Dr. C.V. Raman University
Kargi Road, Kota, BILASPUR, (C. G.),
Ph. : +07753-253801, +07753-253872
E-mail : info@cvru.ac.in | Website : www.cvru.ac.in

Information Technology & Trends

Institute of Open and Distance Education

Faculty of Arts

Information Technology & Trends

4BA8

DR. C.V. RAMAN UNIVERSITY
Chhattisgarh, Bilaspur
A STATUTORY UNIVERSITY UNDER SECTION 2(F) OF THE UGC ACT

4BA8
INFORMATION TECHNOLOGY AND TRENDS

4BA8, Information Technology & Trends

Edition: March 2024

Compiled, reviewed and edited by Subject Expert team of University

1. Dr. Abhivav Shukla

(Associate Professor, Dr. C. V. Raman University)

2. Dr.Vaibhav Sharma

(Associate Professor, Dr. C. V. Raman University)

3. Dr. Ayush Agarwal

(Assistant Professor, Dr. C. V. Raman University)

Published by:

Dr. C.V. Raman University

Kargi Road, Kota, Bilaspur, (C. G.),

Ph. +07753-253801, 07753-253872

E-mail: info@cvru.ac.in

Website: www.cvru.ac.in

TABLE OF CONTENTS

NOTES

1 Introduction

Chapter Includes :

. Introduction

ıanager must understand the broader organization, management and information nology dimensions of systems and their power to provide solutions to challenges problems in business environment.

current challenge before professionals is to use the capabilities of computers to ort intelligent and knowledgeable work including managerial activities and de- n making.

ness applications of information technology now include internet , electronic nerce , intranet , extranet , information systems for business operations , buted systems , e-SCM, e-CRM , expert systems , decision support systems ı and web mining , GIS , mobile business applications etc. .

NOTES

1.2 Parallel and Distributed Systems

A process is a unit of execution managed at the level of the operating system .Each process has its own address space, i.e., no other process can access it. A thread is a sequential flow of control executing within a process .All threads within a process share the same address space, i.e., they share memory.

1.2.1 Parallelism

With ***pseudo-parallelism***, a thread can be interrupted by the system at any time (we say that the system is preemptive) .With ***quasi-parallelism***, a thread can only be interrupted voluntarily, either explicitly or when it performs an input/output system call.

A parallel computing system is a computer with more than one processor for parallel processing. In the past, each processor of a multiprocessing system always came in its own processor packaging, but recently-introduced ***multicore*** processors contain multiple logical processors in a single package.

There are many different kinds of parallel computers. They are distinguished by th kind of interconnection between processors (known as "processing elements" c PEs) and memory.

Flynn's taxonomy: one of the most accepted taxonomies of parallel architecture classifies parallel (and serial) computers according to :

- Whether all processors execute the same instructions at the same tin (*single instruction/multiple data* - SIMD) or
- Each processor executes different instructions (*multiple instruction/mu tiple data* - MIMD).

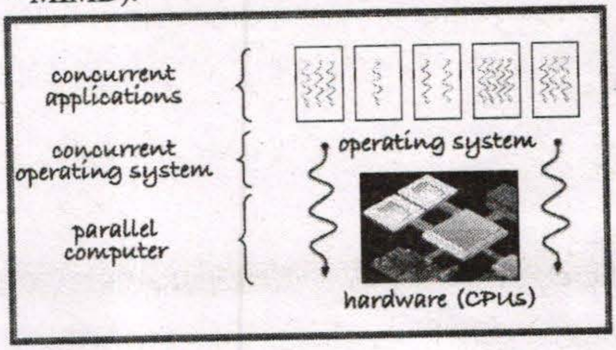


Figure 1.1 :Parallelism

Processors in a parallel computer may communicate with each other in a numb ways, including shared (either multiported or multiplexed) memory, a cross shared bus or an interconnect network of a myriad of topologies including star tree, hypercube, fat hypercube (a hypercube with more than one processc node), an n-dimensional mesh, etc. Parallel computers based on interconne work need to employ some kind of routing to enable passing of messages be nodes that are not directly connected. The communication medium used for cc nication between the processors is likely to be hierarchical in large multipro machines. Similarly, memory may be either private to the processor, shared b a numbers of processors, or globally shared. Systolic array is an examp multiprocessor with fixed function nodes, local-only memory and no message routing.

Parallel programming is the design, implementation, and tuning of parallel computer programs which take advantage of parallel computing systems. It also refers to the application of parallel programming methods to existing serial programs (parallelization).

Parallel programming focuses on partitioning the overall problem into separate tasks, allocating tasks to processors and synchronizing the tasks to get meaningful results. Parallel programming can only be applied to problems that are inherently parallelizable, mostly without data dependence. A problem can be partitioned based on domain decomposition or functional decomposition, or a combination.

A parallel programming model is a computing architecture and language designed to express parallelism in software systems and applications. The software to support these models includes compilers, libraries and other tools that enable the application to use parallel hardware.

1.2.2 Distributed Systems

Distributed systems provide sharing of resources and information over a computer network. A key design issue that makes these systems attractive is that all aspects related to distribution are transparent to users. Unfortunately, general-purpose wide area distributed systems that allow users to share and manage arbitrary resources in a transparent way hardly exist. Two fundamental trends influence the way we conceive and construct new computing and information systems. The first is that information technology of all forms is becoming highly commoditized *i.e.,* hardware ind software artifacts are getting faster, cheaper, and better at a relatively predictible rate. The second is the growing acceptance of a network-centric paradigm, vhere distributed applications with a range of quality of service (QoS) needs are onstructed by integrating separate components connected by various forms of ommunication services.

A distributed system is a collection of independent computers that appears to its ers as a single coherent system. "(***Tanenbaum***)

A distributed system is the one that prevents you from working because of the lure of a machine that you had never heard of." (***Veríssimo & Rodrigues***)

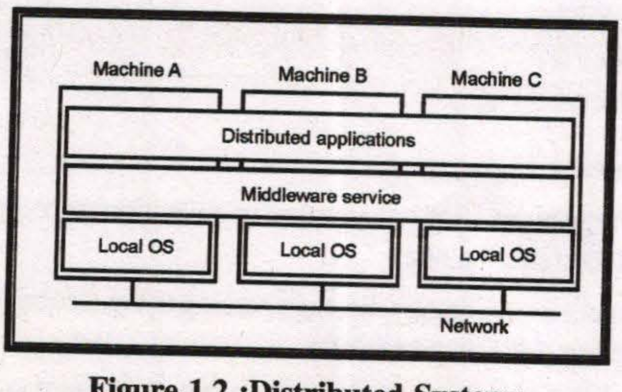


Figure 1.2 :Distributed Systems

NOTES

Distributed computing is a method of computer processing in which different part of a program run simultaneously on two or more computers that are communicating with each other over a network.

Distributed Systems architectures typically fall into following categories:

- Client-server — Smart client code contacts the server for data, then formats and displays it to the user. Input at the client is committed back to the server when it represents a permanent change.
- 3-tier architecture — Three tier systems move the client intelligence to a middle tier so that stateless clients can be used. This simplifies application deployment. Most web applications are 3-Tier.
- N-tier architecture — N-Tier refers typically to web applications which further forward their requests to other enterprise services. This type of application is the one most responsible for the success of application servers.
- Tightly coupled (clustered) — refers typically to a set of highly integrated machines that run the same process in parallel, subdividing the task in parts that are made individually by each one, and then put back together to make the final result.
- Peer-to-peer — an architecture where there is no special machine or machine that provide a service or manage the network resources. Instead all responsi bilities are uniformly divided among all machines, known as peers.

Computer Networks Vs Distributed Systems

A Computer Network: Is a collection of spatially separated, interconnected con puters that exchange messages based on specific protocols. Computers are a dressed by IP addresses.

A Distributed System: Multiple computers on the network working together a system. The spatial separation of computers and communication aspects are hid from users.

1.3 Computer Networks

Common distributed systems are based on widely used computer networks. Following are some examples:

- The Internet
- Intranets
- Wireless networks

1.3.1 The Internet

Consists of a large number of interconnected collection of computer netwo different types. Features include:

- Computers interacting by message passing using a common means of munication (Internet protocol).
- Many different services (applications) (World Wide Web, email, file

NOTES

fer).

- A number of Intranets linked by backbones.
- Internet Service Providers (ISPs), that provide access to the services on the Internet while providing local service such as email and web hosting.
- A backbone network link with high transmission capacity .
- Communication via Satellite, fibre optic cables and other high-bandwidth circuits.

1.3.2 Intranets

It is a portion of the Internet that is separately administered by organizations. Features include:

- A boundary that can be configured to enforce local security policies.
- Several local area connection (LANs) linked by backbone connections.
- A connection to the Internet via a router allowing users within the intranet to access services on the Internet.
- Firewalls to protect an intranet by preventing unauthorized messages leaving or entering by filtering incoming and outgoing messages e.g. by source or destination.

.3.3 Wireless Networks

/ireless networking allows the integration of small and portable computing devices .g. laptop computers, handheld devices (PDAs and mobile phones, video and gital cameras), wearable devices, vice embedded applications (washing machines, refrigerators) into distributed sys-ns.

o popular paradigms that use wireless networks are:

- **Mobile Computing (Nomadic Computing)-** The mobile computing paradigm allows users to perform computing tasks on the move, while being a part of a distributed system. User is normally connected to other resources via a wireless network.
- **Ubiquitous Computing -** Refers to the paradigm where small, cheap, computing devices that are embedded in devices are used as a part of a part of a distributed system, using wireless networks.

Resource Sharing

Service : is a part of distributed system that manages a collection of resources and provides functionality to end users.

Server : is a process that accepts requests from other computers and perform a service.

Client : is a process that requests the service.

NOTES

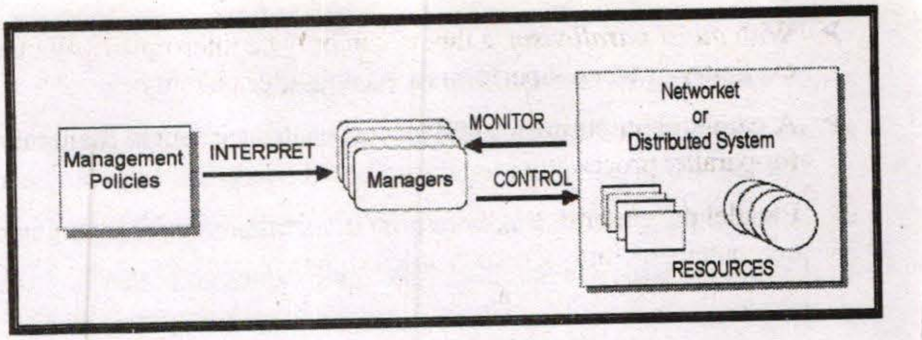


Figure 1.3 :Resource Sharing and Management

- **Distributed database (DDB) :** A logically interrelated collection of share data, physically distributed across multiple servers connected by a networl
- **Distributed database management system (DDBMS) :** The softwa system that facilitates the management of a DDB in such a way that th distribution aspects are transparent to users.

1.4 Modern Business Management

As the Internet evolves, new businesses continually spring up, hoping to take a vantage of the ubiquity and ease of use of this network of networks. Anyone w a computer can traverse the Internet with a standard tool set of browser softw and file transfer protocols. Today's e-business environment enables manufactur to automate and integrate functions as never before—from customer relations management on the sell side, to supply chain management on the buy side, to manufacturing operations on the production floor. Online exchanges began to spr up soon after the emergence of the Web.

According to Arthur Sculley and William Woods, four trading methods are c monly used in today's e-marketplaces:

- Fixed pricing, such as catalog resellers use
- One-to-one negotiation
- Auctions (both seller-driven and buyer-driven, or reverse, auctions)
- Automatic systems (two-way auctions)

SUMMARY

- The current challenge before professionals is to use the capabilities of puters to support intelligent and knowledgeable work including man activities and decision making .
- A process is a unit of execution managed at the level of the op system.
- A thread is a sequential flow of control executing within a proce threads within a process share the same address space, i.e., the memory.
- With *pseudo-parallelism*, a thread can be interrupted by the syster time (we say that the system is preemptive) .

NOTES

- With *quasi-parallelism*, a thread can only be interrupted voluntarily, either explicitly or when it performs a input/output system call.
- A parallel computing system is a computer with more than one processor for parallel processing.
- Parallel programming is the design, implementation, and tuning of parallel computer programs which take advantage of parallel computing systems. It also refers to the application of parallel programming methods to existing serial programs (parallelization).
- Parallel programming focuses on partitioning the overall problem into separate tasks, allocating tasks to processors and synchronizing the tasks to get meaningful results.
- A parallel programming model is a computing architecture and language designed to express parallelism in software systems and applications.
- Distributed systems provide sharing of resources and information over a computer network.
- Distributed computing is a method of computer processing in which different parts of a program run simultaneously on two or more computers that are communicating with each other over a network.
- A Computer Network: Is a collection of spatially separated, interconnected computers that exchange messages based on specific protocols. Computers are addressed by IP addresses.
- Server: is a process that accepts requests from other computers and perform a service.
- Client: is a process that requests the service.
- Distributed database (DDB) : A logically interrelated collection of shared data, physically distributed across multiple servers connected by a network.
- Distributed database management system (DDBMS) : The software system that facilitates the management of a DDB in such a way that the distribution aspects are transparent to users.
- Today's e-business environment enables manufacturers to automate and integrate functions as never before—from customer relationship management on the sell side, to supply chain management on the buy side, to e-manufacturing operations on the production floor. Online exchanges began to spring up soon after the emergence of the Web.

EXERCISE

1. What are different modern Business applications of Information Technology?
2. How can you differentiate between the process and threads?
3. Describe the parallelism with suitable example . Give different types of parallelism.
4. What is Flynn's taxonomy of parallel architectures. Describe in brief.

NOTES

5. What do you understand with parallel computing and parallel processors?
6. How networked and distributed systems are related ? Explain with suitable examples.
7. What is Distributed computing ? Describe its performance features.
8. What are Distributed Systems architectures?
9. Explain how portable computing devices are taking an active participation in modern business management .
10. Describe different advantages of e-business environment

NOTES

2 Distributed Systems

Chapter Includes :

.1 Distributed System

ntroduction

istributed System is collection of computers that are spatially separated and do not hare common memory. The Process executed on these communicates with one nother by exchanging message over communication channel. The messages are elivering after a transmission delay. Distributed System appeares to its user as entralized operating system for a single machine but it run on multiple independent omputers. An identical copy of the operating system may run at every computer. he key concept is transparency (Transparency implies to access resources or rvices without having any knowledge about their location in system)

NOTES

A Distributed Operating System extends the concept of source management and user friendly interface for shared memory computers consisting of several computers connected by a communication network.

A distributed file system gives a new way of accessing files and directories. To implement a distributed file system, we do not have to implement new ways of storing files; instead we use existing implemented file managers to store files.

2.1.1 Characteristics of Distributed Systems

There are three major characteristics of Distribution System –

1. **Concurrency :** In network computer concurrent execution is one of the norms. Adding more resources to the Distribution System can add more capacity to the whole system.
2. **No Global Clock :** When programs need to corporate them co-ordinate their action by exchanging message directly but they turn-out. There are limit to accuracy with which the networks can synchronize their clock. Limits to how closely computers can synchronize their clock.
3. **Independent Failure :** Any components of the network can fail and software must deal with this. For example - Internet and Intranet.

(Table 2.1 : Characteristics of Distributed Systems)

Transparency	Description
Access	Hide differences in data representation and how a resource is accessed
Location	Hide where a resource is located
Migration	Hide that a resource may move to another location
Relocation	Hide that a resource may be moved to another location while in use
Replication	Hide that a resource is replicated
Concurrency	Hide that a resource may be shared by several competitive users
Failure	Hide the failure and recovery of a resource
Persistence	Hide whether a (software) resource is in memory or on disk

2.1.2 Categories of Distributed Systems

- ***Clusters :*** Groups of PCs (ordinary or specialized) brought specifically together to work collectively on problems.

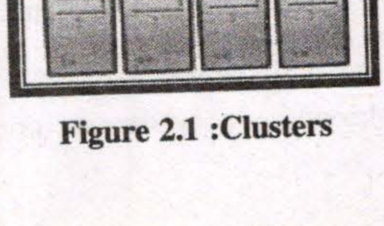


Figure 2.1 :Clusters

NOTES

- *Grids :* Clusters can be combined to form a "Grid", a system of massive collective computing power which is designed to be easily used by "plugging in" to it.

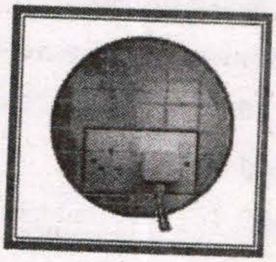


Figure 2.2 : Grids

- *Peer 2 Peer :* A system whereby individual users or nodes can communicate with each other by themselves. Examples of such a system would be Napster.

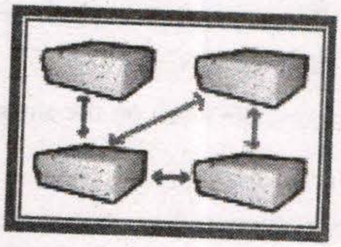


Figure 2.3 : Peer 2 Peer

Others : The WWW is a distributed system! (of information). It is actually peer to peer, but is worth mentioning separately as a good example. CORBA can be used to create a distributed system of programming objects, almost like a distributed developer system.

1.3 Reliability of a Distributed System

ıree components of reliability are availability, security, and fault tolerance deribed below:

Availability

- The fraction of the time the system is usable.
- Try to reduce dependence between servers because this increases the chance that the user can still do something.
- To improve availability, key components should be replicated.

Security

- Must authenticate every message in a distributed OS .
- Can't trust id field in a message .
- Should also encrypt contents to prevent snooping on the network .
- Secure Sockets Layer (SSL) provides this type of security .

Fault Tolerance

- Refers to the general ability of the system to tolerate errors .
- Most often achieved by having multiple instances of a resource .

NOTES

- Server should be able to crash and recover without affecting service
  - May want to reject the request that caused crash.
- should be able to vary number by choice
  - Performance degrades when one crashes or is removed from service.
- Perhaps if one crashes, others should take over its work queue

2.1.4 Hardware for Distributed Systems

(a) Distributed systems are MIMD, i.e., have multiple instruction streams and multiple data streams.

(b) Tightly coupled hardware

- There is a small delay when sending messages .
- Fast rate of data transfer.
- Usually used in parallel systems.
- Parallel algorithm: processors work on the same problem.

(c) Loosely coupled hardware

- There is a longer delay when sending messages.
- Slower rate of data transfer.

(d) Multiprocessor

- A distributed system in which the processors share memory, i.e., they all share a single virtual address space.
- Symmetric multiprocessor (SMP): all processors are of the same type.

(e) Multicomputer

- A distributed system in which the processors do NOT share memory .
- An example: a collection of personal computers connected by a network

(f) Bus system

- A distributed system with a single network, backplane, bus, cable, or othe medium which connects all processors .

(g) Switched system

- A distributed system which has individual wires from machine to machine
- Many different wiring patterns.
- Messages move along wires .
- Switching (routing) decisions are made step by step along the route .

(h) Bus-based multiprocessors

- One memory for several processors .
- Performance drops when bus gets overloaded .(when several CPU's ar attached to the bus)
- Adding cache memory may allow the addition of more CPU's before th bus becomes overloaded

NOTES

2.1.5 Software for Distributed Systems

There are following types of Software for Distributed Systems:

(a) Network Operating System

- Loosely connected software on loosely connected hardware (multicomputer) .
- Network of workstations or PC's .
- Each machine running an OS; the OS may even differ .
- e.g., lab of Linux machines

(b) Integrated Distributed System

- Tightly coupled software on loosely coupled hardware.
- Runs on a collection of shared machines that do not have shared memory yet work like one computer .

(c) Multiprocessor Timesharing System

- Tightly coupled software on tightly coupled hardware .
- e.g., UNIX machine with several processors .
- Shared memory .
- One short-term scheduler queue (kept in shared memory) .
- e.g., Grendel (24 processor SGI super computer)

2.2 Managing Distributed Databases

A **database** is defined as a structured collection of records or data that is stored in a computer so that a program can consult it to answer queries. The records retrieved in answer to queries become information that can be used to make decisions. The computer program used to manage and query a database is known as a database management system (DBMS).

The increasing volume of data to store in databases makes them more and more often huge and permanently growing. Typically, large tables are hash or range partitioned into segments stored over different storage sites. Current RDBMSs, e.g. SQL Server, Oracle or DB2, provide static partitioning only.

Imagine the file users forty years ago in the centralized environment. The Indexed Sequential Access Method was in use for the ordered (range partitioned) files. Likewise, the static hash access methods were the only known for the files. Both approaches required the file reorganization whenever the inserts overflowed the file capacity. The B-trees and the extensible (linear, dynamic) hash methods were invented to avoid the need. They replaced the file reorganization with the dynamic incremental splits of one bucket (page, leaf, segment…) at the time.

Efficient management of distributed data present specific needs. Most large organizations maintain their data in many distinct autonomous databases that have been developed at different times, on different platforms and Data Management Systems, and most often in independent units that have recently merged. It shows a distributed framework of database .

NOTES

Organizing the interaction between each computer is of prime importance. In order to be able to use the widest possible range and types of computers, the protocol or communication channel should not contain or use any information that may not be understood by certain machines. Special care must also be taken that messages are indeed delivered correctly and that invalid messages are rejected which would otherwise bring down the system and perhaps the rest of the network.

Distributed Database (DDB) : A logically interrelated collection of shared data, physically distributed across multiple servers connected by a network.

2.2.1 Categories of DDBMS

There are mainly two categories of DDBMS exists:

- A **homogeneous DDBMS** is one in which the same DBMS is used in all local servers. A homogeneous DDBMS may be administered by a central authority (*non-autonomous*) or support some degree of local autonomy (*autonomous*). All users access the underlying local databases via a global interface described by a global schema and against which user views are defined. Since users access data through the global interface only, the local DBMS may or may not have local schema.
- A **heterogeneous DDBMS** uses different DBMSs at the local server. Integration of the different DBMSs is achieved either through special software that is part of the DDBMS itself, thus creating what is known as a multidatabase management system (MDBMS), or through *gateways*, which convert query languages and data models across different DBMS. This latter approach has serious drawbacks, in that it the gateway can effectively only translate queries and cannot coordinate any form of concurrency control or recovery.

2.2.2 DDB : Pros and Cons

Pros of DDB are :

- ***Improved reliability :*** Replication of data at additional sites improves the overall reliability of the system.
- ***Shareability and local autonomy :*** Local users can control their own data, hold it locally and obtain better performance, while still being responsible to a 'global' data base administrator for centrally defined information.
- ***Improved availability :*** If there is failure at one site, other sites can still access their own data. If data is replicated on another site, then the site that is down can be bypassed completely.
- ***Improved performance :*** Local data is located at the site where demand for it is likely to be greatest, thus opening the possibility of much faster access and less contention for processor time, I/O devices etc. in a centralized system.
- ***Organizational structure :*** Organisations are naturally distributed over many locations, e.g. estate agents or any franchise operation. A database at each branch can store local details for that area and branch staff can make local inquiries, while company headquarters staff make enquiries requiring access to some or all their branches.

NOTES

- ***Economics :*** It is no longer true that centralized processing power in a single piece of hardware is necessarily cheaper than separate smaller units that can be added when and where needed. Local access also reduces communications traffic between centre and periphery.
- ***Modular growth :*** It is simpler to handle expansion in properly design distributed systems. New sites can easily be added without affecting the operations of others. Growth in a centralised system may require changes to both hardware and software.

Cons of DDB are :

- ***Complexity :*** By hiding their distributed nature and trying to ensure optimum performance, reliability and availability, a DDBMS is more complex, particularly if integration of heterogeneous local systems is required.
- ***Cost :*** Procurement and maintenance costs are higher due to complexity and there may be increased ongoing costs to cover communications and local administration.
- ***Security :*** Centralized access to data is easily controlled; multiple locations, replicated data and insecure networks all create additional risks.
- ***Integrity :*** Considerable processing and communication cost may be incurred in order to enforce validity and consistency rules defined on both local and global data. This may be a complex problem affecting performance.
- ***Lack of standards :*** Standard communication and data access protocols are a relative new thing. Also tools and methods to convert from centralised to distributed systems are lacking.
- ***Lack of experience :*** General purpose distributed systems have not yet been widely accepted as replacements for centralised systems; hence there is only a small pool of experience at present.
- ***DDB design more complex :*** In addition to normal design problems, DDB design must also cope with problems of how to subdivide data and place it at suitable sites, the issues known as fragmentation and allocation (of fragments) and data replication.

Distributed databases can suffer from a variety of problems. Networks can be slow; so it is important to minimize the number and volume of messages between sites. Query processing will often offer a number of options whose execution times may vary by an orders of magnitude. Recovery from failure at any site also requires co-ordinated action in order to avoid loss of integrity or consistency.

Distributed databases are likely to become an increasingly important IT component for many commercial organisations as they seek to maximize the value of information held over a network of corporate sites. Distributed databases offer considerable advantages, disadvantages and difficulties, in part the result of the wide variety of possible architectures based on homogeneous and heterogeneous models. When compared with centralized databases, those that are distributed appear far more complex, both because of the new issues introduced by fragmentation and allocation of data and because of new dimensions to old problems such as concurrency and recovery.

NOTES

2.3 Distributing the Processing and Storage Function

In distributed system, multple computers are connected on the network working together as a system. These computers are independent but their collection appears to its users as a single coherent systems. Distributed system provides sharing of resources & information. Processes executed on these systems can communicate with one another by exchanging messages over communication channel.

The Distributed processing refers to LANs designed so that a single program can run simultaneously at various sites. Most distributed processing system contains sophisticated software that detects idle CPUs on the network and parcels out programs to utilize them.

Another form of distributed processing involves distributed data base, databases in which the data stored across two or more computer systems. The database system keeps track of where the data is so that the distributed nature of the database is not apparent to users.

A distributed database consists of two or more data files located at different sites on a computer network. Because the database is distributed, different users can access it without interfering with one another. However, the DBMS must periodically synchronize the scattered databases to make sure that they all have consistent data.

The software system that facilitates the management of a DDB in such a way that the distribution aspects are transparent to users.

A DDBMS running on a different computer at each site can handle local applications autonomously and participates in at least one global application requiring data from other sites. Communication between different sites via a network is essential for any global application.

2.3.1 Functions of DDBMS

The functions of such a DDBMS are :

- To extend communication services to provide access to remote sites and allow the transfer of queries and data across the network.
- To extend the DD to store data distribution details.
- To provide distributed query processing, including optimization and remote data access.
- To extend concurrency control to maintain consistency of replicated data.
- To extend recovery services to take account of failures either of sites or of communications links.

2.3.2 Distributed Data Storage

There are two approaches for storing data in distributed database :

(i) Replication

(ii) Fragmentation

NOTES

(i) Replication : The system creates different similar copy of any data and each copy is stored on different places. There are two types of replication strategy.

(a) Complete Replication

(b) Selective Replication

(a) Complete Replication : The **complete replication** strategy consists of maintaining a complete copy of the database at each site. This maximizes accessibility and reliability, but costs of storage and of the communications needed for up-dates may be high.

(b) Selective Replication : The Selective Replication strategy stores a copy of database in selected places. There is a commonly used strategy, called **selective replication**, which combines degrees of partitioning, replication and centralisation according to usage and cost.

Advantage of Replication

There are four advantage of data replication :

(i) **Reliability :** If one of the sites, containing the relation (or database) fails, a copy can always be found at another site, without network traffic delays. Also available copies can all be updated, as soon as possible transaction occurs, and unavailable nodes will be updated once they return to service.

(ii) **Fast Response :** Each site - that has a full copy can process queries locally, so queries can be processed rapidly.

(iii) **Node decoupling :** Each transaction may proceed without coordination across the network. Thus, if nodes are down, busy, or disconnected, a transaction is handled when the user desires. In the place of real-time, synchronization updates, a behind the scenes process coordination, all data copies.

(iv) **Reduced network traffic at prime time :** Often updating data happens during prime business hours, when network traffic is highest and the demands for rapid response greatest. Replication, with delayed updating of copies of data, moves network traffic for sending updates to other nodes to non prime time hours.

Disadvantages of Replication

(i) **Storage requirement :** Each site has a full copy and there must have the same storage capacity that would be required if the data were stored centrally. Each copy requires storage space, and processing time is required to update each copy - on each node.

(ii) **Complexity and cost of updating :** Whenever a relation is updated, it must be updated at each site that holds a copy.

For these reasons, data replication is used where most process requested are read only, and where data is relatively stated as in catalog telephone directory, train schedules and so on.

Replication is not used in online applications - such as airline reservation, automated teller machine transactions and other financial activities or nonsharable resource.

(ii) Fragmentation : The **partitioned (or fragmented)** strategy splits the database into disjoint fragments, each assigned to a particular site; performance and communication costs should be good if such an allocation is designed carefully using quantitative information about frequency of uses, location of uses and performance of transactions.

There are two types of fragmentation :

(a) Horizontal Fragmentation

(b) Vertical Fragmentation

(a) Horizontal Fragmentation : With horizontal fragmentation, some of the rows of a table are put into a base relation at one site, and other rows are put into a base relation at another site.

Advantages of Horizontal Fragmentation

(i) **Efficiency :** Data are stored, close to where they are used, but separate from other data, used by other users or applications.

(ii) **Local optimization :** Data can be stored to optimize performance for local access.

(iii) **Security :** Data are not relevant to usage if a particular site is not made available.

Disadvantage of Horizontal Fragmentation

(i) **Inconsistent access speed :** When data from several partitions are required, the access time can be significantly different from local data access.

(ii) **Back-up vulnerability :** Since data are not replicated when data at one site become inaccessible or damaged. Usage cannot switch on to another site where a copy exists, and data may be lost if proper backup is not performed at each site.

(b) Vertical Fragmentation : With the vertical fragmentation approach, some of the columns of a relation are projected into a base relation at one of the sites. And other columns are projected into a base relation at another site. The relations at each of the sites must share a common domain, so that the original table can be reconstructed.

2.4 Transactions and Concurrency

Most of the practical databases ("transactional databases") attempt to enforce a database transaction . Ideally, the database software should enforce the **ACID** rules, given below :

- ***Atomicity :*** Either all the tasks in a transaction must be done, or none of them. The transaction must be completed, or else it must be undone (rolled back).

NOTES

- ***Consistency :*** Every transaction must preserve the integrity constraints — the declared consistency rules — of the database. It cannot place the data in a contradictory state.
- ***Isolation :*** Two simultaneous transactions cannot interfere with one another. Intermediate results within a transaction are not visible to other transactions.
- ***Durability :*** Completed transactions cannot be aborted later or their results discarded. They must persist through (for instance) restarts of the DBMS after crashes
- A **cascading rollback** occurs in database systems when a transaction (T1) causes a failure and a rollback must be performed. Other transactions dependent on T1's actions must also be rolled back due to T1's failure, thus causing a cascading effect.

2.5 Advantages and Disadvantages of DS

Advantages of Distributed Systems

(a) Compared to centralized systems:

- Better price/performance than mainframe.
- More total computing power
  - o Sum of the computing power of the processors in the distributed system may be greater than any single processor available (parallel processing).
- Some applications are inherently distributed.
- Improved reliability because system can survive crash of one processor.
- Incremental growth can be achieved by adding one processor at a time.
- Shared ownership facilitated.

(b) Compared to isolated PCs:

- Shared management of system
  - o Backups
  - o Maintenance: download new software.
- Data sharing: many users can access a shared database .
- Device sharing: share expensive peripherals .
- Communication between users easier (e.g., electronic mail) .
- Flexibility: spread workload over machines (the jobs are not necessarily run on the owner's machine) .

Disadvantages of Distributed Systems

- Software may not be available .
- Networks can become saturated (may need additional wiring).
- Security problems protecting data (only an isolated machine is safe).

- Incremental growth is difficult in practice because of changing hardware a software

2.6 Flavors of Distributed Systems

NOTES

Heterogeneity:

Variety and differences in networks, computer hardware, operating systems, prc gramming languages, and implementations by different developers

- ***Middleware :*** Software layer that provides a programming abstraction as wel as masking the heterogeneity of the underlying systems (networks, hardware operating systems, and programming languages): e.g., CORBA and Java RMI.
- ***Mobile code :*** Code that can be sent from one computer to another and run at the destination; e.g., Java applets.

Openness:

System made extensible by publishing the specification and documentation for key software interfaces, i.e. made available to software developers.

Security:

Confidentiality (protection against disclosure to unauthorized individuals), integrity (protection against alteration or corruption), and availability (protection against interference with the means to access the resource).

- ***Denial of service attacks :*** Bombard a service with such a large number of pointless requests that serious users are unable to use it.

Scalability:

System will remain effective if there is a significant increase in the number of resources and users

- Cost of physical resources should O(n) for n users, performance on lookup tasks should be no worse than O(log n) for data of size n, software resources shouldn't run out, and performance bottlenecks should be avoided.
- Techniques for avoiding bottlenecks: distributed algorithms; caching, replication, and other forms of distributed data; e.g., multiple servers.

Failure Handling:

Failure may be partial in a distributed system because some components may fail while others continue processing.

- Techniques are needed for failure detection, failure masking, failure tolerance (by software and users), and recovery from failure.
- Use of redundant components is crucial to failure handling: e.g., multiple routes, Domain Name Servers, and databases.

Concurrency:

Allowing requests from several clients to be processed such that they are in between starting and finishing at the same time

NOTES

- Operations on any object in a distributed environment must be synchronized such that the object's data remains consistent.

Transparency:

Concealment from the user and application programmer of the separation of components in a distributed system , so that the system is perceived as a whole rather than as a collection of independent components.

2.7 Architectures of Distributed Systems

Distributed Systems architectures typically fall into following categories:

- ***Client-server*** — Smart client code contacts the server for data, then formats and displays it to the user. Input at the client is committed back to the server when it represents a permanent change.
- ***3-tier architecture*** — Three tier systems move the client intelligence to a middle tier so that stateless clients can be used. This simplifies application deployment. Most web applications are 3-Tier.
- ***N-tier architecture*** — N-Tier refers typically to web applications which further forward their requests to other enterprise services. This type of application is the one most responsible for the success of application servers.
- ***Tightly coupled (clustered)*** — refers typically to a set of highly integrated machines that run the same process in parallel, subdividing the task in parts that are made individually by each one, and then put back together to make the final result.
- ***Peer-to-peer*** — an architecture where there is no special machine or machines that provide a service or manage the network resources. Instead all responsibilities are uniformly divided among all machines, known as peers.

2.7.1 Client-server Architecture

Client-server(C/S) is computing architecture which separates a client from a server, and is almost always implemented over a computer network. Each client or server connected to a network can also be referred to as a node. The most basic type of client-server architecture employs only two types of nodes: clients and servers. This type of architecture is sometimes referred to as *two-tier*.

Each instance of the client software can send data requests to one or more connected *servers*. In turn, the servers can accept these requests, process them, and return the requested information to the client. Although this concept can be applied for a variety of reasons to many different kinds of applications, the architecture remains fundamentally the same.

These days, clients are most often web browsers, although that has not always been the case. Servers typically include web servers, database servers and mail servers etc.

NOTES

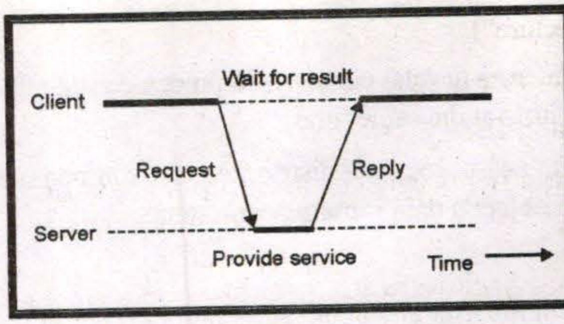


Figure 2.4: General interaction between a client and a server

C/S Characteristics

Characteristics of a client:

- Active (master) .
- Initiates requests .
- Waits for and receives replies .
- Usually connects to a small number of servers at one time .
- Typically interacts directly with human beings using a graphical user interface.

Characteristics of a server:

- Passive (slave) .
- Waits for requests from clients .
- Upon receipt of requests, processes them and then serves replies .
- Usually accepts connections from a large number of clients .
- Typically does not interact directly with human beings .
- Can be stateless or stateful .

2.7.2 Three-tier architecture

Three-tier is a client-server architecture in which the user interface, functional process logic ("business rules"), data storage and data access are developed and maintained as independent modules, most often on separate platforms. The three-tier model is considered to be a software architecture and a software design pattern.

Apart from the usual advantages of modular software with well defined interfaces, the three-tier architecture is intended to allow any of the three tiers to be upgraded or replaced independently as requirements or technology change. For example, a change of operating system from Microsoft Windows to Unix would only affect the user interface code.

Typically, the user interface runs on a desktop PC or workstation and uses a standard graphical user interface, functional process logic may consist of one or more separate modules running on a workstation or application server, and an RDBMS on a database server or mainframe contains the data storage logic. The middle tier may be multi-tiered itself (in which case the overall architecture is called

NOTES

during the computation process. Grids serve to manage the allocation of jobs to computers which will perform the work independently of the rest of the grid cluster. Resources such as storage may be shared by all the nodes, but intermediate results of one job do not affect other jobs in progress on other nodes of the grid.

2.7.5 Peer-to-Peer Architecture

A **peer-to-peer** (or **P2P**) computer network relies primarily on the computing power and bandwidth of the participants in the network rather than concentrating it in a relatively low number of servers. P2P networks are typically used for connecting nodes via largely *ad hoc* connections. Such networks are useful for many purposes. Sharing content files containing audio, video, data or anything in digital format is very common, and real-time data, such as telephony traffic, is also passed using P2P technology.

A pure peer-to-peer network does not have the notion of clients or servers, but only equal *peer* nodes that simultaneously function as both "clients" and "servers" to the other nodes on the network. This model of network arrangement differs from the client-server model where communication is usually to and from a central server. A typical example for a non peer-to-peer file transfer is an FTP server where the client and server programs are quite distinct, and the clients initiate the download/uploads and the servers react to and satisfy these requests.

The earliest peer-to-peer network in widespread use was the Usenet news server system, in which peers communicated with one another to propagate Usenet news articles over the entire Usenet network. Particularly in the earlier days of Usenet, UUCP was used to extend even beyond the Internet. However, the news server system also acted in a client-server form when individual users accessed a local news server to read and post articles. The same consideration applies to SMTP email in the sense that the core email relaying network of Mail transfer agents is a peer-to-peer network while the periphery of Mail user agents and their direct connections is client server.

Some networks and channels such as Napster, OpenNAP and IRC server channels use a client-server structure for some tasks (e.g. searching) and a peer-to-peer structure for others. Networks such as Gnutella or Freenet use a peer-to-peer structure for all purposes, and are sometimes referred to as true peer-to-peer networks, although Gnutella is greatly facilitated by directory servers that inform peers of the network addresses of other peers.

The concept of peer to peer is increasingly evolving to an expanded usage as the relational dynamic active in distributed networks, i.e. not just computer to computer, but human to human. Yochai Benkler has coined the term "commons-based peer production" to denote collaborative projects such as free software. Associated with peer production are the concept of peer governance (referring to the manner in which peer production projects are managed) and peer property (referring to the new type of licenses which recognize individual authorship but not exclusive property rights, such as the GNU General Public License and the Creative Commons License).

NOTES

Classification of peer-to-peer networks

Classification of peer-to-peer networks is according to their degree of centralization:

Pure peer-to-peer :

- Peers act as equals, merging the roles of clients and server.
- There is no central server managing the network.
- There is no central router.

Hybrid peer-to-peer :

- Has a central server that keeps information on peers and responds to requests for that information.
- Peers are responsible for hosting available resources (as the central server does not have them), for letting the central server know what resources they want to share, and for making its shareable resources available to peers that request it.
- Route terminals are used addresses, which are referenced by a set of indices to obtain an absolute address.

Advantages of peer-to-peer networks

An important goal in peer-to-peer networks is that all clients provide resources, including bandwidth, storage space, and computing power. Thus, as nodes arrive and demand on the system increases, the total capacity of the system also increases. This is not true of a client-server architecture with a fixed set of servers, in which adding more clients could mean slower data transfer for all users.

The distributed nature of peer-to-peer networks also increases robustness in case of failures by replicating data over multiple peers, and — in pure P2P systems — by enabling peers to find the data without relying on a centralized index server. In the latter case, there is no single point of failure in the system.

2.8 Security in Distributed Systems

Most computer security uses the access control model , which provides a basis for secrecy and integrity security policies. Figure 2.5 shows the elements of this model:

- ***Principals :*** sources for requests.
- ***Requests :*** to perform operations on objects.
- ***Reference Monitor :*** a guard for each object that examines each request for the object and decides whether to grant it.
- ***Objects :*** resources such as files, devices, or processes.

Advantages of peer-to-peer networks

An important goal in peer-to-peer networks is that all clients provide resources, including bandwidth, storage space, and computing power. Thus, as nodes arrive and demand on the system increases, the total capacity of the system also increases. This is not true of a client-server architecture with a fixed set of servers, in which adding more clients could mean slower data transfer for all users.

NOTES

The distributed nature of peer-to-peer networks also increases robustness in case of failures by replicating data over multiple peers, and — in pure P2P systems — by enabling peers to find the data without relying on a centralized index server. In the latter case, there is no single point of failure in the system.

2.8 Security in Distributed Systems

Most computer security uses the access control model , which provides a basis for secrecy and integrity security policies. Figure 2.5 shows the elements of this model:

- ***Principals :*** sources for requests.
- ***Requests :*** to perform operations on objects.
- ***Reference Monitor :*** a guard for each object that examines each request for the object and decides whether to grant it.
- ***Objects :*** resources such as files, devices, or processes.

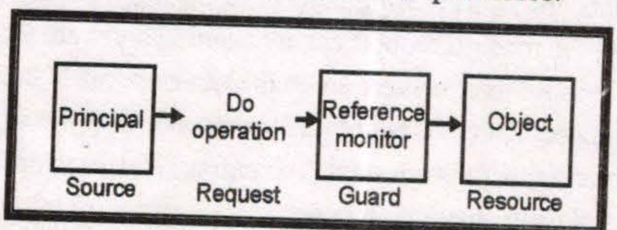


Figure 2.6 The Access Control Model

The reference monitor bases its decision on the principal making the request, the operation in the request, and an access rule that controls which principals may perform that operation on the object.

To do its work the monitor needs a trustworthy way to know both the source of he request and the access rule. Obtaining the source of the request is called authentication'; interpreting the access rule is called 'authorization'.

Thus authentication answers the question "Who said this?", and authorization an-wers the question "Who is trusted to access this?". Usually the access rule is ttached to the object; such a rule is called an access control list or ACL. For each peration the ACL specifies a set of authorized principals, and the monitor grants request if its principal is trusted at least as much as some principal that is autho-ized to do the operation in the request.

request arrives on some channel, such as a wire from a terminal, a network onnection, a pipe, a kernel call from a user process, or the successful decryption f an encrypted message. The monitor must deduce the principal responsible for the equest from the channel it arrives on, that is, it must authenticate the channel. This easy in a centralized system because the operating system implements all the annels and knows the principal responsible for each process. In a distributed stem several things make it harder:

- ***Autonomy :*** The path to the object from the principal ultimately responsible for the request may be long and may involve several machines that are not equally trusted. We might want the authentication to take account of this, say by reporting the principal as "SKPandey working through a remote machine" rather than simply "SKPandey".
- ***Size*** **:** The system may be much larger than a centralized one, and there may be multiple sources of authority for such tasks as registering users.
- ***Heterogeneity*** **:** The system may have different kinds of channels that are secured in different ways. Some examples are encrypted messages, physically secure wires, and interprocess communication done by the operating system.
- ***Fault-tolerance :*** Some parts of the system may be broken, off line, or otherwise inaccessible, but the system is still expected to provide as much service as possible. This is more complicated than a system which is either working or completely broken.

We believe that distributed systems should be designed to make attacks harder and to limit the damage done when attacks succeed.

Virtually all computers and an increasing number of other devices are connected to the Internet. Many of the machines out there are running software that attempts to harm others, either because their owners are malicious or because their owners are careless and have allowed malicious people take control of their machines. To make matters worse, the environment is constantly changing, with machines joining and leaving the system and even changing owners.

The changes in the environment brought about by widespread connectivity mean that builders of distributed systems need to re-examine their assumptions when designing infrastructures for the Internet.

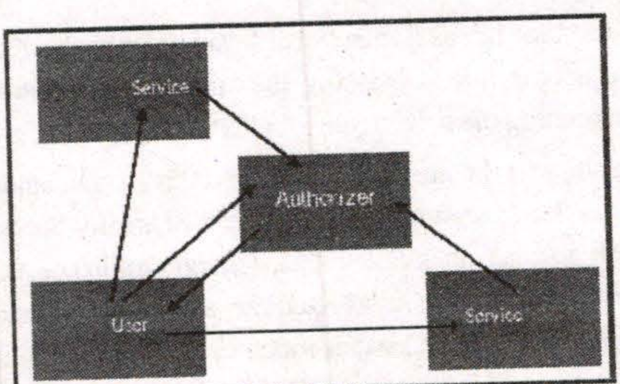


Figure 2.7 Separating Authorization from access control

There is no reason why the mechanism used to grant privileges to a person or a process should be related to the mechanism used to decide whether a particular request should be honored. Security mechanisms that work in the enterprise may be inappropriate in a small business or in a home but it may be costly.

SUMMARY

NOTES

- A *database* is defined as a structured collection of records or data that is stored in a computer so that a program can consult it to answer queries.
- The computer program used to manage and query a database is known as a database management system (DBMS).
- *Distributed Database (DDB) :* A logically interrelated collection of shared data, physically distributed across multiple servers connected by a network.
- *Distributed Database Management System (DDBMS) :* The software system that facilitates the management of a DDB in such a way that the distribution aspects are transparent to users.
- A *homogeneous DDBMS* is one in which the same DBMS is used in all local servers. A homogeneous DDBMS may be administered by a central authority.
- A *heterogeneous DDBMS* uses different DBMSs at the local server.
- Replication of data at additional sites improves the overall reliability of the system.
- Centralized access to data is easily controlled; multiple locations, replicated data and insecure networks all create additional risks.
- Distributed databases can suffer from a variety of problems. Networks can be slow; so it is important to minimize the number and volume of messages between sites. Query processing will often offer a number of options whose execution times may vary by an orders of magnitude. Recovery from failure at any site also requires co-ordinated action in order to avoid loss of integrity or consistency.
- Concurrency: several processes are in between starting and finishing at the same time .
- Parallelism: several processes are executing at the same time .
- Distributed systems are MIMD, i.e., have multiple instruction streams and multiple data streams.
- *Middleware:* Software layer that provides a programming abstraction as well as masking the heterogeneity of the underlying systems (networks, hardware, operating systems, and programming languages): e.g., CORBA and Java RMI.
- *Three-tier* is a client-server architecture in which the user interface, functional process logic ("business rules"), data storage and data access are developed and maintained as independent modules, most often on separate platforms.

NOTES

- The advantages of *n*-tiered architectures is that they are far more scalable, since they balance and distribute the processing load among multiple, often redundant, specialized server nodes.
- A *computer cluster* is a group of tightly coupled computers that work together closely so that in many respects they can be viewed as though they are a single computer.
- Obtaining the source of the request is called 'authentication'; interpreting the access rule is called 'authorization'.

EXERCISE

1. Describe the following in brief:
   - Database
   - Database Management System
   - Distributed Database
   - Distributed Database Management System
2. How distributed systems are different from Centralized systems. Describe different functions of Distributed Database Management System .
3. What is data placement ? Describe different data placement strategies which are in use.
4. How can you categorize the Distributed Database Management System?
5. What are proc and cons behind distributed databases?
6. Describe the role of distributed file system. List its advantages and disadvantages .
7. What is ACID rule for transactional databases ?
8. What are the motivation factors behind the development of distributed systems?
9. Describe Distributed database features and its characteristics .
10. How can you categorize the distributed systems. Describe cluster and grid in detail.
11. Describe different components of reliability in distributed systems .
12. Describe hardware and software requirements for distributed systems.
13. Why there are different architectures exist in distributed systems ? Describe some of the distributed systems architectures in detail.
14. What are High-performance computing (HPC) Clusters? Explain its significance in modern web environment.

NOTES

3 Modern Business Trends : E-Business & E-Commerce

Chapter Includes :

3.1 Introduction

The growth and expansion of the internet as a means for conducting business has indelibly marked the dawn of the 21st Century. For most companies the web serves as a primary source of information and marketing tool for their services. E-catalogues are cost-effective, can be more detailed than paper catalogues, and are friendly to use even for customers with basic Internet skills. Also the web offers opportunities for the 'bundling' of services. Supplier companies, in particular, have the opportunity to enlarge the type and amount of information provided on the web by setting up partnerships with other actors in the value chain or independent companies .In the emerging global economy, E-Commerce and e-business have increasingly become a necessary component of business strategy and a strong catalyst for economic development.

NOTES

The integration of information and communications technology (ICT) in business has revolutionized relationships within organizations and those between and among organizations and individuals. Specifically, the use of ICT in business has enhanced productivity, encouraged greater customer participation, and enabled mass customization, besides reducing costs. Business uses of Information and Communication Technologies (ICTs) to enhance competitiveness and sustain growth, and thereby to create net employment gains, are key expectations of E-Commerce and the new ways of working.

Companies have come to realize that if they're going to respond rapidly and effectively to today's volatile marketplace, they need to do more than web-enable discrete systems, processes or business units. They need to pull together all of the systems they have already got and integrate them securely with their core business activities horizontally, across not just their whole company but their entire value chain, from customers to suppliers. E-Commerce business models are usually based on the network effect. Network effects arise when the value of the e-business application increases with the addition of each participant. Supply and demand do not have the traditional economic relationship. E-business agreements tend to be short in duration with due regard to rapidly changing market conditions and advances in technology. Exponential improvements in technology keep occurring year after year. These differences need to be appreciated.

It is recognized that in the Information Age, Internet commerce is a powerful tool in the economic growth of developing countries. The new technologies are providing another outlet for conducting business, which has its own distinct advantages and drawbacks. E-business technologies can help you become a competitive, Web-enabled business. Growing your business on the Web means rethinking your customer interface and creating a new generation of applications. These applications must allow you to exploit intranets and extranets, providing rapid deployment and easy access to your legacy information— the foundation of your enterprise. They must network your information securely and provide real-time , online transactions.

3.2 Online Shopping

Online shopping is the process, consumers go through, to purchase products or services over the Internet. An online shop, e-shop, internet shop, webshop or online store evokes the physical analogy of buying products or services at a bricks-and-mortar retailer or in a shopping mall.

Online shopping is popular mainly because of its speed and ease of use. Some issues of concern can include fluctuating exchange rates for foreign currencies, local and international laws and delivery methods.

3.2.1 Characteristics of Online Shopping

Product Reviews

Online shoppers can easily learn from previous experiences of different products, mostly by reading user or expert reviews. Reading online product reviews is usually the first step in online shopping, which plays an important role in customers' decisions. Many online stores such as Amazon.com and Newegg today allow customers to comment or rate their items. There are also dedicated review sites (e.g. Epinions, etc.) to host user reviews for different products.

Store Reputation

It is important to do business with reputable online stores to avoid possible Internet fraud and to easily exchange or return when things go wrong. Again, shoppers often read store ratings or reviews by other customers if they are not familiar with some online stores.

Some general guidelines for choosing an online store:

- Honesty in providing product information and availability .
- Speed of orders processing .
- Shipping cost and speed: online package tracking is popular today .
- Return/exchange policy: the time frame customers can return/exchange; who should pay the returning shipping; if there is any restocking fee etc.

Some online stores also provide live chat with their representatives in case customer needs to discuss the product.

Price comparison

An advantage of shopping online is being able to use the power of the internet to seek out the lowest prices or the best deals available for items or services. For example if one is buying a digital camera he/she should enter "digital camera" into a search engine or a price search engine. Most price comparison services have the advantage of store ratings and reviews. Getting the lowest price is important but it is more important to make sure the merchant or store the customer is purchasing from is reputable. There are some online stores that display ridiculously low prices, but when you call them you may find out that the item you're seeking is out of stock. This is called the ***"bait and switch"*** technique. The best way to avoid backorder disappointment is to contact the online store by phone and verify that the product you're seeking is in stock. If you're told that there's only one left, then place your order by phone.

The differences between the online prices and local store prices also rely on the shipping cost and tax. A wise customer would consider the final prices instead of just product prices.

Discounts

There are multiple websites that compile coupon or discount information for most online merchants. Before purchasing online, it is usually worth checking for discounts and coupons. Typing the name of the online merchant along with the word "coupon" in a search engine will provide multiple results. Some of the more popular coupon websites have forums where you can ask for help hunting down a deal if you don't have the time.

Placing online orders

The actual steps of ordering online is simple after the preparing steps of reading reviews and finding a balance between good stores and good prices:

- Browse product categories using a web browser
- Put items into virtual shopping cart (or market basket).

Just as in a physical store viewing the contents of the cart can be done at any time.

NOTES

Quantities of products can be changed or deleted.

- Checkout

  Log in or register by choosing a username and a password. Not all online stores require registration.

  Enter personal data.

  Billing address

  Shipping address (some online stores do not permit a separate shipping address, at least not online)

  Phone number

  E-Mail address (extremely important, as they may have no other way of contacting you if a problem arises)

  Choose means of payment

  Choose delivery speed and method (post, courier and logistics service, etc.)

- Confirm order

  After editing the personal data a confirmation page is displayed so that the online shopper can approve, change or abort the order.

- Logout

Means of Payment

Online shoppers commonly use their credit card for making payments, however some systems enable users to create accounts and pay by alternative means, such as

- Debit card
- Various types of electronic money
- Cash on delivery (C.O.D.) Note: Very few online stores will ship C.O.D.
- Cheque
- Wire transfer/delivery on payment
- Postal money order
- PayPal
- Google Checkout
- Gift Cards
- Direct Debit in some countries

Product Delivery

Once a payment has been accepted the goods or services can be delivered in the following ways.

- Download: This is the method often used for digital media products such as software, music, movies, or images.
- Shipping: The product is shipped to the customer's address.
- In-store pickup: The customer orders online, finds a local store using locator software and picks the product up at the closest store. This is the method often used in the bricks and clicks business model.

Security Issues

- User and payment data is encrypted by SSL (Secure Socket layer)when it is transferred on the Internet.
- Quality seals can be placed on the Shop webpage if it has undergone an independent assessment and meets all requirements of the company issuing the seal. The purpose of these seals is to increase the confidence of the online shoppers; the existence of many different seals foils this effort to a certain extent.
- Privacy of personal information is a big issue. In spite of Privacy Guidelines ,privacy violations still occur and hamper E-Commerce from developing to its full potential.

Setting up a shopping cart system

- Simple systems allow the offline administration of products and categories. The shop is then generated as HTML files and graphics that can be uploaded to a webspace. These systems don't use an online database.
- A high end solution can be bought or rented as a standalone program or as an addition to an ERP program. It is usually installed on the company's own webserver and may integrate very well into the existing supply chain so that ordering, payment, delivery, accounting and warehousing can be automated to a large extent.
- Other solutions allow the user to register and create an online shop on a portal that hosts multiple shops at the same time.
- Open Source solutions can be adapted and installed on a webspace.
- There are also commercial systems that can be tailored to ones needs so that the shop does not have to be created from scratch. By using a framework already existing, software modules for different functionalities required by a webshop can be adapted and combined.
- Following the growth of online retailing activities, there are a number of specialist online retail professional services providers who offer E-Commerce and internet marketing to retailers. This will involve website design and development, integration with order processing and stock systems as well as online advertising services.
- Another option is to open a subscription based store using the technologies from larger companies (eBay or Mega-Bids.com). While they charge for their services, most of the hard work is done for you, they have already lined up payment processors and have tested the store format. They also let you personalize your new store front with logos and by uploading information about your company.

3.3 E-Business

Electronic Business, or "E-Business", may be defined broadly as any business process that relies on an automated information system. Today, this is mostly done with Web-based technologies.

The term "E-Business" was coined by Lou Gerstner, CEO of IBM.

Electronic business methods enable companies to link their internal and external data processing systems more efficiently and flexibly, to work more closely with suppliers

NOTES

and partners, and to better satisfy the needs and expectations of their customers.

In practice, e-business is more than just E-Commerce. While e-business refers to more strategic focus with an emphasis on the functions that occur using electronic capabilities, E-Commerce is a subset of an overall e-business strategy. E-Commerce seeks to add revenue streams using the World Wide Web or the Internet to build and enhance relationships with clients and partners and to improve efficiency using the Empty Vessel strategy. Often, E-Commerce involves the application of knowledge management systems.

E-business involves business processes spanning the entire value chain: electronic purchasing and supply chain management, processing orders electronically, handling customer service, and cooperating with business partners. Special technical standards for e-business facilitate the exchange of data between companies. E-business software solutions allow the integration of intra and inter firm business processes. E-business can be conducted using the Web, the Internet, intranets, extranets, or some combination of these. Following figure shows E-Business cycle :

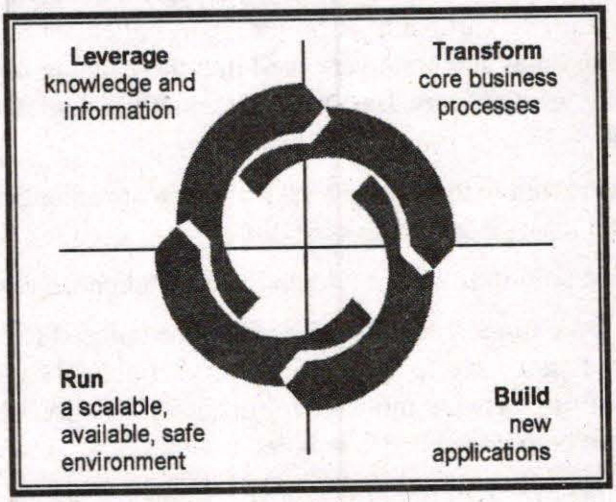


Figure 3.1: E-Business Cycle

When organizations go online, they have to decide which e-business models best suit their goals. A business model is defended as the organization of product, service and information flows, and the source of revenues and benefits for suppliers and customers. The concept of e-business model is the same but used in the online presence. The following is a list of the currently most adopted e-business models:

- E-shops
- E-procurement
- E-malls
- E-auctions
- Virtual Communities
- Collaboration Platforms
- Third-party Marketplaces
- Value-chain Integrators
- Value-chain Service Providers
- Information Brokerage

NOTES

While some use E-Commerce and e-business interchangeably, they are distinct concepts. In E-Commerce, information and communications technology (ICT) is used in inter-business or inter-organizational transactions (transactions between and among firms/organizations) and in business-to-consumer transactions (transactions between firms/organizations and individuals). In e-business, on the other hand, ICT is used to enhance one's business. It includes any process that a business organization (either a for-profit, governmental or non-profit entity) conducts over a computer-mediated network.

A more comprehensive definition of e-business is:

"The transformation of an organization's processes to deliver additional customer value through the application of technologies, philosophies and computing paradigm of the new economy."

Three primary processes are enhanced in e-business:

- **Production processes**, which include procurement, ordering and replenishment of stocks; processing of payments; electronic links with suppliers; and production control processes, among others;
- **Customer-focused processes,** which include promotional and marketing efforts, selling over the Internet, processing of customers' purchase orders and payments, and customer support, among others; and
- **Internal management processes**, which include employee services, training, internal information-sharing, video-conferencing, and recruiting. Electronic applications enhance information flow between production and sales forces to improve sales force productivity. Workgroup communications and electronic publishing of internal business information are likewise made more efficient.

The Internet economy pertains to all economic activities using electronic networks as a medium for commerce or those activities involved in both building the networks linked to the Internet and the purchase of application services such as the provision of enabling hardware and software and network equipment for Web-based/ online retail and shopping malls (or "e-malls"). It is made up of three major segments: physical (ICT) infrastructure, business infrastructure, and commerce.

E-business uses the new family of technologies available on the internet. These technologies enable people to communicate in new ways, provide new business models, permit businesses to operate more efficiently and take advantage of the new global network economy.

The understanding and use of e-business models is essential in an increasingly dynamic and uncertain business environment for the following reasons:

- The process of modeling social systems or an ontology – such as an e-business model – helps identifying and ***understanding*** the relevant elements in a specific domain and the relationships between them.
- The use of formalized e-business models helps managers easily ***communicate and share*** their understanding of an e-business among other stakeholders.
- Mapping and using e-business models as a foundation for discussion facilitates ***change***. Business model designers can easily modify certain elements of an existing e-business model .

NOTES

- A formalized e-business model can help identifying the relevant ***measures*** to follow in an e-business.
- E-Business models can help managers ***simulate*** e-businesses ***and learn*** about them. This is a way of doing risk free experiments, without endangering an organization .

3.3.1 Components of E-Business Model

An E-Business model must have following components :

- A **shared digital business infrastructure**, including digital production and distribution technologies (broadband/wireless networks, content creation technologies and information management systems), which will allow business participants to create and utilize network economies of scale and scope;
- A **sophisticated model for operations,** including integrated value chains-both supplies chains and buys chains;
- An **e-business management model,** consisting of business teams and/or partnerships; and
- **Policy, regulatory and social systems**-i.e., business policies consistent with E-Commerce laws, teleworking/virtual work, distance learning, incentive schemes, among others.

3.3.2 E-Business Trends

As E-Business has taken off over the past few years the different associated domains and technologies have also evolved. We will take a look at some of the major trends associated with E-Business:

- ***Commoditization of Application Servers***

  Application servers provide a base platform for a number of E-Business applications. They provide the plumbing and core set of infrastructure services like session management, transaction management, user management, security, logging, auditing, scalability solutions and ways of encapsulating business services. The J2EE specification is embodied in a number of application servers from vendors like IBM, BEA, Sun, ATG as well as open source solutions like JBoss . Similarly Microsoft has encapsulated a number of services in its .NET framework that provide similar capabilities. The past few years has seen fierce competition in the J2EE application server space with the clear frontrunners being IBM and BEA. However, with the maturity of the underlying specifications, there is very little differentiation that the application server vendors can provide. They are looking to add custom extensions to demonstrate additional value but that could tie you down to a particular vendor. With the recent advances made by the open source frameworks many enterprises are choosing to use JBoss or Tomcat in place of the more costly WebSphere or WebLogic. Understanding the scalability and performance metrics of your E-Business application portfolio is a key factor in deciding which application server to choose.

- **Content Management**

  There is plenty of information residing in any organization. A good content management and access strategy is one of the key E-Business initiatives that an organization should undertake. Such a strategy should involve process, technol-

NOTES

ogy and above all re-use of content. The consumers of the content may be internal, like, employees seeking information on the organization's policies or external, like, customers and partners seeking information about an organization's products. It is all about giving the right information at the right time to the right consumer irrespective of where that information resides in an organization. Many vendors offer a comprehensive content management suite that allows an enterprise to organize the content, define processes and roles in managing the content and provide means for publishing and presenting the content via different channels. Some of the vendors offering such suites include Interwoven, Vignette, Microsoft and IBM. Open source solutions like OpenCMS and Mambo have also found favor amongst organizations that are looking for entry level content management solutions. With majority of the application server vendors realizing the key role of content in any E-Business initiatives, the application servers and content management suites have started to merge over the years. As such vendors like IBM, BEA and Microsoft are offering a comprehensive stack that covers the full series of infrastructure, content and personalization aspects of an E-Business initiative. Taking the content management mantra to an extreme, many organizations end up having multiple repositories for different kinds of content or assets. In such cases, repository consolidation and access to multiple content repositories via an integrated solution becomes a key necessity for new E-Business initiatives.

- **Open Source Frameworks**

  Far from being a collaboration of some geeks around the world, the open source movement has entered the mainstream in the past few years. Many projects have been launched by OSI (Open Source Initiative) and a comprehensive suite of products and solutions are part of the open source environment including (but not limited to):

  - Apache which is one of the most popular Web engines
  - Application server and infrastructure solutions like JBoss and Zope
  - Linux a Unix-like operating system which has been deployed on almost all commercially available platforms. Linux has strong backing by the likes of IBM and HP, which see it as a good counter-balance to the dominance of the Wintel platform.
  - Programming environments like Ruby and Python
  - Integrated development environments like Eclipse
  - Content and collaboration tools like OpenCMS, Red Hat CMS, Wiki and Mambo
  - Database solutions like MySQL

The attraction of free and widely available software may be quite high for an enterprise to refuse. However, before adopting open source solutions as a part of the E-Business initiative, an organization should consider the following aspects:

- ***Support :*** Unlike commercially available software, open source software may not have the requisite support available that is needed for key E-Business initiatives. Things have started to change with vendors like Red Hat and Cova-

NOTES

lent providing support for some of the most popular products like Linux and Apache.

- ***Licensing and Modifications :*** Different open source products come with different licensing mechanisms. Even though most would allow modifications, organizations should be careful in how they modify the software keeping the future upgrades and redistribution of the modifications in mind. Most E-Business initiatives necessitate a comprehensive view of data and functionality across the enterprise. With multiple applications, databases and information silos deployed in an enterprise, integration across the multitude of such applications remains a chief concern for a CIO. Solutions from vendors like webMethods, Tibco, SeeBeyond, Vitria, IBM and Microsoft allow for such integration. A comprehensive enterprise application integration suite would encompass data transformations, reliable messaging, business process mapping, monitoring, workflow and adapters for connecting to various legacy technologies. As this market has matured, various standards have been established . Most vendors are starting to support such standards in addition to the proprietary solutions that they had originally evolved. With the distinct overlap between an integration server and an application server many of the original "pure play" vendors have started offering solutions that encompass the entire infrastructure and integration needs of an E-Business application (Example BEA, IBM and webMethods). Enterprises embarking on an E-Business initiative should evaluate best-of-breed versus full-suite strategy to see which one serves their needs better.
- ***Skills and Training :*** Many open source software products provide only the basic set of documentation and training by 3rd parties is only available for the most popular products.
- ***Service Oriented Architectures and Web Services :*** Services oriented architectures have been around for a while. Web services are one form of implementing a service-oriented architecture (SOA). Due to the widespread agreement around the base Web services standards, SOA has been brought into the mainstream via Web services. Many organizations have already deployed Web services within the enterprise and is now moving to build collaborative applications with business partners using Web services. The applications include a wide range like providing access to legacy systems to front ending business functionality and collaborating with business partners to fulfill orders. Major software vendors like Oracle, SAP and Microsoft are introducing or architecting their products around a SOA-based architecture with Web service implementations. SOA will be a prevailing software engineering practice, ending the 40-year domination of monolithic software. Any new E-Business initiative must encompass SOA as the base architecture for its implementation.
- ***Collaboration - A Key For Successful E-Business Initiatives :*** With teams being distributed around the globe, the different aspects of collaborative software in the form of instant messaging, Web conferencing, application sharing,

white boards, virtual workspaces for teams etc. play an important role in successful delivery of E-Business initiatives. Equipping a team with the right set of collaborative tools ensures that they are able to better share ideas and come out with a more comprehensive and quality solution. The infrastructure for net messaging and Web conferencing improves, team-based real-time collaboration will enter mainstream usage.

- ***Grid Computing :*** Just as the electric power grid led to availability of a cheap and reliable source of power, the computational grid has the potential of changing how we utilize and rely on computational assets as a set of shared resources across a distributed network (like the Internet). Grid computing can be defined as a set of standards and protocols that allow a set of collaborating computing resources in a network to work on a single problem at the same time. An organization may have a set of underutilized resources like Unix servers or File servers; grid computing would allow them to work as a single, highly available unit to solve complex computational problems. Related to the ideas of grid computing is the concept of "utility" computing which promises to deliver on-demand computing capability. Many applications only need high-end computational capability at distinct points in time. The idea would be for an application to be given the amount of resources it requires at a particular point in time via the help of self managing and configurable computational assets. Although grid computing is still in its early adoption phase, it has the potential of fundamentally changing how we utilize computing resources for different applications. Instead of deploying high-end servers to meet a specific application's peak demands, organizations would choose to work with clusters of shared servers and storage devices that are configured at run-time-to serve the needs of multiple applications.
- ***Mainframes :*** Mainframes have played a crucial role in servicing many mission critical applications across a number of organizations. The computational paradigm was generally batch-oriented or tightly controlled by applications that were mostly used by internal users. Many E-Business applications that have been developed still rely on the functionality of the mainframe applications for the business rules. They access the mainframe applications via some gateway or veneers like Web services and provide a friendly user interface to the end user. On the one hand it has given a new lease on life to the Mainframe but it has also exposed a hitherto tightly guarded computational resource to the whims of an Internet user (or a 3rd party application). Consider a hotel reservation system that runs on a mainframe and was mainly used by reservation agents. An E-Business application is developed that exposes the same functionality to users on the Web. This also allows for the number of requests to the mainframe to multiply many-fold and could start impinging on crucial applications sharing the same set of resources. Organizations are dealing with such scenarios by introducing middleware the throttles the number of requests being sent by E-Business applications to such mainframes. It also means that functionalities that were

originally implemented on the mainframes are slowly being migrated out to ***open systems*** that can be more easily scaled to meet the unpredictable demands of E-Business applications.

NOTES

3.4 E-Commerce

Electronic Commerce or E-Commerce refers to a wide range of online business activities for products and services. It also pertains to "any form of business transaction in which the parties interact electronically rather than by physical exchanges or direct physical contact."

E-Commerce is usually associated with buying and selling over the Internet, or conducting any transaction involving the transfer of ownership or rights to use goods or services through a computer-mediated network. Though popular, this definition is not comprehensive enough to capture recent developments in this new and revolutionary business phenomenon.

A more complete definition is:

"E-Commerce is the use of electronic communications and digital information processing technology in business transactions to create, transform, and redefine relationships for value creation between or among organizations, and between organizations and individuals."

Electronic Commerce is exactly analogous to a marketplace on the Internet. Electronic Commerce consists primarily of the distributing, buying, selling, marketing and servicing of products or services over electronic systems such as the Internet and other computer networks. The information technology industry might see it as an electronic business application aimed at commercial transactions; in this context, it can involve electronic funds transfer, supply chain management, e-marketing, online marketing, online transaction processing, electronic data interchange (EDI), automated inventory management systems, and automated data collection systems. Electronic commerce typically uses electronic communications technology of the World Wide Web, at some point in the transaction's lifecycle, although of course electronic commerce frequently depends on computer technologies other than the World Wide Web, such as databases, and e-mail, and on other non-computer technologies, such as transportation for physical goods sold via E-Commerce.

The meaning of the term "electronic commerce" has changed over the last 30 years. Originally, "electronic commerce" meant the facilitation of commercial transactions electronically, usually using technology like Electronic Data Interchange (EDI) and Electronic Funds Transfer (EFT), where both were introduced in the late 1970s, for example, to send commercial documents like purchase orders or invoices electronically.

The 'electronic' or 'e' in E-Commerce refers to the technology/systems; the 'commerce' refers to be traditional business models. E-Commerce is the complete set of processes that support commercial/business activities on a network. In the 1970s and 1980s, this would also have involved information analysis. The growth and acceptance of credit cards, automated teller machines (ATM) and telephone banking in the 1980s were also forms of E-Commerce. However, from the 1990s onwards, this would include enterprise resource planning systems (ERP), data mining and data warehousing.

In the dot com era, it came to include activities more precisely termed "Web commerce" — the purchase of goods and services over the World Wide Web, usually with secure connections (HTTPS, a special server protocol that encrypts confidential ordering data for customer protection) with e-shopping carts and with electronic payment services, like credit card payment authorizations.

Today, it encompasses a very wide range of business activities and processes, from e-banking to offshore manufacturing to e-logistics. The ever growing dependence of modern industries on electronically enabled business processes gave impetus to the growth and development of supporting systems, including backend systems, applications and middleware. Examples are broadband and fiber-optic networks, supply-chain management software, customer relationship management software, inventory control systems and financial accounting software.

When the Web first became well-known among the general public in 1994, many journalists and pundits forecast that E-Commerce would soon become a major economic sector. However, it took about four years for security protocols (like HTTPS) to become sufficiently developed and widely deployed. Subsequently, between 1998 and 2000, a substantial number of businesses in the United States and Western Europe developed rudimentary web sites.

Although a large number of "pure E-Commerce" companies disappeared during the dot-com collapse in 2000 and 2001, many "brick-and-mortar" retailers recognized that such companies had identified valuable niche markets and began to add E-Commerce capabilities to their Web sites. For example, after the collapse of online grocer Webvan, two traditional supermarket chains, Albertsons and Safeway, both started E-Commerce subsidiaries through which consumers could order groceries online.

The emergence of E-Commerce also significantly lowered barriers to entry in the selling of many types of goods; accordingly many small home-based proprietors are able to use the internet to sell goods. Often, small sellers use online auction sites such as EBay, or sell via large corporate websites like Amazon.com, in order to take advantage of the exposure and setup convenience of such sites.

4.1 E-Commerce Transactions

An E-Commerce transaction is concerned with and passes through different parties involved in the payments process . The parties involved in the completion of E-Commerce transaction include:

The User : The party in the E-Commerce transaction who plans to purchase goods or services online .

NOTES

- *The Merchant :* The merchant is the company or business house wishing to sell goods or services to enable customers to by online.
- *The Issuer :* It is a credit card company that uses cards to the users to initiate online buying .
- *The Acquirer :* He is usually the banker of the merchant – i.e. the merchant's financial institution.
- *The Certificate Authority :* It is an independent and third party authority that issues certificates to merchant ,to the issuer and in some cases , the card holder

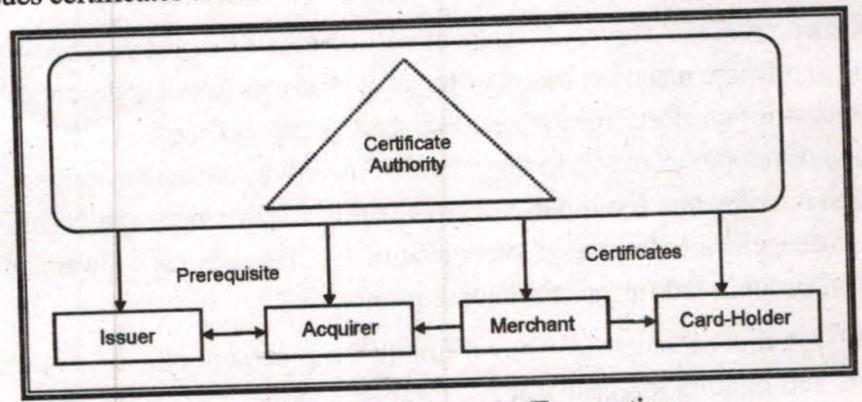


Figure 3.2 : E-Commerce Transaction

3.4.2 Leading with E-Commerce

In many cases, an E-Commerce company will survive not only based on its product, but by having a competent management team, good post-sales services, well-organized business structure, network infrastructure and a secured, well-designed website. A company that wants to succeed will have to perform two things: Technical and organizational aspects and customer-oriented. Following factors will make business of companies succeed in E-Commerce:

Technical And Organizational Aspects

- Sufficient work done in market research and analysis. E-Commerce is not exempt from good business planning and the fundamental laws of supply and demand. Business failure is as much a reality in E-Commerce as in any other form of business.
- A good management team armed with information technology strategy. A company's IT strategy should be a part of the business re-design process.
- Providing an easy and secured way for customers to effect transactions. Credit cards are the most popular means of sending payments on the internet, accounting for 90% of online purchases. In the past, card numbers were transferred securely between the customer and merchant through independent payment gateways. Such independent payment gateways are still used by most small and home businesses. Most merchants today process credit card transactions on site through arrangements made with commercial banks or credit cards companies.
- Providing reliability and security. Parallel servers, hardware redundancy, fail-safe technology, information encryption, and firewalls can enhance this requirement.
- Providing a 360-degree view of the customer relationship, defined as ensuring

NOTES

that all employees, suppliers, and partners have a complete view, and the same view, of the customer. However, customers may not appreciate the big brother experience.

- Constructing a commercially sound business model.
- Engineering an electronic value chain in which one focuses on a "limited" number of core competencies — the opposite of a one-stop shop. (Electronic stores can appear either specialist or generalist if properly programmed.)
- Operating on or near the cutting edge of technology and staying there as technology changes (but remembering that the fundamentals of commerce remain indifferent to technology).
- Setting up an organization of sufficient alertness and activeness to respond quickly to any changes in the economic, social and physical environment.
- Providing an attractive website. The tasteful use of colour, graphics, animation, photographs, fonts, and white-space percentage may aid success in this respect.
- Streamlining business processes, possibly through re-engineering and information technologies.
- Providing complete understanding of the products or services offered, which not only includes complete product information, but also sound advisors and selectors.

Naturally, the E-Commerce vendor must also perform such mundane tasks as being truthful about its product and its availability, shipping reliably, and handling complaints promptly and effectively. A unique property of the Internet environment is that individual customers have access to far more information about the seller than they would find in a brick-and-mortar situation.

Customer-Oriented

A successful E-Commerce organization must also provide an enjoyable and rewarding experience to its customers. Many factors go into making this possible. Such factors include:

- **Providing value to customers.** Vendors can achieve this by offering a product or product-line that attracts potential customers at a competitive price, as in non-electronic commerce.
- **Providing service and performance.** Offering a responsive, user-friendly purchasing experience, just like a flesh-and-blood retailer, may go some way to achieving these goals.
- **Providing an incentive for customers to buy and to return.** Sales promotions to this end can involve coupons, special offers, and discounts. Cross-linked websites and advertising affiliate programs can also help.
- **Providing personal attention.** Personalized web sites, purchase suggestions, and personalized special offers may go some of the way to substituting for the face-to-face human interaction found at a traditional point of sale.
- **Providing a sense of community.** Chat rooms, discussion boards, requesting customer input and loyalty programs (sometimes called affinity programs) can help in this respect.

NOTES

- **Letting customers help themselves.** Provision of a self-serve site, easy to use without assistance, can help in this respect. This implies that all product information is available, cross-sell information, advise for product alternatives, and supplies & accessory selectors.
- **Helping customers do their job of consuming.** E-tailers and online shopping directories can provide such help through sufficient comparative information and good search facilities. Provision of component information and safety-and-health comments may assist e-tailers to define the customers' job.

Problems

Even if a provider of E-Commerce goods and services strictly follows these "key factors" to plan an ideal E-Commerce strategy, problems can still arise. Sources of such problems include:

- **Failure to understand customers, why they buy and how they buy.** Even a product with a sound value proposition can fail if producers and retailers do not understand customer habits, expectations, and motivations. E-Commerce could potentially mitigate this potential problem with proactive and focused marketing research, just as traditional retailers may do.
- **Failure to consider the competitive situation.** One may have the will to construct a viable book e-tailing business model, but lack the capability to compete with others.
- **Inability to predict environmental reaction.** What will competitors do? Will they introduce competitive brands or competitive web sites? Will they supplement their service offerings? Will they try to damage a competitor's site? Will price wars break out? What will the government do? Research into competitors, industries and markets may mitigate some consequences here, just as in non-electronic commerce.
- **Over-estimation of resource competence.** Can staff, hardware, software, and processes handle the proposed strategy? Have e-tailers failed to develop employee and management skills? These issues may call for thorough resource planning and employee training.
- **Failure to coordinate.** If existing reporting and control relationships do not suffice, one can move towards a flat, accountable, and flexible organizational structure, which may or may not aid coordination.
- **Failure to obtain senior management commitment.** This often results in a failure to gain sufficient corporate resources to accomplish a task. It may help to get top management involved right from the start.
- **Failure to obtain employee commitment.** If planners do not explain their strategy well to employees, or fail to give employees the whole picture, then training and setting up incentives for workers to embrace the strategy may assist.
- **Under-estimation of time requirements.** Setting up an E-Commerce venture can take considerable time and money, and failure to understand the timing and sequencing of tasks can lead to significant cost overruns. Basic project planning, critical path, critical chain, or PERT analysis may mitigate such failings. Profitability may have to wait for the achievement of market share.

NOTES

- **Failure to follow a plan.** Poor follow-through after the initial planning, and insufficient tracking of progress against a plan can result in problems. One may mitigate such problems with standard tools: benchmarking, milestones, variance tracking, and penalties and rewards for variances.
- **Becoming the victim of organized crime.** Many syndicates have caught on to the potential of the Internet as a new revenue stream. Two main methods are as follows: (1) Using identity theft techniques like phishing to order expensive goods and bill them to some innocent person, then liquidating the goods for quick cash; (2) Extortion by using a network of compromised "zombie" computers to engage in distributed denial of service attacks against the target Web site until it starts paying protection money.
- **Failure to expect the unexpected.** Too often new businesses do not take into account the amount of time, money or resources needed to complete a project and often find themselves without the necessary components to become successful.

Product Suitability

Certain products or services appear more suitable for online sales; others remain more suitable for offline sales.

Many successful purely virtual companies deal with digital products, (including information storage, retrieval, and modification), music, movies, office supplies, education, communication, software, photography, and financial transactions. Examples of this type of company include: Google, eBay and Paypal. Other successful marketers such as use Drop shipping or Affiliate marketing techniques to facilitate transactions of tangible goods without maintaining real inventory. Examples include numerous sellers on eBay.

Virtual marketers can sell some non-digital products and services successfully. Such products generally have a high value-to-weight ratio, they may involve embarrassing purchases, they may typically go to people in remote locations, and they may have shut-ins as their typical purchasers. Items which can fit through a standard letterbox — such as music CDs, DVDs and books — are particularly suitable for a virtual marketer, and indeed Amazon.com, one of the few enduring dot-com companies, has historically concentrated on this field.

Products such as spare parts, both for consumer items like washing machines and for industrial equipment like centrifugal pumps, also seem good candidates for selling online. Retailers often need to order spare parts specially, since they typically do not stock them at consumer outlets — in such cases, E-Commerce solutions in spares do not compete with retail stores, only with other ordering systems. A factor for success in this niche can consist of providing customers with exact, reliable information about which part number their particular version of a product needs, for example by providing parts lists keyed by serial number.

There are also many disadvantages of E-Commerce, one of the main ones is fraud. This is where your details (name, bank card number, age, national insurance number) are entered into what look to be a safe site but really it is not. These details can then be used to steal money from you and can be used to buy things on line that you are completely unaware of until it is too late. If this information is leaked into the wrong

NOTES

hands. People are able to steal your identity, and commit more fraud crimes under your name.

Products less suitable for E-Commerce include products that have a low value-to-weight ratio, products that have a smell, taste, or touch component, products that need trial fittings — most notably clothing — and products where colour integrity appears important. Nonetheless, Tesco.com has had success delivering groceries in the UK, albeit that many of its goods are of a generic quality, and clothing sold through the internet is big business in the U.S. Also, the recycling program Cheapcycle sells goods over the internet, but avoids the low value-to-weight ratio problem by creating different groups for various regions, so that shipping costs remain low.

Acceptance

Consumers have accepted the E-Commerce business model less readily than its proponents originally expected. Even in product categories suitable for E-Commerce, electronic shopping has developed only slowly. Several reasons might account for the slow uptake, including:

- **Concerns about security.** Many people will not use credit cards over the Internet due to concerns about theft and credit card fraud.
- **Lack of instant gratification with most e-purchases (non-digital purchases).** Much of a consumer's reward for purchasing a product lies in the instant gratification of using and displaying that product. This reward does not exist when one's purchase does not arrive for days or weeks.
- The problem of access to web commerce, mainly for poor households and for developing countries. Low penetration rates of Internet access in some sectors greatly reduce the potential for E-Commerce.
- **The social aspect of shopping.** Some people enjoy talking to sales staff, to other shoppers, or to their cohorts: this social reward side of retail therapy does not exist to the same extent in online shopping.
- Poorly designed, bug-infested E-Commerce web sites that frustrate online shoppers and drive them away.
- Inconsistent return policies among e-tailers or difficulties in exchange/return.

3.4.3 Components of E-Commerce

E-Commerce does not refer merely to a firm putting up a Web site for the purpose of selling goods to buyers over the Internet. For E-Commerce to be a competitive alternative to traditional commercial transactions and for a firm to maximize the benefits of E-Commerce, a number of technical as well as enabling issues have to be considered. A typical E-Commerce transaction loop involves the following major players and corresponding requisites:

- The **Seller** should have the following components:
  - A corporate Web site with E-Commerce capabilities (e.g., a secure transaction server);
  - A corporate intranet so that orders are processed in an efficient manner; and

NOTES

- IT-literate employees to manage the information flows and maintain the E-Commerce system.

o **Transaction partners** include:

- Banking institutions that offer transaction clearing services (e.g., processing credit card payments and electronic fund transfers);
- National and international freight companies to enable the movement of physical goods within, around and out of the country. For business-to-consumer transactions, the system must offer a means for cost-efficient transport of small packages (such that purchasing books over the Internet, for example, is not prohibitively more expensive than buying from a local store); and
- Authentication authority that serves as a trusted third party to ensure the integrity and security of transactions.

o **Consumers** (in a business-to-consumer transaction) who:

- Form a critical mass of the population with access to the Internet and disposable income enabling widespread use of credit cards; and
- Possess a mindset for purchasing goods over the Internet rather than by physically inspecting items.

o **Firms/Businesses** (in a business-to-business transaction) that together form a critical mass of companies (especially within supply chains) with Internet access and the capability to place and take orders over the Internet.

o **Government,** to establish:

- A legal framework governing E-Commerce transactions (including electronic documents, signatures, and the like); and
- Legal institutions that would enforce the legal framework (i.e., laws and regulations) and protect consumers and businesses from fraud, among others.

o The **Internet,** the successful use of which depends on the following:

- A robust and reliable Internet infrastructure; and
- A pricing structure that doesn't penalize consumers for spending time on and buying goods over the Internet (e.g., a flat monthly charge for both ISP access and local phone calls).

For E-Commerce to grow, the above requisites and factors have to be in place. The least developed factor is an impediment to the increased uptake of E-Commerce as a whole. For instance, a country with an excellent Internet infrastructure will not have high E-Commerce figures if banks do not offer support and fulfillment services to E-Commerce transactions. In countries that have significant E-Commerce figures, a positive feedback loop reinforces each of these factors.

NOTES

3.4.4 Internet and E-Commerce

The Internet allows people from all over the world to get connected inexpensively and reliably. As a technical infrastructure, it is a global collection of networks, connected to share information using a common set of protocols. Also, as a vast network of people and information, the Internet is an enabler for E-Commerce as it allows businesses to showcase and sell their products and services online and gives potential customers, prospects, and business partners access to information about these businesses and their products and services that would lead to purchase. Before the Internet was utilized for commercial purposes, companies used private networks-such as the EDI or Electronic Data Interchange-to transact business with each other. That was the early form of E-Commerce. However, installing and maintaining private networks was very expensive. With the Internet, E-Commerce spread rapidly because of the lower costs involved and because the Internet is based on open standards.

3.4.5 Intranet and E-Commerce

An intranet aids in the management of internal corporate information that may be interconnected with a company's E-Commerce transactions (or transactions conducted outside the intranet). As the intranet allows for the instantaneous flow of internal information, vital information is simultaneously processed and matched with data flowing from external E-Commerce transactions, allowing for the efficient and effective integration of the corporation's organizational processes. In this context, corporate functions, decisions and processes involving E-Commerce activities are more coherent and organized. The growth of intranets has caused a shift from a hierarchical command-and control organization to an information-based organization. This shift has implications for managerial responsibilities, communication and information flows, and workgroup structures.

3.4.6 Advantages of E-Commerce

- **E-Commerce serves as an "equalizer".** It enables start-up and small- and medium- sized enterprises to reach the global market.
- **E-Commerce makes "mass customization" possible.** E-Commerce applications in this area include easy-to-use ordering systems that allow customers to choose and order products according to their personal and unique specifications. For instance, a car manufacturing company with an E-Commerce strategy allowing for online orders can have new cars built within a few days (instead of the several weeks it currently takes to build a new vehicle) based on customer's specifications. This can work more effectively if a company's manufacturing process is advanced and integrated into the ordering system.
- **E-Commerce allows "network production."** This refers to the parceling out of the production process to contractors who are geographically dispersed but who are connected to each other via computer networks. The benefits of network production include: reduction in costs, more strategic target marketing, and the facilitation of selling add-on products, services, and new systems when they are needed. With network production, a company can assign tasks within its non-core competencies to factories all over the world that specialize in such tasks (e.g., the assembly of specific components).

NOTES

- **E-Commerce helpful to the consumer**

In C2B transactions, customers/consumers are given more influence over what and how products are made and how services are delivered, thereby broadening consumer choices. E-Commerce allows for a faster and more open process, with customers having greater control. E-Commerce makes information on products and the market as a whole readily available and accessible, and increases price transparency, which enable customers to make more appropriate purchasing decisions.

- **Business relationships through E- Commerce**

E-Commerce transforms old economy relationships (vertical/linear relationships) to new economy relationships characterized by end-to-end relationship management solutions (integrated or extended relationships)

- **Link customers, workers, suppliers, distributors and competitors**

E-Commerce facilitates organization networks, wherein small firms depend on "partner" firms for supplies and product distribution to address customer demands more effectively. To manage the chain of networks linking customers, workers, suppliers, distributors, and even competitors, an integrated or extended supply chain management solution is needed. **Supply chain management (SCM)** is defined as the supervision of materials, information, and finances as they move from supplier to manufacturer to wholesaler to retailer to consumer. It involves the coordination and integration of these flows both within and among companies. The goal of any effective supply chain management system is timely provision of goods or services to the next link in the chain (and ultimately, the reduction of inventory within each link).

There are three main flows in SCM, namely:

✓ The product flow, which includes the movement of goods from a supplier to a customer, as well as any customer returns or service needs;

✓ The information flow, which involves the transmission of orders and the update of the status of delivery; and

✓ The finances flow, which consists of credit terms, payment schedules, and consignment and title ownership arrangements.

Some SCM applications are based on open data models that support the sharing of data both inside and outside the enterprise, called the extended enterprise, and include key suppliers, manufacturers, and end customers of a specific company . Shared data resides in diverse database systems, or data warehouses, at several different sites and companies. Sharing this data "upstream" (with a company's suppliers) and "downstream" (with a company's clients) allows SCM applications to improve the time-to-market of products and reduce costs. It also allows all parties in the supply chain to better manage current resources and plan for future needs.

NOTES

3.4.7 Driving Forces behind E-Commerce

There are at least three major forces fuelling E-Commerce: economic forces, marketing and customer interaction forces, and technology, particularly multimedia convergence.

- **Economic forces**. One of the most evident benefits of E-Commerce is economic efficiency resulting from the reduction in communications costs, low-cost technological infrastructure, speedier and more economic electronic transactions with suppliers, lower global information sharing and advertising costs, and cheaper customer service alternatives. Economic integration is either external or internal. External integration refers to the electronic networking of corporations, suppliers, customers/clients, and independent contractors into one community communicating in a virtual environment (with the Internet as medium). Internal integration, on the other hand, is the networking of the various departments within a corporation, and of business operations and processes. This allows critical business information to be stored in a digital form that can be retrieved instantly and transmitted electronically. Internal integration is best illustrated by corporate intranets. Among the companies with efficient corporate intranets are Procter and Gamble, IBM, Nestle and Intel.
- **Market forces.** Corporations are encouraged to use E-Commerce in marketing and promotion to capture international markets, both big and small. The Internet is likewise used as a medium for enhanced customer service and support. It is much easier for companies to provide their target consumers with more detailed product and service information using the Internet.
- **Technology forces.** The development of ICT is a key factor in the growth of E-Commerce. For instance, technological advances in digitizing content, compression and the promotion of open systems technology have paved the way for the convergence of communication services into one single platform. This in turn has made communication more efficient, faster, easier, and more economical as the need to set up separate networks for telephone services, television broadcast, cable television, and Internet access is eliminated. From the standpoint of firms/businesses and consumers, having only one information provider means lower communications costs. Moreover, the principle of universal access can be made more achievable with convergence. At present the high costs of installing landlines in sparsely populated rural areas is a disincentive to telecommunications companies to install telephones in these areas. Installing landlines in rural areas can become more attractive to the private sector if revenues from these landlines are not limited to local and long distance telephone charges, but also include cable TV and Internet charges. This development will ensure affordable access to information even by those in rural areas and will spare the government the trouble and cost of installing expensive landlines.

3.4.8 Types of E-Commerce

Depending upon the nature of transaction and parties involved, there are many type of E-Commerce exists: The major types of E-Commerce are:

- Business-to-Business (B2B)
- Business-to-Consumer(B2C)

- Consumer-to-Business(C2B)
- Consumer-to-Consumer (C2C)
- Business-to-Government(B2G)
- Government-to-Business(G2B)
- Government-to-Citizen(G2C)
- Mobile Commerce (M-Commerce)

NOTES

Business-to-Business (B2B) E-Commerce

B2B E-Commerce is simply defined as E-Commerce between companies. This is the type of E-Commerce that deals with relationships between and among businesses.Most experts predict that B2B E-Commerce will continue to grow faster than the B2C segment. The B2B market has two primary components:

- E-infrastructure and
- E-markets

E-infrastructure is the architecture of B2B, primarily consisting of the following:

- Logistics - transportation, warehousing and distribution (e.g., Procter and Gamble);
- Application service providers - deployment, hosting and management of packaged software from a central facility (e.g., Oracle and Linkshare);
- Outsourcing of functions in the process of E-Commerce, such as Web-hosting, security and customer care solutions (e.g., outsourcing providers such as eShare, NetSales, iXL Enterprises and Universal Access);
- Auction solutions software for the operation and maintenance of real-time auctions on the Internet (e.g., Moai Technologies and OpenSite Technologies);
- Content management software for the facilitation of Web site content management and delivery (e.g., Interwoven and ProcureNet); and
- Web-based commerce enablers (e.g., Commerce One, a browser-based, XMLenabled purchasing automation software).

E-markets are simply defined as Web sites where buyers and sellers interact with each other and conduct transactions.

The more common B2B examples and best practice models are IBM, Hewlett Packard (HP), Cisco and Dell. Cisco, for instance, receives over 90% of its product orders over the Internet.

Most B2B applications are in the areas of supplier management (especially purchase order processing), inventory management (i.e., managing order-ship-bill cycles), distribution management (especially in the transmission of shipping documents), channel management (i.e., information dissemination on changes in operational conditions), and payment management (e.g., electronic payment systems or EPS).

Benefits of B2B E-Commerce

The impact of B2B markets on the economy of developing countries is evident in the following:

NOTES

- **Transaction costs:** There are three cost areas that are significantly reduced through the conduct of B2B E-Commerce. First is the reduction of search costs, as buyers need not go through multiple intermediaries to search for information about suppliers, products and prices as in a traditional supply chain. In terms of effort, time and money spent, the Internet is a more efficient information channel than its traditional counterpart. In B2B markets, buyers and sellers are gathered together into a single online trading community, reducing search costs even further. Second is the reduction in the costs of processing transactions (e.g. invoices, purchase orders and payment schemes), as B2B allows for the automation of transaction processes and therefore, the quick implementation of the same compared to other channels (such as the telephone and fax). Efficiency in trading processes and transactions is also enhanced through the B2B e-market's ability to process sales through online auctions. Third, online processing improves inventory management and logistics.
- **Disintermediation:** Through B2B e-markets, suppliers are able to interact and transact directly with buyers, thereby eliminating intermediaries and distributors. However, new forms of intermediaries are emerging. For instance, e-markets themselves can be considered as intermediaries because they come between suppliers and customers in the supply chain.
- **Transparency in pricing**: Among the more evident benefits of e-markets is the increase in price transparency. The gathering of a large number of buyers and sellers in a single e-market reveals market price information and transaction processing to participants. The Internet allows for the publication of information on a single purchase or transaction, making the information readily accessible and available to all members of the e-market. Increased price transparency has the effect of pulling down price differentials in the market. In this context, buyers are provided much more time to compare prices and make better buying decisions. Moreover, B2B e-markets expand borders for dynamic and negotiated pricing wherein multiple buyers and sellers collectively participate in price-setting and two-way auctions. In such environments, prices can be set through automatic matching of bids and offers. In the e-marketplace, the requirements of both buyers and sellers are thus aggregated to reach competitive prices, which are lower than those resulting from individual actions.
- **Economies of scale and network effects:** The rapid growth of B2B e-markets creates traditional supply-side cost-based economies of scale. Furthermore, the bringing together of a significant number of buyers and sellers provides the demand-side economies of scale or network effects. Each additional incremental participant in the e-market creates value for all participants in the demand side. More participants form a critical mass, which is key in attracting more users to an e-market.

Business-to-Consumer(B2C) E-Commerce

Business-to-Consumer is a form of electronic commerce in which products or services are sold from a firm or company to a consumer. Business-to-consumer E-Commerce, or commerce between companies and consumers, involves customers gathering information; purchasing physical goods (i.e., tangibles such as books or consumer products) or information goods (or goods of electronic material or digi-

NOTES

tized content, such as software, or e-books); and, for information goods, receiving products over an electronic network.

It is the second largest and the earliest form of E-Commerce. Its origins can be traced to online retailing (or e-tailing). Thus, the more common B2C business models are the online retailing companies such as Amazon.com, Drugstore.com, Beyond. com, Barnes and Noble and ToysRus.

Other B2C examples involving information goods are E-Trade and Travelocity. The more common applications of this type of E-Commerce are in the areas of purchasing products and information, and personal finance management, which pertains to the management of personal investments and finances with the use of online banking tools (e.g., Quicken). B2C E-Commerce reduces transactions costs (particularly search costs) by increasing consumer access to information and allowing consumers to find the most competitive price for a product or service. B2C E-Commerce also reduces market entry barriers since the cost of putting up and maintaining a Web site is much cheaper than installing a "brick-and-mortar" structure for a firm. In the case of information goods, B2C E-Commerce is even more attractive because it saves firms from factoring in the additional cost of a physical distribution network. Moreover, for countries with a growing and robust Internet population, delivering information goods becomes increasingly feasible.

E-tailing

E-tailing (or electronic retailing) is the selling of retail goods on the Internet. It is the most common form of business-to-consumer (B2C) transaction.

Classifications of B2C E-Commerce

- **Online intermediaries**

  Online intermediaries are companies that facilitate transactions between buyers and sellers and receive a percentage of the transaction's value. These firms make up the largest group of B2C companies today. There are two types of online intermediaries: *brokers* and *infomediaries*.

  An infomediary is a Web site that provides specialized information on behalf of producers of goods and services and their potential customers.

- **Advertising-based models**

  In an advertising-based system, businesses' websites have ad inventory, which they sell to interested parties. There are two guiding philosophies for this practice: *high-traffic* or *niche*. Advertisers take a *high-traffic* approach when attempting to reach a larger audience. These advertisers are willing to pay a premium for a site that can deliver high numbers, for example advertisements on Yahoo! or AOL. When advertisers are trying to reach a smaller group of buyers, they take a *niche* approach. These buyers are well-defined, clearly identified, and desirable. The niche approach focuses on quality, not quantity. For example, an advertisement on WSJ.com would chiefly be viewed by business people and executives.

- **Community-based models**

  In a community-based system, companies allow users worldwide access to interact with each other on the basis of similar areas of interest. These firms make money by accumulating loyal users and targeting them with advertising.

NOTES

Example: Yahoo! Groups

- **Fee-based models**

  In a fee-based system, a firm charges a subscription fee to view its content. There are varying degrees of content restriction and subscription types ranging from flat-fees to pay-as-you-go.

Advantages of B2C E-Commerce

B2C E-Commerce has the following advantages:

- Shopping can be faster and more convenient.
- Offerings and prices can change instantaneously.
- Call centers can be integrated with the website.
- Broadband telecommunications will enhance the buying experience.

Challenges for B2C E-Commerce

The two main challenges faced by B2C E-Commerce are *building traffic* and sustaining *customer loyalty*. Due to the winner-take-all nature of the B2C structure, many smaller firms find it difficult to enter a market and remain competitive. In addition, online shoppers are very price-sensitive and are easily lured away, so acquiring and keeping new customers is difficult.

Essentially, these masters of B2C E-Commerce (eBay, Amazon, etc.) remain at the top because of effective communication and value to the customer.

Consumer-to-Business(C2B) E-Commerce

Consumer-to-business (C2B) transactions involve reverse auctions, which empower the consumer to drive transactions. A concrete example of this when competing airlines gives a traveler best travel and ticket offers in response to the traveler's post that he/she wants to fly .

Consumer-to-Consumer (C2C) E-Commerce

Consumer-to-Consumer E-Commerce or C2C is simply commerce between private individuals or consumers. This type of E-Commerce is characterized by the growth of electronic marketplaces and online auctions, particularly in vertical industries where firms/businesses can bid for what they want from among multiple suppliers. It perhaps has the greatest potential for developing new markets.

Consumer-to-Consumer (or C2C) electronic commerce involves the electronically-facilitated transactions between consumers through some third party. A common example is the online auction, in which a consumer posts an item for sale and other consumers bid to purchase it; the third party generally charges a flat fee or commission. The sites are only intermediaries, just there to match consumers. They do not have to check quality of the products being offered.

Advantages of C2C

- Paypal set up directly for this reason .
- Broader market
- Eliminates intermediary
- Constantly changing, updating
- Always there so that consumers can use it whenever they want .

Disadvantages of C2C

- No quality control
- No payment guarantee
- Hard to pay for using cheques, ATM cards, etc. but in the future this is likely to change.

C2C E-Commerce comes in at least three forms:

- Auctions facilitated at a portal, such as eBay, which allows online real-time bidding on items being sold in the Web;
- Peer-to-Peer systems, such as the Napster model (a protocol for sharing files between users used by chat forums similar to IRC) and other file exchange and later money exchange models; and
- Classified ads at portal sites such as Excite Classifieds and eWanted (an interactive, online marketplace where buyers and sellers can negotiate and which features "Buyer Leads & Want Ads").

There is little information on the relative size of global C2C E-Commerce. However, C2C figures of popular C2C sites such as eBay and Napster indicate that this market is quite large. These sites produce millions of dollars in sales every day.

Business-to-Government(B2G) E-Commerce

Business-to-Government E-Commerce or B2G is generally defined as commerce between companies and the public sector. It refers to the use of the Internet for public procurement, licensing procedures, and other government-related operations. This kind of E-Commerce has two features: first, the public sector assumes a pilot/ leading role in establishing E-Commerce; and second, it is assumed that the public sector has the greatest need for making its procurement system more effective. Web-based purchasing policies increases the transparency of the procurement process (and reduces the risk of irregularities). To date, however, the size of the B2G E-Commerce market as a component of total E-Commerce is insignificant, as government e-procurement systems remain undeveloped.

Government-to-Business(G2B) E-Commerce

Government-to-Business (G2B) is the online non-commercial interaction between local and central government and the commercial business sector, rather than private individuals G2C. For example [http://www.dti.gov.uk] is a Government web site where businesses can get information and advice on e-business 'best practice'.

Government-to-Citizen(G2C) E-Commerce

Government-to-Citizen (G2C) is the online non-commercial interaction between local and central Government and private individuals, rather than the commercial business sector G2B. For example Government sectors become visibly open to the public domain via a Web Portal. Thus making public services and information accessible to all. One such web portal is Government Gateway (http://www.gateway.gov.uk) .

Mobile Commerce (M-Commerce)

M-Commerce (Mobile Commerce) is the buying and selling of goods and services through wireless technology-i.e., handheld devices such as cellular telephones and personal digital assistants (PDAs). Japan is seen as a global leader in M-Commerce.

As content delivery over wireless devices becomes faster, more secure, and scalable, some believe that m-commerce will surpass wireline E-Commerce as the method of choice for digital commerce transactions. This may well be true for the Asia-Pacific where there are more mobile phone users than there are Internet users.

NOTES

Industries affected by M-Commerce include:

- **Financial services**, including mobile banking (when customers use their handheld devices to access their accounts and pay their bills), as well as brokerage services (in which stock quotes can be displayed and trading conducted from the same handheld device);
- **Telecommunications,** in which service changes, bill payment and account reviews can all be conducted from the same handheld device;
- **Service/Retail,** as consumers are given the ability to place and pay for orders on-the-fly; and
- **Information services**, which include the delivery of entertainment, financial news, sports figures and traffic updates to a single mobile device.

3.5 Buying and Paying Online

In most developing countries, the payment schemes available for online transactions are the following:

3.5.1 Traditional Payment Methods

- **Cash-on-delivery.** Many online transactions only involve submitting purchase orders online. Payment is by cash upon the delivery of the physical goods.
- **Bank payments.** After ordering goods online, payment is made by depositing cash into the bank account of the company from which the goods were ordered.Delivery is likewise done the conventional way.

3.5.2 Electronic Payment Methods

- **Innovations affecting consumers**, include credit and debit cards, automated teller machines (ATMs), stored value cards, and e-banking.
- **Innovations enabling online commerce** are e-cash, e-checks, smart cards, and encrypted credit cards. These payment methods are not too popular in developing countries. They are employed by a few large companies in specific secured channels on a transaction basis.
- **Innovations affecting companies** pertain to payment mechanisms that banks provide their clients, including inter-bank transfers through automated clearing houses allowing payment by direct deposit.

3.6 Electronic Payment System

An electronic payment system (EPS) is a system of financial exchange between buyers and sellers in the online environment that is facilitated by a digital financial instrument (such as encrypted credit card numbers, electronic checks, or digital cash) backed by a bank, an intermediary, or by legal tender.

EPS plays an important role in E-Commerce because it closes the E-Commerce loop. In developing countries, the underdeveloped electronic payments system is a serious impediment to the growth of E-Commerce. In these countries, entrepreneurs are not able to accept credit card payments over the Internet due to legal and business concerns. The primary issue is transaction security. The absence or inadequacy of legal infrastructures governing the operation of e-payments is also a concern. Hence, banks with e-banking operations employ service agreements between themselves and their clients. The relatively undeveloped credit card industry in many developing countries is also a barrier to E-Commerce. Only a small segment of the population can buy goods and services over the Internet due to the small credit card market base. There is also the problem of the requirement of "explicit consent" (i.e., a signature) by a card owner before a transaction is considered valid-a requirement that does not exist in the U.S. and in other developed countries.

3.6.1 Confidence level of Consumers in EPS

Many developing countries are still cash-based economies. Cash is the preferred mode of payment not only on account of security but also because of anonymity, which is useful for tax evasion purposes or keeping secret what one's money is being spent on. For other countries, security concerns have a lot to do with a lack of a legal framework for adjudicating fraud and the uncertainty of the legal limit on the liability associated with a lost or stolen credit card.

In sum, among the relevant issues that need to be resolved with respect to EPS are: consumer protection from fraud through efficiency in record-keeping; transaction privacy and safety, competitive payment services to ensure equal access to all consumers, and the right to choice of institutions and payment methods. Legal frameworks in developing countries should also begin to recognize electronic transactions and payment schemes.

3.7 Online Publishing

Online publishing is the process of using computer and specific types of software to combine text and graphics to produce Web-based documents such as newsletters, online magazines and databases, brochures and other promotional materials, books, and the like, with the Internet as a medium for publication.

The problems in online publishing can be grouped into two categories: management challenges and public policy issues.

There are two major management issues:

- The **profit question**, which seeks to address how an online presence can be turned into a profitable one and what kind of business model would result in the most revenue; and
- The **measurement issue,** which pertains to the effectiveness of a Web site and the fairness of charges to advertisers.

The most common public policy issues have to do with copyright protection and censorship. Many publishers are prevented from publishing online because of inadequate copyright protection. An important question to be addressed is: How can exist copyright protections in the print environment are mapped onto the online environment? Most of the solutions are technological rather than legal. The more common technological solutions include encryption for paid subscribers, and information usage meters on add-in circuit boards and sophisticated document headers that monitor the frequency and manner by which text is viewed and used.

NOTES

Advantages of online publishing

Among the benefits of using online media are low-cost universal access, the independence of time and place, and ease of distribution. These are the reasons why the Internet is regarded as an effective marketing outreach medium and is often used to enhance information service.

SUMMARY

- The integration of information and communications technology (ICT) in business has revolutionized relationships within organizations and those between and among organizations and individuals. Specifically, the use of ICT in business has enhanced productivity, encouraged greater customer participation, and enabled mass customization, besides reducing costs.
- It is recognized that in the Information Age, Internet commerce is a powerful tool in the economic growth of developing countries.
- E-business technologies can help you become a competitive, Web-enabled business. Growing your business on the Web means rethinking your customer interface and creating a new generation of applications.
- Online shopping is the process consumers go through to purchase products or services over the Internet.
- Online shoppers can easily learn from previous experiences of different products, mostly by reading user or expert reviews.
- User and payment data is encrypted by SSL (Secure Socket layer)when it is transferred on the Internet.
- Electronic Business, or "E-Business", may be defined broadly as any business process that relies on an automated information system. Today, this is mostly done with Web-based technologies.
- E-business involves business processes spanning the entire value chain: electronic purchasing and supply chain management, processing orders electronically, handling customer service, and cooperating with business partners.
- Application servers provide a base platform for a number of E-Business applications.
- Electronic Commerce or E-Commerce refers to a wide range of online business activities for products and services.
- An E-Commerce transaction is concerned with and passes through different parties involved in the payments process .
- In many cases, an E-Commerce company will survive not only based on its product, but by having a competent management team, good post-sales services, well-organized business structure, network infrastructure and a secured, well-designed website.
- A successful E-Commerce organization must also provide an enjoyable and rewarding experience to its customers.
- B2B E-Commerce is simply defined as E-Commerce between companies.

NOTES

- Business-to-Consumer is a form of electronic commerce in which products or services are sold from a firm or company to a consumer.
- E-tailing (or electronic retailing) is the selling of retail goods on the Internet. It is the most common form of business-to-consumer (B2C) transaction.
- Consumer-to-Consumer E-Commerce or C2C is simply commerce between private individuals or consumers.
- Business-to-Government E-Commerce or B2G is generally defined as commerce between companies and the public sector.
- Government-to-Business (G2B) is the online non-commercial interaction between local and central government and the commercial business sector, rather than private individuals G2C.
- Government-to-Citizen (G2C) is the online non-commercial interaction between local and central Government and private individuals, rather than the commercial business sector G2B.
- M-Commerce (Mobile Commerce) is the buying and selling of goods and services through wireless technology-i.e., handheld devices such as cellular telephones and personal digital assistants (PDAs).
- An electronic payment system (EPS) is a system of financial exchange between buyers and sellers in the online environment that is facilitated by a digital financial instrument (such as encrypted credit card numbers, electronic checks, or digital cash) backed by a bank, an intermediary, or by legal tender.
- Online publishing is the process of using computer and specific types of software to combine text and graphics to produce Web-based documents such as newsletters, online magazines and databases, brochures and other promotional materials, books, and the like, with the Internet as a medium for publication.

EXERCISE

1. What is the role of ICT in E-Business and E-Commerce?
2. What is Online Shopping ? Describe its different characteristics .
3. What are Means of Payment in Online shopping?
4. How security issues are playing very important role in online shopping .Describe with suitable example ..
5. Describe the steps in setting up a shopping cart system
6. What is E-Business ? Describe E-Business cycle .
7. Describe in brief the different adopted e-business models .
8. How many primary processes are enhanced in e-business? Describe in brief.
9. Describe different Components of E-Business Model .
10. What are new trends in E-Business.
11. What is E-Commerce . How it is different from E-Business ?
12. Describe E-Commerce transaction steps .

NOTES

13. Describe different components of E-Commerce in brief.
14. Expalin the relationship of internet ,intranet and E-Commerce.
15. Describe different advantages of E-Commerce over tradition business systems .
16. What are deriving forces behind the E-Coomerce?
17. Describe different types of E-Commerce in detail .
18. What is M-Commerce ? What are the industries affected by M-Commerce?
19. How important is an intranet for a business engaging in e-commerce?
20. How are business relationships transformed through e-commerce?
21. How is e-commerce helpful to the consumer?
22. Aside from reducing the cost of doing business, what are the advantages of e-commerce for businesses?
23. How does e-commerce link customers, workers, suppliers, distributors and competitors?
24. What are the existing practices in developing countries with respect to buying and paying online?
25. What is an electronic payment system? Why is it important?
26. What is the confidence level of consumers in the use of an EPS?
27. What is e-banking?
28. What is the status of e-banking in developing countries?
29. What is e-tailing?
30. What are the trends and prospects for e-tailing?
31. What is online publishing? What are its most common applications?
32. What are the benefits and advantages of online publishing to business?
33. What are the problems and issues in online publishing?
34. How is e-commerce useful to developing country entrepreneurs?
35. What is the role of government in the development of e-commerce in developing countries?
36. What is a favorable policy environment for e-commerce?
37. How can government use e-commerce? Are existing legal systems sufficient to protect those engaged in e-commerce?

4 Modern Business Management: E-SCM

Chapter Includes :

4.1 Introduction

Competition is becoming acute and customers are becoming choosy. The emphasis is on building "Value to the customer". Supply Chain concepts are being practiced with increasing frequency by companies, not only in FMCG and Automobile sectors but also in other sectors like steel, cement, etc.

NOTES

Supply chain management - delivering the right product to the right place, at the right time and at the right price - is one of the most powerful engines of business transformation. It is one of the leading cost saving and revenue enhancement strategies in use today.

An important aspect of long-term competitiveness is the ability to adapt to a changing environment. Once new conditions are identified companies need to adapt to stay competitive. When changing conditions demand a shift of technology in important product lines, companies are likely to adopt this new technology. The implementation of new technologies is an always-recurring process.

The information and communication technologies (ICT) have the potential to transform the competitive landscape in many industries, while at the same time, creating whole new industries. Taking into consideration, as an example, the Internet: it is a low cost standard with fast interactivity that exhibits network externalities, moderates time, has a universal reach, acts as a distribution channel, and reduces information asymmetries between transacting parties. Internet plays a critical and profound role in the way firm activities (internal or external) are coordinated, how commerce is conducted, how people and machines communicate, what defines communities and how they interact, and how and when goods are made and delivered.

4.1.1 Components of E-Supply-Chain Management Software

Manufacturing

Engineering and product configuration, production planning and cost management, production execution and quality management.

Logistics

Purchasing and order management, distribution, inventory and warehousing.

Financial

General ledger, payable and receivables, billing, budgets and asset management.

Marketing

Advertising, sales, order management, customer service and support and market research and strategy.

HRM

Payroll management, time and labour management benefits administration, and pension administration.

4.2 Supply Chain Management

Definition

Supply Chain Management (SCM) plays an important role in the management of processes that cut across functional and departmental boundaries. SCM goes beyond organizational boundaries, reaching out to suppliers and customers.

Organizations increasingly find that they must rely on effective supply chains, or networks, to successfully compete in the global market and networked economy. In Peter Drucker's management's new paradigms, this concept of business relationships extends beyond traditional enterprise boundaries and seeks to organize entire business processes throughout a value chain of multiple companies.

In the 21st century, there have been few changes in business environment that have contributed to the development of supply chain networks. First, as an outcome of globalization and proliferation of multi-national companies, joint ventures, strategic alliances and business partnerships were found to be significant success factors.

Supply chain management (SCM) is the process of planning, implementing, and controlling the operations of the supply chain with the purpose to satisfy customer requirements as efficiently as possible. Supply chain management spans all movement and storage of raw materials, work-in-process inventory, and finished goods from point-of-origin to point-of-consumption. The term supply chain management was coined by consultant **Keith Oliver**, of strategy consulting firm Booz Allen Hamilton in 1982.

Supply Chain Management encompasses the planning and management of all activities involved in sourcing, procurement, conversion, and logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers. In essence, Supply Chain Management integrates supply and demand management within and across companies.

4.2.1 SCM Issues

Supply chain management must address the following problems:

- **Distribution Network Configuration:** Number and location of suppliers, production facilities, distribution centers, warehouses and customers.
- **Distribution Strategy:** Centralized versus decentralized, direct shipment, Cross docking, pull or push strategies, third party logistics.
- **Information:** Integrate systems and processes through the supply chain to share valuable information, including demand signals, forecasts, inventory and transportation etc.
- **Inventory Management:** Quantity and location of inventory including raw materials, work-in-process and finished goods.

Supply chain execution is managing and coordinating the movement of materials information and funds across the supply chain. The flow is bi-directional.

4.2.2 SCM Functions

Supply chain management is a cross-functional approach to managing the movement of raw materials into an organization and the movement of finished goods out of the organization toward the end-consumer. As corporations strive to focus on core competencies and become more flexible, they have reduced their ownership of raw materials sources and distribution channels. These functions are increasingly being outsourced to other corporations that can perform the activities better or more cost effectively.

NOTES

The effect has been to increase the number of companies involved in satisfying consumer demand, while reducing management control of daily logistics operations. Less control and more supply chain partners led to the creation of supply chain management concepts. The purpose of supply chain management is to improve trust and collaboration among supply chain partners, thus improving inventory visibility and improving inventory velocity.

Supply chain activities can be grouped into strategic, tactical, and operational levels of activities.

Strategic

- Strategic network optimization, including the number, location, and size of warehouses, distribution centers and facilities.
- Strategic partnership with suppliers, distributors, and customers, creating communication channels for critical information and operational improvements such as cross docking, direct shipping, and third-party logistics.
- Product design coordination, so that new and existing products can be optimally integrated into the supply chain, load management
- Information Technology infrastructure, to support supply chain operations.
- Where to make and what to make or buy decisions
- Align overall organizational strategy with supply strategy

Tactical

- Sourcing contracts and other purchasing decisions.
- Production decisions, including contracting, locations, scheduling, and planning process definition.
- Inventory decisions, including quantity, location, and quality of inventory.
- Transportation strategy, including frequency, routes, and contracting.
- Benchmarking of all operations against competitors and implementation of best practices throughout the enterprise.
- Milestone payments

Operational

- Daily production and distribution planning, including all nodes in the supply chain.
- Production scheduling for each manufacturing facility in the supply chain (minute by minute).
- Demand planning and forecasting, coordinating the demand forecast of all customers and sharing the forecast with all suppliers.
- Sourcing planning, including current inventory and forecast demand, in collaboration with all suppliers.
- Inbound operations, including transportation from suppliers and receiving inventory.
- Production operations, including the consumption of materials and flow of finished goods.

- Outbound operations, including all fulfillment activities and transportation to customers.
- Order promising, accounting for all constraints in the supply chain, including all suppliers, manufacturing facilities, distribution centers, and other customers.

NOTES

4.3 E-Supply Chain Management

E-Supply Chain Management (E-SCM or e-SCM or eSCM) refers to the flow of physical goods and associated information from the source to the consumer with the help of new technologies like internet, intranet, communications technologies etc.. Key eSupply chain activities include purchasing, materials management, distribution, customer service, and inventory forecasting. Effectively managing these processes is critical to the success of any online operation.

The Internet supply chain will be a means of communicating and doing business with suppliers and customers. Fractured, unpredictable supply chains have become less and less tolerable primarily because customers will not absorb the associated costs and long lead times. Its important to bear in mind that the your customer is just a mouse click away from your competitors.

Web-Enabled supply chains assist companies to optimize business processes both within and outside the four walls of the enterprise and to more efficiently deliver the new products customers want, when they want them and where they want them.

Companies are looking at how to provide greater flexibility in moving parts globally. In a quest for greater efficiency and lower costs, there's been growing importance given to managing the supply chain effectively for the benefit of all parties.

The traditional vision of Supply Chain management (SCM) represents only one dimension of a business environment that is growing increasingly multidimensional. Modern day Supply Chain Management is the e-commerce of manufacturing.

Technologies such as the Internet, electronic data interchange, transportation and warehouse management software, including software that manages plant scheduling, demand forecasting, procurement, make SCM a versatile strategy to adopt.

In the new business models, competition is no longer simply viewed as company vs company, but as supply chain vs supply chain. Hence the ability to manage the different models of SCM becomes crucial.

4.3.1 E-SCM Chain

Supply chain has been viewed as an inflexible series of events that somehow managed to get products out the door. It often involved questionable inventory forecasts, rigid manufacturing plans and hypothetical shipping schedules.

The Internet has changed all that. It has transformed this old-fashioned process into something closer to an exact science. An Internet-enabled supply chain helps companies:

- Avoid costly disasters
- Reduce administrative overhead
- Reduce unnecessary inventory (thereby increasing working capital)
- Decrease the number of hands that touch goods on their way to the end customer

- Eliminate obsolete business processes
- Reap cost-cutting and revenue-producing benefits
- Speed up production and responsiveness to consumers
- Garner higher profit margins on finished goods

NOTES

Effective integration of an Organizations supply chain can save millions, improve customer service and reduce inventories.

The key to getting optimum value out of automating your supply chain is to make sure you have your internal systems working well before you start extending them out over the Internet.

4.3.2 Evaluating a E-SCM Initiative

Here are a few tips to bear in mind while evaluating an E-SCM initiative:

- Get Perspective -One should envision the business as a whole including its current strategy and where it wants to go. Supply chain strategy is increasingly being integrated with overall corporate strategy.
- Don't Underestimate Learning Costs - The cost of training people to use new software should not be underestimated. Sending information around the world takes lesser time than it takes to get into someone's mind.
- Link to existing architecture - Supply chain applications must link to existing enterprise resource planning applications. ERP serves as the nerve center of the organization. Ideally, it should be a single point of visibility for inventory and order taking.
- Think Global, Start Local.

4.4 Components of Modern E-SCM

SCM has three key processes :

- Planning
- Execution
- Performance measurement

Firm are shifting from a purely functional view to a more process-oriented view. To take a more process-oriented view, firms must shift from one-time transactions to shared on-line processes.

Planning systems : Planning systems focus on having the right product at the right place and at the right time. These systems facilitate order taking and information gathering from the customer and orchestrate the flow of information along the entire supply chain, from initial to raw material procurement to final consumption.

Execution systems : Execution systems facilitate the physical of goods and services through a supply chain. This focus traditionally includes some application based such as (customer) order fulfilment, inventory control and manufacturing and logistics. Execution systems focus on operational efficiency, which entails finding new way to streamline and automate day-to-day business operations to reduce costs and improve productivity. This first step towards improving operational efficiency is to upgrade key business applications to a single, integrated system that can run the entire business. This enables firm to efficiently move their products through the supply chain.

NOTES

Performance measurement systems : Performance measurement process keeps track of the health of the supply chain. This is necessary in order to make more informed decisions and respond to changing market conditions. Here, accounting and financial management systems are the real focal point. These applications utilize electronic commerce tools such as data warehousing to allow for effective information auditing and analysis.

Supply-chain management takes isolated business functions-marketing, materials management, purchasing, manufacturing and distribution-and allows them to function in tandem.

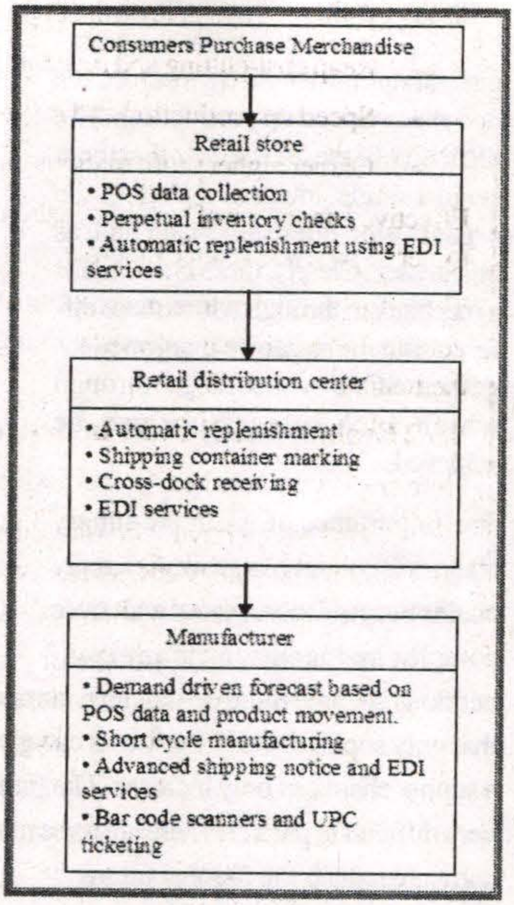


Figure 4.2 : Modern E-SCM

4.5 Major Trends in E-SCM

With the growth of Internet, more and more companies are dedicated to e-business. SCM has been deemed as one of the useful tools for business to promote their competitiveness, and to build up their trust relationship with up and down stream businesses. In order to link the whole value chain, Informations technology (IT) applications has already been essential for enterprises. So, it is possible to adopt a database-oriented approach which uses a server side program, (Eg., Java servlets), to manage a web-based inventory management system. Electronic Supply Chain Management (E-SCM) system to win competitive advantage. Usually, it may take couple of year to build an integrated ERP (Enterprise Resource Planning) system and E-SCM system. It provides a more efficient way to build the system.

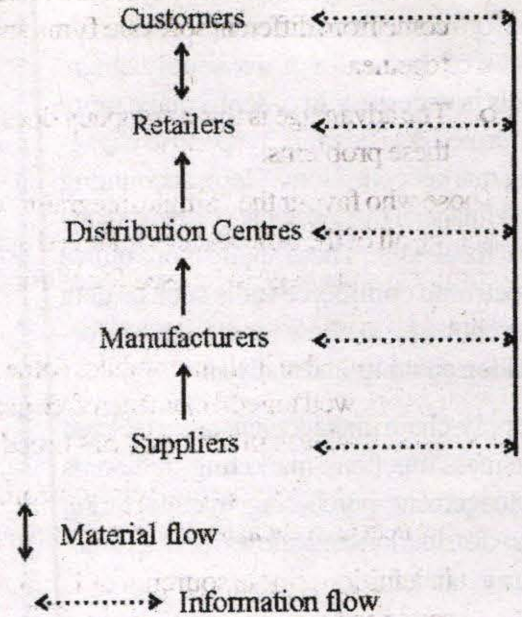
E-commerce has been booming for a while. More and more businesses are dedicated to www business. However, a well designed website is not enough. Competition is more intense in the E-commerce age. Supply Chain Management (SCM) is important issue in this new age. And the true power of E-Commerce rests in the strategy that relationship between a company and its suppliers has been changed since the process of supply and demand became electronic.

Supply Chain Management (SCM) has broad scope that includes suppliers, manufactures, retailers and customers. SCM covers the management of material, information and cash flows.

The Early Stage of SCM : For a single product, the raw material is procured from suppliers, transformed into finished goods from manufacturers in a single step, and then transported to retailers and customers. The whole supply chain involves marketing, manufacturing, distribution, planning and purchasing activities. Traditionally, these

NOTES

activities are generated independently by differnet organizations. These organizations have their own objectives and these are often conflicting. The difficulty of these factors is that there is not a single, integrated plan for the organization, there are many plan as businesses. Clearly, there is a need for a mechanism through which these different functions can be integrated together. SCM is a strategy through which such integration can be achieved.



The importance of IT in the supply chain will continue to grow. As supply chains become more global and more complex and as customers and competitions become more demanding, companies will need the supply chain capabilities that only sophisticated IT systems can give them. Therefore, the importance of IT to a supply chain can only increase. The future role of IT in the supply chain however is very difficult to predict. There are three main trends that will influence the supply chain software industry in the near future.

1. Best of Breed Versus Single Integrator

- A battle is occurring nor just among the developers within a category of software, but also among developers of different types of software.
  - The most vivid example is the conflict between the ERP and analytical applications developers.
  - Developers of analytical applications strive to build a package that is the best at solving problems for that particular function within a stage in the supply chain.
  - ERP providers build a system that allows a single, integrated view of many different functions in a supply chain stage.
  - Analytical applications must integrate with some ERP or legacy system in order To get the data they analyze.
  - The two types of systems thus naturally complement each other. However there is conflict between the groups as the ERP players develop their analytical applications and begin to complete directly with the analytical developers.
- Those who believe that the "best of breed" approach is the preferred solution pick the best application for each individual function or stage in the supply chain.
  - These applications are the best of the breed of that function.
  - Companies then wind up with several different types of systems that they must integrate to achieve global scope of the supply chain.

NOTES

- This integration can be quite difficult because individual parts of the system come from different software firms and they may not be designed to work well together.
- The advantage is that a company does get the best solutions if it can overcome these problems.

- Those who favour the "Single integrator" approach look at the difficulties of integrating all of the different best of breed applications and opt instead for one single company's package to be used for all functions and potentially even all stages of the supply chain.
  - The idea is that if all the modules come from a single firm, they will be designed to work well together and therefore getting them up and operational should be much easier than the best - of - breed approach.
  - Of course, this ease of operation comes at the expense of being able to pick the best analytical tool for a particular problem a firm is facing.
  - In addition, single sourcing of IT could cause the same problems that single sourcing from any supplier can create.
- It is unclear which idea will win this battle or even if a clear winner will emerge.
  - However, analytical applications have proven difficult to develop for those firms with little experience.
  - And with regard to the acquisition strategy it has now worked out well for the ERP players, either.
  - In the end it is likely that both ideas will coexist, each serving those industries that are best suited to their particular attributes.
  - For instance, in industries in which having the best inventory planning can be a competitive advantage, we might expect to see companies choose the best - of - breed solution so that they can take advantage of the absolute best inventory solutions.
  - In industries in which one particular function does not offer a large advantage, we would expect to see companies choose single integrators because the ease of illustration and operation would likely outweigh any gain from a more advanced analytical system.

2. Shifts in platform technology

A supply chain IT systems have moved from legacy to ERP and analytical systems, there has been a shift in technology from mainframes to client / server platforms. Currently, two technologies are taking over the client / server technology that was the dominant platform for supply chain IT applications.

- The first technology is the browser-based internet application.
  - Users of this application only need access to the internet and a browser on their computers.
  - These users utilize the browser to log into the IT system over the internet.
  - Any information or analysis they need is available over the browser.
  - Calculations and storage take place on a central server.

NOTES

- This technology is compelling, as it requires no costly software upgrades on the user's computer.
- Avoiding this expenditure and being able to easily use the latest application version make this option quite attractive..
- At this writing, just about major software player discussed here is trying to take advantage of this technology by modifying products to make them internet accessible.

- The second technology is the whole new business model of Application Service Providers (ASPs).
  - ASPs do not develop software themselves but rather host software developed by others and rent the use of the software to companies
  - The ASP is then responsible for running (generally remotely) the applications for rent.

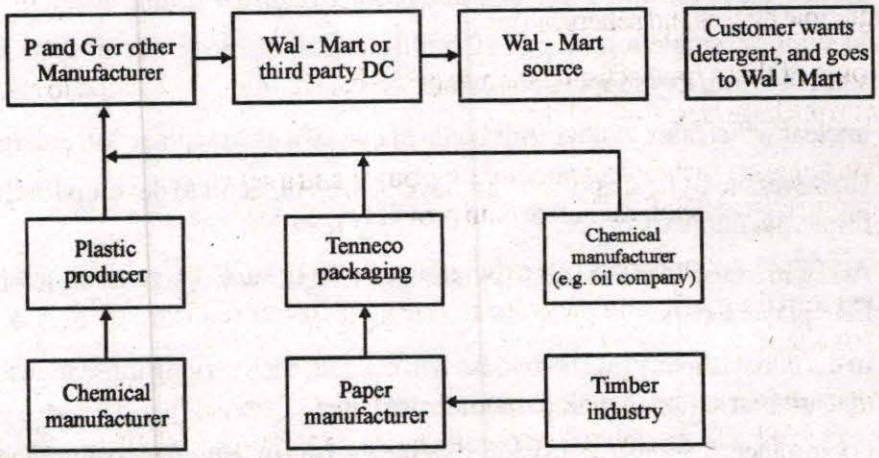


Figure 4.3 : Stages of detergent supply chain

  - The ASP business model turns the traditional software model on its head in that customers pay a monthly subscription fee to use the software rather than a large up-front fee to acquire a license to the software
  - ASPs began by focusing on smaller businesses that lacked the capital to invest to expensive system installations.
  - The benefits of ASPs, such as ease of upgrading and little up front investment, may eventually win over some large companies, although penetration in larger companies is likely to be much lower than with small companies.
  - ASPs exist not only as stand-alone companies but also within software developers (such as ORACLE) that have begun offering their own ASP service.
  - This trend toward a service based model is very likely to continue and will significantly alter the way the software industry is structured.
  - Leading ASPs include US internet working and curio.

3. Convergence of Supply Chain Management and Business-to-Business exchanges

Business-to-business (B2B) exchanges are electronic market places on the internet where suppliers and buyers interact to conduct transactions. Many B2B exchanges sprung in 1999 and 2000, each generally focusing on an industry vertical, such as electronic components, plastics, or auto parts. Most of these B2B sites have focused their energy on developing a web site and pulling in the potential buyers and sellers within their vertical industry.

These exchanges provide the opportunity for huge value creation, through the reduction in transaction costs, improved supply chain visibility and the more efficient allocation of supply and demand.

- These B2B exchanges have fostered many types of transactions.
  - At their most basic level, exchanges create an efficient market place to make spot purchases or sales of goods.
  - These can be done through a variety of transactional means such as auctions (one seller, multiple buyers), reverse auctions (one buyer, multiple sellers), fixed price transactions (one buyer, one seller), and bid/ask auctions (in which multiple buyers and sellers bid).
  - Some of these transactions, such as auctions, are often best suited to short-term relationships.
  - For long-term relationships between a buyer and a seller, in which tight integration is beneficial, exchanges can provide even more value.
  - Exchanges not only produce efficiencies in the transaction process and the creation of a market but they also enable tight collaboration between buyers and sellers.

    For example, exchanges can enable the sharing of demand forecasts between buyers and sellers so that sellers can more accurately manage their production.
  - Exchanges can also allow the efficient combination of purchasing goods with the logistics services needed to fulfill an order.
- As the exchanges have become operational, it has become increasingly clear that the key to exchanges creating value lays in this collaboration between buyers and sellers that an electronic market can enable.
  - What exchanges need to perform these functions are the supply chain IT applications discussed previously.
  - As those involved in B2B exchanges realize they need this capability, they are seeking the services of the supply chain software developers discussed here.
- At the same time, the supply chain software companies are realizing that they hold the key to successful B2B exchanges and in addition to partnering with existing exchanges, are also starting their own exchanges.
  - This is creating a convergence of the B2B exchanges and the supply chain companies, in which each offers a range of products that encompass both B2B and SCM.
  - This convergence creates a whole new set of opportunities for supply chain software developers.

- It also brings a whole new set of competitors into the supply chain software world, which will alter the competitive landscape.

4.6 Example of E-SCM

A supply chain consists of all stages involved, directly or indirectly, in fulfilling a customer request The supply chain not only includes the manufacturer and suppliers, but also Transporters, warehouses, retailers and customers themselves. Within each organization, such as a manufacturer, the supply chain includes all functions involved in filling a customer request These functions include, but are not limited to, new product development, marketing, operations, distribution, finance and customer service.

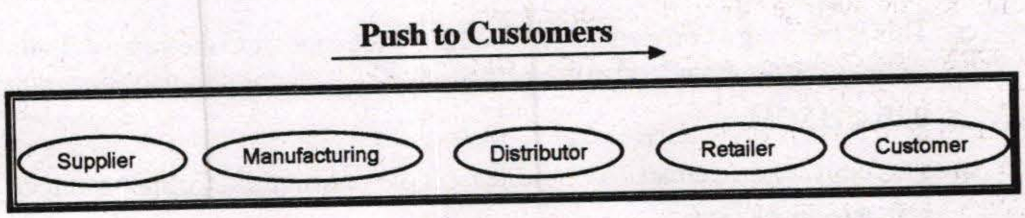
- Consider a customer walking into a Wal-Mart store to purchase detergent.
  - The supply chain begins with the customer and his need for detergent.
  - The next stage of this supply chain is the Wal-Mart retail store that the customer visits.
  - Wal-Mart stocks its shelves using inventory that may have been supplied from finished goods warehouses that Wal-Mart manages or from a distributor using trucks supplied by a third party.
  - The distributor in turn is stocked by the manufactures.
  - The Procter and Gamble (P&G) manufacturing plant receives raw material from a variety of suppliers, who may themselves have been supplied by lower-tier suppliers
  - For example, packaging material may come from Tenneco and Tenneco might receive raw material to manufacture the packaging from other suppliers This supply chain is illustrated in Fig. 4.3.

This example illustrates that the customer is an integral part of the supply chain The primary purpose for existence of any supply chain is to satisfy customer needs, in the process generating profits for itself. Supply chain activities begin with a customer order and end when a satisfied customer has paid for his purchase. The term supply chain conjures up images of product or supply, moving from suppliers to manufacturers to distributors to retailers to customers along a chain.

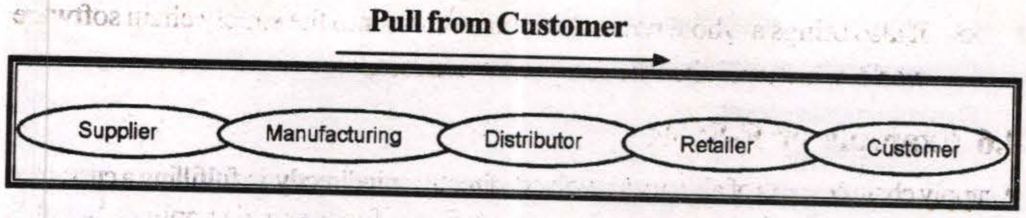
4.7 Architecture of E-Supply Chain Models

The consumer need-based business model is forcing a fundamental shift from traditional manufacturing push-based model (also called build-to-stock) to pull-based model (build-to-order).

Typical Aim : Optimize the production process for cost and efficiency.



NOTES



Typical Characteristics : Manufacturer-led new product development.

Poor data integration through limited use of technology,

Long cycle and response time, and high inventory levels.

Use of IS : Independent data management by supply chain members, limited use of EDI.

Typical Aim : Enhance product and service quality.

Typical Characteristics : Market research driven, technology used to achieve research and data integration, short cycle and response times, low inventory levels.

Use of IS : Integrated internal system, information sharing between supply chain members. Extensive use of EDI and e-commerce, often through B2B exchanges and intermediates.

4.8 E-SCM Process Integration

Companies are using Internet technologies to manage their customers and their supply chain transactions, and find that these applications can offer a great return on investment. Companies realized that to build their own business globally, they had to find better ways to serve their customers, and had to work more efficiently with their suppliers. This is often the first step of a virtual integrated supply chain.

In order to reach this goal a new technology has to be implemented. In other words, a technology has to be absorbed by the organization that is involved in the implementation process. During internet based technologies implementation process users of information and communication technologies (ICT) systems are obliged to overcome the failures and the problems that inhibit the complete "absorption" of the technology, making it difficult to let it become part of the usual managerial routine. Therefore, there are two key entities that characterize the implementation process: the organizations involved and the technology implemented.

Successful E-SCM requires a change from managing individual functions to integrating activities into key supply chain processes. An example scenario: the purchasing department places orders as requirements become appropriate. Marketing, responding to customer demand, communicates with several distributors and retailers, and attempts to satisfy this demand. Shared information between supply chain partners can only be fully leveraged through process integration.

Supply chain business process integration involves collaborative work between buyers and suppliers, joint product development, common systems and shared information. According to Lambert and Cooper - operating an integrated supply chain requires continuous information flows, which in turn assist to achieve the best product flows. However, in many companies, management has reached the conclusion that optimizing the product flows cannot be accomplished without implementing a process approach to the business. The key supply chain processes stated by Lambert are:

NOTES

- Customer relationship management
- Customer service management
- Demand management
- Order fulfillment
- Manufacturing flow management
- Supplier relationship management
- Product development and commercialization
- Returns management

One could suggest other key critical supply business processes combining these processes stated by Lambert such as:

a) Customer service management

b) Procurement

c) Product development and commercialization

d) Manufacturing flow management/support

e) Physical distribution

f) Outsourcing/partnerships

g) Performance measurement

(a) Customer service management process

Customer Relationship Management concerns the relationship between the organization and its customers. Customer service provides the source of customer information. It also provides the customer with real-time information on promising dates and product availability through interfaces with the company's production and distribution operations. Successful organizations use following steps to build customer relationships:

- Determine mutually satisfying goals between organization and customers
- Establish and maintain customer rapport
- Produce positive feelings in the organization and the customers

(b) Procurement process

Strategic plans are developed with suppliers to support the manufacturing flow management process and development of new products. In firms where operations extend globally, sourcing should be managed on a global basis. The desired outcome is a win-win relationship, where both parties benefit, and reduction times in the design cycle and product development is achieved. Also, the purchasing function develops rapid communication systems, such as **electronic data interchange (EDI)** and Internet linkages to transfer possible requirements more rapidly. Activities related to obtaining products and materials from outside suppliers. This requires performing resource planning, supply sourcing, negotiation, order placement, inbound transportation, storage and handling and quality assurance. Also, includes the responsibility to coordinate with suppliers in scheduling, supply continuity, hedging, and research to new sources or programmes.

(c) Product development and commercialization

Here, customers and suppliers must be united into the product development process, thus to reduce time to market. As product life cycles shorten, the appropriate products must be developed and successfully launched in ever shorter time-schedules to remain competitive. According to Lambert and Cooper , managers of the product development and commercialization process must:

- ❖ Coordinate with customer relationship management to identify customer-articulated needs;
- ❖ Select materials and suppliers in conjunction with procurement, and
- ❖ Develop production technology in manufacturing flow to manufacture and integrate into the best supply chain flow for the product/market combination.

(d) Manufacturing flow management process

The manufacturing process is produced and supplies products to the distribution channels based on past forecasts. Manufacturing processes must be flexible to respond to market changes, and must accommodate mass customization. Orders are processes operating on a **just-in-time (JIT)** basis in minimum lot sizes. Also, changes in the manufacturing flow process lead to shorter cycle times, meaning improved responsiveness and efficiency of demand to customers. Activities related to planning, scheduling and supporting manufacturing operations, such as work-in-process storage, handling, transportation, and time phasing of components, inventory at manufacturing sites and maximum flexibility in the coordination of geographic and final assemblies postponement of physical distribution operations.

(e) Physical distribution

This concerns movement of a finished product/service to customers. In physical distribution, the customer is the final destination of a marketing channel, and the availability of the product/service is a vital part of each channel participant's marketing effort. It is also through the physical distribution process that the time and space of customer service become an integral part of marketing, thus it links a marketing channel with its customers (e.g. links manufacturers, wholesalers, retailers).

(f) Outsourcing/Partnerships

This is not just outsourcing the procurement of materials and components, but also outsourcing of services that traditionally have been provided in-house. The logic of this trend is that the company will increasingly focus on those activities in the value chain where it has a distinctive advantage and everything else it will outsource. This movement has been particularly evident in logistics where the provision of transport, warehousing and inventory control is increasingly subcontracted to specialists or logistics partners. Also, to manage and control this network of partners and suppliers requires a blend of both central and local involvement. Hence, strategic decisions need to be taken centrally with the monitoring and control of supplier performance and day-to-day liaison with logistics partners being best managed at a local level.

(g) Performance measurement

Experts found a strong relationship from the largest arcs of supplier and customer integration to market share and profitability. By taking advantage of supplier capabilities and emphasizing a long-term supply chain perspective in customer relationships can be both correlated with firm performance. As logistics competency becomes a more critical factor in creating and maintaining competitive advantage, logistics measurement becomes increasingly important because the difference between profitable

NOTES

and unprofitable operations becomes more narrow. The firms engaging in comprehensive performance measurement realized improvements in overall productivity. According to experts internal measures are generally collected and analyzed by the firm including

- Cost
- Customer Service
- Productivity measures
- Asset measurement, and
- Quality.

External performance measurement is examined through customer perception measures and "best practice" benchmarking, and includes:

1) Customer perception measurement, and

2) Best practice benchmarking.

4.9 Supply chain management components integration

The management components of E-SCM

The E-SCM management components are the third element of the four-square circulation framework. The level of integration and management of a business process link is a function of the number and level, ranging from low to high, of components added to the link. Consequently, adding more management components or increasing the level of each component can increase the level of integration of the business process link. The literature on business process reengineering, buyer-supplier relationships, and SCM suggests various possible components that must receive managerial attention when managing supply relationships. Lambert and Cooper identified the following components which are:

- Planning and control
- Work structure
- Organization structure
- Product flow facility structure
- Information flow facility structure
- Management methods
- Power and leadership structure
- Risk and reward structure
- Culture and attitude

However, a more careful examination of the existing literature will lead us to a more comprehensive structure of what should be the key critical supply chain components, the "branches" of the previous identified supply chain business processes, that is what kind of relationship the components may have that are related with suppliers and customers accordingly. A primary level channel participant is a business that is willing to participate in the inventory ownership responsibility or assume other aspects financial risk, thus including primary level .A secondary level participant (specialized), is a business that participates in channel relationships by performing essential services for primary participants, thus including secondary level components, which are supporting the primary ones. Also, third level channel participants and components may be included, that will support the primary level channel participants, and which are the fundamental branches of the secondary level components.

Consequently, Lambert and Cooper's framework of supply chain components, does not lead us to the conclusion about what are the primary or secondary (specialized) level supply chain components, that is what supply chain components should be viewed as primary or secondary, and how should these components be structured in order to have a more comprehensive supply chain structure and to examine the supply chain as an integrative.

Baziotopoulos suggests the following supply chain components :

- For **customer service management**: Includes the primary level component of customer relationship management, and secondary level components such as benchmarking and order fulfillment.
- For **product development and commercialization**: Includes the primary level component of Product Data Management (PDM), and secondary level components such as market share, customer satisfaction, profit margins, and returns to stakeholders.
- For **physical distribution, Manufacturing support and Procurement**: Includes the primary level component of enterprise resource planning (ERP), with secondary level components such as warehouse management, material management, manufacturing planning, personnel management, and postponement (order management).
- For **performance measurement**: This includes the primary level component of logistics performance measurement, which is correlated with the information flow facility structure within the organization. Secondary level components may include four types of measurement such as: variation, direction, decision and policy measurements. More specifically, in accordance with these secondary level components total cost analysis (TCA), customer profitability analysis (CPA), and Asset management could be concerned as well. In general, information flow facility structure is regarded by two important requirements, which are : (a) Planning and Coordination flows, and (b)Operational requirements.
- For **outsourcing**: This includes the primary level component of management methods and the company's cutting-edge strategy and its vital strategic objectives that the company will identify and adopt for particular strategic initiatives in key the areas of technology information, operations, manufacturing capabilities, and logistics (secondary level components).

4.10 Globalization and E-SCM

Business today is in a global environment. This environment forces companies, regardless of location or primary market base, to consider the rest of the world in their competitive strategy analysis. Firms cannot isolate themselves from or ignore external factors such as economic trends, competitive situations or technology innovation in other countries, if some of their competitors are competing or are located in those countries. Companies are going truly global with Supply-chain Management (SCM).

NOTES

A company can develop a product in the United States, manufacture in India and sell in Europe. Companies have changed the ways in which they manage their operations and logistics activities. Changes in trade, the spread and modernization of transport infrastructures and the intensification of competition have elevated the importance of flow management to new levels.

4.10.1 Driving Forces

The last two decades have seen the evolution of the global manufacturing environment. Majority of the manufacturers have global presence through exports, strategic alliances, joint ventures or as a part of a committed strategy to sell and produce in foreign markets.

The factors shaping the global environment and driving the development of global operations strategies of multinational firms fall into four categories: global market forces, technological forces, global cost forces and political and macroeconomic forces.

(a) Global Market Forces

There is tremendous growth potential in the foreign developing markets which heve resulted in intensified foreign competition in local markets which forces the small - and medium-sized companies to upgrade their operations and even consider expanding internationally. There heve also been growth in foreign demand which necessitates the development of a global network of manufacturing bases and markets. When the markets are global, the production-planning task of the manager becomes difficult on one hand and allows more efficient utilization of resources on the other. Few industries remain today in which the international product life-cycle theory still applies. Product markets, particularly in **technologically intensive industries**, are changing rapidly. Product -cycles are shrinking as customers demand new products faster. In addition, the advances in communication and transportation technology give customers around the world immediate access to the latest available products and technologies. Thus, manufacturers hoping to capture global demand must introduce their new products simultaneously to all major markets. Furthermore, the integration of product design and the development of related manufacturing processes have become the key success factors in many high-technology industries, where fast product introduction and extensive customization determine market success. As a result, companies must maintain production facilities, pilot production plants, engineering resources and even Research and Development (R & D) facilities all over the world. Apple Computer, for example, has built a global manufacturing and engineering infrastructure with facilities in California, Ireland and Singapore. This network allows Apple to introduce new products simultaneously in the American, European and Asian markets. Companies use the state-of-the-art markets as learning grounds for product development and effective production management, and then transfer this knowledge to their other production facilities worldwide. This rationale explains why Mercedes-Benz decided recently to locate a huge manufacturing plant in Vance, Alabama. The company recognizes that the United States is the state-of-the-art market for sport utility vehicles. It plans to produce those vehicles at the Vance plant and introduce them worldwide.

(b) Technological Forces

A peculiar trend which was prevalent in the last decade, besides globalization, was a limited number of producers which emerged due to diversity among products and uniformity across national markets. Product diversity has increased as products have grown more complex and differentiated and product life cycles have shortened. The share of the US market for high-technology goods supplied by imports from foreign-based companies rose from a negligible 5 per cent to more than 20 percent with the last decade. Moreover, the sources of such imports expanded beyond Europe to include Japan and the newly industrialized countries of Hong Kong, Singapore, South Korea and Taiwan. There has been diffusion of technological knowledge and global low-cost manufacturing locations have emerged. In response to this diffusion of technological capability, multinational firms need to improve their ability to tap multiple sources of technology located in various countries. They also must be able to absorb quickly, and commercialize effectively, new technologies that, in many cases, were invented outside the firm thus overcoming the destructive and pervasive 'not-invented-here' attitude and resulting inertia. There has been technology sharing and inter-firm collaborations. The well-known joint ventures in the auto industry between US and Japanese firms (GM-Toyota, Chrysler-Mitsubishi, Ford-Mazda) followed a similar pattern. US firms needed to obtain first-hand knowledge of Japanese production methods and accelerated product development cycles, while the Japanese producers were seeking ways to overcome US trade barriers and gain access to the vast American auto market. As competitive priorities in global products markets shift more towards product customization and fast new product development, firms are realizing the importance of co-location of manufacturing and product design facilities abroad. In certain product categories, such as Application Specific Integrated Circuits (ASICs), this was the main motivation for establishing design centres in foreign countries. Other industries such as pharmaceuticals and consumer electronics also have taken this approach.

(c) Global Cost Forces

New competitive priorities in manufacturing industries, that is product and process conformance quality, delivery reliability and speed, customization and responsiveness to customers, have forced companies to reprioritize the cost factors that drive their global operations strategies. The Total Quality Management (TQM) revolution brought with it a focus on total quality costs, rather than just direct labour costs. Companies realized that early activities such as product design and worker training substantially impact production costs. They began to emphasize prevention rather than inspection. In addition, they quantified the costs of poor design, low input quality and poor workmanship by calculating internal and external failure costs. All these realizations placed access to skilled workers and quality suppliers high on the priority list for firms competing on quality. Similarly, Just-in-time (JIT) manufacturing methods, which companies widely adopted for the management of mass production systems, emphasized the importance of frequent deliveries by nearby suppliers. A number of high-technology industries have experienced dramatic growth in the capital intensity of production facilities. A state-of-the-art semiconductor factory, for instance, costs close to half a billion dollars. When R & D costs are included, the cost of production facilities for a new generation of electronic products can easily exceed $ 1 billion. Similarly, huge numbers apply for the development and production of new drugs in the pharmaceutical industry. Such high costs drive firms to adopt an economies-of-scale strategy that concentrates production in a single location, typically in a country that has the required labour and supplier infrastructures. They then achieve high-capacity utilization of the capital-intensive facility by aggressively pursuing the global market. Besides this the host government subsidies also become an important consideration.

NOTES

(d) Political and Macroeconomic Forces

Getting hit with unexpected or unreasonable currency devaluations in the foreign countries in which they operate is a nightmare for global operations managers. Managing exposure to changes in nominal and real exchange rates is a task which the global operations manager must master. If the economics are favourable, the firm may even go so far as to establish a supplier in a foreign country where one does not yet exist. For example, if the local currency is chronically undervalued, it is to the firm's advantage to shift most of its sourcing to local vendors. In any case, the firm may still want to source a limited amount of its inputs from less favourable suppliers in other countries if it feels that maintaining an ongoing relationship may help in the future when strategies need to be reversed. Philips, Thomson, Electrolux and Ford are in the process of creating pan-European networks of factories (producing both components and finished goods). The trade protection mechanisms which exist in the form of tariff and non-tariff barriers effect the global operation strategy; but these are readily losing importance in the new borderless trade regime.

4.11 E-Supply Chain Network

Supply chain consultants(SCC)stand between the big management consultants and the organizational consultants, and between institutes affiliated to universities and manufacturers and general managers.

- Their niche lies between organisational strategy and supply chain strategy.
- They are not research oriented, yet they are innovative.
- They are able to realise ideas but are independent of both manufacturers and product.

In short, supply chain consultants are placed beside and behind their customer in the function of temporary supply chain managers. They are trustees of the company's interests, yet they are at the same time more independent than company members are. Their relationship to their client is one of absolute loyalty. One can find supply chain consultants in a wide variety of functions.

Some of them are illustrated below:

(A) Consultants arc competitors, sparing partners and spare capacity for company management:

Competition stimulates business and the quality of a solution. Large organisations are allowing us more and more opportunities to compete against their own company management. This type of planned competition with their own people is often an important instrument in testing and improving the capabilities of their staff. In case of large investments of strategic decisions, planned competition with the outsider, i.e. the consultant, brings new ideas to the situation, gained through his broad experience of other industries. Finally, consultants act as a pool to assist at times of peak demand. They can be used as a free reserve that does not cause any\ fixed costs.

(B) The consultant supports the positioning of supply chain in the organization:

Is supply chain one of the central areas of competence of a company? Would one contract out supply chain management? What does supply chain contributes to the company's profit? Many companies relieve themselves of burden of supply chain management by simply contracting out. With regard to central areas of competence, this is often very sensible. Yet this decision should only be taken if it is workable in the long-term. Supply chain consultants can help, acting here as independent external business partners to answer the question of central competence of supply chain management. For example, they can calculate exactly the contribution of supply chain towards the profit and the cost of supply chain itself, i.e. calculate carefully, make or buy alternatives.

(C) Consultants integrate supply chain as a business management system:

Material flow in the warehouse cannot be regarded as an isolated process. Both the systems theorist and also the supply chain practitioner know how much one process depends on the other. For example, the warehouse dimensions depend on the structure of procurement and on stock; the distribution routes depend upon stock location; delivery service and commissioning depend on Last-Order times and agreed delivery dates. Often, the wishes of the sales strategist and the customer coincide, yet these are not always cost effective.

The level of service that is regarded by supply chain personnel as possible is often oriented towards daily business and is dominated by considerations of cost. Supply chain consultants can help sales and logistics personnel to reach a compromise.

(D) Supply chain consultants help to formulate strategic goals in logistics and to draw up investment and rationalisation programmes:

Those organisations seeking orientation, which want to invest in the future of their supply chain, must clarify first of all their strategic goals. Ordering methods, distribution point, and delivery service requirements must be defined in accord with the aims of the company.

Supply chain consultants are in this instance helpful business partners because they know both the strategic and also the technical side of a distribution system, that is the technical aspect involving knowledge of investment. They can calculate exactly at what cost and with how many employees an imaginary warehouse "X" can be operated at location "Y". Hypothetical examples that are unrealistic are virtually impossible with an experienced consultant. As soon as an investment programme is strategically fixed, the next step is to work out the technical details. Whoever desires now well-established standards will work in the main directly together with a manufacturer or general manager. On the other hand, whoever is seeking competitive advantage within a strongly contested market, for example by means of exceptional commitment to schedule and to low rate of failure, whoever finally, is aiming at exceptional growth, then let tailor-made supply chains be recommended to him over and above standards.

(E) Supply chain consultants are independent and unbiased towards product and manufacturer :

Manufacturers, too frequently offer planning and development resources. It is advisable to use them when optimising sub-systems for which the manufacturer is a specialist. An investor, who is certain that, for example, only a new shelf system is required, or the connection of two warehouse areas by conveyors, will be in safe hands with the producer of material flow components. If he wants to know, however, whether he needs a warehouse or rather a service provider or if a new system, conveyor belt or shelf system is at all necessary, then firstly, he should think and plan independently of both the product and manufacturer. One makes a better choice of system.

NOTES

4.12 E-Supply Chain Management Framework

In the last decade the strategic role of operations has increased, which led to fast improvements in management and methodology. In parallel, two basic principles have been strengthened. One of the principles is the idea to concentrate on processes instead of organizational function based task orientation. The other principle is the extension of coordination and integration of value creating processes. Changes did not stop at the border of companies, the nature of inter-company relations has also changed according to the principles above. On the basis of these dimensions of supply chain management the companies go through three development stages:

1) Transaction dominated companies,

2) Internally integrated companies,

3) Externally integrated companies.

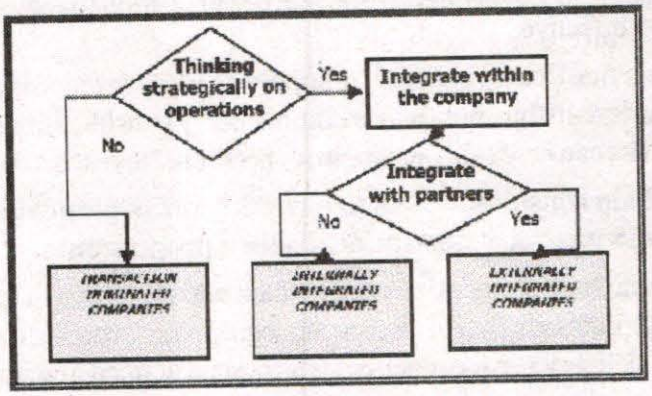


Figure 4.1 : E-SCM View of Development stages

The three stage model of E-SCM is described as:

- *Intrafunctional* coordination (administration of the activities and processes *within* the function of the firm);
- Coordination of *interfunctional* activities, such as between finance, logistics and production, logistics and marketing as they take place *among* the functional areas of the firm.
- Coordination of *inter-organizational* supply chain activities that take place between legally separate firms within the product-flow channel, such as *between* firms and their suppliers

Model Description

Four dimensions that are common in today's E-SCM. These are as follows:

- o *Increasing strategic role* of operations and operations management.
- o *Increasing level of coordination and integration* of material processes.
- o *Process view* getting more and more accepted and applied.
- o Building stronger *relationships* between collaborating partners.

The three operating models outlined :

- *Transaction dominated companies* representing a starting point from which development has been started.

- *Internally integrated companies* having recognized the strategic importance of the field and having a company-focused orientation while managing it.
- *Externally integrated companies* realizing the strategic importance of materials management and operations and extending their management efforts beyond company's borders.

For **transaction dominated companies** the strategic importance of materials management and operations is low and as a consequence they do not stress those development aspects that are suitable for increasing competitiveness of this field. These companies do not invest in coordinating an integrating material processes, manage activities instead of company-wide processes and do not strive for building closer connections with collaborating companies.

Internally integrated companies can be characterized with the recognition of strategic importance of operations and with an effort to manage them in order to gain competitive advantage. This means that companies start to apply different coordination and integration techniques, among them heavily invest in developing planning procedures and information technology. They can be characterized with a process focus. Most of their improvement efforts focus on their own operation. Although they start to build long-term relations with some partners, this is not accompanied by intensive use of relation specific investments. The basic difference between the internally and the externally integrated operating models lies in the dimension of managing partner relationships. Both model stress the strategic importance of operations, improve their coordination and integration with a process focus.

The **externally oriented companies** spread this improvement efforts and management attitude beyond the company and improve operation and collaboration with partners significantly. This can be achieved only by investing heavily in important relationships and build strategic alliances. This operating model contains the basic elements of the SCM concept.

SUMMARY

- Supply chain management - delivering the right product to the right place, at the right time and at the right price - is one of the most powerful engines of business transformation. It is one of the leading cost saving and revenue enhancement strategies in use today.
- *Supply chain management (SCM)* is the process of planning, implementing, and controlling the operations of the supply chain with the purpose to satisfy customer requirements as efficiently as possible. Supply chain management spans all movement and storage of raw materials, work-in-process inventory, and finished goods from point-of-origin to point-of-consumption. The term supply chain management was coined by consultant *Keith Oliver*, of strategy consulting firm Booz Allen Hamilton in 1982.
- Supply Chain Management integrates supply and demand management within and across companies.
- Supply chain management is a cross-functional approach to managing the movement of raw materials into an organization and the movement of finished goods out of the organization toward the end-consumer.

NOTES

- *E-Supply Chain Management (E-SCM or e-SCM or eSCM)* refers to the flow of physical goods and associated information from the source to the consumer with the help of new technologies like internet, intranet , communications technologies etc. . Key eSupply chain activities include purchasing, materials management, distribution, customer service, and inventory forecasting. Effectively managing these processes is critical to the success of any online operation.
- Web-Enabled supply chains assist companies to optimize business processes both within and outside the four walls of the enterprise and to more efficiently deliver the new products customers want, when they want them and where they want them.
- Companies are using Internet technologies to manage their customers and their supply chain transactions, and find that these applications can offer a great return on investment.
- Companies are going truly global with E-Supply-chain Management .

EXERCISE

1. What is supply chain management, explain with an Example? How it is growing with information and communication technologies(ICT)?
2. How traditional CRM is different from modern CRM ? What is E-CRM?
3. What are the problems which the Supply chain management must address?
4. Explain different functions of SCM.
5. Compare and Contrast the tactical and operational functions of SCM.
6. How can you define and E-Supply chain?
7. Describe Pull and Push E-Supply Chain Models in brief.
8. How E-SCM process are integrated ?Explain with suitable example.
9. Describe the management components of E-SCM.
10. What are the impacts of E-CRM in Globalization era?
11. What are driving forces behind modern E-SCM?
12. Describe in brief the E-SCM framework.
13. What are different element of E-SCM?
14. Describe basic components of E-SCM softwares.
15. Prove the statement with suitable example "Supply Chain Management integrates supply and demand management within and across companies".

5 Modern Business Management: E-CRM

Chapter Includes :

5.1 Customer Relationship Management Concepts

Customer Relationship Management (CRM)

can be defined as the strategies, processes, people and technologies used by companies to successfully attract and retain customers for maximum corporate growth and profit. CRM initiatives are designed with the goal of meeting customer expectations and needs in order to achieve maximum customer lifetime value and return to the enterprise. The use of customer relationship management products, CRM software and CRM solutions enhance the effective implementation of CRM in an organization.

NOTES

Customer Relationship Management (CRM) is a way to identify, acquire, and retain customers - a business' greatest asset. By providing the means to manage and coordinate customer interactions, CRM helps companies maximize the value of every customer interaction and in turn improve corporate performance.

Understanding the concept of CRM helps in decisions relating Customer relationship management product, CRM software and CRM solutions.

In simpler terms, CRM is the technique of establishing and maintaining long-term business relationships with your customers. CRM involves utilizing the data collected during your customer interactions to determine the demographics and future needs of each customer. Customers and relationships with them have always existed right from the birth of the 'buying and selling concept'.

Competition, driven by globalization and the internet has turned things around. Customers have a variety of choices and most importantly, they have become more knowledgeable and demanding. It is no myth that they are King. With this scenario, companies have realized that it's not just enough to satisfy and delight them but also build genuine relationships in a way that would benefit them.

CRM can help gain a greater share of a loyal customer's business.

Customers can be divided into three zones:

1. *Zone of defection* where customers are extremely hostile and have the lowest level of satisfaction .
2. *Zone of indifference* where customers are not sure. They have a medium level of customer satisfaction and customer loyalty towards the company.
3. The third levels of customers are in the *zone of affection* described as "Apostles". CRM focuses on bringing customers from level 1 to level 3 and retaining apostle customers.

5.1.1 Electronic Customer Relationship Management

Electronic Customer Relationship Management (E-CRM or e-CRM or eCRM) has become the latest paradigm in the world of Customer Relationship Management. E-CRM is becoming more and more necessary as businesses take to the web. No longer can web-enabled companies rely on the traditional brick & mortar strategies that have gotten them to where they are today. Such organizations have to evolve with the market instead of behind it.

Electronic commerce relies on customer interactions via a computer and telecommunications infrastructure for the purpose of advertising, promoting, and selling products and services online. Electronic commerce replicates most of the physical activities that take place in the market place to the point where increasing electronic commerce usage are shifting companies from those traditional market places to new market spaces. The traditional market places emphasize "customer satisfaction" as a way to earn consumer loyalty and attract new customers.

E-CRM is a business strategy that utilizes the power of technology to tie together all aspects of a company's business with the goal of building long-term customer loyalty. E-CRM, in practical terms, is the management of customer interactions at all levels, channels, and media.

According to Hansen : E-CRM is defined as "a process of acquiring, retaining and growing profitable customers. It requires a clear focus on the service attributes that represent value to the customer ant that create loyalty."

E-CRM, or Electronic Customer Relationship Management, is an integrated online sales, marketing and service strategy that are used to identify, attract and retain an organization's customers. It describes improved and increased communication between an organization and its clients by creating and enhancing customer interaction through innovative technology.

Electronic-CRM (E-CRM)is not just customer service, self-service web applications, sales force automation tools or the analysis of consumer buying behavior on the internet. E-CRM is all of these initiatives working together to enable an organization to more effectively respond to its customers' needs and to market to them on a one-to-one basis.

E-CRM gives the internet users the ability to interact with the business through their preferred communication channel and allows the business to offset expensive customer service agents with technology. So the value is the largest one of improved customer satisfaction and reduced cost through improved efficiency.

Some of the major players in the Customer Relationship Management (CRM) industry are ICT Group, Oracle, PeopleSoft, Amdoc, Converges, CSG Systems, SAP, Pivotal, Siebel Systems etc.

5.1.2 E-CRM Software

E-CRM software provides profiles and histories of each interaction the organization has with its customers, making it an important tool for all small and medium businesses.

E-CRM software systems may contain a selection of the following features:

- **Customer management** – Provides access to all customer information including enquiry status and correspondence.
- **Knowledge management** – A centralized knowledge base that handles and shares customer information.
- **Account management** – Access to customer information and history, allowing sales teams and customer service teams to function efficiently.
- **Case management** – Captures enquiries, escalates priority cases and notifies management of unresolved issues.
- **Back-end integration** – Blends with other systems such as billing, inventory and logistics through relevant customer contact points such as websites and call centres .

NOTES

- **Reporting and analysis** – Report generation on customer behaviour and business criteria.

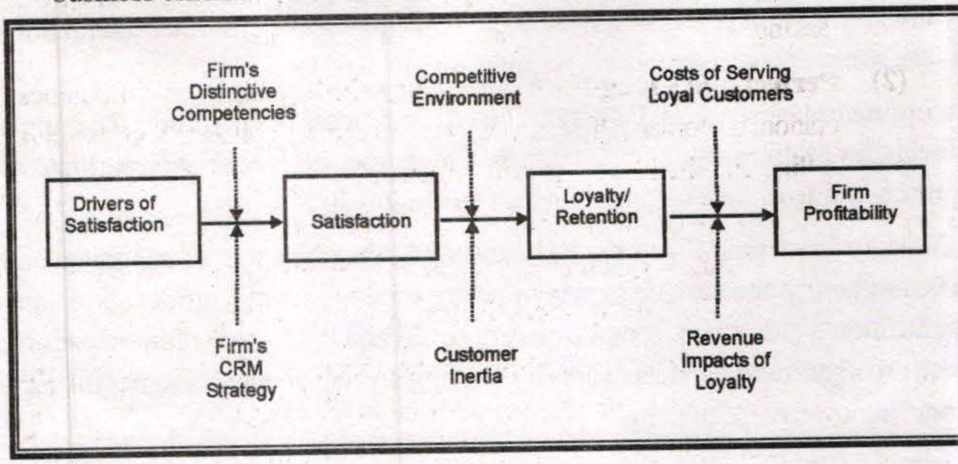


Figure 5.1 : E-CRM Goals

5.1.3 E-CRM Goals

The goals of E-CRM can be categorized in two parts:

Transactional Goals

- Facilitation of consumer search
- Decision-making support for standardized products
- Convenience in ordering
- Delivery tracking
- Limited personalization
- Asynchronous electronic communication (E-mail)
- Technical support for routine issues

Relational Goals

- Understanding customer and segment needs unobtrusively (data mining)
- Interactivity (synchronous communication, etc.)
- Third party E-CRM
- Enterprise Resource Planning

The link between loyalty and profitability is explained by the lower cost of serving loyal customers as well as the increases in revenue through increased patronage and positive word-of-mouth of such loyal customers. At the same time, it is also assumed that loyal customers are not price sensitive, thereby allowing firms to charge higher prices for added benefits and services provided to them.

5.1.4 E-CRM Business Models

Technology of DOT COM interfaces, such as order fulfillment, organizational business processes, and incentives, must be exploited by enterprises to remain competitive. There are six *value imperatives which* are vital for companies to execute

NOTES

(4) **Trust:** *strong branding*; sensitive use of customer profiles; security precautions.

(5) **Virtual communities:** information exchange about products and interests; market segment profiling.

(6) **Processing:** cuts across blocks from 1 to 5; internal and external; interface and ease of use.

(7) **Controlling:** cuts across all other blocks.

5.1.5 Technologies for e-CRM

Passive

- Cookies
- Chat rooms
- Bulletin boards and fan clubs
- Mailing lists
- News groups
- Observation studies through virtual reality and simulated environments
- Product-related discussion groups and lists

Active

- Chat rooms (hosted by Seller)
- Bulletin boards (hosted by Seller)
- Forums (hosted by the Seller)
- Internet surveys
- Product-related discussion groups and lists
- Recommender software

Interactive

- E-mail
- Forums
- Online focus groups
- Interactive online interviews
- Survey panels
- Auctions
- Online trade shows
- Shopping agent

Following figure presents our input-process-output model that serves as a theoretical framework for the study of some of the variables that influence e-CRM processes and outcomes.

(1) **Perfect one's logistics:** supply chain management; operating resource management; win–win trading partner collaborations and electronic out-tasking.

(2) **Perfect one's long-term customer relationships:** repeat business generation; customer self-management; community collaboration; massive cross-selling and lifetime relationship-focus.

(3) **Harmonize one's channels:** "seamless" links between the Internet, call centers, and physical channels; and strengthening distribution channels, while simultaneously strengthening your own brand.

(4) **Build a powerful portal/hub brand:** incentives for customers to routinely "park" on sites; aggressive customization and personalization; and revenue generation through hosting and selective use of give-it-away-free.com.

(5) **Transform capital and cost structures:** move toward negative working capital; slashing general selling and administrative costs; leverage cash flow generation on minimal physical balance sheet "assets"; reduce cost of capital by building a price/vision premium in market evaluation; and use the valuation advantage to buy needed capabilities at low capital dilution.

(6) **Build value-adding intermediation:** provide a hub with reliable information and advice to link buyers with sellers; offer more efficient transaction processing between trading partners; and build win–win relationships along an entire business chain.

These above imperatives involve developing a business model that stresses E-CRM.

Management of Customer Relationship in Business Media (MCR-BM)-concept:

This concept is defined as "*The Management of Customer Relationship in Business Media comprises the design, development and application of holistic concepts in order to manage relationships to economically valuable current or future customers*".

The MCR-BM concept therefore offers a basic framework for the design of business models to meet the challenges posed by the digital economy. The MCR-BM model consists of seven interrelated building blocks listed below with some key associated issues:

(1) **Customer interaction:** informational content and channels; value-added through non-standard information; pull and push mechanisms; customer communication channel choice.

(2) **Customer added value:** mass customization and personalization; economic incentives.

(3) **Customer profiling:** collection and analysis of customer information; value-added exchange for information.

NOTES

IP protocols, XML, and Internet telephony are not only cost-efficient but also enable better contacts with customers.

But the core concern over technology-based CRM remains the inattention to human factors in the development and nurture of the buyer-seller relationship. These human factors go beyond the user-friendly design and deployment of CRM. It involves building and sustaining trust in the relationship, developing emotional and structural bonds with customers, and demonstrating sincerity and commitment in the relationship-all of which find only partial and imperfect solutions when left to technology alone. For example, Internet based CRM approaches, while recognizing the importance of trust, view trust in the very limited context of providing and guaranteeing privacy and security when dealings with customers through websites CRM strategies must be based on a detailed evaluation of the loyalty-profitability link as well as the real nature of the bond between the firm and the customer.

5.3 E-CRM Solutions

Once a CRM plan has been developed, with key objectives and goals identified, eCRM deployment can begin. There are two main options available for small and medium size businesses to deploy eCRM:

5.3.1 Contact Management Software

Easy to install CRM software for small organisations which comprise a directory of customer details and allows sales and activity reports to be generated. The following organisations provide this type of eCRM Software

- **Contact Business Communications** *(www.contactsoftware.com)*
- **ACT** *(www.act.com)*
- **Act Today** *(www.acttoday.com.au)*
- **Maximizer** *(www.maximizer.com.au)*
- **Legrand CRM** *(www.legrandsoftware.com.au)*
- **Vital Software** *(www.vitalsoftware.net)*

5.3.2 Customized CRM solutions

Customised CRM solutions for medium size businesses often require enterprise software, service and support from external providers. There are a wide range of software vendors and consultants that provide CRM solutions. These vary from simple applications to the implementation of comprehensive software, hardware and customer relationship ideologies. Some key providers include:

- **Siebel**(*www.siebel.com)*
- **Frontrange Solutions (Goldmine)** *(www.frontrange.com.au)*
- **SalesLogix** (*www.saleslogix.com)*
- **SalesForce** (*www.salesforce.com/au)*

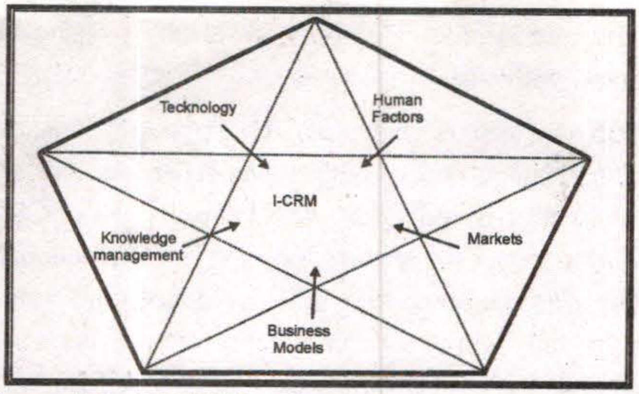


Figure 5.2 : E-CRM ->Input-process-output model

The model consists of five input variables, representing each of the five areas, and e-CRM processes and outcomes. The combination of these five input variable categories captures the human participation, economic environment, strategic considerations, technical infrastructure, and intellectual capital components of e-CRM. e-CRM processes are continuous and evolutionary in nature, just as are the relationships they attempt to foster and manage for mutual gain. Feedback loops illustrate that e-CRM outcomes can result in learning for both customers and enterprises. Both types of outcomes – performance and non-performance – are important. Non-performance outcomes may have significant negative or positive impacts on performance outcomes.

5.2 How Technology can help in CRM

The Internet has been both a boon and a bane for customer relationship management. On one hand, the lowered costs of market entry (or the creation of an additional communication and/or sales channel) increases the competition for customer attention and sales, while concurrently reducing seller margins through reductions in buyer search costs as well. This places additional pressures on firms to seek out their most valuable customers and devise programs and strategies to retain them. Fortunately, the Internet also helps firms pursue such objectives. Firms can now understand customer needs better, develop more customer-centric programs for satisfying needs, and offer enhanced value through managing customer information and needs, and providing customized products and services.

The Internet as a technological tool adds greater value to CRM, primarily through making the various stages/strategies of CRM more cost-efficient as well as enabling a host of other activities that would have otherwise been either impossible or arduous.

The Internet enables ready identification of the customer/visitor, cost-efficient data collection, personalization, customization, and interactivity in the CRM process. These enhanced value-creating activities also expand the abilities of firms to "establish, nurture, and sustain long-term customer relationships than ever before" While traditional CRM activities remained distinct from web-enabled CRM, or E-CRM, in the early years of the "dot-com" era, it is being increasingly recognized that web-enabled CRM is now the norm rather than the exception. This is due to the fact that most firms now view the Internet as an additional channel either for communication or sales, or both, and that firms are recognizing that open-standard Internet TCP/

The present research draws on the theory of dynamic capabilities, the resource based view and the E-CRM literature to develop a research model which conceptualizes E-CRM in an SME context.

NOTES

Dynamic capabilities describes an organization's ability to use ICT and complementary organizational resources for develop unique, change - oriented capabilities that enable organizations to respond to the market. ICT provides the building blocks to form capabilities, and an organization's ability to enhance these capabilities leads to superior performance integrating dynamic capabilities theory with the CRM.

The capabilities of E-CRM can be categorised in two forms -

Traditional CRM Capabilities

- Load Management
- Task Management
- Surveys Management
- Literature Fulfillment
- Knowledge Base
- Analytics
- Compaign Management
- List Management
- Marketing Encyclopedia
- Sales Process Automation
- Forecast
- Pipeline Management
- Service Automation
- Support Automation
- Product History
- Quality Assurance

Web CRM Capabilities

- Online lead capture
- Online surveys
- Online literature fulfillment
- E-commerce
- Online configuration
- Web self-help
- Email-based support
- Online product registration
- Online profile management

5.4 Advantages of E-CRM

Implementation of an E-CRM system enables an organization to streamline processes and provide sales, marketing and service personnel with better, more complete customer information. The result is that E-CRM allows organisations to build more profitable customer relationships and decrease operating costs.

Direct benefits of an E-CRM system include:

- **Service level improvements** – Using an integrated database to deliver consistent and improved customer responses.
- **Revenue growth** – Decreasing costs by focusing on retaining customers and using interactive service tools to sell additional products.
- **Productivity** – Consistent sales and service procedures to create efficient work processes.
- **Customer satisfaction** – Automatic customer tracking and detection will ensure enquiries are met and issues are managed. This will improve the customer's overall experience in dealing with the organization.
- **Automation** - E-CRM software helps automate campaigns including:
  - ✓ Telemarketing
  - ✓ Telesales
  - ✓ Direct mail
  - ✓ Lead tracking and response
  - ✓ Opportunity management
  - ✓ Quotes and order configuration

Across every sector and industry, effective CRM is a strategic imperative for corporate growth and survival:

- Sales organisations can shorten the sales cycle and increase key sales-performance metrics such as revenue per sales representative, average order size and revenue per customer.
- Marketing organisations can increase campaign response rates and marketing driven revenue while simultaneously decreasing lead generation and customer acquisition costs.
- Customer service organisations can increase service agent productivity and customer retention while decreasing service costs, response times and request-resolution times.

5.5 E-CRM Capabilities

Customer Relationships have been extensively investigated in the marketing and strategic management literature.

5.6 Data Mining and E-CRM

CRM is a process that manages the interactions between a company and its customers. The primary users of CRM software applications are database marketers who are looking to automate the process of interacting with customers.

To be successful, database marketers must first identity market segments containing customers or propects with high-profit potential. They then build and execute compaigns that favorably impact the behaviour of these individuals.

The first task, identifying market segments, requires significant data about prospective customers and their buying behaviour. In theory, the more data the better. In practice, however, massive data store often imped marketers, who struggle to sift through the minutiae to find the nuggets of valuable information.

Recently, marketers have added a new class of software to their targetting arsenal. Data mining applications automate the process of searching the mountains of data to find patterns that are good predictors of purchasing behaviors.

After mining the data, marketers must feed the results into compaign management software that, as the name implies, manages the compaign directed at the defined market segments.

In the past, the link between data mining and CRM was mostly manual. In the worst cases it involved "Sneakar net", crating a physical file on tape or disks, which someone then carried to another computer and loaded into the marketing database.

This separation of data mining and CRM introduces considerable inefficiency and opens the door for human errors. Tightly integrating the two disciplines presents opportunity for competitive advantages.

5.7 Example of E-CRM

5.7.1 Information Management

- CRM software helps sales marketing, and service professionals fetch and track relevant data about every past and planned contact with prospects and customers, as well as other business and life cycle events of customers.
  - Information is fetched from all customer touch-points, such as telephone, fax, e-mail, the company's website, retail stores, kiosks and personal contact.
  - CRM systems store the data in a common customer database that integrates all customer account information and makes it available throughout the company via Internet, Intranet, or other network links for sales marketing, service and other CRM application.

5.7.2 Sales

- A CRM system support sales representatives to manage their sales activities, and optimise cross selling and up selling.

NOTES

For example : Sales prospect and product information, product configuration and sales quote generation capabilities are provided.

- CRM also gives them real time access to a single common view of the customer enabling them to check on all aspects of a customer's account status and history before scheduling their sales calls.

  For example : A CRM system would alert a bank sales person to call customers who make large deposits to sell them premier credit or investment services. Or it would alert a salesperson of unresolved service, delivery, or payment problems that could be resolved through a personal contact with a customer ·

5.7.3 Marketing

- CRM systems help marketing professionals accomplish direct marketing campaigns by automating such tasks as qualifying leads for targeted marketing and scheduling and tracking direct marketing mailings, then the CRM software helps marketing professionals capture and manage prospect and customer response data in the CRM database, and analyse the customer and business value of a company's direct marketing campaigns.
- CRM also assists in the fulfilment of prospect and customer responses and requests by quickly scheduling sales contacts and providing appropriate information on products and services to them, while capturing relevant information for the CRM database.

5.7.4 Customer Service and Support

A CRM system provides common customer database shared by sales and marketing professionals CRM helps customer service managers create, assign and manage requests for service by customers Call Centre software :

1. Routes calls to customer,
2. Supports agents based on their skills and authority to handle specific kinds of service requests.
3. Web-based self-service enables customers to easily access personalized support information at the company website, while giving them an option to receive further assistance online or by phone from customer service personnel.

5.7.5 Customer Retention

- More efforts are required to sell to a new customer than to sell to an existing one.
- A typical dissatisfied customer will do bad publicity about his or her experience.
- A company can boost its profits 85 percent by increasing its annual customer retention by only 5 percent.
- The odds of selling product to a new customer are 15 percent, whereas the odds of selling a product to an existing customer are 50 percent.

That's why enhancing and optimising customer retention and loyalty is a major business strategy and primary objective of customer relationship management.

5.8 Implementing an E-CRM System

NOTES

When approaching the development and implementation of eCRM there are important considerations to keep in mind:

- **Define customer relationships** – Generate a list of key aspects of your customer relationships and the importance of these relationships to your business.
- **Develop a plan** – Create a broad Relationship Management program that can be customised to smaller customer segments. A suitable software solution will help deliver this goal.
- **Focus on customers** - The focus should be on the customer, not the technology. Any technology should have specific benefits in making customers' lives easier by improving support, lowering their administrative costs, or giving them reasons to shift more business to your company.
- **Save money** – Focus on aspects of your business that can contribute to the bottom line. Whether it is through cutting costs or increasing revenue, every capability you implement should have a direct, measurable impact on the bottom line.
- **Service and support** - By tracking and measuring the dimensions of the relationship, organizations can identify their strengths and weaknesses in the relationship management program and continually fine tune it based on ongoing feedback from customers.

5.9 E-CRM Framework

In today's world, customers interact with an organization via multiple communication channels—the World Wide Web, call centres, field salespeople, dealers and partner networks. Many organisations also have multiple lines of business that interact with the same customers. eCRM systems enable customers to do business with the organisation the way the customer wants - any time, via any channel, in any language or currency—and to make customers feel that they are dealing with a single, unified organisation that recognizes them every step of the way.

The eCRM system does this by creating a central repository for customer records and providing a portal on each employee's computer system allowing access to customer information by any member of the organisation at any time.

Through this system, eCRM gives you the ability to know more about customers, products and performance results using real time information across your business.

NOTES

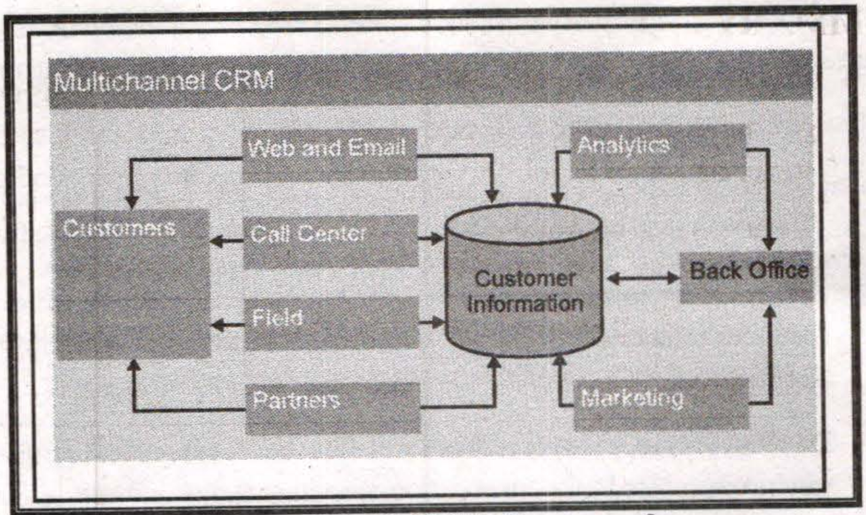


Figure 5.3 : E-CRM Framework

5.10 Next Generation CRM

Current CRM implementation through the Internet could also be evaluated through an examination of strategies/tactics that are well enabled by technology and others that may require more human and personal intervention. In this context, it is useful to make a distinction between a firm's transactional goals and its relational goals.

Incentive-based programs, such as frequency and loyalty programs, may serve to provide only transactional benefits .However, technology-based CRM programs may quite adequate for achieving such transactional goals. For example, customers in B2C markets may find Internet-based personalization for obtaining loyalty rewards, easy ordering, delivery tracking, and routine pre- and post sales service support to be of value. Similarly, customers in B2B markets may find cost-efficiencies and value enhancements in routine purchases through B2B exchanges. On the other hand, a pure technology solution may be inadequate for several customer issues, including communication on critical issues, handling complaints, service recovery, and for developing trust and deeper structural ties.

For example, in both work and non-work contexts, computer-mediated communications, such as e-mails, are perceived by users to be less valuable for developing social relationships as compared to offline communication methods, such as face-to-face contact and telephone conversations.

Given that the primary CRM objective is to enable a long-term relationship with customers, it follows that the relational dimension and even some of the transactional dimensions of relationship building and maintenance would require the use of more personal, human contacts.

For technology-based CRM to be deployed successfully, it must be secondary to the personal, human relationships formed between the firm and its customers, at least until such time as a value migration occurs between the offline and online domains.

SUMMARY

NOTES

- Customer Relationship Management (CRM) can be defined as the strategies, processes, people and technologies used by companies to successfully attract and retain customers for maximum corporate growth and profit.
- The use of customer relationship management products, CRM software and CRM solutions enhance the effective implementation of CRM in an organization.
- E-CRM is a business strategy that utilizes the power of technology to tie together all aspects of a company's business with the goal of building long-term customer loyalty.
- According to Hansen : E-CRM is defined as "a process of acquiring, retaining and growing profitable customers. It requires a clear focus on the service attributes that represent value to the customer ant that create loyalty."
- Implementation of an E-CRM system enables an organization to streamline processes and provide sales, marketing and service personnel with better, more complete customer information.
- E-CRM software provides profiles and histories of each interaction the organization has with its customers, making it an important tool for all small and medium businesses.
- The Internet as a technological tool adds greater value to CRM, primarily through making the various stages/strategies of CRM more cost-efficient as well as enabling a host of other activities that would have otherwise been either impossible or arduous.
- Management of Customer Relationship in Business Media (MCR-BM)-concept: This concept is defined as "The Management of Customer Relationship in Business Media comprises the design, development and application of holistic concepts in order to manage relationships to economically valuable current or future customers".

EXERCISE

1. What is CRM and how it is playing a vital role in modern business management ?
2. Prove the statement with suitable example "By providing the means to manage and coordinate customer interactions, CRM helps companies maximize the value of every customer interaction and in turn improve corporate performance".
3. How can you categorize the customers on the basis of different zones ? What are its significance ?
4. What is E-CRM ? How it different from traditional CRM ?

NOTES

5. Give some evidence to prove this statement "E-CRM is a business strategy that utilizes the power of technology to tie together all aspects of a company's business with the goal of building long-term customer loyalty".
6. Describe the business benefits of E-CRM to the business organizations .
7. Describe the features of E-CRM software systems .
8. "Internet is the backbone of E-CRM " prove it .
9. What are different goals of E-CRM?
10. Describe the significance of E-CRM Business Models ?
11. What is Management of Customer Relationship in Business Media (MCR-BM)-concept ?
12. What are technologies behind the E-CRM ?
13. What are major considerations before implementing the E-CRM in organizations?
14. Describe the working and E-CRM framework in brief?
15. What are different popular E-CRM products available in the market ?

6 Virtual Reality

Chapter Includes :

- INTRODUCTION
- HISTORY OF VIRTUAL REALITY
- VIRTUAL REALITY
- VIRTUAL REALITY APPLICATION
- IMPACT OF VR
- SIMULATED REALITY
- VIRTUAL REALITY : HARDWARE
- LEVELS OF VR HARDWARE SYSTEMS
- VR SOFTWARE SYSTEMS
- ASPECTS OF VR PROGRAM
- WORLD SPACE
- WORLD COORDINATES
- WORLD DATABASE
- CONTROL PANELS
- TYPES OF VR SYSTEMS
- VR CHALLENGES

6.1 Introduction

The term virtual reality was coined in 1987 by Jaron Lanier, whose research and engineering contributed a number of products to the nascent VR industry. A common thread linking early VR research and technology development in the United States was the role of the federal government, particularly the Department of Defense, the National Science Foundation, and the National Aeronautics and Space Administration (NASA). Projects funded by these agencies and pursued at university-based research laboratories yielded an extensive pool of talented personnel in fields such as computer graphics, simulation, and networked environments and established links between academic, military, and commercial work.

NOTES

Artists, performers, and entertainers have always been interested in techniques for creating imaginative worlds, setting narratives in fictional spaces, and deceiving the senses. Illusionary spaces created by paintings or views have been constructed for residences and public spaces since antiquity, culminating in the monumental panoramas of the 18th and 19th centuries. Panoramas blurred the visual boundaries between the two-dimensional images displaying the main scenes and the three-dimensional spaces from which these were viewed, creating an illusion of immersion in the events depicted. This image tradition stimulated the creation of a series of media—from futuristic theatre designs, stereopticons, and 3-D movies to IMAX movie theatres—over the course of the 20th century to achieve similar effects. For example, the Cinerama widescreen film format, originally called Vitarama when invented for the 1939 New York World's Fair by Fred Waller and Ralph Walker, originated in Waller's studies of vision and depth perception. Waller's work led him to focus on the importance of peripheral vision for immersion in an artificial environment, and his goal was to devise a projection technology that could duplicate the entire human field of vision. The Vitarama process used multiple cameras and projectors and an arc-shaped screen to create the illusion of immersion in the space perceived by a viewer. Though Vitarama was not a commercial hit until the mid-1950s (as Cinerama), the Army Air Corps successfully used the system during World War II for anti-aircraft training under the name Waller Flexible Gunnery Trainer—an example of the link between entertainment technology and military simulation that would later advance the development of virtual reality.

An important area of application for VR systems has always been training for real-life activities. The appeal of simulations is that they can provide training equal or nearly equal to practice with real systems, but at reduced cost and with greater safety.

By 1985, Fisher had also left Atari to join NASA's Ames Research Center at Moffett Field, California, as founding director of the Virtual Environment Workstation (VIEW) project. The VIEW project put together a package of objectives that summarized previous work on artificial environments, ranging from creation of multisensory and immersive "virtual environment workstations" to telepresence and teleoperation applications. Influenced by a range of prior projects that included Sensorama, flight simulators, and arcade rides, and surprised by the expense of the air force's Darth Vader helmets, Fisher's group focused on building low-cost, personal simulation environments. While the objective of NASA was to develop telerobotics for automated space stations in future planetary exploration, the group also considered the workstation's use for entertainment, scientific, and educational purposes. The VIEW workstation, called the Virtual Visual Environment Display when completed in 1985, established a standard suite of VR technology that included a stereoscopic head-coupled display, head tracker, speech recognition, computer-generated imagery, data glove, and 3-D audio technology.

Today, virtual reality is poised to change the way we interact with and control computers. Like the introduction of computers more than 50 years ago, its impacts are unknown. Will there be VR in every house, classroom, and office? Will immersing oneself in a computer-generated world be as commonplace as watching a movie?

About the only thing that does seem certain about VR is that it will grow and develop. And as the technology matures, it will become better, cheaper, and more accessible. Furthermore, the networks that link computers will expand, making it possible for VR to weave its way into our daily lives. Clearly, the future of VR is limited only by our imaginations.

6.2 History of Virtual Reality

NOTES

The thought of Virtual Reality has been around since 1965, when Ivan sutherland expressed his ideas of creating virtual or imaginary world. AT MIT, he conducted experiments with three dimensional displays. In 1969, he developed the first system to surround people in three dimensional displays of information. Between the 70's and late 80's, the concept of Virtual Reality was mainly used by the united states. The military used it as flight simulators to train pilots. The other countries in the world did not show any interest in this technology until the late 1980's. Since then, Virtual Reality developed in many ways to become an emerging technology of our time.

Virtual Reality may have popped into the headlines only in the past few years, but its roots reach back four decades. It was in the late 1950, Just as the nation was shaking off stale traces of McCarthyism and was shaking to the sounds of Elvis, that an idea arose that would change the way people interacted with computers and Make possible Virtual Reality.

At this, computers were locked in A.C. rooms and used only secret programming languages. But a young electrical engineer noval radar technician named Douglas Engelbart viewed them differently. Engelbart, Imagin them as tools for digital display. He knew from his days that with radar that any digital information could be viewed on a screen.

At first, Engelbart's ideas were dismissed, but by the early 1960 other people started thinking in the same way. Communications technology was intersecting with computing and graphics technology. The first computer based on transistors rather than vacuum tubes became available.

Fear of nuclear attack prompted the U.S. military to commission a new radar system that would process large amounts of information and immediately display it in a form that humans could readily understand. The resulting radar defense system was the first "real time", or simulation of data.

In 1962 Ivan Sutherland developed a light pen with which images could be sketched on a computer. Sutherland's first computer-aided design program, called sketchpad, opened the way for designers to use computers to create blueprints of automobiles, cities, and industrial products. By 1970, Sutherland also uses scientific visualization in computer graphics to transform columns of data into images.

A goal of scientific visualization is to capture the dynamic qualities of systems or processes in its images. In the 1980, borrowing and as well as creating many of the special effects techniques of Hollywood, scientific visualization moved into animation. In 1990, NCSA's award-winning animation of Smog descending upon Los Angeles influenced air pollution legislation in the state.

Scientists wanted interactivity. So did the military, industry, business and entertainment. The demand for interactivity pushed computer visualization to the limits, towards virtual reality.

6.3 Virtual Reality

Definition

Virtual reality (VR) refers to computer simulations of real-world "environments" that use 3-D graphics and external devices like a dataglove or helmet to allow users

NOTES

to interact with the simulation. Users move through virtual reality environments as though they were navigating in real worlds — walking through structures and interacting with objects in the environment.

"Virtual Reality is a way for humans to visualize, manipulate and interact with computers and extremely complex data" .

The visualization part refers to the computer generating visual, auditory or other sensual outputs to the user of a world within the computer. This world may be a CAD model, a scientific simulation, or a view into a database. The user can interact with the world and directly manipulate objects within the world. Some worlds are animated by other processes, perhaps physical simulations, or simple animation scripts. Interaction with the virtual world, at least with near real time control of the viewpoint, in my opinion, is a critical test for a 'virtual reality'.

Virtual reality (VR) is a technology which allows a user to interact with a computer-simulated environment, be it a real or imagined one. Most current virtual reality environments are primarily visual experiences, displayed either on a computer screen or through special stereoscopic displays, but some simulations include additional sensory information, such as sound through speakers or headphones. Some advanced, haptic systems now include tactile information, generally known as force feedback, in medical and gaming applications. Users can interact with a virtual environment or a virtual artifact (VA) either through the use of standard input devices such as a keyboard and mouse, or through multimodal devices such as a wired glove, the Polhemus boom arm, and omnidirectional treadmill. The simulated environment can be similar to the real world, for example, simulations for pilot or combat training, or it can differ significantly from reality, as in VR games. In practice, it is currently very difficult to create a high-fidelity virtual reality experience, due largely to technical limitations on processing power, image resolution and communication bandwidth. However, those limitations are expected to eventually be overcome as processor, imaging and data communication technologies become more powerful and cost-effective over time.

The origin of the term virtual reality is uncertain. *The Judas Mandala*, a 1982 science fiction novel by Damien Broderick where the context of use is somewhat different from that defined above. The VR developer Jaron Lanier claims that he coined the term. A related term coined by Myron Krueger, "artificial reality", has been in use since the 1970s. The concept of virtual reality was popularized in mass media by movies such as *Brainstorm* and *The Lawnmower Man* , and the VR research boom of the 1990s was motivated in part by the non-fiction book *Virtual Reality* by Howard Rheingold. The book served to demystify the subject, making it more accessible to less technical researchers and enthusiasts, with an impact similar to what his book *The Virtual Community* had on virtual community research lines closely related to VR.

In the late 1980s the term "virtual reality" was popularized by Jaron Lanier, one of the modern pioneers of the field. Lanier had founded the company VPL Research (from "Virtual Programming Languages") in 1985, which developed and built some of the seminal "goggles n' gloves" systems of that decade.

6.4 Virtual Reality Application (Present uses of Virtual Reality)

NOTES

- **Games**

In 1991, the company (originally W Industries, later renamed) Virtuality licensed the Amiga 3000 for use in their VR machines and released a VR gaming system called the 1000CS. This was a stand-up immersive HMD platform with a tracked 3D joystick. The system featured several VR games including *Dactyl Nightmare* (shoot-em-up), *Legend Quest* (adventure and fantasy), *Hero* (VR puzzle), *Grid Busters* (shoot-em-up). *Virtual Reality 1 Glasses Personal Display System* is a visor and headphones headset that is compatible with any video input including 3D broadcasting, and usable with most game systems (Nintendo, PlayStation, etc.). *Virtual Reality World 3D Color Ninja* game comes with headset visor and ankle and wrist straps that sense the player's punches and kicks. *Virtual Reality Wireless TV Tennis Game* comes with a toy tennis racket that senses the player's swing, while *Wireless TV Virtual Reality Boxing* includes boxing gloves that the player wears and jabs with. Nintendo's Virtual Boy was sold for only one year, 1995. Aura Interactor Virtual Reality Game Wear is a chest and back harness through which the player can feel punches, explosions, kicks, uppercuts, slam-dunks, crashes, and bodyblows. It works with Sega Genesis and Super Nintendo.

In *Kingdom Hearts II*, the character Roxas lives in a virtual Twilight Town until he merges with Sora. In *System Shock*, the player has implants making him able to enter into a kind of cyberspace. Its sequel, *System Shock 2* also features some minor levels of VR.

- **Education, training, and life-long learning**

Interactive computing and communications technology are ushering in a new era of education in which students and teachers, separated by a distance, are conducting research and performing experiments through high-speed connections that will eventually incorporate VR.

- **Manufacturing and industrial design**

Major companies are maintaining their global competitiveness by designing products better and faster with virtual reality. General Motors, for instance, is using a virtual prototype rather than a physical model of one of its new model cars to evaluate the interior design for aesthetics, engineering, safety, and ergonomic features. Caterpillar refined its new backhoe and wheel loaders electronically with virtual reality technology.

- **Medicine**

To study the circulation of blood simulated a heart by treating the heart wall as a set of fibers immersed in fluid and responding to both fluid forces and tension forces. The fluid, in turn, experiences a force field in the neighborhood of the fibers that prevents flow through the gaps in the fiber network, allowing the heart to pump the fluid.

- **Material sciences**

Researchers are modeling the more than 400 different hydrogen-nitrogen chemical reactions in an internal combustion engine to design cooler, more efficient car engines. Other researchers are analyzing the properties of compounds in a race

NOTES

to discover the next generation of superconducting material.

- **Relativity**

  By simulating the gravitational ripples that would be generated if two black holes collided, researchers hope to confirm the existence of these elusive objects predicted as a consequence of Einstein's famous General Theory of Relativity. Should the simulated ripples precisely match gravitational waves detected by LIGO,an array of sensing devices that will become operational in 2000, not only could the existence of black holes be confirmed but also Einstein's 80-year-old theory finally will be vindicated.

- **Weather forecasting**

  When Hurricane Emily approached the Atlantic coast in 1993, a new hurricane model accurately predicted 48 hours in advance that the hurricane would turn sharply back out to sea off Cape Hatteras without making landfall. Predicting long-term weather patterns is one of the outcomes of the new monitoring and instrumentation tools being developed at part of the HPCC.

  Atmospheric scientists are also turning to advanced computing tools and virtual environments like the CAVE to calculate the behavior of more local disturbances, particularly thunderstorms .

- **Molecular biosciences**

  VR models of some of life's smallest structures are helping scientists decipher the precise mechanisms through which proteins communicate with each other, for instance, to bind antibody to antigen or signal a cell membrane to dilate. Knowledge of this kind is speeding the development of biological and industrial catalysts as well as therapeutic drugs.

- **Mass media**

  Mass media has been a great advocate and perhaps a great hindrance to its development over the years. During the research "boom" of the late 1980s into the 1990s the news media's prognostication on the potential of VR — and potential overexposure in publishing the predictions of anyone who had one (whether or not that person had a true perspective on the technology and its limits) — built up the expectations of the technology so high as to be impossible to achieve under the technology then or any technology to date. Entertainment media reinforced these concepts with futuristic imagery many generations beyond contemporary capabilities.

- **Television**

  Perhaps the earliest example of virtual reality on television is a *Doctor Who* serial "The Deadly Assassin". This story, first broadcast in 1976, introduced a dream-like computer-generated reality known as the Matrix (no relation to the film). The first major television series to showcase virtual reality was *Star Trek: The Next Generation*. They featured the holodeck, a virtual reality facility on starships, that enabled its users to recreate and experience anything they wanted. One difference from current virtual reality technology, however, was that replicators, force fields, holograms, and transporters were used to actually recreate and place objects in the holodeck, rather than relying solely on the illusion of physical objects, as is done today.

In Japan and Hong Kong, the first anime series to use the idea of virtual reality was *Video Warrior Laserion* (1984).

An anime series known as *Lain:Serial Experiments* included a virtual reality world known as "The Wired" that eventually co-existed with the real world.

Channel 4's *Gamesmaster* (1992–1998) also used a VR headset in its "tips and cheats" segment.

In 2005, Brazilian's Globo TV features a show where VR helmets are used by the attending audience in a space simulation called *Conquista de Titã*, broadcasted for more than 20 million viewers weekly.

- **Marketing**

  A side effect of the chic image that has been cultivated for virtual reality in the media is that advertising and merchandise have been associated with VR over the years to take advantage of the buzz. This is often seen in product tie-ins with cross-media properties, especially gaming licenses, with varying degrees of success. The NES Power Glove by Mattel from the 1980s was an early example as well as the U-Force and later, the Sega Activator. Marketing ties between VR and video games are not to be unexpected, given that much of the progress in 3D computer graphics and virtual environment development (traditional hallmarks of VR) has been driven by the gaming industry over the last decade.

- **Real estate**

  The real estate sector has used the term "virtual reality" for websites that offer panoramic images laced into a viewer such as QuickTime player in which the viewer can rotate to see all 360 degrees of the image.

A Phoenix-based research and design company has launched what they call a VuPOD. Using this device, a prospective buyer can request the images of any property for sale and see them in the VuPOD. The images immerse the viewer in a full 360 degree image that surrounds them, complete with floor plans of the property so the viewer can tell where the room is in relation to the rest of the property.

6.5 Impact of VR

There has been increasing interest in the potential social impact of new technologies, such as virtual reality . Mychilo S. Cline, in his book, *Power, Madness, and Immortality: The Future of Virtual Reality*, argues that virtual reality will lead to a number of important changes in human life and activity. He argues that:

- Virtual reality will be integrated into daily life and activity and will be used in various human ways.
- Techniques will be developed to influence human behavior, interpersonal communication, and cognition (i.e., virtual genetics).
- As we spend more and more time in virtual space, there will be a gradual "migration to virtual space," resulting in important changes in economics, worldview, and culture.
- The design of virtual environments may be used to extend basic human rights into virtual space, to promote human freedom and well-being, and to promote

NOTES

social stability as we move from one stage in socio-political development to the next.

6.6 Simulated Reality

Simulated reality(SR) is the idea that reality could be simulated — often computer-simulated — to a degree indistinguishable from 'true' reality. It could contain conscious minds which may or may not know that they are living inside a simulation. In its strongest form, the "Simulation Hypothesis" claims we actually are living in such a simulation.

This is different from the current, technologically achievable concept of virtual reality. Virtual reality is easily distinguished from the experience of 'true' reality; participants are never in doubt about the nature of what they experience. Simulated reality, by contrast, would be hard or impossible to distinguish from 'true' reality.

The idea of a simulated reality raises several questions:

- Is it possible, even in principle, to tell whether we are in a simulated reality?
- Is there any difference between a simulated reality and a 'real' one?
- How should we behave if we knew that we were living in a simulated reality?

6.6.1 Types of Simulation

- **Brain-computer interface**

  In a brain-computer interface simulation, each participant enters from outside, directly connecting their brain to the simulation computer. The computer transfers sensory data to them and reads their desires and actions back; in this manner they interact with the simulated world and receive feedback from it. The participant may even receive adjustment in order to temporarily forget that they are inside a virtual realm (e.g. "passing through the veil"). While inside the simulation, the participant's consciousness is represented by an avatar, which could look very different from the participant's actual appearance.

- **Virtual people**

  In a virtual-people simulation, every inhabitant is a native of the simulated world. They do not have a 'real' body in the 'outside' reality. Rather, each is a fully simulated entity, possessing an appropriate level of consciousness that is implemented using the simulation's own logic (i.e. using its own physics). As such, they could be downloaded from one simulation to another, or even archived and resurrected at a later date. It is also possible that a simulated entity could be moved out of the simulation entirely by means of mind transfer into a synthetic body (an example of this is when SID 6.7 escapes his simulated reality in the movie Virtuosity).

  This category subdivides into two further types:

  - **Virtual people-virtual world**, in which an external reality is simulated separately to the artificial consciousnesses; and
  - **Solipsistic simulation** in which consciousness is simulated and the "world" participants perceive exists only within their minds.

- **Emigration**

  In an emigration simulation, the participant enters the simulation from the outer reality, as in the brain-computer interface simulation, but to a much greater degree. On entry, the participant uses mind transfer to temporarily *relocate* their mental processing into a virtual-person. After the simulation is over, the participant's mind is transferred back into their outer-reality body, along with all new memories and experience gained within.

- **Intermingled**

  An intermingled simulation supports both types of consciousness: players from the outer reality who are visiting (as a brain-computer interface simulation) or emigrating, and virtual-people who are natives of the simulation and hence lack any physical body in the outer reality.

  The Matrix movies feature an intermingled type of simulation: they contain not only human minds (with their physical brains remaining outside), but also the 'agents', who are sovereign software programs indigenous to the computed realm.

6.6.2 Recursive Simulations

A simulated reality could contain a computer that is running a simulated reality. The 'parent' simulator would be simulating all of the atoms of the computer, atoms which happen to be calculating a 'child' simulation. By way of illustration: imagine that a human is playing a game of The Sims in which one of the player's Sims (simulated people) is playing a computer game *in the game*.

This recursion could continue to infinitely many levels — a simulation containing a computer running a simulation containing a computer running a simulation and so on. The recursion is subject only to one constraint: each 'nested' simulation must be:

- **smaller** than its parent reality, because its own memory must be a subset of the parent's;

...**and** must be at least one of the following:

- **slower** than its parent reality, because its own calculations must be a subset of the parent's; **or**
- **less complex** than its parent reality, via simplifications of processes that are computationally intensive in the parent reality; **or**
- **less complete** than its parent reality, via approximations of objects that nobody is observing.

The latter is the basis of the idea that quantum uncertainties are circumstantial evidence that our own reality is a simulation. However, this assumes that there is a finite limitation somewhere in the chain. Assuming an infinite number of simulations within simulations, there need not be any noticeable difference between any of the subsets.

6.5.3 Processing Power

A computer simulation would be limited to the processing power of its host computer, and so there may be aspects of the simulation that are not computed at a fine-grained (e.g. subatomic) level. This might show up as a limitation on the accuracy of information that can be obtained in particle physics.

However, this argument, like many others, assumes that accurate judgments about the simulating computer can be made from within the simulation. If we are being simulated, we might be misled about the nature of computers.

Taken one step further, the "fine grained" elements of our world could themselves be simulated since we never see the sub-atomic particles due to our inherent physical

NOTES

limitations. In order to see such particles we rely on other instruments which appear to magnify or translate that information into a format our limited senses are able to view: computer print out, lens of a microscope, etc. Therefore, we essentially take on faith that they're an accurate portrayal of the fine grained world which appears to exist in a realm beyond our natural senses. Assuming the sub-atomic could also be simulated then the processing power required to generating a realistic world would then be greatly reduced.

6.7 Virtual Reality : Hardware

There are a number of specialized types of hardware devices that have been developed and used for Virtual Reality applications.

6.7.1 Image Generators

One of the most time consuming tasks in a VR system is the generation of the images. Fast computer graphics opens a very large range of applications aside from VR, so there has been a market demand for hardware acceleration for a long while. There are currently a number of vendors selling image generator cards for PC level machines, many of these are based on the Intel i860 processor. The simulator market has produced several companies that build special purpose computers designed expressly for real time image generation. These computers often cost several hundreds of thousands of dollars.

6.7.2 Manipulation and Control Devices

One key element for interaction with a virtual world, is a means of tracking the position of a real world object, such as a head or hand. There are numerous methods for position tracking and control. Ideally a technology should provide 3 measures for position(X, Y, Z) and 3 measures of orientation (roll, pitch, yaw). One of the biggest problem for position tracking is latency, or the time required to make the measurements and preprocess them before input to the simulation engine.

The simplest control hardware is a conventional mouse, trackball or joystick. While these are two dimensional devices, creative programming can use them for 6D controls. There are a number of 3 and 6 dimensional mice/trackball/joystick devices being introduced to the market at this time. These add some extra buttons and wheels that are used to control not just the XY translation of a cursor, but its Z dimension and rotations in all three directions. The Global Devices 6D Controllers is one such 6D joystick .It looks like a racket ball mounted on a short stick. You can pull and twist the ball in addition to the left/right & forward/back of a normal joystick. Other 3D and 6D mice, joystick and force balls are available from Logitech, Mouse System Corp. among others.

One common VR device is the instrumented glove. The use of a glove to manipulate objects in a computer is covered by a basic patent in the USA. Such a glove is outfitted with sensors on the fingers as well as an overall position/orientation tracker. There are a number of different types of sensors that can be used. VPL (holders of the patent) made several DataGloves, mostly using fiber optic sensors for finger bends and magnetic trackers for overall position. Mattel manufactured the PowerGlove for use with the Nintendo game system, for a short time. This device is easily adapted to interface to a personal computer. It provides some limited hand location and finger position data using strain gauges for finger bends and ultrasonic position sensors.

NOTES

The concept of an instrumented glove has been extended to other body parts. Full body suits with position and bend sensors have been used for capturing motion for character animation, control of music synthesizers, etc. in addition to VR applications.

Mechanical armatures can be used to provide fast and very accurate tracking. Such armatures may look like a desk lamp (for basic position/orientation) or they may be highly complex exoskeletons (for more detailed positions). The drawbacks of mechanical sensors are the encumbrance of the device and its restrictions on motion. Exos Systems builds one such exoskeleton for hand control. It also provides force feedback. Shooting Star system makes a low cost armature system for head tracking. Fake Space Labs and LEEP Systems make much more expensive and elaborate armature systems for use with their display systems.

Ultrasonic sensors can be used to track position and orientation. A set of emitters and receivers are used with a known relationship between the emitters and between the receivers. The emitters are pulsed in sequence and the time lag to each receiver is measured. Triangulation gives the position. Drawbacks to ultrasonics are low resolution, long lag times and interference from echoes and other noises in the environment. Logitech and Transition State are two companies that provide ultrasonic tracking systems.

Magnetic trackers use sets of coils that are pulsed to produce magnetic fields. The magnetic sensors determine the strength and angles of the fields. Limitations of these trackers are a high latency for the measurement and processing, range limitations, and interference from ferrous materials within the fields. However, magnetic trackers seem to be one of the preferred methods. The two primary companies selling magnetic trackers are Polhemus and Ascension.

Optical position tracking systems have been developed. One method uses a ceiling grid LEDs and a head mounted camera. The LEDs are pulsed in sequence and the cameras image is processed to detect the flashes. Two problems with this method are limited space (grid size) and lack of full motion (rotations). Another optical method uses a number of video cameras to capture simultaneous images that are correlated by high speed computers to track objects. Processing time (and cost of fast computers) is a major limiting factor here. One company selling an optical tracker is Origin Instruments.

Inertial trackers have been developed that are small and accurate enough for VR use. However, these devices generally only provide rotational measurements. They are also not accurate for slow position changes.

6.7.3 Stereo Vision

Stereo vision is often included in a VR system. This is accomplished by creating two different images of the world, one for each eye. The images are computed with the viewpoints offset by the equivalent distance between the eyes. There are a large number of technologies for presenting these two images. The images can be placed side-by-side and the viewer asked (or assisted) to cross their eyes. The images can be projected through differently polarized filters, with corresponding filters placed in front of the eyes. Anaglyph images user red/blue glasses to provide a crude (no color) stereovision.

The two images can be displayed sequentially on a conventional monitor or projection display. Liquid Crystal shutter glasses are then used to shut off alternate eyes in synchronization with the display. When the brain receives the images in rapid enough succession, it fuses the images into a single scene and perceives depth. A

NOTES

fairly high display swapping rate (min. 60hz) is required to avoid perceived flicker. A number of companies made low cost LC shutter glasses for use with TVs (Sega, Nintendo, Toshiba, etc.). There are circuits and code for hooking these up to a computer available on many of the On-line systems, BBSs and Internet FTP sites mentioned later. However, locating the glasses themselves is getting difficult as none are still being made or sold for their original use. Stereographics sells a very nice commercial LC shutter system called CrystalEyes.

Another alternative method for creating stereo imagery on a computer is to use one of several split screen methods. These divide the monitor into two parts and display left and right images at the same time. One method places the images side by side and conventionally oriented. It may not use the full screen or may otherwise alter the normal display aspect ratio. A special hood viewer is placed against the monitor who helps the position the eyes correctly and may contain a divider so each eye e sees only its own image. Most of these hoods, such as the one for the V5 of Rend386, use fresnel lenses to enhance the viewing. An alternative split screen method orients the images so the top of each points out the side of the monitor. A special hood containing mirrors is used to correctly orient the images.

6.7.4 Head Mounted Display (HMD)

One hardware device closely associated with VR is the Head Mounted Device (HMD).

These use some sort of helmet or goggles to place small video displays in front of each eye, with special optics to focus and stretch the perceived field of view. Most HMDs use two displays and can provide stereoscopic imaging. Others use a single larger display to provide higher resolution, but without the stereoscopic vision.

Most lower cost HMDs use LCD displays, while others use small CRTs, such as those found in camcorders. The more expensive HMDs use special CRTs mounted along side the head or optical fibers to pipe the images from non-head mounted displays. A HMD requires a position tracker in addition to the helmet. Alternatively, the display can be mounted on an armature for support and tracking .

6.7.5 Interface Devices

Datagloves, wands, stairsteppers. These and other interface devices used in virtual environments serve as portals into virtual reality.

- **Datagloves**

  Data gloves offer a simple means of gesturing commands to the computer. Rather than punching in commands on a keyboard, which can be tricky if you're wearing a head-mounted display , you program the computer to change modes in response to the gestures you make with the datagloves.

  One type of dataglove has a web of fiber optic cables along its back. Changes in the amount of light transmitted to the computer by the cables signal how the joints of your fingers are bent. Once the dataglove has been calibrated to your hand, your gestures trigger pre-programmed commands.

  Other gloves use strain sensors over the joints to detect movement. Yet others rely on mechanical sensors to measure your hand movements.

  Some computer users have elaborated on the dataglove concept by creating facial sensors, even body suits. Not many scientists have climbed into these get ups, but animators have. Already, facial movement sensors hooked to computers are simplifying their job: animating cartoons.

- **Wands**

  Wands, the simplest of the interface devices, come in all shapes and variations. Most incorporate on-off buttons to control variables in a simulation or in the display of data. Others have knobs, dials, or joy sticks. Their design and manner of response a retailored to the application. For example, biologists sometimes use wands like scalpels to slice tissue samples from virtual brains.

  Most wands operate with six degrees of freedom; that is, by pointing a wand at an object, you can change its position and orientation in any of six directions: forward or backward, up or down, or left or right. This versatility coupled with simplicity is the reasons for the wand's popularity.

- **Other Input Devices**

  Almost anything can be converted into a sensing device for simulation in virtual reality. Stairsteppers are an example of the limitless manifestations of interface devices. As part of a simulated battlefield terrain, engineers from an army research lab outfitted a stairstepper with sensing devices to detect the speed, direction, and intensity of a soldier's movements in response to the battlefied scenes projected onto a head-mounted display. The stairstepper provided feedback to the soldier by making the stairs easier or more difficult to climb.

6.8 Levels of VR Hardware Systems

The following defines a number of levels of VR hardware systems. These are not hard levels, especially towards the more advanced systems.

6.8.1 Entry VR (EVR)

The 'Entry Level' VR system takes a stock personal computer or workstation and implements a WoW system. The system may be based on an IBM clone (MS-DOS/Windows) machine or an Apple Macintosh, or perhaps a Commodore Amiga. The DOS type machines (IBM PC clones) are the most prevalent. There are Mac based systems, but few very fast rendering ones. Whatever the base computer it includes a graphic display, a 2D input device like a mouse, trackball or joystick, the keyboard, hard disk & memory.

6.8.2 Basic VR (BVR)

The next step up from an EVR system adds some basic interaction and display enhancements. Such enhancements would include a stereographic viewer (LCD Shutter glasses) and an input/control device such as the Mattel PowerGlove and/or a multidimensional (3D or 6D) mouse or joystick.

6.8.3 Advanced VR (AVR)

The next step of the VR technology ladder is to add a rendering accelerator and/or frame buffer and possibly other parallel processors for input handling, etc. The simplest enhancement in this area is a faster display card. For the PC class machines, there are a number of new fast VGA and SVGA accelerator cards. These can make a dramatic improvement in the rendering performance of a desktop VR system. Other more sophisticated image processors based on the Texas Instruments TI34020 or Intel i860 processor can make even more dramatic improvements in rendering capabilities. The i860 in particular is in many of the high end professional

NOTES

systems. The Silicon Graphics Reality Engine uses a number of i860 processors i addition to the usual SGI workstation hardware to achieve stunning levels of realism in real time animation.

An AVR system might also add a sound card to provide mono, stereo or true 3D audio output. Some sound cards also provide voice recognition. This would be an excellent additional input device for VR applications.

6.8.4 Immersion VR (IVR)

An Immersion VR system adds some type of immersive display system: a HMD, a Boom, or multiple large projection type displays (Cave).

An IVR system might also add some form of tactile, haptic and touch feedback interaction mechanisms. The area of Touch or Force Feedback (known collectively as Haptics) is a very new research arena.

A common variation on VR is to use a Cockpit or Cab compartment to enclose the user. The virtual world is viewed through some sort of view screen and is usually either projected imagery or a conventional monitor. The cockpit simulation is very well known in aircraft simulators, with a history dating back to the early Link Flight Trainers. The cockpit is often mounted on a motion platform that can give the illusion of a much larger range of motion. Cabs are also used in driving simulators for ships, trucks, tanks and 'battle mechs'. The latter are fictional walking robotic devices (i.e. the Star Wars films). The BattleTech location based entertainment (LBE) centers use this type of system.

6.8.5 SIMNET: Defense Simulation Internet

One of the biggest VR projects is the Defense Simulation Internet. This project is a standardization being pushed by the USA Defense Department to enable diverse simulators to be interconnected into a vast network. It is an outgrowth of the Defense Advanced Research Projects Administration (DARPA) SIMNET project of the later 1980s. SIMNET was/is a collection of tank simulators (Cab type) that are networked together to allow unit tactical training. Simulators in Germany can operate in the same virtual world as simulators in the USA, partaking of the same battle exercise.

The basic Distributed Interactive Simulation (DIS) protocol has been defined by the Orlando Institute for Simulation & Training. It is the basis for the next generation of SIMNET, the Defense Simulation Internet (DSI).

The basic DIS protocol has been adopted as a standard for communication between distributed simulations by the IEEE.

6.9 VR Software Systems

There are currently quite a number of different efforts to develop VR technology. Each of these projects has different goals and approaches to the overall VR technology. Large and small University labs have projects underway (UNC, Cornell, U.Rochester, etc.). ARPA , NIST, National Science Foundation and other branches of the US Government are investing heavily in VR and other simulation technologies. There are industry supported laboratories too, like the Human Interface Technologies Laboratory (HITL) in Seattle and the Japanese NTT project. Many existing and startup companies are also building and selling world building tools (Autodesk, IBM', Sense8, VREAM).

NOTES

There are two major categories for the available VR software: toolkits and authoring systems. Toolkits are programming libraries, generally for C or C++ that provide a set of functions with which a skilled programmer can create VR applications. Authoring systems are complete programs with graphical interfaces for creating worlds without resorting to detailed programming. These usually include some sort of scripting language in which to describe complex actions, so they are not really non-programming, just much simpler programming. The programming libraries are generally more flexible and have faster renders than the authoring systems, but you must be a very skilled programmer to use them.

6.9.1 Freeware VR Programs

There are currently a few fast rendering programs that have been released with source code and no charge. These programs are generally copyrighted freeware, which means that the original creators retain the copyright and commercial use is restricted. They are not polished commercial programs, and are often written by students. However, these programs exist to give people a very low cost entry into the VR world.

Rend386 is one such freeware library and world player written for 386/486 DOS systems. It was written by Dave Stampe and Bernie Roehl at the University of Waterloo, Canada. It creates images at a resolution of 320x200x256 and supports various extra devices such as the Mattel PowerGlove, LC shutter glasses, Split Screen stereo viewers etc. Rend386 is provided both as a complete world player and as a C source code. It does not provide a full authoring environment for world and object building.

ACK3D is a freeware C programming library developed by Lary Meyer that provides a fast 'raycasting' renderer for PC systems.

Gossamer, a freeware VR package for the Apple Macintosh system, written by Jon Blossom. Source code has not been released yet, but Jon has released a demo and a Think C library.

Multiverse is a freeware UNIX based client/server system written by Robert Grant. It is a multi-user, non-immersive, X-Windows based Virtual Reality system, primarily focused on entertainment/research. It includes capabilities for setting up multi-person worlds and a client/server type world simulation over a local or long haul network. Multiverse source and binaries for several flavors of UNIX are available.

The MRToolkit is a programming library for UNIX systems that is available at no cost from the University of Alberta, but the licensing agreement stipulates no commercial products may be made with it.

VEOS is another programming toolkit that provides a basis for VR development on networked UNIX machines. Source code is available from the Human Interface Technology Lab (HITL) at University of Washington. (ftp.u.washington.edu)

6.9.2 Paid VR Programs

- Virtual Reality Studio (aka VR Studio, VRS) is a very low cost VR authoring system that does allow the user to define their own virtual worlds. This program is also known as "3D Construction Kit" in Europe. The program has a fairly nice graphical interface and includes a simple scripting language.
- Another entrant into the low cost market is the Lepton VR Data Modeling Toolkit. This package is a collection of C programming libraries for real-time 3d data modeling on DOS systems.

NOTES

- For the Macintosh market there are the Qd3d, 3dPane, and SmartPane C++ libraries from ViviStar Consulting . These provide a full suite of 3D graphics functions for popular Macintosh C++ compilers as well as Think C 6.0.
- VREAM is a complete VR authoring package for MS-DOS systems from VREAM, Inc.. It provides a nice GUI environment for creation of objects and worlds, as well as a fairly powerful scripting language. VREAM supports a very wide variety of input and output devices, including HMDs. Two versions of the runtime system are available at a much lower cost to provide only the playback ability. The lower cost runtime will work only with standard VGA display and mouse/joystick. The advanced runtime system supports more devices.
- Virtus Walkthrough, from Virtus Corp., is available for both Mac and Windows systems. It provides a nice 3D modeling package and the ability to interactively control the viewpoint within the created worlds. However, it does not allow for interaction with the objects. The latest version Walkthrough Pro supports texture maps, including QuickTime movies.
- The Sense8 World Tool Kit (WTK) is probably the most widely used product of this type. It runs on a wide variety of platforms from i860 assisted PCs to high end SGI boxes. It has won several awards for excellence.
- The Autodesk Cyberspace Development kit is a C++ library for MSDOS systems using the Metaware HighC/C++ compiler and Pharlap DOS 32bit extender. It supports VESA displays as well as several rendering accelerator boards (SPEA Fireboard, FVS Sapphire, Division's dView).
- Straylight Corp. makes a package called PhotoVR that uses special rendering boards (Intel ActionMedia cards) to provide excellent texture mapped walkthrough environments.
- Dimension International's Superscape VRT3 is a very powerful authoring system for virtual worlds. It provides both a graphical environment for object and world creation and a lower level C library.
- Division Ltd. sells a programming environment for VR called dVS. This package runs on SGI systems, IBM RS/6000 workstations and a proprietary Division workstation.

6.10 Aspects of VR Program

The basic parts of the system can be broken down into an Input Processor, a Simulation Processor, a Rendering Process, and a World Database. All these parts must consider the time required for processing. Every delay in response time degrades the feeling of 'presence' and reality of the simulation.

The **Input Processes** of a VR program control the devices used to input information to the computer. There are a wide variety of possible input devices: keyboard, mouse, trackball, joystick, 3D & 6D position trackers (glove, wand, head tracker, body suit, etc.). A networked VR system would add inputs received from net. A voice recognition system is also a good augmentation for VR, especially if the user's hands are being used for other tasks. Generally, the input processing of a VR system is kept simple. The object is to get the coordinate data to the rest of the system with minimal lag time. Some position sensor systems add some filtering and data smoothing processing. Some glove systems add gesture recognition. This processing step examines the glove inputs and determines when a specific gesture has been made. Thus it can provide a higher level of input to the simulation.

6.10.1 Simulation Process

The core of a VR program is the simulation system. This is the process that knows about the objects and the various inputs. It handles the interactions, the scripted object actions, simulations of physical laws (real or imaginary) and determines the world status. This simulation is basically a discrete process that is iterated once for each time step or frame. A networked VR application may have multiple simulations running on different machines, each with a different time step. Coordination of these can be a complex task.

It is the simulation engine that takes the user inputs along with any tasks programmed into the world such as collision detection, scripts, etc. and determines the actions that will take place in the virtual world.

6.10.2 Rendering Processes

The Rendering Processes of a VR program are those that create the sensations that are output to the user. A network VR program would also output data to other network processes. There would be separate rendering processes for visual, auditory, haptic (touch/force), and other sensory systems. Each renderer would take a description of the world state from the simulation process or derive it directly from the World Database for each time step.

- **Visual Renderer**

  The visual renderer is the most common process and it has a long history from the world of computer graphics and animation. The reader is encouraged to become familiar with various aspects of this technology.

  The major consideration of a graphic renderer for VR applications is the frame generation rate. It is necessary to create a new frame every 1/20 of a second or faster. 20 frames per second (fps) is roughly the minimum rate at which the human brain will merge a stream of still images and perceive a smooth animation. 24fps is the standard rate for film, 25fps is PAL TV, 30fps is NTSC TV. 60fps is Showscan film rate. This requirement eliminates a number of rendering techniques such as raytracing and radiosity. These techniques can generate very realistic images but often take hours to generate single frames.

  Visual renderers for VR use other methods such as a 'painter's algorithm', a Z-Buffer, or other Scanline oriented algorithm. There are many areas of visual rendering that have been augmented with specialized hardware. The Painter's algorithm is favored by many low end VR systems since it is relatively fast, easy to implement and light on memory resources. However, it has many visibility problems.

  The visual rendering process is often referred to as a rendering pipeline. This refers to the series of sub-processes that are invoked to create each frame. A sample rendering pipeline starts with a description of the world, the objects, lighting and camera (eye) location in world space. A first step would be eliminate all objects that are not visible by the camera. This can be quickly done by clipping the object bounding box or sphere against the viewing pyramid of the camera. Then the remaining objects have their geometry's transformed into the eye coordinate system (eye point at origin). Then the hidden surface algorithm and actual pixel rendering is done.

  The pixel rendering is also known as the 'lighting' or 'shading' algorithm. There are a number of different methods that are possible depending on the realism

NOTES

and calculation speed available. The simplest method is called flat shading and simply fills the entire area with the same color. The next step up provides some variation in color across a single surface. Beyond that is the possibility of smooth shading across surface boundaries, adding highlights, reflections, etc.

An effective short cut for visual rendering is the use of "texture" or "image" maps. These are pictures that are mapped onto objects in the virtual world. Instead of calculating lighting and shading for the object, the renderer determines which part of the texture map is visible at each visible point of the object. The resulting image appears to have significantly more detailed than is otherwise possible. Some VR systems have special 'billboard' objects that always face towards the user. By mapping a series of different images onto the billboard, the user can get the appearance of moving around the object.

- **Auditory Rendering**

  A VR system is greatly enhanced by the inclusion of an audio component. This may produce mono, stereo or 3D audio. The latter is a fairly difficult proposition. It is not enough to do stereo-pan effects as the mind tends to locate these sounds inside the head. Research into 3D audio has shown that there are many aspects of our head and ear shape that effect the recognition of 3D sounds. It is possible to apply a rather complex mathematical function (called a Head Related Transfer Function or HRTF) to a sound to produce this effect. The HRTF is a very personal function that depends on the individual's ear shape, etc. However, there has been significant success in creating generalized HRTFs that work for most people and most audio placement. There remains a number of problems, such as the 'cone of confusion' wherein sounds behind the head are perceived to be in front of the head.

  Sound has also been suggested as a means to convey other information, such as surface roughness. Dragging your virtual hand over sand would sound different than dragging it through gravel.

- **Haptic Rendering**

  Haptics is the generation of touch and force feedback information. This area is a very new science and there is much to be learned. There have been very few studies done on the rendering of true touch sense (such as liquid, fur, etc.). Almost all systems to date have focused on force feedback and kinesthetic senses. These systems can provide good clues to the body regarding the touch sense, but are considered distinct from it. Many of the haptic systems thus far have been exo-skeletons that can be used for position sensing as well as providing resistance to movement or active force application.

6.11 World Space

The virtual world itself needs to be defined in a 'world space'. By its nature as a computer simulation, this world is necessarily limited. The computer must put a numeric value on the locations of each point of each object within the world. Usually these 'coordinates' are expressed in Cartesian dimensions of X, Y, and Z (length, height, depth). It is possible to use alternative coordinate systems such as spherical but Cartesian coordinates are the norm for almost all applications. Conversions between coordinate systems are fairly simple (if time consuming).

6.12 World Coordinates

A major limitation on the world space is the type of numbers used for the coordinates. Some worlds use floating point coordinates. This allows a very large range of numbers to be specified, with some precision lost on large numbers. Other systems used fixed point coordinates, which provides uniform precision on a more limited range of values. The choice of fixed versus floating point is often based on speed as well as the desire for a uniform coordinate field.

One method of dealing with the limitations on the world coordinate space is to divide a virtual world up into multiple worlds and provide a means of transiting between the worlds. This allows fewer objects to be computed both for scripts and for rendering. There should be multiple stages (also called rooms, areas, zones, worlds, multiverses, etc.) and a way to move between them (Portals).

6. 13 World Database

The storage of information on objects and the world is a major part of the design of a VR system. The primary things that are stored in the World Database (or World Description Files) are the objects that inhabit the world, scripts that describe actions of those objects or the user (things that happen to the user), lighting, program controls, and hardware device support.

There are a number of different ways the world information may be stored: a single file, a collection of files, or a database. The multiple file method is one of the more common approaches for VR development packages. Each object has one or more files (geometry, scripts, etc.) and there is some overall 'world' file that causes the other files to be loaded. Some systems also include a configuration file that defines the hardware interface connections.

Sometimes the entire database is loaded during program startup, other systems only read the currently needed files. A real database system helps tremendously with the latter approach. An Object Oriented Database would be a great fit for a VR system.

The data files are most often stored as ASCII (human readable) text files. However, in many systems these are replaced by binary computer files. Some systems have all the world information compiled directly into the application.

Objects in the virtual world can have geometry, hierarchy, scripts, and other attributes. The capabilities of objects have a tremendous impact on the structure and design of the system.

6.14 Control Panels

A VR system often needs to have some sort of control panels available to the user. The world database may contain information on these panels and how they are integrated into the application. Alternatively, they may be a part of the program code.

There are several ways to create these panels. There could be 2D menus that surround a WoW display, or are overlaid onto the image. An alternative is to place control devices inside the virtual world. The simulation system must then note user interaction with these devices as providing control over the world.

One primary area of user control is control of the viewpoint (moving around within the virtual world). Some systems use the joystick or similar device to move. Others use gestures from a glove, such as pointing, to indicate a motion command.

NOTES

6.15 Types of VR Systems

A major distinction of VR systems is the mode with which they interface to the user. This section describes some of the common modes used in VR systems.

6.15.1 Window on World Systems (WoW)

Some systems use a conventional computer monitor to display the visual world. This sometimes called Desktop VR or a Window on a World (WoW). This concept traces its lineage back through the entire history of computer graphics.

6.15.2 Video Mapping

A variation of the WoW approach merges a video input with a 2D computer graphic. The user watches a monitor that shows his body's interaction with the world.

6.15.3 Immersive Systems

The ultimate VR systems completely immerse the user's personal viewpoint inside the virtual world. These "immersive" VR systems are often equipped with a Head Mounted Display (HMD). This is a helmet or a face mask that holds the visual and auditory displays. The helmet may be free ranging, tethered, or it might be attached to some sort of a boom armature.

A nice variation of the immersive systems use multiple large projection displays to create a 'Cave' or room in which the viewer(s) stand. An early implementation was called "The Closet Cathedral" for the ability to create the impression of an immense environment within a small physical space.

6.15.4 Telepresence

Telepresence is a variation on visualizing complete computer generated worlds. This a technology links remote sensors in the real world with the senses of a human operator. The remote sensors might be located on a robot, or they might be on the ends of WALDO like tools. Fire fighters use remotely operated vehicles to handle some dangerous conditions. Surgeons are using very small instruments on cables to do surgery without cutting a major hole in their patients. The instruments have a small video camera at the business end. Robots equipped with telepresence systems have already changed the way deep sea and volcanic exploration is done. NASA plans to use telerobotics for space exploration. There is currently a joint US/Russian project researching telepresence for space rover exploration.

6.15.5 Mixed Reality

Merging the Telepresence and Virtual Reality systems gives the Mixed Reality or Seamless Simulation systems. Here the computer generated inputs are merged with telepresence inputs and/or the users view of the real world. A surgeon's view of a brain surgery is overlaid with images from earlier CAT scans and real-time ultra-sound. A fighter pilot sees computer generated maps and data displays inside his fancy helmet visor or on cockpit displays.

6.16 VR Challenges

Virtual reality has been heavily criticized for being an inefficient method for navigating non-geographical information. At present, the idea of ubiquitous computing is very popular in user interface design, and this may be seen as a reaction against VR and its problems. In reality, these two kinds of interfaces have totally different goals and are complementary. The goal of ubiquitous computing is to bring the computer into the user's world, rather than force the user to go inside the computer. The current trend in VR is actually to merge the two user interfaces to create a fully immersive and integrated experience.Another obstacle is the headaches due to eye strain, caused by VR headsets. RSI can also result from repeated use of the handset gloves.

SUMMARY

- Virtual reality (VR) refers to computer simulations of real-world "environments" that use 3-D graphics and external devices like a dataglove or helmet to allow users to interact with the simulation.
- Virtual Reality is a way for humans to visualize, manipulate and interact with computers and extremely complex data.
- Simulated reality(SR) is the idea that reality could be simulated — often computer-simulated — to a degree indistinguishable from 'true' reality. It could contain conscious minds which may or may not know that they are living inside a simulation.
- A computer simulation would be limited to the processing power of its host computer.
- One of the most time consuming tasks in a VR system is the generation of the images. Fast computer graphics opens a very large range of applications aside from VR, so there has been a market demand for hardware acceleration for a long while.
- One hardware device closely associated with VR is the Head Mounted Device (HMD). These use some sort of helmet or goggles to place small video displays in front of each eye, with special optics to focus and stretch the perceived field of view. Most HMDs use two displays and can provide stereoscopic imaging. Others use a single larger display to provide higher resolution, but without the stereoscopic vision.
- There are two major categories for the available VR software: toolkits and authoring systems.
- Multiverse is a freeware UNIX based client/server system written by Robert Grant. It is a multi-user, non-immersive, X-Windows based Virtual Reality system, primarily focused on entertainment/research.
- The virtual world itself needs to be defined in a 'world space'.
- The storage of information on objects and the world is a major part of the design of a VR system. The primary things that are stored in the World Database (or World Description Files) are the objects that inhabit the world, scripts that describe actions of those objects or the user (things that happen to the user), lighting, program controls, and hardware device support.

➤ Some systems use a conventional computer monitor to display the visual world. This sometimes called Desktop VR or a Window on a World (WoW).

EXERCISE

NOTES

1. Describe the historical background of Virtual Reality.
2. What is virtual reality ? What are different impacts of virtual reality?
3. What is simulated reality ? Compare and contrast the simulated reality and virtual reality.
4. Describe different types of simulation .What is recursive simulation and where it is beneficial?
5. Explain the types of hardwares used in virtual reality applications with proper Example.
6. What are interfacing devices used in virtual reality applications?
7. What are different Levels of VR Hardware Systems ?
8. Describe in brief the Virtual Reality Software Systems. Also describe its types.
9. What are aspects of Virtual Reality program?
10. Define the following VR terms:
    - World space
    - World coordinates
    - World database
    - Control Panel
11. What are different types of Virtual Reality systems ?
12. What are the major challenges that the present VR systems are facing?
13. Describe in detail the major applications of virtual reality .

NOTES

7 Artificial Intelligence

Chapter Includes :

7.1 Introduction

There has been much recent success for Artificial Intelligence(AI) systems undertaking creative tasks in scientific domains such as astronomy, biology, medicine, chemistry, physics and mathematics. In many scientific domains, we can build on the wealth of philosophical and computational studies into creative aspects of human intelligence, and use the abstract nature of the data to derive specialist algorithms for discovery. To achieve high level scientific creativity, the computational techniques employed are often domain specific.

NOTES

We must create a vision for the future which is exciting and challenging .Fortunately for us, any significant demonstration of intelligent systems is exciting. But we must go one step further. Whenever possible, we must identify and work on problems of relevance to the nation—bold national initiatives that capture the imagination of the public.

Like any other science, AI has the potential to do a lot of good and some bad. Sharing the knowledge and know how in the form of information products is surely the only way to reduce this ever-widening gap between the have and have-nots. The current technological revolution provides a new hope and new understanding. The computer and communication technologies will make it possible for a rapid and inexpensive sharing of knowledge.

Success in AI depends on advances in all of computer science. Success in AI requires advances in all of its disparate parts including cognitive science, logic, and connectionism.

7.2 Concept of AI

7.2.1 Cognitive Science

Cognitive Science is the interdisciplinary study of the cognitive processes underlying the acquisition and use of knowledge. It draws from converging evidence and methodology of diverse fields, including psychology, neuroscience, philosophy, computer science, anthropology and linguistics. The term *cognitive science* was coined by Christopher Longuet-Higgins in 1973.

The term "cognitive" in "cognitive science" is "used for any kind of mental operation or structure that can be studied in precise terms. This conceptualization is very broad, and should not be confused with how "cognitive" is used in some traditions of analytic philosophy, where "cognitive" has to do only with formal rules and truth conditional semantics.

Cognitive science is a large field, and covers a wide array of topics on cognition. However, it should be recognized that cognitive science is not equally concerned with every topic that might bear on the nature and operation of the mind or intelligence. Social and cultural factors, emotion, consciousness, animal cognition, comparative and evolutionary approaches are frequently de-emphasized or excluded outright, often based on key philosophical conflicts. Some within the cognitive science community, however, consider these to be vital topics, and advocate the importance of investigating them.

In any way, the essential questions of cognitive science seem to be: *What is intelligence?* and *How is it possible to model it computationally?*. These questions will be answered later in this chapter .

7.2.1 Approaches

There are several approaches to the study of cognitive science. These approaches may be classified broadly as symbolic, connectionist, and dynamic systems.

- Symbolic - holds that cognition can be explained using operations on symbols, by means of explicit computational theories and models of mental (but not brain) processes analogous to the workings of a digital computer.

- Connectionist (subsymbolic) - holds that cognition can only be modeled and explained by using artificial neural networks on the level of physical brain properties.
- Dynamic Systems - holds that cognition can be explained by means of a continuous dynamical system in which all the elements are interrelated.

7.2.2 Levels of analysis

One of the central principles of systemics applied in the symbolic approach to cognitive science is that:

(1) There are different Levels of Analysis (LOA) from which the brain and mind can be studied, and

(2) Mental phenomena are best studied from *multiple* levels of abstraction. For example, these levels are broken into three groups:

- Computational (Behavioral) level: describes the directly observable output (or behavior) of a system.
- Algorithmic (Functional) level: describes how information is processed to produce the behavioral output.
- Implementational (Physical) level: describes the physical substrate that the system consists of (e.g. the brain; neurons).

A simple analogy often used to describe LOA is to compare the brain to a computer. The physical level would consist of the computer's hardware, the behavioral level represents the computer's software, and the functional level would be the computer's operating system, which allows the software and hardware components to communicate.

A central tenet of cognitive science is that a complete understanding of the mind/brain cannot be attained by studying only a single level. For example, consider the problem of remembering a phone number and recalling it later. How does this process occur? One approach would be to study behavior through direct observation. A person could be presented with a phone number, asked to recall it after some delay. Then the accuracy of the response could be measured. Another approach would be to study the firings of individual neurons while a person is trying to remember the phone number. Neither of these experiments on their own would fully explain *how* the process of remembering a phone number works. Even if the technology to map out every neuron in the brain in real-time were available, and it was known when each neuron was firing, it would still be impossible to know how a particular firing of neurons translates into the observed behavior. Thus an understanding of how these two levels relate to each other is needed. This can be provided by a functional level account of the process. Studying a particular phenomenon from multiple levels creates a better understanding of the processes that occur in the brain to give rise to a particular behavior.

7.2.3 Interdisciplinary Nature

Cognitive science is an interdisciplinary field with contributors from various fields, including psychology, neuroscience, linguistics, philosophy, computer science, anthropology, biology, and physics. Cognitive science tends to view the world outside

NOTES

the mind much as other sciences do. Thus it too has an objective, observer-independent existence. The field is usually seen as compatible with the physical sciences, and uses the scientific method as well as simulation or modeling, often comparing the output of models with aspects of human behavior. Still, there is much disagreement about the exact relationship between cognitive science and other fields, and the interdisciplinary nature of cognitive science is largely both unrealized and circumscribed.

Many, but not all, who consider themselves cognitive scientists have a functionalist view of the mind—the view that mental states are classified functionally, such that any system that performs the proper function for some mental state is considered to be in that mental state. Thus, according to functionalism about the mind, even non-human systems, such as other animal species, alien life forms, or advanced computers can, in principle, have mental states. This perspective is one of the reasons the term "cognitive science" is not exactly coextensive with neuroscience, psychology, or some combination of the two.

From the external point of view, the largest interdisciplinary context of cognitive science is systemics. It includes the socio-cognitive extension of the cognition models and theories over different social environments social systems, with the emphasis on distributed cognition and intelligence.

7.3 AI Applications

Typical problems to which AI methods are applied :

- Computer vision, Virtual reality and Image processing .
- Diagnosis (artificial intelligence) .
- Game theory and Strategic planning.
- Game artificial intelligence and Computer game bot.
- Natural language processing, Translation and Chatterbots .
- Non-linear control and Robotics.
- Pattern recognition
  - Optical character recognition
  - Handwriting recognition
  - Speech recognition
  - Face recognition
- Artificial Creativity

Other fields in which AI methods are implemented :

- **Robotics**

  The main aspect of robotics today is mobility. For example how can a mechanical device are controlled to move its body parts in a planned fashion, or navigate around a room? This can be done by learning the task in a virtual simulation, and then applying it to the real robot. If specific conditions of training are respected, the problem has a high probability of working in real life, but this is no guarantee.

In practice when moving robotic arms, the arm has a few movement possibilities: the shoulder allows rotations according to two axis, and the elbow also allow two basic rotations. Each of these possibilities is called one degree of freedom. Usually, one controller is assigned to provide movement for one DOF. The task at hand is to learn the optimal combination of controllers, where they can successfully cooperate to perform a given task.

Most of Artificial Intelligence will eventually lead to robotics, the field of computer science and engineering concerned with creating robots, devices that can move and react to sensory input. Most neural networking, natural language processing, image recognition, speech recognition/synthesis research aims at eventually incorporating their technology into the epitome of robotics - the creation of a fully humanoid robot.

What is a Robot ?

"A robot is an apparently human automaton, intelligent and obedient but impersonal machine"

The word robot comes from robota, Czech for 'forced labour'. Many robots required human operators, or precise guidance throughout their missions. However, robots have started to employ a degree of Artificial Intelligence in their work recently and they are gradually becoming more and more autonomous.

- o Behavior-based robotics
- o Cognitive robotics
- o Cybernetics
- o Developmental robotics
- o Epigenetic robotics
- o Evolutionary robotics

- Hybrid intelligent system
- Intelligent agent
- Intelligent control
- Litigation
- Artificial life
- Automated reasoning
- Automation
- Biologically-inspired computing
- Colloquis
- Concept mining
- Data mining
- Knowledge representation
- Semantic Web
- E-mail spam filtering

NOTES

7.4 Intelligence

One popular definition comes from the " Encyclopedia Britannica ":

"Ability to adapt effectively to the environment, either by making a change in oneself or by changing the environment or finding a new one".

Intelligence is independent from the hardware , and can be described at an abstract level: cognition is computation (computationalism).

Thinking at the speed of light, manipulating matter at the atomic scale, liberating ourselves from the constraints of body, the networked successor of humanity will become the master of the universe. It will discover new ways to avert its own ultimate extinction. It will recreate lost worlds and resurrect the dead. It will close the gap between imagination and reality. And here we see the great temptation of artificial life: It offers both a *critique of human limitations* and a *promise of future power*. The limits create the desire for power; the promise of power makes the limits seem all the less acceptable.

The dictionary defines intelligence with phrases such as "The capacity to acquire and apply knowledge", and "The faculty of thought and reason."

General intelligence implies an ability to acquire and apply knowledge, and to reason and think, in a variety of domains, not just in a single area like, say, chess or game-playing or languages or mathematics.

Gardner's theory of multiple intelligences argues that human intelligence largely breaks down into a number of specialized-intelligence components (including linguistic, logical-mathematical, musical, bodily-kinesthetic, spatial, interpersonal, intra-personal, naturalist and existential).

Taking a broad view, it is clear that, in fact, human intelligence is not all that general. A huge amount of our intelligence is focused on situations that have occurred in our evolutionary experience: social interaction, vision processing, motion control, and so forth.

Our intelligence is general "in principle", but in order to solve many sorts of problems, we need to resort to cumbersome and slow methods such as mathematics and computer programming. Whereas we are vastly more efficient at solving problems that make use of our in-built specialized neural circuitry for processing vision, sound, language, social interaction data, and so forth.

In principle, a human with poor social intelligence but strong logical-mathematical intelligence could solve a difficult problem regarding social interactions, but might have to do so in a very slow and cumbersome over-intellectual way, whereas an individual with strong innate social intelligence would solve the problem quickly and intuitively.

Taking a somewhat different approach, psychologist Robert Sternberg distinguishes three aspects of intelligence: componential, experiential and contextual .

- **Componential intelligence** refers to the specific skills people have that make them intelligent;
- **Experiential intelligence** refers to the ability of the mind to learn and adapt through experience;

NOTES

- **Contextual intelligence** refers to the ability of the mind to understand and operate within particular contexts, and select and modify contexts.

7.5 Artificial Intelligence

Artificial intelligence (AI) involves the study of cognitive phenomena in machines. One of the practical goals of AI is to implement aspects of human intelligence in computers. Computers are also widely used as a tool with which to study cognitive phenomena. Computational modeling uses simulations to study how human intelligence may be structured.

There is some debate in the field as to whether the mind is best viewed as a huge array of small but individually feeble elements (i.e. neurons), or as a collection of higher-level structures such as symbols, schemas, plans, and rules. The former view uses connectionism to study the mind, whereas the latter emphasizes symbolic computations. One way to view the issue is whether it is possible to accurately simulate a human brain on a computer without accurately simulating the neurons that make up the human brain.

Attention is the selection of important information. The human mind is bombarded with millions of stimuli and it must have a way of deciding which of this information to process. Attention is sometimes seen as a spotlight, meaning one can only shine the light on a particular set of information.

The term **Artificial Intelligence (AI)** was first used by **John McCarthy** who used it to mean "the science and engineering of making intelligent machines". It can also refer to intelligence as exhibited by an artificial (*man-made, non-natural, manufactured*) entity. While AI is the generally accepted term others, including both Computational Intelligence and Synthetic Intelligence have been proposed as potentially being "more accurate." The terms strong and weak AI can be used to narrow the definition for classifying such systems. AI is studied in overlapping fields of computer science, psychology, philosophy, neuroscience, and engineering, dealing with intelligent behavior, learning, and adaptation and usually developed using customized machines or computers.

Research in AI is concerned with producing machines to automate tasks requiring intelligent behavior. Examples include control, planning and scheduling, the ability to answer diagnostic and consumer questions, handwriting, natural language, speech, and facial recognition. As such, the study of AI has also become an engineering discipline, focused on providing solutions to real life problems, knowledge mining, software applications, strategy games like computer chess and other video games. One of the biggest difficulties with AI is that of comprehension. Many devices have been created that can do amazing things, but critics of AI claim that no actual comprehension by the AI machine has taken place.

7.5.1 AI Timeline

The field of artificial intelligence dawned in the 1950s. Since then, there have been many achievements in the history of artificial intelligence; some of the more notable moments include:

NOTES

Year	Development
1950	Alan Turing introduces the Turing test intended to test a machine's capability to participate in human-like conversation.
1951	The first working AI programs were written to run on the Ferranti Mark I machine of the University of Manchester: a checkers-playing program written by Christopher Strachey and a chess-playing program written by Dietrich Prinz.
1956	John McCarthy coined the term "artificial intelligence" as the topic of the Dartmouth Conference.
1958	John McCarthy invented the Lisp programming language.
1965	Joseph Weizenbaum built ELIZA, an interactive program that carries on a dialogue in English language on any topic.
1965	Edward Feigenbaum initiated Dendral, a 10-yr effort to develop software to deduce the molecular structure of organic compounds using scientific instrument data. It was the first expert system.
1965	J.A. Robinson introduced resolution as an inference method in logic.
1966	Machine Intelligence workshop at Edinburgh - the first of an influential annual series organized by Donald Michie and others.
1968	Work on MACSYMA was initiated at MIT by Carl Englaman, William Martin and Joel Moses, it is a large interactive program written in LISP.
1972	The Prolog programming language was developed by Alain Colmerauer.
1973	Edinburgh *Freddy* Assembly Robot: a versatile computer-controlled assembly system.
1974	Ted Shortliffe's PhD dissertation on the MYCIN program (Stanford) demonstrated a very practical rule-based approach to medical diagnoses, even in the presence of uncertainty. While it borrowed from DENDRAL, its own contributions strongly influenced the future of expert system development, especially commercial systems.
1991	AI logistics systems deployed in the first Gulf War save the US more money than spent on all AI research since 1950.
1997	The Deep Blue chess machine (IBM) beats the world chess champion, Garry Kasparov.
1999	Sony introduces the AIBO, an artificially intelligent pet.
2004	DARPA introduces the DARPA Grand Challenge requiring competitors to produce autonomous vehicles for prize money.

7.5.2 AI Mechanisms

AI systems are built around automated inference engines. Based on certain conditions ("if") the system infers certain consequences ("then"). AI applications are generally divided into two types, in terms of consequences: classifiers ("if shiny then diamond") and controllers ("if shiny then pick up"). Controllers do however also classify conditions before inferring actions and therefore classification form a central part of most AI systems.

Classifiers make use of pattern recognition for condition matching. In many cases this does not imply absolute, but rather the closest match. Techniques to achieve this divides roughly into two schools of thought: Conventional AI and Computational intelligence (CI).

Classifiers

Classifiers are functions that can be tuned according to examples, making them very attractive for use in AI. These examples are known as observations or patterns. In supervised learning, each pattern belongs to a certain predefined class. A class can be seen as a decision that has to be made. All the observations combined with their class labels are known as a data set.

When a new observation is received, the observation is classified based on previous experience. A classifier can be trained in various ways, there are mainly statistical and machine learning approaches.

A wide range of classifiers are available, each with its strengths and weaknesses. Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems; this is also referred to as the "no free lunch" theorem. Various empirical tests have been performed to compare classifier performance and to find the characteristics of data that determine classifier performance. Determining a suitable classifier for a given problem is however still more an art than science.

The most widely used classifiers are the neural network, support vector machine, k-nearest neighbor algorithm, Gaussian mixture model, naive Bayes classifier, and decision tree.

Conventional AI

Conventional AI mostly involves methods now classified as machine learning, characterized by formalism and statistical analysis. This is also known as symbolic AI, logical AI, neat AI and Good Old Fashioned Artificial Intelligence (GOFAI). Methods include:

- **Expert systems:** apply reasoning capabilities to reach a conclusion. An expert system can process large amounts of known information and provide conclusions based on them.
- **Case based reasoning:** stores a set of problems and answers in an organized data structure called cases. A Case Based Reasoning system upon being presented with a problem finds a case in its knowledge base that is most closely related to the new problem and presents its solutions as an output with suitable modifications.

NOTES

- **Bayesian networks :** a complex and efficient problem solving technique.
- **Behavior based AI :** a modular method building AI systems by hand.

Computational Intelligence

Computational intelligence involves iterative development or learning (e.g. parameter tuning e.g. in connectionist systems). Learning is based on empirical data and is associated with non-symbolic AI, scruffy AI and soft computing. Methods mainly include:

- **Neural networks:** systems with very strong pattern recognition capabilities.
- **Fuzzy systems:** techniques for reasoning under uncertainty, have been widely used in modern industrial and consumer product control systems.
- **Evolutionary computation:** applies biologically inspired concepts such as *populations*, *mutation* and *survival of the fittest* to generate increasingly better solutions to the problem. These methods most notably divide into evolutionary algorithms (e.g. genetic algorithms) and swarm intelligence (e.g. ant algorithms).

With hybrid intelligent systems attempts are made to combine these two groups. Expert inference rules can be generated through neural network or production rules from statistical learning. It is thought that the human brain uses multiple techniques to both formulate and cross-check results. Thus, systems integration is seen as promising and perhaps necessary for true AI.

Computational Modeling

Computational models require a mathematically and logically formal representation of a problem. Computer models are used in the simulation and experimental verification of different specific and general properties of intelligence. Computational modeling can help us to understand the functional organization of a particular cognitive phenomenon.

- **Symbolic modeling,** is based on the technologies of expert systems or more general Knowledge-Based Systems. They are especially used in information engineering and, recently, in generalized systemics.
- **Subsymbolic modeling** which includes *Connectionist/neural network models*. Connectionism relies on the idea that the mind/brain is composed of simple nodes and that the power of the system comes primarily from the existence and manner of connections between the simple nodes. *Neural nets* are textbook implementations of this approach. Some critics of this approach feel that while these models approach biological reality as a repetition of how the system works, they lack explanative powers as complicated systems of connections with even simple rules are extremely complex and often less interpretable than the system they model.

All the above approaches tend to be generalized to the form of integrated computational models of a synthetic/abstract intelligence, in order to be applied to the explanation and improvement of individual and social/organizational decision-making.

Brain Imaging

Brain imaging involves analyzing activity within the brain while performing various cognitive tasks. This allows us to link behavior and brain function to help understand how information is processed. Different types of imaging techniques vary in their temporal (time-based) and spatial (location-based) resolution. Brain imaging is often used in cognitive neuroscience.

- **Single photon emission computed tomography(SPECT) and Positron emission tomography(PET).** SPECT and PET use radioactive isotopes, which are injected into the subject's bloodstream and taken up by the brain. By observing which areas of the brain take up the radioactive isotope, we can see which areas of the brain are more active than other areas. PET has similar spatial resolution to fMRI, but it has extremely poor temporal resolution.
- **Electroencephalography(EEG)** measures the electrical fields generated by large populations of neurons in the cortex by placing a series of electrodes on the scalp of the subject. This technique has an extremely high temporal resolution, but a relatively poor spatial resolution.
- **Functional magnetic resonance imaging(fMRI)** measures the relative amount of oxygenated blood flowing to different parts of the brain. More oxygenated blood in a particular region is assumed to correlate with an increase in neural activity in that part of the brain. This allows us to localize particular functions within different brain regions. fMRI has moderate spatial and temporal resolution.
- **Optical imaging.** This technique uses infrared transmitters and receivers to measure the amount of light reflectance by blood near different areas of the brain. Since oxygenated and deoxygenated blood reflects light by different amounts, we can study which areas are more active (i.e., those that have more oxygenated blood). Optical imaging has moderate temporal resolution, but poor spatial resolution. It also has the advantage that it is extremely safe and can be used to study infants' brains.
- **Magnetoencephalography(MEG)** measures magnetic fields resulting from cortical activity. It is similar to EEG, except that it has improved spatial resolution since the magnetic fields it measures are not as blurred or attenuated as the electrical activity measured in EEG . MEG uses SQUID sensors to detect tiny magnetic fields.

7.5.3 Modern AI Methodologies

The Constructionist design methodology (CDM, or 'Constructionist A.I.') is a formal methodology proposed in the year 2004, for use in the development of cognitive robotics, communicative humanoids and broad AI systems. The creation of such systems requires integration of a large number of functionalities that must be carefully coordinated to achieve coherent system behavior. CDM is based on iterative design steps that lead to the creation of a network of named interacting modules, communicating via explicitly typed streams and discrete messages. The OpenAIR message protocol was inspired by the CDM, and has frequently been used to aid in development of intelligent systems using CDM.

NOTES

One of the first projects to use CDM was **Mirage**, an embodied, graphical agent visualized through augmented reality which could communicate with human users and talk about objects present in the user's physical room. Mirage was created by Kristinn R. Thórisson, the creator of CDM, and a number of students at Columbia University in 2004. The methodology is actively being developed at Reykjavik University.

7.5.4 AI Languages and tools

AI research has led to many advances in programming languages including the first list processing language by Allen Newell *et al.*, Lisp dialects, Planner, Actors, the Scientific Community Metaphor, production systems, and rule-based languages.

GOFAI TEST research is often done in programming languages such as Prolog or Lisp. Matlab and Lush (a numerical dialect of Lisp) include many specialist probabilistic libraries for Bayesian systems. AI research often emphasis rapid development and prototyping, using such interpreted languages to empower rapid command-line testing and experimentation. Real-time systems are however likely to require dedicated optimized software.

Many expert systems are organized collections of if-then such statements, called productions. These can include stocastic elements, producing intrinsic variation, or rely on variation produced in response to a dynamic environment.

OpenAIR was created to allow software components that serve their own purpose to communicate with each other in order to produce large scale, overall behavior of an intelligent systems. A simple example would be to have a speech recognition system, and a speech synthesizer communicate with an expert system through OpenAIR messages, to create a system that can hear and answer various questions through spoken dialogue. CORBA is an older but similar architecture that can be used for comparison, but OpenAIR was specifically created for A.I. research, while CORBA is a more general standard.

Psyclone is a software platform, or an AI operating system (AIOS), developed by Communicative Machines Laboratories for use in creating large, multi modal A.I. systems. The system is an implementation of a blackboard system that supports the OpenAIR message protocol. Psyclone is available for free for non-commercial purposes and has therefore often been used by research institutes on low budgets and novice A.I. developers.

The *OOA* is a hybrid architecture that relies on a special inter-agent communication language (ICL) – a logic-based declarative language which is good for expressing high-level, complex tasks and natural language expressions.

The Messaging Open Service Interface Definition (**MOSID**) provides a means of sending, subscribing and receiving messages. **MOSIDs** are programmatic interfaces which comprise a Service Oriented Architecture for designing and building reusable and interoperable software.

7.6 Intelligent Systems

The investigation and construction of intelligent systems is part of the science of artificial intelligence. The science of artificial intelligence tries to build computer programs that have some function of a human or animal brain. The investigation of intelligent systems tries to explain the functioning of a brain. Artificial intelligent systems are computer programs that have the majority and eventually all the functions of a brain.

The expression **intelligent system** is sometimes used for incomplete intelligent systems, for instance for an intelligent house or an expert system. Here we talk about complete intelligent systems. Such a system has senses to gather information from its environment. It can act and has a memory of the results of its actions. It has an objective and by inspecting its memory it can learn from experience, how to better reach its objectives.

For a complete intelligent system all these following capabilities have to be present:

- **Intelligence**

  Many definitions of "intelligence" are used. For practical purposes we use the following: Intelligence is the systems level of performance in reaching its objectives.

- **System**

  A system is part of the universe, with a limited extension in space and time. Stronger or more correlations exist between one part of the system and another, than between this part of the system and parts outside of the system.

- **Sense**

  A sense organ is that part of the intelligent system that can receive information from its environment. Sense organs are needed so that the system can know its environment and act accordingly

- **Objective**

  An objective is a certain situation that some systems try to reach. Normally there are many levels of objectives, there can be a main objective and sub-objectives.

- **Concept**

  A concept is the basic element of thought. It is a physical, material storage of information (in neurons or electronics). All the concepts in memory are interrelated, they form a web, a net.

- **Situation**

  The situation is a series of concepts that the intelligent system uses to express the information extracted from the environment through its senses.

- **Action rules**

  An action rule is the result of an experience or the result of observing its memory. It is the physical storage by the intelligent system of a situation and a good corresponding action.

NOTES

- **Memory**

  The memory is the physical storage of concepts and action rules. This includes the situations that the intelligent system has experienced.

- **Learning**

  Learning is probably the most important capacity of an intelligent system. It learns concepts from concurrent sense information. It learns action rules from its experience. Acting, possibly haphazard acting, in a certain situation, is stored with a value. An action rule has a higher value if it helped to reach an objective. Learning includes creating abstract concepts based on concrete examples. It includes learning composite concepts containing the parts of an object. Learning is also the detection or relationships (patterns) between the situation part of an action rule and the resulting future situation.

Human beings and animals are intelligent systems. At present, research is advancing in building primitive artificial intelligent systems. Intelligent systems are usually characterized by their ability to adapt to dynamic situations in uncertain environments. This is a high bar that at present only living systems can match. For example, a refrigerator is not an intelligent system since it cannot learn.

7.7 Knowledge-based Systems

A **knowledge-based system** is a program for extending and/or querying a knowledge base.

The Computer User High-Tech Dictionary defines **a knowledge-based system** as a computer system that is programmed to imitate human problem-solving by means of artificial intelligence and reference to a database of knowledge on a particular subject.

Knowledge-based systems are systems based on the methods and techniques of Artificial Intelligence. Their core components are the knowledge base and the inference mechanisms.

While for some authors' expert systems, case-based reasoning systems and neural networks are all three particular types of knowledge-based systems, there are other approaches considering that expert systems and neural networks are different and cannot be included in this category. KBS is a frequently used abbreviation for knowledge-based system.

Knowledge based system get their power from the expert knowledge that has been coded into facts, rules and Procedures. The knowledge is saved in a knowledge Base and separated by components.

"In the knowledge is lies the power!"

This message was learned by Farsighted Researcher at Stanford University in 1960 end early 1970.

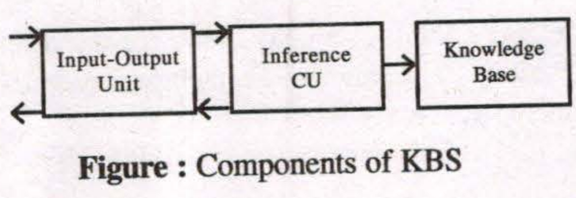


Figure : Components of KBS

NOTES

7.8 Knowledge-based Engineering

Knowledge-based engineering (KBE) is a discipline with roots in computer-aided design (CAD) and knowledge-based systems but has several definitions and roles depending upon the context. An early role was support tool for a design engineer generally within the context of product design. Success of early KBE prototypes was remarkable; eventually this led to KBE being considered as the basis for generative design with many expectations for hands-off performance where there would be limited human involvement in the design process.

KBE can have a wide scope that covers the full range of activities related to Product Lifecycle Management and Multidisciplinary design optimization. KBE's scope would include design, analysis (computer-aided engineering – CAE), manufacturing, and support. In this inclusive role, KBE has to cover a large multi-disciplinary role related to many computer aided technologies (CAx).

KBE also has more general overtones. One of its roles is to bridge knowledge management and design automation. Knowledge processing is a recent advance in computing. It has played a successful role in engineering and is now undergoing modifications (to be explained). An example of KBE's role is generative mechanical design. There are other. KBE can be thought of as an advanced form of computer applications (in some forms with an extreme end-user computing flavor) that support CAx.

7.8.1 KBE and CAX

CAx crosses many disciplinary bounds and provides a sound basis for PLM. In a sense, CAx is a form of applied science that uses most of the disciplines of engineering and their associated fields. Materials science comes to mind.

KBE's support of CAx may have some similarities with its support of PLM but, in a sense, the differences are going to be larger.

The KBE flavor at the CAx level may assume a strong behavioral flavor. Given the underlying object oriented focus, there is a natural use of entities possessing complicated attributes and fulfilling non-trivial roles. One vendor's approach provides a means via workbenches to embed attributes and methods within sub-parts (object) or within a joining of sub-parts into a part.

As an aggregate, the individual actions, that are event driven, can be fairly involved. This fact identifies one major problem, namely control of what is essentially a non-deterministic mixture. This characteristic of the decision problem will get more attention as the KBE systems subsume more levels and encompasses a broader scope of PLM.

7.8.2 KBE and Knowledge Management

KBE is related to knowledge management which has many levels itself. Some approaches to knowledge are reductionistic, as well they ought to be given the pragmatic focus of knowledge modeling. However, due to KBE dealing with aggregates that can be quite complicated both in structure and in behavior, some holistic notions (note link to complexity theory) might be apropos.

NOTES

Also, given all the layers of KBE and given the fact that one part of an associated space is heavily mathematical (namely, manifold in nature), KBE is extremely interesting from the knowledge viewpoint .

All one has to do is note that the KBE process's goal is to produce results in the 'real world' via artifacts and to do so using techniques that are highly computational. That, in essence, is the epitome of applied science/engineering, and it could never be non-interesting.

7.8.3 KBE Methodology

The development of KBE applications concerns the requirements to identify, capture, structure, formalize and finally implement knowledge. Many different so-called KBE platforms support only the implementation step which is not always the main bottleneck in the KBE development process. In order to limit the risk associated with the development and maintenance of KBE application there is a need to rely on an appropriate methodology for managing the knowledge and maintaining it up to date. As example of such KBE methodology the EU project MOKA "Methodology and tools Oriented to Knowledge based Applications" propose solutions which focus on the structuration and formalization steps as well as links to the implementation.

7.8.4 Languages for KBE

Some questions can be asked in regard to KBE implementation: can we represent knowledge in a vendor-neutral format? can the knowledge in our designs be retained for decades, long after a vendor system has disappeared? These questions are addressed in a 2005 Aerospace COE presentation "A Proposal for CATIA V6" by Walter Wilson of Lockheed Martin.

Mr. Wilson advocates using a type of programming language to define design data — operations, parameters, formulas, etc. — instead of a proprietary file format (such as Dassault's CATIA). One's data would no longer be tied to a specific CAD system. Unlike STEP, which inevitably lags commercial CAD systems in the features it supports, programmability would allow the definition of new design features.

A logic programming language is proposed as the basis for the engineering design language because of its simplicity and extensibility. The geometric engine for the language features would be open source to give engineers control over approximation algorithms and to better guarantee long-term accessibility of the data.

7.8.5 KBE futures

KBE is a multi-disciplinary framework that has more than practical considerations. Not only will KBE require successful handling of issues of the computational (Ontology, Artificial Intelligence, Entscheidungsproblem, Interactive computation, Category Theory, ...) and logic (non-monotonic issues related to the qualification, frame, and ramification problems)), it will touch upon all sciences that deal with matter, its manipulations, and the related decisions. What better framework is there to explore the "increasingly complicated mappings between the human world and the computational"?

NOTES

In terms of methodology and their associated means, KBE offers support via several paradigms. These range from the home-grown all the way to strategically defined and integrated tools that cover both breadth and depth. A continuing theme will be resolving the contextual definitions for KBE into a coherent discipline (or at least attempting this) and keeping a handle on managing the necessary quantitative comparisons. One issue of importance considers what limits there may be to the computational; this study requires a multi-disciplinary focus and an understanding of the quasi-empirical. Given the knowledge focus of KBE, another issue involves what limits there might be to a computational basis for knowledge and whether these are overcome with the more advanced types of human-machine interface.

SUMMARY

- *Artificial intelligence (AI)* involves the study of cognitive phenomena in machines. One of the practical goals of AI is to implement aspects of human intelligence in computers. Computers are also widely used as a tool with which to study cognitive phenomena. Computational modeling uses simulations to study how human intelligence may be structured.
- Success in AI depends on advances in all of computer science. Success in AI requires advances in all of its disparate parts including cognitive science, logic, and connectionism.
- *Cognitive Science* is the interdisciplinary study of the cognitive processes underlying the acquisition and use of knowledge.
- The term *cognitive science* was coined by Christopher Longuet-Higgins in 1973.
- Cognitive science is an interdisciplinary field with contributors from various fields, including psychology, neuroscience, linguistics, philosophy, computer science, anthropology, biology, and physics.
- The dictionary defines intelligence with phrases such as "The capacity to acquire and apply knowledge", and "The faculty of thought and reason."
- Gardner's theory of multiple intelligences argues that human intelligence largely breaks down into a number of specialized-intelligence components (including linguistic, logical-mathematical, musical, bodily-kinesthetic, spatial, inter-personal, intra-personal, naturalist and existential).
- *Componential intelligence* refers to the specific skills people have that make them intelligent.
- *Experiential intelligence* refers to the ability of the mind to learn and adapt through experience.
- *Contextual intelligence* refers to the ability of the mind to understand and operate within particular contexts, and select and modify contexts.

NOTES

- The term *Artificial Intelligence (AI)* was first used by *John McCarthy* who used it to mean "the science and engineering of making intelligent machines".
- AI systems are built around automated inference engines. Based on certain conditions ("if") the system infers certain consequences ("then"). AI applications are generally divided into two types, in terms of consequences: classifiers ("if shiny then diamond") and controllers ("if shiny then pick up").
- Conventional AI mostly involves methods now classified as machine learning, characterized by formalism and statistical analysis. This is also known as symbolic AI, logical AI, neat AI and Good Old Fashioned Artificial Intelligence *(GOFAI)*.
- *Expert systems*: apply reasoning capabilities to reach a conclusion. An expert system can process large amounts of known information and provide conclusions based on them.
- *Case based reasoning*: stores a set of problems and answers in an organized data structure called cases. A Case Based Reasoning system upon being presented with a problem finds a case in its knowledge base that is most closely related to the new problem and presents its solutions as an output with suitable modifications.
- Computational intelligence involves iterative development or learning (e.g. parameter tuning e.g. in connectionist systems). Learning is based on empirical data and is associated with non-symbolic AI, scruffy AI and soft computing.
- Neural networks: systems with very strong pattern recognition capabilities.
- Fuzzy systems: techniques for reasoning under uncertainty, have been widely used in modern industrial and consumer product control systems.
- Evolutionary computation: applies biologically inspired concepts such as *populations*, *mutation* and *survival of the fittest* to generate increasingly better solutions to the problem. These methods most notably divide into evolutionary algorithms (e.g. genetic algorithms) and swarm intelligence (e.g. ant algorithms).
- Computational models require a mathematically and logically formal representation of a problem. Computer models are used in the simulation and experimental verification of different specific and general properties of intelligence. Computational modeling can help us to understand the functional organization of a particular cognitive phenomenon.
- Brain imaging involves analyzing activity within the brain while performing various cognitive tasks. This allows us to link behavior and brain function to help understand how information is processed.

NOTES

- The Constructionist design methodology (CDM, or 'Constructionist A.I.') is a formal methodology proposed in the year 2004, for use in the development of cognitive robotics, communicative humanoids and broad AI systems.
- *Psyclone* is a software platform, or an AI operating system (AIOS), developed by Communicative Machines Laboratories for use in creating large, multi modal A.I. systems.
- The science of artificial intelligence tries to build computer programs that have some function of a human or animal brain. The investigation of intelligent systems tries to explain the functioning of a brain. Artificial intelligent systems are computer programs that have the majority and eventually all the functions of a brain.
- Learning is probably the most important capacity of an intelligent system. It learns concepts from concurrent sense information. It learns action rules from its experience.
- A *knowledge-based system* is a program for extending and/or querying a knowledge base.
- *Knowledge-based engineering (KBE)* is a discipline with roots in computer-aided design (CAD) and knowledge-based systems but has several definitions and roles depending upon the context.

EXERCISE

1. What is Cognitive Science? What are different approaches to the study of cognitive science?
2. What are levels of analysis in Cognitive science?
3. Explain the Interdisciplinary Nature of Cognitive science with suitable examples.
4. What is intelligence ? Explain its different types .
5. What is artificial intelligence? How is it giving a new direction to the computer science developments?
6. Explain following AI related terms:
   - Conventional AI
   - Case based reasoning
   - Computational Intelligence
   - Evolutionary computation
   - Computational Modeling
7. Differentiate between the Symbolic modeling and Subsymbolic modeling.
8. What is brain imaging ? What are different approaches for brain imaging ?
9. Describe following :

NOTES

- Single photon emission computed tomography(SPECT) and Positron emission tomography(PET)
- Electroencephalography(EEG)
- Functional magnetic resonance imaging(fMRI)
- Magnetoencephalography(MEG)

10. Explain Modern AI Methodologies in brief .
11. What are different AI languages and tool available for AI system

NOTES

8 Expert Systems

Chapter Includes :

8.1 Introduction

Expert systems apply reasoning capabilities to reach a conclusion. An expert system can process large amounts of known information and provide conclusions based on them.

Expert systems are most valuable to organizations that have a high-level of know-how experience and expertise that cannot be easily transferred to other members. They are designed to carry the intelligence and information found in the intellect of experts and provide this knowledge to other members of the organization for problem-solving purposes.

Typically, the problems to be solved are of the sort that would normally be tackled by a medical or other professional. Real experts in the problem domain (which will typically be very narrow, for instance "diagnosing skin conditions in human teenagers") are asked to provide "rules of thumb" on how they evaluate the problems,

NOTES

either explicitly with the aid of experienced systems developers, or sometimes implicitly, by getting such experts to evaluate test cases and using computer programs to examine the test data and (in a strictly limited manner) derive rules from that. Generally expert systems are used for problems for which there is no single "correct" solution which can be encoded in a conventional algorithm — one would not write an expert system to find shortest paths through graphs, or sort data, as there are simply easier ways to do these tasks.

Simple systems use simple true/false logic to evaluate data, but more sophisticated systems are capable of performing at least some evaluation taking into account real-world uncertainties, using such methods as fuzzy logic. Such sophistication is difficult to develop and still highly imperfect.

8.2 Background History

Expert system first emerged from the research laboratories of a few leading U.S. Universities during the 1960s and 1970s. They were developed as specialized problem solvers which emphasized the use of knowledge rather than algorithms and general search methods.

The first Expert System developed in 1960 was DENDRAL at Stanford University. This is capable of determining the structure of chemical compounds and mass spectrometry rules from positive examples. After DENDRAL was completed, the development of MYCIN began at Stanford University. The MYCIN, which diagnosis infection, blood diseases and determines a list of therapies for patient. The MYCIN involve many projects directly such as - THEIRESIUS, GUIDON and EMYCIN (Essential MYCIN) EMYCIN was used to build other diagstic systems including PUFF.

Researchers then took the opposite approach in the development of the DENDRAL. They applied the knowledge of analytical chemists to infer the molecular structure from the information provided by a mass spectrometer. DENDRAL holds a significant place in the history of expert/knowledge-based systems because it was the first system to use the expertise of human problem solvers and translate that knowledge into a large numbers of special purpose rules, known as a rule -based system.

The work on DENDRAL leads to many others successful applications of this new technology known as expert systems. Feigenbaum and others at Stanford began the Heuristic Programming Project (HPP) to investigate other problem domains that could benefit from this new technology. The next major effort was in the area of edical diagnosis. Bruce Buchanan and Dr. Edward Shortliffe developed MYCIN to diagnose blood infections. Using about 450 rules, MYCIN was able to perform as well as some experts, and considerably better than some junior doctors were.

MYCIN is one of the most widely known of all expert system applications developed. And this despite the fact that it has never been put into practice. However, MYCIN is significant to the history of expert/knowledge-based systems for two particular reasons. First, unlike DENDRAL, which used a model of a particular molecule as the basis for its reasoning, MYCIN was constructed from interviews with various doctors in the particular domain. Therefore, MYCIN contains a number of heuristic rules that are used by physicians in the identification of certain infections. The second major contribution of MYCIN was the later development of EMYCIN (Empty MYCIN). EMYCIN was the first expert/knowledge-based system shell. It took approximately 20 man-years to develop the MYCIN. The researchers realized that if expert systems were to become a viable problem solving technique this development, time must be cut. In an effort to reduce the time to develop an expert system the researchers developed EMYCIN by taking all the rules out of the system and leaving just an empty "shell" in which other developers in other domains could then just "plug-in" their new knowledge base. We discuss expert/knowledge-based systems shells in the development section.

There were other significant expert system applications that were also developed in the early days of expert systems. These systems include PUFF, which used EMYCIN in the domain of pulmonary disorders, DELTA/CATS, which was developed at General Electric Company to assist railroad personnel in the maintenance of GE's diesel-electric locomotives (10). Also at this time, researchers at CMU developed the first truly successful commercial application of expert systems. The system, developed for Digital Equipment Corporation (DEC), was used for computer configuration and known as XCON (R1). XCON, originally titled R1, was developed by John McDermott at CMU for aiding in the configuration of VAX and PDP-11 computer systems at DEC. There exist an enormous number of configurations for VAX and PDP-11 computer system—DEC attempts to configure each system to meet specific customer needs. XCON was originally developed as a 500-rule prototype that examined the specific needs of the customer and decided the exact configuration of components necessary to meet the customer requirements. In particular, XCON's function was to select and arrange the components of a computer systems including: the CPU, the memory, the terminals, the tape and disk drives, and any other peripherals attached to the system.

The development effort began in 1978 and by September 1979, XCON was able to configure more than 75 percent of all customer orders that it was given. By 1981, XCON was being used by DEC on a regular basis and DEC estimates that its cost savings in 1983, 1984, and 1985 were a combined $83 million. Today, XCON is still being used by DEC to configure all VAX orders. There is a development team dedicated to keeping the rules in XCON current and keeping the users of XCON trained on the latest updates. A new copy of XCON is released practically every 3 months and the latest version handles nearly 12,000 different computer components that could possibly configured into a customer order. XCON is one of the major, early success stories in the field of expert systems, for its high visibility domain, its continued use and expansion, and its tremendous impact on the bottom line (profit) at DEC.

Both MYCIN and XCON point out two different functions that are viewed as highly favorable for expert/knowledge-based system development. MYCIN mainly deals with the diagnosis of a disease has given a set of symptoms and patient information. XCON, on the other hand, is a synthesis-based (design) configuration expert system. It takes as its input the needs of the customer and builds a feasible arrangement of components to meet the need. Both of these systems solve different generic "types" of problems.

8.2.1 Expert Systems Shells

Knowledge-based expert systems, or simply expert systems, use human knowledge to solve problems that normally would require human intelligence. These expert systems represent the expertise knowledge as data or rules within the computer. These rules and data can be called upon when needed to solve problems. Books and manuals have a tremendous amount of knowledge but a human has to read and interpret the knowledge for it to be used. Conventional computer programs perform tasks using conventional decision-making logic -- containing little knowledge other

NOTES

than the basic algorithm for solving that specific problem and the necessary boundary conditions. This program knowledge is often embedded as part of the programming code, so that as the knowledge changes, the program has to be changed and then rebuilt. Knowledge-based systems collect the small fragments of human know-how into a knowledge-base which is used to reason through a problem, using the knowledge that is appropriate. A different problem, within the domain of the knowledge-base, can be solved using the same program without reprogramming. The ability of this system to explain the reasoning process through back-traces and to handle levels of confidence and uncertainty provides an additional feature that conventional programming does't handle.

Most expert systems are developed via specialized software tools called shells. These shells come equipped with an inference mechanism (backward chaining, forward chaining, or both), and require knowledge to be entered according to a specified format (all of which might lead some to categorize OPS5 as a shell). They typically come with a number of other features, such as tools for writing hypertext, for constructing friendly user interfaces, for manipulating lists, strings, and objects, and for interfacing with external programs and databases. These shells qualify as languages, although certainly with a narrower range of application than most programming languages. For more detailed information on expert system shells.

8.3 Concept of Expert Systems

Definition : An **expert system(ES)**, also known as a knowledge based system, is a computer program that contains some of the subject-specific knowledge, and contains the knowledge and analytical skills of one or more human experts. This class of program was first developed by researchers in artificial intelligence during the 1960s and 1970s and applied commercially throughout the 1980s. The most common form of expert systems is a program made up of a set of rules that analyze information (usually supplied by the user of the system) about a specific class of problems, as well as providing mathematical analysis of the problem(s), and, *depending upon their design*, recommend a course of user action in order to implement corrections. It is a system that utilizes what appear to be reasoning capabilities to reach conclusions.

A related term is wizard. A wizard is an interactive computer program that helps a user solves a problem. Originally the term wizard was used for programs that construct a database search query based on criteria supplied by the user. However some rule-based expert systems are also called wizards. Other "Wizards" are a sequence of online forms that guide users through a series of choices, such as the ones which manage the installation of new software on computers, and these are not Expert Systems.

An expert system is a computer based information system in which knowledge is represented in data in which the processing of the knowledge is directed, primarily by computer programs .The development of systems that process the knowledge of a true human expert has proved to be difficult and expensive. The term expert system originated because these systems were aimed, initially, at replicating the abilities of true human experts. The goal was to create a system that would be able, for example to diagnose human diseases as well as or better than a human expert. Many of the expert systems, we encounter in business, will not have the abilities of a true human expert.

NOTES

ES employs human knowledge captured in a computer to solve problems that ordinarily require human expertise. These can be used by non-experts to improve their problem solving abilities. ES becomes knowledgeable assistant to human experts. They are used to propagate search knowledge resources or improved consistent results. Such systems could function better than any single expert, in making judgments in a specific, usually narrow area of expertise, termed as "Domain". This possibility may have a significant impact both on advisory professionals (financial analysts etc.) and on organizations and management.

8.3.1 Expert Systems Vs. Problem solving Systems

The principal distinction between expert systems and traditional problem solving programs is the way in which the problem related expertise is coded. In traditional applications, problem expertise is encoded in both program and data structures.

In the expert system approach all of the problem related expertise is encoded in data structures only; none is in programs. This organization has several benefits.

An example may help contrast the traditional problem solving program with the expert system approach. The example is the problem of tax advice. In the traditional approach data structures describe the taxpayer and tax tables, and a program in which there are statements representing an expert tax consultant's knowledge, such as statements which relate information about the taxpayer to tax table choices. It is this representation of the tax expert's knowledge that is difficult for the tax expert to understand or modify.

In the expert system approach, the information about taxpayers and tax computations is again found in data structures, but now the knowledge describing the relationships between them is encoded in data structures as well. The programs of an expert system are independent of the problem domain (taxes) and serve to process the data structures without regard to the nature of the problem area they describe. For example, there are programs to acquire the described data values through user interaction, programs to represent and process special organizations of description, and programs to process the declarations that represent semantic relationships within the problem domain and an algorithm to control the processing sequence and focus.

The general architecture of an expert system involves two principal components: a problem dependent set of data declarations called the *knowledge base* or *rule base*, and a problem independent (although highly data structure dependent) program which is called the *inference engine*.

8.3.2 People involved in Expert Systems

There are generally three individuals having an interaction with expert systems. Primary among these is the end-user; the individual who uses the system for its problem solving assistance. In the building and maintenance of the system there are two other roles: the problem domain expert who builds and supplies the knowledge base providing the domain expertise, and a knowledge engineer who assists the experts in determining the representation of their knowledge, enters this knowledge into an explanation module and who defines the inference technique required to obtain useful problem solving activity. Usually, the knowledge engineer will represent the problem solving activity in the form of rules which is referred to as a rule-based expert system. When these rules are created from the domain expertise, the knowledge base stores the rules of the expert system.

NOTES

8.3.2.1 The End User

The end-user usually sees an expert system through an interactive dialog. In expert systems, dialogs are not pre-planned. There is no fixed control structure. Dialogs are synthesized from the current information and the contents of the knowledge base. Because of this, not being able to supply the answer to particular questions does not stop the consultation.

It is very difficult to implement a general explanation system (answering questions like Why and How) in traditional systems. The response of the expert system to the question WHY is an exposure of the underlying knowledge structure. It is a rule; a set of antecedent conditions which, if true, allow the assertion of a consequent. The rule references values, and tests them against various constraints or asserts constraints onto them. This, in fact, is a significant part of the knowledge structure. There are values, which may be associated with some organizing entity. For example, the individual diner is an entity with various attributes (values) including whether they drink soup and the kind of soup. There are also rules, which associate the currently known values of some attributes with assertions that can be made about other attributes. It is the orderly processing of these rules that dictates the dialog itself.

8.3.2.2 The Knowledge Engineer

Knowledge engineers are concerned with the representation chosen for the expert's knowledge declarations and with the inference engine used to process that knowledge. He / she can use the knowledge acquisition component of the expert system to input the several characteristics known to be appropriate to a good inference technique including:

- A good inference technique is independent of the problem domain. In order to realize the benefits of explanation, knowledge transparency, and reusability of the programs in a new problem domain, the inference engine must not contain domain specific expertise.
- Inference techniques may be specific to a particular task, such as diagnosis of hardware configuration. Other techniques may be committed only to a particular processing technique.
- Inference techniques are always specific to the knowledge structures.
- Successful examples of rule processing techniques include: (a) Forward chaining (b) Backward chaining

8.4 Features of Expert System

Following characteristic involve in Expert System.

1. Expert System use knowledge to control the solution process.
2. The knowledge is maintained as an entity separate from the control program.
3. Expert systems are capable of explaining how a particular conclusion was reached, and why requested information is needed during a consultation. This is important as it gives the user a chance to access and understand the system's reasoning ability, thereby improving the user's confidence in the system.
4. Expert systems use symbolic representations for knowledge (rules, networks, or frames).

5. Expert systems often reason with metaknowledge; that is, they reason with knowledge about themselves, and their own knowledge limits and capabilities.

8.5 Building of the Expert System

NOTES

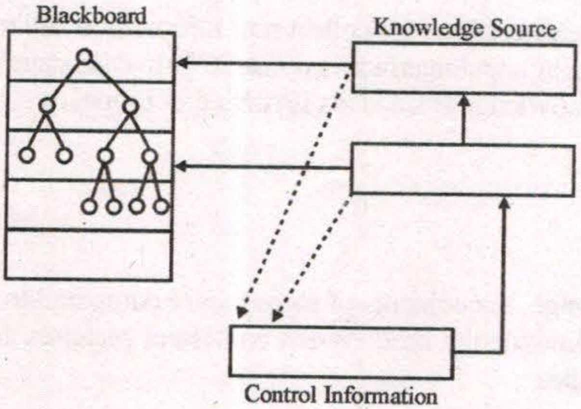
The components of Expert System can be categorised as -

1. **Blackboard System –** It refers to a special type of KBS which uses a form of opportunistic reasoning. The Blackboard system are composed of three components -

   (a) They have number of knowledge sources which are separate set of code. Every knowledge source is specialist in limited area to solve a problem. This source contains knowledge in the form of rules, schemes and procedures.

   (b) A Blackboard contains the current information needed by the knowledge sources.

   (c) The third component of Blackboard system is control information. The Control knowledge monitors the changes to the blackboard and determines what the immediate focus of attention should be in solving the problems.



2. **Knowledge Acquisition –** This program is used by someone who has expertise in solving the problem to create add to or change the knowledge Base. Knowledge for expert systems must be derived from expert sources like journal articles, texts, reports, and information of database. Experts do not know what specific knowledge is being applied or just how it is applied in the solution of a given problem.

   An experienced knowledge engineer Penny Nii, at Stanford University, has described some useful practices to follow in solving acquisition problems.

3. **Knowledge-based Systems :** It contains rules, facts, descriptions of objects etc. The information in the knowledge base is everything that is necessary for understanding and formulating the problem and then solving it.

   It includes two basic elements :

   (i) Facts (for example, the problem situation and the problem area).

   (ii). Heuristics (rules that direct the use of knowledge to solve problems in a particular domain).

   A knowledge-based system is a program for extending and/or querying a knowledge base.

   The Computer User High-Tech Dictionary defines a knowledge-based system as a computer system that is programmed to imitate human problem-solving by

NOTES

means of artificial human problem-solving by means of artificial intelligence and reference to a database of knowledge on a particular subject.

Knowledge-based systems are systems based on the methods and techniques of Artificial Intelligence. Their core components are the knowledge base and the inference mechanisms.

While for some authors' expert systems, case-based reasoning systems and neural networks are all three particular types of knowledge-based systems, there are other approaches considering that expert systems and neural networks are different and cannot be included in this category. KBS is a frequently used abbreviation for knowledge-based system.

Knowledge based system get their power from the expert knowledge that has been coded into facts, rules and Procedures. The knowledge is saved in a knowledge Base and separated by components.

"In the knowledge is lies the power !"

This message was learned by Farsighted Researcher at Stanford University in 1960 and early 1970.

4. **The Inference engine :** It's collect user information and responses to question through Input/Output interface in dynamic form with static knowledge which is stored in knowledge Base. This is carried out repeatively in three stages -

(a) Match

(b) Select

(c) Run

In match stage, the contents of memory are compared to facts and rules contained in Knowledge base. When consistent matches are found, the corresponding rules

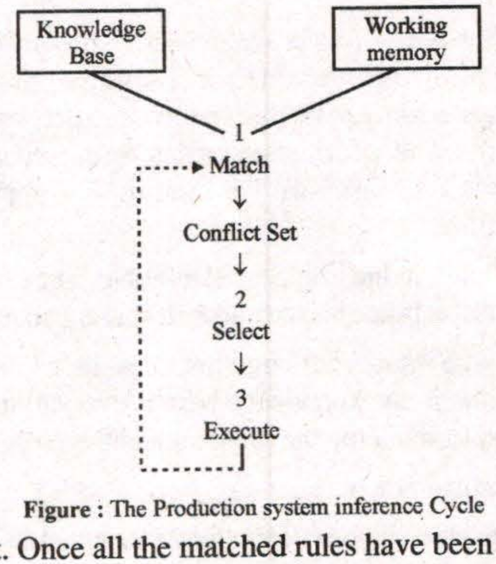


Figure : The Production system inference Cycle

are store in a set. Once all the matched rules have been added to the conflict set during a given cycle, one of the rules is selected for execution. The criteria for selection may be most recent use, rule condition specificity, (the number of conjuncts on the left), or simply the smallest rule number. The selected rule is then executed and the right hand side or part of the rule is then carried out. In figure illustrates this match - select - execute cycle. As an example, suppose the working memory contains the two clauses :

NOTES

(father bob sam)

(mother sue sam)

5. **The User Interface :** ES contains a language processor friendly problem oriented communications between the manager-user and the computer. The user Interface gives the user to communicate with the system in a normal way as simple selection menus or language which is related to natural language. This means that the system must have special prompts or a specialized vocabulary which encompasses the terminology of the given domain of expertise. For example MYCIN can recognize many medical terms in addition to many common words needed to communicate. For this, MYCIN has a vocabulary of some 2000 words.

   Personal Consultant Plus, a commercial PC version of the MYCIN architecture, uses menus and english prompts to communicate with the user. The prompts, written is standard english, are provided by the developer during the system building stage. How and why explanations are also given in natural language form.

   The learning module and history file are not common components of expert systems. When they are provided, they are used to assist in building and refining the knowledge base.

8.6 Advantages and Disadvantages of ES

Advantages

- Provide consistent answers for repetitive decisions, processes and tasks.
- Hold and maintain significant levels of information.
- Reduces creating entry barriers to competitors.
- Review transactions that human experts may overlook .

Disadvantages

- The lack of human common sense needed in some decision makings .
- The creative responses human experts can respond to in unusual circumstances.
- Domain experts not always being able to explain their logic and reasoning.
- The challenges of automating complex process.
- The lack of flexibility and ability to adapt to changing environments as questions are standard and cannot be changed .
- Not being able to recognize when no answer is available .

8.7 ES Applications

Expert systems are designed and created to facilitate tasks in the fields of accounting, medicine, process control, financial service, production, human resources etc. Indeed, the foundation of a successful expert system depends on a series of technical procedures and development that may be designed by certain technicians and related experts. When a corporation begins to develop and implement an expert system project, it will use selfsourcing, insourcing and/or outsourcing techniques.

While expert systems have distinguished themselves in AI research in finding practical application, their application has been limited. Expert systems are notoriously narrow in their domain of knowledge—as an amusing example, a researcher used the "skin disease" expert system to diagnose his rustbucket car as likely to have developed measles—and the systems were thus prone to making errors that humans would easily spot. Additionally, once some of the mystique had worn off, most programmers realized that simple expert systems were essentially just slightly more elaborate versions of the decision logic they had already been using. Therefore, some of the techniques of expert systems can now be found in most complex programs without any fuss about them.

An example, and a good demonstration of the limitations of, an expert system used by many people is the Microsoft Windows operating system troubleshooting software located in the "help" section in the taskbar menu. Obtaining expert / technical operating system support is often difficult for individuals not closely involved with the development of the operating system. Microsoft has designed their expert system to provide solutions, advice, and suggestions to common errors encountered throughout using the operating systems.

Another 1970s and 1980s application of expert systems — which we today would simply call AI — was in computer games. For example, the computer baseball games Earl Weaver Baseball and Tony La Russa Baseball each had highly detailed simulations of the game strategies of those two baseball managers. When a human played the game against the computer, the computer queried the Earl Weaver or Tony La Russa Expert System for a decision on what strategy to follow. Even those choices where some randomness was part of the natural system (such as when to throw a surprise pitch-out to try to trick a runner trying to steal a base) were decided based on probabilities supplied by Weaver or La Russa. Today we would simply say that "the game's AI provided the opposing manager's strategy."

8.8 Expert System Problem domains

The expert system can be used to address problems in a number of different problem domains. These domains are :

- Procedural
- Diagnosis
- Monitoring
- Configuration and Design
- Scheduling and Planning

It has been found that the type of information that is solicited by the system from the user by means of questions or classes should be tailored to the level of knowledge of the user. In many applications, the group of prospective uses is nicely defined and the knowledge level can be estimated so that the questions can be presented at a level which corresponds generally to the average user. However, in other applications, knowledge of the specific domain of the expert system might vary considerably among the group of prospective users.

8.9 Benefits and Limitations of Expert Systems

Benefits of Expert Systems

- Working with incomplete and uncertain information
- Capturing expertise
- Cost reduction

NOTES

- Increased output
- Improved quality
- Reduced down time
- Equipment operation
- Operation in hazardous environments
- Use of less expensive equipment
- Reliability
- Response time is faster
- Enhancing problem solving
- Solving complex problem in a narrow domains.

Limitations of Expert Systems :

- ES works well only in a narrow domain.
- Knowledge is not always readily available.
- Expertise is hard to extract from humans.
- The approach of each expert to situation assessment may be different, yet correct.
- Most experts have no independent means of checking.
- The vocabulary that experts use for expressing facts and relations is frequently limited.
- Help is frequently required from knowledge engineers who are rare and expensive.

8.10 Examples of Expert Systems

- *Dendral* analyse mass spectra .
- *Dipmeter* Advisor analysis of data gathered during oil exploration.
- *Mycin* diagnose infectious blood diseases and recommend antibiotics (by Stanford University)
- *CADUCEUS (expert system)* blood-borne infectious bacteria.
- *R1 (expert system)/XCon* order processing .
- *Jess:* A CLIPS engine implemented in Java.
- *Nexpert* Early general-purpose commercial backwards-chaining product.
- *SHINE Real-Time expert system* - Spacecraft Health INference Engine.

SUMMARY

- Expert systems apply reasoning capabilities to reach a conclusion. An expert system can process large amounts of known information and provide conclusions based on them.
- An *expert system(ES)*, also known as a knowledge based system, is a computer program that contains some of the subject-specific knowledge, and contains the knowledge and analytical skills of one or more human experts.
- An expert system is a computer based information system in which knowledge is represented in data in which the processing of the knowledge is directed, primarily by computer programs .

NOTES

- ES employs human knowledge captured in a computer to solve problems that ordinarily require human expertise. These can be used by non-experts to improve their problem solving abilities.
- In the expert system approach all of the problem related expertise is encoded in data structures only; none is in programs.
- The general architecture of an expert system involves two principal components: a problem dependent set of data declarations called the *knowledge base* or *rule base*, and a problem independent (although highly data structure dependent) program which is called the *inference engine*.
- Knowledge engineers are concerned with the representation chosen for the expert's knowledge declarations and with the inference engine used to process that knowledge.
- The expert system development environment is used by the ES builder to build the components and introduce expert knowledge into the ES knowledge base.
- The consultation environment is used by non-experts to obtain the expert knowledge and advice.
- The Inference Engine : This applies knowledge in a systematic way. It is the brain of the ES. It applies rules and other forms of knowledge in an attempt to make sequences of logical conclusions that will solve a user's problem.
- Expert systems are designed and created to facilitate tasks in the fields of accounting, medicine, process control, financial service, production, human resources etc.

EXERCISE

1. Describe the historical background of Expert systems.
2. What is an Expert System ? How it is different from traditional problem solving systems?
3. What are different categories of people involved in expert systems?
4. Differentiate between end users and knowledge engineers .
5. What are different components/elements of Expert systems? Describe in detail.
6. Describe the expert systems problem domain.
7. Describe benefits and limitations of expert systems.
8. Give some examples of expert systems which are leading in the market.
9. What are different advantages and disadvantages of expert systems?
10. Describe some major applications of expert systems.

NOTES

9 Data Warehousing

Chapter Includes :

9.1 Introduction

Data Warehouses became a distinct type of computer database during the late 1980s and early 1990s. They were developed to meet a growing demand for management information and analysis that could not be met by operational systems. Operational systems were unable to meet this need for a range of reasons:

- The processing load of reporting reduced the response time of the operational systems,
- The database designs of operational systems were not optimized for information analysis and reporting,

NOTES

- Most organizations had more than one operational system, so company-wide reporting could not be supported from a single system, and
- Development of reports in operational systems often required writing specific computer programs which were slow and expensive .

As a result, separate computer databases began to be built that were specifically designed to support management information and analysis purposes. These data warehouses were able to bring in data from a range of different data sources, such as mainframe computers, minicomputers, as well as personal computers and office automation software such as spreadsheet, and integrate this information in a single place. This capability, coupled with user-friendly reporting tools and freedom from operational impacts, has led to a growth of this type of computer system.

As technology improved (lower cost for more performance) and user requirements increased (faster data load cycle times and more features), data warehouses have evolved through several fundamental stages:

- **Offline Operational Databases** : Data warehouses in this initial stage are developed by simply copying the database of an operational system to an off-line server where the processing load of reporting does not impact on the operational system's performance.
- **Offline Data Warehouse** : Data warehouses in this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented data structure
- **Real Time Data Warehouse:** Data warehouses at this stage are updated on a transaction or event basis, every time an operational system performs a transaction (e.g. an order or a delivery or a booking etc.)
- **Integrated Data Warehouse:** Data warehouses at this stage are used to generate activity or transactions that are passed back into the operational systems for use in the daily activity of the organization.

9.2 Data Warehouse

Definition : A data warehouse(DW) is the main repository of an organization's historical data, its corporate memory. It contains the raw material for management's decision support system. The critical factor leading to the use of a data warehouse is that a data analyst can perform complex queries and analysis, such as data mining, on the information without slowing down the operational systems.

By W.H. Inmon "A Datawarehouse is a subject -oriented, integrated, time variant, non-volatile collection of data in support of management decision."

W.H. Inmon, an early and influential practitioner, has formally defined a data warehouse in the following terms . These are also special characteristics of dataware houses:

- **Subject-oriented,** meaning that the data in the database is organized so that all the data elements relating to the same real-world event or object are linked together;

NOTES

- **Time-variant,** meaning that the changes to the data in the database are tracked and recorded so that reports can be produced showing changes over time;
- **Non-volatile,** meaning that data in the database is never over-written or deleted, once committed, the data is static, read-only, but retained for future reporting; and
- **Integrated,** meaning that the database contains data from most or all of an organization's operational applications, and that this data is made consistent.

While operational systems are optimized for simplicity and speed of modification through heavy use of database normalization and an entity-relationship model, the data warehouse is optimized for reporting and analysis (online analytical processing, or OLAP). Frequently data in data warehouses are heavily denormalised, summarized or stored in a dimension-based model. This is not always required to achieve acceptable query response times, however.

9.2.1 DW Issues

The following issues are concerned with data warehouses:

- Extracting, transforming and loading data consumes a lot of time and computational resources.
- Data warehousing project scope must be actively managed to deliver a release of defined content and value.
- Compatibility problems with systems already in place.
- Security could develop into a serious issue, especially if the data warehouse is web accessible.
- Data Storage design controversy warrants careful consideration and perhaps prototyping of the data warehouse solution for each project's environments.

9.2.2 DW Storage

In OLTP - online transaction processing systems relational database design use the discipline of data modeling and generally follow the Codd rules of data normalization in order to ensure absolute data integrity. Less complex information is broken down into its most simple structures (a table) where all of the individual atomic level elements relate to each other and satisfy the normalization rules. Codd defines 5 increasingly stringent rules of normalization and typically OLTP systems achieve a 3rd level normalization. Fully normalized OLTP database designs often result in having information from a business transaction stored in dozens to hundreds of tables. Relational database managers are efficient at managing the relationships between tables and result in very fast insert/update performance because only a little bit of data is affected in each relational transaction.

OLTP databases are efficient because they are typically only dealing with the information around a single transaction. In reporting and analysis, thousands to billions of transactions may need to be reassembled imposing a huge workload on the relational database. Given enough time the software can usually return the requested results, but because of the negative performance impact on the machine and all of its hosted applications, data warehousing professionals recommend that reporting databases be physically separated from the OLTP database.

NOTES

In addition, data warehousing suggests that data be restructured and reformatted to facilitate query and analysis by novice users. OLTP databases are designed to provide good performance by rigidly defined applications built by programmers fluent in the constraints and conventions of the technology. Add in frequent enhancements, and too many a database is just a collection of cryptic names, seemingly unrelated and obscure structures that store data using incomprehensible coding schemes. All factors that while improving performance, complicate use by untrained people. Lastly, the data warehouse needs to support high volumes of data gathered over extended periods of time and are subject to complex queries and need to accommodate formats and definitions of inherited from independently designed package and legacy systems.

Designing the data warehouse data Architecture synergy is the realm of Data Warehouse Architects. The goal of a data warehouse is to bring data together from a variety of existing databases to support management and reporting needs. The generally accepted principle is that data should be stored at its most elemental level because this provides for the most useful and flexible basis for use in reporting and information analysis. However, because of different focus on specific requirements, there can be alternative methods for design and implementing data warehouses. There are two leading approaches to organizing the data in a data warehouse: the dimensional approach advocated by Ralph Kimball and the normalized approach advocated by Bill Inmon. Whilst the dimension approach is very useful in data mart design, it can result in a rats nest of long term data integration and abstraction complications when used in a data warehouse.

In the "dimensional" approach, transaction data is partitioned into either a measured "facts" which are generally numeric data that captures specific values or "dimensions" which contain the reference information that gives each transaction its context. As an example, a sales transaction would be broken up into facts such as the number of products ordered, and the price paid, and dimensions such as date, customer, product, geographical location and salesperson. The main advantages of a dimensional approach are that the data warehouse is easy for business staff with limited information technology experience to understand and use. Also, because the data is pre-joined into the dimensional form, the data warehouse tends to operate very quickly. The main disadvantage of the dimensional approach is that it is quite difficult to add or change later if the company changes the way in which it does business.

The "normalized" approach uses database normalization. In this method, the data in the data warehouse is stored in third normal form. Tables are then grouped together by **subject areas** that reflect the general definition of the data (customer, product, finance, etc.). The main advantage of this approach is that it is quite straightforward to add new information into the database — the primary disadvantage of this approach is that because of the number of tables involved, it can be rather slow to produce information and reports. Furthermore, since the segregation of facts and dimensions is not explicit in this type of data model, it is difficult for users to join the required data elements into meaningful information without a precise understanding of the data structure.

Subject areas are just a method of organizing information and can be defined along any lines. The traditional approach has subjects defined as the subjects or nouns within a problem space. For example, in a financial services business, you might have customers, products and contracts. An alternative approach is to organize around the business transactions, such as customer enrollment, sales and trades.

9.3 Data Warehousing

Data Warehousing is open to an almost limitless range of definitions. Simply put, Data Warehouses store an aggregation of a company's data.

Data Warehouses are an important asset for organizations to maintain efficiency, profitability and competitive advantages. Organizations collect data through many sources - Online, Call Center, Sales Leads, Inventory Management. The data collected have degrees of value and business relevance. As data is collected, it is passed through a 'conveyor belt', call the Data Life Cycle Management.

An organization's data life cycle management's policy will dictate the data warehousing design and methodology.

The data warehouse architecture describes the overall system from various perspectives such as data, process, and infrastructure needed to communicate the structure, function and interrelationships of each component. The infrastructure or technology perspective details the various hardware and software products used to implement the distinct components of the overall system. The data perspective typically diagrams the source and target data structures and aid the user in understanding what data assets are available and how they are related. The process perspective is primarily concerned with communicating the process and flow of data from the originating source system through the process of loading the data warehouse, and often the process that client products use to access and extract data from the warehouse.

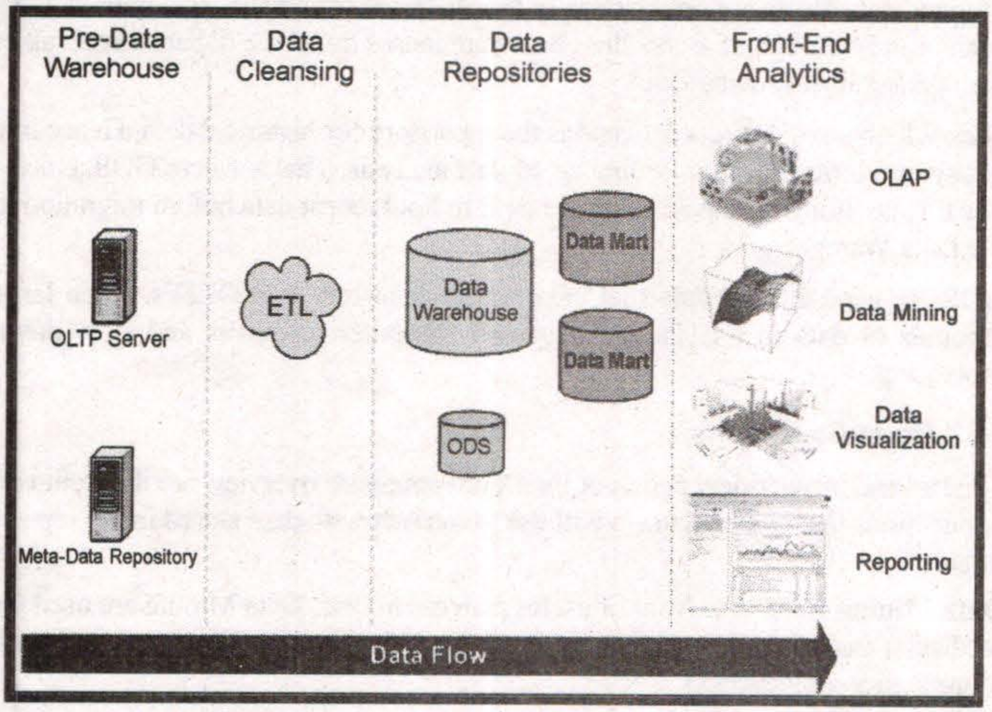


Figure 9.1 : Overview of Data Warehousing Infrastructure

The goal of Data Warehousing is to generate front-end analytics that will support business executives and operational managers.

9.3.1 Pre-Data Warehouse

The pre-Data Warehouse zone provides the data for data warehousing. Data Warehouse designers determine which data contains business value for insertion.

NOTES

OLTP databases are where operational data are stored. OLTP databases can reside in transactional software applications such as Enterprise Resource Management (ERP), Supply Chain, Point of Sale, Customer Serving Software. OLTPs are design for transaction speed and accuracy.

Metadata ensures the sanctity and accuracy of data entering into the data lifecycle process. Meta-data ensures that data has the right format and relevancy. Organizations can take preventive action in reducing cost for the **Extract, transform, and load** (ETL) stage by having a sound Metadata policy. The commonly used terminology to describe meta data is "data about data".

9.3.2 Data Cleansing

Before data enters the data warehouse, the extraction, transformation and cleaning (ETL) process ensures that the data passes the data quality threshold.

ETLs are also responsible for running scheduled tasks that extract data from OLTPs.

9.3.3 Data Repositories

The Data Warehouse repository is the database that stores active data of business value for an organization. The Data Warehouse modeling design is optimized for data analysis.

There are variants of Data Warehouses - Data Marts and ODS (Operational Data Stores). Data Marts are not physically any different from Data Warehouses. Data Marts can be though of as smaller Data Warehouses built on a departmental rather than on a company-wide level.

Data Warehouses collects data and is the repository for historical data. Hence it is not always efficient for providing up-to-date analysis. This is where ODS, Operational Data Stores, come in. ODS are used to hold recent data before migration to the Data Warehouse.

ODS are used to hold data that have a deeper history that OLTPs. Keep large amounts of data in OLTPs can tie down computer resources and slow down processing.

9.3.4 Front-End Analysis

The last and most critical potion of the Data Warehouse overview are the front-end applications that business users will use to interact with data stored in the repositories.

Data Mining is the discovery of useful patterns in data. Data Mining are used for prediction analysis and classification - e.g. what is the likelihood that a customer will migrate to a competitor.

OLAP, Online Analytical Processing, is used to analyze historical data and slice the business information required. OLAPs are often used by marketing managers.

Reporting tools are used to provide reports on the data. Data are displayed to show relevancy to the business and keep track of key performance indicators (KPI).

Data Visualization tools is used to display data from the data repository. Often data visualization is combined with Data Mining and OLAP tools. Data visualization can allow the user to manipulate data to show relevancy and patterns.

9.4 Some Key Concepts About Data

NOTES

9.4.1 Operational Data Store

According to W.H. Inmon, an **operational data store (ODS)** is a subject-oriented, integrated, volatile, current-valued, detailed-only collection of data in support of an organization's need for up-to-the-second, operational, integrated, collective information.

An **operational data store** (or **"ODS"**) is a database designed to integrate data from multiple sources to facilitate operations, analysis and reporting. Because the data originates from multiple sources, the integration often involves cleaning, redundancy resolution and business rule enforcement. An ODS is usually designed to contain low level or atomic (indivisible) data such as transactions and prices as opposed to aggregated or summarized data such as net contributions. Aggregated data is usually stored in the Data warehouse.

9.4.2 Multidimensional Database

Multidimensional databases are variously (depending on the context) data aggregators which combine data from a multitude of data sources; databases which offer networks, hierarchies, arrays and other data formats difficult to model in SQL; or databases which give a high degree of flexibility in the definition of dimensions, units, and unit relationships, regardless of data format.

Multi-dimensional databases are especially useful in a sales and marketing applications that involve time series. Large volumes of sales and inventory data can be stored to ultimately be used for logistics and executive planning. For example, data can be more readily segregated by sales region, product, or time period. While many of the major database vendors have recognized and implemented at least a partial solution, most frequently they rely upon a Star schema database design. However, The Star database design does not account for "sparse data" - basically, wasted empty space. Database strategies to manage sparse data result in the compression of large blocks of empty data elements and improves the performance of the database as a whole.

The data cube is a conceptual representation of database which can be implemented in a variety of ways, including top-down, bottom-up, and arrays. Multi-dimensional databases for time-series or other data vector analysis is preferable over relational databases. On the other hand, dimensionality becomes problematic when working with greater than four dimensions - there is often a resultant amount of sparse or empty data. Removing empty or sparse data poses risks as it can ruin the context and more specifically the vector coordinates of the data.

This is an active area of database development, in which the set of desired features is somewhat vague, but better-defined than the set of known or proposed solutions. Defining and implementing a database which allows people at each level of an organization to define tables and data formats in the way that is most useful to them, yet which supports a single clear query language and consistent infrastructure, remains an open problem.

9.4.3 Data Life Cycle Management

NOTES

Data becomes active as soon as it is of interest to an organization. Data life cycle begins with a business need for acquiring data. Active data are referenced on a regular basis during day-to-day business operations. Over time, this data loses its importance and is accessed less often, gradually losing its business value, and ending with its archival or disposal.

Active Data

Active data is of business use to an organization. The ease of access for business users to active data is an absolute necessity in order to run an efficient business.

The simple, but critical principle, that all data moves through life-cycle stages is key to improving data management. By understanding how data is used and how long it must be retained, companies can develop a strategy to map usage patterns to the optimal storage media, thereby minimizing the total cost of storing data over its life cycle.

The same principles apply when data is stored in a relational database, although the challenge of managing and storing relational data is compounded by complexities inherent in data relationships. Relational databases are a major consumer of storage and are also among the most difficult to manage because they are accessed on a regular basis. Without the ability to manage relational data effectively, relative to its use and storage requirements, runaway database growth will result in increased operational costs, poor performance, and limited availability for the applications that rely on these databases. The ideal solution is to manage data stored in relational databases as part of an overall enterprise data management solution.

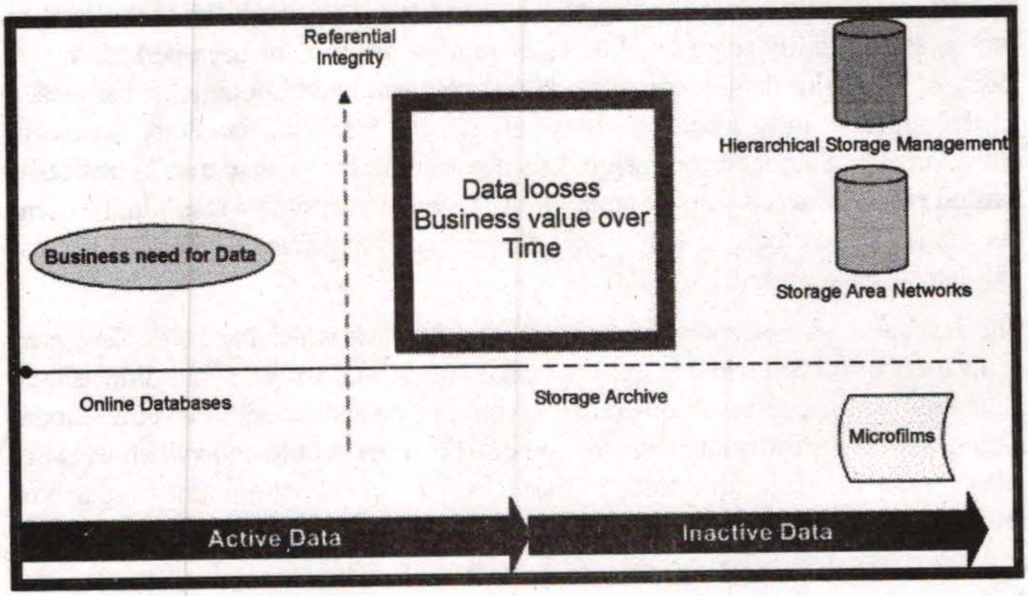


Figure 9.2 : Data Life Cycle in Enterprises

Inactive Data

Data are put out to pasture once they are no longer active. i.e. there are no longer needed for critical business tasks or analysis. Prior to the mid-nineties, most enterprises achieved data in Microfilms and tape back-ups. There are now technologies for data archival such as **Storage Area Networks (SAN), Network Attached Storage (NAS)** and **Hierarchical Storage Management**. These storage systems can maintain referential integrity and business context.

NOTES

9.4.4 Master Data Management

Master Data Management *(***MDM***)*, also known as Reference Data Management, is a sub-discipline of Data architecture within Information Technology (IT) that focuses on the management of reference or master data that is shared by several disparate IT systems and groups. MDM is required to enable consistent computing between diverse system architectures and business functions.

Large companies often have IT systems that are used by diverse business functions (e.g., finance, sales, R&D, etc.) and span across multiple countries. These diverse systems usually need to share key data that is relevant to the parent company (e.g., products, customers, and suppliers). It is critical for the company to consistently use these shared data elements through various IT systems.

MDM also becomes important when two or more companies want to share data across corporate boundaries. In this case, MDM becomes an industry issue such as is the case with the Finance industry and the required STP (Straight Through Processing) or T+1.

Other computing paradigms relevant for the core business of large companies are e.g. OLTP transactional computing (used for ERP and other cases where it is vital to make sure that not even a small part of a data set is lost) and DSS (Decision Support Systems, as one of the basic sources for management decisions); their tasks range from operational reporting to EIS (Executive Information Systems). Master data management is not only required to coordinate different ERP systems, but also necessary to supply meta-data for aggregating and integrating transactional data. This use of MDM is necessary for Data Warehouse projects typically incorporated in Decision Support Systems. For this reason, MDM systems sometimes provide a meta-data abstraction layer.

Master Data Management solutions tackle the rapidly escalating issues of dimension coordination, integration and reconciliation that enterprises face with each incremental year of systems evolution. An easy to use, web-based master data management system is the missing link between operational/transactional systems, business intelligence and Performance Management systems.

Most enterprises have implemented their own special blends of ERP, CRM, Business Intelligence, Financial Reporting, Planning, and other operational systems in order to accommodate unique business requirements. Each of these systems requires the same dimensional information to operate, yet they manage this information independently, which directly affects dimensional synchronization and reconciliation across systems. Master data management software aims to integrate dimensional and master data across BI, data warehouse, financial & operational systems, providing for accurate, consistent and compliant enterprise reporting.

Master Data Management software empowers business users with a best-practice business management process to centralize and directly manage the structure of corporate data.

Many organizations are looking to empower their business users to create and maintain Master Data and with the appropriate security release it to their business systems in minutes - without adding the burdens of time and money on their IT divisions.

NOTES

9.4.5 Data Farming

Data Farming is the process of using a high performance computer or computing grid to run a simulation thousands or millions of times across a large parameter and value space. The result of Data Farming is a "landscape" of output that can be analyzed for trends, anomalies, and insights in multiple parameter dimensions.

The term Data Farming comes from the idea of planting data in the simulation and parameter/value space, and then harvesting the data that results from the simulation runs.

Data Farming was originally used in the Marine Corp's Project Albert. Small agent-based distillation models (simulations) were created to capture a specific military challenge. These models were run thousands or millions of times at the Maui High Performance Computer Center and other facilities. Project Albert analysts would work with the military subject matter experts to refine the models and interpret the results. The Naval Post Graduate School also worked closely with Project Albert in model generation, output analysis, and the creation of new experimental designs to better leverage the computing capabilities at Maui and other facilities.

9.5 Data Warehouse Components

1. **Data Warehouse Database :** The central data warehouse database is a corner stone of the data warehouse environment. Certain data warehouse attribute such as very large database size, ad-hoc query processing and the need for flexible user view creation including aggregates multi-table joins and drill down have become drivers for different technological approaches to data warehouse database. This approch includes.

   (a) Parallel relational database design.

   (b) New indexed structure to bypass relational table.

   (c) Multidimensional database.

This approach is highly coupled with the OLAP (Online Analytical Processing) tools.

2. **Sourcing, Acquisition, Cleanup and Transformation Tools :** A data sourcing, cleanup, transformation, migrátion tools perform all the conversion, and summarization and key changes, structural changes and condensation needed to transform data into information that can be used by decision support tools.

   It produces the program and control statement including the COBOL program, Job Control Language (JCL), Unix scripts, and SQL data definition languages. The function includes -

   (a) Removing unwanted data

   (b) Converting to common data name and identification

   (c) Calculating summarize and drive data.

   (d) Establishing default for missing data

NOTES

(e) Accommodating source data definition changes.

Some Significance is used as follows :

(i) Data heterogeneity

(ii) Database heterogeneity

(iii) Meta Data

3. **Access Tools –** These users interact with Data warehouses using front end tools. Many of these tools require information specialist although may end-user develop expertise in the tools. These tools divided into main four categories :

(a) **Query & Reporting Tools –** This type of tool is divided into two category -

(i) **Reporting Tool :** It can be divided into production reporting tools, desktop report writers. Production reporting tools will let computer generate regular operational report on other hand report writers are inexpensive key tool designed for end-user.

(ii) **Managed Query Tools :** Managed query tool is used for SQL and database structure by inserting a Meta-layer between user and database.

(b) **Application -** These applications development environment includes power builder, visual basic, Business objects or the same application tools etc.

(c) **OLAP -** These tools are based on the concept of multidimensional databases and allow a sophisticated user to analyze the data using multidimensional views.

(d) **Data Mining -** A critical success factor for any business today is its ability to use the information effectively. This strategic use of data can result from opportunities presented by discovering hidden. Previsouly undeleted and frequently valuable facts about the consumer.

4. **Data Marts :** A data mart stores only that information which is needed to address a particular subject area and this can be translated as supporting the needs of a specific group of users. Compared with a data warehouse and data marts.

(a) has a smaller data model

(b) A shorter implementation curve

(c) Less Data

(d) Fewer users

A Data Mart will normally fit within the entry level.

If a Data Mart is implemented in a stand-alone fashion with its contents being extracted directly from production systems it can also be called an 'Independent Data Mart'. Sometimes a number of Data Marts are linked, together using distributes database techniques and become essentially a'Distributed DWH'.

5. **DWH Administrator and Management**

(a) Security management

NOTES

(b) Multiple sources
(c) Data quality checks
(d) Managing and updating Metadata
(e) Editing and Reporting data
(f) Delete unuseful data
(g) Backup and Recovery
(h) Storage and Management

6. **Information Delivery System :** The Information delivery system distributes warehouse stored data and other information objects to other data warehouses and end-user products such as spreadsheets and local databases. Delivery of information may be based on time of day. The rationale for the delivery system component is based on the fact that once the DWH is installed and operational, its users do not have to be aware of its location and maintenance.

9.6 Structure of Datawarehouse

The structure of a data warehouse is shown in figure. The structure consists of the following physical data warehouse, logical data warehouse and data marts.

(a) Physical Data warehouse : Physical database in which all the data for the data warehouse are stored, along with metadata and processing logic for scrubbing, organizing, packaging and processing the detail data.

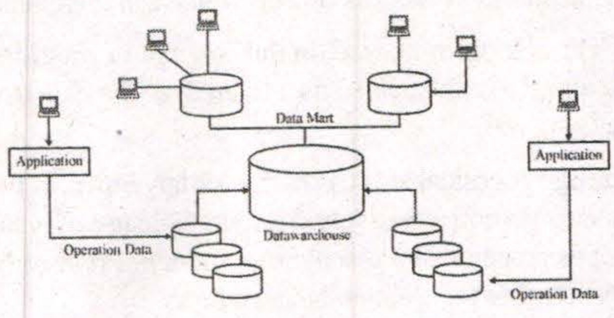


Figure 9.3 : Structure of Datawarehouse

(b) Logical Data warehouse – The logical data warehouse contains meta data, including enterprise rules and processing logic for scrubbing, organizing, packaging and processing the data, but does not contain actual data, instead, it contains the information necessary to access the data wherever they reside. This structure is effective only when there is a single source for the data and they are known to be accurate and timely.

9.7 Advantage of Data Warehouse

There are many advantages to using a data warehouse, some of them are:

- Enhances end-user access to a wide variety of data.
- Decision support system users can obtain specified trend reports, e.g. the item with the most sales in a particular area/country within the last two years.
- A data warehouse can be a significant enabler of commercial business applications, most notably customer relationship management (CRM).

9.8 Summarized Data

NOTES

According to W.H Inmon "a data warehouse is "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process".

Data that are being dealt with in a data warehouse are of very high volume as it contains organization's historical data. So if you we want to know about the current status of the company then accessing the whole data is irrelevant.

Summarized data is data obtained from whole set of data on the basis of some important characteristics. It is a mechanism that is used to divide the whole data into many parts and then store them on different databases. Now data can be accessed from these databases as per requirement.

Summarized data are taken from evaluational databases. The evaluational database is the data store to support OLAP function. This summarized data is used by the company or any data consumer requesting the information.

The use of highly summarized data is very useful for quick meeting reports and presentations or during times of very hectic schedules as in having stiff competition and an instant relevant and high quality information is needed as a basis for wise decision under pressure.

9.9 Current Details

The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the "business data warehouse". In essence, the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to decision support environments. The concept attempted to address the various problems associated with this flow - mainly, the high costs associated with it. In the absence of a data warehousing architecture, an enormous amount of redundancy was required to support multiple decision support environments. In larger corporations it was typical for multiple decision support environments to operate independently. Each environment served different users but often required much of the same data. The process of gathering, cleaning and integrating data from various sources, usually long existing operational systems (usually referred to as legacy systems), was typically in part replicated for each environment. Moreover, the operational systems were frequently reexamined as new decision support requirements emerged. Often new requirements necessitated gathering, cleaning and integrating new data from the operational systems that were logically related to prior gathered data.

Based on analogies with real-life warehouses, data warehouses were intended as large-scale collection/storage/staging areas for corporate data. Data could be retrieved from one central point or data could be distributed to "retail stores" or "data marts" that were tailored for ready access by users.

Key developments in early years of data warehousing were:

- 1960s - General Mills and Dartmouth College, in a joint research project, develop the terms *dimensions* and *facts*.
- 1970s - ACNielsen and IRI provide dimensional data marts for retail sales.
- 1983 - Teradata introduces a database management system specifically designed for decision support.

NOTES

- 1988 - Barry Devlin and Paul Murphy publish the article *An architecture for a business and information systems* in *IBM Systems Journal* where they introduce the term "business data warehouse".
- 1990 - Red Brick Systems introduces Red Brick Warehouse, a database management system specifically for data warehousing.
- 1991 - Prism Solutions introduces Prism Warehouse Manager, software for developing a data warehouse.
- 1991 - W.H. Inmon publishes the book *Building the Data Warehouse*.
- 1996 - Ralph Kimball publishes the book *The Data Warehouse Toolkit*.
- 1997 - Oracle 8, with support for star queries, is released.

In 1998, Microsoft released an OLAP (Online Analytical Processing) solution which manages multidimensional data. It includes multidimensional OLAP (MOLAP) and relational OLAP (ROLAP) capabilities. Many customers who deploy data warehousing decision support systems discover that a portion of their decision-support needs are constrained by the two-dimensional data structure of relational database management systems. The Microsoft OLAP product is designed with the following capabilities:

- Microsoft's OLAP solution works with any data provider that exposes ODBC or OLE DB.
- Microsoft offers a very unique and compelling integration of its OLAP and SQL servers.
- Microsoft's OLAP solution exposes an object model that is programmable using Microsoft Visual Basic, Java and C++ programming languages.
- Microsoft is working with many third-party applications providers to help them develop specialized front end products.

Some of the recent developments of data warehouse are as follows:

Clickstream Data

- Results from clicks at web sites.
- A dialog manager handles user interactions. An ODS helps to custom tailor the dialog.
- The clickstream data is filtered and parsed and sent to a data warehouse where it is analyzed.
- Software is available to analyze the clickstream data.

Further Automation of ETL Processes

- MetaRecon from Metagenix reverse engineers data into information.
- Analyzes and profiles source systems.
- Uncovers problems in source systems.
- Recommends primary and secondary keys, dimensions and measures, etc.
- Generates ETL scripts.

Meta Data Integration

- A growing realization that meta data is critical to data warehousing success.
- Progress is being made on getting vendors to agree on standards and to incorporate the sharing of meta data among their tools.

NOTES

- Vendors like Microsoft, Computer Associates, and Oracle have entered the meta data marketplace with significant product offerings.

Growing Dominance of MS SQL Server 7.0 with OLAP Services

- Low cost, integration of bundled DSS components from one vendor, and extended SQL for OLAP.
- Competitors are either leaving the market or are repositioning their products to be complimentary.

Enterprise Intelligence Portals

- Offers users an effective way to access information scattered across networked enterprise systems through a simple and personalized Web interface.
- Provides access to structured and unstructured data.
- Potentially integrates data warehousing and knowledge management.

9.10 System of Records

The opinion of the system of record has long been one of the cornerstones of the data warehouse environment. The system of record is the place where there is a definitive value for some unit of data. It is intuitive and dates back to the days of transaction processing systems. The system of record for the banking environment states that you have your balance for your account in exactly one place. If you have no system of record for your bank account or if you have multiple systems of record for the same account, something is fundamentally wrong. The same concept applies to many other systems, such as insurance policies. The insurance company has one and only one place where there is a system of record for your insurance policy. The company that sends you catalogs in the mail has only one place where you are known as a customer, and so forth.

Where there is a single system of record, there is integrity of data. Where there is no system of record, there is no integrity of data. For this simple reason, the system of record is one of the most basic concepts of the information systems environment.

Consider what happens when there is a global data warehouse. A global data warehouse is a data warehouse with multiple supporting local data warehouses. For example, there might be a headquarters data warehouse in New York City and different country data warehouses in France, the U.S., Hong Kong, Saudi Arabia and Canada. Or, there might be a central global data warehouse in Chicago, and a data warehouse for manufacturing in Dallas, a data warehouse for sales in Detroit and a data warehouse for distribution in Duluth. In each of these cases, there is a need for central data and a need for local data. In turn, there is a need for integrity of data across both local and global warehouses.

The question naturally arises: Where is the system of record? Is it at the global data warehouse, at the local data warehouse or somewhere in between? If there are multiple data warehouses, doesn't that destroy the concept of the system of record? Not at all. In the face of global data warehouses, the system of record distributes over multiple locations in accordance with the business function being served by the global data warehouse.

In order to show how the system of record becomes distributed in the face of local and global data warehouses, an example is in order. Suppose there is a global data warehouse for multiple lines of business. The global data warehouse resides in

NOTES

Chicago, the manufacturing data warehouse in Dallas, the sales data warehouse in Detroit and the distribution data warehouse in Duluth. The data found in the global data warehouse in Chicago is probably only financial. The only business connection between manufacturing, sales and distribution is at the dollar level. There is no attempt to create a commonality of data between sales and manufacturing. Therefore, the information found in the global data warehouse does not challenge the system of record because the system of record for different kinds of data resides at the local level.

However, there is a need for financial data to be collected at the global level. The financial data is first collected at the local level. Then, after collection, the data is passed to the global level. However, it is worth noting that in passing from the local level to the global level, the definition of data and its classification often change. For example, the global data warehouse may have a different definition of revenue than the local data warehouse. There are many reasons why a basic transformation of data takes place as the data passes from the local to the global data warehouse. This basic transformation explains why data in the global data warehouse does not add up to the simple sum of data from the local data warehouse; there are different systems of record. There is the local system of record, which is based on the understanding of business as the local operation sees business. Then, there is the global system of record, which operates on a different set of business rules. Additionally, there is the transformation, which defines how one system of record relates to another.

In many cases, the granularity of data changes as the system of record changes. The granularity of data at the local level is lower than that at the global level. For example, at the local level, "sales" means a record of each individual sale. However, at the global level, "sales" means sales by country accumulated across all of the individual records. If you want to find the details of a particular sale, you go to the local data warehouse. If you want a global perspective of all sales, you go to the global warehouse. In this sense, the system of record for a global data warehouse/ local data warehouse environment is split.

A local/global mapping is required to keep track of the transformations that occur as data is passed from one level to another. The local/global mapping is one that identifies the system of record locally and globally, and the transformations of the data as it passes from one level to the next. In addition, the mapping is time-variant; the different states of the mapping are kept as they change over time. The relationship between local mapping and global mapping is kept in this manner.

9.11 Integration and Transformation Program

Transformation of data usually involves code resolution with mapping tables for example changing the variables gender to:

(i) 0 if the value is female

(ii) 1 if the value is male

It involves changing the resolution of hidden. Business rules in data fields such as account numbers. Also, the structure and the relationships of the data are adjusted to the analysis domain. Transformations occur throughout the population process usually is more than one step. In the early stages of the process, the transformation are used more to consolidate the dates from different sources.

NOTES

Whereas in the later stages, data is transformed to satisfy a specific analysis problem on a tool requirement. Data Warehousing turns data into information; on the other hand, the data integration process that must be followed to physically load the data into the Data Warehousing . These steps are universally well understood since they are typically done regardless of whether the target system is a Data Warehouse or any system that require data migration. Experienced developers are generally available to perform these tasks since they represent standard activities performed daily in every data center. The only difference in Data Warehousing environment is that these steps are performed repeatedly for change requests to add or modify sources to populate the Data Warehouse.

The following summarizes the remaining steps of the data integration process that must be followed to physically load the data into the Data Warehouse.

Data Population represents the steps that need to be performed to physically load the source data into the integrated target data structure. The deliverable is the conversion programs run to load the source data into the target database. These are the steps for data population.

1. **Write Conversion Programs :** Once the conversion specifications have been signed off, developers are able to write the programs that will load the target data warehouse. Once the programs have been completed, they should be signed off and put into change management under the same version number as the assoicated conversion specifications they were written against.
2. **Perform Testing :** Once unit testing is completed, it will be necessary to perform system. testing by running the conversion programs to load data into the data warehouse test environment. Test plans and tools should be made available to review the data as it has been populated to determine of the business rules were written correctly
3. **Collect Statistics :** Generate log files to gather statistics regarding the load, e.g. how many records processed, and start and end times.
4. **Perform Quality Assurance :** Use the conversion specifications as a basis to determine the accuracy of the data that has populated the data warehouse. Review the exception file and repeat the necessary steps in the data integration process as needed.

9.12 Archives (Store House)

Archives means, a place for keeping records. The repository provides a place to store metadata. There are several implications that can, and should be made regarding the characteristics of a meta data repository. In addition to the obvious requirements of a physical area to accommodate meta data, the repository should be capable of managing and maintaing this meta data such that it can be used to better understand the contents of the data warehouse.

It should be architected such that the contents of the meta data repository are continually synchronized with the data in the enterprise data warehouse. This means that any structural change to the actual data is also reflected in the repository. Structural changes could include changing a calculation or size for a given column.

The scope of the meta data repository should include any data that is contained in the enterprise data warehouse. The meta data repository and the data warehouse

NOTES

should be tightly coupled to gain maximum value. All data in the warehouse should be accounted for in the repository. This should include both internal and external data, structured and unstructured.

From an IBM perspectire, there are two places where warehouse meta data is being stored. The first is the **Data warehouse centre control database,** which keeps the warehouse technical meta data generated during the definition of the warehouse and the warehouse transformations. It can also receive meta data from third-party ETL tools.

The second place, the **information catalog**, can be viewed as the central business meta data repository. However, this repository does not generate or refresh any meta data in and of itself. It is merely a persistent database for the storage, maintenance, and publication of meta data. It is fed meta data by other IBM tools, as well as by third party applications. The most significant contributor to the catalog repository from the IBM world is the data warehouse center, which controls much of the data propagation throughout the warehouse environment.

9.13 Meta Data

Meta Data is arguably the only ubiquitous warehouse component covering the entire landscape of core components and issues. Even through it touches virtually everything. It is probably the least understood and most neglected component. Often more difficult to implement then to talk about, Meta data is a constant challenge for data architects. Even so, a warehouse is rich with important information on the data being acquire, transformed, stored, accessed and analyzed.

Types of Meta Data

1. **Technical :** If any Meta-Data is captured by your organization, it will likely be this category. This is primarily because technical meta data is generated by the CASE (Computer Aided Software Engineering) tools used in warehouse efforts, as well as the DDL (Data Definition Language) associated with relational database. Examples of technical meta data include table names, column types and sizes, indexes and system names.

2. **Business :** A layer of meta data that we should gather and keep in repositories has to do with the business-related aspects of the data being stored and processed in the warehouse. Some samples of business meta data include hierarchies, drived data calculations and business names, for examples-P_id may be the technical name of the Product key in the product table, but the business name of the same column could be Product ID, which is easier for a user to understand.

9.14 Uses of a Datawarehouse

DW appliances provide solutions for many analytic application uses, including:

- Enterprise data warehousing .
- Super-sized sandboxes isolate power users with resource intensive queries .
- Pilot projects or projects requiring rapid prototyping and rapid time-to-value .

NOTES

- Off-loading projects from the enterprise data warehouse; ie large analytical query projects that affect the overall workload of the enterprise data warehouse .
- Applications with specific performance or loading requirements .
- Data marts that have outgrown their present environment .
- Turnkey data warehouses or data marts .
- Solutions for applications with high data growth and high performance requirements.
- Applications requiring data warehouse encryption .

9.14.1 DWA Trends

The DW appliance market is shifting trends in many areas as it evolves:

- Vendors are moving toward using commodity technologies rather than proprietary assembly of commodity components.
- Implemented applications show usage expansion from tactical and data mart solutions to strategic and enterprise data warehouse use.
- Mainstream vendor participation is now apparent.
- With a lower total cost of ownership, reduced maintenance and high performance to address business analytics on growing data volumes, most analysts believe that DW appliances will gain market share.

9.15 Standard Reports and Query

This type of tool divided into two category.

(1) Reporting Tool

(2) Managed Query Tool

(1) Reporting Tool – It can be divided into production reporting tools, desktop report writers, Productions reporting tools will get computer generate regular operational report on other hand report writers are inexpensive desktop tools designed for end-user.

The reports to look exactly the way that adheres to the corporate standard. It makes the analysts jobs much easier, and the time savings are tremendous.

Changing the number of dimensions or the number of aggregation levels should not significantly change reporting performance.

The most common export needs are to excel, to a flat file, and to PDF, and a good report tool must be able to export to all three formats. For excel, if the situation warrants it you will want to verify that the reporting format, not just the data itself, will be exported out to excel. This can often be a time-saver.

(2) Managed Query Tools – Managed query tools shield end users from the complexities of SQL and database structures by inserting a metalayer between users and the database. Metalayer is the software that provides subject-oriented view of a database and support point-and-click creation of SQL.

NOTES

Managed query tools have been extremely popular because they make it possible for knowledge workers to access corporate data without are intervention. Managed query tools ventors are racing to embed support for OLAP and data mining features.

Leading managed query tools are IQ software's IT objects, Andyne Computing Ltd's GQL, IBM's Decision Server, Speedware Corp's Esperant, and Oracle Corp.'s Discover/2000.

Most managed query tools have embraced three- tiered architectures to improve scalability. They support asynchronous query execution and integrate with web servers.

9.16 Queries Against Summarised Data

This unit describes how to use queries that group and aggregate data. These queries can be used to summarise data and calculate descriptive statistics based on the data values. At the end of this unit you should:

- understand why totals queries are useful for producing information from a database.
- be able to create a totals query that summarises data.
- understand the purpose and behaviour of each of the aggregate functions that are available in totals queries.
- understand the different ways to filter data by using criteria in a totals query.

9.16.1 Why use a totals query?

Management often requires summarised information regarding the operations of an organisation. These requests may take the form of a question such as how many staff members have attended the staff induction course so far this year, or what is the total number of staff in each department? In traditional information systems a great deal of programmer effort was expended developing reports that answered these types of questions. Microsoft Access provides a type of query called a totals query that simplifies the task of answering questions such as these.

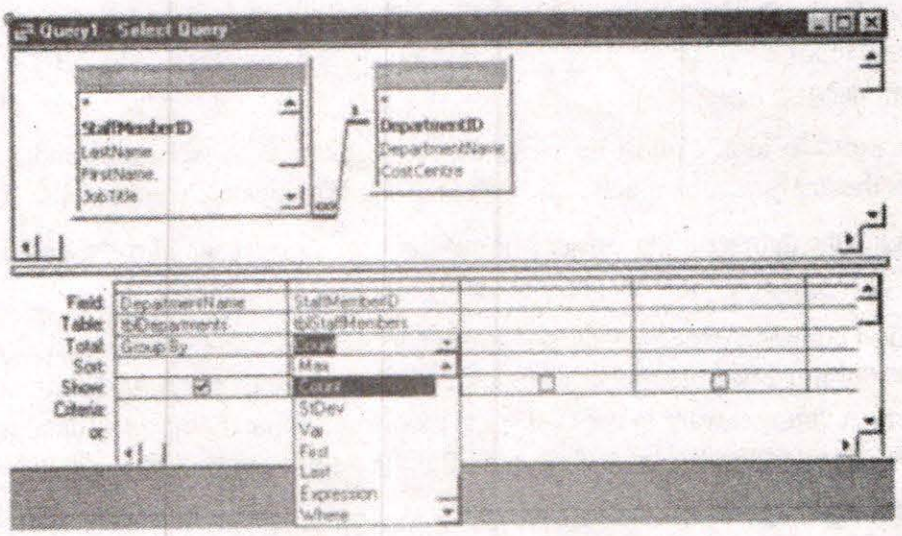


Figure 9.4 : Choosing a summary function in a totals query. This query will tell you how many staff are in each department

NOTES

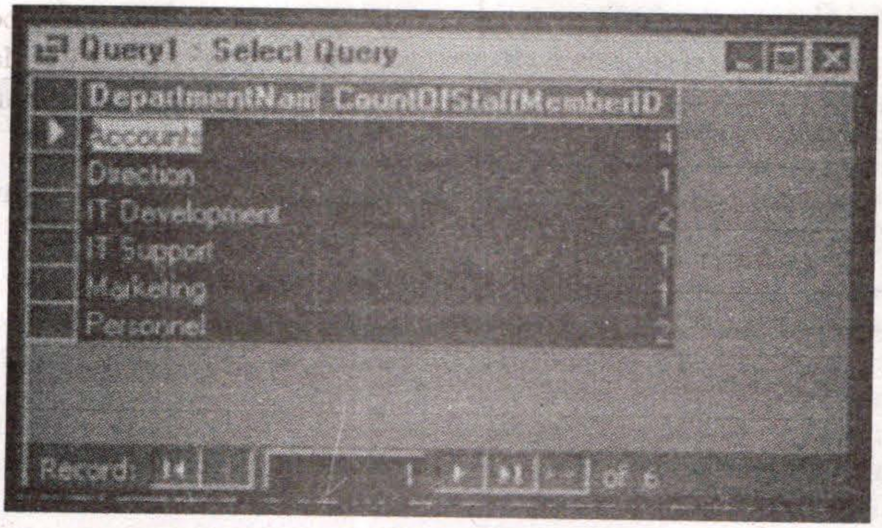


Figure 9.5 : The results of the totals query in a datasheet window

Totals queries allow you to partition your data into groups and to calculate summary information for each group. The values in the tabic or tables determine the groupings. Each group in a totals query is mutually exclusive.

To create a locals query, follow these steps.

1. Create a new query as described in previous topic. Add the fields that you need to include in the query to the query grid.
2. Now click on the View menu and choose Totals (or click on the Totals button on the Query design toolbar). An additional row will appear on the query grid with the word Total to the left. To begin with each column in the query will have the words "Group By" in this row.
3. If you want to group by the values in a particular column, leave the words "Group By" in the Total row. If you want to summarise the data, you need to change this value from "Group By" to a summary function in each column that you want to summarise. To do this click in the Total row under the appropriate column. Now click on the button with the down arrow and select the summary function you require (these functions are described in detail below).
4. When you have chosen a summary function for each column you want To summarise, run the query in the normal way. The results of a totals query-are displayed in a datasheet like other queries. You will see one row for each data. value in the last "Group By" column.

The column heading for the summarised columns will reflect the calculation that is being performed. As you can see in Figure. a count of a column named StaffMemberID will result in a column called CountofStaffMemberID.

Please note, totals queries cannot be updated. Each row in a totals query displays data that is derived from several rows in the underlying table. Because the rows returned by a totals query do not relate to individual rows in the table on which they are based. Access is unable to determine which row or rows to update.

9.16.2 Grouping data

When you apply grouping to the values in a particular column, you are partitioning the complete set of values in that column into sub-sets. Each sub-set is separate to each other sub-set (there are no overlaps) and you can emble the original set by combining all the sub-sets. You can picture this as being like the slices in a sliced loaf of bread.

NOTES

9.16.3 Groups and Nulls

If a column that you are grouping by contains Null values. Access will create a single group for all the null values. This suggests that all the Null values are treated as being equal, which we have seen previously is not the case. The alternative would be to create a separate group (and therefore a separate row) for each individual Null value. In most cases this is not what you want.

9.16.4 Sorting and groups

In order to perform the grouping in a totals query. Access will sort the values in the group by columns. As a result the values in the data set returned by the query will be sorted, even though you have not specified any sort order. If you specify sorting on a column other than a column on which you are grouping, it may lead to poor performance. In some cases this may be inevitable - if you want to create a ranking you will need to sort by one of the columns being summarised, not a group by column.

9.16.5 Summarising data

Microsoft Access provides a number of aggregate functions that can be used for summarising data. These aggregate functions measure a characteristic of a set of values and return it as a single number. The sections below describe each function in detail.

Please note, the Access implementation of these aggregate functions differs from the same functions in other database products. It also differs from the definition of these functions in the international database standard called ISO SQL. This is mainly because the functions have been designed in such as way that they can be used not only in queries, but also for calculations in forms and reports.

9.17 Interface With Other Warehouse

Data Warehouse enables you to access underlying historical data from various Tivoli and customer applications. The infrastructure provides a set of extract, transform and load (ETL) processes that you use to extract and move data from Tivoli application data stores to a central data warehouse database.

Tivoli Enterprise Data Warehouse provides the following capabilities:

- An open architecture for storing, aggregating, and correlating historical data. In addition to the data collected by IBM Tivoli software, Tivoli Enterprise Data Warehouse has the flexibility and extensibility to enable you to integrate your own application data.
- Database optimizations for the efficient storage of large amounts of historical data and for fast access to data for analysis and report generation.
- The infrastructure and tools necessary for maintaining and viewing the collected data. These include the Tivoli Enterprise Data Warehouse application, IBM DB2 Universal Database(TM) Enterprise Edition, the Data Warehouse Center, DB2 Warehouse Manager, and a user interface for creating and viewing reports.
- The ability to use your choice of data analysis tools to examine your historical data. In addition to the report interface, you can analyze your data using other products such as online analytical processing (OLAP), planning, trending, analysis, accounting, and data mining tools.
- The ability to control access to your historical data. You can keep data about multiple customers and data centers in one central data warehouse, but restrict access so that customers can see and work with data and reports based only on their data and not any other customer's data. You can also restrict an

individual user's ability to access data.

- A zero-footprint client. Users can access Tivoli Enterprise Data Warehouse reports from any system by using a Web browser. No special software is required on the user's system.
- Internationalization support. Not only is the report interface localized, application programmers can localize the data stored in the central data warehouse.

NOTES

Tivoli Enterprise Data Warehouse consists of the following components:

- Control server
- Central data warehouse
- Data marts
- Report interface

Control server

The control server contains the control database for Tivoli Enterprise Data Warehouse from which you manage your data warehouse. The control server has the following subcomponents:

- A server that controls communication between the control server, the central data warehouse server, the data mart server, and the report server.
- The control database, which contains metadata for Tivoli Enterprise Data Warehouse.

The control server uses the following parts of the IBM DB2 product, which you must install manually before installing the control server. These parts are all automatically installed when you install IBM DB2 Universal Database Enterprise Edition on a Microsoft Windows system:

- DB2 Server
- The Data Warehouse Center, a component that automates data warehouse processing. You can use the Data Warehouse Center to define the ETL processes that move and transform data into the central data warehouse and the star schemas used by the data marts. Then, you can use the Data Warehouse Center to schedule, maintain, and monitor these processes.
- The warehouse agent, part of DB2 Warehouse Manager.

Central Data Warehouse

The *central data warehouse* is a DB2 database that contains the historical data for your enterprise. The system that hosts the central data warehouse is called the *central data warehouse server*. The central data warehouse component uses IBM DB2 Universal Database Enterprise Edition, which you must install manually before installing the control server.

Data marts

A separate DB2 database contains the data marts for your enterprise. Each data mart contains a subset of the historical data from the central data warehouse to satisfy the analysis and reporting needs of a specific department, team, customer, or application. The system that hosts this DB2 database is called the *data mart server*. Although you can have many data marts, you can have only one data mart server.

The data mart component requires IBM DB2 Universal Database Enterprise Edition, which you must install manually before installing the control server.

The warehouse pack for each component of IBM Tivoli Monitoring for Business Integration creates a data mart whose structure is suitable for the report interface.

NOTES

The warehouse pack provides an extract, transform, and load (ETL) process called a data mart ETL that populates the star schemas associated with the data mart with data from the central warehouse.

You can modify an existing data mart, or create new data marts that contain slightly different data, to address a reporting need specific to your situation. To modify or create a data mart, you must be familiar with database ETL processes and with the internal representation of a data mart as star schemas in the Tivoli Enterprise Data Warehouse databases. For more information, refer to the *Enabling an Application for Tivoli Enterprise Data Warehouse* document.

Report interface

The Tivoli Enterprise Data Warehouse report interface (RPI) provides tools and a graphical user interface that other Tivoli software products use to create and display reports. You can use the Tivoli Enterprise Data Warehouse report interface to customize the reports that warehouse packs provide and to create new reports. You also use the report interface to control access to data marts and to the reports associated with a data mart. The system on which you install the report interface is called the report server.

Use the Work with Reports task group in the report interface to manage users, groups, and data marts for Tivoli Enterprise Data Warehouse or to run, create, and view Tivoli Enterprise Data Warehouse reports. The following sections describe how to perform these tasks:

- Working with users and user groups
- Managing Tivoli Enterprise Data Warehouse reports

9.18 Common Warehouse Metamodel

The **Common Warehouse Metamodel** (**CWM**) is a specification for modeling metadata for relational, non-relational, multi-dimensional, and most other objects found in a data warehousing environment. The specification is released and owned by the Object_Management_Group, which also claims a trademark in the use of "CWM" .

The active version of the CWM specification is v1.1 with a supplementary specification, Common Warehouse Metamodel (CWM), Metadata Interchange Patterns (MIP) that further refines the requirements for tools to smoothly interoperate.

Purpose of CWM

The Common Warehouse metamodel specifies interfaces that can be used to enable interchange of warehouse and business intelligence metadata between warehouse tools, warehouse platforms and warehouse metadata repositories in distributed heterogeneous environments. CWM is based on three standards:

- UML - Unified Modeling Language, an OMG modeling standard
- MOF - Meta Object Facility, an OMG metamodeling and metadata repository standard
- XMI - XML Metadata Interchange, an OMG metadata interchange standard

CWM models enable users to trace the lineage of data – CWM provides objects that describe where the data came from and when and how the data was created. Instances of the metamodel are exchanged via XML Metadata Interchange (XMI) documents.

NOTES

Vendors Supporting CWM

These vendors have been identified as having a CWM implementation or have active projects to support CWM.

- **IKAN** CWD4ALL, a CWM compliant database modeling and design tool
- **SAS** ,SAS adheres to the Object Management Group's Common Warehouse Metamodel (CWM) as the interoperability and interchange standard. An alliance between SAS and Meta Integration Technology Inc. (MITI) enables SAS to provide bridges for sharing and exchanging metadata with more than 40 design tool and repository vendors.
- **Oracle_Corporation** Oracle Warehouse Builder
- **Informatica** produces Superglue, an ETL tool with a Metadata extension (Metadata Manager formerly known as SuperGlue). Informatica is one of the members of the OMG .
- **PrudSys** - XELOPES library for embedded data mining

9.19 ETL

Extract, transform, and load ***(ETL)*** is a process in data warehousing that involves extracting data from outside sources, transforming it to fit business needs, and ultimately loading it into the data warehouse. ETL is important, as it is the way data actually gets loaded into the warehouse.

The processes inside ETL are:

Extract

The first part of an ETL process is to extract the data from the source systems. Most data warehousing projects consolidate data from different source systems. Each separate system may also use a different data organization / format. Common data source formats are relational databases and flat files, but may include non-relational database structures or other data structures. Extraction converts the data into a format for transformation processing.

Transform

The transform stage applies a series of rules or functions to the extracted data to derive the data to be loaded. Some data sources will require very little manipulation of data. In other cases, one or more of the following transformations types may be required:

- Selecting only certain columns to load (or selecting null columns not to load)
- Translating coded values (*e.g.*, if the source system stores M for male and F for female, but the warehouse stores 1 for male and 2 for female)
- Encoding free-form values (*e.g.*, mapping "Male" and "M" and "Mr" onto 1)
- Deriving a new calculated value (*e.g.*, sale_amount = qty * unit_price)
- Joining together data from multiple sources (*e.g.*, lookup, merge, *etc.*)
- Summarizing multiple rows of data (*e.g.*, total sales for each region)
- Generating surrogate key values

- Transposing or pivoting (turning multiple columns into multiple rows or vice versa)
- Splitting a column into multiple columns (*e.g.*, putting a comma-separated list specified as a string in one column as individual values in different columns)

NOTES

Load

The load phase loads the data into the data warehouse. Depending on the requirements of the organization, this process ranges widely. Some data warehouses merely overwrite old information with new data. More complex systems can maintain a history and audit trail of all changes to the data.

9.19.1 ETL Challenges

ETL processes can be quite complex, and significant operational problems can occur with improperly designed ETL systems.

The range of data values or data quality in an operational system may be outside the expectations of designers at the time validation and transformation rules are specified. **Data profiling** of a source during data analysis is recommended to identify the data conditions that will need to be managed by transform rules specifications.

The scalability of an ETL system across the lifetime of its usage needs to be established during analysis. This includes understanding the volumes of data that will have to be processed within Service Level Agreements, (SLAs). The time available to extract from source systems may change, which may mean the same amount of data may have to be processed in less time. Some ETL systems have to scale to process terabytes of data to update data warehouses with tens of terabytes of data. Increasing volumes of data may require designs that can scale from daily batch to intra-day micro-batch to integration with message queues for continuous transformation and update.

A recent development in ETL software is the implementation of **parallel processing**. This has enabled a number of methods to improve overall performance of ETL processes when dealing with large volumes of data.

There are **3 main types of parallelisms** as implemented in ETL applications:

- **Data**: By splitting a single sequential file into smaller data files to provide parallel access.
- **Pipeline**: Allowing the simultaneous running of several components on the same data stream. An example would be looking up a value on record 1 at the same time as adding together two fields on record 2.
- **Component**: The simultaneous running of multiple processes on different data streams in the same job. Sorting one input file while performing a reduplication on another file would be an example of component parallelism.

All three types of parallelism are usually combined in a single job.

An additional difficulty is making sure the data being uploaded is relatively consistent. Since multiple source databases all have different update cycles (some may be updated every few minutes, while others may take days or weeks), an ETL system may be required to hold back certain data until all sources are synchronized. Likewise, where a warehouse may have to be reconciled to the contents in a source system or with the general ledger, establishing synchronization and reconciliation points is necessary.

9.19.2 ETL Tools

NOTES

While an ETL process can be created using almost any programming language, creating them from scratch is quite complex. Increasingly, companies are buying ETL tools to help in the creation of ETL processes.

A good ETL tool must be able to communicate with the many different relational databases and read the various file formats used throughout an organization. ETL tools have started to migrate into Enterprise Application Integration, or even Enterprise Service Bus, systems that now cover much more than just the extraction, transformation and loading of data. Many ETL vendors now have data profiling, data quality and metadata capabilities.

9.20 Data Warehouse Appliance

A data warehouse appliance(DWA) is an integrated set of servers, storage, OS, DBMS and software specifically pre-installed and pre-optimized for data warehousing. DW appliances provide solutions for the mid-to-large volume data warehouse market, offering low-cost performance on data volumes in the terabyte to petabyte range. Data warehouse appliance vendors provide and service all the parts within the appliance.

As companies recognize the benefits gained from analyzing data and business intelligence, the demand for business analytics explodes. Data warehouses that service analytical queries must reliably support increased workloads as well as rapid data growth. The average data warehouse expands between 50% and 100% each year. The high cost to meet new workload and capacity requirements has allowed DW appliance vendors to enter the market with data warehouse platforms that offer reduced administration and attractive price/performance value.

In a nutshell, DW appliances provide low-cost packaged solutions for parallel data warehouse performance unmatched in many alternate solutions.

Most DW appliance vendors use massively parallel processing (MPP) architectures to provide high query performance and platform scalability. MPP architectures consist of independent processors or servers executing in parallel. Most MPP architectures implement a "shared nothing architecture" where each server is self-sufficient and controls its own memory and disk. Shared nothing architectures have a proven record for high scalability and little contention. DW appliances distribute data onto dedicated disk storage units connected to each server in the appliance. This distribution allows DW appliances to resolve a relational query by scanning data on each server in parallel. The divide-and-conquer approach delivers high performance and scales linearly as new servers are added into the architecture.

MPP database architectures are not new. Teradata, Tandem, Britton Lee and Sequent designed the first MPP SQL-based architectures. The re-emergence of MPP data warehouses is due to open source and commodity components. Advances in technology have reduced costs and improved performance in storage devices, multi-core CPUs and networking components. Open source RDBMS products, such as

NOTES

Ingres and Postgres, reduce software licenses and allow DW appliance vendors to focus on optimization rather than providing basic database functionality. Open source Linux provides a stable, well-implemented OS for DW appliances.

9.20.1 DWA Products

Many consider Teradata's initial product as the first DW appliance. Since then, Teradata has changed its architecture to a more software-centric multi-platform solution. While Teradata retains some appliance-like benefits, analysts now consider Netezza Performance Server as the forerunner in today's data warehouse appliance market. Netezza entered the market in 2000 with a data warehouse appliance offering fast table scans at low costs. Internally, Netezza combines the Postgres DBMS and its own record storage manager (RSM).

More recently, a second generation of DW appliances has emerged, marking the move to mainstream vendor integration. These DW appliance vendors have partnered with major hardware vendors to bring their appliances to market. DATAllegro partners with EMC and Dell and implements open source Ingres on Linux. Greenplum has a partnership with Sun Microsystems and implements Bizgres (a form of Postgres) on Linux. HP Neoview has a wholly owned solution and uses HP NonStop SQL.

Kognitio and ParAccel are "virtual" data warehouse appliances. These solutions provide software-only solutions deployed on clusters of commodity hardware. Kognitio's homegrown WX2 database runs on several blade configurations. Other players in the DW appliance space include Calpont and column-based Vertica.

The newest entry into the data warehouse appliance market is Dataupia with a product accelerating the performance of Oracle and Microsoft SQL Server Recently, the market has seen the emergence of data warehouse bundles where vendors combine their hardware and database software together as a data warehouse platform. IBM bundled DB2 Warehousing edition with its i-Series servers to create the BCU (Balance Configuration Unit). Some analysts classify IBM's BCU as a bundle, while others classify it as a semi-appliance.

9.20.2 DWA Benefits

- **Reduction in Costs**

  The total cost of ownership (TCO) of a data warehouse consists of initial entry costs, on-going maintenance costs and the cost of increasing capacity as the data warehouse grows. DW appliances offer low entry and maintenance costs. Initial costs range from \$18,000 to \$150,000 per terabyte, depending on the size of the DW appliance installed.

  The resource cost for monitoring and tuning the data warehouse makes up a large part of the TCO, often as much as 80%. DW appliances reduce administration for day-to-day operations, setup and integration. Many also offer low costs for expanding processing power and capacity.

  With the increased focus on controlling costs combined with tight IT Budgets, data warehouse managers need to reduce and manage expenses while leveraging their technology as much as possible making DW appliances a natural solution.

NOTES

- **Parallel Performance**

  DW appliances provide a compelling price/performance ratio. Many support mixed-workloads where a broad range of ad-hoc queries and reports run simultaneously with loading. DW appliance vendors use several distribution and partitioning methods to provide parallel performance. Some DW appliances scan data using partitioning and sequential I/O instead of index usage. Other DW appliances use standard database indexing.

  With high performance on highly granular data, DW appliances are able to address analytics that previously could not meet performance requirements.

- **Reduced Administration**

  DW appliances provide a single vendor solution and take ownership for optimizing the parts and software within the appliance. This eliminates the customer's costs for integration and regression testing of the DBMS, storage and OS on a terabyte scale and avoids some of the compatibility issues that arise from multi-vendor solutions. A single support point also provides a single source for problem resolution and a simplified upgrade path for software and hardware.

  The care and feeding of DW appliances is less than many alternate data warehouse solutions. DW appliances reduce administration through automated space allocation, reduced index maintenance and in most cases, reduced tuning and performance analysis.

- **Built-in High Availability**

  DW appliance vendors provide built-in high availability through redundancy on components within the appliance. Many offer warm-standby servers, dual networks, dual power supplies, disk mirroring with robust failover and solutions for server failure.

- **Scalability**

  DW appliances scale for both capacity and performance. Many DW appliances implement a modular design that database administrators can add to incrementally, eliminating up-front costs for over-provisioning. In contrast, architectures that do not support incremental expansion result in hours of production downtime, during which database administrators export and re-load terabytes of data. In MPP architectures, adding servers increases performance as well as capacity. This is not always the case with alternate solutions.

- **Rapid Time-to-Value**

  Companies increasingly expect to use business analytics to improve the current cycle. DW appliances provide fast implementations without the need for regression and integration testing. Rapid prototyping is possible because of reduced tuning and index creation, fast loading and reduced needs for aggregation in some cases.

9.21 Data Marts

A **data mart (DM)** is a specialized version of a data warehouse (DW). Like data warehouses, data marts contain a snapshot of operational data that helps business people to strategize based on analyses of past trends and experiences. The key

NOTES

difference is that the creation of a data mart is predicated on a specific, predefined need for a certain grouping and configuration of select data. A data mart configuration emphasizes easy access to relevant information.

There can be multiple data marts inside a single corporation; each one relevant to one or more business units for which it was designed. DMs may or may not be dependent or related to other data marts in a single corporation. If the data marts are designed using conformed facts and dimensions, then they will be related. In some deployments, each department or business unit is considered the *owner* of its data mart including all the *hardware*, *software* and *data*. This enables each department to use, manipulate and develop their data any way they see fit; without altering information inside other data marts or the data warehouse. In other deployments where conformed dimensions are used, this business unit ownership will not hold true for shared dimensions like customer, product, etc.

9.21.1 Need for Data Marts

- Easy access to frequently needed data .
- Creates collective view by a group of users.
- Improves end-user response time.
- Ease of creation.
- Lower cost than implementing a full Data warehouse .
- Potential users are more clearly defined than in a full Data warehouse .

9.21.2 Dependent Data Mart

A **dependent data mart** is a logical subset (view) or a physical subset (extract) of a larger data warehouse, isolated for one of the following reasons:

- A need for a special data model or schema: e.g., to restructure for OLAP .
- **Performance:** to offload the data mart to a separate computer for greater efficiency or to obviate the need to manage that workload on the centralized data warehouse.
- **Security:** to separate an authorized data subset selectively.
- **Expediency:** to bypass the data governance and authorizations required to incorporate a new application on the Enterprise Data Warehouse.
- **Proving Ground:** to demonstrate the viability and ROI (return on investment) potential of an application prior to migrating it to the Enterprise data warehouse

  .
- **Strategy:** a coping strategy for IT (Information Technology) in situations where a user group has more influence than funding or is not a good citizen on the centralized data warehouse.
- **Politics:** a coping strategy for consumers of data in situations where a data warehouse team is unable to create a usable data warehouse.

Tradeoffs inherent with data marts include limited scalability, duplication of data, data inconsistency with other silos of information, and inability to leverage enterprise sources of data.

SUMMARY

NOTES

- An operational data store (ODS) is a subject-oriented, integrated, volatile, current-valued, detailed-only collection of data in support of an organization's need for up-to-the-second, operational, integrated, collective information.
- Multidimensional databases are variously (depending on the context) data aggregators which combine data from a multitude of data sources; databases which offer networks, hierarchies, arrays and other data formats difficult to model in SQL; or databases which give a high degree of flexibility in the definition of dimensions, units, and unit relationships, regardless of data format.
- Multi-dimensional databases are especially useful in a sales and marketing applications that involve time series.
- The data cube is a conceptual representation of database which can be implemented in a variety of ways, including top-down, bottom-up, and arrays.
- Data becomes active as soon as it is of interest to an organization. Data life cycle begins with a business need for acquiring data. Active data are referenced on a regular basis during day-to-day business operations. Over time, this data loses its importance and is accessed less often, gradually losing its business value, and ending with its archival or disposal.
- *Master Data Management (MDM),* also known as Reference Data Management, is a sub-discipline of Data architecture within Information Technology (IT) that focuses on the management of reference or master data that is shared by several disparate IT systems and groups.
- Master Data Management software empowers business users with a best-practice business management process to centralize and directly manage the structure of corporate data.
- *Data Farming* is the process of using a high performance computer or computing grid to run a simulation thousands or millions of times across a large parameter and value space.
- A *data warehouse(DW)* is the main repository of an organization's historical data, its corporate memory. It contains the raw material for management's decision support system.
- The goal of Data Warehousing is to generate front-end analytics that will support business executives and operational managers.
- The *Common Warehouse Metamodel (CWM)* is a specification for modeling metadata for relational, non-relational, multi-dimensional, and most other objects found in a data warehousing environment.
- *Extract, transform, and load (ETL)* is a process in data warehousing that involves extracting data from outside sources, transforming it to fit business needs, and ultimately loading it into the data warehouse.

NOTES

- A data warehouse appliance(DWA) is an integrated set of servers, storage, OS, DBMS and software specifically pre-installed and pre-optimized for data warehousing.
- A data mart (DM) is a specialized version of a data warehouse (DW). Like data warehouses, data marts contain a snapshot of operational data that helps business people to strategize based on analyses of past trends and experiences. The key difference is that the creation of a data mart is predicated on a specific, predefined need for a certain grouping and configuration of select data.

EXERCISE

1. What are the role of operational data stores?
2. What are Multidimensional databases ? How are these gaining popularity now a days?
3. How can you describe the Data Life Cycle Management ?
4. Differentiate between the active and inactive data ? Compare it with historical and operational data.
5. Describe the Master Data Management (MDM) ?
6. What is data farming ?
7. What is data warehouse and data warehousing ?
8. Describe some special characteristics of data warehouses .
9. Whay Operational systems are becoming unable to meet the present business needs ?
10. Describe several fundamental stages from where the data warehouses are evolved.
11. Describe the following :
    - Offline Operational Databases
    - Offline Data Warehouse
    - Real Time Data Warehouse
    - Integrated Data Warehouse
12. What are different issues associated with data warehouses ?
13. Describ the detail architecture of data warehouse and data warehousing framework .
14. What are the advatages of data warehouses ?
15. Describe the Common Warehouse Metamodel (CWM) .
16. What is ETL ? Explain its steps .
17. What are data warehouse appliances(DWA) ?
18. What are data marts ? How these are different from Data warehouses?

NOTES

10 Data Mining

Chapter Includes :

10.1 Introduction

The large amounts of data collected and stored might contain some information, which could be useful, but it is not obvious to recognize, nor trivial to obtain it. Huge amount of data collected in various processes, either manufacturing or business, (often as a side effect of computerization) should be thoroughly analyzed as they might contain some precious information for decision support. There is nothing new about analyzing data, but it is in the amount of data, where traditional methods are becoming inefficient.

Modern databases contain a huge amount of data. However raw data are hardly of any utility. To gain information and knowledge, data must be first analyzed and processed. The automated analysis of data in search of some information is called data mining.

NOTES

The growth of information resources along with the accelerating rate of technological change has produced huge amounts of information that often exceed the ability of managers and employees to assimilate and use it productively. Data must be categorized in some manner if it is to be accessed, re-used, organized, or synthesized to build a picture of the company's competitive environment or solve a specific business problem.

In recent years, the need to extract knowledge automatically from very large databases has grown. In response, the closely related fields of knowledge discovery in databases (KDD) and data mining have developed processes and algorithms that attempt to intelligently extract interesting and useful information from vast amounts of raw data.

10.2 Data Mining & Evolution of DM

Introduction : Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).

As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a *difference of kind rather than degree*. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. For example, a user might hypothesize that a customer who buys a hammer, will also buy a box of nails. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, a hardware store may compare their customers' tool purchases with home ownership, type of automobile driven, age, occupation, income, and/or distance between residence and the store. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.

Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation.

NOTES

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source.

"Data Mining is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data" (Srikant & Agrawal).

That is a very brief definition that implies the purposes of doing mining and extracting new information from data:

- It is "valid" because it looks to be well grounded in logic patterns. In order to be valid the processes can be automatic or semiautomatic and there are many tools that are used to make the used algorithms and the resulting patterns as valid as possible;
- It is a "novel" because data mining a lot of research needs to be done yet;
- It is "potentially useful" because the results can be used in the decision making process of any organization, such as health, education, marketing, etc.
- "Understandable" patterns because the results should be capable of being understood or interpreted by users from different backgrounds and not only for researchers.
- "Patterns" from previous data because a perceptual structure has been created as a model that can be applied to new data.
- "Data" refers to the digitalized information in databases first and data warehouses later that can be accessed by data mining tools.

Data Mining is an information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results . The main points of this definition are the term "facts" because data mining works with "real" data and "hidden facts" because data mining shows the behavior and performance that is not easily discovered.

NOTES

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in NewEngland last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Table 10.1. Steps in the Evolution of Data Mining

Commercial databases are growing at unprecedented rates. In the evolution from business data to business information, each new step has built upon the previous one. According to Kurt Thearling , these are the steps in the evolution of data mining:

10.3 Data mining: Verification vs. Discovery

Decision support systems (DSS), executive information systems, and query/report writing tools are used to produce reports about data, usually aggregating it through any number of dimensions. Another use of these tools is to detect trends and patterns in customer data that will help answer some questions about the business. When used in this mode, a query is created to access the records relevant to the question(s) being formulated. After the data is retrieved, it is examined to detect the existence of patterns or other useful information that can be used in answering the original question(s). We call this the verification mode. In this mode, the user of a DSS generates a hypothesis about the data, issues a query against the data and examines the results of the query looking for affirmation or negation of the hypothesis. In the first case, the process ends; in the latter case, a new query is reformulated and the process iterates until the resulting data either verifies the hypothesis or the user decides that the hypothesis is not valid for his data.

Consider the following example. A sales executive has a limited budget to do a mailing campaign for a new product. In order to optimize the use of this money, the marketing executive wants to identify the largest set of people that are the most likely candidates to buy the new product and which can be reached within the budget limitation. To identify these customers and to verify that the customer set has been adequately narrowed to match the available promotional budget, the executive makes a hypothesis about the potential customer set. Issuing a query against the databases that contain historical data about customer purchases and demographic information respectively, the set of customers that have made significant purchases of competitive products can be obtained. Furthermore, to limit the number of customers found to a reasonable number, the executive requests to only get information about those customers that are characterized by having ages between 30 and 45 years, being heads of household with combined incomes between $25,000 and

$50,000 and living in some specific zip code regions. If the result of this query returns a number of customers that match the available budget for mailing promotions, the process ends. However, if either significantly more (or less) customers are found than the number that can be reached with the given budget, a new query limiting (or expanding) the set of customer addresses requested must be issued.

In the above example, the hypotheses used in formulating the queries where quite explicit (e.g., incomes between certain amounts.) Even when the hypotheses are implicit, the process of finding useful trends or patterns using queries can be described by the above behavior, as shown in the following example involving a query drill down process.

After a report about company sales shows that the last quarter sales where significantly lower than expected, the financial officer of the company wants to discover what caused this situation. A query is first issued to return the sales figures, by region, for the last quarter. The results of this query show that all the sales are up, except for one particular region. The financial officer begins to suspect that the problem may have occurred in some localized store. To better understand the nature of the problem, another query is issued that will return sales results for all the cities in the offending region. A result showing one city significantly lower that the rest reinforces the officers' suspicion; a result showing that sales were uniformly lower among all cities in this region requires that the initial guess about what caused the problem (i.e., the implicit hypothesis) be modified. New queries continuing to drill down looking for the results by store within an offending city follow the previous query; totally new queries need to be devised if the results of the last query contradict the implicit hypothesis.

Queries, such as those used in the previous two examples, always return records that satisfy the query predicates. Thus, little new information is created in this retrieval process: either the hypothesis is verified or it is negated. The process of information finding is done by the user by successive iterations upon examining the results of query after query and linking the verified and refined hypotheses. This is the essence of a verification model.

Many times, while performing a query, a request is made to compute functions related to the records being inspected during the query (e.g., count the number of records, find the average of a given field of the records, etc.) All these operations result in additional information being returned together with the query. For the purposes of this discussion, these derived facts are not considered.

Notice that, from the user perspective, he/she is discovering facts about the data. The use of queries to extract facts from databases is a common practice. There are other tools that, like query generators, are used in a mode that follows the verification model described above. Examples of these other tools are multidimensional analysis tools and visualization tools. Multidimensional tools make it easier for the user to formulate drill down queries such as those shown in the last example. Visualization tools are used, as their name implies, to present data in a visual manner and to allow the user to easily interact with the data in search of hidden patterns. The user of a visualization tool takes advantage of the human's visual perception capabilities to discern patterns. The three types of tools discussed above, queries, multidimensional analysis and visualization, all have in common that the user is essentially "guiding" the exploration of the data being inspected.

NOTES

Data mining uses a different model for the creation of information about data. We call this the discovery model. In the next section we will describe methodologies that can sift through the data in search of frequently occurring patterns, can detect trends, produce generalizations about the data, etc. These tools can discover these types of information with very little (or no) guidance from the user. The discovery of these facts is not a consequence of a haphazard event. Quite to the contrary, a well designed data mining tool is one that is architected and built so that the exploration of the data is done in such a way as to yield as large a number of useful facts about the data as possible in the shortest amount of time.

Comparing the process of finding information in a collection of data to that of mining diamonds in a diamond mine, we can say that "verification" is like drilling individual holes in a lode with the expectation of finding diamonds. Finding all (or many) diamonds in this way can be very inefficient. "Discovery", on the other hand, is similar to scooping out all the material in the lode and dumping it on a plain field so that all the glittering stones are thrown up into the open. Diamonds are then separated from the quartz by further inspection. In Data Mining, large amounts of data are inspected, facts are discovered and brought to the attention of the person doing the mining. Unlike diamonds, which are easily distinguishable from quartz, business judgment must be used to separate the useful facts from those which are not. Because this last step does not involve sifting through the raw data, Data Mining is a more efficient mode of finding useful facts about data.

10.4 Data Mining Technology

Extracting knowledge hidden in large volumes of raw data

Competitive advantage requires abilities. Abilities are built through knowledge. Knowledge comes from data. The process of extracting knowledge from data is called Data Mining.

Typical tasks addressed by data mining include:

- Rate customers by their propensity to respond to an offer
- Identify cross-sell opportunities
- Detect fraud and abuse in insurance and finance
- Estimate probability of an illness re-occurrence or hospital re-admission
- Isolate root causes of an outcome in clinical studies
- Determine optimal sets of parameters for a production line operation
- Predict peak load of a network

Without proper analytical tools, discovering useful knowledge hidden in huge volumes of raw data represents a formidable task. The exponential growth in data, diverse nature of data and analysis objectives, the complexity of analyzing mixed structured data and text are among the factors that turn knowledge discovery into a real challenge.

Data Mining provides tools for automated learning from historical data and developing models to predict outcomes of future situations. The best data mining software tools provide a variety of machine learning algorithms for modeling, such as Regression, Neural Network, Decision Tree, Bayesian Network, CHAID, Support Vector Machine, and Random Forest, to name a few.

Yet, data mining requires far more than just machine learning. Data mining additionally involves data pre-processing, and results delivery. Data pre-processing includes loading and integrating data from various data sources, normalizing and cleansing data, and carrying out exploratory data analysis. Results delivery includes model application in production environment and generating reports summarizing the results of the analysis in a simple form for business users.

Megaputer's flagship data mining tool PolyAnalyst supports all steps of data pre-processing and modeling, and results delivering. PolyAnalyst enables you to solve tasks of predicting, classification, clustering, affinity grouping, link analysis, multi-dimensional analysis, and interactive graphical reporting.

10.5 Advantages of Data Mining

Marking/Retailing

Data mining can aid direct marketers by providing them with useful and accurate trends about their customers' purchasing behavior. Based on these trends, marketers can direct their marketing attentions to their customers with more precision. For example, marketers of a software company may advertise about their new software to consumers who have a lot of software purchasing history. In addition, data mining may also help marketers in predicting which products their customers may be interested in buying. Through this prediction, marketers can surprise their customers and make the customer's shopping experience becomes a pleasant one.

Retail stores can also benefit from data mining in similar ways. For example, through the trends provide by data mining, the store managers can arrange shelves, stock certain items, or provide a certain discount that will attract their customers.

Banking/Crediting

Data mining can assist financial institutions in areas such as credit reporting and loan information. For example, by examining previous customers with similar attributes, a bank can estimated the level of risk associated with each given loan. In addition, data mining can also assist credit card issuers in detecting potentially fraudulent credit card transaction. Although the data mining technique is not a 100% accurate in its prediction about fraudulent charges, it does help the credit card issuers reduce their losses.

Law enforcement

Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.

Researchers

Data mining can assist researchers by speeding up their data analyzing process; thus, allowing those more time to work on other projects.

10.6 Disadvantages of Data Mining Privacy Issues

Personal privacy has always been a major concern in this country. In recent years, with the widespread use of Internet, the concerns about privacy have increase

NOTES

tremendously. Because of the privacy issues, some people do not shop on Internet. They are afraid that somebody may have access to their personal information and then use that information in an unethical way; thus causing they harm.

Although it is against the law to sell or trade personal information between different organizations, selling personal information have occurred. For example, according to Washing Post, in 1998, CVS had sold their patient's prescription purchases to a different company. In addition, American Express also sold their customers' credit care purchases to another company. What CVS and American Express did clearly violate privacy law because they were selling personal information without the consent of their customers? The selling of personal information may also bring harm to these customers because you do not know what the other companies are planning to do with the personal information that they have purchased.

Security issues

Although companies have a lot of personal information about us available online, they do not have sufficient security systems in place to protect that information. For example, recently the Ford Motor credit company had to inform 13,000 of the consumers that their personal information including Social Security number, address, account number and payment history were accessed by hackers who broke into a database belonging to the Experian credit reporting agency. This incidence illustrated that companies are willing to disclose and share your personal information, but they are not taking care of the information properly. With so much personal information available, identity theft could become a real problem.

Misuse of information/inaccurate information

Trends obtain through data mining intended to be used for marketing purpose or for some other ethical purposes, may be misused. Unethical businesses or people may used the information obtained through data mining to take advantage of vulnerable people or discriminated against a certain group of people. In addition, data mining technique is not a 100 percent accurate; thus mistakes do happen which can have serious consequence

10.7 Data Mining Processes

Data mining can do basically six tasks. The first three are all examples of **directed data mining**, where the goal is to use the available data to build a model that describes one particular variable of interest in terms of the rest of the available data. For example, analyzing bankruptcy, the target variable is a binary variable that describes if a client was declared on bankruptcy or not. In directed data mining, we try to find patterns that will make that variable have that value: 0 or 1. The next three tasks are examples of **undirected data mining** where no variable is singled out as a target and the goal is to establish some relationship among all the variables.

These are the types of information can be obtained by data mining, summarizing from two different sources: Turban & Aronson and Berry & Linoff :

- **Classification**: consists of examining the features of a newly presented object and assigning to it a predefined class or group. The task is to build a model that can be applied to unclassified data in order to classify it using the defined characteristics of a certain group (e.g., classifying credit applicants as low, medium of high risk).

NOTES

- **Estimation**: Given some input data, we use estimation to come up with a value for some unknown continuous variable such as income, height, or credit card balance. (e.g. a bank trying to decide to whom they should offer a home equity loan based on the probability that the person will respond positively to an offer).
- **Prediction**: records are classified according to some predicted future behavior or estimated future values based on patterns within large sets of data (e.g. demand forecasting or predicting which customers will leave within the next six months).
- **Association**: identifies relationships between events that occur at one time, determines which things go together (e.g., the contents of a shopping basket: fruits with snacks)
- **Clustering**: identifies groups of items that share a particular characteristic segmenting a diverse group into a number of more similar subgroups or clusters. Clustering differs from classification in that it does not rely on predefined classes or characteristics for each group. (e.g. as a first step in a market segmentation effort, we can divide the customer base into clusters of people with similar buying habits, and ten ask what kind of promotion works best for each cluster or group).
- **Description and Visualization**: the purpose is to describe what is going on in a complicated database in a way that increases our understanding of the people, products, or processes that produced the data in the firs place. A good description suggests where to start looking for an explanation. (e.g., repeat visits to a supermarket).

Data mining is conducted against data accumulated in OLTP repositories, data warehouses, data marts and archived data. The steps for data mining follows the following pattern:

- Data extraction
- Data cleansing
- Modeling data
- Applying data mining algorithm
- Pattern discovery
- Data visualization

Data extraction and data cleansing can be eased with good data lifecycle management policies. Very often a data warehousing project will ensure that data extraction and meta-data standards are pre-defined in an organization.

Data models for operational and archived data are different from data mining models. Data stored referentially in operational systems are designed for transactional speed.

In data mining a unified table view is created where data of interest is stored. Most data mining vendors offer the ability to extract data from repositories and transfer to the data mining database.

Not all of the data found in the data mining table view will have relevance. Other data may hold hidden patterns that can be discovered after relevancy is captured, often with external data sources.

NOTES

10.8 Data Mining Techniques

The most commonly used techniques in data mining are (By Thearling):

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset . Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

10.9 Data Mining tools

Data mining software is based in mathematical algorithms and statistics. Developers have been working in data mining software tools to make them more users friendly and the different products available on the market have advantages and disadvantages basically related to their interface and available techniques. Today, the market offers a variety of products. The most preferred tools are:

- Clementine
- SAS EM (Enterprise Miner)
- EasyMiner
- Oracle's Darwin
- Megaputer
- GainSmarts and
- SGI's MineSet

10.10 Conducting a Data Mining Project

The following are the necessary steps for conducting a data-mining project:

- Understand the purpose of the project together with managers of the organization. Define the objectives and how the results will be applied in the organization.
- Create the dataset that is going to be used. This may include creating a data warehouse and selecting the databases that probably need to be joined.
- Define the tools that are going to be used: Hardware (e.g. powerful computers with a high speed processor), Software (e.g. Database management software and data mining tool like Clementine, SAS or SPSS) and most important, people who are going to work in the project.
- Analyze the dataset together with the managers. It is important to define each variable. Sometimes the names of the variables do not have the real meaning or they can have different interpretations.
- Check if data is ready to apply data mining methods. That means for example, to clean the data, and input missing data. Create dummy variables if necessary.

- Reduce the size of the dataset. For example, summarize some variables that can have same meaning. Reduce the format for binary variables; they can be reduced to one bit instead of one byte. That reduces the size of the file, so it can be easier to manipulate.
- Identify the relevance of variables. For instance, Wal-mart may require primarily transaction variables instead frequency of transactions.

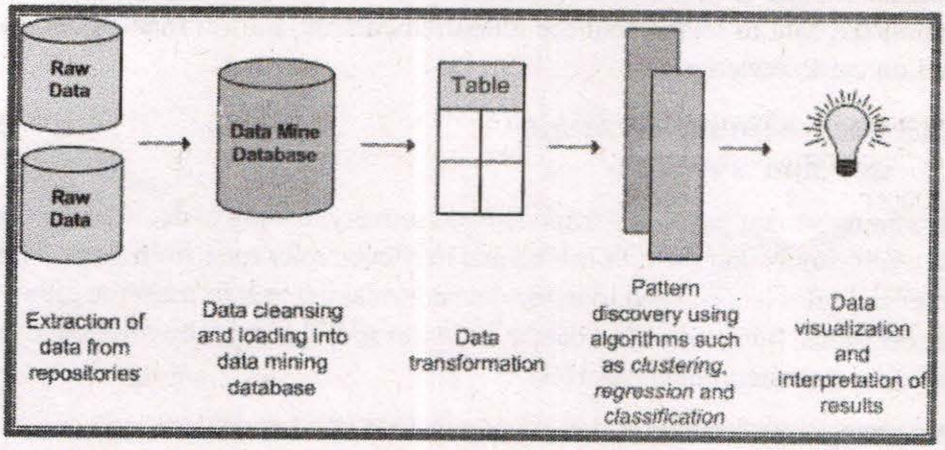


Figure 10.1 : The Steps in Data Mining

- Take sample data and identify the best rules for decomposition of data. For example, to define what percent of available data is going to be used for learning, validation and testing data.
- Use data mining software tools and data mining techniques to search for the models that have better performance.
- Test the results in a different set of data.
- Develop an interpretation of the results, so they can be understandable not only by the analysts and better understandable by the managers.
- Use the results in the management decision-making process. For example, choosing a better packing strategy after mining CRM data.

10.11 Data Mining Issues

10.11.1 Privacy

Data mining is both a powerful and profitable tool, but it poses challenges to the protection of individual privacy. Many critics wonder whether companies should be allowed to collect such detailed information about individuals. Consumer groups, privacy activists and persons associated with the Federal Trade Commission are concerned about the privacy issues that arise with the vast amount of information that companies acquire when data mining. In order to protect privacy of customers, some measures were taken in the Prepared Statement of The Federal Trade Commission.

10.11.2 . Database Requirements

In order to apply data mining techniques they require to be fully integrated with a data warehouse. The data warehouse grows and the organization can continually mine the best practices and apply them to future decisions. However, if the data

NOTES

warehouse is not ready to mine, preparing data could take big percentage of the available resources for the project.

10.11.3 Result Interpretation

Sometimes the results that are obtained after mining data can be difficult to interpret. data mining has a variety of tools to make it easier to interpret and explain, but an intelligent human is still required to (a) structure de data in the first place; (b) interpret the data to understand the identified pattern; and (c) make a decision based on the knowledge .

10.12 Data Mining strategic decision support

Data mining's main purpose is knowledge discovery leading to decision support. Data mining tools find patterns in data and may even infer rules from them. These patters and rules can be used to guide decision-making and forecast the effect of these decisions. It is proved that data mining can speed analysis by focusing attention on the most important variables.

Some strategic applications of data mining include (By Laudon):

- Identifying individuals or organization most likely to respond to a direct mailing.
- Determining which products or services are commonly purchased together.
- Predicting which customers are likely to switch to competitors.
- Identifying which transactions are likely to be fraudulent.
- Identifying common characteristics of customers who purchase the same product.
- Predicting what each visitor to a Web site is most interesting in seeing.

Because competitor plans are never fully known ahead of time, it is essential to leverage all available information about customers' reactions to potential offers. Sometimes the information comes from the specific markets where more aggressive marketing initiatives are anticipated; sometimes the information comes from sources external to that market. In either case, effective knowledge management (KM) requires seeking diverse data about customer and competitor activities and capitalizing on these data.

Creating new knowledge for competitive situations requires openness to an enlarged array of data sources and the ability to capitalize on developments in data modeling and mining. Interpretation and analysis that might otherwise lead only to directional guidance can result in more specific decision parameters.

Data mining can also be used to locate individual customers with specific interests or determine the interests of a specific group of customers.

10.13 Limitations of Data Mining

While data mining products can be very powerful tools, they are not self sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel related , rather than technology-related.

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations).

10.14 Knowledge Discovery

Knowledge Discovery is a concept that describes the process of automatically searching large volumes of data for patterns that can be considered knowledge *about* the data. The most well-known application of *Knowledge Discovery* is data mining also known as Knowledge Discovery in Databases (KDD).

Another promising application of *Knowledge discovery* is in the area of software modernization which involves understanding existing software artifacts. This process is related to a concept of reverse engineering. Usually the knowledge obtained from existing software is presented in the form of models to which specific queries can be made when necessary. An entity relationship is a frequent format of representing knowledge obtained from existing software. Object Management Group (OMG) developed specification Knowledge Discovery Metamodel (KDM) which defines an ontology for the software assets and their relationships for the purpose of performing knowledge discovery of existing code. Knowledge discovery from existing software systems, also known as software mining is closely related to data mining, since existing software artifacts contain enormous business value, key for the evolution of software systems. Instead of mining individual data sets, software mining focuses on metadata, such as database schemas.

Knowledge Discovery is the process of *deriving* knowledge from the input data. Some forms of Knowledge Discovery create abstractions of the input data. In some scenarios, the *knowledge* obtained through the process of Knowledge Discovery becomes further *data* that can be used for continuous discovery.

Knowledge Discovery is a complex topic that can be further categorized according to

- What kind of *data* is searched; and
- In what form is the result of the search represented.

NOTES

Knowledge discovery is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data". Under their conventions, the knowledge discovery process takes the raw results from **data mining** (the process of extracting trends or patterns from data) and carefully and accurately transforms them into useful and understandable information.

The most sophisticated definition is one according to Fayyad et al., where authors have determined that knowledge discovery in databases is interactive and iterative process with several steps and data mining is a part of this process. Process of KDD is defined as:

The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

The terms of above definition are explained as follows:

Pattern

- models or structure in data (traditional sense)
- expression in some language describing a subset of the data or a model applicable to that subset (data comprises a set of facts)

Process

- implies there are many steps repeated in multiple iterations

Nontrivial (process)

- it must involve search for structure, models, patterns, or parameters

Valid

- discovered patterns should be valid for new data with some degree of certainty

Novel

- at least to the system and preferably to the user

Potentially useful

- for the user or task

Understandable

- discovered patterns should be understandable - if not immediately, then after some post processing.

10.14.1 KDD Process

According to definition above, the KDD is an interactive and iterative process. The first two steps of the KDD process, namely *task discovery* and *data discovery,* produce the first input (goal of the KDD process). The next steps in the KDD process are *data cleaning, model development, data analysis* and *output generation*.

- **Task Discovery** is one of first steps of KDD. Client has to state the problem or goal, which often seems to be clear. Further investigation is recommended such as to get acquainted with customer's organisation after spending some time at the place and to sift through the raw data (to understand its form, content, organizational role and sources of data). Then the real goal of the discovery will be found.
- **Data Discovery** is complementary to step of task discovery. In the step of data discovery, we have to decide whether quality of data is satisfactory for the goal (what data does or does not cover).

- **Domain Model** plays an important role in the KDD process, though it often remains in the mind of the expert. A data dictionary, integrity constraints and various forms of metadata from the DBMS can possibly contribute to retrieval of the background knowledge for the KDD purposes as well as some analysis techniques. Those can take advantage of formally represented knowledge when fitting data to a model (for example ML techniques such as explanation-based learning integrated with inductive learning techniques).
- **Data Cleaning** is often necessary though it may happen that something removed by cleaning can be indicator of some interesting domain phenomenon (outlier or key data point?). Analyst's background knowledge is crucial in data cleaning provided by comparisons of multiple sources. Other way is to clean data before loaded into database by editing procedures. Recently, the data for KDD are coming form data warehouses which contain data already cleaned on some way.

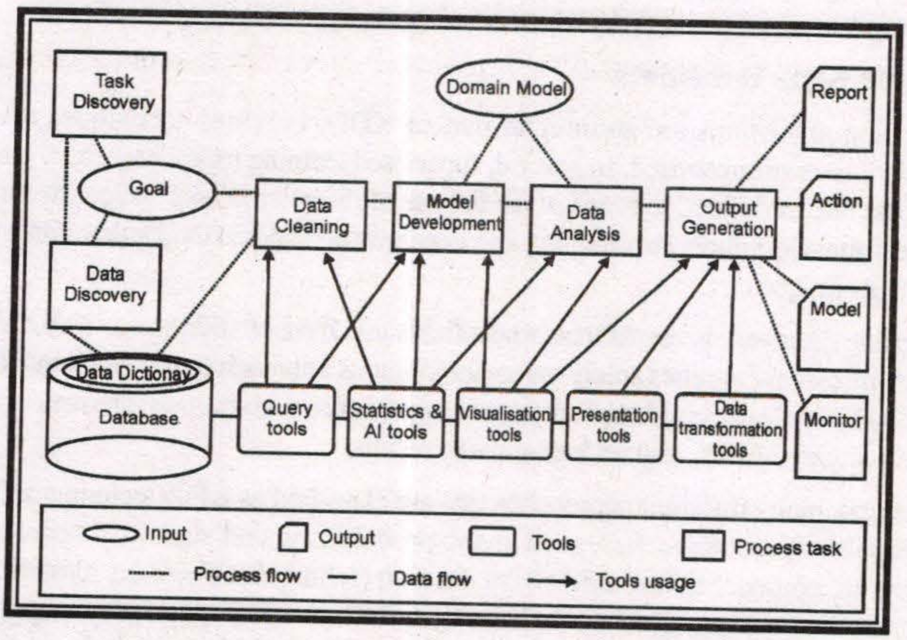


Figure 10.2 : The Framework of KDD Process

- **Model Development** is an important phase of KDD that must precede actual analysis of the data. Interaction with the data leads analysts to formation of hypothesis (it is often based on experiènce and background knowledge). Sub-processes of model development are:
  - *Data segmentation* (unsupervised learning techniques, for example clustering);
  - *Model selection* (choosing the best type of model after exploring several different types);
  - *Parameter selection* (parameters of chosen model).
- **Data Analysis** is in general an ambition to understand why certain groups of entities are behaving on the way they do, it is search for laws or rules of such behaviour. As first should be analyzed those parts where such a groups are already identified. Sub-processes in data analysis are:
  - *Model specification* - some formalism is used to denote specific model;

NOTES

- *Model fitting* - when necessary the specific parameters are determined (in some cases the model is independent from data in other cases the model has to be fitted to training data);
- *Evaluation* - model is evaluated against the data;
- *Model refinement* - model is refined in iterations according to the evaluation results.

  As mentioned above the model development and data analysis are complementary so it often leads to oscillation between those two steps.

- **Output Generation** - output can be in various forms. The simplest form is a report with analysis results. The other, more complicated forms, are graphs or in some cases it is desirable to obtain action descriptions which might be taken directly as outputs. Or there should be a monitor as the output, which should trigger an alarm or action under some certain condition. Output requirements might determine task of designed KDD application.

10.14.2 KDD Techniques

Learning algorithms are an integral part of KDD. Learning techniques may be supervised or unsupervised. In general, supervised learning techniques enjoy a better success rate as defined in terms of usefulness of discovered knowledge. According to, learning algorithms are complex and generally considered the hardest part of any KDD technique.

Machine discovery is one of the earliest fields that have contributed to KDD. While machine discovery relies solely on an autonomous approach to information discovery, KDD typically combines automated approaches with human interaction to assure accurate, useful, and understandable results.

There are many different approaches that are classified as KDD techniques. There are quantitative approaches, such as the probabilistic and statistical approaches. There are approaches that utilize visualization techniques. There are classification approaches such as Bayesian classification, inductive logic, data cleaning/pattern discovery, and decision tree analysis. Other approaches include deviation and trend analysis, genetic algorithms, neural networks, and hybrid approaches that combine two or more techniques.

Because of the ways that these techniques can be used and combined, there is a lack of agreement on how these techniques should be categorized. For example, the Bayesian approach may be logically grouped with probabilistic approaches, classification approaches, or visualization approaches. For the sake of organization, each approach described here is included in the group that it seemed to fit best. However, this selection is not intended to imply a strict categorization.

The major techniques are explained below:

- **Probabilistic Approach**

  This family of KDD techniques utilizes graphical representation models to compare different knowledge representations. These models are based on probabilities and data independencies. They are useful for applications involving uncertainty and applications structured such that a probability may be assigned to each "outcome" or bit of discovered knowledge. Probabilistic techniques may be used in diagnostic systems and in planning and control systems. Automated probabilistic tools are available both commercially and in the public domain.

- **Statistical Approach**

  The statistical approach uses rule discovery and is based on data relationships. An "inductive learning algorithm can automatically select useful join paths and attributes to construct rules from a database with many relations". This type of induction is used to generalize patterns in the data and to construct rules from the noted patterns. Online analytical processing (OLAP) is an example of a statistically-oriented approach. Automated statistical tools are available both commercially and in the public domain.

  An example of a statistical application is determining that all transactions in a sales database that start with a specified transaction code are cash sales. The system would note that of all the transactions in the database only 60% are cash sales. Therefore, the system may accurately conclude that 40% are collectibles.

- **Classification Approach**

  Classification is probably the oldest and most widely-used of all the KDD approaches. This approach groups data according to similarities or classes. There are many types of classification techniques and numerous automated tools available.

  The **Bayesian Approach** to KDD "is a graphical model that uses directed arcs exclusively to form a directed acyclic graph". Although the Bayesian approach uses probabilities and a graphical means of representation, it is also considered a type of classification.

  Bayesian networks are typically used when the uncertainty associated with an outcome can be expressed in terms of a probability. This approach relies on encoded domain knowledge and has been used for diagnostic systems. Other pattern recognition applications, including the Hidden Markov Model, can be modeled using a Bayesian approach. Automated tools are available both commercially and in the public domain.

  Pattern Discovery and Data Cleaning is another type of classification that systematically reduces a large database to a few pertinent and informative records. If redundant and uninteresting data is eliminated, the task of discovering patterns in the data is simplified. This approach works on the premise of the old adage, "less is more". The pattern discovery and data cleaning techniques are useful for reducing enormous volumes of application data, such as those encountered when analyzing automated sensor recordings. Once the sensor readings are reduced to a manageable size using a data cleaning technique, the patterns in the data may be more easily recognized. Automated tools using these techniques are available both commercially and in the public domain.

  The **Decision Tree Approach** uses production rules, builds a directed acyclical graph based on data premises, and classifies data according to its attributes. This method requires that data classes are discrete and predefined.The primary use of this approach is for predictive models that may be appropriate for either classification or regression techniques. Tools for decision tree analysis are available commercially and in the public domain.

NOTES

NOTES

- **Deviation and Trend Analysis**

  Pattern detection by filtering important trends is the basis for this KDD approach. Deviation and trend analysis techniques are normally applied to temporal databases. A good application for this type of KDD is the analysis of traffic on large telecommunications networks.

  AT&T uses such a system to locate and identify circuits that exhibit deviation (faulty behavior). The sheer volume of data requiring analysis makes an automated technique imperative. Trend-type analysis might also prove useful for astronomical and oceanographic data, as they are time-based and voluminous. Public domain tools are available for this approach.

- **Other Approaches**

  Neural networks may be used as a method of knowledge discovery. Neural networks are particularly useful for pattern recognition, and are sometimes grouped with the classification approaches. There are tools available in the public domain and commercially. Genetic algorithms, also used for classification, are similar to neural networks although they are typically considered more powerful. There are tools for the genetic approach available commercially.

  A hybrid approach to KDD combines more than one approach and is also called a multi-paradigmatic approach. Although implementation may be more difficult, hybrid tools are able to combine the strengths of various approaches. Some of the commonly used methods combine visualization techniques, induction, neural networks, and rule-based systems to achieve the desired knowledge discovery. Deductive databases and genetic algorithms have also been used in hybrid approaches. There are hybrid tools available commercially and in the public domain.

10.15 Knowledge Discovery Metamodel

Knowledge Discovery Metamodel *(***KDM***)* is publicly available specification from the Object Management Group (OMG). KDM is a common intermediate representation for existing software systems and their operating environments, that defines common metadata required for deep semantic integration of Application Lifecycle Management tools. KDM was designed as the OMG's foundation for software modernization, IT portfolio management and software assurance. KDM uses OMG's Meta-Object Facility to define an XMI interchange format between tools that work with existing software as well as an abstract interface (API) for the next-generation assurance and modernization tools. KDM standardizes existing approaches to knowledge discovery in software engineering artifacts, also known as software mining.

The goal of KDM is to ensure interoperability between tools for maintenance, evolution, assessment and modernization. KDM is defined as a metamodel that can be also viewed as an ontology for describing the key aspects of knowledge related to the various facets of enterprise software. KDM support means investment into the KDM ecosystem - a growing open-standard based cohesive community of tool vendors, service providers, and commercial components.

KDM represents entire enterprise software systems, not just code. KDM is a wide-spectrum entity-relationship representation for describing existing software. KDM represents structural and behavior elements of existing software systems. The key concept of KDM is a *container*: an entity that owns other entities. This allows KDM to represent existing systems at various degrees of granularity.

NOTES

KDM defines precise semantic foundation for representing behavior, the so-called *micro-KDM*. It provides a high-fidelity intermediate representation which can be used, for example, for performing static analysis of existing software systems. micro-KDM is similar in purpose to a Virtual machine for KDM, although KDM is not an executable model, or a constraint model, but a representation of existing artifacts for analysis purposes.

KDM facilitates incremental analysis of existing software systems, where the initial KDM representation is analyzed and more pieces of knowledge are extracted and made explicit as KDM to KDM transformation performed entirely within the KDM technology space. The steps of the knowledge extraction process can be performed by tools, and may involve the analyst.

KDM is the uniform language- and platform- independent representation. Its extensibility mechanism allows addition of domain-, application- and implementation-specific knowledge.

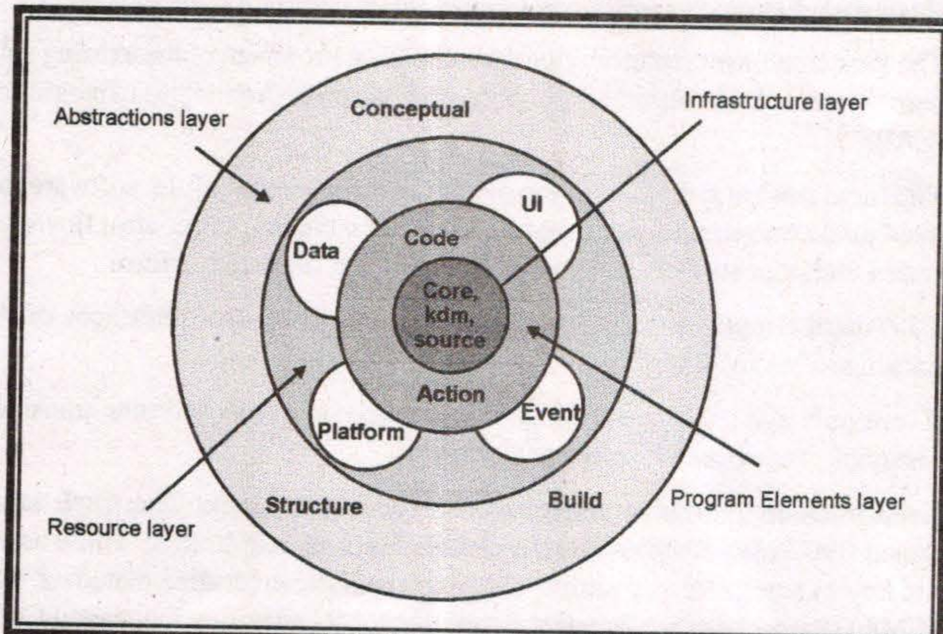


Figure 10.3 : The Architecture of KDM

10.15.1 Architecture of KDM

Knowledge Discovery Metamodel consists of 12 packages arranged into 4 layers. KDM packages are arranged into the following four layers:

- **Infrastructure Layer**

  The KDM Infrastructure Layer consists of the **Core**, **kdm**, and **Source** packages which provide a small common core for all other packages, the inventory model of the artifacts of the existing system and full traceability between the meta-model elements as links back to the source code of the artifacts, as well as the uniform extensibility mechanism. The Core package determines several of

NOTES

patterns that are reused by other KDM packages. Although KDM is a meta-model that uses Meta-Object Facility, there is an alignment between the KDM Core and Resource Description Framework (RDF).

- **Program Elements Layer**

  The Program Elements Layer consists of the Code and Action packages.

  The **Code package** represents programming elements as determined by programming languages, for example data types, procedures, classes, methods, variables, etc. This package is similar in purpose to the Common Application Meta-model (CAM) from another OMG specification, called Enterprise Application Integration (EAI). KDM Code package provides greater level of detail and is seamlessly integrated with the architecturally significant views of the software system. Representation of data types in KDM is aligned with ISO standard ISO/IEC 11404.

  The **Action package** captures the low level behavior elements of applications, including detailed control- and data flow between statements. Code and Action package in combination provide a high-fidelity intermediate representation of each component of the enterprise software system

- **Resource Layer**

  The Resource Layer represents the operational environment of the existing software system. It is related to the area of Enterprise Application Integration (EAI).

  Platform package represents the operating environment of the software, related to the operating system, middleware, etc. including the control flows between components as they are determined by the runtime platform .

  UI package represents the knowledge related to the user interfaces of the existing software system .

  Event package represents the knowledge related to events and state-transition behavior of the existing software system .

  Data package represents the artifacts related to persistent data, such as indexed files, relational databases, and other kinds of data storage. These assets are key to enterprise software as they represent the enterprise metadata. The KDM Data package is aligned with another OMG specification, called Common Warehouse Metamodel (CWM).

- **Abstractions Layer**

  The Abstraction Layer represents domain and application abstractions.

  Conceptual package represent business domain knowledge and business rules, insofar as this information can be mined from existing applications. These packages are aligned with another OMG specification, called Semantics of Business Vocabulary and Rules (SBVR)

  Structure package describes the meta-model elements for representing the logical organization of the software system into subsystems, layers and components

NOTES

Build package represents the engineering view of the software system

10.16 On Line Analytical Processing

On Line Analytical Processing, or OLAP, is an approach to quickly providing answers to analytical queries that are multidimensional in nature. OLAP is part of the broader category business intelligence, which also includes Extract transform load (ETL), relational reporting and data mining. The typical applications of OLAP are in business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas. The term OLAP was created as a slight modification of the traditional database term OLTP (**On Line Transaction Processing**).

Databases configured for OLAP employ a multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time. Nigel Pendse has suggested that an alternative and perhaps more descriptive term to describe the concept of OLAP is **Fast Analysis of Shared Multidimensional Information** (FASMI). They borrow aspects of navigational databases and hierarchical databases that are speedier than their relational kin.

The output of an OLAP query is typically displayed in a matrix (or pivot) format. The dimensions form the row and column of the matrix; the measures, the values.

10.17 OLAP Types

OLAP systems have been traditionally categorized using the following taxonomy:

10.17.1 MOLAP

MOLAP stands for Multidimensional Online Analytical Processing. **MOLAP** is the 'classic' form of OLAP and is sometimes referred to as just OLAP. MOLAP uses database structures that are generally optimal for attributes such as time period, location, product or account code. The way that each dimension will be aggregated is defined in advance by one or more hierarchies.

MOLAP is an alternative to the ROLAP (Relational OLAP) technology. While both ROLAP and MOLAP analytic tools are designed to allow analysis of data through the use of a multidimensional data model, MOLAP differs significantly in that it requires the pre-computation and storage of information in the cube - the operation known as *processing*. MOLAP stores this data in an optimized multi-dimensional array storage, rather than in a relational database (i.e. in ROLAP)

MOLAP Advantages

- Fast query performance due to optimized storage, multidimensional indexing and caching.
- Smaller on-disk size of data compared to data stored in relational database due to compression techniques.
- Automated computation of higher level aggregates of the data.
- It is very compact for low dimension data sets.
- Array model provides natural indexing .
- Effective data extract achieved through the pre-structuring of aggregated data.

NOTES

MOLAP Disadvantages

- The processing step (data load) can be quite lengthy, especially on large data volumes. This is usually remedied by doing only incremental processing, i.e., processing only the data which has changed (usually new data) instead of re-processing the entire data set.
- MOLAP tools traditionally have difficulty querying models with dimensions with very high cardinality (i.e., millions of members).
- Certain MOLAP tools (e.g., Essbase) have difficulty updating and querying models with more than ten dimensions. This limit differs depending on the complexity and cardinality of the dimensions in question. It also depends on the number of facts or measures stored. Other MOLAP tools (e.g., Microsoft Analysis Services or Applix TM1) can handle hundreds of dimensions.
- MOLAP approach introduces data redundancy.

MOLAP Trends

Most commercial OLAP tools now use a "Hybrid OLAP" (HOLAP) approach, which allows the model designer to decide which portion of the data will be stored in MOLAP and which portion in ROLAP.

MOLAP Products

Examples of commercial products that use MOLAP are Microsoft Analysis Services, Essbase, MIS Alea and TM1. There is also an open source MOLAP server Palo.

10.17.2 ROLAP

ROLAP stands for Relational Online Analytical Processing. **ROLAP** works directly with relational databases. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information. Depends on a specialized schema design.

ROLAP is an alternative to the MOLAP (Multidimensional OLAP) technology. While both ROLAP and MOLAP analytic tools are designed to allow analysis of data through the use of a multidimensional data model, ROLAP differs significantly in that it does not require the pre-computation and storage of information. Instead, ROLAP tools access the data in a relational database and generate SQL queries to calculate information at the appropriate level when an end user requests it. With ROLAP, it is possible to create additional database tables (*summary tables* or *aggregations*) which summarize the data at any desired combination of dimensions.

While ROLAP uses a relational database source, generally the database must be carefully designed for ROLAP use. A database which was designed for OLTP will not function well as a ROLAP database. Therefore, ROLAP still involves creating an additional copy of the data. However, since it is a database, a variety of technologies can be used to populate the database.

Advantages of ROLAP

- ROLAP is considered to be more scalable in handling large data volumes, especially models with dimensions with very high cardinality (ie. millions of members).

NOTES

- With a variety of data loading tools available, and the ability to fine tune the ETL code to the particular data model, load times are generally much shorter than with the automated MOLAP loads.
- The data is stored in a standard relational database and can be accessed by any SQL reporting tool (the tool does not have to be an OLAP tool).
- ROLAP tools are better at handling *non-aggregatable facts* or (e.g. textual descriptions). MOLAP tools tend to suffer from slow performance when querying these elements.
- By decoupling the data storage from the multi-dimensional model, it is possible to successfully model data that would not otherwise fit into a strict dimensional model.
- The ROLAP approach can leverage database authorization controls such as row-level security; whereby the query results will be filtered depending on a preset criteria applied for example to a given user or group of users (SQL WHERE clause).

Disadvantages of ROLAP

- There is a general consensus in the industry that ROLAP tools have slower performance than MOLAP tools.
- The loading of *aggregate tables* must be managed by custom ETL code. The ROLAP tools do not help with this task. This means additional development time and more code to support.
- When the step of creating aggregate tables is skipped, the query performance then suffers because the larger detailed tables must be queried. This can be partially remedied by adding additional aggregate tables, however it is still not practical to create aggregate tables for all combinations of dimensions/attributes.
- ROLAP relies on the general purpose database for querying and caching, and therefore several special techniques employed by MOLAP tools are not available (such as special hierarchical indexing). However, modern ROLAP tools take advantage of latest improvements in SQL language such as CUBE and ROLLUP operators, DB2 Cube Views, as well as other SQL OLAP extensions. These SQL improvements can mitigate the benefits of the MOLAP tools.
- Since ROLAP tools rely on SQL for all of the computations, they are not suitable when the model is heavy on calculations which don't translate well into SQL. Examples of such models include budgeting, allocations, financial reporting and other scenarios.

ROLAP Trends

The undesirable trade-off between additional ETL cost and slow query performance has ensured that most commercial OLAP tools now use a "Hybrid OLAP" (HOLAP) approach, which allows the model designer to decide which portion of the data will be stored in MOLAP and which portion in ROLAP.

ROLAP Products

Examples of commercial products using ROLAP include Microsoft Analysis Services, Microstrategy and Business Objects, Oracle BI (the former Siebel Analytics). There is also an open source ROLAP server - Mondrian.

NOTES

10.17.3 HOLAP

HOLAP (**H**ybrid **Onl**ine **A**nalytical **P**rocess) is a combination of ROLAP and MOLAP which are other possible implementation of OLAP. There is no clear agreement across the industry as to what constitutes "Hybrid OLAP", except that a database will divide data between relational and specialized storage. For example, for some vendors, a HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more-aggregate or less-detailed data.

HOLAP allows to store part of the data in the MOLAP store and another part of the data in ROLAP store. The degree of control that cube designer has over this partitioning varies from product to product:

- **Vertical partitioning**

  In this mode HOLAP stores *aggregations* in MOLAP for fast query performance, and detailed data in ROLAP to optimize time of cube *processing*.

- **Horizontal partitioning**

  In this mode HOLAP stores some slice of data, usually the more recent one (i.e. sliced by Time dimension) in MOLAP for fast query performance, and older data in ROLAP.

HOLAP Products

Examples of commercial products which support HOLAP storage mode are Microsoft Analysis Services, MicroStrategy and SAP AG BI Accelerator.

Comparison

Each type has certain benefits, although there is disagreement about the specifics of the benefits between providers.

Some MOLAP implementations are prone to database explosion. Database explosion is a phenomenon causing vast amounts of storage space to be used by MOLAP databases when certain common conditions are met: high number of dimensions, pre-calculated results and sparse multidimensional data. The typical mitigation technique for database explosion is not to materialize all the possible aggregation, but only the optimal subset of aggregations based on the desired performance vs. storage trade off.

MOLAP generally delivers better performance due to specialized indexing and storage optimizations. MOLAP also needs less storage space compared to ROLAP because the specialized storage typically includes compression techniques.

ROLAP is generally more scalable. However, large volume pre-processing is difficult to implement efficiently so it is frequently skipped. ROLAP query performance can therefore suffer.

Since ROLAP relies more on the database to perform calculations, it has more limitations in the specialized functions it can use.

HOLAP encompasses a range of solutions that attempt to mix the best of ROLAP and MOLAP. It can generally pre-process quickly, scale well, and offer good function support.

NOTES

10.17.4 Other Types

The following acronyms are also used sometimes, although they are not as widespread as the ones above:

- **WOLAP** - Web-based OLAP
- **DOLAP** - Desktop OLAP
- **RTOLAP** - Real-Time OLAP
- **SOLAP** - Spatial OLAP

10.17.5 OLAP APIs and Query languages

Unlike relational databases - which had SQL as the standard query language, and wide-spread APIs such as ODBC, JDBC and OLEDB - there was no such unification in the OLAP world for a long time. The first real standard API was OLE DB for OLAP (ODBO) specification from Microsoft which appeared in 1997 and introduced the MDX query language. Several OLAP vendors - both server and client - adopted it. In 2001 Microsoft and Hyperion announced the XML for Analysis specification, which was endorsed by most of the OLAP vendors. Since this also used MDX as a query language, MDX became the de-facto standard in the OLAP world.

OLE DB for OLAP (ODBO)

OLE DB for OLAP (abbreviated **ODBO**) is a Microsoft published specification and an industry standard for multi-dimensional data processing. ODBO is the standard application programming interface (API) for exchanging metadata and data between an OLAP server and a client on a Windows platform. ODBO extends the ability of OLE DB to access multi-dimensional (OLAP) data stores.

ODBO is the most widely supported, multi-dimensional API to date. Platform specific to Microsoft Windows, ODBO was specifically designed for Online Analytical Processing (OLAP) systems by Microsoft as an extension to Object Linking and Embedding Database (OLE DB). ODBO uses Microsoft's Component Object Model.

ODBO permits independent software vendors (ISVs) and corporate developers to create a single set of standard interfaces that allow OLAP clients to access multi-dimensional data, regardless of vendor or data source. ODBO is currently supported by a wide spectrum of server and client tools.

10.18 Selecting an OLAP Application

While selecting an OLAP tool that can fit your organization's need requires a closer inspection of your requirements.

NOTES

- **Platform Choice**

  The first step in narrowing the list of vendors is the strength of the OLAP tool on the platform that your organization will be using. Some vendors tend to be only Microsoft centric or have products that are better supported on a particular platform.

- **Vendor History**

  There are many new OLAP vendors that have cropped up since the mid-nineties. Some of the newer companies have rewarding and rich OLAP tools that can be cost effective. Investigate how many installs the vendor has made and, if possible, interview some of the vendor's clients.

- **OLAP Cube Size and Transaction Speed**

  There are products that are optimized to your organization's needs. Fine tune your vendor list to your operational needs. How large is your cube? What would be a tolerable usage of storage space and transaction speed? How will the OLAP cube be created - i.e. HOLAP, MOLAP, ROLAP?

- **User Interface (GUI)**

  Will you prefer a thin client or desktop interface to interact with the OLAP cube? What are the trade-offs? List potential security concerns.

- **Consulting**

  How much consulting would be require to install the product? If changes are needed after installation, will this require the need of external consultants?

- **Integration to Database**

  How tightly integrated is the OLAP tool to your database? Can you perform SQL queries within the OLAP tool?

- **Integration to other Analysis Tools**

  Does your organization have or plan to have other analysis tools such as data mining, reporting and data visualization? How will the OLAP application integrate with the other analysis tools?

- **Price**

  It is often challenging to gauge the additional hidden cost from the actual purchase price. Some of the points above can help make a good financial forecast. Discuss with other users in your industry vertical or hire an independent consultant who is familiar with the process.

Sample list of OLAP vendors are Hyperion, Business Objects, Cognos, Oracle, SAS, IBM.

SUMMARY

- Modern databases contain a huge amount of data. However raw data are hardly of any utility. To gain information and knowledge, data must be first analyzed and processed. The automated analysis of data in search of some information is called data mining.

NOTES

- In recent years, the need to extract knowledge automatically from very large databases has grown. In response, the closely related fields of knowledge discovery in databases (KDD) and data mining have developed processes and algorithms that attempt to intelligently extract interesting and useful information from vast amounts of raw data.
- *Knowledge discovery* is defined as "the non-trivial extraction of implicit, unknown, and potentially useful information from data". Under their conventions, the knowledge discovery process takes the raw results from *data mining* (the process of extracting trends or patterns from data) and carefully and accurately transforms them into useful and understandable information.
- Pattern detection by filtering important trends is the basis for this KDD approach. Deviation and trend analysis techniques are normally applied to temporal databases. A good application for this type of KDD is the analysis of traffic on large telecommunications networks.
- *Knowledge Discovery Metamodel (KDM)* is publicly available specification from the Object Management Group (OMG). KDM is a common intermediate representation for existing software systems and their operating environments, that defines common metadata required for deep semantic integration of Application Lifecycle Management tools. KDM was designed as the OMG's foundation for software modernization, IT portfolio management and software assurance.
- Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees).
- Data Mining is a non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data.
- Data extraction and data cleansing can be eased with good data lifecycle management policies.
- Data mining softwares are based in mathematical algorithms and statistics.
- Data mining's main purpose is knowledge discovery leading to decision support. Data mining tools find patterns in data and may even infer rules from them. These patters and rules can be used to guide decision-making and forecast the effect of these decisions.
- On Line Analytical Processing, or OLAP, is an approach to quickly providing answers to analytical queries that are multidimensional in nature. OLAP is part of the broader category business intelligence, which also includes Extract transform load (ETL), relational reporting and data mining.
- The typical applications of OLAP are in business reporting for sales, marketing, management reporting, business process management (BPM), budgeting and forecasting, financial reporting and similar areas. The term OLAP was created as a slight modification of the traditional database term OLTP (On Line Transaction Processing).

NOTES

- MOLAP stands for **M**ultidimensional **O**nline **A**nalytical **P**rocessing. **MOLAP** is the 'classic' form of OLAP and is sometimes referred to as just OLAP.
- ROLAP stands for **R**elational **O**nline **A**nalytical **P**rocessing. **ROLAP** works directly with relational databases. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information.
- HOLAP (**H**ybrid **O**nline **A**nalytical **P**rocess) is a combination of ROLAP and MOLAP which are other possible implementation of OLAP. There is no clear agreement across the industry as to what constitutes "Hybrid OLAP", except that a database will divide data between relational and specialized storage. For example, for some vendors, a HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more-aggregate or less-detailed data.
- *OLE DB for OLAP* (abbreviated ODBO) is a Microsoft published specification and an industry standard for multi-dimensional data processing. ODBO is the standard application programming interface (API) for exchanging metadata and data between an OLAP server and a client on a Windows platform. ODBO extends the ability of OLE DB to access multi-dimensional (OLAP) data stores.

EXERCISE

1. What is data mining ? How is it related to KDD ?
2. Describe different steps in the evolution of data mining.
3. What are different data mining processes ?
4. briefly explain the verification and discovery of data mining.
5. How many steps are involved in data mining ?
6. Describe different data mining techniques.
7. List the most popular data mining tools .
8. What are the necessary steps for conducting a data-mining project?
9. Explain different issues in data mining .
10. Explain how data mining supports strategic decisions ?
11. Describe limitations of data mining .
12. What is knowledge discovery ? Describe different KDD terms.

11 Mobile Commerce

NOTES

Chapter Includes :

11.1 Mobile Commerce

Today's market is moving quickly toward the promise of communications, commerce, and content available anytime, anywhere, and on any device. Service providers are not only network operators, but are evolving to become providers of rich, interactive media, productivity services, and retail experiences. Consumer expectations have also evolved to where the term wireless implies voice service and represents a complex infrastructure to which one connects in real time, from any place to be informed, entertained or engaged in visual or audio communication or to buy goods and services.

Recently a new method of commerce named "M-Commerce" is used. It is defined as: Performing commerce using mobile and handheld devices such as Personal Digital Assistants (PDAs), Mobile phone, Smartphone and other emerging mobile equipments equipped with Web-ready micro-browsers, through wireless technology, is called Mobile Commerce. It also involves the transfer of ownership or rights to use goods and services. It is Ubiquitous in nature for its services and therefore

NOTES

also known as M-Commerce or U-Commerce. In the beginning it was believed that M-Commerce will be at priori as a choice for digital commerce and will go beyond E-Commerce as message passing through wireless technology is faster, secure and scalable. It may well be factual for the Asia-Pacific as there are more mobile phone users than Internet users.

M-Commerce is one whose time has truly come. Because there have been so many recent advances in wireless technology the number of individuals who use mobile devices has increased rapidly. E-commerce is often conducted on these devices and this is also moving at a rapid pace. There are several new types of e-commerce transactions and many of these use wireless telecommunications networks and other technologies to conduct business through mobile devices. This has been called mobile commerce and is increasingly known as either m-commerce or mobile e-commerce. Many constraints and special characteristic related to mobile devices allow mobile commerce operating in an environment that is much different from transactions conducted over the Internet. There are so many more market opportunities with mobile commerce because it can be personalized to an individual and it is much more flexible

Different services affected by M-Commerce include; Financial services, Telecommunications, Service/Retail and Information services. These all are discussed below:

Financial services include mobile banking; when customers use their handheld devices to access their accounts and pay their bills, as well as brokerage services; in which stock quotes can be displayed and trading conducted from the same handheld device.

Telecommunications, in which service changes, bill payment and account reviews can all be conducted from the same handheld device. **Service/Retail,** as consumers are given the ability to place and pay for orders on-the-fly. **Information services** include the delivery of entertainment, financial news, sports figures and traffic updates.

M-commerce solutions should provide open door between consumer and their desired products with authentic purchase and provisioning. This scenario can be given by the following Figure 11.1.

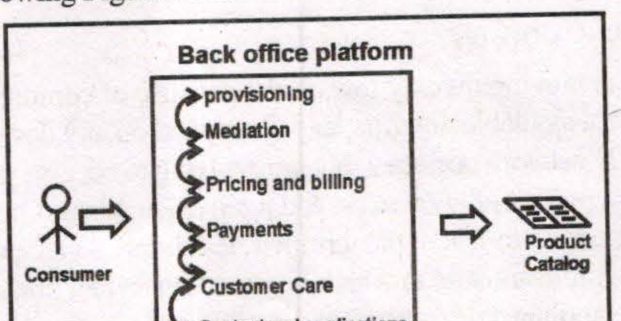


Figure 11.1 Back office platform of M-commerce

Internet and wireless technology has changed the way we perceive, the way we interact and the way we do commerce. The emergence of M-Commerce is resulting in highly technical disputes, doubts and complicated questions of law which have a direct bearing on the future growth of mobile commerce. As such the discipline of M-Commerce Law is the latest to be emerging in the field of Cyberlaw. This is a highly specialized and nische field of Cyberlaw which is developing through out the world, as till date there has no uniformity of approach as to the adoption of Wireless Application Protocol (WAP) as the universal M-Commerce technology platform and protocol.

NOTES

11.2 Technology for Mobile Commerce

The computing environment for mobile commerce is mobile and moves along with the user. Therefore the technical infrastructure required for mobile commerce is the same as required for mobile computing. Computing environment is defined as mobile if it supports one or more of the following characteristic:

- User, Device, Bearer and Network mobility: User should be able to move from one physical location to another location, from one device to another, from one bearer to another, from one network to another and use the same service. For example a user moves from US to UK and uses Internet to access its business application in the same way the user uses in the home office. In US he uses this service over GPRS (General packet Radio Service) whereas in UK he accesses it over GSM network. There can be the possibility that WAP (Wireless Application protocol) bearer is providing the services in US while SMS (Short Message Service) bearer in UK.
- Session, Service and Host mobility: A user should be able to move from one user-agent environment to another and from one service to another. For example for session mobility a user was using his service through a CDMA IX network. The user entered into the basement to park the car and got disconnected from his CDMA network. User goes to home office and starts using the desktop. The interrupted session in the CDMA device moves from the mobile device to the desktop computer. An example for service mobility is that when a user is writing a mail, to complete the mail user needs to refer to some other information like user opens another service and moves between them using the task bar. For the host mobility user device can be either client or server.

Technologies required for M-Commerce include the wireless technologies like required for mobile computing e.g. wireless communication technologies, Wireless Application Protocol (WAP), WAP programming model etc. The following sections discuss all these in detail:

11.3 Wireless Communication and its Generations

Wireless communication systems have been around for quite some time and their obvious use in garage door openers and cordless telephones has gone noticed until recently. The introduction of affordable priced wireless telephones has made them attractive form the general population. Their main usefulness is their capability to maintain the same contact number even if the user moves from one location to another. With the success of wireless telephony and messaging services like paging, wireless communication is beginning to be applied to realm of personal and business computing in the domain of local area networks. The Wireless Communication service can be provided by various methods such as radio, cellular and satellite based wireless systems.

11.3.1 Cellular systems

Early wireless systems had a high-power transmitter, covering the entire service area. This required a huge amount of power and was not suitable for many practical reasons. The cellular system replaced a large zone with a number of smaller hexagonal cells with a single powerful transmitting station called Base Station (BS)

NOTES

covering a fraction of the area. A Base Station consists of a base transceiver system (BTS) and a BS Controller (BSC). Both tower and antenna are a part of the BTS, while all associated electronics are contained in the BSC. The Home Location Register (HLR) and Visitor Location Register (VLR)are two sets of pointers that support mobility and the use of the same telephone numbers worldwide. HLR is located at the Mobile Switching Centre (MSC) where the mobile device like laptop called Mobile System (MS) is registered and where the initial home location for billing and access information is maintained. In simple words, any incoming call, based on the called number, is directed to HLR of the home MSC and then HLR redirects the call to the MSC (and the BS) where the MS is currently located. VLR basically contains information about all visiting MSs in that particular MSC area.

The cellular radio is grouped into the following three types systems:

1. **Increasing Capacity and widespread coverage for cordless telephones**

   Cellular technology is evolving toward taller base station antennas and larger cells to provide economical coverage of sparsely populated areas. The extreme case of this large cell evolution is the development of satellite systems.

2. **Decreasing Cell Size and Power levels for hand-held and vehicular cellular radio**

   Cellular cells are also moving towards smaller cells having low base station antennas, or even antennas within buildings to provide higher overall capacity and to provide better coverage to the lower power cellular pocket phones that are now in increasingly widespread usage. Cordless telephones have evolved from primarily home usage, then free roaming applications towards widespread "Universal" low power system aimed at pedestrians, exemplified by the Personal Communication Systems (PCS).

3. **Specialized wireless data systems**

   Specialized systems are also emerging although they are not yet as prominent as wireless telephony. These data-oriented wireless systems are also emerging through cellular packet data systems and smaller area coverage. High-data rate wireless LANs using infra-red and radio technology.

The cellular communication works through dynamic switching to facilitate a smooth transaction, the cells would be terminated as the user crosses the boundary of a cell. The cellular communication process works through a cycle consisting of Log on, Monitoring, Incoming calls and outgoing calls

Log On

Each cellular handset is assigned a unique identity or Numeric Arrangement Module (NAM). This identity is based on its home area. Messages are often sent to the handset through separate control channel to verify that it is operating with home area. As the mobile unit moves across cells, it must continuously send messages to the mobile telephone switching office to confirm its location so that new calls can be directed to that traffic area.

Monitoring

Once the handset is powered on, it monitors the control channels to get information on local paging channels. It selects an available channel and goes into an idle state. In the idle state it listens to the data being transmitted over that channel.

Incoming Calls

When an incoming calls is received by the MTSO (Mobile Telephone Switching Office), it sends a signal to all the cell in that traffic area. The set receives the signals and responds to the MTSO. The MTSO will inform the set to use a specific channel to receive the call, and the set will retune itself to the new frequency.

Outgoing Calls

To make a call, the user enters the telephone number and transmits it to the MTSO (Mobile Telephone Switching Office) by an available access channel. The MTSO either agrees to the handset request or asks the set to change frequency to another available channel.

NOTES

There are many issues involved in wireless communication and extensive signal processing is required before any signals are transmitted. The major steps are shown in Figure 11.2 and 11.3. Many of the signal processing operations are beyond the scope of this book so are not discussed here.

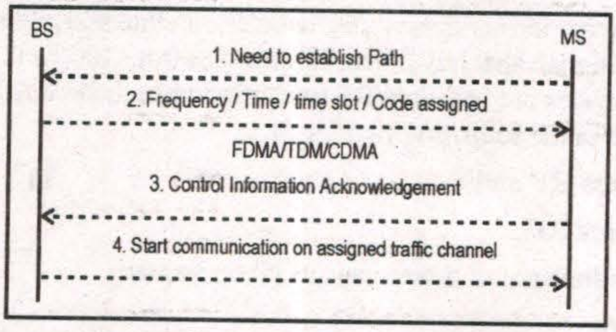


Figure 11.2 Steps for a call setup from MS to BS

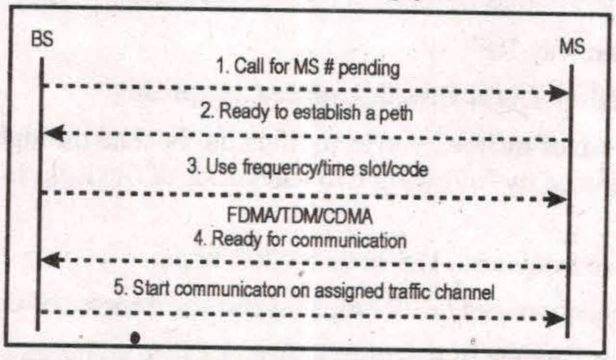


Figure 11.3 Steps for a call setup from BS to MS

11.3.2 Satellite systems

Satellite system networks have been in use for several decades. Satellites, which are far away from the surface of the earth, can cover large area, with several satellite beams being controlled and operated by one satellite. Large areas can be covered due to the rotation of satellites around the earth. Satellite networks facilitate global communications and cater to thousands of locations across the globe. Though it's possible to develop an equivalent land-based network using leased lines, local telephone company facilities, and/or public data networks' yet logistics of constructing, managing and maintaining such a network is a massive task. Satellite networks are quite useful in any situation where data need to be dispersed to or gathered from many remote nodes, and where end-to-end delay is not a primary concern. Satellite networks are divided into two types:

(i) **Very Small Aperture Terminals (VSATs) :** The communication between two ground stations via the satellite need not be of same size or transmit data with the same amount of power. Various satellite networks use a large number of small dishes, referred as VSATs for outlaying nodes and one central hub with a big dish that can transmit very powerful signals and is very sensitive to incom-

NOTES

ing ones. VSAT networks are useful where telephone links are overloaded, unreliable or difficult to obtain.

(ii) Paging and Satellite Networks : Paging is a wide area wireless communications method, through which brief alphanumeric messages are transmitted using radio frequencies to an electronic pager. When a subscriber's designated telephone number is dialed, the paging switch sends information to a radio transmitter that broadcasts a signal in the service area, which in turn delivers a tone, or numeric, alpha numeric, or voice message to the subscriber's specific pager. The Pager technology is the old form of mobile Telecommunication technology. It started in 1949 for sending one way mobile communication.

11.3.3 Radio Based System

Wireless communication services: radio, cellular, satellite and paging involve Radio Frequency (RF) signals that travel over frequencies from 100 KHz to 20GHz. Most wireless data services are based on RF transmission in the 800 to 900 MHz range. RF signals come in the following two forms:

1. Local area RF and
2. Wide Area RF.

RF Communications are of three types as given below:

1. One way
2. Two way
3. Local one way RF

11.3.4 Infra-red or Light Based Mobile Computing

User can make use of infrared waves to transmit the data through infrared equipment. These are used by following two categories of devices, as per the speed of data transfer:

(1) Low speed devices (115 kbps to 200 kbps)

(2) High speed devices (1.25 mbps) speed for transfer of data)

Infrared works by sending pulses of light from LED to a photo sensor that decodes the signals. Unlike rudimentary infrared devices like - TV remote controller, which can only send signals, computing devices typically can both send and receive infrared signals. Because the information is carried by light waves, the system will not work if any physical obstruction is placed between the sending and receiving devices. The infrared based computing is very cost effective.

Generations in Wireless Communications

Wireless systems have evolved over time and the chronological development of first, second, third and fourth generation is discussed below:

First, Second and Third Generation wireless systems (1G, 2G and 3G)

The first generation (1G) of wireless technology uses the analog technology. It uses frequency division multiple access (FDMA) for modulation. Systems of 1G were primarily developed for voice communication using frequency division multiplexing. To have efficient use of communication channels, time division multiplexing was used in second generation systems so that data can also be processed. The third generation systems evolved due to the need for transmitting integrated voice, data and multimedia traffic. The channel capacity is still limited and attempts are being made to compress the amount of information without compromising the quality of received signals.

NOTES

In second generation (2G) technology voice is digitized over a circuit. In 1G and 2G networks, data is transacted over circuits. This technology is called Circuit Switched Data (CSD). The 2G wireless systems have been designed for both indoor and vehicular environments with an emphasis on voice communication. An increased acceptance of mobile communication networks for convenient services has led to demands for high bandwidth wireless multimedia services. These ever growing demands require a new generation of high speed mobile infrastructure networks that can provide the capacity needed for high traffic volumes as well as flexibility in communication bandwidth and services. There is a need for frequent Internet access and multimedia data transfer, both of which may also involve the use of satellite communication. Thus the third generation (3G) systems (IMT-2000: International Mobile Telecommunications 2000) need to support real time data communication while maintaining compatibility with second generation systems.

Third generation or 3G is more of a generic term to mean mobile networks with high bandwidth. There are two schools of thought on the 3G systems. In the United States, people are inclined to cdms2000 as the basic technology, while in Europe and Japan, W-CDMA is being considered as future scheme. In principle both these schemes are similar but there are differences in their implementations. There are subtle differences in wireless and mobile systems e.g. a system could be immobile but wireless, or a system could be mobile but not wireless.

While 1G, 2G or 3G were making their marks in the Metropolitan Area Wireless networks (MAN), wireless technology has been getting popular in Local Area Networks (LAN) and Personal Area Networks (PAN) as well.

Fourth Generation Systems (4G)

International Mobile Telecommunications-Advanced (IMT Advanced), better known as 4G, 4th Generation or Beyond 3G, is the next technological strategy in the field of wireless communications. A 4G system will upgrade existing communication networks and is expected to provide a comprehensive and secure IP based solution where facilities such as voice, data and streamed multimedia will be provided to users on an "Anytime, Anywhere" basis and at much higher data rates compared to previous generations.

4G is being developed to accommodate the Quality of Service (QoS of a network is defined in different parameters such as bandwidth, latency, jitter, packet loss and packet delay) and rate requirements set by forthcoming applications like wireless broadband access, Multimedia Messaging Service (MMS), video chat, mobile TV, HDTV content, Digital Video Broadcasting (DVB), minimal services like voice and data, and other services that utilize bandwidth.

The 4G working group has defined the following as objectives of the 4G wireless communication standard:

- A spectrally efficient system (in bits/s/Hz and bits/s/Hz/site),
- High network capacity: more simultaneous users per cell,
- A nominal data rate of 100 Mbit/s while the client physically moves at high speeds relative to the station, and 1 Gbit/s while client and station are in relatively fixed positions as defined by the ITU-R,
- A data rate of at least 100 Mbit/s between any two points in the world,
- Smooth handoff across heterogeneous networks,
- Seamless connectivity and global roaming across multiple networks,
- High quality of service for next generation multimedia support (real time audio, high speed data, HDTV video content, mobile TV, etc)

NOTES

- Interoperability with existing wireless standards, and
- All IP, packet switched network.

In summary, the 4G system should dynamically share and utilize network resources to meet the minimal requirements of all the 4G enabled users.

As described in 4G consortia including WINNER (Wireless-World-Initiative-New-Radio), WINNER - Towards Ubiquitous Wireless Access and WWRF, a key technology based approach is summarized as follows, where WINNER is a consortium to enhance mobile communication systems. Considerable points for 4G systems are:

- Coverage, radio environment, spectrum, services, business models and deployment types, users Principal technologies.
- Baseband technique includes the following:
  1) OFDM: To exploit the frequency selective channel property.
  2) MIMO: To attain ultra high spectral efficiency.
  3) Turbo Principle: To minimize the required SNR at the reception side.
- Adaptive radio interface.
- Modulation, spatial processing including multi-antenna and multi-user MIMO.
- Relaying including fixed relay networks (FRNs) and the cooperative relaying concept, known as multi-mode protocol.

According to the members of the 4G working group, the infrastructure and the terminals of 4G will have almost all the standards from 2G to 4G implemented. Although legacy systems are in place to adopt existing users, the infrastructure for 4G will be only packet-based (all-IP). Some proposals suggest having an open Internet platform. Technologies considered to be early 4G include: Flash-OFDM, the 802.16e mobile version of WiMax (also known as WiBro in South Korea), and HC-SDMA (see iBurst). 3GPP Long Term Evolution may reach the market 1-2 years after Mobile WiMax is released. An even higher speed version of WiMax is the IEEE 802.16m specification. LTE Advanced will be the later evolution of the 3GPP LTE standard.

As the wireless standards evolved, the access techniques used also exhibited increase in efficiency, capacity and scalability. The first generation wireless standards used plain Time Division Multiple Access (TDMA) and Frequency Division Multiple Access (FDMA). In the wireless channels, TDMA proved to be less efficient in handling the high data rate channels as it requires large guard periods to alleviate the multipath impact. Similarly, FDMA consumed more bandwidth for guard to avoid inter carrier interference. So in second generation systems, one set of standard used the combination of FDMA and TDMA and the other set introduced a new access scheme called Code Division Multiple Access (CDMA). Usage of CDMA increased the system capacity and also placed a soft limit on it rather than the hard limit. Data rate is also increased as this access scheme is efficient enough to handle the multipath channel. This enabled the third generation systems to use CDMA as the access scheme IS-2000, UMTS, HSXPA, 1xEV-DO, TD-CDMA and TD-SCDMA. The only issue with CDMA is that it suffers from poor spectrum flexibility and scalability.

Recently, new access schemes like Orthogonal FDMA (OFDMA), Single Carrier FDMA (SC-FDMA), Interleaved FDMA and Multi-carrier code division multiple access (MC-CDMA) are gaining more importance for the next generation systems. WiMax is using OFDMA in the downlink and in the uplink. For the next generation UMTS, OFDMA is being considered for the downlink. By contrast, IFDMA is being considered for the uplink since OFDMA contributes more to the PAPR related issues and results in nonlinear operation of amplifiers. IFDMA provides less power fluctuation and thus avoids amplifier issues. Similarly, MC-CDMA is in the proposal for the IEEE 802.20 standard. These access schemes offer the same efficiencies as older technologies like CDMA. Apart from this, scalability and higher data rates can be achieved.

The other important advantage of the above mentioned access techniques is that they require less complexity for equalization at the receiver. This is an added advantage especially in the MIMO environments since the spatial multiplexing transmission of MIMO systems inherently requires high complexity equalization at the receiver.

In addition to improvements in these multiplexing systems, improved modulation techniques are being used. Whereas earlier standards largely used Phase-shift keying, more efficient systems such as 64QAM are being proposed for use with the 3GPP Long Term Evolution standards.

Internet Protocol version 6 (IPv6) Support

By the time that 4G is deployed, the process of IPv4 address exhaustion is expected to be in its final stages. Therefore, in the context of 4G, IPv6 support is essential in order to support a large number of wireless-enabled devices. By increasing the number of IP addresses, IPv6 removes the need for Network Address Translation (NAT), a method of sharing a limited number of addresses among a larger group of devices, although NAT will still be required to communicate with devices that are on existing IPv4 networks. IPv6 Unlike 3G, is based on two parallel infrastructures consisting of circuit switched and packet switched network nodes respectively, 4G will be based on packet switching only. This will require low-latency data transmission. As of June 2009, Verizon has posted specifications that require any 4G devices on its network to support IPv6.

Advanced Antenna Systems

The performance of radio communications obviously depends on the advances of an antenna system; refer to smart or intelligent antenna. Recently, multiple antenna technologies are emerging to achieve the goal of 4G systems such as high rate, high reliability, and long range communications. In the early 90s, to cater the growing data rate needs of data communication, many transmission schemes were proposed. One technology, spatial multiplexing, gained importance for its bandwidth conservation and power efficiency. Spatial multiplexing involves deploying multiple antennae at the transmitter and at the receiver. Independent streams can then be transmitted simultaneously from all the antennae. This increases the data rate into multiple folds with the number equal to minimum of the number of transmit and receive antennae. This is called MIMO (as a branch of intelligent antenna).

Apart from this, the reliability in transmitting high speed data in the fading channel can be improved by using more antennae at the transmitter or at the receiver. This is called transmit or receive diversity. Both transmit/receive diversity and transmit spatial multiplexing are categorized into the space-time coding techniques, which does not necessarily require the channel knowledge at transmit. The other category is closed-loop multiple antenna technologies which use the channel knowledge at the transmitter.

Software-Defined Radio (SDR)

SDR is one form of open wireless architecture (OWA). Since 4G is a collection of wireless standards, the final form of a 4G device will constitute various standards.

NOTES

This can be efficiently realized using SDR technology, which is categorized to the area of the radio convergence.

4G wireless standards

3GPP is currently standardizing LTE Advanced as future 4G standard. A first set of 3GPP requirements on LTE Advanced has been approved in June 2008. The working groups are currently evaluating various proposals for standardization. LTE Advanced will be standardized as part of the Release 10 of the 3GPP specification.

Development of 4G

The Japanese company NTT DoCoMo has been testing a 4G communication system prototype with 4x4 MIMO called VSF-OFCDM at 100 Mbit/s while moving, and 1 Gbit/s while stationary. In February 2007, NTT DoCoMo completed a trial in which they reached a maximum packet transmission rate of approximately 5 Gbit/s in the downlink with 12x12 MIMO using a 100MHz frequency bandwidth while moving at 10 km/h, and is planning on releasing the first commercial network in 2010.

Digiweb, an Irish fixed and wireless broadband company, has announced that they have received a mobile communications license from the Irish Telecoms regulator, ComReg. This service will be issued the mobile code 088 in Ireland and will be used for the provision of 4G Mobile communications. Digiweb launched a mobile broadband network using FLASH-OFDM technology at 872 MHz.

Pervasive networks are an amorphous and at present entirely hypothetical concept where the user can be simultaneously connected to several wireless access technologies and can seamlessly move between them (See vertical handoff, IEEE 802.21). These access technologies can be Wi-Fi, UMTS, EDGE, or any other future access technology. Included in this concept is also smart-radio (also known as cognitive radio technology) to efficiently manage spectrum use and transmission power as well as the use of mesh routing protocols to create a pervasive network.

Verizon Wireless announced on September 20, 2007 that it plans a joint effort with the Vodafone Group to transition its networks to the 4G standard LTE. On December 9, 2008, Verizon Wireless announced that they intend to build and begin to roll out a LTE network by the end of 2009.

Telus and Bell Canada, the major Canadian cdmaOne and EV-DO carriers, have announced that they will be cooperating towards building a fourth generation (4G) LTE wireless broadband network in Canada. As a transitional measure, they are implementing 3G UMTS to go live by early 2010. Sprint announced it will be offering a 3g/4g connection plan for $79.99 it is only currently available in Baltimore.

At the present rates of 15-30 Mbit/s, 4G is capable of providing users with streaming high-definition television. At rates of 100 Mbit/s, the content of a DVD-5 (for example a movie), can be downloaded within about 5 minutes for offline access.

11.4 Wireless Application Protocol

It is a protocol that supports distributed, mobile computing and commerce. Mobile computing and commerce requires a wireless information transmission method with many approaches of communication without wires. WAP is designed for access to Internet and advanced telephony services from mobile phones. It is a common effort of many companies and organizations to set up a framework for wireless and mobile web access using many different transport systems. It integrates several communication layers for security mechanism, transaction oriented protocols and application support. It is a data bearer service over HTTP protocol. It combines the telephone network and the internet by integrating telephony applications into the web using its own Wireless Markup Language (WML) and scripting language (WMLScript).

WAP is defined as a set of specifications, developed by the WAP forum that lets developer using Wireless Mark-up Language (WML) built networked application designed for hand held wireless devices. WAP is a de-facts standard with support from more than 200 vendors. Net-net using WAP, a mobile user will be able to access the same wealth of information from a pocket-sized device as they can from a desktop.

WAP pays proper sensitivity to the constraints of mobile devices like:

Small display,

Limited keys on the keyboard,

Limited memory,

Low bandwidth,

No pointer device like mouse etc.

Networks of WAP

Though WAP can be used from a variety of networks, GPRS and 3G networks are more suited for these applications. As a part of WAP Forum's goals, WAP can be accessible from, but not limited to, the following more networks:

- GSM-900, GSM-1800, GSM-1900
- GPRS
- CDMA IS-95, CDMA2000

The WAP Architecture

WAP is based on layered architecture. The WAP protocol stack is similar to the OSI network model shown in Figure 11.4. These layers consist of:

1. Wireless Application Environment (WAE)
2. Wireless Session Protocol (WSP)
3. Wireless Transactional Protocol (WTP)
4. Wireless Transport Layer Security (WTLS)
5. Wireless Datagram Layer (WDP).

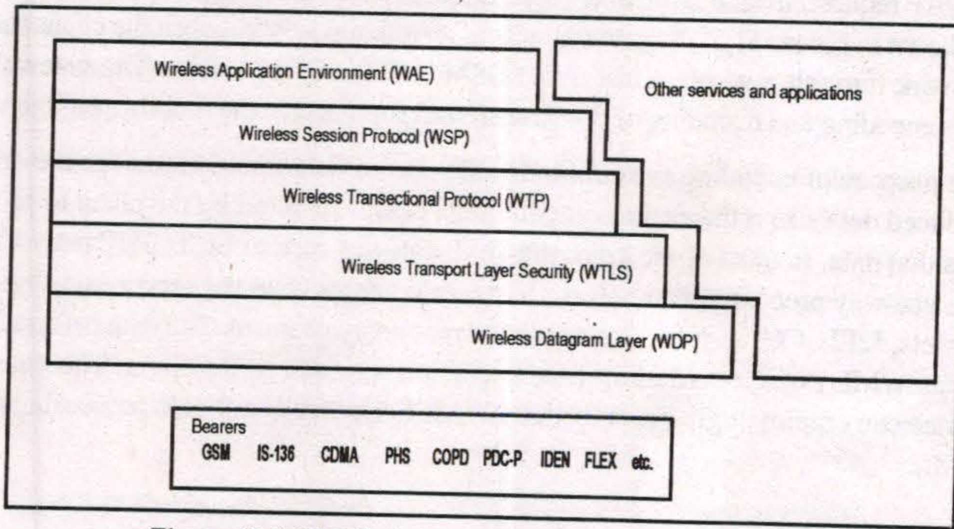


Figure 11.4 WAP layered Architecture and protocol stack

NOTES

The application environment of WAE comprises multiple components to provide facilities like:

- User agent: the browser or a client program.
- Wireless Mark-up Language: a lightweight markup language, similar to HTML, but optimized for use in wireless devices.
- WML Script: a lightweight client side scripting language, similar JavaScript in web.
- Wireless Telephony Application (WTA): telephony services and programming interfaces.
- WAP Push Architecture: mechanism to allow origin servers to deliver content to the terminal without the terminal requesting for it.
- Content Formats: a set of well defined data formats, including images, phone book records and calendar iformation.

WAP supports different types bearer networks. These are GSM, IS-136, CDMA, PHS (Personal Handy phones System), CDPD (Cellular Digital packet Data), etc.

11.4.1 WAP Programming Model

It includes WAE (WAP Application Environment) logical model and understanding of Internet model. Following sections discusses both:

WAE Logical model

The primary objective of the WAP application environment is to provide an interoperable environment to build services in wireless space. It covers system architecture relating to the user agents, networking schemes, content formats, programming languages and shared services based on World Wide Web (WWW) technologies. Contents are transported using standard protocol in the WWW domain and an optimized HTTP-like protocol in the wireless domain.WAE architecture allows all contents and services to be hosted on standard web servers. All contents are located using WWW standard URL's.

A WAP request, from the browser (user agent), is routed through a WAP gateway as shown in Figure 11.5.The gateway acts as an intermediary between the client and network through a wireless last mile (GSM, GPRS, CDMA etc.). The gateway does encoding and decoding of data transferred from and to the mobile user agent.

The purpose of encoding is to minimize the size of data transferred over-the-air. Reduced data size reduces the computational power required by the client to process that data. In most of the cases the WAP gateway resides on TCP/IP network. The gateway processes the request, Retrieves contents from the server using Java servlets, J2EE, CGI scripts or some other dynamic mechanism. The data is formatted as WML (Wireless Markup Language) and returned to the client. The client device can employ logic via embedded WMLScript for client-side processing of WML.

NOTES

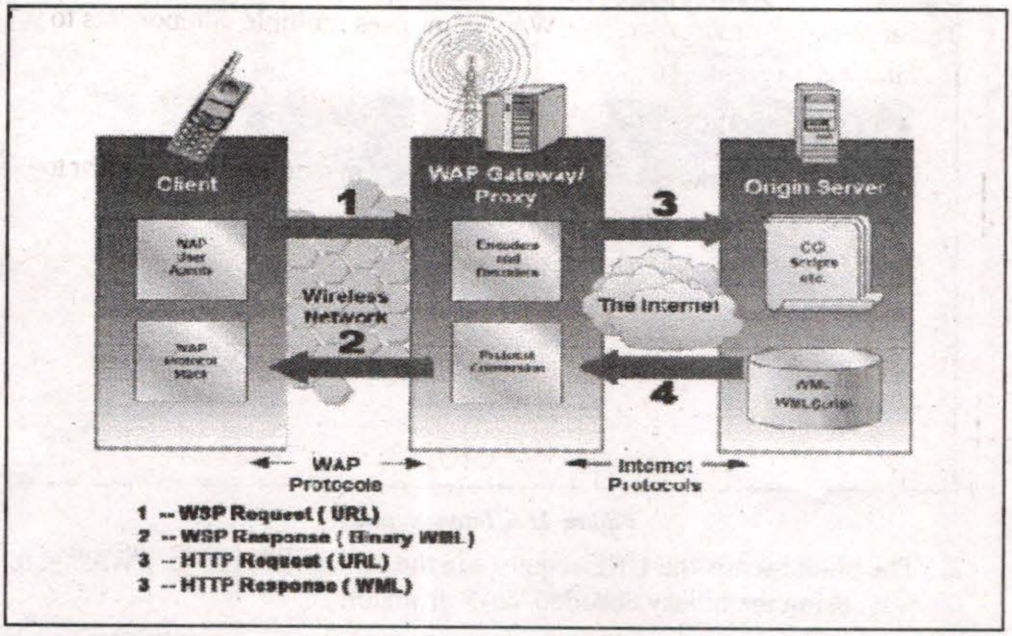


Figure 11.5 WAE Logical Model

WAP Gateway/Proxy is the entity that connects the wireless domain with the Internet. Note that the request that is sent from the wireless client to the WAP Gateway/ Proxy uses the Wireless Session Protocol (WSP). In its essence, WSP is a binary version of HTTP.

The Internet Model

The Internet model makes it possible for a client to reach services on a large number of origin servers; each addressed by a unique Uniform Resource Locator (URL). The contents stored on the servers are of various formats, while HTML is prominent. HTML provides the content developer with a means to describe the appearance of a service in a flat document structure. If more advanced features like procedural logic is needed then scripting languages such as JavaScript or VB Script may be utilized. The Figure 11.6 shows how a WWW client request a resource stored on a web server. On the Internet, standard communication protocols like HTTP and Transmission Control Protocol/Internet Protocol (TCP/IP) are used.

The contents available at the web server may be static or dynamic. Static content is produced once and not changed or updated very often, e.g. a company presentation. Dynamic content is needed when the information provided by the service changes more often, e.g. Timetables, News, Stock quotes and Account information. Technologies such as Active Server Pages (ASP), Common Gateway Interface (CGI) and Servlets allow content to be generated dynamically.

Working steps of WAP model

Working of WAP model can be given by the following steps:

1. The user selects an option on their mobile device that has a URL with Wireless Markup language (WML) content assigned to it.

NOTES

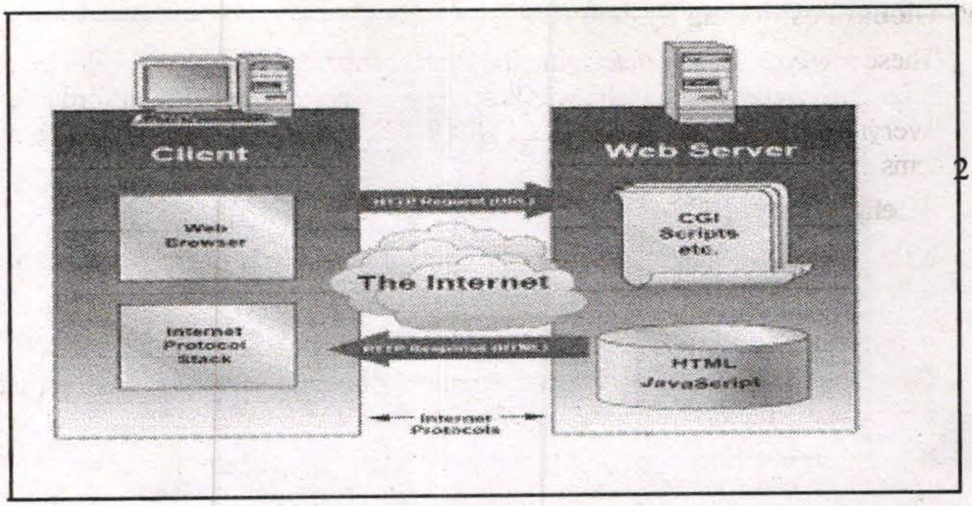


Figure 11.6 Internet model

2. The phone sends the URL request via the phone network to a WAP gateway, using the binary encoded WAP protocol.
3. The gateway translates this WAP request into a conventional HTTP request for the specified URL, and sends it on to the Internet.
4. The appropriate Web server picks up the HTTP request.
5. The server processes the request, just as it would process any other request. If the URL refers to a static WML file, the server delivers it. If a CGI script is requested, it is processed and the content is returned as usual.
6. The Web server adds the HTTP header to the WML content and returns it to the gateway.
7. The WAP gateway compiles the WML into binary form.
8. The gateway then sends the WML response back to the phone.
9. The phone receives the WML via the WAP protocol.
10. The micro-browser processes the WML and displays the content on the screen.

11.5 Other Wireless Technologies

A wireless technology is expected to provide "anytime anywhere" type of service and this characteristic has made it a very attractive technology. This kind of feature is essential for military and defense areas as well as to a limited class of potentially life-threatening applications like nuclear power, aviation and medical emergencies. Other then the wireless systems discussed in section 11.3, some other different wireless technologies and their associated characteristics are given below:

1. **Cellular systems network**

   It provides data and voice through handheld phones. There can be continuous coverage limited to metropolitan regions. It has its limitations like available bandwidth is very low for most data intensive applications. Examples of these types of systems are Cellular phones and Personal digital assistants.

2. **Wireless local area network**

   These are Traditional LAN extended with wireless interface. It can be used in local environments only with limited range. Examples of these types of systems are NCR's wavelan, Motorola's ALTAIR, Proxim's range LAN and Telesystem's ARLAN.

NOTES

3. Global Positioning System (GPS)

These systems help to determine the three dimensional position, velocity and time. Its coverage area is the whole surface of earth. It is still not affordable by everyone. GNSS, NAVSTAR and GLONASS are the examples of such systems

4. Satellite-based PCS

It provides voice paging and messaging. It covers almost anyplace on earth. It is costly. Iridium and teledisc are its examples.

5. Ricochet

These systems provide high speed, secure mobile access to the desktop (data) from outside the office.

6. Home Networking

These are used to connect different PCs in the house to share files and devices such as printers. They can cover all the places of a house. Netgear Phoneline 10X, Intel anyPoint, Phoneline Home Network are some examples of these types of systems.

7. Ad hoc networks

In these networks group of people come together for a shore time to share data. Its coverage area is equal to LAN, but without fixed infrastructure. It has a limited range. These networks are in use for defense applications.

8. WPAN (Bluetooth)

Using this technology all digital devices can be connected without any cable. It provides private ad hoc grouping away from fixed network infrastructure. Its range is limited due to the short range radio link. Some home devices have this technology.

9. Sensor networks

These networks provide a large number of tiny sensors with wireless capabilities. There coverage area is relatively terrain i.e. very limited range. These networks are in use for defense civilian applications.

11.6 GSM/CDMA Security Issues

The radio medium is open to everybody and anybody. Anybody who can get hold of a radio receiver can access GSM signal or data. Therefore, it is necessary and important that the communication over the wireless radio media is secured. The first step to the GSM security is the authentication. The authentication of a user is done to ensure that the user is really the person who claims he is. Authentication involves two functional entities, the SIM card in the mobile phone and the Authentication Centre (AUC). Authentication is done by using an algorithm by name A3. Following the authentication a key is generated for encryption. An algorithm by the name A8 is used to generate the key. A different algorithm called A5 is used for both ciphering and deciphering procedures. The ciphering is done on signals, voice and data. In other words this means that SS7 signal, voice, data and SMS within GSM are ciphered over the wireless radio interface.

The GSM specifications for security were designed by the GSM Consortium in secrecy and distributed only on a need-to-know basis to hardware and software manufacturers and to GSM network operators. The specifications were never exposed to the public. The GSM Consortium relies on Security of Obscurity, i.e. the algorithms would be harder to crack if they were not publicly available.

NOTES

Authentication algorithm A3

During the authentication process the MSC challenges the MS a RANDom number (RAND) as shown in Figure 11.7. The SIM card uses this RAND received from the MSC and a secret key Ki stored within the SIM as input. Both RAND and the Ki secret are 128 bits long. Using the A3 algorithm with RAND and Ki a 32-bit output called Signature RESponse (SRES) is generated in the MS. This SRES is then sent back to the MSC as the response to the challenge. Using the same set of algorithm, the AUC also generates a SRES. The SRES from MS (SIM) and the SRES generated by the AUC are compared. If they are same, the MS is authenticated. The idea is that no keys will be transacted over the air. However, if the SRES values calculated independently by the SIM and the AUC are the same, the Ki has to be same. If Ki is same, SIM card is genuine.

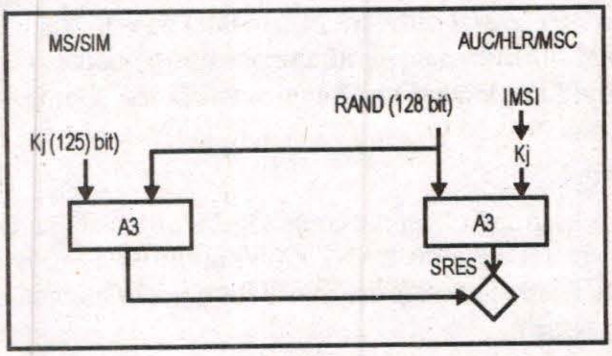


Figure 11.7 The flow of steps for authentication

The voice privacy key generation algorithm A8

For any type of cipher, we need a key. If the key is random and difficult to guess, the cipher is relatively secured. In the GSM security model, A8 algorithm is the key generation algorithm, shown in Figure 11.7. A8 generates a session key, Kc, from the random challenge, RAND, received from the MSC and from the secret key Ki. The inputs for A8 are the same set of 128-bit Ki and RAND as used in A3. The A8 algorithm takes these inputs and generates a 64-bit output. The keys are generated at both the MS (SIM) end. The BTS received the Kc from the MSC. The session key Kc, is used for ciphering, till the time MSC decides to authenticate the MS once again. This might sometime take days.

The Strong Over the Air Voice Privacy algorithm A5/1

In the GSM security model, A5 is the stream cipher algorithm used to encrypt over the air transmissions, as shown in Figure 11.8. There are several cryptographic algorithms in this series used for security. The A5/1 and A5/2 stream ciphers are used for ensuring over-the-air voice privacy. A5/1 was developed first and is a stronger algorithm used within Europe and the United States; A5/2 is weaker and used in other countries. A large security advantage of GSM over earlier systems is that the Key, the crypto variable stored on the SIM card that is the key to any GSM ciphering algorithm, is never sent over the air interface. Serious weaknesses have been found in both algorithms, and it is possible to break A5/2 in real-time in a cipher text only attack. The system supports multiple algorithms so operators may replace that cipher with a stronger one.

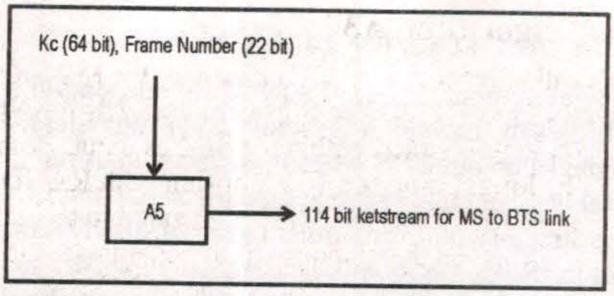


Figure 11.8 Keystream generation

NOTES

CDMA

CDMA is a specific modulation technique of Spread Spectrum technology. A number of different terms are used to refer to CDMA implementations. The original U.S. standard defined by 'Qualcomm' was known as IS-95, the IS referring to an Interim Standard of the Telecommunications Industry Association (TIA). IS-95 is often referred to as 2G or second generation cellular. The 'Qualcomm' brand name 'cdmaOne' may also be used to refer to the 2G CDMA standard. CDMA has been submitted for approval as a mobile air interface standard to the International Telecommunication Union (ITU).

Whereas the Global System for Mobile Communications (GSM) standard is a specification of an entire network infrastructure, the CDMA interface relates only to the air interface the radio part of the technology. For example, GSM specifies an infrastructure based on internationally approved standard while CDMA allows each operator to provide the network features as it finds suited. On the air interface, the signaling suite (GSM: ISDN SS7) work has been progressing to harmonies these.

After a couple of revisions, IS-95 was superseded by the IS-2000 standard. This standard was introduced to meet some of the criteria laid out in the IMT-2000 specification for 3G. It is also referred to as 1xRTT which simply means "1 times Radio Transmission Technology" and indicates that IS-2000 uses the same 1.25 MHz carrier shared channel as the original IS-95 standard. A related scheme called 3xRTT uses three 1.25 MHz carriers for a 3.75 MHz bandwidth that would allow higher data burst rates for an individual user, but the 3xRTT scheme has not been commercially deployed. More recently, 'Qualcomm' has led the creation of a new CDMA-based technology called 1xEV-DO or IS-856, which provides the higher packet data transmission rates required by IMT-2000 and desired by wireless network operators.

This CDMA system is frequently confused with a similar but incompatible technology called Wideband Code Division Multiple Access (W-CDMA) which forms the basis of the W-CDMA air interface. The W-CDMA air interface is used in the global 3G standard UMTS and the Japanese 3G standard FOMA, by NTT DoCoMo and Vodafone. However, the CDMA family of US national standards (including cdmaOne and CDMA2000) are not compatible with the W-CDMA family of ITU standards. Another important application of CDMA predating and entirely distinct from CDMA cellula is the Global Positioning System (GPS).

The 'Qualcomm CDMA' system includes highly accurate time signals (usually referenced to a GPS receiver in the cell base station), so cell phone CDMA-based clocks are an increasingly popular type of radio clock for use in computer networks. The main advantage of using CDMA cell phone signals for reference clock purposes is that they work better inside buildings, thus often eliminating the need to mount a GPS antenna on the outside of a building.

NOTES

Authentication and security of CDMA

The electronic serial number (ESN) of a IS-95 mobile station is a 32 bits binary number that identifies the mobile. It is factory and is not alterable in the field. All mobiles are assigned a unique ESN when manufactured. A mobile station also has a unique 15 digital number called Mobile Identification Number (MIN). This is the mobile's 10 digit directory number similar to the MSISDN number in GSM. The difference is that in IS-95, the mobile is assigned a number similar to the North American numbering scheme. For example, a fixed line number in Allahabad may have a directory number of 090-3678-2468 whereas a mobile phone may have a directory number of a GSM phone like 35467-87654. On the contrary in USA if fixed line number is 1-456-879-5676, the mobile number can be 11-456-340-6500.

Whenever a mobile is turned on, it registers with the network. During the authentication process the network throws a global challenge to the mobile. The AC (Authentication centre) transmits a random number to the mobile station. The mobile station encrypts it using a key shared between the mobile station and the AC. The encrypted random message is sent back to the network. The AC checks this result with the result it calculates using the same number and shared key. If these match the mobile is authenticated. The global challenge in IS-95 is more frequent than n GSM. In IS-95, authentications even take place following successful handoffs. Following successful authentication, the VLR (Visitor Location Register) assigns a TMSI (Temporary International Mobile Subscriber Identity). TMSI provides anonymity since it is a transient identity only the mobile and the network are aware of. The spreading PN (Pseudorandom Noise) sequence also play a role in security. For anyone to impersonate the CDMA traffic, the eavesdropper needs to know the PN sequence.

11.7 Growth, Success Stories and Applications of M-Commerce

Analysts say that the rapid commercialization of 3G services is likely to open up new opportunities in developed markets for M-Commerce. In order to develop the M-Commerce market potential, handset manufactures such as - Nokia, Ericson, Motorola and Qualcomm are working with carriers such as AT & T wireless and sprint to develop Wireless Application Protocol-enabled smart phones and ways to reach them. Using Bluetooth technology, smart phones offer fax, e-mail and phone capabilities all in one, concreting the way for M-Commerce to be accepted by an increasingly mobile workforce.

PDAs and cellular phones have becomes so popular that many businesses are beginning to use M-Commerce as a more efficient method of reaching and communicating with their customers. After technological trends and advances, Asia, Europe, Canada and the United States are also beginning to experiment with early-stage M-Commerce. Japan is seen as an international chief in M-Commerce.

M-Commerce was born in 1997, when the first two mobile phones enabled Coca Cola vending machines were installed in the Helsinki area in Finland. They used SMS text messages to send the payment to the vending machines. In 1997 also the first mobile phone based banking service was launched by Merita bank of Finland also using SMS.

In 1998 the first digital content sales were made possible as downloads to mobile phones when the first commercial downloadable ringing tones were launched in Finland by Radionlinja, now is part of Elisa.

NOTES

In 1999 two major national commercial platforms for M-Commerce were launched with the introduction of a national M-Payment system by Smart money in the philippines and the launch of the first mobile internet platform by NTT DoCoMo in Japan, called I-Mode. I-Mode was revolutionary also in offering a revenue-sharing deal where NTT DoCoMo only kept 9% of the content Payment and returned 91% to the content owner.

Mobile Commerce related services spread rapidly in early 2000 from Norway launching mobile parking, Austria offering mobile tickets to trains, and Japan offering mobile purchases of airline tickets.

The first conference dedicated to mobile commerce was held in London in July 2001 and the first book to cover M-Commerce was Tomi Ahonen's M-profiles in 2002. The first university short course to discuss M-Commerce was held at the University of Oxford in 2003 with Tomi Ahonen and Steve Jones lecturing.

Applications of M-Commerce

There are many applications of M-commerce in present busy and mobile world, when everyone wants to work even during movements. Some existing applications have been discussed below:

Mobile Ticketing

Tickets can be sent to mobile phones using a variety of technologies. User is then able to use their tickets immediately by presenting their phones at the venue. Tickets can be booked and cancelled on the mobile with the help of simple application downloads or by accessing WAP portals of various travel agents or direct service providers.

Mobile ticketing for airports, ballparks and train stations e.g. will not only streamline unexpected metropolitan traffic surges, but also help users remotely secure parking spot and greatly facilitate mass surveillance at transport hubs.

Mobile ticketing technology can also be used for the distribution of vouchers, coupons and loyalty cards. The voucher, coupon, or loyalty card is represented by a virtual token that is sent to the mobile phone. Presenting a mobile phone with one of these tokens at the point of sale allows the customer to receive the same benefits as another customer who has a loyalty card or other paper coupon/voucher. Coupons may be sent to a customer utilizing location based services when he is in a certain physical proximity.

Mobile delivery enables

Following are some benefits of the mobile message delivery:

- Economy of scale.
- Quicker and easier delivery.
- Effective target marketing.
- Privacy friendly data mining and customer behavior.
- Environment friendly and resource saving efficacy.

Content Purchase and delivery

Mobile Content purchase and delivery technologies mainly consist of the sale of ring tones, wallpapers and games for mobile phones. The convergence of mobile phones, mp3 players and video players into a single device will result in an increase in the purchase and delivery of full-length must and video download speeds, if increased to 4G levels, will make it possible to buy a movie on a mobile device in a couple of seconds, while on the go.

NOTES

Location Based Services

The location of the mobile phone user is an important piece of information used for mobile commerce transaction. Knowing the location of the user allows for location based service such as:

1. Local Maps
2. Local offers
3. People tracking and monitoring

Information Technology

A wide variety of information technology can be delivered to mobile phone users in much the same way as it is delivered to PCs. This service includes the following:

- News services
- Stock data
- Sports results
- Financial records
- Traffic data and information.

Mobile Banking

Banks and other financial institutions are exploring the use of M-Commerce to allow their customers to not only access account information but also make transaction, e.g. Purchasing stock, remitting money via mobile phones and other mobile equipment. This service is often referred to as mobile banking or M-Banking. More negative issues like ID theft, phishing and farming are lurking when it comes to mobile banking, particularly done on the mobile web net-security technology free from redundancy and paradigm shifts away from mobile web-based banking will be an optional solution to mobile banking in the near future.

Mobile Banking in the world

This part of the mobile commerce is very popular in countries where most of their population is unbanked. Countries like Sudan, Ghana and South Africa reacted well to this new type of commerce. In the latin America countries like Argentina, Brazil, Venezuela, Colombia recently Mexico started with a huge success. In colombia it was released with Redeban. Guatemala has the support of Banco industrial. Mexico released the mobile commerce with omnilife, Bancomer and a private company.

Mobile brokerage

Stock market services offered via mobile devices have also become more popular and are known as mobile brokerage. They allow the subscriber to react to market development in a timely fashion and irrespective of their physical location.

Auction

Over the past three years mobile reverse auction solution have grown in popularity unlike traditional auction, the reverse auction bills the consumer's phone each time they place a bid. Many mobile PSMS commerce solutions rely on a one-time purchase or one-time subscription, however, reverse auction are high return applications as they allow the consumer to transact over a long period of time.

11.8 M-Commerce in India

M-Commerce is taking off in India and one of the companies leading the way is paymate, a Mumbai-based wireless-transaction platform provider. It is not surprising that India has become a thriving center for M-Commerce. The Telecom Regulatory authority of India reports that the number of mobile subscribers has already reached 185 million, for surpassing the country's 9 million online customers (23 Oct. 2007).

NOTES

Ajay Adiseshann, Managing Director and founder of paymate, says that his country is leap fogging the internet revolution and going straight to the mobile experience. "The first experience people will have of digital shopping, or e-shopping, will probably not be over the internet but rather on the mobile," he says "That is where the opportunity lies and that is why businesses are now looking at the mobile phone as the next channel for providing their services to customers."

Paymate launched its payment service last year, after spending two year developing an "Eco-system" that connects banks, retailers and customers using SMS. The service has been recognized both in India and Abroad for its simplicity and ease of use. It works like: You register for paymate with your bank by calling or sending a text message. Once your phone number has been verified by the bank, you are a given a unique Personal Identification Number (PIN) that you use whenever you make a payment online, through your phone or over the counter.

SUMMARY

- Performing commerce using mobile and handheld devices such as Personal Digital Assistants (PDAs), Mobile phone, Smartphone and other emerging mobile equipments equipped with Web-ready micro-browsers, through wireless technology, is called Mobile Commerce.
- Different services affected by M-Commerce include; Financial services, Telecommunications, Service/Retail and Information services.
- Technologies required for M-Commerce include the wireless technologies like required for mobile computing e.g. wireless communication technologies, Wireless Application Protocol (WAP), WAP programming model etc.
- The Wireless Communication service can be provided by various methods such as radio, cellular and satellite based wireless systems.
- The first generation (1G) of wireless technology uses the analog technology.
- In second generation (2G) technology voice is digitized over a circuit.
- Third generation or 3G is more of a generic term to mean mobile networks with high bandwidth.
- A 4G system will upgrade existing communication networks and is expected to provide a comprehensive and secure IP based solution.
- The 4G system should dynamically share and utilize network resources to meet the minimal requirements of all the 4G enabled users.
- Wireless Application Protocol is a protocol that supports distributed, mobile computing and commerce.
- WAP Programming Model includes WAE (WAP Application Environment) logical model and understanding of Internet model.

NOTES

- M-Commerce was born in 1997, when the first two mobile phones enabled Coca Cola vending machines were installed in the Helsinki area in Finland. They used SMS text messages to send the payment to the vending machines. In 1997 also the first mobile phone based banking service was launched by Merita bank of Finland also using SMS.
- M-Commerce is taking off in India and one of the companies leading the way is paymate, a Mumbai-based wireless-transaction platform provider.

EXERCISE

1. What do you understand by Mobile Commerce ? Give its applications also.
2. Briefly describe the success stories of M-Commerce.
3. Briefly explain the technology for M-Commerce.
4. Explain the Role of M-Commerce in India.
5. Explain the WAP Model and its working.
6. What is Wirless Communication.
7. Briefly explain the security issues of GSM and CDMA.
8. Explain the generations in Wireless Communication.

12 Geographic Information Systems

Chapter Includes :

12.1 Introduction

When introduced in the 1960s, the use of Geographic Information Systems (GIS) was limited to a small number of research and applications users. Today, GIS is one of the fastest growing technologies; it has applications in public safety, natural resource management, environmental analysis, utilities, and government, and is moving quickly into many other areas. The motivation for the tremendous growth in the use of GIS is clearly linked to both the increasing demand for information and the ever-increasing ability of computer technology to provide effective, cost-efficient data processing and management capabilities.

Conceptually, a GIS can be envisioned as a stacked set of map layers, where each layer is aligned or registered to all other layers. Typically, each layer will contain a unique geographic theme or data type. These themes might include, for example,

NOTES

topography, soils, land-use, cadastral (land ownership) information, or infrastructure such as roads, pipelines, power lines, or sewer networks. By sharing mutual geography, all layers in the GIS can be combined or overlaid in any user-specified combination. In some cases the GIS may be defined by the type of data that the system is designed to handle. For example, the term "Land Information System" or "LIS" is often applied to a type of GIS used by counties, cities, and municipalities to manage land parcel information.

In some of its simpler operations, GIS provides an automated version of traditional map analysis. Map overlay is probably the most common GIS function; this function has long been performed manually or optically using maps registered over light tables or via photographic techniques. In GIS, however, the number of registered map layers that can be collected and stored is theoretically infinite. The user can quickly retrieve, overlay, manipulate, and analyze any number or combination of layers. The user can then assess the results of the analysis on a computer screen or on a hard copy paper map produced by the GIS, or can summarize the results in a tabular format. However simple or complex the user's purpose, a GIS is used to access an integrated, geographically referenced database of maps that can be overlaid, combined, and analyzed to user specifications.

Because GIS is a rapidly growing technology, many different definitions of GIS now exist. A very useful definition of GIS is a true systems-based definition: *a GIS is a computerized, integrated system used to compile, store, manipulate, and output mapped spatial data.*

12.2 Geographic Information System

Definition : A geographic information system (GIS) is a "computer-based information system that enables capture, modeling, manipulation, retrieval, analysis, and presentation of geographically referenced data". The main technology components, geographic information systems (GIS), consist of software packages that are capable of integrating spatial and non-spatial data to yield the spatial information that is used in decision making. They are computer-based equipment, procedures and techniques for manipulating spatial or map data.

The advent of cheap and powerful computers over the last few decades has allowed for the development of innovative software applications for the storage, analysis, and display of geographic data. Many of these applications belong to a group of software known as **Geographic Information Systems (GIS)**. Many definitions have been proposed for what constitutes a GIS. Each of these definitions conforms to the particular task that is being performed. Thus, the activities normally carried out on a GIS include:

- The measurement of natural and human made phenomena and processes from a spatial perspective. These measurements emphasize three types of properties commonly associated with these types of **systems**: **elements**, **attributes**, and **relationships**.
- The storage of measurements in digital form in a computer database. These measurements are often linked to features on a digital map. The features can be of three types: points, lines, or areas (polygons).

- The analysis of collected measurements to produce more data and to discover new relationships by numerically manipulating and modeling different pieces of data.
- The depiction of the measured or analyzed data in some type of display - maps, graphs, lists, or summary statistics.

NOTES

The first computerized GIS began its life in 1964 as a project of the Rehabilitation and Development Agency Program within the government of Canada. The **Canada Geographic Information System** (**CGIS**) was designed to analyze Canada's national land inventory data to aid in the development of land for agriculture. The CGIS project was completed in 1971 and the software is still in use today. The CGIS project also involved a number of key innovations that have found their way into the feature set of many subsequent software developments.

From the mid-1960s to 1970s, developments in GIS were mainly occurring at government agencies and at universities. In 1964, Howard Fisher established the **Harvard Lab for Computer Graphics** where many of the industries early leaders studied. The Harvard Lab produced a number of mainframe GIS applications including: SYMAP (Synagraphic Mapping System),CALFORM, SYMVU, GRID, POLYVRT, and ODYSSEY. ODYSSEY was first modern vector GIS and many of its features would form the basis for future commercial applications. Automatic Mapping System was developed by the United States Central Intelligence Agency (CIA) in the late 1960s. This project then spawned the CIA's **World Data Bank**, a collection of coastlines, rivers, and political boundaries, and the **CAM** software package that created maps at different scales from this data. This development was one of the first systematic map databases. In 1969, Jack Dangermond, who studied at the Harvard Lab for Computer Graphics, co-founded **Environmental Systems Research Institute (ESRI)** with his wife Laura. ESRI would become in a few years the dominate force in the GIS marketplace and create **ArcInfo** and **ArcView** software. The first conference dealing with GIS took place in 1970 and was organized by Roger Tomlinson (key individual in the development of **CGIS**) and Duane Marble (professor at Northwestern University and early GIS innovator). Today, numerous conferences dealing with GIS run every year attracting thousands of attendants.

In the 1980s and 1990s, many GIS applications underwent substantial evolution in terms of features and analysis power. Many of these packages were being refined by private companies who could see the future commercial potential of this software. Some of the popular commercial applications launched include: **ArcInfo, ArcView, MapInfo, SPANS GIS, PAMAP GIS, INTERGRAPH,** and **SMALLWORLD**. It was also during this period that many GIS applications moved from expensive minicomputer workstations to personal computer hardware.

12.3 Components of a GIS

A **Geographic Information System** combines computer cartography with a database management system. GIS consists of three subsystems:

(1) An input system that allows for the collection of data to be used and analyzed for some purpose;

NOTES

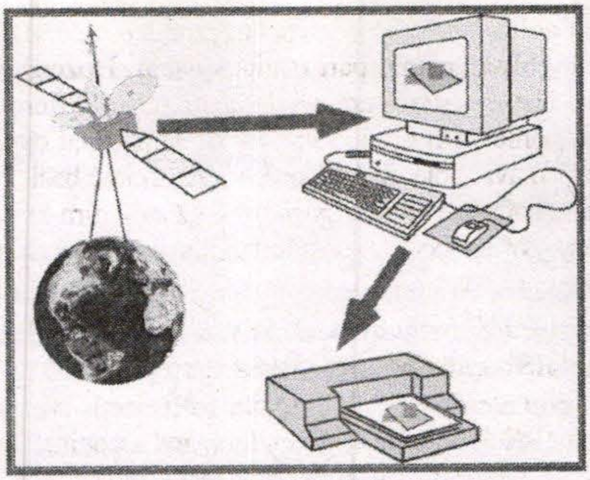


Figure 12.1: GIS Subsystems

(2) Computer hardware and software systems that store the data, allow for data management and analysis, and can be used to display data manipulations on a computer monitor;

(3) An output system that generates hard copy maps, images, and other types of output.

A common misconception is that data alone constitute a GIS. Similarly, it is sometimes assumed that someone who has purchased GIS software has acquired a "GIS." There are four components of GIS: (1) **data**, (2) **hardware**, (3) **software**, and (4) **users**.

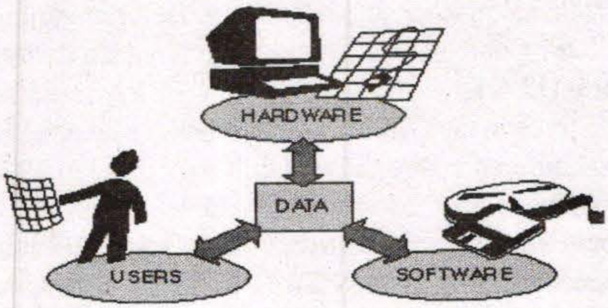


Figure 12.2: GIS Components

12.3.1 Hardware

Computer hardware used to support GIS is a highly variable part of the overall system. Users will customize their hardware environment to best meet their own individual needs. In all cases, however, a fully-functional GIS must contain hardware to support data input, output, storage, retrieval, display, and analysis. During its infancy, GIS processing was supported only on large mainframe computers. Today, a variety of platforms support GIS processing, ranging from large mainframe computers to mini-computers to scientific work stations to personal computers. In many cases, hardware used to support other applications (e.g., payroll or accounting or digital image processing) is also used for GIS. Hardware configurations for GIS span a tremendous range in terms of start-up and acquisition costs. All systems carry associated costs for maintenance and support.

NOTES

12.3.2 Software

Software is also a highly dynamic part of the system. Dozens of GIS software packages now exist. These systems are available on many different types of hardware platforms and come with a wide variety of functional capabilities. Public domain software is also available, though on a more limited basis than commercial software. The range in software options goes from generic turn-key systems that are ready for use "right out of the box" to customized installations designed to support specific user needs.

Given the sometimes bewildering array of choices for hardware and software, selection and use of a GIS should be approached strategically. In all cases, anyone considering a GIS software package should consider needs carefully and consult various references, including other users, vendors, and technical publications.

12.3.3 Data

GIS components are dynamic; there is rapid change in the computer industry as well as turnover of personnel involved in GIS projects. For this reason, GIS developers are often encouraged to adopt a data-centered approach. Simply stated, a data-centered approach views data as the central resource in the GIS. Though data may be shared among multiple users and multiple hardware/software environments, the data are collected and compiled by a person or organization to support the goal of that user. The other components provide the support needed to process that data.

All data in a GIS are either **spatial** data or **attribute** data. As discussed above, spatial data tells us **where** something occurs. Attribute data tells **what** occurs; it tells us the nature or characteristics of the spatial data. For example, we might describe the location of a municipal water well as a point with the coordinates "45 degrees 17 minutes 20 seconds north latitude, 94 degrees 7 minutes 48 seconds west longitude." Furthermore, we can observe and report several attributes of that well, including its depth, yield, water quality, and proximity to a pumping station. Every GIS provides the ability to store and manipulate both the spatial data and the associated attribute data.

12.3.4 People

The final component required for a true GIS is users. The term "user" may refer to any individual who will use GIS to support project or program goals, or to an entire organization that will employ GIS in support of its overall mission.

GIS users are often envisioned as hands-on computer processing people. While this is in part true, we choose to define a broader spectrum of GIS users. One classification scheme classifies users into two groups: **system users** and **end users**.

System users are those persons who have actual hands-on use of the GIS hardware and software. These persons have advanced technical skills in the application of GIS to problem solving. System users tend to be responsible not only for the day-to-day use of the system, but also for system maintenance and upkeep.

End users are those persons who do not have actual hands-on use of the system but who do make use of the information products generated via the GIS. End users do not necessarily have to possess hands-on technical skills. However, they must be able to communicate effectively and interact with system users in order to make requests for information products, and must also understand the limitations and requirements of GIS-based processing. GIS must not be a "black box" to end users. They should always be aware of the costs and benefits of a GIS approach to problem solving in order to determine if the requests they make of GIS are

NOTES

reasonable and to judge the quality of the output they receive.

Yet another type of user can be called **data generators**. These are people who collect the raw data to be entered into the GIS. Because a GIS requires such a strict structuring of the data (refer to Maps and Map Data Handling), data generators must often predicate data collection and compilation on the demands of the GIS. Thus, these persons must be made aware that the intended use of the GIS will play a major role in the way that they collect and package data.

In sum, we can see that there is a broad spectrum of GIS users, a spectrum that extends beyond the image of all users as "hands-on" GIS experts. GIS has the capacity to affect people at all stages of the data collection, management, manipulation, and interpretation process.

12.3.5 Method

The right method is a key for successful operation of GIS technology. The well-designed implementation plan and business rules are unique to each organization.

Generally GIS is considered to be expensive and difficult but with the advent of new technology like graphical user interface, powerful and affordable hardware and software it is gaining grounds and is slowly getting included in mainstream use.

A successful GIS operates according to a well-designed plan and business rules, which are the models and operating practices unique to each organization.

Methods include how the data will be retrieved, input into the system, stored, managed, transformed, analyzed, and finally presented in a final output. The procedures are the steps taken to answer the question need to be resolved. The ability of a GIS to perform spatial analysis and answer these questions is what differentiates this type of system from any other information systems. The transformation processes includes such tasks as adjusting the coordinate system, setting a projection, correcting any digitized errors in a data set, and converting data from vector to raster or raster to vector. (Carver, 1998)

12.4 Working of GIS

A GIS stores information about the world as a collection of thematic layers that can be linked together by geography. This simple but extremely powerful and versatile concept has proven invaluable for solving many real-world problems from tracking delivery vehicles, to recording details of planning applications, to modeling global atmospheric circulation.

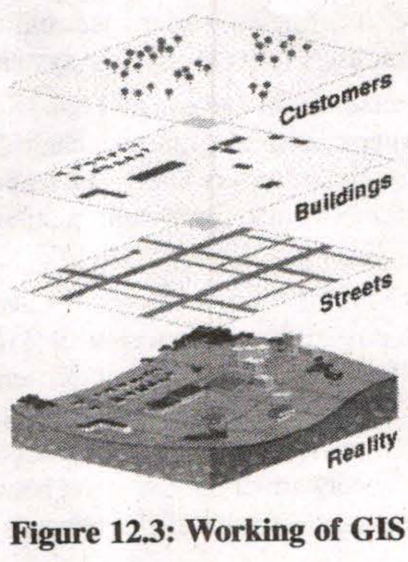


Figure 12.3: Working of GIS

NOTES

12.4.1 Geographic References

Geographic information contains either an explicit geographic reference, such as a latitude and longitude or national grid coordinate, or an implicit reference such as an address, postal code, census tract name, forest stand identifier, or road name. An automated process called geocoding is used to create explicit geographic references (multiple locations) from implicit references (descriptions such as addresses). These geographic references allow you to locate features, such as a business or forest stand, and events, such as an earthquake, on the earth's surface for analysis.

12.4.2 Vector and Raster Models

Geographic information systems work with two fundamentally different types of geographic models--the "vector" model and the "raster" model. In the vector model, information about points, lines, and polygons is encoded and stored as a collection of x,y coordinates. The location of a point feature, such as a bore hole, can be described by a single x,y coordinate. Linear features, such as roads and rivers, can be stored as a collection of point coordinates. Polygonal features, such as sales territories and river catchments, can be stored as a closed loop of coordinates.

The vector model is extremely useful for describing discrete features, but less useful for describing continuously varying features such as soil type or accessibility costs for hospitals. The raster model has evolved to model such continuous features. A raster image comprises a collection of grid cells rather like a scanned map or picture. Both the vector and raster models for storing geographic data have unique advantages and disadvantages. Modern GISs are able to handle both models.

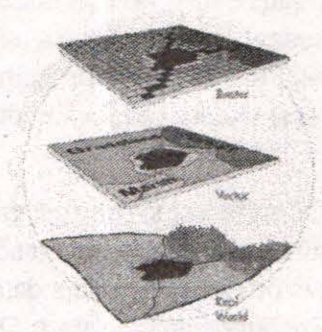


Figure 12.4: Vector & Raster Model

General purpose geographic information systems essentially perform six processes or tasks:

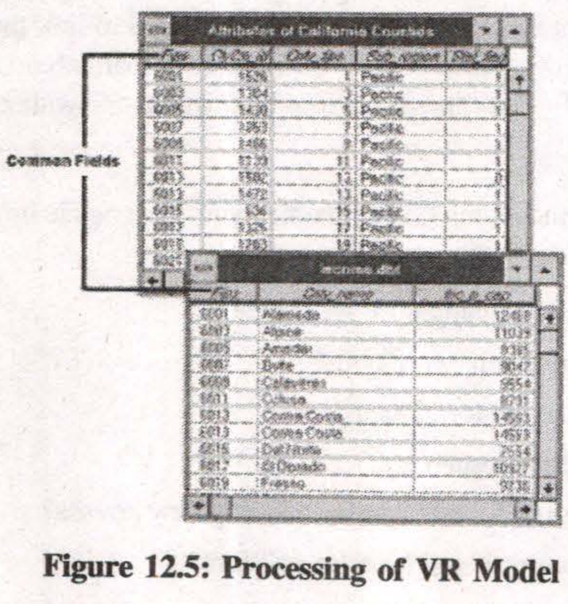


Figure 12.5: Processing of VR Model

NOTES

- Input
- Manipulation
- Management
- Query and Analysis
- Visualization

Input

Before geographic data can be used in a GIS, the data must be converted into a suitable digital format. The process of converting data from paper maps into computer files is called digitizing.

Modern GIS technology can automate this process fully for large projects using scanning technology; smaller jobs may require some manual digitizing (using a digitizing table). Today many types of geographic data already exist in GIS-compatible formats. These data can be obtained from data suppliers and loaded directly into a GIS.

Manipulation

It is likely that data types required for a particular GIS project will need to be transformed or manipulated in some way to make them compatible with your system. For example, geographic information is available at different scales (detailed street centerline files; less detailed census boundaries; and postal codes at a regional level). Before this information can be integrated, it must be transformed to the same scale (degree of detail or accuracy). This could be a temporary transformation for display purposes or a permanent one required for analysis. GIS technology offers many tools for manipulating spatial data and for weeding out unnecessary data.

Management

For small GIS projects it may be sufficient to store geographic information as simple files. However, when data volumes become large and the number of data users becomes more than a few, it is often best to use a database management system (DBMS) to help store, organize, and manage data.A DBMS is nothing more than computer software for managing a database.

There are many different designs of DBMSs, but in GIS the relational design has been the most useful. In the relational design, data are stored conceptually as a collection of tables. Common fields in different tables are used to link them together. This surprisingly simple design has been so widely used primarily because of its flexibility and very wide deployment in applications both within and without GIS.

Query and Analysis

Once you have a functioning GIS containing your geographic information, you can begin to ask simple questions such as

- Who owns the land parcel on the corner?
- How far is it between two places?
- Where is land zoned for industrial use?

And analytical questions such as

- Where are all the sites suitable for building new houses?
- What is the dominant soil type for oak forest?

NOTES

- If I build a new highway here, how will traffic be affected?

GIS provides both simple point-and-click query capabilities and sophisticated analysis tools to provide timely information to managers and analysts alike. GIS technology really comes into its own when used to analyze geographic data to look for patterns and trends and to undertake "what if" scenarios. Modern GISs have many powerful analytical tools, but two are especially important.

Proximity Analysis

- How many houses lie within 100 m of this water main?
- What is the total number of customers within 10 km of this store?
- What proportion of the alfalfa crop is within 500 m of the well?

To answer such questions, GIS technology uses a process called buffering to determine the proximity relationship between features.

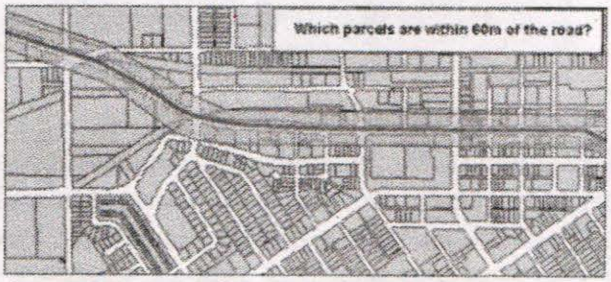


Figure 12.6: Proximity Relationship

Overlay Analysis

The integration of different data layers involves a process called overlay. At its simplest, this could be a visual operation, but analytical operations require one or more data layers to be joined physically. This overlay, or spatial join, can integrate data on soils, slope, and vegetation, or land ownership with tax assessment.

Visualization

For many types of geographic operation the end result is best visualized as a map or graph. Maps are very efficient at storing and communicating geographic information. While cartographers have created maps for millennia, GIS provides new and exciting tools to extend the art and science of cartography. Map displays can be integrated with reports, three-dimensional views, photographic images, and other output such as multimedia.

12.5 Data for GIS

Two basic types of data are normally entered into a GIS. The first type of data consists of real world phenomena and features that have some kind of spatial dimension. Usually, these data **elements** are depicted mathematically in the GIS as either points, lines, or polygons that are referenced geographically (or **geocoded**) to some type of coordinate system. This type data is entered into the GIS by devices like scanners, digitizers, GPS, air photos, and satellite imagery. The other type of data is sometimes referred to as an **attribute**. Attributes are pieces of data that are connected or related to the points, lines, or polygons mapped in the GIS. This attribute data can be analyzed to determine patterns of importance. Attribute

NOTES

data is entered directly into a database where it is associated with element data.

Now lets see what type of data is required by maps.

If you are unfamiliar with map data, think first about how you want to use map data. Many project needs are met with the following common map data types. Then explore these links to learn more about map data!

Base Maps : Include streets and highways; boundaries for census, postal, and political areas; rivers and lakes; parks and landmarks; place names; and USGS raster maps.

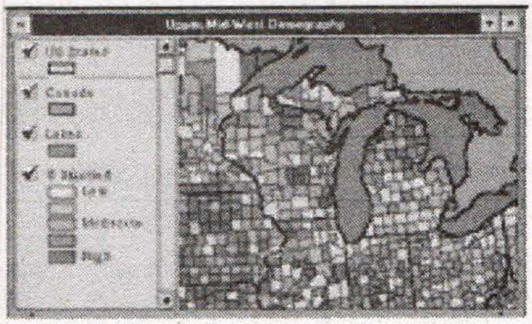


Figure 12.7: Basemaps

Business Maps and Data : Include data related to census/demography, consumer products, financial services, health care, real estate, telecommunications, emergency preparedness, crime, advertising, business establishments, and transportation.

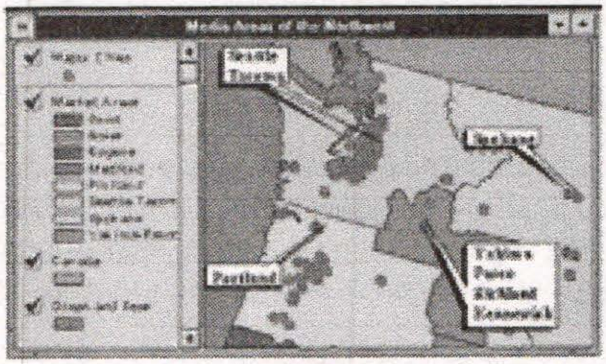


Figure 12.8: Business Maps & Data

Environmental Maps and Data : Include data related to the environment, weather, environmental risk, satellite imagery, topography, and natural resources.

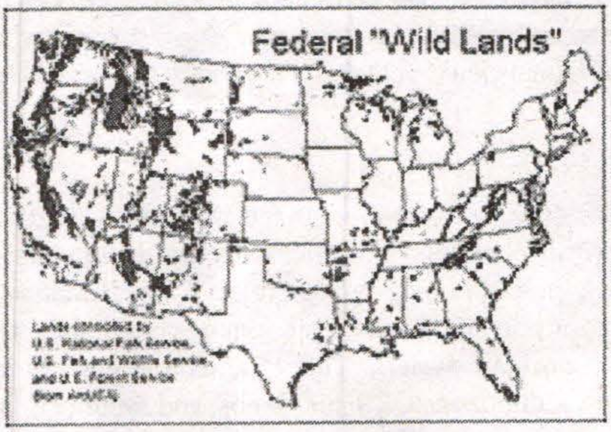


Figure 12.7: Environmental Maps

NOTES

resentative shape, and a symbol that represents one or more of its characteristics. Because features on maps are organized according to relative location or position, maps are good for showing the relationships among feature locations. These relationships, called spatial relationships, are important because understanding them helps us solve problems.

Maps use three basic shapes-- points, lines and areas to represent real-world objects. Points represent objects that have discrete locations and are too small to be depicted as areas. Lines represent objects that have length but are too narrow to be depicted as areas. Areas represent objects too large to be depicted as points or lines. Shapes alone do not give you enough information, so maps use graphic symbols to help identify features and provide information about them.

Most features can be represented as more than one shape. The scale of a map tells how the size of the map features compares with the size of the geographic objects they represent. Map scales vary from small-scale to large-scale.For example, on a small-scale map a city may be represented as a point.

With desktop GIS, you are not limited to the amount of information you can get about what you see on the map. Desktop GIS stores all the information about map features in a GIS database and links the features on the map to the information about them. This means that you can access all the information about a feature by simply clicking on it.For example, in Figure 3 (shown below) a table is shown on the left (highlighted in yellow) and a window titled a 'View' (highlighted in blue) is shown on the right.'Views' are made of 'themes'.'Themes' are used to create the map shown in Figure 3 and highlighted in red.

Figure 3.Window from a GIS that shows the 'View' area with 'Themes' shown on the right side of the picture.On the left side of the picture, a table is shown.This table contains data about the map.

The information that a desktop GIS stores about map features is referred to as attribute information, or attributes. The attributes of a river, for example, might include its name, length, average depth, rate of flow, water quality, how many dams are on it, and how many bridges cross it. Desktop GIS formats attributes in rows and columns, and stores them as tables. Each column stores a different attribute and each row relates to a single feature. The link between map features and their attributes is the basic principle behind how a desktop GIS works, and is the source of its power. Once the map features and attributes are linked, you can access the attributes for any map feature or locate any feature from its attributes in a table. GIS can also display features based on any attribute in the table. Because the link between features and attributes is a two-way relationship, changing an attribute in the table automatically results in a change on a map. Desktop GIS links sets of features and their attributes and manages them together in units called themes. A theme contains a set of related features, such as roads, streams, parcels, or wildlife habitat areas, along with the attributes for those features.

12.6.2 CAD

CAD systems evolved to create designs and plans of buildings and infrastructure. This activity required that components of fixed characteristics be assembled to create the whole structure. These systems require few rules to specify how com-

NOTES

General Reference Maps : World and country maps and data that can be a foundation for your database.

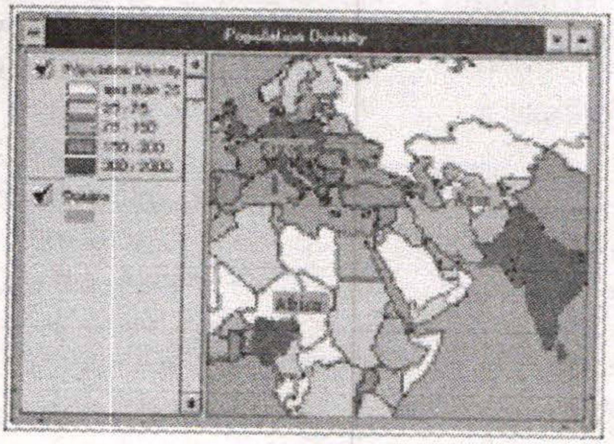


Figure 12.8: Geographic Display

How To Get Map Data?

Fortunately, volumes of existing geographic data are readily available. Through the ArcData Publishing Program, ESRI has established a partnership with leading commercial data vendors to provide a wealth of information in a plug-n-play format for use with ArcView GIS. ESRI's GIS Store and ArcData Online both offer a convenient way to get the most popular geographic data.

And, a variety of useful geographic data come bundled with ArcView GIS to help you get started quickly. These data sets can be used as the foundation for your GIS projects or to supplement your existing data.

12.6 GIS and Related Technologies

GISs are closely related to several other types of information systems, but it is the ability to manipulate and analyze geographic data that sets GIS technology apart. Although there are no hard and fast rules about how to classify information systems, the following discussion should help differentiate GIS from desktop mapping, computer-aided design (CAD), remote sensing, DBMS, and global positioning systems (GPS) technologies.

12.6.1 Desktop Mapping

A desktop mapping system uses the map metaphor to organize data and user interaction. The focus of such systems is the creation of maps: the map is the database. Most desktop mapping systems have more limited data management, spatial analysis, and customization capabilities. Desktop mapping systems operate on desktop computers such as PCs, Macintoshes, and smaller UNIX workstations.

Desktop GIS represents the real world on a computer similar to the way maps represent the world on paper.Both GIS and paper maps convey information about places. However, desktop GIS has power and flexibility that paper maps lack. The scale of the map influences the size of what appears on it. With GIS, however, you can store and link huge amounts of information about the objects represented on maps. These objects are called features. Each map feature has a location, a rep-

NOTES

ponents can be assembled and very limited analytical capabilities. CAD systems have been extended to support maps but typically have limited utility for managing and analyzing large geographic databases.

Computer-aided design (CAD) is the use of computer technology for the design of objects, real or virtual. The design of geometric models for object shapes, in particular, is often called computer-aided geometric design (CAGD).

However CAD often involves more than just shapes. As in the manual drafting of technical and engineering drawings, the output of CAD often must convey also symbolic information such as materials, processes, dimensions, and tolerances, according to application-specific conventions.

CAD may be used to design curves and figures in two-dimensional ("2D") space; or curves, surfaces, or solids in three-dimensional ("3D") objects.

CAD is an important industrial art extensively used in many applications, including automotive, shipbuilding, and aerospace industries, industrial and architectural design, prosthetics, and many more. CAD is also widely used to produce computer animation for special effects in movies, advertising, technical manuals. The modern ubiquity and power of computers means that even perfume bottles and shampoo dispensers are designed using techniques unheard of by shipbuilders of 1960s. Because of its enormous economic importance, CAD has been a major driving force for research in computational geometry, computer graphics (both hardware and software), and discrete differential geometry

The Effects of CAD

Starting in the late 1980s, the development of readily affordable Computer-Aided Design programs that could be run on personal computers began a trend of massive downsizing in drafting departments in many small to mid-size companies. As a general rule, one CAD operator could readily replace at least three to five drafters using traditional methods. Additionally, many engineers began to do their own drafting work, further eliminating the need for traditional drafting departments. This trend mirrored that of the elimination of many office jobs traditionally performed by a secretary as word processors, spreadsheets, databases, etc. became standard software packages that "everyone" was expected to learn.

Another consequence had been that since the latest advances were often quite expensive, small and even mid-size firms often could not compete against large firms who could use their computational edge for competitive purposes. Today, however, hardware and software costs have come down. Even high-end packages work on less expensive platforms and some even support multiple platforms. The costs associated with CAD implementation now are more heavily weighted to the costs of training in the use of these high level tools, the cost of integrating a CAD/CAM/CAE PLM using enterprise across multi-CAD and multi-platform environments and the costs of modifying design work flows to exploit the full advantage of CAD tools.

CAD vendors have effectively lowered these training costs. These methods can be split into three categories:

1. Improved and simplified user interfaces. This includes the availability of "role" specific tailorable user interfaces through which commands are presented to

NOTES

users in a form appropriate to their function and expertise.

2. Enhancements to application software. One such example is improved design-in-context, through the ability to model/edit a design component from within the context of a large, even multi-CAD, active digital mockup.
3. User oriented modeling options. This includes the ability to free the user from the need to understand the design intent history of a complex intelligent model.

12.6.3 Remote Sensing and GPS

Remote sensing is the art and science of making measurements of the earth using sensors such as cameras carried on airplanes, GPS receivers, or other devices. These sensors collect data in the form of images and provide specialized capabilities for manipulating, analyzing, and visualizing those images. Lacking strong geographic data management and analytical operations, they cannot be called true GISs.

In the broadest sense, **remote sensing** is the short or large-scale acquisition of information of an object or phenomenon, by the use of either recording or real-time sensing device(s) that is not in physical or intimate contact with the object (such as by way of **aircraft**, **spacecraft**, **satellite**, **buoy**, or **ship**). In practice, remote sensing is the stand-off collection through the use of a variety of devices for gathering information on a given object or area. Thus, **Earth observation** or **weather satellite** collection platforms, ocean and atmospheric observing **weather buoy** platforms, monitoring of a pregnancy via **ultrasound**, **Magnetic Resonance Imaging** (MRI), **Positron Emission Tomography** (PET), and **space probes** are all examples of remote sensing. In modern usage, the term generally refers to the use of imaging sensor technologies including but not limited to the use of instruments aboard aircraft and spacecraft, and is distinct from other imaging-related fields such as **medical imaging**.

There are two kinds of remote sensing. **Passive** sensors detect natural energy (radiation) that is emitted or reflected by the object or surrounding area being observed. Reflected sunlight is the most common source of radiation measured by passive sensors. Examples of passive remote sensors include film photography, infra-red, charge-coupled devices and radiometers. **Active** collection, on the other hand, emits energy in order scan objects and areas whereupon a passive sensor then detects and measures the radiation that is reflected or backscattered from the target. RADAR is an example of active remote sensing where the time delay between emission and return is measured, establishing the location, height, speed and direction of an object.

Remote sensing makes it possible to collect data on dangerous or inaccessible areas. Remote sensing applications include monitoring deforestation in areas such as the Amazon Basin, the effects of global warming on glaciers and Arctic and Antarctic regions, and depth sounding of coastal and ocean depths. Military collection during the cold war made use of stand-off collection of dangerous border areas. Remote sensing also replaces costly and slow collection on the ground, ensuring in the process that areas or objects are not disturbed.

Orbital platforms collect and transmit data from different parts of the electromagnetic spectrum, in conjunction with smaller scale aerial or ground-based sensing and analysis provides researchers with enough information to monitor trends .Other uses include different areas of earth sciences, to include natural resource management, agricultural fields such as land usage and conservation, and national security, both overhead, ground-based and stand-off collection on border areas.

12.6.4 DBMS

Database management systems specialize in the storage and management of all types of data including geographic data. DBMSs are optimized to store and retrieve data and many GISs rely on them for this purpose. They do not have the analytic and visualization tools common to GIS.

The elements in a vector based GIS are then the DBMS (Data Base Management System) for the attributes and the system that manages the topological data. In some GIS packages, the DBMS is based in an existing software, i.e. dBASE.

Entity-Rrelation Model

Three elements are considered in this approach: (a) Entities as the relevant objects for the data base. In a GIS, an entity is any fact that can be localised spatially. (b) Attributes or characteristics attached to the entities. Each attribute has a limited domain of possible values, i.e. the quality of a road can be bad, average, good, very good. (c) Relations or mechanisms that allow to relate entities. Some examples are: 'located in', 'contained in', 'crossed with', etc.

The data bases used in GIS are most commonly relational. Nevertheless, Object Oriented data bases are progressively incorporated.

Relational Data Bases

In a relational data base, data is stored in tables where rows represent the objects or entities and columns the attributes or variables. A data base is usually composed of several tables and the relations between them is possible through a common identifier that is unique for each entity. Most of the relational data bases in GIS present two variables with identifiers; one of them is unique and correlative, it could be numeric or alphabetic, and the second one might be repeated and helps to organise the attribute table.

The advantages of using this kind of data base are:

- The design is based in a methodology with heavy theoretical basis, which offers confidence in its capacity to evolve.
- It is very easy to implement it, specially in comparison with other models such as hierarchical, network, and object oriented.
- It is very flexible. New tables can be appended easily.
- Finally, many powerful DBMS using this approach contains query languages (like SQL) which makes easy to include this tool in a GIS. Thus, some commercialised GIS packages include a DBMS pre- existent.

NOTES

Object Oriented Databases

Based on objects, it can be defined as an entity with a localisation represented by values and by a group of operations. Thus, the advantage in comparison with relational data bases is based on the inclusion, in the definition of an object, not only its attributes but also the methods or operations that act on this object. In addition, the objects belong to classes that can have their own variables and these classes can belong to super-classes.

12.7 Spatial Data Infrastructures

Infrastructure Characteristics

A dictionary definition of an infrastructure is "the underlying foundation or basic framework (e.g., of a system or organization)" . It is the basic structural foundation of a society or enterprise, a substructure on which other components are based.

An infrastructure has the following characteristics:

- Its users are not conscious of its "ownership". Users are aware that 'somebody' maintains the infrastructure, but do not regard this maintainer as an owner. For example, nobody regards the Roads Department as the owner of the highways, rather, users think of them as custodians or maintainers of that infrastructure.
- As of result of the removal of the concept of ownership from the infrastructure, users take it for granted. They expect it to always be available, even if there is a few or other consideration for its use. For example, though we pay for electric power, we expect it to be there always.
- The user is responsible for any equipment required for exploiting or benefiting from the infrastructure.
- The delivery or provision of the service is standardized. A service that is delivered as infrastructures is delivered in that manner because it is the most effective and efficient way to deliver it. Imagine supplying water to a community without the usual water infrastructure of reservoirs, pumps and pipes. This would involve going to the source with a container to collect water, as in fact is done in rural villages where the infrastructure for water distribution has not been installed. Other services can only be delivered in the form of infrastructure and there is no imaginable alternative, e.g., transportation and electricity supplies.

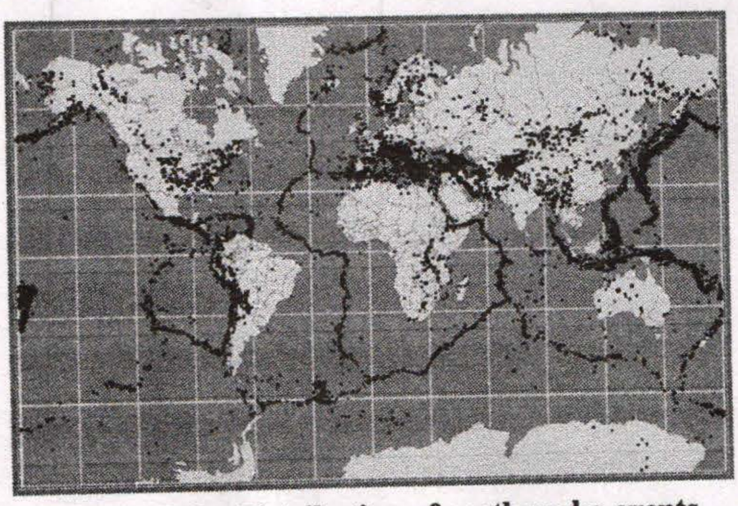


Figure 12.9 : Distribution of earthquake events

NOTES

13.8 GIS Terms

NAME	ACRONYM	USES
Computer Assisted Drawing and Drafting	CADD	Drawing packages used to generate and produce digitally compiled drawings. CADD data are often not geographic (mapped). Some data developed in CADD systems can be incorporated into a GIS.
Automated Mapping/ Facilities Management	AM/FM	Computerized map generation software that is most often utilized for infrastructure analysis and management. Similar to LIS technology.
Database Management System	DBMS	Software packages used to manage and analysis attribute data. Not used specifically for map analysis. Some GIS packages use DBMS to handle data management tasks.
Computer/Automated Cartography	C/AC	Computer software used automatically generate maps. Does not include integrated database or data analysis capabilities of GIS
Image Processing	IP	Sets of computerized routines used for information extraction (e.g., pattern recognition and classification) from remotely sensed images. Output from IP systems is often used as input to a GIS
Land Information System	LIS	Type of GIS that manages and analyzes data related to land ownership (e.g., tax parcels, urban infrastructure, property assessment).
Spatial Decision Support System	SDSS	A customized computer-based information system that utilizes decision rules and models and incorporates spatial data. These systems are often built from GIS databases, and may employ a subset of GIS functionality to meet user requirements.

12.9 Maps and Map Data Handling

Maps are fundamental tools used to portray spatial or geographic data. Data shown on maps vary with location. Map data are organized, classified, and depicted in a manner chosen by the map maker to optimize the map's effectiveness in communicating the nature of the data. When we read a map, we are looking for patterns, linkages, or relationships in the map's data. In some cases, we may look at several maps simultaneously to determine relationships between geographic data depicted on those separate maps.

All maps are simplifications of the real world. The true earth is infinitely complex and it is not possible to depict on a map all of the real earth features that we might be interested in. As a simplified image of the earth, maps can be called models of reality. Though they are simplifications, these models are quite sophisticated, and the science of map making, called cartography, is a formal geographic science.

Cartographers design and produce maps using very rigid rules and guidelines. A true map must accurately show not only what the nature of the mapped variable is (in correct proportion), but also must correctly place all mapped data in their true geographic locations. The geographic location of spatial data can be classified into two types: **absolute** and **relative**. Absolute location refers to a unique and stan-

NOTES

dardized place or position, while relative location defines position based on the location of other variables or phenomena. Absolute geographic position is specified using a universal coordinate system such as Latitude/Longitude or Universal Transverse Mercator(UTM) coordinates. These universal coordinate systems allow both the map maker and the map user to specify a unique and definite position for every location on both the earth and maps of the earth. This unique location, which is "tagged" to all spatial data, is critical to being able to store and analyze data in a GIS. It is this geographic tag or characteristic that distinguishes GIS as a technology that focuses on mapped data.

When we read a map, we rarely depend on absolute geographic information to assess patterns or relationships. In consulting a road map, for example, to decide on a route between Delhi to Mumbai, the traveler would probably not need to know the coordinates (e.g., latitude/longitude) of each city to select a travel route. Rather, the traveler would visualize the optional routes and come to some decision based on a personal interpretation of the "best" route. In selecting a route by looking at the geographical network of roads connecting the two cities, a casual map user would not seek a formal interpretation of the map. However, even a casual user would assume that all roads and other data depicted on the map are shown in their correct relative positions. Also, this type of user would assume that all mileage and other ancillary data such as the locations of rest stops were accurately reported on the map legend.

For a map to be geographically correct and correctly portray relationships between, for example, roads and cities, the cartographer must follow precise guidelines when compiling the map. For our purposes, we will simplify that set of rules to three parameters: **coordinate systems**, **projections**, and **scale**. Only when these three parameters are specified can a correct map be generated. In turn, an accurate GIS database depends on the availability and use of correct original map manuscripts.

These mapping guidelines prove essential when we convert traditional maps into computerized, GIS-compatible data sets. By specifying locations according to coordinate systems, a robust and accurate GIS can be established. Without a common set of geographic coordinates, layers in a GIS will not register (i.e., overlay precisely), thus precluding any type of multiple map operation such as overlaying two or more maps.

In turn, the ability to generate good location data is based on the accurate projection of the original map data. Projection is the process of transforming locations on the curved earth surface to locations on a two-dimensional plane (Figure 3). This mathematical step is performed by the cartographer to ensure geographic fidelity. While most of us never consider such a seemingly irrelevant source of error as the curvature of the earth, it is a critical consideration to the cartographer, and ultimately but perhaps unknowingly, to the conscientious map user.

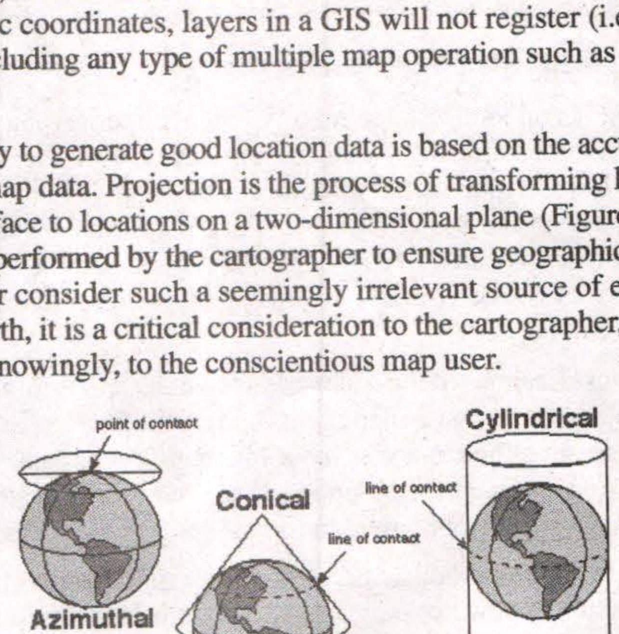


Figure 12.10 : Map Projections

NOTES

Generally there are three classes or families of map projections. Each geometric shape is used to transform the globe (a curved surface) to a plane (the map surface). These three families are called "developable" surfaces because planes, cones, and cylinders can be "flattened" without distortion. **Azimuthal or planar projections** use a flat two-dimensional surface to develop the map, **conical projections** are transformed onto a cone wrapped around the globe which is then flattened, and the **cylindrical** family of maps are projected onto a cylinder wrapped around the globe. Different projections are selected to minimize specific types of distortions in distance, direction, shape and area found on all flat maps.

Few GIS users will actually perform map projection or create a map from scratch. It is more likely that users will compile data from an existing map base, such as that provided by the U.S. Geological Survey's (USGS) National Mapping Program. Fortunately, these standard map series conform to rigid accuracy standards. Furthermore, the margin of each map contains information regarding the projection and coordinate systems that were used. Thus, USGS maps often provide a basis for the generation of other GIS data layers. As GIS developers build their system, and as users access and manipulate data from multiple maps, each must be aware of the need to standardize these parameters. Common projection and coordinate systems for all spatial data must be established if the map data in any GIS are to be fully integrated and registered.

GIS developers and users must also concern themselves with map scale. Scale is the mathematical relationship of real earth distance (ground distance) to that same distance as it is shown on a map. This relationship is often stated as a ratio of the two distances. As mentioned above, all maps are models or simplifications of reality. Maps also are reductions of reality. That is, the ratio of ground to map distance is normally much less than one. Maps to be stored in a GIS must be similar in scale if they are to be manipulated together. Maps with widely varying scales cannot be accurately combined. Thus, the user of either traditional paper maps or computerized GIS maps is fundamentally restricted by the degree of scale difference between map manuscripts. Maps with large differences in scale (e.g., 1:250,000 vs. 1:9,600) cannot be registered and overlaid without serious distortion and probable error.

A GIS converts traditional (usually paper) maps into a computer-compatible form. In order that the spatial data stored in the GIS is accurate, the map maker and the GIS user must consider issues of coordinate systems, projections, and scale. Every GIS requires that data be compatible with respect to these basic issues used in compiling mapped data.

As suggested by the example of the traveler selecting a route between two cities, often our analysis of the data shown on maps is informal. In other cases map users want specific measurements of the relationships between map features. Traditional map analysis is based on manual methods such as the use of a dot grid to measure area or a planimeter to measure distance or area. One of the most significant advantages afforded by a GIS is the ability to automate map analysis functions. That

NOTES

is, a GIS can be used to quantitatively assess the nature of spatial data stored in a GIS map database.

12.10 Traditional maps Vs. GIS

Traditional maps suffer some disadvantages over the maps that have been entered (automated) into a GIS. Traditional maps are static and fixed with respect not only to the data, but also with respect to projection, scale, and coordinate system. When a user needs to integrate maps of different scales, for example, it is often necessary to re-draft the maps. Map updates may similarly require a tedious manual process. Users also face practical limits on the number of maps that can be manually overlaid, using a light table for example. Similarly, it is difficult to combine multiple map sheets together into a seamless mosaic of maps which covers an area of interest that extends to more than one map sheet or manuscript. In short, a GIS is able to overcome some of the disadvantages of traditional map management and analysis.

Table 12.1

TRADITIONAL MAPS SYSTEMS	GEOGRAPHIC INFORMATION
Data are static	Data easy to update
Fixed projection, scale, coordinate system nate system	Can convert to new, scale, scale or coordi-
Quantitative analyses often tedious maps	Many manipulation options for analyzing
Difficult to combine multiple map sheets	Easy to combine multiple map layers
Overlays are restricted to a few layers database	Can overlay as many maps as contained in
Updates require re-drafting drafting	Tools allow for map updates without re-
Difficult to copy and share between many users	Multiple, simultaneous user access available
Lower overhead costs	Higher overhead costs
Paper maps usable in present form ment	Must convert map data to a digital environ-
Few Changes in technology	Technology changes rapidly

12.11 Functions of GIS

Another productive way to study GIS is through our original definition: a GIS is a computerized, integrated system used to compile, store, manipulate, and output mapped data. This section examines each of these functions.

12.11.1 Compilation

Data compilation involves assembling all of the spatial and attribute data that are to be stored in a computerized format within the GIS. Map data with common projections, scales, and coordinate systems must be pulled together in order to establish the centralized GIS database. Data must also be examined for compatibility in terms of content and time of data collection. Ultimately, the data will be stored in a GIS according to the specific format requirements set by both the user and the chosen GIS software/hardware environment.

When all of the common data requirements are set by the GIS user, a "base map" has been established. A base map is a set of standard requirements for data. It provides accurate standards for geographic control, and also defines a model or template that is used to shape all data into a compatible form. A base map is not necessarily a map ; rather, it is a comprehensive set of standards established and enacted to ensure quality control for the spatial and attribute data contained in the GIS.

Once the data are assembled and base map parameters are set, the user must translate the map and attribute data into computer-compatible form. This conversion process, referred to as "conversion" or "digitizing," converts paper maps into numerical digits that can be stored in the computer. Digitizing can be performed using various techniques. Scanning is one technique. Another technique is line digitizing which uses a tablet and a tracing stylus . Digitizing simplifies map data into sets of points, lines, or cells that can be stored in the GIS computer. Each GIS software package will impose a specific form and design on the way that these sets of points, lines, and cells are stored as digital map files.

Digitization is a simplification process that converts all spatial data to either a point (e.g., a well), a line (e.g., a stream), a polygon formed by a closed, complex line (e.g., a lake), or a grid cell. Digitization reduces all spatial entities to these simple forms because they are easy to store in the computer. A GIS database cannot readily recognize features or entities as human map users do. For example, we cannot enter the entity "lake" into a GIS. Rather, we enter the spatial data coordinates for the lake's shoreline as a polygon. Later, the attributes of the lake will be entered into the GIS database and will be associated with the polygon.

Following the digitization of map features, the user completes the compilation phase by relating all spatial features to their respective attributes, and by cleaning up and correcting errors introduced as a result of the data conversion process. The end result of compilation is a set of digital files, each accurately representing all of the spatial and attribute data of interest contained on the original map manuscripts. These digital files contain geographic coordinates for spatial objects (points, lines, polygons, and cells) that represent mapped features. Although we conceptualize the GIS as a set of registered map layers, the GIS actually stores these data at a much more primitive level.

12.11.2 Storage

Once the data have been digitally compiled, digital map files in the GIS are stored on magnetic or other (e.g., optical) digital media. Again, different GIS software packages will employ different storage formats. In most cases, however, data storage will be based on a generic data model that is used to convert map data into a digital form. The two most common types of data models are **raster** and **vector**.

NOTES

Both types are used to simplify the data shown on a map into a more basic form that can be easily and efficiently stored in the computer.

- **Raster**

  Raster approaches to storing map data in a GIS are perhaps the most intuitive. First, a girded matrix is registered to and overlaid on the original map manuscript. Location in the grid is defined by the row and column coordinates of each cell. To encode the map data for each cell in the raster format, three pieces of data are recorded: the **row coordinate**, the **column coordinate**, and the **attribute**. Thus a triplet of data is recorded for each cell in the array, which is termed a raster. After map data are stored in a raster format, each cell in the raster corresponds to a location on the map and each cell's location in the raster is identified by row and column coordinates. By assigning a value to each cell, the corresponding attribute data for that location are also stored. The end result of this conversion process is a set of cells, each with a specified location and an attribute value. These data can then be entered into a computer-compatible file and stored in the GIS database. Perhaps the most critical issue in using a raster GIS is the selection of an appropriate grid cell size. The user is forced to examine the trade-off between data resolution (how small grid cells are in the raster array) and storage requirements (increasing the number of rows and columns causes exponential increases in storage requirements). The use of smaller cells records greater detail in the GIS, so the user would normally attempt to select the smallest practical cell size. The choice of cell size depends on many factors, including the resolution of the original map data, the degree of resolution needed in the GIS analysis, the time and money available for data compilation, available storage space on the GIS computer, and cell sizes already employed for previously existing raster data that the user may want to incorporate into the GIS database. Cell size is critical because it is one of the base map parameters that must be standard for all of the layers in a raster GIS. While it is technically feasible to store data at different cell sizes, all analysis and multi-map manipulation operations require that cell sizes be the same (as well as projection, scale, and coordinate system). Typically, a GIS allows the user to adjust cell sizes within the database. The user must be aware of the nature of the data and not violate the limits of those data. For example, when attempting to combine two data layers that have been compiled and stored at two cell sizes, the user must increase the cell size on the layer having smaller cells to match that of the layer using larger cells. For this reason, the highest level of detail stored in such a database is only equal to the map layer having the largest cell size. This cell size becomes, in essence, the lowest common geography for the raster-based GIS.

- **Vector**

  A vector data structure is very different from a raster data model. Whereas the raster data model uses sets of grid cells to record all data, a vector model stores all spatial data as either a point, line, or polygon. These three types of spatial data are referred to as features, and a vector GIS can be termed a "feature-based technology." . When a vector model represents an entity as a point (e.g., a well), a single coordinate pair is used to specify its location. A feature represented as a line (e.g., a stream) uses a linked set of coordinates, and a feature represented as a polygon, which is an alternative form of a line (e.g., a lake), must have the same beginning and ending point coordinates. In a raster model, a point is a single cell, a line is a linked set of cells, and a polygon is a group or neighborhood of similarly encoded cells. For all three types of features stored in a vector GIS, an attribute code is entered into the GIS files to identify the object. For example, a lake would be vectorized (added to the database as a vector map) as a polygon by storing the linked set of coordinates for the shoreline and the attribute code "lake" would be entered in the GIS files to identify that group of cells.

12.11.3 Manipulation

Once data are stored in a GIS, many retrieval, analysis, and output options are available to users. These functions are often available in the form of "toolkits." A toolkit is a set of generic functions that a GIS user can employ to manipulate and analyze geographic data. Toolkits provide processing functions such as data retrieval, measuring area and perimeter, overlaying maps, performing map algebra, and reclassifying map data. A GIS usually includes a basic set of computer programs or "tools." The functions provided by the toolkit vary with the software package.

12.11.4 Output

The final functional task of a GIS is to generate output; usually a map. GIS-generated maps are compiled from the many data sets contained in the digital GIS and match exact user specifications. Map output may employ several color and symbology schemes, and will be sized and scaled to meet user needs. These output products resemble hand-drafted maps and fulfill essentially the same purposes. However, it is incorrect to refer to GIS simply as a mapping system. Although GIS is able to generate high-quality map output, its ability to perform analysis and management sets it apart from the more limited computer-mapping packages.

Another form of output from a GIS is tabular or report information. Data summarized according to user-defined classes or within user-defined areas can readily be generated in a textual format. This output may also be routed to another computer application such as a statistical analysis package or a graphing package for subsequent analysis and display.

The digital data themselves are often overlooked as a type of GIS output. Data files can be readily shared between users or systems. Because the data are in a digital format they can easily be copied, transmitted via cable or phone line, or distributed on media such as diskettes, computer-compatible tapes, or optical media such as compact disks. This greatly facilitates data sharing and provides increased access to data and information across the entire user spectrum.

Having completed a review of the basic elements and functions of a GIS, you can understand what GIS is designed to do. A few of the relative advantages and disadvantages of a GIS approach can now be assessed from a foundation of understanding.

12.12 Planning for GIS

NOTES

Because of the advantages afforded by GIS, its use is increasing rapidly. Many different types of organizations and users are interested in getting started with this new and exciting technology. Yet the developers of new GIS projects often find that selecting and implementing a GIS solution is a complex and demanding task. In all cases, new and existing GIS users are advised to engage in a strategic planning process prior to system acquisition, start-up, or the initiation of a new project on an existing system. Private consultants and vendors can assist with the planning process. It is also recommended that organizations planning to adopt GIS discuss the utility of the technology with existing GIS users who are performing similar tasks. The following phases are involved:

12.12.1 Planning

A planning process is the first stage in the life cycle. This phase involves a systematic review of users, their data, and their information needs. This is the time to educate decision makers about the costs and benefits of GIS and to include potential users in the planning process so that they receive an overview of the technology. Once a thorough understanding of user needs is established, the second life cycle phase, system design, can begin.

12.12.2 System Design

The design phase matches user needs to GIS functionality. Design includes not only selection of hardware and software, but also the design of the GIS spatial and attribute database. Part of that database design will include base map specifications for scale, projection, and coordinate systems. Data should be tracked using a "data dictionary." A data dictionary catalogs all data entered into the integrated GIS database and maintains current records on the source, accuracy, the time of data collection, and overall nature of each and every data element stored in the GIS. Establishment and maintenance of a robust data dictionary are essential to the overall utility of any GIS.

During the design phase an incremental plan is often anticipated for implementation of the technology. Incremental implementation means that users will build a GIS piece-by-piece. Oftentimes a prototype or pilot project is implemented so that lessons learned can be used to streamline the development of a fully-implemented system.

12.12.3 Implementation

During the implementation phase, attention to all user needs must be provided through training and education. Hands-on users must be trained to utilize and maintain the system and the database. All types of users should be made cognizant of how the GIS will affect them and their data processing tasks. They must also be made aware of the changes that GIS will introduce in the area of information generation and decision making.

12.12.4 Maintenance

Finally, a GIS application must be maintained and kept current in terms of data and user support. In some cases, a GIS is designed to meet the needs of a specific, finite project. In other instances, GIS is used to support an on-going mission or program. In the former case, the GIS application will terminate once the project is completed and maintenance will probably not be an issue. However, even if the initial GIS application is no longer being utilized, the data generated for the initial project may be useful to other projects or users. In those instances, a current data dictionary will be vital for determining the utility of the existing digital data for other uses.

In the case of an on-going GIS effort the system must be kept up-to-date in order to fulfill its design goals. Maintenance includes updating hardware and software, adding new data and updating existing data records, and keeping users current in terms of system functionality.

12.13 Implications of GIS

In any situation where it is applied, GIS may fundamentally change the way spatial data are managed and analyzed. When potential users are unaware of how GIS will affect them and their duties, the changes may be disturbing. GIS developers have a responsibility to initiate and maintain a focus on user issues and concerns. Ultimately, the success of a GIS venture is determined by the ability of the system to fulfill design goals. Those design goals are determined in the planning phase of the system life cycle, during which an understanding of data issues and user needs must be paramount. Potential users must be made cognizant of the role that GIS will play in their jobs, and should know that the responsibility for the success of the GIS effort ultimately resides on how well the system responds to their needs and demands.

An organization adopting GIS will need to make adjustments in user access to, and control of, information. Because of the dynamic nature of GIS, it is often difficult to track data and maintain up-to-date records on data status. Data can be copied, shared, and modified with ease in automated systems. Not all data transactions can be monitored; user access must often be restricted so that the integrity of the system is not jeopardized. Indeed, system security is a significant issue, particularly in a networked computing environment. A current data dictionary is very helpful in monitoring data integrity.

The organization must also adapt to changes in the cost of information. Because GIS requires such a strict structuring of spatial and attribute data, system developers and users must be prepared to deal with the overhead costs associated with a GIS approach. As mentioned previously, the process of converting existing mapped data into a computer-compatible format is time-consuming and costly. Conversely, a very robust data environment can be established via GIS when proper stipulations and standards are imposed on data development and use. In a GIS, data can be shared efficiently and effectively, and formerly diverse and incompatible databases can be integrated.

NOTES

The GIS that results from development efforts is ultimately only as good as the data contained in the system. A GIS will not fundamentally change the nature of spatial and attribute data, nor will it correct basic inaccuracies in data. GIS provides a rigorous structure for spatial and attribute data, but bad or incorrect data of both types can be made to fit a base map standard. The bottom line is, just because data are in a GIS does not mean that they are correct.

Users should be made aware of the limitations of the GIS approach. For example, a typical GIS can technically adjust the scale of any map to user specifications. It is up to the user not to violate the precision of the data, even if the software will allow this function to be performed. For example, if data are mapped at a relatively small scale of 1:250,000, the user must decide if rescaling the map to a larger scale of 1:15,840 presumes more precision or accuracy than really exists in the data.

The powerful output capabilities afforded by GIS in producing maps or reports on paper or as displays on computer monitors provide unparalleled opportunities for visualizing spatial data. Users who generate these products must exercise care in presenting information accurately and honestly. Persons who produce information from a GIS should be aware of the intended audience and the type of information that needs to be communicated. Similarly, users of GIS output should exercise care in interpreting and drawing conclusions from those information products. In every case, knowing something about the data is critical to proper and complete use of GIS output.

12.14 Virtual Globe

A **virtual globe** is a **3D software** model or representation of the **Earth** or another world. A virtual globe provides the user with the ability to freely move around in the virtual environment by changing the viewing angle and position. Compared to a conventional **globe**, virtual globes have the additional capability of representing many different views on the surface of the Earth. These views may be of geographical features, man-made features such as roads and buildings or abstract representations of demographic quantities such as population.

Virtual Globe Types

Virtual globes may be used for study or navigation (by connecting to a **GPS** device) and their design varies considerably according to their purpose. Those wishing to portray a visually accurate representation of the Earth often use satellite image servers and are capable not only of rotation but also zooming and sometimes horizon tilting. Very often such virtual globes aim to provide as true a representation of the world as is possible with worldwide coverage up to a very detailed level. When this is the case the interface often has the option of providing simplified graphical overlays to highlight man-made features since these are not necessarily obvious from a photographic aerial view. The other issue raised by such detail available is that of security with some governments having raised concerns about the ease of access to detailed views of sensitive locations such as airports and military bases.

Another type of virtual globe exists whose aim is not the accurate representation of the planet but instead a simplified graphical depiction. Most early computerized atlases were of this type and, while displaying less detail, these simplified interfaces are still widespread since they are faster to use because of the reduced graphics content and the speed with which the user can understand the display.

Online virtual globes

As more and more high-resolution **satellite imagery** and **aerial photography** become accessible for free, many of the latest online virtual globes are built to fetch and display these images. They include:

- **Norkart Virtual Globe**, first version released summer 2001 as the **SINTEF** Virtual Globe.
- **NASA World Wind**, USGS **topographic maps** and several satellite and aerial image datasets .
- **Google Earth**, satellite & aerial photos dataset (including commercial **Digital Globe** images) with international road dataset .
- **Virtual Earth 3D**, 3D interface to the Windows Live Local maps and photos, runs inside a web browser (**IE** only) .
- **ESRI ArcGIS Explorer** a lightweight client for **ArcGIS Server**, supports **WMS** and many other **GIS file formats** .
- **SkylineGlobe** virtual globe that runs inside a web browser .
- **Earthsim**, from **Servan Keondjian** which displays real-time 3D rendered views and information about the Earth.
- **Software MacKiev's** 3D Weather Globe & Atlas, 3D views based on the **Blue Marble** imagery, near-real-time cloud coverage and weather forecast from **CustomWeather**, time zones, day/night views.

12.15 GIS Softwares

Open source software

Most widely used **open source** applications:

- **GRASS** – Originally developed by the **U.S. Army Corps of Engineers**, open source: a complete GIS .
- **MapServer** – Web-based mapping server, developed by the **University of Minnesota**.
- **Chameleon** – Environments for building applications with MapServer.
- **GeoTools** – Open source GIS toolkit written in **Java**, using **Open Geospatial Consortium** specifications.
- **gvSIG** – Open source GIS written in Java.
- **JUMP GIS** – Java Unified Mapping Platform.
- **MapWindow GIS** – Free, open source GIS desktop application and programming component.

NOTES

- **PostGIS** – Spatial extensions for the open source **PostgreSQL** database, allowing geospatial queries.
- **Quantum GIS** – QGIS is a user friendly Open Source GIS that runs on Linux, Unix, Mac OSX, and Windows.
- **TerraView** – GIS desktop that handles vector and raster data stored in a relational or geo-relational database, including ACCESS, PostGreSQL, MySQL and Oracle Spatial.

Commercial GIS softwares

Most widely used **notable proprietary software** applications and providers:

- **Autodesk** – Products include **MapGuide** and other products that interface with its flagship **AutoCAD** software package.
- **Cadcorp** – Developers of world leading GIS software and OpenGIS standard (e.g. Read/Write Open Source PostGIS database).
- **ESRI** – Products include **ArcView 3.x**, **ArcGIS**, **ArcSDE**, **ArcIMS**, and **ArcWeb** services.
- **IDRISI** – Proprietary GIS product developed by Clark Labs.
- **Intergraph** – Products include **GeoMedia** GIS software, as well as
- **photogrammetry** and other mapping-related software. **MapInfo** – Products include **MapInfo Professional** and **MapXtreme**. integrates GIS software, data and services.
- **MapPoint** – Proprietary GIS product developed by Microsoft.
- **Caliper** – Products include **Maptitude**, **TransCAD** and **TransModeler**. Develops GIS and the only GIS for transportation.
- **CARIS** (Computer Aided Resource Information System) – GIS systems for **hydrography** and **cadastral** systems.
- **CartaVista** – Geographic Information Visualization (GIV) product developed by A3Dt.
- **DeLorme** – Producer **XMap** and other GIS tools, data, and GPS hardware.
- **GMS** – Three-dimensional environment for building geologic and groundwater models .
- **Maria** – A Windows-based GIS product developed by Teleplan Globe AS.
- **Manifold System** – Low-cost GIS software package.
- **Map Maker** – Low-cost Windows-based GIS software.
- **Oracle Spatial** – Product allows users to perform basic geographic operations and store common spatial data types in a native Oracle environment.
- **Safe Software** – Spatial ETL products including **FME**, SpatialDirect and the **ArcGIS** Data Interoperability Extension.
- **Smallworld** – GIS product developed in Cambridge, England (Smallworld, Inc.) and purchased by **General Electric** in 2000 and used primarily by **public utilities** and other related industries.

- **TatukGIS** – Products include: TatukGIS Developer Kernel (a comprehensive GIS development toolkit), Internet Map Server, GIS Editor, free GIS Viewer, Aerial Imagery Corrector.
- **GeoBase** – Geospatial platform developed by Telogis. A particular focus is placed on real-time processing for reverse-geocoding, geofencing, etc.
- **LandSerf** – Free GIS written in Java. Source available but not strictly open source.
- **Panorama** – Russian GIS for military uses.
- **SPRING** – GIS software developed at **INPE - Instituto Nacional de Pesquisas Espaciais** and available free of charge.
- **TerraLib** – GIS class and functions library, available from the Internet as open source, allowing a collaborative environment and its use for the development of multiple GIS tools.
- **TNTmips** – Geospatial analysis system providing a GIS, RDBMS, and automated image processing system with CAD, TIN, surface modeling, and data publishing tools.
- **SavGIS** – Free and complete GIS software available in French, English and Spanish, developed since 1984 by the Development Research French Institute **(IRD)**

SUMMARY

- A very useful definition of GIS is a true systems-based definition: *a GIS is a computerized, integrated system used to compile, store, manipulate, and output mapped spatial data.*
- A geographic information system (GIS) is a "computer-based information system that enables capture, modeling, manipulation, retrieval, analysis, and presentation of geographically referenced data". The main technology components, geographic information systems (GIS), consist of software packages that are capable of integrating spatial and non-spatial data to yield the spatial information that is used in decision making. They are computer-based equipment, procedures and techniques for manipulating spatial or map data.
- All data in a GIS are either spatial data or attribute data.
- Every GIS provides the ability to store and manipulate both the spatial data and the associated attribute data.
- CADD (Computer Assisted Drawing and Drafting): Drawing packages used to generate and produce digitally compiled drawings. CADD data are often not geographic (mapped). Some data developed in CADD systems can be incorporated into a GIS.
- The geographic location of spatial data can be classified into two types: absolute and relative.
- A GIS converts traditional (usually paper) maps into a computer-compatible form. In order that the spatial data stored in the GIS be accurate, the map maker and the GIS user must consider issues of coordinate systems, projections, and scale. Every GIS requires that data be compatible with respect to these basic issues used in compiling mapped data.

NOTES

- Data compilation involves assembling all of the spatial and attribute data that are to be stored in a computerized format within the GIS.
- Digitization is a simplification process that converts all spatial data to either a point (e.g., a well), a line (e.g., a stream), a polygon formed by a closed, complex line (e.g., a lake), or a grid cell. Digitization reduces all spatial entities to these simple forms because they are easy to store in the computer.
- The two most common types of data models are raster and vector. Both types are used to simplify the data shown on a map into a more basic form that can be easily and efficiently stored in the computer.
- During the implementation phase, attention to all user needs must be provided through training and education.
- A virtual globe is a 3D software model or representation of the Earth or another world. A virtual globe provides the user with the ability to freely move around in the virtual environment by changing the viewing angle and position.
- Remote sensing is the short or large-scale acquisition of information of an object or phenomenon, by the use of either recording or real-time sensing device(s) that is not in physical or intimate contact with the object (such as by way of aircraft, spacecraft, satellite, buoy, or ship).

EXERCISE

1. What is geographic information system (GIS) ?
2. What are the activities normally carried out on a GIS ?
3. Describe the historical background of GIS ?
4. Explain the spatial data infrastructure .
5. What are different components of GIS ?
6. Explain the GIS data format .
7. Describe the map data handling .
8. What is map projection ? Describe its types .
9. Compare and contrast the traditional map and GIS .
10. What are the functions of GIS ?
11. Describe the different phases of planning GIS .
12. What are the implications of GIS ?
13. What is virtual globe ? Describe its types .
14. Briefly explain the Related technology of GIS and working of GIS.
15. List most popular GIS softwares.

13 Introduction and Basic Concepts of Modern Communication and Telephony Technology

Chapter Includes :

- INTRODUCTION
- CODE DIVISION MULTIPLE ACCESS
- WIRELESS LOCAL LOOP
- GSM (GLOBAL SYSTEM FOR MOBILE COMMUNICATION)
- VOICE OVER IP
- BLUETOOTH
- WI-FI
- ISDN

13.1 Introduction

Communication allows people to exchange information by one of several methods. There are auditory means, such as speaking or singing, and nonverbal, physical means, such as body language, sign language, paralanguage, touch or eye contact.

Communication is a process by which information is exchanged between or among individuals through a common system of symbols, signs, and behavior. As a process, communication has synonyms such as expressing feelings, conversing, speaking, corresponding, writing, listening and exchanging. People communicate to satisfy needs in both their work and non-work lives. People want to be heard, to be appreciated and to be wanted. They also want to accomplish tasks and to achieve goals. Obviously, then, a major purpose of communication is to help people feel good about themselves and about their friends, groups, and organizations. For true communication, there must be a transmission of thoughts, ideas and feelings from one mind to another. However, human language is very different from plant communication.

NOTES

The beginning of human communication through artificial channels, i.e. not vocalization or gestures, goes back to ancient cave paintings, drawn maps, and writing.

Our indebtedness to the Ancient Romans in the field of communication does not end with the Latin root "communicate". They devised what might be described as the first real mail or postal system in order to centralize control of the empire from Rome. This allowed for personal letters and for Rome to gather knowledge about events in its many widespread provinces.

In the last century, a revolution in telecommunications has greatly altered communication by providing new media for long distance communication. The first transatlantic two-way radio broadcast occurred on July 25, 1920 and led to common communication via analogue and digital media:

- Analog telecommunications include traditional telephony, radio, and TV broadcasts.
- Digital telecommunications allow for computer-mediated communication, telegraphy, and computer networks.

Communications media impact more than the reach of messages. They impact content and customs; for example, Thomas Edison had to discover that *hello* was the least ambiguous greeting by voice over a distance; previous greetings such as *hail* tended to be garbled in the transmission. Similarly, the terseness of e-mail and chat rooms produced the need for the emoticon.

Modern communication media now allow for intense long-distance exchanges between larger numbers of people (many-to-many communication via e-mail, Internet forums). On the other hand, many traditional broadcast media and mass media favor one-to-many communication (television, cinema, radio, newspaper, magazines).

The latest trend in communication, termed smartmobbing, involves ad-hoc organization through mobile devices, allowing for effective many-to-many communication and social networking.

13.2 Code Division Multiple Access

Codes with certain characteristics are applied to the transmission to enable the use of code division multiplexing (CDM). CDMA systems use exactly these codes to separate different users in code space and to enable access to a shared medium without interference. The main problem is how to find "good" codes and how to separate the signals from noise generated by other signal and the environment. The code directly controls the chipping sequence. Now, good code means a code for a certain user should have a good auto correlation and should be orthogonal to the other codes. Auto correlation means the absolute value of the inner product of a vector multiplied with itself should be large. The inner Product of two vectors a and b with $a = (a_1, a_2 \ldots a_n)$ and $b = (b_1, b_2 \ldots b_n)$ is defined as –

$$a * b = \sum_{i=1}^{n} a_i b_i$$

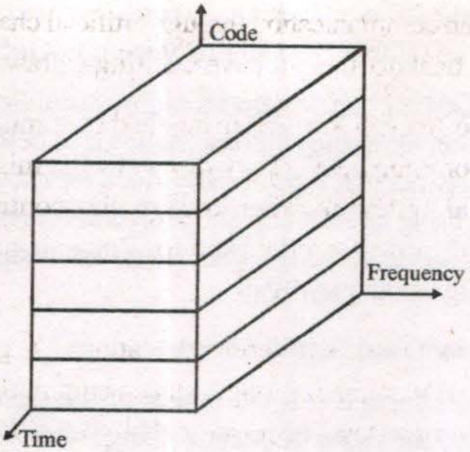


Figure 13.1 : Signal of CDMA

13.2.1 CDMA in Mobile Telephony

A number of different terms are used to refer to CDMA implementations. The original U.S. standard defined by QUALCOMM was known as IS-95, the IS referring to an Interim Standard of the Telecommunications Industry Association (TIA). IS-95 is often referred to as 2G or second generation cellular. The QUALCOMM brand name cdmaOne may also be used to refer to the 2G CDMA standard. CDMA has been submitted for approval as a mobile air interface standard to the ITU International Telecommunication Union.

Whereas the Global System for Mobile Communications (GSM) standard is a specification of an entire network infrastructure, the CDMA interface relates only to the air interface—the radio part of the technology. For example, GSM specifies an infrastructure based on internationally approved standard while CDMA allows each operator to provide the network features as it finds suited. On the air interface, the signalling suite (GSM: ISDN SS7) work has been progressing to harmonise these.

After a couple of revisions, IS-95 was superseded by the IS-2000 standard. This standard was introduced to meet some of the criteria laid out in the IMT-2000 specification for 3G, or third generation, cellular. It is also referred to as 1xRTT which simply means "1 times Radio Transmission Technology" and indicates that IS-2000 uses the same 1.25 MHz carrier shared channel as the original IS-95 standard. A related scheme called 3xRTT uses three 1.25 MHz carriers for a 3.75 MHz bandwidth that would allow higher data burst rates for an individual user, but the 3xRTT scheme has not been commercially deployed. More recently, QUALCOMM has led the creation of a new CDMA-based technology called 1xEV-DO, or IS-856, which provides the higher packet data transmission rates required by IMT-2000 and desired by wireless network operators.

This CDMA system is frequently confused with a similar but incompatible technology called Wideband Code Division Multiple Access (W-CDMA) which forms the basis of the W-CDMA air interface. The W-CDMA air interface is used in the global 3G standard UMTS and the Japanese 3G standard FOMA, by NTT DoCoMo and Vodafone; however, the CDMA family of US national standards (including cdmaOne and CDMA2000) are not compatible with the W-CDMA family of International Telecommunication Union (ITU) standards.

NOTES

Another important application of CDMA — predating and entirely distinct from CDMA cellular — is the Global Positioning System or GPS.

The QUALCOMM CDMA system includes highly accurate time signals (usually referenced to a GPS receiver in the cell base station), so cell phone CDMA-based clocks are an increasingly popular type of radio clock for use in computer networks. The main advantage of using CDMA cell phone signals for reference clock purposes is that they work better inside buildings, thus often eliminating the need to mount a GPS antenna on the outside of a building.

13.2.2 CDMA Coverage and Applications

The size of a given cell depends on the power of the signal transmitted by the handset, the terrain, and the radio frequency being used. Various algorithms can reduce the noise introduced by variations in terrain, but require extra information be sent to validate the transfer. Hence, the radio frequency and power of the handset effectively determine the cell size. Long wavelengths need less energy to travel a given distance vs. short wavelengths, so lower frequencies generally result in greater coverage while higher frequencies result in lesser coverage. These characteristics are used by mobile network planners in determining the size and placement of the cells in the network. In cities, many small cells are needed; the use of high frequencies allows sites to be placed more-closely together, with more subscribers provided service. In rural areas with a lower density of subscribers, use of lower frequencies allows each site to provide broader coverage.

Various companies use different variants of CDMA to provide fixed-line networks using Wireless local loop (WLL) technology. Since they can plan with a specific number of subscribers per cell in mind, and these are all stationary, this application of CDMA can be found in most parts of the world.

CDMA is suited for data transfer with bursty behaviour and where delays can be expected. It is therefore used in Wireless LAN applications; the cell size here is 150 m (500 feet) because of the high frequency (2.4 GHz) and low power. The suitability for data transfer is the reason for why W-CDMA seems to be "winning technology" for the data portion of third-generation (3G) mobile cellular networks.

13.2.3 Asynchronous CDMA

The mobile-to-base links cannot be precisely coordinated, particularly due to the mobility of the handsets, and require a somewhat different approach. Since it is not mathematically possible to create signature sequences that are orthogonal for arbitrarily random starting points, unique "pseudo-random" or "pseudo-noise" (PN) sequences are used in *Asynchronous* CDMA systems. These PN sequences are statistically uncorrelated, and the sum of a large number of PN sequences results in Multiple Access Interference (MAI) that is approximated by a Gaussian noise process (following the "central limit theorem" in statistics). If all of the users are received with the same power level, then the variance (e.g., the noise power) of the MAI increases in direct proportion to the number of users.

All forms of CDMA use spread spectrum process gain to allow receivers to partially discriminate against unwanted signals. Signals encoded with the specified PN sequence (code) are received, while signals with different codes (or the same code but a different timing offset) appear as wideband noise reduced by the process gain.

Since each user generates MAI, controlling the signal strength is an important issue with CDMA transmitters. A CDM (Synchronous CDMA), TDMA or FDMA receiver can in theory completely reject arbitrarily strong signals using different codes, time slots or frequency channels due to the orthogonality of these systems. This is not true for Asynchronous CDMA; rejection of unwanted signals is only partial. If any or all of the unwanted signals are much stronger than the desired signal, they will overwhelm it. This leads to a general requirement in any Asynchronous CDMA system to approximately match the various signal power levels as seen at the receiver. In CDMA cellular, the base station uses a fast closed-loop power control scheme to tightly control each mobile's transmit power.

13.2.4 Advantages of Asynchronous CDMA

Asynchronous CDMA's main advantage over CDM (*Synchronous* CDMA), TDMA and FDMA is that it can use the spectrum more efficiently in mobile telephony applications. (In theory, CDMA, TDMA and FDMA have exactly the same spectral efficiency. When it comes to practical application, each has its own challenges. Timing in the case of TDMA, power control in the case of CDMA and frequency generation/filtering in the case of FDMA.). TDMA systems must carefully synchronize the transmission times of all the users to ensure that they are received in the correct timeslot and do not cause interference. Since this cannot be perfectly controlled in a mobile environment, each timeslot must have a guard-time, which reduces the probability that users will interfere, but decreases the spectral efficiency. Similarly, FDMA systems must use a guard-band between adjacent channels, due to the random doppler shift of the signal spectrum which occurs due to the user's mobility. The guard-bands will reduce the probability that adjacent channels will interfere, but decrease the utilization of the spectrum.

Most importantly, Asynchronous CDMA offers a key advantage in the flexible allocation of resources. There are a fixed number of orthogonal codes, timeslots or frequency bands that can be allocated for CDM, TDMA and FDMA systems, which remain underutilized due to the bursty nature of telephony and packetized data transmissions. There is no strict limit to the number of users that can be supported in an *Asynchronous* CDMA system, only a practical limit governed by the desired bit error probability, since the SIR (Signal to Interference Ratio) varies inversely with the number of users. In a bursty traffic environment like mobile telephony, the advantage afforded by Asynchronous CDMA is that the performance (bit error rate) is allowed to fluctuate randomly, with an average value determined by the number of users times the percentage of utilization. Suppose there are 2N users that only talk half of the time, then 2N users can be accommodated with the same *average* bit error probability as N users that talk all of the time. The key difference here is that the bit error probability for N users talking all of the time is constant, whereas it is a *random* quantity (with the same mean) for 2N users talking half of the time.

In other words, Asynchronous CDMA is ideally suited to a mobile network where large numbers of transmitters each generate a relatively small amount of traffic at irregular intervals. CDM (*Synchronous* CDMA), TDMA and FDMA systems cannot recover the underutilized resources inherent to bursty traffic due to the fixed number of orthogonal codes, time slots or frequency channels that can be assigned

NOTES

to individual transmitters. For instance, if there are N time slots in a TDMA system and 2N users that talk half of the time, then half of the time there will be more than N users needing to use more than N timeslots. Furthermore, it would require significant overhead to continually allocate and deallocate the orthogonal code, time-slot or frequency channel resources. By comparison, Asynchronous CDMA transmitters simply send when they have something to say, and go off the air when they don't, keeping the same PN signature sequence as long as they are connected to the system.

13.2.5 Features of CDMA

Many users of a CDMA system share the same frequency. Either TDD or FDD may be used.

1. CDMA has a soft capacity limit increasing the number of users in a CDMA system.
2. Multipart fading may be substantially reduced because the signals is spread over a large spectrum. If the spread spectrum bandwidth is greater than the coherence bandwidth of the channel the inherent frequency diversity effect of small scale fading.
3. In CDMA system, channel data rate are very high.
4. In CDMA system, self Jamming is a Problem.
5. The near-far problem occurs at a CDMA receiver if an undesired user has a high detected power as compared to the desired user.

13.3 Wireless Local Loop

Wireless local loop (WLL), also called Broadband Wireless Access (BWA) radio in the loop (RITL) or fixed-radio access (FRA) or fixed-wireless access (FWA) or Fixed Wireless Terminal (FWT), is the use of a wireless communications link as the "last mile / first mile" connection for delivering plain old telephone service (POTS) and broadband Internet to telecommunications customers. Various types of WLL systems and technologies exist.

13.3.1 Fixed Wireless Service

Fixed Wireless Terminals or FWT units differ from conventional mobile terminal units operating within cellular networks -such as GSM- in that a fixed wireless terminal or deskphone will be limited to an almost permanent location with almost no roaming or *find-me anywhere* facilities.

WLL and FWTs are generic terms for radio based telecommunications technologies and the respective devices which can be implemented using a number of different wireless and radio technologies.

WiMAX (or IEEE 802.16)

Currently more operators are running on the 802.11 MAC at 2 and 5 GHz. 802.16 is unlikely to outperform 802.11 until at least late 2007. It may become the dominant medium for wireless local loop. Intel is promoting this standard, while Atheros and Broadcom are still focused largely on 802.11. Atheros, using its highly successful 802.11 OFDM chipsets, will likely be able to deliver comparable service levels to Intel's 802.16 TDM OFDM chipsets for the foreseeable future.

Mobile Technologies

These are available in Code Division Multiple Access(CDMA), Digital Enhanced Cordless Telecommunications - DECT (TDMA/DCA), Global System for Mobile Communications(GSM), IS136 Time Division Multiple Access(TDMA) as well as analog access technologies such as Advanced Mobile Phone System(AMPS), for which there will be independent standards defining every aspect of modulation, protocols, error handling, etc.

Deployment

The Wireless Local Loop market is currently an extremely high growth market, offering Internet Service Providers immediate access to customer markets without having to either lay cable through a metropolitan area MTA, or work through the ILEC's, reselling the telephone, cable or satellite networks, owned by companies that prefer to largely sell direct.

This trend revived the prospects for local and regional ISPs, as those willing to deploy fixed wireless networks were not at the mercy of the large telecommunication monopolies. They were at the mercy of unregulated re-use of unlicensed frequencies upon which they communicate.

Due to the enormous quantity of 802.11 "Wi-Fi" equipment and software, coupled with the fact that spectrum licensed are not requires in the ISM and UNII bands, the Industry has moved well ahead of the regulators and the standards bodies.

Sprint and ClearWire are preparing to roll out massive WiMAX networks in the United States.

13.3.2 WLL Standards

Mobile:

- CDMA (USA).
- TDMA (USA).
- GSM (ITU - Worldwide).
- UMTS 3rd Generation (World).
- Personal Handy-phone System (PHS in Japan, PAS/Xiaolingtong in China)

Fixed or local area network:

- DECT, for local loop
- corDECT (variant of DECT originates from India)
- LMDS
- 802.11, originally designed for short range mobile internet and network access service, it has emerged as the facto standard for Wireless Local Loop.

WiMAX or IEEE 802.16 may become the dominant medium for wireless local loop. Currently more operators are running on the 802.11 MAC at 2 and 5 GHz. 802.16 is unlikely to outperform 802.11 until at least late 2007. Intel is promoting this standard, while Atheros and Broadcom are still focused largely on 802.11.

13.4 GSM (Global System for Mobile Communication)

GSM is a second generation cellular system or it is world most popular 2G technology. It was developed to solve the fragmentation problems of the first cellular

NOTES

systems in Europe. GSM also specify digital modulation and network level architecture and services.

The work of specifying a common mobile communication system for europe in the 900 MHz band was taken up on the mid-1980, by the GSM which was a working group of the CEPT (Conference European of Postal and Telecommunication Administrations). In 1992, GSM changed its name to the Global system for mobile communications for marketing reasons. By the and of 1993, several non-european countries is South America, Asia and Australia had Adopted GSM.

GSM define three different categories of services –

(1) Bearer Services, (2) Tele Services, (3) Supplementary Services

(1) Bearer Services – Bearer services permit transparent and nontransparent, synchronous or aynchronous data transmission. Transparent bearer services only use the functions of the physical layer to transmit data. No transmission error occur of data transmission has a constant delay and throughput. Transparent bearer services do not try to recover last data. Non-transparent bearer services use protocols of layers to implement error correction, and flow control. GSM also specifies several bearer services for interworking with PSTN, ISDN and packets switched public data networks. Data transmission can be full-duplex, synchronous with data rates of 1.2, 7.4, 4.8 and 9.6 kbit/sec. asynchronous from 300 to 9600 bit/sec.

(ii) Tele Services – GSM Mainly focuses on voice-oriented tele-services - It provides high quality digital voice transmission of 3.1 Khz of analog phone system. Another service offered by GSM is the emergency number. i.e. Same number can be used throughout Europe. This service is provide for all providers, free of cost.

Another useful service of GSM is SMS (Short Message Service). This service offers transmission of message of 160 characters. SMS Messages do not use the standard data channels of GSM but exploit unused capacity in the signalling channel.

(iii) Supplementary Services – Supplementary Services offer many enhancements for the standard telephony service. It may vary from provider to provider. Typically services are user identification, call redirection or forwarding of ongoing call.

13.4.1 Radio Interface

GSM is a cellular network, which means that mobile phones connect to it by searching for cells in the immediate vicinity. GSM networks operate in four different frequency ranges. Most GSM networks operate in the 900 MHz or 1800 MHz bands. Some countries in the Americas (including the United States and Canada) use the 850 MHz and 1900 MHz bands because the 900 and 1800 MHz frequency bands were already allocated.

The rarer 400 and 450 MHz frequency bands are assigned in some countries, notably Scandinavia, where these frequencies were previously used for first-generation systems.

In the 900 MHz band the uplink frequency band is 890-915 MHz, and the downlink frequency band is 935-960 MHz. This 25 MHz bandwidth is subdivided into 124 carrier frequency channels, each spaced 200 kHz apart. Time division multiplexing is used to allow eight full-rate or sixteen half-rate speech channels per radio frequency channel. There are eight radio timeslots (giving eight burst periods) grouped into what is called a TDMA frame. Half rate channels use alternate frames in the same timeslot. The channel data rate is 270.833 kbit/s, and the frame duration is 4.615 ms.

The transmission power in the handset is limited to a maximum of 2 watts in GSM850/900 and 1 watt in GSM1800/1900.

GSM has used a variety of voice codecs to squeeze 3.1 kHz audio into between 6 and 13 kbit/s. Originally, two codecs, named after the types of data channel they were allocated, were used, called "Full Rate" (13 kbit/s) and "Half Rate" (6 kbit/s). These used a system based upon linear predictive coding (LPC). In addition to being efficient with bitrates, these codecs also made it easier to identify more important parts of the audio, allowing the air interface layer to prioritize and better protect these parts of the signal.

GSM was further enhanced in 1997 with the GSM-EFR codec, a 12.2 kbit/s codec that uses a full rate channel. Finally, with the development of UMTS, EFR was refactored into a variable-rate codec called AMR-Narrowband, which is high quality and robust against interference when used on full rate channels, and less robust but still relatively high quality when used in good radio conditions on half-rate channels.

There are four different cell sizes in a GSM network - macro, micro, pico and umbrella cells. The coverage area of each cell varies according to the implementation environment. Macro cells can be regarded as cells where the base station antenna is installed on a mast or a building above average roof top level. Micro cells are cells whose antenna height is under average roof top level; they are typically used in urban areas. Picocells are small cells whose diameter is a few dozen meters; they are mainly used indoors. Umbrella cells are used to cover shadowed regions of smaller cells and fill in gaps in coverage between those cells.

Cell horizontal radius varies depending on antenna height, antenna gain and propagation conditions from a couple of hundred meters to several tens of kilometers. The longest distance the GSM specification supports in practical use is 35 km or 22 miles. There are also several implementations of the concept of an extended cell, where the cell radius could be double or even more, depending on the antenna system, the type of terrain and the timing advance.

Indoor coverage is also supported by GSM and may be achieved by using an indoor picocell base station, or an indoor repeater with distributed indoor antennas fed through power splitters, to deliver the radio signals from an antenna outdoors to the separate indoor distributed antenna system. These are typically deployed when a lot of call capacity is needed indoors, for example in shopping centers or airports. However, this is not a prerequisite, since indoor coverage is also provided by in-building penetration of the radio signals from nearby cells.

The modulation used in GSM is Gaussian minimum shift keying (GMSK), a kind of continuous-phase frequency shift keying. In GMSK, the signal to be modulated onto the carrier is first smoothed with a Gaussian low-pass filter prior to being fed to a frequency modulator, which greatly reduces the interference to neighboring channels (adjacent channel interference).

A nearby GSM handset is usually the source of the "dit dit dit, dit dit dit, dit dit dit" signal that can be heard from time to time on home stereo systems, televisions, computers, and personal music devices. When these audio devices are in the near

NOTES

NOTES

field of the GSM handset, the radio signal is strong enough that the solid state amplifiers in the audio chain function as a detector. The clicking noise itself represents the power bursts that carry the TDMA signal. These signals have been known to interfere with other electronic devices, such as car stereos and portable audio players. This is a form of RFI, and could be mitigated or eliminated by use of additional shielding and/or bypass capacitors in these audio devices, however, the increased cost of doing so is difficult for a designer to justify.

13.4.2 Structure of a GSM network

The network behind the **GSM** system seen by the customer is large and complicated in order to provide all of the services which are required. It is divided into a number of sections :

- The Base Station Subsystem (the base stations and their controllers).
- The Network and Switching Subsystem (the part of the network most similar to a fixed network). This is sometimes also just called the core network.
- The GPRS Core Network (the optional part which allows packet based Internet connections).
- All of the elements in the system combine to produce many GSM services such as voice calls and SMS.

13.4.3 Subscriber Identity Module

One of the key features of GSM is the Subscriber Identity Module (SIM), commonly known as a **SIM card**. The SIM is a detachable smart card containing the user's subscription information and phonebook. This allows the user to retain his or her information after switching handsets. Alternatively, the user can also change operators while retaining the handset simply by changing the SIM. Some operators block this by allowing the phone to use only a single SIM, or only a SIM issued by them; this practice is known as SIM locking, and is illegal in some countries.

In the United States, Canada, Europe and Australia, many operators lock the mobiles they sell. This is done because the price of the mobile phone is typically subsidised with revenue from subscriptions and operators want to try to avoid subsidising competitor's mobiles. A subscriber can usually contact the provider to remove the lock for a fee, utilize private services to remove the lock, or make use of ample software and websites available on the Internet to unlock the handset themselves. While most web sites offer the unlocking for a fee, some do it for free. The locking applies to the handset, identified by its International Mobile Equipment Identity (IMEI) number, not to the account (which is identified by the SIM card). It is always possible to switch to another (non-locked) handset if such other handset is available.

Some providers will unlock the phone for free if the customer has held an account for a certain period. Third party unlocking services exist that are often quicker and lower cost than that of the operator. In most countries removing the lock is legal. Cingular and T-Mobile provide free unlock services to their customers after 3 months of subscription.

In countries like India, Pakistan, Indonesia, Belgium, etc., all phones are sold unlocked. However, in Belgium, it is unlawful for operators there to offer any form of subsidy on the phone's price. This was also the case in Finland until April 1, 2006, when selling subsidized combinations of handsets and accounts became legal though operators have to unlock phone free of charge after a certain period (at most 24 months).

13.4.4 GSM Security

GSM was designed with a moderate level of security. The system was designed to authenticate the subscriber using shared-secret cryptography. Communications between the subscriber and the base station can be encrypted. The development of UMTS introduces an optional USIM, that uses a longer authentication key to give greater security, as well as mutually authenticating the network and the user - whereas GSM only authenticated the user to the network (and not vice versa). The security model therefore offers confidentiality and authentication, but limited authorization capabilities, and no non-repudiation.

GSM uses several cryptographic algorithms for security. The A5/1 and A5/2 stream ciphers are used for ensuring over-the-air voice privacy. A5/1 was developed first and is a stronger algorithm used within Europe and the United States; A5/2 is weaker and used in other countries. A large security advantage of GSM over earlier systems is that the Key, the crypto variable stored on the SIM card that is the key to any GSM ciphering algorithm, is never sent over the air interface. Serious weaknesses have been found in both algorithms, and it is possible to break A5/2 in real-time in a ciphertext-only attack. The system supports multiple algorithms so operators may replace that cipher with a stronger one.

13.4.5 Features of GSM

From the user's point of view, one of the most remarkable features of GSM is the subcriber Identity Module (SIM), which is a memory device that stores information such as the subcribers identification number, the networks and countries where the subcriber is entitled to service, privacy key etc.

A subcriber uses the SIM with a four digit personal ID number to activate service from any GSM Phone. All GSM mobiles are identical and non-operational without SIM. It is the SIM that gives GSM subcriber units their identity.

A second remarkable feature of GSM is the Air privacy which is provided by the system. Unlike analog FM cellular phone systems, it continously monitor the mobile node.

13.5 Voice Over IP

Voice over Internet Protocol, also called **VoIP**, **IP Telephony**, **Internet telephony**, **Broadband telephony**, **Broadband Phone** *and* **Voice over Broadband** is the routing of voice conversations over the Internet or through any other IP-based network.

Companies providing VoIP service are commonly referred to as providers, and protocols which are used to carry voice signals over the IP network are commonly referred to as **Voice over IP** or **VoIP** protocols. They may be viewed as commercial realizations of the experimental Network Voice Protocol (1973) invented for the ARPANET providers. Some cost savings are due to utilizing a single network .VoIP to VoIP phone calls are sometimes free, while VoIP to public switched telephone networks, PSTN, may have a cost that's borne by the VoIP user.

NOTES

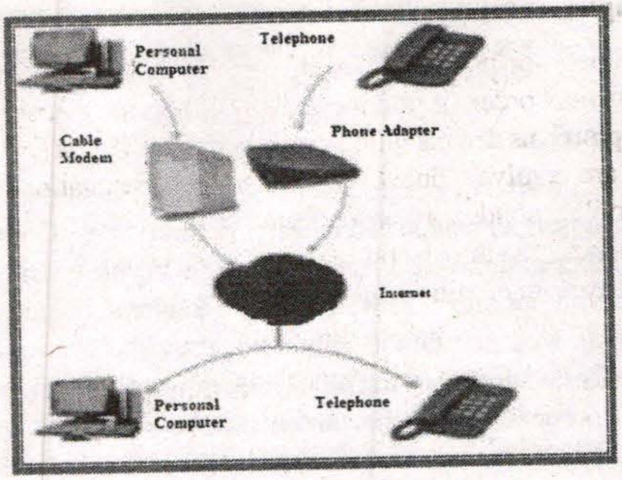


Figure 13.2 : VoIP Framework

There are two types of PSTN to VoIP services: DID (Direct Inward Dialing) and **access numbers**. DID will connect the caller directly to the VoIP user while access numbers require the caller to input the extension number of the VoIP user. Access numbers are usually charged as a local call to the caller and free to the VoIP user while DID usually has a monthly fee. There are also DIDs that are free to the VoIP user but chargeable to the caller.

13.5.1 VoIP Functionality

VoIP can facilitate tasks that may be more difficult to achieve using traditional networks:

- Ability to transmit more than one telephone call down the same broadband-connected telephone line. This can make VoIP a simple way to add an extra telephone line to a home or office.
- Incoming phone calls can be automatically routed to your VoIP phone, regardless of where you are connected to the network. Take your VoIP phone with you on a trip, and wherever you connect to the Internet, you can receive incoming calls.
- Free phone numbers for use with VoIP are available in the USA, UK and other countries from organizations such as VoIP User.
- Call center agents using VoIP phones can work from anywhere with a sufficiently fast and stable Internet connection.
- Many VoIP packages include PSTN features that most telcos (telecommunication companies) normally charge extra for, or may be unavailable from your local telco, such as 3-way calling, call forwarding, automatic redial, and caller ID.
- VoIP is location independent, only an internet connection is needed to get a connection to a VoIP provider.
- VoIP phones can integrate with other services available over the Internet, including video conversation, message or data file exchange in parallel with the conversation, audio conferencing, managing address books and passing information about whether others (e.g. friends or colleagues) are available online to interested parties.

13.5.2 VoIP Implementation

Because UDP does not provide a mechanism to ensure that data packets are delivered in sequential order, or provide Quality of Service guarantees, VoIP implementations face problems dealing with latency and jitter. This is especially true when satellite circuits are involved, due to long round trip propagation delay (400 milliseconds to 600 milliseconds for geostationary satellite). The receiving node must restructure IP packets that may be out of order, delayed or missing, while ensuring that the audio stream maintains a proper time consistency. This functionality is usually accomplished by means of a jitter buffer.

Another challenge is routing VoIP traffic through firewalls and address translators. Private Session Border Controllers are used along with firewalls to enable VoIP calls to and from a protected enterprise network. Skype uses a proprietary protocol to route calls through other Skype peers on the network, allowing it to traverse symmetric NATs and firewalls. Other methods to traverse firewalls involve using protocols such as STUN or ICE.

13.5.3 VoIP Challenges

- Availसइंसमइदंकूपकजी
- क्मसंलधछ्मजूवता स्ंजमदबल
- च्ंबामजसवे
- श्रपजजमत
- म्बीव
- ैमबनतपजल
- त्मसपइंपसपजल
- च्नसेमdialing to DTMF translation

Many VOIP providers do not translate pulse dialing from older phones to DTMF. The VoIP user may use a VOIP Pulse to Tone Converter, if needed.

Fixed delays cannot be controlled but some delays can be minimized by marking voice packets as being delay-sensitive .

The principal cause of packet loss is congestion, which can be controlled by congestion management and avoidance. Carrier VoIP networks avoid congestion by means of traffic engineering.

Variation in delay is called jitter. The effects of jitter can be mitigated by storing voice packets in a buffer (called a play-out buffer) upon arrival, before playing them out. This avoids a condition known as buffer underrun, in which the playout process runs out of voice data to play because the next voice packet has not yet arrived, but increases delay by the length of the buffer.

Common causes of echo include impedance mismatches in analog circuitry, and acoustic coupling of the transmit and receive signal at the receiving end.

13.5.4 VoIP Reliability

Conventional phones are connected directly to telephone company phone lines, which in the event of a power failure are kept functioning by back-up generators or batteries located at the telephone exchange. However, household VoIP hardware uses broadband modems and other equipment powered by household electricity,

NOTES

which may be subject to outages dictating the use of an uninterruptible power supply or generator to ensure availability during power outages. Early adopters of VoIP may also be users of other phone equipment, such as PBX and cordless phone bases, that rely on power not provided by the telephone company. Even with local power still available, the broadband carrier itself may experience outages as well. While the PSTN has been matured over decades and is typically extremely reliable, most broadband networks are less than 10 years old, and even the best are still subject to intermittent outages.. Furthermore, consumer network technologies such as cable and DSL often are not subject to the same restoration service levels as the PSTN or business technologies such as T-1 connection.

13.5.5 VoIP Quality of Service

Some broadband connections may have less than desirable quality. Where IP packets are lost or delayed at any point in the network between VoIP users, there will be a momentary drop-out of voice. This is more noticeable in highly congested networks and/or where there is long distances and/or interworking between end points. Technology has improved the reliability and voice quality over time and will continue to improve VoIP performance as time goes on.

It has been suggested to rely on the packetized nature of media in VOIP communications and transmit the stream of packets from the source phone to the destination phone simultaneously across different routes (multi-path routing). In such a way, the temporary failures have less impact on the communication quality. In capillary routing it has been suggested to use at the packet level Fountain codes or particularly raptor codes for transmitting extra redundant packets making the communication more reliable.

A number of protocols have been defined to support the reporting of QoS/QoE for VoIP calls. These include RTCP XR (RFC3611), SIP RTCP Summary Reports, H.460.9 Annex B (for H.323), H.248.30 and MGCP extensions. The RFC3611 VoIP Metrics block is generated by an IP phone or gateway during a live call and contains information on packet loss rate, packet discard rate (due to jitter), packet loss/discard burst metrics (burst length/density, gap length/density), network delay, end system delay, signal / noise / echo level, MOS scores and R factors and configuration information related to the jitter buffer.

RFC3611 VoIP metrics reports are exchanged between IP endpoints on an occasional basis during a call, and an end of call message sent via SIP RTCP Summary Report or one of the other signaling protocol extensions. RFC3611 VoIP metrics reports are intended to support real time feedback related to QoS problems, the exchange of information between the endpoints for improved call quality calculation and a variety of other applications.

13.5.6 VoIP Difficulty with sending faxes

The support of sending faxes over VoIP is still limited. The existing voice codecs are not designed for fax transmission. An effort is underway to remedy this by defining an alternate IP-based solution for delivering Fax-over-IP, namely the T.38 protocol. Another possible solution to overcome the drawback is to treat the fax system as a message switching system which does not need real time data transmission - such as sending a fax as an email attachment or remote printout. The end system can completely buffer the incoming fax data before displaying or printing the fax image.

13.5.7 VoIP Emergency calls

The nature of IP makes it difficult to locate network users geographically. Emergency calls, therefore, cannot easily be routed to a nearby call center, and are impossible on some VoIP systems. Sometimes, VoIP systems may route emergency calls to a non-emergency phone line at the intended department. In the US, at least one major police department has strongly objected to this practice as potentially endangering the public.

Moreover, in the event that the caller is unable to give an address, emergency services may be unable to locate them in any other way. Following the lead of mobile phone operators, several VoIP carriers are already implementing a technical work-around. For instance, one large VoIP carrier requires the registration of the physical address where the VoIP line will be used. When you dial the emergency number for your country, they will route it to the appropriate local system. They also maintain their own emergency call center that will take non-routable emergency calls (made, for example, from a software based service that is not tied to any particular physical location) and then will manually route your call once learning your physical location.

The United States government had set a deadline, requiring VoIP carriers to implement E911; however, the deadline is being appealed by several of the leading VoIP companies.

This is a different situation with IPBX systems, where these corporate systems often have full E911 capabilities built into the system.

13.5.8 VoIP Integration into global telephone number system

While the traditional Plain Old Telephone System (POTS) and mobile phone networks share a common global standard (E.164) which allocates and identifies any specific telephone line, there is no widely adopted similar standard for VoIP networks. Some allocate an E.164 number which can be used for VoIP as well as incoming/external calls. However, there are often different, incompatible schemes when calling between VoIP providers which use provider specific short codes.

13.5.9 VoIP Single point of calling

With hardware VoIP solutions it is possible to connect the VoIP router into the existing central phone box in the house and have VoIP at every phone already connected. Software based VoIP services require the use of a computer, so they are limited to single point of calling, though handsets are now available, allowing them to be used without a PC. Some services provide the ability to connect WiFi SIP phones so that service can be extended throughout the premises, and off-site to any location with an open hotspot. However, note that many hotspots require browser-based authentication, which most SIP phones do not support.

13.5.10 VoIP Mobile Phones & Handheld Devices

Telcos and consumers have invested billions of dollars in mobile phone equipment. In developed countries, mobile phones have achieved nearly complete market penetration, and many people are giving up landlines and using mobiles exclusively. Given this situation, it is not entirely clear whether there would be a significant higher demand for VoIP among consumers until either public or community wireless networks have similar geographical coverage to cellular networks (thereby enabling mobile VoIP phones, so called WiFi phones) or VoIP is implemented over legacy

NOTES

3G networks. However, "dual mode" handsets, which allow for the seamless handover between a cellular network and a WiFi network, are expected to help VoIP become more popular.

Phones like the Nokia E60, E61 have been the first "dual mode" handsets capable of delivering mobile VoIP. With more and more mobile phones and handheld devices using VOIP, the nicknames of "MoIP" and MVoip (Mobile VoIP)have been attributed to these mobile applications.

Handheld Devices are another type of medium whereby you can use VoIP services. Since most of these devices are limited to using GSM/GPRS type of communication mediums, almost all of the handheld devices use WiFi of some sort.

Another addition to handheld devices are ruggedized barcode type devices that are used in warehouses and retail environments. These types of devices rely on "inside the 4 walls" type of VoIP services that do not connect to the outside world and are solely to be used from employee to employee communications.

VoIP Pre-Paid Phone Cards

VoIP has become an important technology for phone services to travelers, migrant workers and ex-pats, who either, due to not having a fixed or mobile phone or high overseas roaming charges, choose instead to use VoIP services to make their phone calls. Pre-paid phone cards can be used either from a normal phone or from Internet cafes that have phone services. Developing countries and areas with high tourist or immigrant communities generally have a higher uptake.

Caller ID

Caller ID support among VoIP providers varies, although the majority of VoIP providers now offer full Caller ID with name on outgoing calls. When calling a traditional PSTN number from some VoIP providers, Caller ID is not supported.

In a few cases, VoIP providers may allow a caller to spoof the Caller ID information, making it appear as though they are calling from a different number. Business grade VoIP equipment and software often makes it easy to modify caller ID information. Although this can provide many businesses great flexibility, it is also open to abuse.

Mass-market telephony

A major development starting in 2004 has been the introduction of mass-market VoIP services over broadband Internet access services, in which subscribers make and receive calls as they would over the PSTN. Full phone service VoIP phone companies provide inbound and outbound calling with Direct Inbound Dialing. Many offer unlimited calling to the U.S., and some to Canada or selected countries in Europe or Asia as well, for a flat monthly fee.

These services take a wide variety of forms which can be more or less similar to traditional POTS. At one extreme, an analog telephone adapter (ATA) may be connected to the broadband Internet connection and an existing telephone jack in order to provide service nearly indistinguishable from POTS on all the other jacks in the residence. This type of service, which is fixed to one location, is generally offered by broadband Internet providers such as cable companies and telephone companies as a cheaper flat-rate traditional phone service. Often the phrase "VoIP" is not used in selling these services, but instead the industry has marketed the phrases "Internet Phone", "Digital Phone" or "Softphone" which is aimed at typical phone users who are not necessarily tech-savvy. Typically, the provider touts the advantage of being able to keep one's existing phone number.

At the other extreme are services like Gizmo Project and Skype which rely on a software client on the computer in order to place a call over the network, where one user ID can be used on many different computers or in different locations on a laptop. In the middle lie services which also provide a telephone adapter for connecting to the broadband connection similar to the services offered by broadband providers (and in some cases also allow direct connections of SIP phones) but which are aimed at a more tech-savvy user and allow portability from location to location. One advantage of these two types of services is the ability to make and receive calls as one would at home, anywhere in the world, at no extra cost. No additional charges are incurred, as call diversion via the PSTN would, and the called party does not have to pay for the call. For example, if a subscriber with a home phone number in the U.S. or Canada calls someone else within his local calling area, it will be treated as a local call regardless of where that person is in the world. Often the user may elect to use someone else's area code as his own to minimize phone costs to a frequently called long-distance number.

For some users, the broadband phone complements, rather than replaces, a PSTN line, due to a number of inconveniences compared to traditional services. VoIP requires a broadband Internet connection and, if a telephone adapter is used, a power adapter is usually needed. In the case of a power failure, VoIP services will generally not function. Additionally, a call to the U.S. emergency services number 9-1-1 may not automatically be routed to the nearest local emergency dispatch center, and would be of no use for subscribers outside the U.S. This is potentially true for users who select a number with an area code outside their area. Some VoIP providers offer users the ability to register their address so that 9-1-1 services work as expected.

Another challenge for these services is the proper handling of outgoing calls from fax machines, TiVo/ReplayTV boxes, satellite television receivers, alarm systems, conventional modems or FAXmodems, and other similar devices that depend on access to a voice-grade telephone line for some or all of their functionality. At present, these types of calls sometimes go through without any problems, but in other cases they will not go through at all. And in some cases, this equipment can be made to work over a VoIP connection if the sending speed can be changed to a lower bits per second rate. If VoIP and cellular substitution becomes very popular, some ancillary equipment makers may be forced to redesign equipment, because it would no longer be possible to assume a conventional voice-grade telephone line would be available in almost all homes in North America and Western-Europe. The TestYourVoIP website offers a free service to test the quality of or diagnose an Internet connection by placing simulated VoIP calls from any Java-enabled Web browser, or from any phone or VoIP device capable of calling the PSTN network.

In India, it is legal to use VoIP, but it is illegal to have VoIP gateways inside India. This effectively means that people who have PCs can use them to make a VoIP call to any number, but if the remote side is a normal phone, the gateway that converts the VoIP call to a POTS call should not be inside India.

Interconnection is sometimes charged. (Sometimes, it's free of charge.) In case of free of charge, mostly, the traffics are exchanged via P2P connection with the same VoIP standard. Otherwise, certain conversion is needed at the point of VoIP gateway, which needs running costs.

NOTES

13.5.11 VoIP Security

The majority of consumer VoIP solutions do not support encryption yet. As a result, it is relatively easy to eavesdrop on VoIP calls and even change their content. There are several open source solutions that facilitate sniffing of VoIP conversations. A modicum of security is afforded due to patented audio codecs that are not easily available for open source applications, however such security through obscurity has not proven effective in the long run in other fields. Some vendors also use compression to make eavesdropping more difficult. However, real security requires encryption and cryptographic authentication which are not widely available at a consumer level. The existing secure standard SRTP and the new ZRTP protocol is available on Analog Telephone Adapters(ATAs) as well as various softphones. It is possible to use IPsec to secure P2P VoIP by using opportunistic encryption. Skype does not use SRTP, but uses encryption which is transparent to the Skype provider.

The Voice VPN solution provides secure voice for enterprise VoIP networks by applying IPSec encryption to the digitized voice stream.

13.6 Bluetooth

Blue Tooth technology, also called ad-hoc piconets. In 1998 five companies (Ericsion, Intel, IBM, Nokia, Toshiba) founded the blue-tooth with the goal of developing a single chip, low cost, radio based wireless network technology. Many other companies and research Institutions joined the special Internet group around Bluetooth (2002), whose goal was the development of mobile phones, Laptops, notebooks etc.

Bluetooth is an industrial specification for wireless personal area networks (PANs). Bluetooth provides a way to connect and exchange information between devices such as mobile phones, laptops, PCs, printers, digital cameras, and video game consoles over a secure, globally unlicensed short-range radio frequency. The Bluetooth specifications are developed and licensed by the Bluetooth Special Interest Group.

Bluetooth is a radio standard and communications protocol primarily designed for low power consumption, with a short range (power-class-dependent: 1 meter, 10 meters, 100 m) based on low-cost transceiver microchips in each device.

Bluetooth lets these devices communicate with each other when they are in range. The devices use a radio communications system, so they do not have to be in line of sight of each other, and can even be in other rooms, as long as the received transmission is powerful enough.

Class	Maximum Permitted (Power (mW/dBm)	Range (approximate)
Class 1	100 mW (20 dBm)	~100 meters
Class 2	2.5 mW (4 dBm)	~10 meters
Class 3	1 mW (0 dBm)	~1 meter

13.6.1 Bluetooth Applications

More prevalent applications of Bluetooth include:

- Wireless control of and communication between a cell phone and a hands-free headset or car kit. This was one of the earliest applications to become popular.

- Wireless networking between PCs in a confined space and where little bandwidth is required.
- Wireless communications with PC input and output devices, the most common being the mouse, keyboard and printer.
- Transfer of files between devices with OBEX.
- Transfer of contact details, calendar appointments, and reminders between devices with OBEX.
- Replacement of traditional wired serial communications in test equipment, GPS receivers, medical equipment and traffic control devices.
- For controls where infrared was traditionally used.
- Sending small advertisements from Bluetooth enabled advertising hoardings to other, discoverable, Bluetooth devices.
- Seventh-generation game consoles—Nintendo Wii, Sony PlayStation 3—use Bluetooth for their respective wireless controllers.

Bluetooth is implemented in a variety of new products such as phones, printers, modems, and headsets. Bluetooth is acceptable for situations when two or more devices are in proximity to each other and don't require high bandwidth. Bluetooth is most commonly used with phones and hand-held computing devices, either using a Bluetooth headset or transferring files from phones/PDAs to computers.

Bluetooth also simplifies the discovery and setup of services. Bluetooth devices advertise all services they provide. This makes the utility of the service that much more accessible, without the need to worry about network addresses, permissions and all the other considerations that go with typical networks.

Wi-Fi is more analogous to the traditional Ethernet network and requires configuration to set up shared resources, transmit files, set up audio links (for example, headsets and hands-free devices). It uses the same radio frequencies as Bluetooth, but with higher power output resulting in a stronger connection. Wi-Fi is sometimes called "wireless Ethernet." Although this description is inaccurate, it provides an indication of its relative strengths and weaknesses. Wi-Fi requires more setup, but is better suited for operating full-scale networks because it enables a faster connection, better range from the base station, and better security than Bluetooth.

One method for comparing the efficiency of wireless transmission protocols such as Bluetooth and Wi-Fi is spatial capacity, or bits per second per square meter.

A personal computer must have a Bluetooth adapter in order to be able to communicate with other Bluetooth devices (such as mobile phones, mice and keyboards).While some portable computers and fewer desktop computers already contain an internal Bluetooth dongle, most PCs require an external USB Bluetooth dongle. Most Macs come with built-in Bluetooth adapters.

Unlike its predecessor, IrDA, in which each device requires a separate dongle, multiple Bluetooth devices can communicate with a computer over a single dongle.

13.6.2 Operating system support

Linux provides two Bluetooth stacks, with the BlueZ stack included with most Linux kernels. It was originally developed by Qualcomm and Affix. BlueZ supports all core Bluetooth protocols and layers.

NOTES

Only Microsoft Windows XP Service Pack 2 and later versions of Windows have native support for Bluetooth. Previous versions required the users to install their Bluetooth adapter's own drivers, which was not directly supported by Microsoft. Microsoft's own Bluetooth dongles (that are packaged with their Bluetooth computer devices) have no external drivers and thus require at least Windows XP Service Pack 2.

Mac OS X has supported Bluetooth since version 10.2 released in 2002.

13.6.3 Specifications and Features

The Bluetooth specification was developed in 1994 by Sven Mattisson and Jaap Haartsen, who were working for Ericsson Mobile Platforms in Lund, Sweden. The specifications were formalized by the Bluetooth Special Interest Group (SIG). The SIG was formally announced on May 20, 1998. Today it has over 7000 companies worldwide. It was established by Ericsson, Sony Ericsson, IBM, Intel, Toshiba, and Nokia, and later joined by many other companies. Bluetooth is also known as IEEE 802.15.1.

Bluetooth 1.0 and 1.0B

Versions 1.0 and 1.0B had many problems, and manufacturers had difficulties making their products interoperable. Versions 1.0 and 1.0B also had mandatory Bluetooth hardware device address (BD_ADDR) transmission in the Connecting process, rendering anonymity impossible at a protocol level, which was a major setback for services planned to be used in Bluetooth environments, such as Consumerium.

Bluetooth 1.1

- Many errors found in the 1.0B specifications were fixed.
- Added support for non-encrypted channels.
- Received Signal Strength Indicator (RSSI).

Bluetooth 1.2

This version is backward-compatible with 1.1 and the major enhancements include the following:

- Faster Connection and Discovery
- *Adaptive frequency-hopping spread spectrum (AFH)*, which improves resistance to radio frequency interference by avoiding the use of crowded frequencies in the hopping sequence.
- Higher transmission speeds in practice, up to 721 kbit/s, as in 1.1.
- Extended Synchronous Connections (eSCO), which improve voice quality of audio links by allowing retransmissions of corrupted packets.
- Host Controller Interface (HCI) support for three-wire UART.

Bluetooth 2.0

This version, specified in November 2004, is backward-compatible with 1.1. The main enhancement is the introduction of an enhanced data rate (EDR) of 3.0 Mbit/s. This has the following effects:

- Three times faster transmission speed—up to 10 times in certain cases (up to 2.1 Mbit/s).

- Lower power consumption through a reduced duty cycle.
- Simplification of multi-link scenarios due to more available bandwidth.
- Further improved (bit error rate) performance.

Bluetooth 2.1

Bluetooth Core Specification Version 2.1 + EDR, is fully backward-compatible with 1.1, and will be adopted by the Bluetooth SIG once interoperability testing has completed. This specification includes the following features:

- Extended inquiry response: provides more information during the inquiry procedure to allow better filtering of devices before connection. This information includes the name of the device, a list of services the device supports, as well as other information like the time of day, and pairing information.
- Sniff subrating: reduces the power consumption when devices are in the sniff low-power mode, especially on links with asymmetric data flows. Human interface devices (HID) are expected to benefit the most, with mouse and keyboard devices increasing the battery life from 3 to 10 times those currently used.
- Encryption Pause Resume: enables an encryption key to be refreshed, enabling much stronger encryption for connections that stay up for longer than 24 hours.
- Secure Simple Pairing: radically improves the pairing experience for Bluetooth devices, while increasing the use and strength of security. It is expected that this feature will significantly increase the use of Bluetooth.
- NFC cooperation: automatic creation of secure Bluetooth connections when NFC radio interface is also available. For example, a headset should be paired with a Bluetooth 2.1 phone including NFC just by bringing the two devices close to each other (a few centimeters). Another example is automatic uploading of photos from a mobile phone or camera to a digital picture frame just by bringing the phone or camera close to the frame .

13.6.4 Future of Bluetooth

- **Broadcast Channel:** enables Bluetooth information points. This will drive the adoption of Bluetooth into cell phones, and enable advertising models based around users pulling information from the information points, and not based around the object push model that is used in a limited way today.
- **Topology Management:** enables the automatic configuration of the piconet topologies especially in scatternet situations that are becoming more common today. This should all be invisible to the users of the technology, while also making the technology just work.
- **Alternate MAC PHY:** enables the use of alternative MAC and PHY's for transporting Bluetooth profile data. The Bluetooth Radio will still be used for device discovery, initial connection and profile configuration, however when lots of data needs to be sent, the high speed alternate MAC PHY's will be used to transport the data. This means that the proven low power connection models of Bluetooth are used when the system is idle, and the low power per bit radios are used when lots of data needs to be sent.
- QoS improvements: enable audio and video data to be transmitted at a higher quality, especially when best effort traffic is being transmitted in the same piconet.

NOTES

Bluetooth technology already plays a part in the rising Voice over IP (VOIP) scene, with Bluetooth headsets being used as wireless extensions to the PC audio system. As VOIP becomes more popular, and more suitable for general home or office users than wired phone lines, Bluetooth may be used in cordless handsets, with a base station connected to the Internet link.

The next version of Bluetooth after v2.1, code-named Seattle, that will be called **Bluetooth 3.0**, has many of the same features, but is most notable for plans to adopt ultra-wideband (UWB) radio technology. This will allow Bluetooth use over UWB radio, enabling very fast data transfers of up to 480 Mbit/s, while building on the very low-power idle modes of Bluetooth.

13.6.5 High speed Bluetooth

On 28 March 2006, the Bluetooth Special Interest Group announced its selection of the WiMedia Alliance Multi-Band Orthogonal Frequency Division Multiplexing (MB-OFDM) version of UWB for integration with current Bluetooth wireless 4technology.

UWB integration will create a version of Bluetooth wireless technology with a high-speed/high-data-rate option. This new version of Bluetooth technology will meet the high-speed demands of synchronizing and transferring large amounts of data, as well as enabling high-quality video and audio applications for portable devices, multi-media projectors and television sets, and wireless VOIP.

At the same time, Bluetooth technology will continue catering to the needs of very low power applications such as mice, keyboards, and mono headsets, enabling devices to select the most appropriate physical radio for the application requirements, thereby offering the best of both worlds.

13.6.6 Ultra Low Power Bluetooth

On 12 June 2007, Nokia and Bluetooth SIG announced that Wibree will be a part of the Bluetooth specification as an ultra low power Bluetooth technology. Expected user cases includes watches displaying Caller ID information, sports sensors monitoring your heart rate during exercise, as well as medical devices. The Medical Devices Working Group is also creating a medical devices profile and associated protocols to enable this market.

13.6.7 Bluetooth Communication

A master Bluetooth device can communicate with up to seven devices. This network group of up to eight devices is called a piconet.

A piconet is an ad-hoc computer network, using Bluetooth technology protocols to allow one master device to interconnect with up to seven active devices. Up to 255 further devices can be inactive, or parked, which the master device can bring into active status at any time.

At any given time, data can be transferred between the master and one other device. However, the master switches rapidly from device to another in a round-robin fashion. (Simultaneous transmission from the master to multiple other devices is possible, but not used much.) Either device can switch roles and become the master at any time.

Bluetooth specification allows connecting two or more piconets together to form a scatternet, with some devices acting as a bridge by simultaneously playing the master role and the slave role in one piconet.

Setting up Connections

Any Bluetooth device will transmit the following sets of information on demand:

- Device name.
- Device class.
- List of services.
- Technical information, for example, device features, manufacturer, Bluetooth specification, clock offset.

Any device may perform an inquiry to find other devices to which to connect, and any device can be configured to respond to such inquiries. However, if the device trying to connect knows the address of the device, it always responds to direct connection requests and transmits the information shown in the list above if requested. Use of device services may require pairing or acceptance by its owner, but the connection itself can be started by any device and held until it goes out of range. Some devices can be connected to only one device at a time, and connecting to them prevents them from connecting to other devices and appearing in inquiries until they disconnect from the other device.

Every device has a unique 48-bit address. However these addresses are generally not shown in inquiries. Instead, friendly Bluetooth names are used, which can be set by the user. This name appears when another user scans for devices and in lists of paired devices.

Most phones have the Bluetooth name set to the manufacturer and model of the phone by default. Most phones and laptops show only the Bluetooth names and special programs that are required to get additional information about remote devices. This can be confusing as, for example, there could be several phones in range named T610.

Pairing

Pairs of devices may establish a trusted relationship by learning (by user input) a shared secret known as a *passkey*. A device that wants to communicate only with a trusted device can cryptographically authenticate the identity of the other device. Trusted devices may also encrypt the data that they exchange over the air so that no one can listen in. The encryption can, however, be turned off, and passkeys are stored on the device file system, not on the Bluetooth chip itself. Since the Bluetooth address is permanent, a pairing is preserved, even if the Bluetooth name is changed. Pairs can be deleted at any time by either device. Devices generally require pairing or prompt the owner before they allow a remote device to use any or most of their services. Some devices, such as Sony Ericsson phones, usually accept OBEX business cards and notes without any pairing or prompts.

Certain printers and access points allow any device to use its services by default, much like unsecured Wi-Fi networks. Pairing algorithms are sometimes manufacturer-specific for transmitters and receivers used in applications such as music and entertainment.

NOTES

Air interface

The protocol operates in the license-free ISM band at 2.45 GHz. To avoid interfering with other protocols that use the 2.45 GHz band, the Bluetooth protocol divides the band into 79 channels (each 1 MHz wide) and changes channels up to 1600 times per second. Implementations with versions 1.1 and 1.2 reach speeds of 723.1 kbit/s. Version 2.0 implementations feature Bluetooth Enhanced Data Rate (EDR) and reach 2.1 Mbit/s. Technically, version 2.0 devices have a higher power consumption, but the three times faster rate reduces the transmission times, effectively reducing power consumption to half that of 1.x devices (assuming equal traffic load).

Bluetooth differs from Wi-Fi in that the latter provides higher throughput and covers greater distances, but requires more expensive hardware and higher power consumption. They use the same frequency range, but employ different multiplexing schemes. While Bluetooth is a cable replacement for a variety of applications, Wi-Fi is a cable replacement only for local area network access. Bluetooth is often thought of as wireless USB, whereas Wi-Fi is wireless Ethernet, both operating at much lower bandwidth than the cable systems they are trying to replace. However, this analogy is not entirely accurate since any Bluetooth device can, in theory, host any other Bluetooth device—something that is not universal to USB devices.

Many USB Bluetooth adapters are available, some of which also include an IrDA adapter. Older (pre-2003) Bluetooth adapters, however, have limited services, offering only the Bluetooth Enumerator and a less-powerful Bluetooth Radio incarnation. Such devices can link computers with Bluetooth, but they do not offer much in the way of services that modern adapters do.

13.6.8 Bluetooth Consortium

In 1998, Ericsson, IBM, Intel, and Nokia, formed a consortium and adopted the code name Bluetooth for their proposed open specification. In December 1999, 3Com, Lucent Technologies, Microsoft, and Motorola joined the initial founders as the promoter group. Since that time, Lucent Technologies transferred their membership to their spinoff Agere Systems, and 3Com has left the promoter group.

13.7 Wi-Fi

Wi-Fi, popularly known as an acronym for **wi**reless **fi**delity , but, in actuality is simply a play on the term "Hi-Fi," was originally a brand licensed by the Wi-Fi Alliance to describe the embedded technology of wireless local area networks (WLAN) based on the IEEE 802.11 specifications. Use of the term has now broadened to generically describe the wireless interface of mobile computing devices, such as laptops in LANs. Wi-Fi is now increasingly used for more services, including Internet and VoIP phone access, gaming, and basic connectivity of consumer electronics such as televisions, DVD players, and digital cameras. More standards are in development that will allow Wi-Fi to be used by cars on highways in support of an Intelligent Transportation System to increase safety, gather statistics, and enable mobile commerce. **Wi-Fi** and the **Wi-Fi CERTIFIED** logo are registered trademarks of the **Wi-Fi Alliance** - the trade organization that tests and certifies equipment compliance with the 802.11x standards.

13.7.1 Wi-Fi Uses

A person with a Wi-Fi enabled device such as a PC, cell phone or PDA can connect to the Internet when in proximity of an access point. The region covered by one or several access points is called a hotspot. Hotspots can range from a single room to many square miles of overlapping hotspots. Wi-Fi can also be used to create a mesh network. Both architectures are used in community networks.

Wi-Fi also allows connectivity in peer-to-peer (wireless ad-hoc network) mode, which enables devices to connect directly with each other. This connectivity mode is useful in consumer electronics and gaming applications.

When the technology was first commercialized there were many problems because consumers could not be sure that products from different vendors would work together. The Wi-Fi Alliance began as a community to solve this issue so as to address the needs of the end user and allow the technology to mature. The Alliance created the branding *Wi-Fi CERTIFIED* to show consumers that products are interoperable with other products displaying the same branding.

Major Applications include:

- **Wi-Fi at home**

  Home Wi-Fi clients come in many shapes and sizes, from stationary PCs to digital cameras. The trend today is to incorporate wireless into every electronic where mobility is desired.

  Wi-Fi **devices** in home or consumer-type environments connect in the following ways:

  § Via a broadband Internet connection into a single router which can serve both wired and wireless clients .

  - Ad-hoc mode for client to client connections .
  - Built into non-computer devices to enable wireless connectivity to other devices or the Internet .

- **Wi-Fi in Business**

  Business and industrial Wi-Fi has taken off, with the trends in implementation varying greatly over the years. Current technology trends in the corporate wireless world are:

  - Dramatically increasing the number of Wi-Fi Access Points in an environment, in order to provide redundancy,support fast roaming and increasing overall network capacity by using more channels and/or creating smaller cells
  - Designing for wireless voice applications (VoWLAN or WVOIP)
  - Moving toward 'thin' Access Points, with more of the network intelligence housed in a centralized network appliance; relegating individual Access Points to be simply 'dumb' radios
  - Outdoor applications utilizing true mesh topologies
  - A proactive, self-managed network that functions as a security gateway, firewall, DHCP server, intrusion detection system, and a myriad of other features not previously considered relevant to a wireless network.

- **Wi-Fi at Hotspots**

  The most publically visible use of Wi-Fi is at hotspots. These trends include:

NOTES

- Free Wi-Fi at venues like Panera Bread, It's a Grind Coffee House, and over 100,000 locations in the USA has been growing in popularity. According to a door-to-door survey in San Jose, CA, the number of venues and users is growing fast.
- Paid Wi-Fi at venues like Starbucks, McDonalds, and at hotels. This trend is growing rapidly at venues that require a higher rate of customer churn, such as sit-down restaurants.
- According to Muni Wireless, metropolitan-wide WiFi (Mu-Fi) already has more than 300 projects in process.

13.7.2 Generations of Wi-Fi

Wi-Fi technology has gone through several generations since its inception in 1997.

802.11. The original version of the standard IEEE 802.11 released in 1997 specifies two raw data rates of 1 and 2 megabits per second (Mbps) to be transmitted via infrared (IR) signals or by either frequency hopping or direct-sequence spread spectrum in the Industrial Scientific Medical frequency band at 2.4 GHz. IR remains a part of the standard but has no actual implementations.

802.11a. The 802.11a amendment to the original standard was ratified in 1999. The 802.11a standard uses the same core protocol as the original standard and yields realistic throughput in the mid-20 Mbps. Since the 2.4 GHz band is heavily used, using the 5 GHz band gives 802.11a the advantage of less interference. However, this high carrier frequency also brings disadvantages. It restricts the use of 802.11a to almost line of sight, necessitating the use of more access points.

802.11b. The 802.11b amendment to the original standard was ratified in 1999. 802.11b has a maximum raw data rate of 11 Mbps and uses the same CSMA/CA media access method defined in the original standard. The dramatic increase in throughput of 802.11b (compared to the original standard) along with substantial price reductions led to the rapid acceptance of 802.11b as the definitive wireless LAN technology.

802.11g. In June 2003, a third standard was ratified: 802.11g. This works in the 2.4 GHz band (like 802.11b) but operates at a maximum raw data rate of 54 Mbps, or about 24.7 Mbps net throughputs (like 802.11a). Despite its major acceptancem, 802.11g suffers from the same interference as 802.11b in the already crowded 2.4 GHz range. Devices operating in this range include microwave ovens, Bluetooth devices, and cordless telephones.

802.11n. 802.11n builds upon previous standards by adding MIMO (multiple-input multiple-output). MIMO uses multiple transmitter and receiver antennas to allow for increased data throughput through spatial multiplexing and increased range by exploiting the spatial diversity, through coding. On January 19, 2007, the IEEE 802.11 Working Group unanimously approved 802.11n to issue a new Draft 2.0 of the proposed standard.

13.7.3 Wi-Fi Manufacturers

There are many vendors now manufacturing and selling 802.11n products. They include:

- Acer

- Airgo Networks
- Apple
- Asus
- Atheros
- Broadcom
- Buffalo Technology
- Dell
- Intel
- Linksys
- Netgear
- U.S. Robotics

13.7.4 Wi-Fi: How it Works

Wi-Fi networks use radio technologies called IEEE 802.11 to provide secure, reliable, fast wireless connectivity. A typical Wi-Fi setup contains one or more Access Points (APs) and one or more clients. An AP **broadcasts** its SSID (Service Set Identifier, "Network name") via packets that are called beacons, which are usually broadcast every 100 ms. The beacons are transmitted at 1 Mbit/s, and are of relatively short duration and therefore do not have a significant effect on performance. Since 1 Mbit/s is the lowest rate of Wi-Fi it assures that the client that receives the beacon can communicate at at least 1 Mbit/s. Based on the settings (e.g. the SSID), the client may decide whether to connect to an AP. If two APs of the same SSID are in range of the client, the client firmware might use signal strength to decide with which of the two APs to make a connection.

The Wi-Fi standard leaves connection criteria and roaming totally open to the client. This is a strength of Wi-Fi, but also means that one wireless adapter may perform substantially better than another. Since Wi-Fi transmits in the air, it has the same properties as a non-switched wired Ethernet network, and therefore collisions can occur. Unlike a wired Ethernet, and like most packet radios, Wi-Fi cannot do collision detection, and instead uses an acknowledgment packet for every data packet sent. If no acknowledgement is received within a certain time a retransmission occurs. Also, a medium reservation protocol can be used when excessive collisions are experienced or expected (Request ToSend/ClearToSend used for Collision Avoidance or CA) in an attempt to try to avoid collisions.

A Wi-Fi network can be used to connect computers to each other to the internet and to wired networks (which use IEEE 802.3 or Ethernet). Wi-Fi networks operate in the unlicensed 2.4 (802.11b/g) and 5 GHz (802.11a/h) radio bands, with an 11 Mbit/s (802.11b) or 54 Mbit/s (802.11a or g) data rate or with products that contain both bands (dual band). They can provide real world performance similar to the basic 10BaseT wired Ethernet networks.

13.7.5 Wi-Fi Channels

Except for 802.11a/h, which operates at 5 GHz, Wi-Fi devices historically primarily use the spectrum in 2.4 GHz, which is standardized and *unlicensed* by international agreement, although the exact frequency allocations and maximum permitted power

NOTES

vary slightly in different parts of the world. Channel numbers, however, are standardized by frequency throughout the world, so authorized frequencies can be identified by channel numbers. The 2.4 GHz band is also used by microwave ovens, cordless phones, baby monitors and Bluetooth devices.

The maximum number of available channels for Wi-Fi enabled devices are:

- 13 for Europe
- 11 for North America. Only channels 1, 6, and 11 are recommended for 802.11b/g to minimize interference from adjacent channels.
- 14 for Japan

13.7.6 Advantages of Wi-Fi

- Allows LANs to be deployed without cabling for client devices, typically reducing the costs of network deployment and expansion. Spaces where cables cannot be run, such as outdoor areas and historical buildings, can host wireless LANs.
- Built into most modern laptops, getting a laptop without a built in WiFi has become an exception.
- Wi-Fi chipset pricing continues to come down, making Wi-Fi a very economical networking option and driving inclusion of Wi-Fi in an ever-widening array of devices.
- Wi-Fi products are widely available in the market. Different competitive brands of access points and client network interfaces are inter-operable at a basic level of service. Products designated as Wi-Fi CERTIFIED by the Wi-Fi Alliance are backwards inter-operable.
- Wi-Fi is a global set of standards. Unlike cellular carriers, the same Wi-Fi client works in different countries around the world.
- Widely available in more than 250,000 public hot spots and tens of millions of homes and corporate and university campuses worldwide.
- As of 2007, WPA is not easily cracked if strong passwords are used and WPA2 encryption has no known weaknesses.
- New protocols for Quality of Service (WMM) and power saving mechanisms (WMM Power Save) make Wi-Fi even more suitable for latency-sensitive applications (such as voice and video) and small Form-Factor

13.7.7 Disadvantages of Wi-Fi

- Spectrum assignments and operational limitations are not consistent worldwide; most of Europe allows for an additional 2 channels beyond those permitted in the US (1-13 vs 1-11); Japan has one more on top of that (1-14) - and some countries, like Spain, prohibit use of the lower-numbered channels. Furthermore some countries, such as Italy, used to require a 'general authorization' for any Wi-Fi used outside an operator's own premises, or require something akin to an operator registration.

- Equivalent isotropically radiated power (EIRP) in the EU is limited to 20 dBm (0.1 W).
- Power consumption is fairly high compared to some other low bandwidth standards (Zigbee and Bluetooth), making battery life a concern.
- The most common wireless encryption standard, Wired Equivalent Privacy or WEP, has been shown to be easily breakable even when correctly configured. Wi-Fi Protected Access (WPA and WPA2) which began shipping in 2003 aims to solve this problem and is now available on most products.
- Wi-Fi Access Points typically default to an open (encryption-free) mode. Novice users benefit from a zero configuration device that works out of the box but without security enabled providing open wireless access to their LAN. To turn security on requires the user to configure the device, usually via a software GUI.
- Many 2.4 GHz 802.11b and 802.11g Access points default to the same channel on initial start up, contributing to congestion on certain channels. To change the channel of operation for an access point requires the user to configure the device.
- Wi-Fi networks have limited range. A typical Wi-Fi home router using 802.11b or 802.11g with a stock antenna might have a range of 45 m (150 ft) indoors and 90 m (300 ft) outdoors. Range also varies with frequency band. Wi-Fi in the 2.4 GHz frequency block has slightly better range than Wi-Fi in the 5 GHz frequency block. Outdoor range with improved antennas can be several kilometres or more with line-of-sight.
- Wi-Fi pollution, of an excessive number of an access point with other access points in the area, especially on the same or neighboring channel, can prevent access and interfere with the use of other access points by others caused by overlapping channels in the 802.11g/b spectrum as well as with decreased signal-to-noise ratio (SNR) between access points. This can be a problem in high-density areas such as large apartment complexes or office buildings with many Wi-Fi access points. Additionally, other devices use the 2.4 GHz band: microwave ovens, cordless phones, baby monitors, security cameras, and Bluetooth devices can cause significant additional interference.
- It is also an issue when municipalities or other large entities such as universities seek to provide large area coverage. Everyone is considered equal for the base standard without 802.11e/WMM when they use the band. This openness is also important to the success and widespread use of 2.4 GHz Wi-Fi, but makes it unsuitable for "must have" public service functions or where reliability is required. Users sometimes suffer network "frustrations" or a total network breakdown if gaming because a neighbour microwaves some pop corn.
- Interoperability issues between brands or proprietary deviations from the standard can disrupt connections or lower throughput speeds on other user's devices that are within range. And, Wi-Fi devices do not presently pick channels to avoid interference.

NOTES

- Wi-Fi networks that are open (unencrypted) can be monitored and used to read and copy data (including personal information) transmitted over the network unless another security method is used to secure the data like a VPN or a secure web page.

13.7.8 Wi-Fi Standard Devices

Wireless Access Point (WAP)

A wireless access point connects a group of wireless devices to an adjacent wired LAN. An access point is similar to an ethernet hub, relaying data between connected wireless devices in addition to a (usually) single connected wired device, most often an ethernet hub or switch, allowing wireless devices to communicate with other wired devices.

Wireless Adapter

A wireless adapter allows a device to connect to a wireless network. These adapters connect to devices using various external or internal interconnects such as PCI, miniPCI, USB, ExpressCard, Cardbus and PC card. Most newer laptop computers are equipped with internal adapters. Internal cards are generally more difficult to install.

Wireless Router

A wireless router integrates a WAP, ethernet switch, and internal Router firmware application that provides IP Routing, NAT, and DNS forwarding through an integrated WAN interface. A wireless router allows wired and wireless ethernet LAN devices to connect to a (usually) single WAN device such as cable modem or DSL modem. A wireless router allows all three devices (mainly the access point and router) to be configured through one central utility. This utility is most usually an integrated web server which serves web pages to wired and wireless LAN clients and often optionally to WAN clients. This utility may also be an application that is run on a desktop computer such as Apple's AirPort.

Wireless Ethernet Bridge

A wireless Ethernet bridge connects a wired network to a wireless network. This is different from an access point in the sense that an access point connects wireless devices to a wired network at the data-link layer. Two wireless bridges may be used to connect two wired networks over a wireless link, useful in situations where a wired connection may be unavailable, such as between two separate homes.

Range Extender

A wireless range extender or wireless repeater can extend the range of an existing wireless network. Range extenders can be strategically placed to elongate a signal area or allow for the signal area to reach around barriers such as those created in L-shaped corridors. Wireless devices connected through repeaters will suffer from an increased latency for each hop. Additionally, a wireless device at the end of chain of wireless repeaters will have a throughput that is limited by the weakest link within the repeater chain.

Antenna connectors

Most commercial devices (routers, access points, bridges, repeaters) designed for home or business environments use either RP-SMA or RP-TNC antenna connectors. PCI wireless adapters also mainly use RP-SMA connectors.

Most PC card and USB wireless only have internal antennas etched on their printed circuit board while some have MMCX connector or MC-Card external connections in addition to an internal antenna. A few USB cards have a RP-SMA connector.

Most Mini PCI wireless cards utilize Hirose U.FL connectors, but cards found in various wireless appliances contain all of the connectors listed.

Many high-gain (and homebuilt antennas) utilize the Type N connector more commonly used by other radio communications methods.

13.7.9 Long Range Wi-Fi

Recently, long range Wi-Fi kits have begun to enter the market. Companies like RadioLabs and BroadbandXpress offer long range, inexpensive kits that can be setup with limited knowledge. These kits utilize specialized antennas which increase the range of Wi-Fi dramatically, in the case of the world record 137.2 miles (220 km). These kits are commonly used to get Broadband internet to a place that cannot access the service itself.

The longest link ever achieved was by the Swedish space agency. They attained 310 km, but used 6 watt amplifiers to reach an overhead stratospheric balloon.

13.7.10 Wi-Fi OS Support

There are two sides to Wi-Fi support under an operating system: driver level support, and configuration and management support.

Driver support is usually provided by the manufacturer of the hardware or, in the case of Unix clones such as Linux and FreeBSD, sometimes through open source projects.

Configuration and management support consists of software to enumerate, join, and check the status of available Wi-Fi networks. This also includes support for various encryption methods. These systems are often provided by the operating system backed by a standard driver model. In most cases, drivers emulate an ethernet device and use the configuration and management utilities built into the operating system. In cases where built in configuration and management support is non-existent or inadequate, hardware manufacturers may include their own software to handle the respective tasks.

Microsoft Windows

Microsoft Windows has comprehensive driver-level support for Wi-Fi, the quality of which depends on the hardware manufacturer. Hardware manufactures almost always ship Windows drivers with their products. Windows ships with very few Wi-Fi drivers and depends on the original equipment manufacturers (OEMs)and device

NOTES

manufacturers to make sure users get drivers. Configuration and management depend on the version of Windows.

- Earlier versions of Windows, such as 98, ME and 2000 do not have built-in configuration and management support and must depend on software provided by the manufacturer
- Microsoft Windows XP has built-in configuration and management support. The original shipping version of Windows XP included rudimentary support which was dramatically improved in Service Pack 2. Support for WPA2 and some other security protocols require updates from Microsoft. There are still problems with XP support of Wi-Fi. (One simple interface problem is that if the user makes a mistake in the (case sensitive) passphrase, XP keeps trying to connect but never tells the user that the passphrase is wrong. A second problem is not allowing the user to see different BSSID's for the same ESSID; that is, it provides no way for the user to differentiate access points with the same name.) To make up for Windows' inconsistent and sometimes inadequate configuration and management support, many hardware manufacturers include their own software and require the user to disable Windows' built-in Wi-Fi support.
- Microsoft Windows Vista has improved Wi-Fi support over Windows XP. The original betas automatically connected to unsecured networks without the user's approval. The release candidate (RC1 or RC2) does not continue to display this behavior, requiring user permissions to connect to an unsecured network, as long as the user account is in the default configuration with regards to User Account Control.

Apple Mac OS

Apple was an early adopter of Wi-Fi, introducing its AirPort product line, based on the 802.11b standard, in July 1999. Apple then introduced AirPort Extreme as an implementation of 802.11g. All Macs starting with the original iBook included AirPort slots for which an AirPort card can be used, connecting to the computer's internal antenna. All Intel-based Macs either come with built-in Airport Extreme or a slot for an AirPort card. In late 2006, Apple began shipping Macs with Broadcom Wi-Fi chips that also supported the Draft 802.11n standard which can be unlocked through buying a $2 driver released by Apple at the January 2007 Macworld Expo. The driver is also included for free with Apple's 802.11n AirPort Extreme.

Apple makes the Mac OS operating system, the computer hardware, the accompanying drivers, AirPort WiFi base stations, and configuration and management software, simplifying Wi-Fi integration. The built-in configuration and management is integrated throughout many of the operating system's applications and utilities. Mac OS X has Wi-Fi support, including WPA2, and ships with drivers for Apple's Broadcom-based AirPort cards. Many third-party manufacturers make compatible hardware along with the appropriate drivers which work with Mac OS X's built-in configuration and management software. Other manufacturers distribute their own software.

Apple's older Mac OS 9 does not have built in support for Wi-Fi configuration and management nor does it ship with Wi-Fi drivers, but Apple provides free drivers and configuration and management software for their AirPort cards for OS 9, as do a few other manufacturers. Versions of Mac OS before OS 9 predate Wi-Fi and do not have any Wi-Fi support, although some third-party hardware manufacturers have made drivers and connection software that allows earlier OSes to use Wi-Fi.

Open source Unix-like systems

Linux, FreeBSD and similar Unix-like clones have much coarser support for Wi-Fi. Due to the open source nature of these operating systems, many different standards have been developed for configuring and managing Wi-Fi devices. The open source nature also fosters open source drivers which have enabled many third party and proprietary devices to work under these operating systems.

- Linux has patchy Wi-Fi support. Native drivers for many Wi-Fi chipsets are available either commercially or at no cost, although some manufacturers don't produce a Linux driver, only a Windows one. Consequently, many popular chipsets either don't have a native Linux driver at all, or only have a half-finished one. For these, the freely available NdisWrapper and its commercial competitor DriverLoader allow Windows x86 and 64 bit variants NDIS drivers to be used on x86-based Linux systems but not on other architectures. As well as the lack of native drivers, some Linux distributions do not offer a convenient user interface and configuring Wi-Fi on them can be a clumsy and complicated operation compared to configuring wired Ethernet drivers[!]. This is changing with Network Manager, a utility that allows users to automatically switch between networks without using the command line.
- FreeBSD has Wi-Fi support similar to Linux. Support under FreeBSD is best in the 6.x versions, which introduced full support for WPA and WPA2, although in some cases this is driver dependent. FreeBSD comes with drivers for many wireless cards and chipsets, including those made by Atheros, Ralink, Cisco, D-link, Netgear, and many Centrino chipsets, and provides support for others through the ports collection. FreeBSD also has "Project Evil", which provides the ability to use Windows x86 NDIS drivers on x86-based FreeBSD systems as NdisWrapper does on Linux, and Windows amd64 NDIS drivers on amd64-based systems.
- NetBSD, OpenBSD, and DragonFly BSD have Wi-Fi support similar to FreeBSD. Code for some of the drivers, as well as the kernel framework to support them, is mostly shared among the 4 BSDs.

13.8 ISDN

Integrated Services Digital Network or **Isolated Subscriber Digital Network (ISDN)**, originally **"Integriertes Sprach- und Datennetz"** (German for "Integrated Speech and Data Net"), is a telephone system network. Prior to the ISDN, the phone system was viewed as a way to transport voice, with some special services available for data. The key feature of the ISDN is that it integrates speech and data on the same lines, adding features that were not available in the classic telephone system. There are several kinds of access interfaces to the ISDN defined: Basic Rate Interface (BRI), Primary Rate Interface (PRI) and Broadband-ISDN (B-ISDN).

NOTES

ISDN is a circuit-switched telephone network system, that also provides access to packet switched networks, designed to allow digital transmission of voice and data over ordinary telephone copper wires, resulting in better voice quality than an analog phone. It offers circuit-switched connections (for either voice or data), and packet-switched connections (for data), in increments of 64 kbit/s. Another major use case is Internet access, where ISDN typically provides a maximum of 128 kbit/s in both upstream and downstream directions (which can be considered to be broadband speed, since it exceeds the narrowband speeds of standard analog 56k telephone lines). ISDN channels may use bonding to achieve a greater data rate, typically 3 or 4 BRIs (6 to 8 64 kbit/s channels) are bonded.

13.8.1 ISDN ELEMENTS

- *Integrated Services* refers to ISDN's ability to deliver at minimum two simultaneous connections, in any combination of data, voice, video, and fax, over a single line. Multiple devices can be attached to the line, and used as needed. That means an ISDN line can take care of most people's complete communications needs at a much higher transmission rate, without forcing the purchase of multiple analog phone lines.
- *Digital* refers to its purely digital transmission, as opposed to the analog transmission of plain old telephone service (POTS). Use of an analog telephone modem for Internet access requires that the Internet service provider's (ISP) modem converts the digital content to analog signals before sending it and the user's modem then converts those signals back to digital when receiving. When connecting with ISDN there is no digital to analog conversion.
- *Network* refers to the fact that ISDN is not simply a point-to-point solution like a leased line. ISDN networks extend from the local telephone exchange to the remote user and includes all of the telecommunications and switching equipment in between.

The purpose of the ISDN is to provide fully integrated digital services to the users. These services fall under three categories: bearer services, supplementary services and teleservices.

13.8.2 ISDN : Consumer and industry perspectives

There are two points of view into the ISDN world. The most common viewpoint is that of the end user, who wants to get a digital connection into the telephone/data network from home, whose performance would be better than an ordinary analog modem connection. The typical end-user's connection to the Internet is related to this point of view, and discussion on the merits of various ISDN modems, carriers' offerings and tarriffing (features, pricing) are from this perspective. Much of the following discussion is from this point of view, but it should be noted that as a data connection service, ISDN has been mostly superseded by DSL.

There is a second viewpoint: that of the telephone industry, where ISDN is a core technology. A telephone network can be thought of as a collection of wires strung between switching systems. The common electrical specification for the signals on these wires is T1 or E1. Between telephone company switches, the signaling is performed via SS7. Normally, a corporate or other PBX is connected via a T1, and the signalling was done with A&B bits to indicate on-hook or off-hook conditions and MF and DTMF tones to encode the destination number. ISDN is much better because messages can be sent much more quickly than by trying to encode numbers as long (100 ms per digit) tone sequences. This translated to much faster call setup times. Also, a greater number of features are available and fraud is reduced.

ISDN is also used as a smart-network technology intended to add new services to the public switched telephone network (PSTN) by giving users direct access to end-to-end circuit-switched digital services and as a backup or failsafe circuit solution for critical use data circuits.

13.8.3 ISDN : From India Perspective

India

In India, ISDN was very popular until the introduction of ADSL. Bharat Sanchar Nigam Limited, the largest communication service provider in India and a state owned company, is offering both ISDN BRI and PRI services across the country over its ISDN network. After the introduction of ADSL broadband technology with static IPs, the data transfer load is taken up by ADSL. But ISDN still plays a very big role as a backup network for point-to-point leased line customers and low cost reliable data network for organisations located all over India, such as banks, E-seva centres, Life Insurance Corporation of India, and so on.

13.8.4 Advantages of ISDN

Speed

The modem was a big breakthrough in computer communications. It allowed computers to communicate by converting their digital information into an analog signal to travel through the public phone network. There is an upper limit to the amount of information that an analog telephone line can hold. Currently, it is about 56 kb/s bidirectionally. Commonly available modems have a maximum speed of 56 kb/s, but are limited by the quality of the analog connection and routinely go about 45-50 kb/s. Some phone lines do not support 56 kb/s connections at all. There were currently 2 competing, incompatible 56 kb/s standards (X2 from U S Robotics (recently bought by 3Com), and K56flex from Rockwell/Lucent). This standards problem was resolved when the ITU released the V.90, and later V.92, standard for 56 kb/s modem communications.

ISDN allows multiple digital channels to be operated simultaneously through the same regular phone wiring used for analog lines. The change comes about when the telephone company's switches can support digital connections. Therefore, the same physical wiring can be used, but a digital signal, instead of an analog signal, is transmitted across the line. This scheme permits a much higher data transfer rate than analog lines. BRI ISDN, using a channel aggregation protocol such as BONDING or Multilink-PPP, supports an uncompressed data transfer speed of 128 kb/s, plus bandwidth for overhead and signaling. In addition, the latency, or the amount of time it takes for a communication to begin, on an ISDN line is typically about half that of an analog line. This improves response for interactive applications, such as games.

NOTES

Multiple Devices

Previously, it was necessary to have a separate phone line for each device you wished to use simultaneously. For example, one line each was required for a telephone, fax, computer, bridge/router, and live video conference system. Transferring a file to someone while talking on the phone or seeing their live picture on a video screen would require several potentially expensive phone lines.

ISDN allows multiple devices to share a single line. It is possible to combine many different digital data sources and have the information routed to the proper destination. Since the line is digital, it is easier to keep the noise and interference out while combining these signals. ISDN technically refers to a specific set of digital services provided through a single, standard interface. Without ISDN, distinct interfaces are required instead.

Signaling

Instead of the phone company sending a ring voltage signal to ring the bell in your phone ("In-Band signal"), it sends a digital packet on a separate channel ("Out-of-Band signal"). The Out-of-Band signal does not disturb established connections, no bandwidth is taken from the data channels, and call setup time is very fast. For example, a V.90 or V.92 modem typically takes 30-60 seconds to establish a connection; an ISDN call setup usually takes less than 2 seconds.

The signaling also indicates who is calling, what type of call it is (data/voice), and what number was dialed. Available ISDN phone equipment is then capable of making intelligent decisions on how to direct the call.

SUMMARY

- Communication allows people to exchange information by one of several methods. There are auditory means, such as speaking or singing, and non-verbal, physical means, such as body language, sign language, paralanguage, touch or eye contact.
- Modern communication media now allow for intense long-distance exchanges between larger numbers of people (many-to-many communication via e-mail, Internet forums). On the other hand, many traditional broadcast media and mass media favor one-to-many communication (television, cinema, radio, newspaper, magazines).
- The *Global System for Mobile communications (GSM*: originally from *Groupe Spécial Mobile*) is the most popular standard for mobile phones in the world. GSM service is used by over 2 billion people across more than 212 countries and territories.
- GSM is a cellular network, which means that mobile phones connect to it by searching for cells in the immediate vicinity. GSM networks operate in four different frequency ranges. Most GSM networks operate in the 900 MHz or 1800 MHz bands.
- The modulation used in GSM is Gaussian minimum shift keying (GMSK), a kind of continuous-phase frequency shift keying.
- One of the key features of GSM is the Subscriber Identity Module (SIM),

commonly known as a SIM card. The SIM is a detachable smart card containing the user's subscription information and phonebook. This allows the user to retain his or her information after switching handsets.

- Wireless local loop (WLL), also called Broadband Wireless Access (BWA) radio in the loop (RITL) or fixed-radio access (FRA) or fixed-wireless access (FWA) or Fixed Wireless Terminal (FWT), is the use of a wireless communications link as the "last mile / first mile" connection for delivering plain old telephone service (POTS) and broadband Internet to telecommunications customers.
- *Code division multiple access (CDMA)* is a form of multiplexing and a method of multiple access that divides up a radio channel not by time (as in time division multiple access), nor by frequency (as in frequency-division multiple access), but instead by using different pseudo-random code sequences for each user. CDMA is a form of "spread-spectrum" signaling, since the modulated coded signal has a much higher bandwidth than the data being communicated.
- *CDMA* also refers to digital cellular telephony systems that make use of this multiple access scheme, such as those pioneered by QUALCOMM, and W-CDMA by the International Telecommunication Union or ITU.
- *Voice over Internet Protocol*, also called *VoIP*, *IP Telephony, Internet telephony, Broadband telephony, Broadband Phone and Voice over Broadband* is the routing of voice conversations over the Internet or through any other IP-based network.
- *Bluetooth* is an industrial specification for wireless personal area networks (PANs). Bluetooth provides a way to connect and exchange information between devices such as mobile phones, laptops, PCs, printers, digital cameras, and video game consoles over a secure, globally unlicensed short-range radio frequency. The Bluetooth specifications are developed and licensed by the Bluetooth Special Interest Group.
- Bluetooth is a radio standard and communications protocol primarily designed for low power consumption, with a short range (power-class-dependent: 1 meter, 10 meters, 100 m) based on low-cost transceiver microchips in each device.
- *Wi-Fi*, popularly known as an acronym for wireless fidelity , but, in actuality is simply a play on the term "Hi-Fi," was originally a brand licensed by the Wi-Fi Alliance to describe the embedded technology of wireless local area networks (WLAN) based on the IEEE 802.11 specifications.
- ISDN is a circuit-switched telephone network system, that also provides access to packet switched networks, designed to allow digital transmission of voice and data over ordinary telephone copper wires, resulting in better voice quality than an analog phone.

EXERCISE

1 What are communications technologies ?

2. What is Global System for Mobile communications ? Explain it radio interface.

NOTES

3. Describe the structure of GSM network .
4. What is SIM ?
5. Explain the security features of GSM .
6. What is WLL ? How it is related to FWT ?
7. What is WiMAX ?
8. Describe different WLL standards .
9. What is CDMA technology ? Why it is popular ?
10. Differentiate between synchronous and asynchronous CDMA .
11. What is VoIP ? Explain its functionality .
12. What is Bluetooth ? Describe its applications .
13. Describe the OS support for Bluetooth .
14. Describe different specifications of Bluetooth .
15. What is high speed bluetooth and Ultra Low Power Bluetooth ?
16. What is Wi-Fi technologies ? Describe its uses .
17. Explain the WiFi communication and connection .
18. Describe different generations of Wi-Fi .
19. Explain the working framework of Wi-Fi .
20. Describe the advantages and disadvantages of Wi-Fi .
21. Explain ISDN. Give its advantages also.

NOTES

14 Electronic Data Interchange

Chapter Includes :

- ✦ INTRODUCTION
- ✦ ELECTRONIC DATA INTERCHANGE
- ✦ THE STRUCTURE OF EDI SYSTEMS
  - ARCHITECTURE FOR EDI
- ✦ EDI STANDARDS
  - EDIFACT
  - COMPONENTS OF EDI STANDARDS
- ✦ FEATURES OF EDI
- ✦ EDI TECHNOLOGY
  - VALUE-ADDED NETWORK
  - EDI SPECIFICATIONS
  - EDI REQUIREMENTS
  - EDI TRANSMISSION
  - INTERPRETING DATA IN EDI
- ✦ ADVANTAGES OF EDI
- ✦ BARRIERS IN ADOPTING EDI
- ✦ DRAWBACKS OF EDI
- ✦ NEW TRENDS IN EDI
  - XML - EDI
  - EDI AND EBXML

14.1 Introduction

Among the most pervasive and impacting changes has been the growth of electronic commerce (EC). Innovations such as electronic data interchange (EDI), on-line banking, and web-based commerce have changed the way business information moves through the supply chain, thus making the need for advanced "EC-friendly" business software a must.

Electronic Data Interchange(EDI) makes it feasible for computer systems that store data in disparate proprietary data formats to effectively communicate with one another in an efficient manner. EDI enables a commonly understood and standardized format of the relevant data to be transmitted from one computer system to another with minimal human intervention. EDI transactions are structured for highly automated processing.

The early applications of what became known as EDI were undertaken in the United States. The idea's origins have an international flavour, however, being trace-

NOTES

able back to the 1948 Berlin Airlift, where the task of co-ordinating air freighted consignments of food and consumables (which arrived with differing manifests, languages and numbers of copies) was addressed by devising a standard manifest. Electronic transmission commenced during the 1960s, initially in the rail and road transport industries. The standardization of documents was a necessary concomitant to that change. In 1968 the United States Transportation Data Coordinating Committee (TDCC) was formed, to coordinate the development of translation rules among four existing sets of industry-specific standards. A further significant move towards standardization came with the X12 standards of the American National Standards Institute (ANSI), which gradually extended and replaced those created by the TDCC.

At about the same time, the U.K. Department of Customs and Excise, with the assistance of SITPRO (the British Simplification of Trade Procedures Board), was developing its own standards for documents used in international trade, called Tradacoms. These were later extended by the United Nations Economic Commission for Europe (UNECE) into what became known as the GTDI (General-purpose Trade Data Interchange standards), and were gradually accepted by some 2,000 British exporting organizations.

Problems created by the trans-Atlantic use of two different (and largely incompatible) sets of standardized documents have been addressed by the formation of a United Nations Joint European and North American working party (UN-JEDI), which began the development of administration, commerce and transport document translation standards & was named as Electronic Data Interchange for Administration, Commerce and Transport (EDIFACT). A full range of business documents is in the process of being developed.

EDI's direct impact is to reduce the amount of data capture and transcription. This generally results in a decreased incidence of errors, less time spent on exception-handling, and fewer data-caused delays in the business process. Benefits can be attained in such areas as inventory management, transport and distribution, administration and cash management.

The earliest implementation of a system compliant with EDI is the British Automated Clearing Service (1970).

- Many organisations that had a significant amount of regular payments used the BACS.
- BACS users would record information they would otherwise have printed as cheques, on to magnetic tape reels and courier them to the BACS center.
- Later an online submission facility was added

Other early examples include LACES (1971-1981), a freight clearing system used at Heathrow and WMO, the World Meteorological Office system used to share weather information.

Preliminary initiatives of trading groups increased and standardization in Europe soon became needed.

- The first effort is TRADACOM, a UK standard developed by ANA in 1982 for general trade.
- Other European countries developed similar standards such as SEDAS in Germany, and GENCOD in France. These standards have since migrated to EDIFACT.

NOTES

North America had similar problems with industry groups making up their own standards rapidly.

- In the beginning of the 80s ANSI took up the task of standardizing the EDI messages, in order to make cross-industry trade possible. The resulting standard is called ANSI X12.

EDI is used in a variety of industries. In fact over 80,000 companies have made the switch to EDI to improve their efficiencies. Many of these companies require all of their partners should use EDI.

14.2 Electronic Data Interchange

Electronic Data Interchange(EDI) is intended to handle all aspects of business transactions such as ordering, acknowledgements, pricing, status, scheduling, shipping, receiving, invoices, payments, and financial reporting. One of the principal aims of EDI has been to develop electronic surrogates for the myriad of paper forms used in commercial transactions such as purchase orders, bills of lading, and invoices.

Basically, the electronic data interchange process is the computer-to-computer exchange of business documents between companies. EDI replaces the faxing and mailing of paper documents. EDI documents use specific computer record formats that are based on widely accepted standards. However, each company may use the flexibility allowed by the standards in a unique way that fits their business needs.

Electronic Data Interchange is a set of standards for structuring information to be electronically exchanged between and within businesses, organizations, government entities and other groups. The standards describe structures that emulate documents, for example purchase orders to automate purchasing. The term EDI is also used to refer to the implementation and operation of systems and processes for creating, transmitting, and receiving EDI documents.

Despite being relatively unheralded, in this era of technologies such as XML services, the Internet and the World Wide Web, EDI is still the data format used by the vast majority of electronic commerce transactions in the world.

Electronic Data Interchange is the electronic transmission of business documents between vendor and customer computer systems. The use of EDI guarantees optimal efficiency and accuracy of the purchasing cycle.

EDI is widely used in industry to transmit traditional "documents," such as invoices or purchase orders, between companies. The standardized transaction set has been refined, expanded and developed so there are now hundreds of different "documents" that can be electronically exchanged between multiple trading partners. The electronic transmission of these transactions is an efficient means of conducting business.

Figure 14.1 : Electronic Data Interchange

NOTES

EDI is the exchange of documents in standardized electronic form, between organizations, in an automated manner, directly from a computer application in one organization to an application in another .The EDI process provides many benefits. Computer-to-computer exchange of information is much less expensive than handling paper documents. Much less labor time is required. Fewer errors occur because computer systems process the documents rather than processing by hand.

EDI transactions between companies flow faster and more reliably than paper documents. Faster transactions support reduction in inventory levels, better use of warehouse space, fewer out-of-stock occurrences and lower freight costs through fewer emergencies expedites.

The Internet has enabled EDI transactions to be transmitted between trading partners in an even more efficient manner. The Internet provides business and government agencies with an environment that is open, fast, cost effective, and widely accepted and used.

The concept of a value-added network (VAN) is a mechanism that facilitates the transfer of electronic data between trading partners. A VAN can be thought of as a post office that allows an entity to send EDI formatted data to one of their trading partners at any time. The VAN will hold the file of transmitted transactions until the trading partner to whom it is addressed retrieves it at a later time.

14.3 The Structure of EDI Systems

The basic EDI structure is given below.

- Each partner has their own internal software systems.
- Special EDI adapters have to be implemented which will be able to interface with the internal system as well as communicate with the value added network.

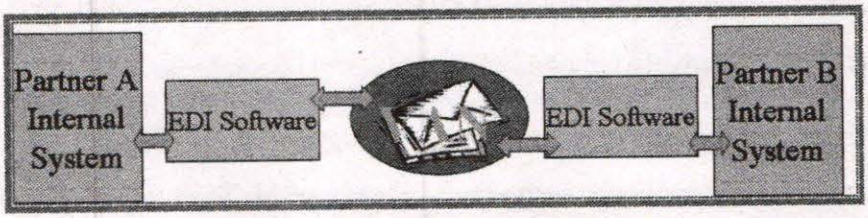
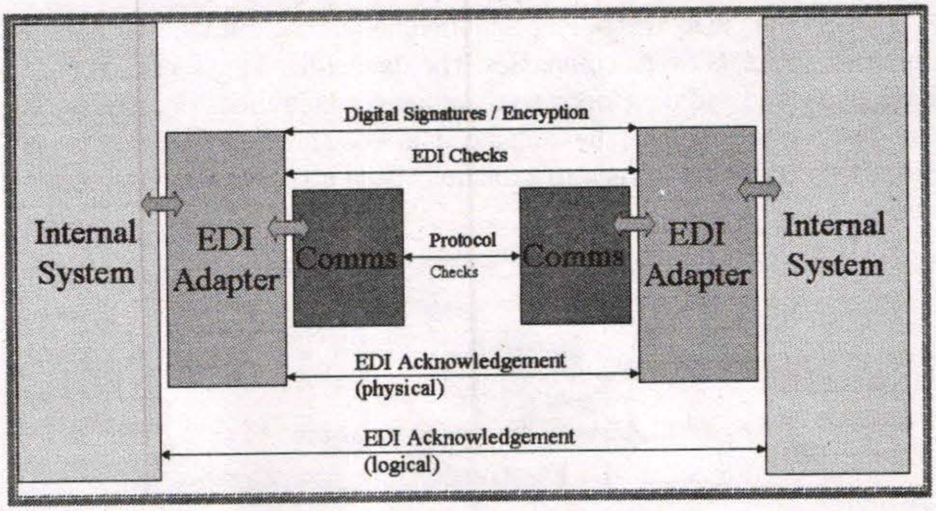


Figure 14.2 : EDI Framework

-



- The particulars of the message syntax and interaction process are negotiated between partners in advance. Sometimes a dominant partner will impose its standards on smaller partners.

14.3.1 Architecture for EDI

EDI can be compared and contrasted with electronic mail (email). Email enables free-format, textual messages to be electronically transmitted from one person to another. EDI, on the other hand, supports structured business messages (those which are expressed in hard-copy, pre-printed forms or business documents), and transmits them electronically between computer applications, rather than between people.

The essential elements of EDI are:

- The use of **an electronic transmission medium** (originally a value-added network, but increasingly the open, public Internet) rather than the dispatch of physical storage media such as magnetic tapes and disks;
- The use of **structured, formatted messages based on agreed standards** (such that messages can be translated, interpreted and checked for compliance with an explicit set of rules);
- **Relatively fast delivery** of electronic documents from sender to receiver (generally implying receipt within hours, or even minutes); and
- **Direct communication between applications** (rather than merely between computers).

EDI depends on a moderately sophisticated information technology(IT) infrastructure. This must include data processing, data management and networking capabilities, to enable the efficient capture of data into electronic form, the processing and retention of data, controlled access to it, and efficient and reliable data transmission between remote sites.

A common connection point is needed for all participants, together with a set of electronic mailboxes (so that the organizations' computers are not interrupted by one another), and security and communications management features.

It is entirely feasible for organizations to implement EDI directly with one another, but it generally proves advantageous to use a third-party network services provider.

14.4 EDI Standards

EDI is considered to be a technical representation of a business conversation between two entities, either internal or external. There is a perception that "EDI" consists of the entire electronic data interchange paradigm, including the transmission, message flow, document format, and software used to interpret the documents. EDI is considered to describe the rigorously standardized format of electronic documents.

The EDI (Electronic Data Interchange) standards were designed to be independent of communication and software technologies. EDI can be transmitted using any methodology agreed to by the sender and recipient. This includes a variety of technologies, including modem (asynchronous, and bisynchronous), FTP, Email, HTTP, AS1, AS2, MQ, etc. It is important to differentiate between the EDI documents and

NOTES

the methods for transmitting them. While comparing the bisynchronous 2400 bit/s modems, CLEO devices and value-added networks used to transmit EDI documents to transmitting via the Internet some people equated the non-Internet technologies with EDI and predicted erroneously that EDI would be replaced along with the non-Internet technologies. These non-internet transmission methods are being replaced by Internet Protocols such as FTP, telnet and email but the EDI documents themselves still remain.

As more trading partners use the Internet for transmission, standards have emerged. In 2002, the IETF published RFC 3335, offering a standardized, secure method of transferring EDI data via e-mail. On July 12th, 2005, an IETF working group ratified RFC4130 for MIME-based HTTP EDIINT (aka. AS2) transfers, and is preparing similar documents for FTP transfers (aka. AS3). While some EDI transmission has moved to these newer protocols the providers of the value-added networks remain active.

EDI documents generally contain the same information that would normally be found in a paper document used for the same organizational function. For example an EDI 940 ship-from-warehouse order is used by a manufacturer to tell a warehouse to ship product to a retailer. It typically has a ship to address, bill to address, a list of product numbers (usually a UPC code) and quantities. It may have other information if the parties agree to include it. However, EDI is not confined to just business data related to trade but encompasses all fields such as medicine (e.g., patient records and laboratory results), transport (e.g., container and modal information), engineering and construction, etc. In some cases, EDI will be used to create a new business information flow (that was not a paper flow before). This is the case in the Advanced Shipment Notification (856) which was designed to inform the receiver of a shipment, the goods to be received and how the goods are packaged.

There are two major sets of EDI standards: the United Nations recommended UN/EDIFACT is the only international standard and is predominant outside of North America; and the U.S. standard ANSI ASC X12 (X12) is predominant in North America. These standards prescribe the formats, character sets, and data elements used in the exchange of business documents and forms. The complete X12 Document List includes all major business documents, including purchase orders (called "ORDERS" in UN/EDIFACT and an "850" in X12) and invoices (called "INVOIC" in UN/EDIFACT and an "810" in X12).

The EDI standard says which pieces of information are mandatory for a particular document, which pieces are optional and give the rules for the structure of the document. The standards are like building codes. Just as two kitchens can be built "to code" but look completely different, two EDI documents can follow the same standard and contain different sets of information. For example a food company may indicate a product's expiration date while a clothing manufacturer would choose to send color and size information.

14.4.1 EDIFACT

- EDIFACT stands for EDI for administration, commerce and transportation.
- It has been introduced by the UN center for the facilitation of administration, commerce and transportation (UN/CEFACT) in the mid 1980s.

- Older European EDI standards such as TRADACOM, GENCOD, SEDAS and ODETTE have all migrated to EDIFACT.
- EDIFACT has furthermore has become an international standard as ANSI has stopped all work on X12 since 1997 and X12 systems are migrating to EDIFACT.

EDIFACT: Structure

- EDIFACT Interchanges consist of messages which are in turn composed of data segments. The segments themselves consist of data elements.

Standards are generally updated each year.

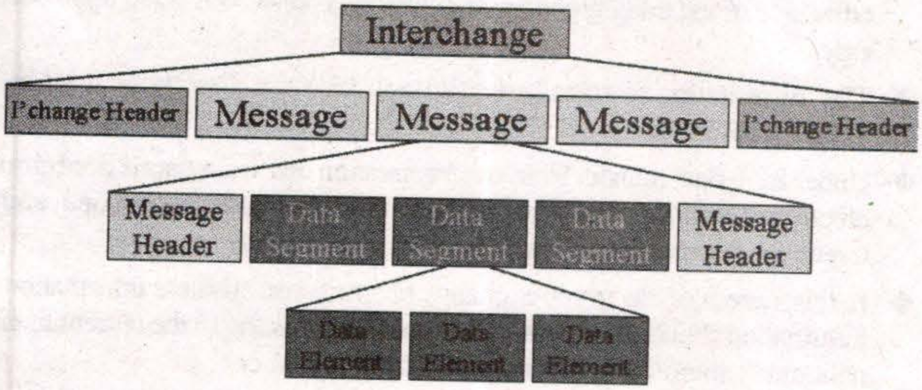


Figure 14.4 : EDIFACT Structure

14.4.2 Components of EDI Standards

The following outlines the three basic structural components of EDI standards. Both the ANSI X12 and EDIFACT standards are based on these components.

1. **A syntax and encoding scheme for messages which specifies the structure of data.** The data should be independent of systems, machine and media constraints and should allow for human interpretation of the data transferred. As well, the data elements or groupings which are part of standard messages should be independent of each other so that one part may be changed without affecting any other part.
2. **A data dictionary.** This component of EDI standards defines the standard business data elements, such as date, time, delivery address, and currency used to create messages.
3. **Combinations of data elements to be used for standard messages.** A paper invoice, for instance, normally consists of a header portion stating the name and address of the billing party, the name and address of the paying party, the date of the invoice, an account number, etc. There is then a detail portion which consists of a series of invoice lines, each giving details of a billed transaction such as date, order number, number of units, item number, item description, unit price, and total price. There may also be a summary portion which gives totals. Each of these sections has an equivalent in EDI format with data elements combined into "segments" and segments combined into "messages".

NOTES

14.5 Features of EDI

The fundamental difference between EDI and electronic mail and fax is that the data exchanged through the EDI is meant for direct use by the destination recipient computer.

There are following special features of EDI :

- The formats of various business documents like invoices, purchase orders, acknowledgement and other documents are agreed upon by the trading partners availing the facility of EDI.
- There is no need in EDI that the information received should be read by the other side of the trading partner and then interpreted and acted upon accordingly.
- The information so transferred through EDI can directly be used by the recipient computer.
- Under EDI, data transfer from one business house to another is done through electronic documents which have to be electronically forwarded and entered into the destination trading partner's computer.
- In this system of electronic exchange of structured business information, the information should be in standard formats conforming to the current industry, national or international format standards.
- The EDI establishes an electronic link between companies.
- Its special feature is that it has reduced data entry links, avoids paper-based systems and improves business cycle time.
- In EDI, the concept of trading partners is very wide which may include intra company business links or external companies, internal means the groups of departments and external companies are referred to as suppliers, customers and subsidiaries.
- It pertains to the fact that the information transferred through EDI will not have to be re-typed, re-keyed or printed. But the condition is that the information should be in pre-determined and pre-defined structure agreed by the trading partners who send and receive business data.
- It pertains to communicate with the trading partners by exchange of data with several trading partners directly and interactive with multiple companies through central information clearing houses.
- The third party's computer systems under EDI allows the sender to communicate with an unlimited number of trading partners.
- The translation service of the EDI translates the into standard formats for transmission and for reverse translation in data it receives.
- The application services, translation service and communication services of the EDI put together make the working EDI complete and give essence to online business.
- The mechanism is to pull down the forms, documents, process description and capture noting etc. into predefined and agreed standard formats.
- It may be noted that the aim of EDI is not completely to eliminate paper work, but rather to eliminate processing delays, data re-entry, re-typing and errors and discrepancies.

- In EDI, the data is transferred electronically in machine readable form and the message is mechanically processed by the receiving computer without any human intervention, interpretation or re-keying delay.

EDI Drivers

- Ability to strengthen partnerships
- Improve business processes
- A communication tool to allow new ways to do business
- A preferred way of doing business among Fortune 500 companies
- A business basic for the industry

14.6 EDI Technology

Electronic data interchange (EDI) is an automated method of placing electronic transactions. The transactions involve transmitting a standardized message from one computer to another . The components involved in the EDI transactions are the transaction software and an EDI value added network (VAN). The transaction software is a computer application that sends and receives messages that adhere to the American National Standard Institute (ANSI) X.12 transaction sets. The ANSI X.12 standard dictates the messages' syntax in order to allow interoperability among EDI capable devices. The United Nations Standard Messages (UNSMs) are international EDI standards developed by the United Nations Electronic Data Interchange For Administration, Commerce and Trade (UN/EDIFACT) . The UNSMs are analogous to the ANSI X.12 transaction sets.

When an EDI transaction is executed, the transaction software will transform the data into the format defined by the ANSI X.12 standard and transmit this information to an EDI VAN. The EDI VAN is an EDI service provider that transmits messages to their destinations. Destinations are identified by an EDI address that points to an EDI mailbox. The mailboxes store transactions for a subscriber until the translation software connects to the EDI VAN and downloads EDI transactions from their mailboxes. When a transaction is downloaded, the EDI VAN will send a "functional acknowledgment" message to the sender in order to let the sender know that the recipient has received the transaction. The translation software will send any outward-bound EDI transactions to trading partners during the time when the translation software is connected to the EDI network.

Because there are many EDI VANs, such as MCI and GE Information Service (GEIS), they are interconnected in order to allow trading partners to have EDI mailboxes residing on different VANs. These interconnections allow EDI messages to traverse the EDI VAN of the sender and the EDI VAN of the subscriber. An EDI address will identify a subscriber's EDI VAN and EDI mailbox. In order to establish an electronic relationship, the EDI address of the trading partners and the transaction set(s) for those partners must be programmed in the translation software

The high end EDI VANs allow customers to connect to EDI services using many types of protocols such as asynchronous connections, bisynchronous connections , SNA , Organization for Data Exchange through Tele-Transmission in Europe File Transfer Protocol (ODETTE FTP), X.25, X.400 and X.435. Most EDI VANs also provide access using switched and dedicated connections.

The EDI transaction goes beyond simple computer-to- computer communication and involves application to application communication. Specific transactions such as

NOTES

exchanging purchase orders and invoices, allow companies to automate their business processes.

14.6.1 Value-Added Network

A **value-added network (VAN)** is a specialized application service provider (ASP) that acts as an intermediary between trading partners sharing data or business processes. VANs traditionally transmitted data formatted as Electronic Data Interchange (EDI) but increasingly they also transmit data formatted as XML. VANs usually service a given vertical or industry and provide value-added services such as data transformation between formats (EDI?XML, EDI?EDI, etc.). At one extreme a VAN hosts only horizontal business-to-business (B2B) application integration services, hosting general-purpose integration services for any process or industry.

At the other extreme a VAN also hosts process-specific or industry-specific predefined integration capabilities (e.g., data synchronization services as part of the Global Data Synchronization Network (GDSN)) and applications (e.g., supply chain order visibility). Traditionally, most VANs primarily only supported general-purpose B2B integration capabilities focused on EDI but these service providers are quickly evolving to become more process- and industry-specific over time, particularly in industries such as retail and hi-tech manufacturing.

VANs today require a global footprint with capabilities, tools and people to service supply chains that extend from Shanghai to New York, Thailand to Hungary. Modern Value-Added Networks today are also referred to as trading grids. A VAN not only receives, stores, forwards messages but also adds audit information to the messages, it modifies the data (automatic error detection and correction, protocol conversion) and then transports the information.

Examples of VAN companies include Sterling Commerce, IBM, INOVIS, IBMIE, GXS and BT*EDINET.

14.6.2 EDI Specifications

Organizations that send or receive documents from each other are referred to as "trading partners" in EDI terminology. The trading partners agree on the specific information to be transmitted and how it should be used. This is done in human readable specifications (also called EDI Implementation Guidelines). While the standards are analogous to building codes, the specifications are analogous to blue prints. (The specification may also be called a mapping but the term mapping is typically reserved for specific machine readable instructions given to the translation software.) Larger trading "hubs" have existing EDI Implementation Guidelines which mirror their business processes for processing EDI and they are usually unwilling to modify their EDI business practices to meet the needs of their trading partners. Often in a large company these EDI guidelines will be written to be generic enough to be used by different branches or divisions and therefore will contain information not needed for a particular business document exchange. For other large companies, they may create separate EDI guidelines for each branch/division.

14.6.3 EDI Requirements

Each trading partner has unique EDI requirements. These will include the specific kinds of EDI documents to be processed. The fact is that most any business document that one company would exchange with another company can be sent via EDI. However each EDI document must be exchanged with the partner in exactly the format they specify.

Many partners will have an EDI implementation guide or kit that explains their specific requirements. Maps are required to translate the EDI documents from the trading partner's format into the format that is usable by the receiving party. EDI capability involves either buying or outsourcing the following components:

- Software for communications
- VAN service for EDI transmission
- Mailboxing of EDI transactions
- Mapping
- Translation

VAN, ASYNC, BISYNC and direction connection and Internet communications will be required by various partners. A server or PC, communication devices and peripherals will be needed as well as secured office space, monitored security, backups and redundant power. Additional software will be needed if integration of the EDI transactions with back office systems is desired. A VAN will need to be contracted for transmissions. Personnel must be trained in how to use the software and communication devices. Maps will then need to be developed.

14.6.4 EDI Transmission

Trading partners are free to use any method for the transmission of documents. In the past one of the more popular methods was the usage of a bisync modem to communicate through a "Value Added Network" (VAN). Some organizations have used direct modem to modem connections, "Bulletin Board System" (BBS), and recently there has been a move towards using the some of the many Internet protocols for transmission, but most EDI is still transmitted using a VAN. In the healthcare industry, a VAN is referred to as a "Clearinghouse".

Value Added Networks

In the most basic form, a VAN acts as a regional post office. They receive transactions, examine the 'From' and the 'To' information, and route the transaction to the final recipient. VAN's provide a number of additional services, e.g. retransmission of documents, provide third party audit information, and act as a gateway for different transmission methods, handling telecommunications support, etc. Because of these and other services VAN's provide, businesses frequently use a VAN even when both trading partners are using Internet-based protocols. Healthcare clearinghouses perform many of the same functions as a VAN, but have additional legal restrictions that govern protected healthcare information.

Internet

Until recently the Internet transmission was handled by nonstandard methods between trading partners usually involving FTP or email attachments. There are also standards for embedding EDI documents into XML. Many organizations are migrating to this protocol to reduce costs. For example, Wal-Mart is now requiring its trading partners to switch to the AS2 protocol.

14.6.5 Interpreting data in EDI

Often missing from the EDI specifications (referred to as EDI Implementation Guidelines) are real world descriptions of how the information should be interpreted by the business receiving it. For example, suppose candy is packaged in a large box that contains 5 display boxes and each display box contains 24 boxes of candy packaged for the consumer. If an EDI document says to ship 10 boxes of candy

NOTES

it may not be clear whether to ship 10 consumer packaged boxes, 240 consumers packaged boxes or 1200 consumer packaged boxes. It is not enough for two parties to agree to use a particular qualifier indicating case, pack, box or each; they must also agree on what that particular qualifier means.

EDI translation software provides the interface between internal systems and the EDI format sent/received. For an "inbound" document the EDI solution will receive the file (either via a Value Added Network or directly using protocols such as FTP or AS2), take the received EDI file (commonly referred to as a "mailbag"), validate that the trading partner who is sending the file is a valid trading partner, that the structure of the file meets the EDI standards and that the individual fields of information conforms to the agreed upon standards. Typically the translator will either create a file of either fixed length, variable length or XML tagged format or "print" the received EDI document (for a non-integrated EDI environments). The next step is to convert/transform the file that the translator creates into a format that can be imported into a company's back-end business systems or ERP. This can be accomplished by using a custom program, an integrated proprietary "mapper" or to use an integrated standard based graphical "mapper" using a standard data transformation language such as XSLT. The final step is to import the transformed file (or database) into the company's back-end ERP.

For an "outbound" document the process for integrated EDI is to export a file (or read a database) from a company's back-end ERP, transform the file to the appropriate format for the translator. The translation software will then "validate" the EDI file sent to ensure that it meets the standard agreed upon by the trading partners, convert the file into "EDI" format (adding in the appropriate identifiers and control structures) and send the file to the trading partner (using the appropriate communications protocol.

Another critical component of any EDI translation software is a complete "audit" of all the steps to move business documents between trading partners. The audit ensures that any transaction (which in reality is a business document) can be tracked to ensure that they are not lost. In case of a retailer sending a Purchase Order to a supplier, if the Purchase Order is "lost" anywhere in the business process, the effect is devastating to both businesses. To the supplier, they do not fulfill the order as they have not received it thereby losing business and damaging the business relationship with their retail client. For the retailer, they have a stock outage and the effect is lost sales, reduced customer service and ultimately lowers profits.

In EDI terminology "inbound" and "outbound" refer to the direction of transmission of an EDI document in relation to a particular system, not the direction of merchandise, money or other things represented by the document. For example, an EDI document that tells a warehouse to perform an outbound shipment is an inbound document in relation to the warehouse computer system. It is an outbound document in relation to the manufacturer or dealer that transmitted the document.

14.7 Advantages of EDI

EDI saves unnecessary re-capture of data. This leads to faster transfer of data, far fewer errors, less time wasted on exception-handling, and hence a more streamlined business process. Benefits can be achieved in such areas as inventory management, transport and distribution, administration and cash management. EDI offers the prospect of easy and cheap communication of structured information throughout the government community, and between government agencies and their suppliers and clients.

EDI can be used to automate existing processes. In addition, the opportunity can be taken to rationalize procedures, and thereby reduce costs, and improve the speed and quality of services. Because EDI necessarily involves business partners, it can be used as a catalyst for gaining efficiencies across organisational boundaries. This strategic potential inherent in EDI is expected to be, in the medium term, even more significant that the short-term cost, speed and quality benefits.

EDI and other similar technologies save a company money by providing an alternative to or replacing information flows that require a great deal of human interaction and materials such as paper documents, meetings, faxes, email, etc. Even when paper documents are maintained in parallel with EDI exchange, e.g. printed shipping manifests, electronic exchange and the use of data from that exchange reduces the handling costs of sorting, distributing, organizing, and searching paper documents. EDI and similar technologies allows a company to take advantage of the benefits of storing and manipulating data electronically without the cost of manual entry or scanning.

The benefits and advantages of EDI can be summarized as :

- Improved reporting performance
- Time savings
- Cost savings
- Improved accuracy
- Enhanced flexibility

Improved reporting performance

Electronic submissions are a much more efficient way to transmit the legally required information . Typically, the electronically submitted EDI data will be received, processed and acknowledged within hours of when it was submitted, rather than the multiple days it would take through the postal system.EDI allows the department's trading partners to meet their reporting deadlines in a timely manner.

Time savings

EDI claim submission provides an efficient means of getting the correct information to the department as quickly as possible. EDI saves time by eliminating the overhead of the paper handling that is required and is otherwise necessary for both the trading partner and the department.

Cost savings

Although there are initial costs involved with designing, developing and implementing a new EDI system, these costs can be recouped and the system can pay for itself many times over by the efficiencies garnered by the use of EDI. The cost of mailing and handling paper documents is completely avoided when the documents are sent electronically.

Personnel at both ends of the electronic transaction that would otherwise be involved in the handling of paper generated information can be redeployed to other tasks. There are fewer people required to monitor and administer the EDI system than is necessary to process paper documents.

Improved accuracy

EDI reduces the number of times the same data needs to be redundantly entered into multiple computer systems. There is also the inherent efficiency and improved

NOTES

accuracy from the electronic acknowledgment process that allows for the senders' transactions to be verified and validated immediately upon receipt. The acknowledgment process allows the trading partner to submit more timely and accurate information while, at the same time, reducing the amount of time it takes to correct invalid or inaccurate information.

Enhanced flexibility

Electronic data can be sent anytime, day or night, to ensure the most accurate and timely information is delivered in an efficient manner. The EDI submission of first report transactions can be scheduled to run when the computing resources are at a lower demand.

14.8 Barriers in adopting EDI

There are a few barriers to adopting electronic data interchange. One of the most significant barriers is the accompanying business process change. Existing business processes built around slow paper handling may not be suited for EDI and would require changes to accommodate automated processing of business documents. For example, a business may receive the bulk of their goods by 1 or 2 day shipping and all of their invoices by mail. The existing process may therefore assume that goods are typically received before the invoice. With EDI, the invoice will typically be sent when the goods ship and will therefore require a process that handles large numbers of invoices whose corresponding goods have not yet been received.

Another significant barrier is the cost in time and money in the initial set-up. The preliminary expenses and time that arise from the implementation, customization and training can be costly and therefore may discourage some businesses. The key is to determine what method of integration is right for your company which will determine the cost of implementation. For a business that only receives one P.O. per year from a client, fully integrated EDI may not make economic sense. In this case, businesses may implement inexpensive "rip and read" solutions or use outsourced EDI solutions provided by EDI "Service Bureaus". For other businesses, the implementation of an integrated EDI solution may be necessary as increase in trading volumes brought on by EDI force them to re-implement their order processing business processes.

The key hindrance to a successful implementation of EDI is the perception many businesses have of the nature of EDI. Many view EDI from the technical perspective that EDI is a data format; it would be more accurate to take the business view that EDI is a system for exchanging business documents with external entities, and integrating the data from those documents into the company's internal systems. Successful implementations of EDI take into account the effect externally generated information will have on their internal systems and validate the business information received. For example, allowing a supplier to update a retailer's Accounts Payables system without appropriate checks and balances would be a recipe for disaster. Businesses new to the implementation of EDI should take pains to avoid such pitfalls.

Increased efficiency and cost savings drive the adoption of EDI for most trading partners. But even if a company would not choose to use EDI on their own, pressures from larger trading partners (called hubs) often force smaller trading partners to use EDI.

14.9 Drawbacks of EDI

- EDI adapter software is too expensive for most organizations.
- The software has to be practically rewritten for different combinations of VANs, internal hardware and trading scenarios.

- The software is also subject to change when there is a revision in EDIFACT
- VAN subscription costs and dedicated line costs are prohibitive for most SMEs.
- The EDI system is highly static and every business process has to be meticulously negotiated between partners.
- Since there is no common registry or discovery mechanism, partners have to retain information on institution codes, product codes, up-to-date catalogs etc. associated with everybody they do business with.

14.10 New Trends in EDI

The future of EDI involves using the Internet instead of an EDI VAN for transporting EDI messages. This enhancement will allow trading partners to conduct electronic transactions without the added cost of paying for an EDI mailbox residing on an EDI VAN. However, problems that need to be overcome include those involving customer service, standards, and security.

Traditional EDI customer service handles many types of problems, specifically, those that result when messages do not arrive at their destination. Customer service personnel and technicians have limited visibility into the Internet because the Internet is a public network consisting of over 40,000 different networks. Therefore, it is very difficult to troubleshoot.

Problems associated with the standardization of EDI transaction sets may also hinder wide spread use of EDI over the Internet because transaction software must be re-written to operate using the Internet as a transport medium. However, the Internet Engineering Task Force (IETF), the governing body for Internet standard, has devised a standard for sending EDI documents over the Internet .

Security of the data traversing the Internet is a concern because the Internet is a public network that can send data over many other public networks. Data may be compromised by individuals who have control over a particular node in the network. Those individuals can easily view data that traverses the node. Many vendors are reacting to this threat by offering Internet EDI products that contain encryption and key management software . Individuals must be aware that all trading partners must have software containing common encryption algorithms and keys in order to conduct electronic transactions.

14.10.1 XML - EDI

XML-EDI is the collection of five core technologies:

- **XML:** Provides the foundation and replaces EDI segment identifiers with XML tokens.
- **EDI:** Provides the business methods and existing process data formats and specifications.
- **Templates:** Define the rules by which the data is to be processed. They provide the glue which holds the process together.
- **Agents:** Interpret the Templates to perform the work needed, and also interact with the transaction and the user to create new templates for each new specific task, or look up and attach the right template for existing jobs.

NOTES

- **Repository:** This component provides the semantic foundation for business transactions and the underpinning that the Agents need to correctly cross-reference entities.

14.10.2 EDI and ebXML

- Both the ANSI X12 group and the EDIFACT group have joined in the ebXML initiative as of July 2001
  - The EDI groups will be work in the development of the core components of ebXML dealing with business process integration.
- ANSI has about 300 different business processes and EDIFACT contains about 1200 processes.
  - The business processes outlined in EDI use standardized message segments which are re-usable.
  - The main problem is defining the EDI business processes currently in use with an unambiguous format. When this is done, the processes will become syntax independent.
  - Once this is done, EDI and ebXML will be able to use the same processes and be able to communicate regardless of syntax.

In order to meet the forthcoming demands and needs of the business community, two major organizations of India, viz., National Informatics Center (NIC) and Videsh Sanchar Nigam Limited (VSNL) have established EDI-VAN services. Other VAN service providers include Satyam Online and Global Telecom services. Other organizations that have been active in developing and promoting various aspects of e-commerce in India include the Federation of Indian Export Organisations (FIEO), the All India Management Association (AIMA) and NASSCOM.

Systems analysts should evaluate the use of EDI technology in their enterprise infrastructure because EDI has the ability to automate many business processes. Traditionally, systems analysts had an inward focus in developing network and computer strategies that included a limited set of intra-company applications. Systems analysts must now look outward towards inter-company communication technologies, like EDI, to improve logistical performance and operation efficiencies.

SUMMARY

- Electronic Data Interchange is a set of standards for structuring information to be electronically exchanged between and within businesses, organizations, government entities and other groups. The standards describe structures that emulate documents, for example purchase orders to automate purchasing.
- EDI enables a commonly understood and standardized format of the relevant data to be transmitted from one computer system to another with minimal human intervention. EDI transactions are structured for highly automated processing.
- The Internet has enabled EDI transactions to be transmitted between trading partners in an even more efficient manner. The Internet provides business and government agencies with an environment that is open, fast, cost effective, and widely accepted and used.

- The concept of a value-added network (VAN) is a mechanism that facilitates the transfer of electronic data between trading partners. A VAN can be thought of as a post office that allows an entity to send EDI formatted data to one of their trading partners at any time. The VAN will hold the file of transmitted transactions until the trading partner to whom it is addressed retrieves it at a later time.
- EDI can be compared and contrasted with electronic mail (email). Email enables free-format, textual messages to be electronically transmitted from one person to another. EDI, on the other hand, supports structured business messages (those which are expressed in hard-copy, pre-printed forms or business documents), and transmits them electronically between computer applications, rather than between people.
- EDI depends on a moderately sophisticated information technology(IT) infrastructure. This must include data processing, data management and networking capabilities, to enable the efficient capture of data into electronic form, the processing and retention of data, controlled access to it, and efficient and reliable data transmission between remote sites.
- The EDI (Electronic Data Interchange) standards were designed to be independent of communication and software technologies. EDI can be transmitted using any methodology agreed to by the sender and recipient.
- EDIFACT stands for EDI for administration, commerce and transportation.
- The application services, translation service and communication services of the EDI put together make the working EDI complete and give essence to online business.
- In EDI, the data is transferred electronically in machine readable form and the message is mechanically processed by the receiving computer without any human intervention, interpretation or re-keying delay.
- The components involved in the EDI transactions are the transaction software and an EDI value added network (VAN).
- The high end EDI VANs allow customers to connect to EDI services using many types of protocols such as asynchronous connections, bisynchronous connections.
- *A* **value-added network (VAN)** is a specialized application service provider (ASP) that acts as an intermediary between trading partners sharing data or business processes.
- VANs usually service a given vertical or industry and provide value-added services such as data transformation between formats (EDI?XML, EDI?EDI, etc.). At one extreme a VAN hosts only horizontal business-to-business (B2B) application integration services, hosting general-purpose integration services for any process or industry.
- Examples of VAN companies include Sterling Commerce, IBM, INOVIS, IBMIE, GXS and BT*EDINET.
- The future of EDI involves using the Internet instead of an EDI VAN for transporting EDI messages.
- Both the ANSI X12 group and the EDIFACT group have joined in the ebXML.
- National Informatics Center (NIC) and Videsh Sanchar Nigam Limited (VSNL) have established EDI-VAN services in India .

NOTES

- EDI reduces the number of times the same data needs to be redundantly entered into multiple computer systems.
- In EDI terminology "inbound" and "outbound" refer to the direction of transmission of an EDI document in relation to a particular system, not the direction of merchandise, money or other things represented by the document.
- EDI Benefits:
  - Builds closer business partnerships
  - Reduces/eliminates manual handling of data, errors and rework
  - Transfers information faster and more accurately
  - Automates routine transactions
  - Improves productivity and business controls
  - Reduces costs
  - Shortens transaction processing cycles
  - Enhances data accuracy
  - Lowers inventory levels
  - Contributes to better in-stock positions
  - Lowers freight costs
  - Provides Quick Response capability
  - Improves cash-flow management
  - Creates a competitive advantage

EXERCISE

1. What is EDI ? Describe its role in modern business transactions .
2. Describe the origin and history of EDI.
3. What is the structure of EDI systems? Describe the security and privacy framework of EDI.
4. What are the essential components of EDI systems?
5. What are EDI standards and why these are required ?
6. What is EDIFACT? Describe its structure.
7. What are different components of EDI standards.
8. Describe different features and benefits of EDI.
9. What are the drivers behind EDI ?
10. What are different technologies working behind the EDI systems?
11. What is VAN ? Describe its role in EDI .
12. What are major requirements for EDI systems ?
13. Describe the complete EDI transmission cycle .
14. How are the data interpreted in EDI ?
15. What are advantages and drawbacks of EDI?
16. What are the role of modern XML – EDI and ebXML.