Mapping the data warehouse architecture to Multiprocessor architecture

T.R.Lekhaa

AP – IT

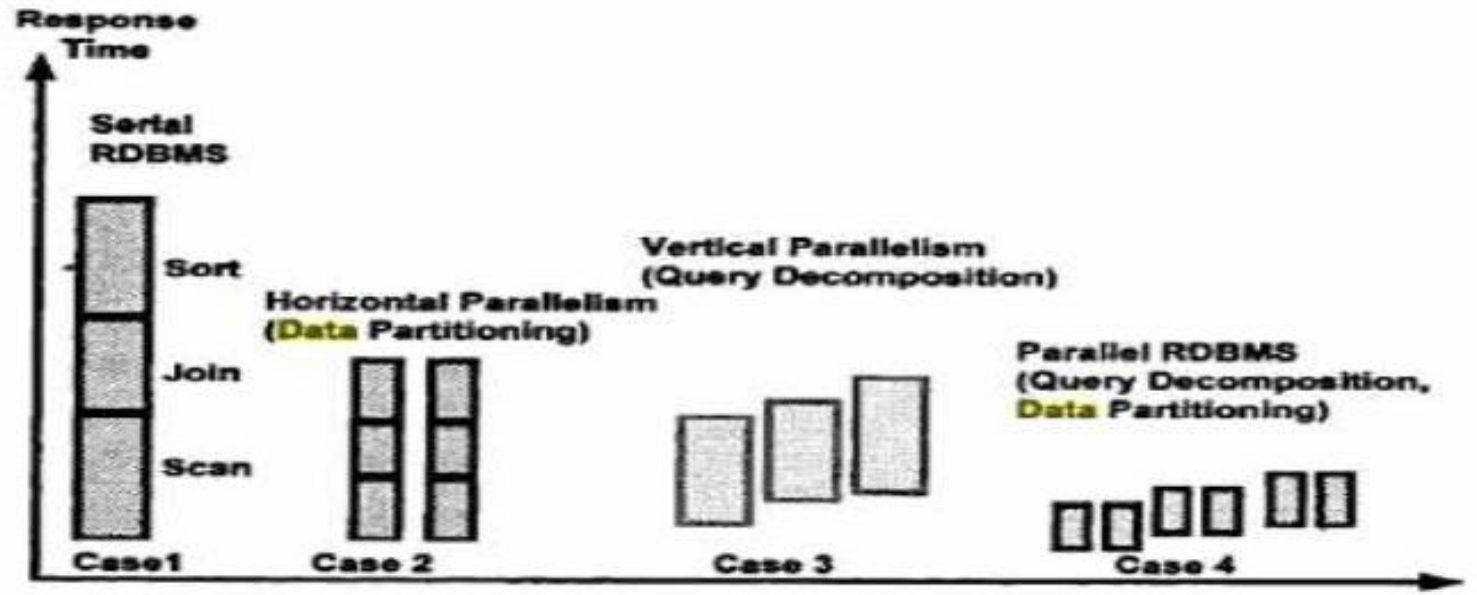
SNSCE

- The functions of data warehouse are based on the relational data base technology. The relational data base technology is implemented in parallel manner. There are two advantages of having parallel relational data base technology for data warehouse:
- **Linear Speed up: refers the ability to increase the number of processor to reduce response time**
- **Linear Scale up: refers the ability to provide same performance on the same requests as the database size increases**

Types of parallelism

- There are two types of parallelism:
- Inter query Parallelism: In which different server threads or processes handle multiple requests at the same time.
- Intra query Parallelism: This form of parallelism decomposes the serial SQL query into lower level operations such as scan, join, sort etc.

- Intra query parallelism can be done in either of two ways:
- Horizontal parallelism: which means that the data base is partitioned across multiple disks and parallel processing occurs within a specific task that is performed concurrently on different processors against different set of data
- Vertical parallelism: This occurs among different tasks. All query components such as scan, join, sort etc are executed in parallel in a pipelined fashion.



Types of DBMS parallelism.

Data partitioning

- Data partitioning is the key component for effective parallel execution of data base operations. Partition can be done randomly or intelligently.
- Random portioning includes random data striping across multiple disks on a single server. Another option for random portioning is round robin fashion partitioning in which each record is placed on the next disk assigned to the data base.
- Intelligent partitioning assumes that DBMS knows where a specific record is located and does not waste time searching for it across all disks.
- The various intelligent partitioning include:
- Hash partitioning: A hash algorithm is used to calculate the partition number based on the value of the partitioning key for each row

- Key range partitioning: Rows are placed and located in the partitions according to the value of the partitioning key.

- Schema portioning: an entire table is placed on one disk; another table is placed on different disk etc.

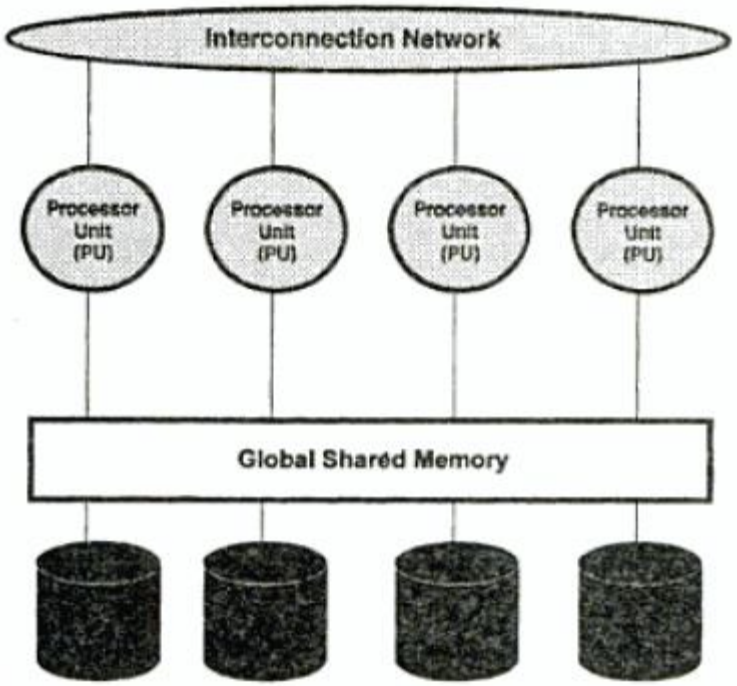
- User defined portioning: It allows a table to be partitioned on the basis of a user defined expression.

Data base architectures of parallel processing

- There are three DBMS software architecture styles for parallel processing:
- 1. Shared memory or shared everything Architecture
- 2. Shared disk architecture
- 3. Shared nothing architecture

Shared Memory Architecture

- Tightly coupled shared memory systems, illustrated in following figure have the following characteristics:
- Multiple PUs share memory.
- Each PU has full access to all shared memory through a common bus.
- Communication between nodes occurs via shared memory.
- Performance is limited by the bandwidth of the memory bus.
- Symmetric multiprocessor (SMP) machines are often nodes in a cluster.



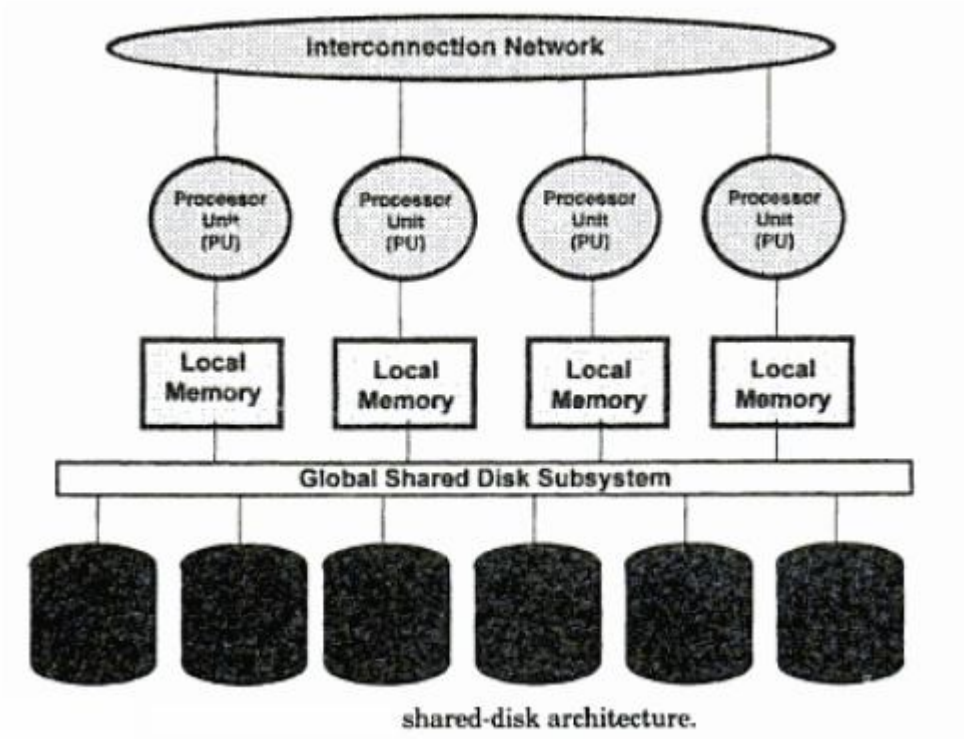
Shared-memory architecture.

- Parallel processing advantages of shared memory systems are these:
- Memory access is cheaper than inter-node communication.

- Shared memory systems are easier to administer than a cluster.
- A disadvantage of shared memory systems for parallel processing is as follows:
- Scalability is limited by bus bandwidth and latency, and by available memory.

Shared Disk Architecture

- Shared disk systems are typically loosely coupled. Such systems, illustrated in following figure, have the following characteristics:
- Each node consists of one or more PUs and associated memory.
- Memory is not shared between nodes.
- Communication occurs over a common high-speed bus.
- Each node has access to the same disks and other resources.
- A node can be an SMP if the hardware supports it.
- Bandwidth of the high-speed bus limits the number of nodes (scalability) of the system.

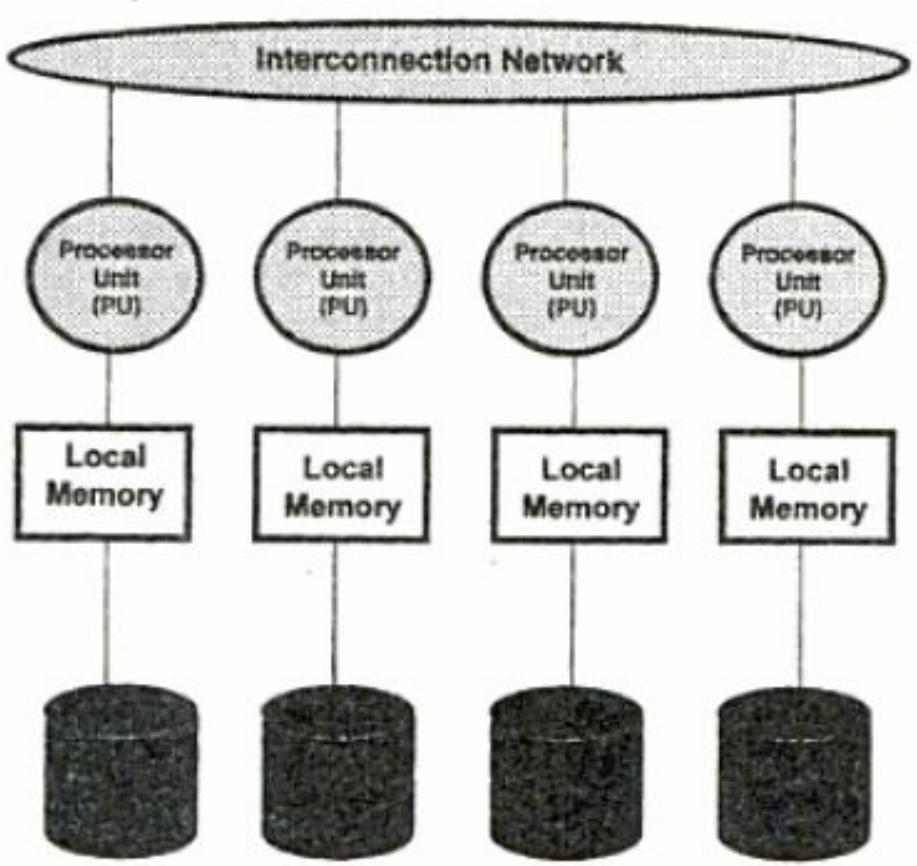


shared-disk architecture.

- **Parallel processing advantages of shared disk systems are as follows:**
- Shared disk systems permit high availability.
-
- Shared disk systems provide for incremental growth.
- Parallel processing disadvantages of shared disk systems are these:
- Inter-node synchronization is required, involving DLM overhead and greater dependency on high-speed interconnect.
- If the workload is not partitioned well there may be high synchronization overhead.
- There is operating system overhead of running shared disk software.

Shared Nothing Architecture

- Shared nothing systems are typically loosely coupled. In shared nothing systems only one CPU is connected to a given disk. If a table or database is located on that disk, access depends entirely on the PU which owns it. Shared nothing systems can be represented as follows:



Distributed memory architecture.

- **Shared nothing systems have advantages and disadvantages for parallel processing:**
- **Advantages**
- Shared nothing systems provide for incremental growth.
- System growth is practically unlimited.
- MPPs are good for read-only databases and decision support applications.
- Failure is local: if one node fails, the others stay up.
- **Disadvantages**
- More coordination is required.
- More overhead is required for a process working on a disk belonging to another node.
- If there is a heavy workload of updates or inserts, as in an online transaction processing system, it may be worthwhile to consider data-dependent routing to alleviate contention.