vulnerabilities in Internet routers and other network devices.

Figure 1.1 b shows the number of security-related incidents reported to CERT. These include denial of service attacks; IP spoofing, in which intruders create packets with false IP addresses and exploit applications that use authentication based on IP; and various forms of eavesdropping and packet sniffing, in which attackers read transmitted information, including logon information and database contents.

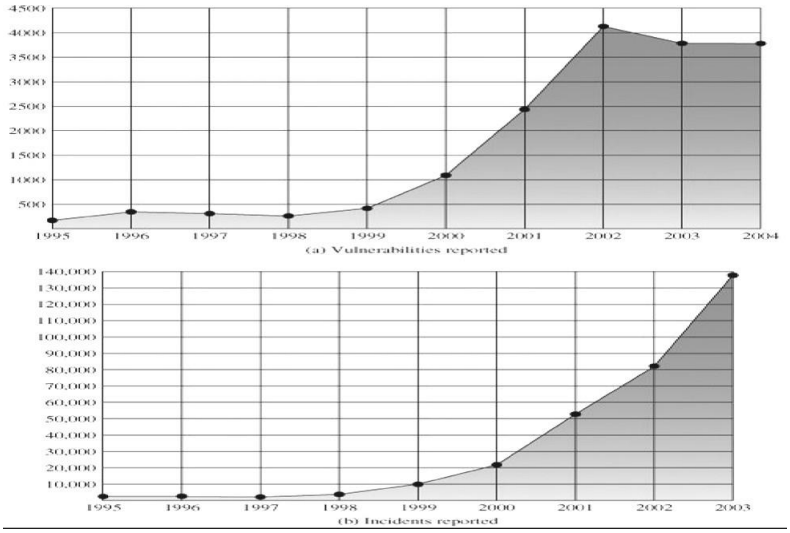


Figure 1.1 a - Trend in Internet-related vulnerabilities reported to CERT
1.1.b - Number of security-related incidents reported to CERT

Over time, the attacks on the Internet and Internet-attached systems have grown more sophisticated while the amount of skill and knowledge required to mount an attack has declined (Figure 1.2).

Attacks have become more automated and can cause greater amounts of damage. This increase in attacks coincides with an increased use of the Internet and with increases in the complexity of protocols, applications, and the Internet itself. Critical infrastructures increasingly rely on the Internet for operations. Individual users rely on the security of the Internet, email, the Web, and Web-based applications to a greater extent than ever. Thus, a wide range of technologies and tools are needed to counter the growing threat.

At a basic level, cryptographic algorithms for confidentiality and authentication assume greater importance. As well, designers need to focus on Internet-based protocols and the vulnerabilities of attached operating systems and applications.

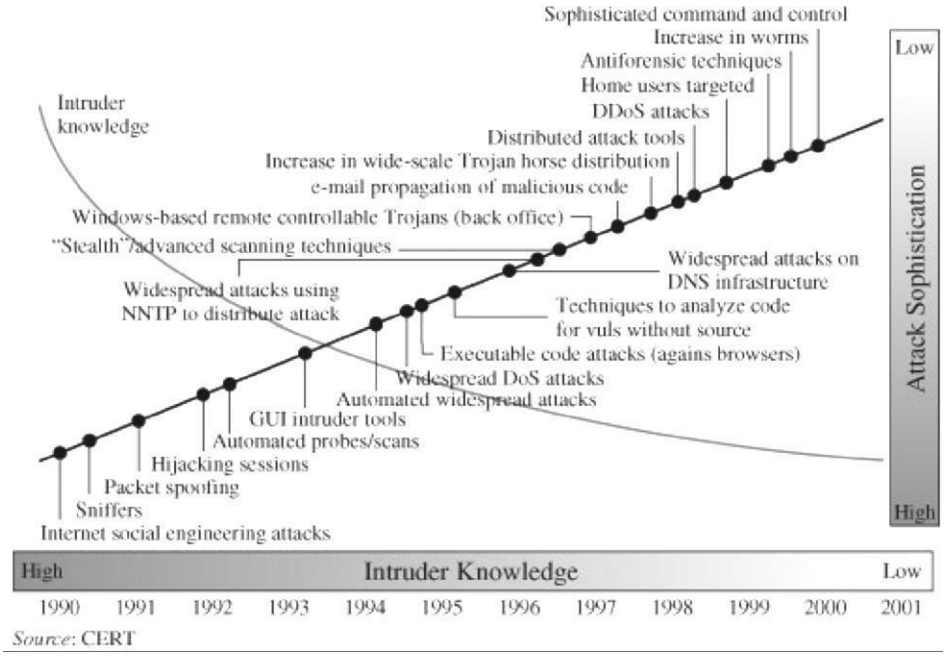


Figure 1.2 - CERT Statistics

1.2. LEGAL, ETHICAL AND PROFESSIONAL ASPECTS OF SECURITY

1.3. NEED FOR SECURITY AT MULTIPLE LEVELS

Multilevel security or multiple levels of security (MLS) is the application of a computer system to process information with incompatible classifications (i.e., at different security levels), permit access by users with different security clearances and needs-to-know, and prevent users from obtaining access to information for which they lack authorization. There are two contexts for the use of multilevel security. One is to refer to a system that is adequate to protect itself from subversion and has robust mechanisms to separate information domains, that is, trustworthy. Another context is to refer to an application of a computer that will require the computer to be strong enough to protect itself from subversion and possess adequate mechanisms to separate information domains, that is, a system we must trust. This distinction is important because systems that need to be trusted are not necessarily trustworthy.

1.4. SECURITY POLICIES

1.5. MODEL OF NETWORK SECURITY

- A message is to be transferred from one party to another across some sort of Internet service.

- A logical information channel is established by defining a route through the Internet from source to destination and by the cooperative use of communication protocols (e.g., TCP/IP) by the two principals.
- Security aspects come into play when it is necessary or desirable to protect the information transmission from an opponent who may present a threat to confidentiality, authenticity, and so on. All the techniques for providing security have two components:
    - A ***security-related transformation*** on the information to be sent.
    - Some ***secret information shared by the two principals*** and, it is hoped, unknown to the opponent. A trusted third party may be needed to achieve secure transmission.

This general model shows that there are four basic tasks in designing a particular security service:

1. **Design an algorithm** for performing the security-related transformation. The algorithm should be such that an opponent cannot defeat its purpose.
2. **Generate the secret information** to be used with the algorithm.
3. **Develop methods** for the distribution and sharing of the secret information.
4. **Specify a protocol** to be used by the two principals that makes use of the security algorithm and the secret information to achieve a particular security service.

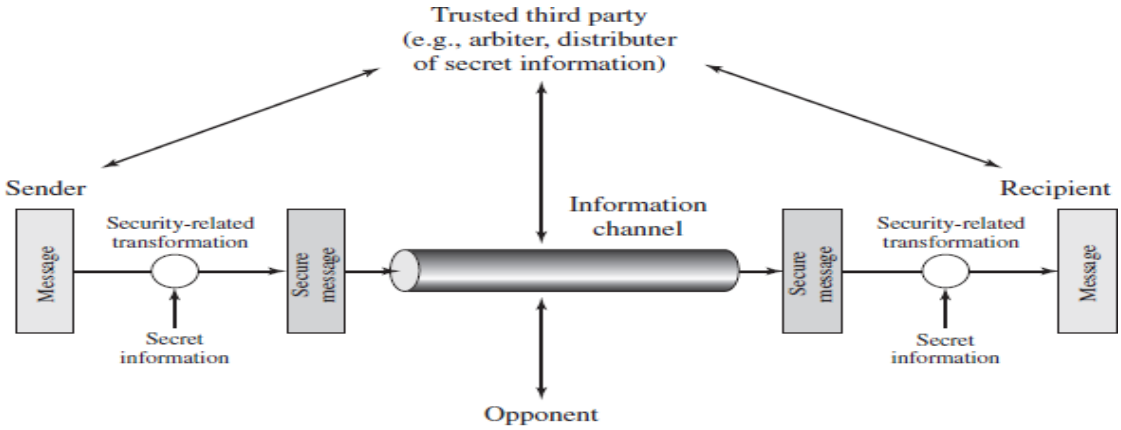


Figure 1.2 Model for Network Security

- A general model of these other situations is illustrated in Figure 1.3, which reflects a concern for protecting an information system from unwanted access.
- The hacker can be someone who, with no malign intent, simply gets satisfaction from breaking and entering a computer system.

Programs can present two kinds of threats:
Information access threats: Intercept or modify data on behalf of users who should not have access to that data.
Service threats: Exploit service flaws in computers to inhibit use by legitimate users.

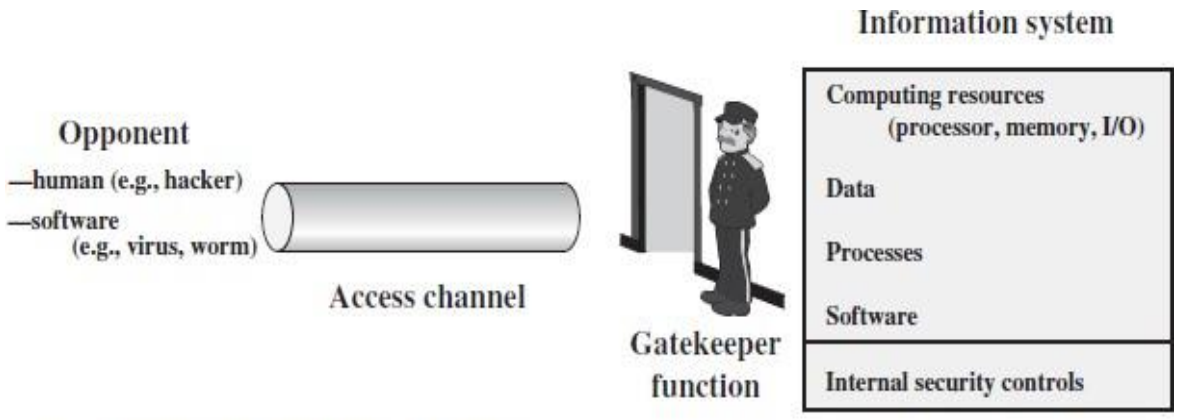


Figure 1.3 Network Access Security Model

- ***Viruses* and *worms*** are two examples of software attacks. They can also be inserted into a system across a network.

- The security mechanisms needed to cope with unwanted access fall into two broad categories (see Figure 1.3).
  - The first category might be termed **a gatekeeper function.** It includes password-based login procedures and screening logic that is designed to detect and reject worms, viruses.
  - The second line of defense consists of a **variety of internal controls** that monitor activity and analyze stored information in an attempt to detect the presence of unwanted intruders.

1.6. SECURITY ATTACKS

Contents

• **Introduction**
• **Passive Attacks**
  ○ the release of message contents and
  ○ traffic analysis.
• **Active Attacks**
  ○ masquerade,
  ○ replay,
  ○ modification of messages, and
  ○ denial of service.

Introduction

- security attacks, uses both in X.800 and RFC 2828, is in terms of ***passive attacks* and *active attacks.***
- A passive attack attempts to learn or make use of **information from the system but does not affect system resources.**
- An active attack attempts to alter **system resources or affect their operation.**

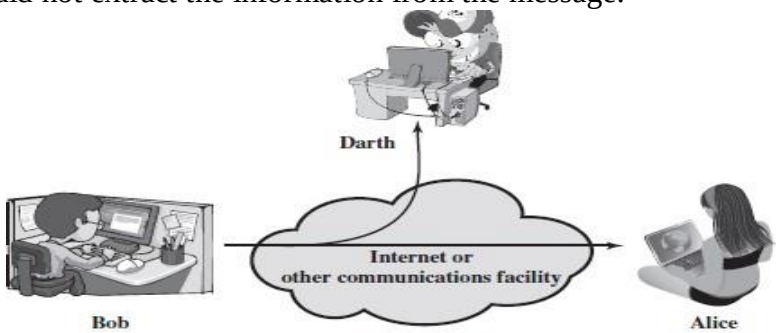
Passive Attacks

- Passive attacks (Figure 1.1) are in the nature of eavesdropping on, or monitoring of, transmissions.
- The goal of the opponent is to obtain information that is being transmitted.
- Two types of passive attacks are :

- the release of message contents and
- traffic analysis.

❖ The **release of message contents** is easily understood. A telephone conversation, an electronic mail message, and a transferred file may contain sensitive or confidential information.

❖ A **traffic analysis**, is subtler. Suppose that we had a way of masking the contents of messages or other information traffic so that opponents, even if they captured the message, could not extract the information from the message.



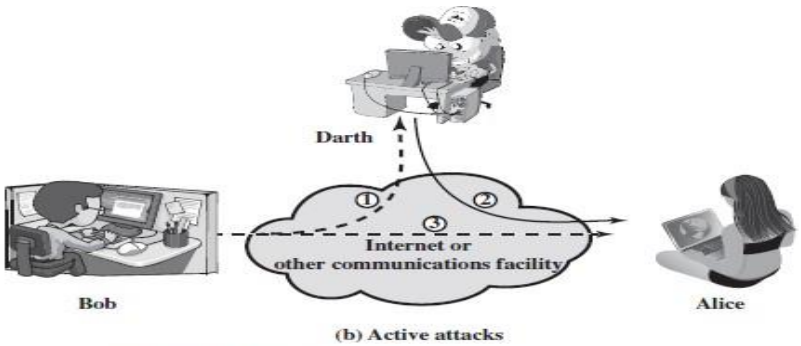
(a) Passive attacks

Active Attacks

❖ Active attacks (Figure 1.1b) involve some modification of the data stream or the creation of a false stream and can be subdivided into **four categories:**

- masquerade,
- replay,
- modification of messages, and
- denial of service.

- A **masquerade -** A masquerade attack usually includes one of the other forms of active attack.
- **Replay** involves the passive capture of a data unit and its subsequent retransmission to produce an unauthorized effect.
- **Modification of messages** simply means that some portion of a legitimate message is altered, or that messages are delayed or reordered, to produce an unauthorized effect.
- The **denial of service** prevents or inhibits the normal use or management of communications facilities.



(b) Active attacks

Figure 1.1 Security Attacks

1.7. SECURITY SERVICES

Contents
1. Introduction

2. Authentication
- Peer Entity Authentication
- Data Origin Authentication

3. Access Control

4. Data Confidentiality
- Connection Confidentiality
- Connectionless Confidentiality
- Selective-Field Confidentiality
- Traffic Flow Confidentiality

5. Data Integrity
- Connection Integrity with Recovery
- Connection Integrity without Recovery
- Selective-Field Connection Integrity
- Connectionless Integrity
- Selective-Field Connectionless Integrity

6. Nonrepudiation
- Nonrepudiation, Origin
- Nonrepudiation, Destination

7. Availability Service

Introduction

- Security services is defined as a processing or communication service that is provided by a system to give a specific kind of protection to system resources.
- X.800 divides these services into five categories and fourteen specific services.

1. Authentication

- The assurance that the communicating entity is the one that it claims to be.

Two specific authentication services are defined in X.800:

- **Peer Entity Authentication**
  Used in association with a logical connection to provide confidence in the identity of the entities connected.
- **Data Origin Authentication**
  In a connectionless transfer, provides assurance that the source of received data is as claimed.

2. Access Control

- The prevention of unauthorized use of a resource.

3. Data Confidentiality:

- The protection of data from unauthorized disclosure.
  - **Connection Confidentiality**
    The protection of all user data on a connection.
  - **Connectionless Confidentiality**
    The protection of all user data in a single data block
  - **Selective-Field Confidentiality**
    The confidentiality of selected fields within the user data on a connection or in a single data block.
  - **Traffic Flow Confidentiality**
    The protection of the information that might be derived from observation of traffic flows.

4. Data Integrity:

- ❖ The assurance that data received are exactly as sent by an authorized entity (i.e.,contain no modification, insertion, deletion, or replay).
  - o **Connection Integrity with Recovery**

  Provides for the integrity of all user data on a connection and detects any modification, with recovery attempted.
  - o **Connection Integrity without Recovery**

  As above, but provides only detection without recovery.
  - o **Selective-Field Connection Integrity**

  Provides for the integrity of selected fields within the user data of a data block transferred over a connection.
  - o **Connectionless Integrity**

  Provides for the integrity of a single connectionless data block.
  - o **Selective-Field Connectionless Integrity**

  Provides for the integrity of selected fields within a single connectionless data block;

5. Nonrepudiation

- ❖ Provides protection against denial by one of the entities involved in a communication of having participated in all or part of the communication.
  - o **Nonrepudiation, Origin**

    Proof that the message was sent by the specified party.
  - o **Nonrepudiation, Destination**

    Proof that the message was received by the specified party.

1.8. SECURITY MECHANISMS

Contents
• **Introduction** • **Encipherment** • **Digital Signature** • **Access Control** • **Data Integrity** • **Authentication Exchange** • **Traffic Padding** • **Routing Control** • **Notarization** • **Pervasive Security Mechanisms** • **Trusted Functionality** • **Security Label** • **Event Detection** • **Security Audit Trail**

Introduction

- ❖ The mechanisms are divided into those that are implemented in a specific protocol layer, such as TCP or an application-layer protocol, and those that are not specific to any particular protocol layer or security service.

Encipherment

- ❖ The use of mathematical algorithms to transform data into a form that is not readily intelligible.

Digital Signature

- Data appended to, or a cryptographic transformation of, a data unit that allows a recipient of the data unit to prove the source and integrity of the data unit and protect against forgery (e.g., by the recipient).

Access Control

- A variety of mechanisms that enforce access rights to resources.

Data Integrity

- A variety of mechanisms used to assure the integrity of a data unit or stream of data units.

Authentication Exchange

- A mechanism intended to ensure the identity of an entity by means of information exchange.

Traffic Padding

- The insertion of bits into gaps in a data stream to frustrate traffic analysis attempts.

Routing Control

- Enables selection of particular physically secure routes for certain data and allows routing changes, especially when a breach of security is suspected.

Notarization

- The use of a trusted third party to assure certain properties of a data exchange.

Pervasive Security Mechanisms

- Mechanisms those are not specific to any particular OSI security service or protocol layer.

Trusted Functionality

- That which is perceived to be correct with respect to some criteria (e.g., as established by a security policy).

Security Label

- The marking bound to a resource (which may be a data unit) that names or designates the security attributes of that resource.

Event Detection

- Detection of security-relevant events.

Security Audit Trail

- Data collected and potentially used to facilitate a security audit, which is an independent review and examination of system records and activities.
- A reversible encipherment mechanism is simply an encryption algorithm that allows data to be encrypted and subsequently decrypted.
- Irreversible encipherment mechanisms include hash algorithms and message authentication codes, which are used in digital signature and message authentication applications.

1.9. OSI SECURITY ARCHITECTURE

To assess effectively the security needs of an organization and to evaluate and choose various security products and policies, the manager responsible for security needs some systematic way of defining the requirements for security and characterizing the approaches to satisfying those requirements. This is difficult enough in a centralized data processing environment; with the use of local and wide area networks, the problems are compounded.

ITU-T3 Recommendation X.800, *Security Architecture for OSI*, defines such a systematic approach.4 The OSI security architecture is useful to managers as a way of organizing the task of providing security. Furthermore, because this architecture was developed as an international standard, computer and communications vendors have developed security

features for their products and services that relate to this structured definition of services and mechanisms.

For our purposes, the OSI security architecture provides a useful, if abstract, overview of many of the concepts that this book deals with. The OSI security architecture focuses on security attacks, mechanisms, and services. These can be defined briefly as

■ **Security attack:**

Any action that compromises the security of information owned by an organization.

■ **Security mechanism:**

A process (or a device incorporating such a process) that is designed to detect, prevent, or recover from a security attack.

■ **Security service:**

A processing or communication service that enhances the security of the data processing systems and the information transfers of an organization. The services are intended to counter security attacks, and they make use of one or more security mechanisms to provide the service.

1.10. CLASSICAL ENCRYPTION TECHNIQUES

Symmetric encryption, also referred to as conventional encryption or single-key encryption, was the only type of encryption in use prior to the development of public key encryption in the 1970s. It remains by far the most widely used of the two types of encryption.

Terminologies:

- An original message is known as the **plaintext**, while the coded message is called the **ciphertext**.
- The process of converting from plaintext to ciphertext is known as **enciphering** or **encryption**;
- Restoring the plaintext from the ciphertext is **deciphering** or **decryption**.
- The many schemes used for encryption constitute the area of study known as **cryptography**. Such a scheme is known as a **cryptographic system** or a **cipher**.
- Techniques used for deciphering a message without any knowledge of the enciphering details fall into the area of **cryptanalysis**.
- Cryptanalysis is what the layperson calls "breaking the code."
- The areas of cryptography and cryptanalysis together are called **cryptology**.

1.10.1. SUBSTITUTION TECHNIQUES

Contents
• **Introduction**
• **Caesar Cipher**
• **Monoalphabetic Ciphers**
• **Playfair Cipher**
• **Hill Cipher**
• **Polyalphabetic Ciphers**
• **One-Time Pad**

Introduction

- ❖ The two basic building blocks of all encryption techniques are ***substitution*** and ***transposition***.
- ❖ A **substitution technique** is one in which the ***letters of plaintext are replaced by other letters*** or by numbers or symbols.

Caesar Cipher

- The Caesar cipher involves replacing each letter of the alphabet with the letter standing three places further down the alphabet. **For example**,

```
plain:  meet me after the toga party
cipher: PHHW PH DIWHU WKH WRJD SDUWB
```

- The algorithm can be expressed as follows.

$$C = \mathrm{E}(k, p) = (p + k) \bmod 26 \qquad \textbf{(2.1)}$$

- The decryption algorithm is simply

$$p = \mathrm{D}(k, C) = (C - k) \bmod 26 \qquad \textbf{(2.2)}$$

- If it is known that a given ciphertext is a Caesar cipher, then a brute-force cryptanalysis is easily performed: simply try all the 25 possible keys.

MonoalphabeticCiphers

- With only 25 possible keys, the Caesar cipher is far from secure.
- Before proceeding, we define the term *permutation*. A **permutation** of a finite set of elements *S* is an ordered sequence of all the elements of *S*, with each element appearing exactly once.
- **For example, if** *S* = {a, b, c}, there are six permutations of *S*:
  abc, acb, bac, bca, cab, cba
- If, instead, the "cipher" line can be any permutation of the 26 alphabetic characters, then there are 26! or greater than 4 * 1026 possible keys.

Playfair Cipher

- The best-known ***multiple-letter encryption*** cipher is the Playfair.
- The Playfair algorithm is based on the use of a 5 * 5 matrix of letters constructed using a keyword. Here is an example,

M	O	N	A	R
C	H	Y	B	D
E	F	G	I/J	K
L	P	Q	S	T
U	V	W	X	Z

- In this case, the keyword is *monarchy*.
- The matrix is constructed by filling in the letters of the keyword (minus duplicates) from left to right and from top to bottom, and then filling in the remainder of the matrix with the remaining letters in alphabetic order. The letters I and J count as one letter.
- **The rules to be followed are:**
  - Repeating plaintext letters that come in the same pair are separated with a filer letter, such as x.
  - Plaintext letters that fall in the same row are replaced by the letter to the right, with the first element of the row circularly following the first.
  - Plaintext letters that fall in the same column are replaced by the letter beneath, with the top element circularly following the last.
  - Otherwise each letter is replaced by the letter that lies in its own row and the column occupied by the other plaintext.

Hill Cipher

- ***The Hill Algorithm*** This encryption algorithm takes *m* successive plaintext letters and substitutes for them *m* ciphertext letters.

In general terms, the Hill system can be expressed as

$$\mathbf{C} = \mathrm{E}(\mathbf{K}, \mathbf{P}) = \mathbf{PK} \bmod 26$$
$$\mathbf{P} = \mathrm{D}(\mathbf{K}, \mathbf{C}) = \mathbf{CK}^{-1} \bmod 26 = \mathbf{PKK}^{-1} = \mathbf{P}$$

Polyalphabetic Ciphers

- Another way to improve on the simple monoalphabetic technique is to use different monoalphabetic substitutions as one proceeds through the plaintext message.
- The general name for this approach is **polyalphabetic substitution cipher**.
- A general equation of the ***encryption process*** is

$$C_i = (p_i + k_{i \bmod m}) \bmod 26 \qquad \textbf{(2.3)}$$

- Similarly, ***decryption is***

$$p_i = (C_i - k_{i \bmod m}) \bmod 26 \qquad \textbf{(2.4)}$$

- To encrypt a message, a key is needed that is as long as the message.
- For example, if the keyword is **deceptive**, the message "**we are discovered save yourself"** is encrypted as

```
key:          deceptivedeceptivedeceptive
plaintext:    wearediscoveredsaveyourself
ciphertext:   ZICVTWQNGRZGVTWAVZHCQYGLMGJ
```

Vignere cipher

- Simplest polyalphabetic substitution cipher is the Vigenère Cipher.

- effectively multiple caesar ciphers
- key is multiple letters long K = k1 k2 ... kd
- i^{th} letter specifies i^{th} alphabet to use
- use each alphabet in turn
- repeat from start after d letters in message
- decryption simply works in reverse

Example:

- write the plaintext out
- write the keyword repeated above it
- use each key letter as a caesar cipher key
- encrypt the corresponding plaintext letter
- eg using keyword *deceptive*

key: deceptivedeceptivedeceptive

plaintext: wearediscoveredsaveyourself

ciphertext:ZICVTWQNGRZGVTWAVZHCQYGLMGJ

Security:

- have multiple ciphertext letters for each plaintext letter
- hence letter frequencies are obscured
- but not totally lost
- start with letter frequencies
  - see if look monoalphabetic or not
- if not, then need to determine the 'number of alphabets' in the key string (aka. the *period* of the key), since then can attach each

Kasisky Method:

- method developed by Babbage / Kasiski
- repetitions in ciphertext give clues to period
- so find same plaintext an exact period apart
- which results in the same ciphertext
- e.g., repeated "VTW" in previous example
- suggests size of 3 or 9
- then attack each monoalphabetic cipher individually using same techniques as before

Autokey cipher

- ideally want a key as long as the message
- Vigenère proposed the autokey cipher
- with keyword is prefixed to message as key
- knowing keyword can recover the first few letters
- use these in turn on the rest of the message
- but still have frequency characteristics to attack
- e.g., given key '*deceptive'*

key: deceptivewearediscoveredsav

plaintext: wearediscoveredsaveyourself

ciphertext:ZICVTWQNGKZEIIGASXSTSLVVWLA

One-Time Pad

- using a random key that is as long as the message, so that the key **need not be repeated**.
- In addition, the key is to be used to encrypt and decrypt a single message, and then is **discarded**.
- Each new message requires a new key of the same length as the new message. Such a scheme, known as a **one-time pad**, is **unbreakable.**

1.10.2. TRANSPOSITION TECHNIQUES

- A very different kind of mapping is achieved by performing **some sort of permutation** on the plaintext letters. This technique is referred to as a transposition cipher.
- The simplest such cipher is the **rail fence** technique, in which the plaintext is written down as a sequence of diagonals and then read off as a sequence of rows.
- For example, to encipher the message **"meet me after the toga party**" with a rail fence of depth 2, we write the following:

```
m e m a t r h t g p r y
e t e f e t e o a a t
```

The encrypted message is

MEMATRHTGPRYETEFETEOAAT

1.10.3. STEGANOGRAPHY

Contents
• **Techniques** ✓ **Character marking** ✓ **Invisible ink** ✓ **Pin punctures** ✓ **Typewriter correction ribbon**

Techniques

- The methods of steganography **conceal the existence of the message**.
- **For example**, the sequence of first letters of each word of the overall message spells out the hidden message.
- **Various other techniques have been used historically; some examples are the following:**

  • **Character marking:** Selected letters of printed or typewritten text are overwritten in pencil. The marks are ordinarily not visible unless the paper is held at an angle to bright light.

  • **Invisible ink:** A number of substances can be used for writing but leave no visible trace until heat or some chemical is applied to the paper.

  • **Pin punctures:** Small pin punctures on selected letters are ordinarily not visible unless the paper is held up in front of a light.

  • **Typewriter correction ribbon:** Used between lines typed with a black ribbon, the results of typing with the correction tape are visible only under a strong light.

Advantages:

- **The advantage of steganography** is that it can be employed by parties who have something to lose should the fact of their secret communication (not necessarily the content) be discovered.
- Encryption flags traffic as important or secret or may identify the sender or receiver as someone with something to hide. Steganography has a number of drawbacks when compared to encryption.
- It requires a lot of overhead to hide a relatively few bits of information, although using a scheme like that proposed in the preceding paragraph may make it more effective.

1.11. FOUNDATIONS OF MODERN CRYPTOGRAPHY

Modern cryptography is the cornerstone of computer and communications security. Its foundation is based on various concepts of mathematics such as number theory, computational-complexity theory, and probability theory.

Characteristics of Modern Cryptography

There are three major characteristics that separate modern cryptography from the classical approach.

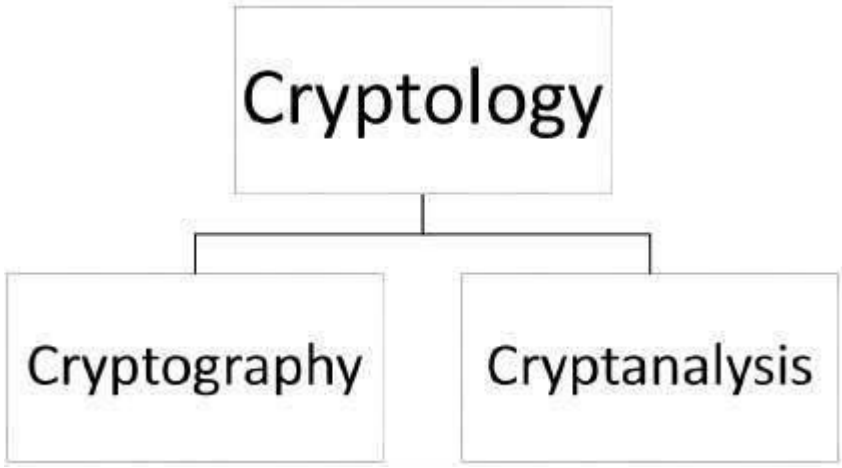
Classic Cryptography	Modern Cryptograph
It manipulates traditional characters, i.e., letters and digits directly.	It operates on binary bit sequences.
It is mainly based on 'security through obscurity'. The techniques employed for coding were kept secret and only the parties involved in communication knew about them.	It relies on publicly know mathematical algorithms for coding the information. Secrecy is obtained through a secrete key which is used as the seed fo the algorithm The computationa difficulty of algorithms, absence of secret key, etc., make it impossible for an attacker to obtain the original information

	even if he knows the algorithm use for coding.
It requires the entire cryptosystem for communicating confidentially.	Modern cryptography requires parties interested in secure communicatio to possess th secret key only.

Context of Cryptography

Cryptology, the study of cryptosystems, can be subdivided into two branches −

- Cryptography
- Cryptanalysis



What is Cryptography?

Cryptography is the art and science of making a cryptosystem that is capable of providing information security.

Cryptography deals with the actual securing of digital data. It refers to the design of mechanisms based on mathematical algorithms that provide fundamental information security services. You can think of cryptography as the establishment of a large toolkit containing different techniques in security applications.

What is Cryptanalysis?

The art and science of breaking the cipher text is known as cryptanalysis.

Cryptanalysis is the sister branch of cryptography and they both co-exist. The cryptographic process results in the cipher text for transmission or storage. It involves the study of cryptographic mechanism with the intention to break them. Cryptanalysis is also used during the design of the new cryptographic techniques to test their security strengths.

Note − Cryptography concerns with the design of cryptosystems, while cryptanalysis studies the breaking of cryptosystems.

1.11.1. PERFECT SECURITY

1.11.2. INFORMATION THEORY

1.11.3. PRODUCT CRYPTOSYSTEM

Another innovation introduced by Shannon in his 1949 paper was the idea of combining cryptosystems by forming their "product." In cryptography, a product cipher combines two or more transformations in a manner intending that the resulting cipher is more secure than the individual components to make it resistant to cryptanalysis. The product cipher combines a sequence of simple transformations such as substitution (S-box), permutation (P-box), and modular arithmetic. The concept of product ciphers is due to Claude Shannon.

This idea has been of fundamental importance in the design of present-day cryptosystems such as the Data Encryption Standard,

For simplicity, we will confine our attention in this section to cryptosystems in which $\mathcal{C} = \mathcal{P}$: cryptosystems of this type are called *endomorphic*.

Suppose $\mathbf{S_1} = (\mathcal{P}, \mathcal{P}, \mathcal{K}_1, \mathcal{E}_1, \mathcal{D}_1)$ and $(\mathbf{S_2} = (\mathcal{P}, \mathcal{P}, \mathcal{K}_2, \mathcal{E}_2, \mathcal{D}_2)$ are two endomorphic cryptosystems which have the same plaintext (and ciphertext) spaces. Then the product of S_1 and S_2, denoted by $S_1 \times S_2$, is defined to be the cryptosystem

$$(\mathcal{P}, \mathcal{P}, \mathcal{K}_1 \times \mathcal{K}_2, \mathcal{E}, \mathcal{D}).$$

A key of the product cryptosystem has the form $K = (K_1, K_2)$,

where $K_1 \in \mathcal{K}_1$ and $K_2 \in \mathcal{K}_2$. The encryption and decryption rules of the product cryptosystem are defined as follows: For each $K = (K_1, K_2)$, we have an encryption rule e_K defined by the formula

$$e_{(K_1,K_2)}(x) = e_{K_2}(e_{K_1}(x)),$$

and a decryption rule defined by the formula

$$d_{(K_1,K_2)}(y) = d_{K_1}(d_{K_2}(y)).$$

That is, we first encrypt x with e_{K_1}, and then "re-encrypt" the resulting ciphertext with e_{K_2}. Decrypting is similar, but it must be done in the reverse order:

$$\begin{aligned} d_{(K_1,K_2)}(e_{(K_1,K_2)}(x)) &= d_{(K_1,K_2)}(e_{K_2}(e_{K_1}(x))) \\ &= d_{K_1}(d_{K_2}(e_{K_2}(e_{K_1}(x)))) \\ &= d_{K_1}(e_{K_1}(x)) \\ &= x. \end{aligned}$$

Recall also that cryptosystems have probability distributions associated with their keyspaces. Thus we need to define the probability distribution for the keyspace $\mathcal{K}$ of the product cryptosystem. We do this in a very natural way:

$$p_{\mathcal{K}}(K_1, K_2) = p_{\mathcal{K}_1}(K_1) \times p_{\mathcal{K}_2}(K_2).$$

1.11.4. CRYPTANALYSIS