UNIT-1

DISTRIBUTED SYSTEM

DISTRIBUTED SYSTEM:

→It is an interconnection of autonomous

→Computers that operate parallel execution of related task

→A computer system is distributed in 3 ways

- Hardware –single CPU with many systems
- Control-control of whole system
- Database-multiple copies of data

DISTRIBUTED DATA PROCESSING (DDP):

→DDP is defined as, centralized site, processing in organization around multiple processing elements.

→Processing Element (PE) is a computer or other device capable of doing functions

→PE is organized on a functional or geographical basic.

→Distributed elements co-operate in support with users requirements.

→Connection of PE is through network carriers or private link.

→The dispersion of hardware, software, data to multiple Processing Elements (PE).

WHY GO DISTRIBUTED?

→Machines perform functions in cost-effective system maintenance.

→Data is processed easily.

→It co-ordinates the organizations

→Users can communicate with each other

→It requires top management involvement user participation planning

Eg: Hotel reservation, Banking

PROS AND CONS OF DISTRIBUTED PROCESSING:

DISADVANTAGES:

→Loss of control → provide proper audit among departments

→Duplication of software resources

→Duplication of data → duplication of data in corporate database

→ Redundancy

→Hardware problems

→Revision of past efficiencies- project control programming

→Maintenance to remote site

→Incompatibility of hardware & software

ADVANTAGES:

→Reduction of cost

-local processing is done easily

-use of less expensive computers

-decreased complexity of applications

Eg: Payroll, Accounts receivable

-better utilization

→Response time improvements

-improve response time for applications

(DIAGRAM)

→User control

-very flexible and cost effective

-if a particular site fails, backup is transferred to another site

DISTRIBUTED DATABASE:

REASONS FOR DISTRIBUTING DATA:

→Placing data at the organizations department's personal computers can reduce the amount of data transferred among the host computers, resulting in reduced communication cost.

→**Local Access of Data** can improve response time

→**Distributed databases** can give reliability to a system because data are located at more than one site

→The provision for local storage gives users more control over the data

→Distributed data bases provide management control of the process for an organization

REASONS FOR DISTRIBUTING DATA:

→Placing data on organizations departments on personal computers reduce the amount of data transferred results in reduced communication cost

TYPES OF COMMUNICATION:

- **Centralized**
- **Partitioned**
- **Replicated**

CENTRALIZED:

→It is not distributed

→It is used for comparison

→All data reside at one site

→All data queries and updates from the remote site are transmitted to the central site

(DIAGRAM)

PARTITIONED:

→Data is split into pieces or partitioned and assigned selected department or sites in a network

→Partitioned database may be of the same data structure format or access method called as **homogeneous partitioned system**

→Partitioned database may be of different format called as **heterogeneous partitioned system**

REPLICATED:

→Multiple copies of the data reside at different departments or different sites in a network

→Same data among replicated data are called homogeneous replicated data

→Duplicated data are placed under different access method called heterogeneous

Distributed Database

(i) Vertically
(ii) Horizontally

Vertically:

- Vertical topology
- Detailed data are stored locally

Horizontally:

- Horizontal topology
- Data is distributed among departments(or) peer sites in the network

Database can be placed in network in various combinations

- ❖ Centralization
- ❖ Partitioning
- ❖ Replication
- ❖ Homogeneity
- ❖ Heterogeneity
- ❖ Vertical distribution
- ❖ Horizontal distribution

DISTRIBUTED DATA TERMS & CONCEPTS:

1) **CONSISTENT STATE:**
   →**A**ll data in a network database are correct & accurate
   →The replicated copies contain the same value in the data field
   Eg: deposit =$100 saving=$250 total=$350 if data is consistent
2) **TRANSACTION:**

   →A sequence of operations on a database

   Eg: transferring fund from one account to another

3) **TEMPORARY INCONSISTENCY:**

   →State of during execution of transaction

   SEQUENCE OF OPERATION BEGIN

Π – 1Read customer A account

Π – 2 Read customer B account

Π – 3 write customer A account- $100 to customer A account

Π – 4 write customer B account- $100 to customer B account

4) **CONFLICT:**

→Occurs when two (or) move transactions are involved in the update of same data

Eg: Transaction 1 transfer $100 from A to B

Transaction 2 transfer $75 from A to B

→After the execution of transaction the result is conflict the data are inconsistent. The total deposit of two customers should be $100 with $125 in customer A account and $475 in customer B account.

CHALLENGES OF DISTRIBUTED DATA:

→A distributed data is deceptively complex. This will describe the problems and complexity happened in distributed database

→Some of the techniques are available to avoid some of the complexities and problems.

→The complexities and problems can arise due to inconsistency and an updation when the two users simultaneously accessed or update the data item at the same time. Only one updation can be possible. The other user's data will be lost.

→The temporary inconsistency may occur when a user shares a particular database and it can be stored into the local databases

→User A and B simultaneously update an item in database in the absence of control machine.
→The Database updates only one the other lost.

→It happens when both users retrieve the data item change it and revised value back into the database.

(DIAGRAM)

→User C retrieves the data and stores in a replicated local database

→Then the other series consistency problem and it is not a temporary inconsistency

→The databases are in conflict

→The various problems and complexities can be avoided using different methods. They are

1) Lock outs & deadly embrace
2) Lock outs & sharing with personal computers
3) Update & retrieval overhead
4) Failure & recovery
5) Major factors in distributed data decisions

1) **LOCK OUTS & DEADLY EMBRACE:**

   →Common solution is preventing the sides from simultaneously execution on the same data

   →Lock out can work well in centralized database but in distributed network

   →Mutual Lockout or Deadly Embrace can arise, consider A&B are two users, X&Y are two databases

   →A locks Y from B and B locks Y from A. After completing the transaction both the users wants to access the transaction locked Database but neither can access the locked sites

   →So the sides are entered into mutual exclusion state

   (DIAGRAM)

2) **LOCK OUTS & SHARING WITH PERSONAL COMPUTERS:**

   →The personal network provides two functions are used to access the database by multiple users. They are

   - File Sharing
   - User Query Read Direction

   →The share function allows the initial user to stipulate that the database can only be read by any users who subsequently attempt to access the data. This technique is called read or write axis deny write sharing mode

   →The user query redirection can reroute a user's transaction over the network to a database at another computer after getting an acknowledgement signal any user can access in particular database

3) **UPDATE AND RETRIEVAL OVERHEAD:**

   →The use of locks provide effective consistency

   →When properly implemented with serialized scheduling locks provide valuable method for maintaining database integrity but it is a costlier method

   →The following communication messages can be used while transmitting data between two sides. They are

   - Lock request
   - Lock grant
   - Update
   - Update acknowledgement

- Lock release

→A sent lock request message to B & C

→B & C sent lock grant message to A if it accepts

→A transmits the update transaction

→B & C update database and transmit to A. Acknowledgement of an update

→A receive and transmit message to release the lock

(DIAGRAM)

→A locking algorithm requires 5(n-1)

→Each of the messages require Data Link Control (DLC) messages to ensure that the database messages are properly received

→A DLC requires four DLC control messages to every data message

→The number varies from 3 to 9, so the number 4 is conservative

→The locking algorithm is actually D[5(n-1)] where D is the number of overhead messages

Eg: in the figure assuming a communication overhead of factor 4, 40 messages are transported to update at 2 sides 40=4*[5*(3-1)]

→It also takes additional cost of environments that have heterogeneous data systems

→It also requires cost for protocol conversions for using various data structures

→Retrieval overhead problems arises during partitioning the database

→A data retrieval request may require the data needed at several databases in the network

→While retrieving data the following concepts have to be considered

1) Data Link Control
2) Protocol Conversion
3) Database Structure Translation
4) Database Software Execution
5) Operating System Efficiency
6) Mission Processing Power

FAILURE & RECOVERY:

→To achieve resiliency in a distributed system is different from other environments

→If there is any failure the OS suspend the execution of the program and store the query registers and blocks during which the problem components does not change

→In a DDP system, the problem may not be suspended in a centralized system. Since the distribution system have horizontal topologies or near autonomous computers

MAJOR FACTORS IN DISTRIBUTED DATA DECISIONS:

→The general guidelines for deciding how to distribute data in a network area are as follows

1) Frequency of use at each site
2) User control of data
3) Real time update requirements
4) Backup requirements
5) Class of data
6) Cost to store locally Vs. cost to transmit remotely
7) Security consideration
8) Time of use of data
9) Volume of data accessed
10) Retrieval response time requirements
11) Location of data users
12) Retrieve access Vs. update accesses

DISTRIBUTED DATA TERMS AND CONCEPTS

SCHEDULE:

→A schedule is ordering of events

SERIALIZABLE SCHEDULE:

→Transactions are executed serially (i.e.,) consistent state

LOCKING:

→It is used to prevent multiple transactions creating conflict in the database

→Database is responsible for locking

→Lock is released after transaction is complete

RESILIENCY:

→Ability to return to the original form

→Data stored to the original form

→Failure should not affect the another system

MANAGING THE DISTRIBUTED RESOURCES:

→The advantages in distributed processing are remarkable

→Many fully tested hardware and software components are now available

Eg: Distributed network architecture managers load leveling or resource sharing among the sites

→Distribute Data Processing (DDP) leads computerized resources to various departments in an organizational structure

→Distributed Data Processing (DDP) management effectively administers hardware, software and data resources that exist throughout the network

→The use of DDP requires organization management control and guidance

→A distributed system may have millions of source statements (program instructions), hundreds of programs, thousands of data item and scores of databases residing in many sites

→Many of these resources will be duplicated at various network sites and departments in the organization

→Different departments develop, maintain and update their data files and programs

→It is not uncommon for a site to develop programs or create data for its own needs

→Some organizations permit users to modify the code or files that were originally developed for general use

→Companies do not always track and record those changes. So redundant programs performs similar functions

(DIAGRAM)

→Site I develop a program or database resources passes it to site II

→To meet its specific needs site II modifies the core program

→At a later date, site I changes its version of the program and then passes modified program to site II

→The alter program does not fulfill site II needs and in some instances may not function in second environment

→Site II cannot accept the change to maintain the core program and enhanced program

→Thus distributed system become inaccurate

→The ideal solution is to give the distributed sites freedom to manage their resources and develop information systems in a distributed network

MANAGEMENT MODEL:

→**Distributed Automation Management (DAM)** is a management and control mode for a distributed network

→DAM is an architecture containing detailed information about the automated resources existing in an organization

→It can be applied to an organization with distributed departments for data processing needs

→The advantages of DAM is that all dispersed sites can participate in the decisions and implementations of a developing system, management can monitor the process

(DIAGRAM)

→When a company chooses DAM approach it must establish well defined standards for structuring the information in sub systems and software modules

For eg: Distributed modules will require more complex interfaces and leads to probability error

→In the fig. DAM contains information on Project management, Data/Database Administration, Software Inventory, Equipment Inventory & Sources and uses of company reports

→These responsibilities are broken down into functions, which are handled by the individual elements and sub elements

DAM ELEMENTS:

ELEMENTS AND SUB ELEMENTS	FUNCTIONS AND STORED ELEMENTS
System Life Cycle Management (SLC)	1)identification of project phases 2)designation of site responsibility for project deliverables 3)description of end products 4)milestones due date for deliverables 5)exception reporting procedures 6)description of host implementation review Milestones 7)designation of site for acceptance testing

Node Responsibility Matrix (NRM)	1)designation of site responsibility for ongoing Operations 2)identification of sites security responsibilities 3)description of rules for change to common Components 4)designation of audit sites & milestones for Audit
Internode Data Administration (IDA)	1)information on organizations data, data files, databases and data items across all distributed sites 2)description of data 3)node awareness and access to data 4)description of identifier schema and naming conventions 5)data relationship between local and remote Sites
Physical Data Model (PDM)	1)physical location of data 2)physical device view of data 3)internal representation of data 4)access methods to data 5)data usage statistics
Logical Data Model	1)application program view of data 2)description of program interface to physical Data model
Database/Files & Data Item Inventories (DFII)	1)description of files and databases for distributed sharing 2)description of data groupings 3)relationship among data groupings 4)description of unique data characteristics 5)relationship between data items and databases & files
Equipment Inventory & Use (EIU)	1)description of network hardware components 2)repair history of components 3)vendor information 4)maintenance statistics & dates 5)financial information
Internode Configuration Management (ICM) **Software System Modules Inventories (SSMI)**	1)information on available programs 2)general description of program functions 3)identification of which sites use program 4)detailed description of each modules functions 5)description of hardware architecture, compilers, job control language used to run a software 6)listing of parameters for module use

Sources/ User Repository (SUR)	1)description of input forms (sources) 2)description of output reports (users) 3)relationship between data source & users and distributed software and database components

→In DAM architecture **System Life Cycle Management (SLCM)** in which system development is divided into phases such as analysis, design, testing & cut over

→The difference between SLC & centralized method is that SLC process all distributed departments and sites

→SLC includes **mile stones, due dates, exception report, description of end products, commitment for equipment installation, testing plan & other project information**

→SLC identifies acceptance sites for software development projects

→**Node Responsibility Matrix (NRM)** identifies the sites, ongoing responsibility for distributed programs, equipment, data files

→The **NRM** is a logical tool for functions, rules & identify the sites that receive altered components

→**NRM** contains technical and audit information (an important component of DAM architecture)

→To review teams in accessing nodes design and project control

→A Company can place security related entries in the NRM

INTERNODE DATA ADMINISTRATION (IDA):

→It is a management organizing company's**data files, data items** across all departments and sites

→It determines the data changes in the network database where the data should reside and controlling the replication of databases

→**IDA** has sub elements **Physical Data Model, Logical Data Model, Database File and Database, Data Item Inventories** which includes data description, data definition, type of database, location of data

→**IDA** provides naming conventions for data items and databases

→Physical Data Model → provides physical data characteristics, location of data format, physical record layout, data item relationship, access method

→Logical Data Model → provides view of data

→Interfaces to software program

→Virtual access to data from local to remote sites

Eg: if an application issues a request for data the Logical Data Model examines the request and passes to Physical Data Model, PDM examines the home locator information and initiates a request to network protocols for data

→The requested data are located and passed to logical model

DATABASE FILES AND DATA ITEM INVENTORY:

→It defines the relationship between data items, other data items, databases, data files and distributed nodes

→**DDI** standardized names for data elements in the distributed network

ADVANTAGES:

- ✓ Naming standards → allows users programs to exchange data
- ✓ Layers → easily interface with one another
- ✓ Defining data elements and establish names
- ✓ Handle derived data and time dependent data

EQUIPMENT INVENTORY:

→It describes network hardware components and holds information on vendor names, model number, service date, unique characteristics, repair history

USING DISTRIBUTED AUTOMATION MANAGEMENT:

→If a company decides a new automated system

→The functional requirements, data needs and output report request or review by a co-coordinating group which examines DAM to determine the new requirements can be fulfilled by automated resources

→The database/ file inventory determines if the data are already available

→The software system inventory provides a planner the option of selecting existing code for the use of modification to support new system

→DAM provides information on all forms, reports, files, software modules, software components and sites that will be altered by the new system

Eg: if there is a failure in the hardware DAM contains information that helps the management how to reconfigure the equipment

→DAM controls the distributed components

DIVISION OF RESPONSIBILITEIES:

→The departments and distributed sites must be given some responsibility and control their resources

→The below table provides a list of responsibilities that are assigned either to the distributed sites or central headquarters

CENTRALIZED	**DISTRIBUTED**
(i)Definition of central and local responsibilities	(i)participated in definition but policies are established at head quarters
(ii)selection and design of applications to multiple locations	(ii)participate in selection and design of application to serve multiple location
(iii)maintenance of data dictionary review of local sites input to dictionary	(iii)input of local data to corporate dictionary
(iv)network standard, data description language and database software	(iv)participation in standard, DDL, database software
(v)maintenance of master database	(v)input to local copies of database and up line loading to master database
(vi)selection of applications for sites	(vi)development of site applications
(vii)review of site application	(vii)follow up reviews
(viii)review of local application development	(viii)local application development
(ix)review of local database	(ix)designing of local database
(x)review of local sub schema	(x)design of sub schema
(xi)consulting services	(xi)selection of used equipment
(xii)consulting services	(xii)receives consulting services
(xiii)system security policies	(xiii)participation in development of security policies
(xiv)auditing policies and procedures	(xiv)participation in auditing procedures
(xv)coordination of operational review	(xv)conducting operational review
(xvi)coordination of secondary site testing	(xvi)secondary site testing
(xvii)PC acquisition policy	(xvii)participation of policy
(xviii) Global Data Base Administrator	(viii) Local Data Base Administrator