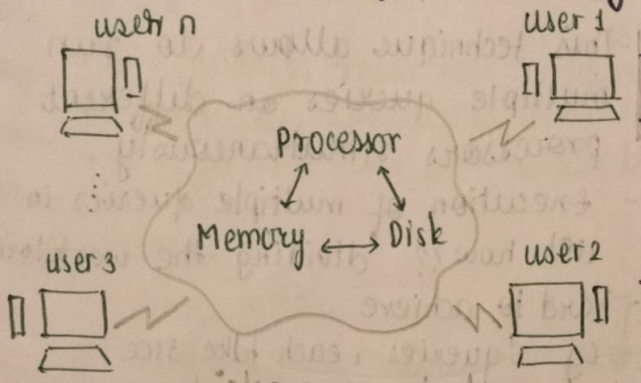
Parallel Database System.

Database that uses multiple CPU & Disk to perform db operations simultaneously is called Parallel database sys.

Goals ? Improved performance.
not there? ✗ Increased availability.
fail? ✗ Increase reliability.
Improve speed of operation.
ie high data transfer rate & huge amount of data handling.



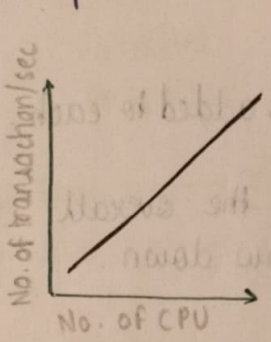
Measure of performance?

① Throughput (output efficiency):
no. of task completed in given time

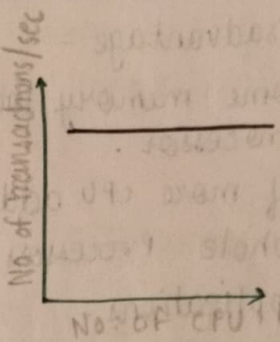
② Response time :
Amount of time taken to complete a single task from time alloted

③ Speed up :
- running task in less time by increasing degree of parallelism
- Time for processing task $\propto \frac{1}{\text{no. of resources}}$
(inversely proportional).

④ Scale up :
- handling larger task in same amount of time by increasing degree of parallelism (ie provide more resources)



Speed up



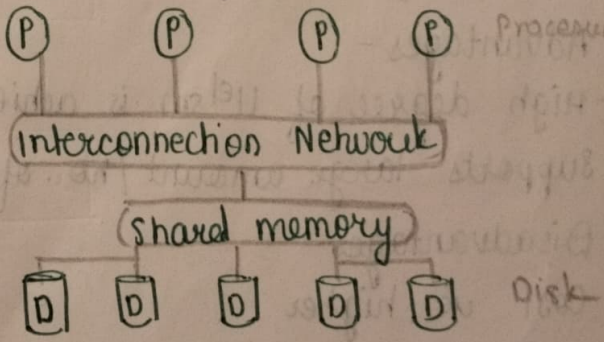
Scale up.

ARCHITECTURE OF PARALLEL DATABASE
(Types of parallel db systems)

① Shared Memory system
→ Memory - shared → Disk - pvt/shared
→ Scalability - low to high → Complexity - low

Advantages
- Effective communication betn processors.
- Data can be accessed by any processor without being moved from one place to another

Disadvantages
- Not scalable beyond 32 or 64 processor
- More no of processor can increase waiting time of processors.

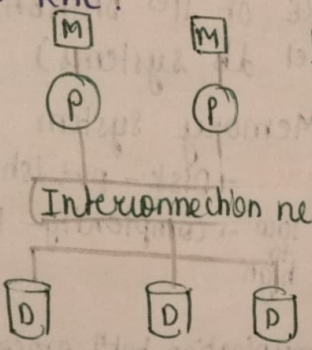


② Shared Disk System
Memory - pvt
Disk - Shared
Scalability - medium
Complexity - medium

Advantages -
- Each CPU has its own local memory, so memory bus will not face bottleneck
- If one Processor fails, other will take over its task.

Disadvantage -
- Some memory load is added to each processor.
- If more CPU added, the overall whole processors slow down.

Applications -
DEC (Digital Equipment Corporation)
Oracle RAC.

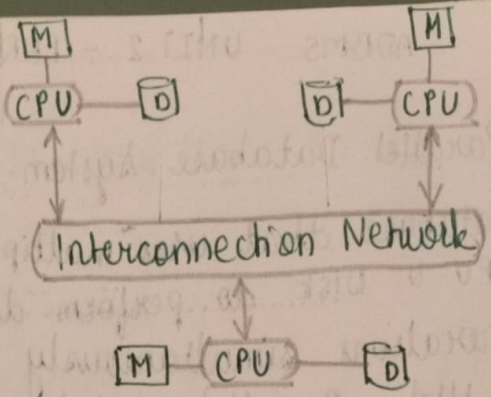


③ Shared Nothing system:
Memory - pvt
Disk - pvt
Scalability - ~~variable~~ high
Complexity - high

Advantages -
- High degree of ||elsm is achieved
- Supports large amount / no. of CPU's

Disadvantages -
- Cost is higher

Application -
Google BigQuery
Amazon Redshift
Tera data database machines



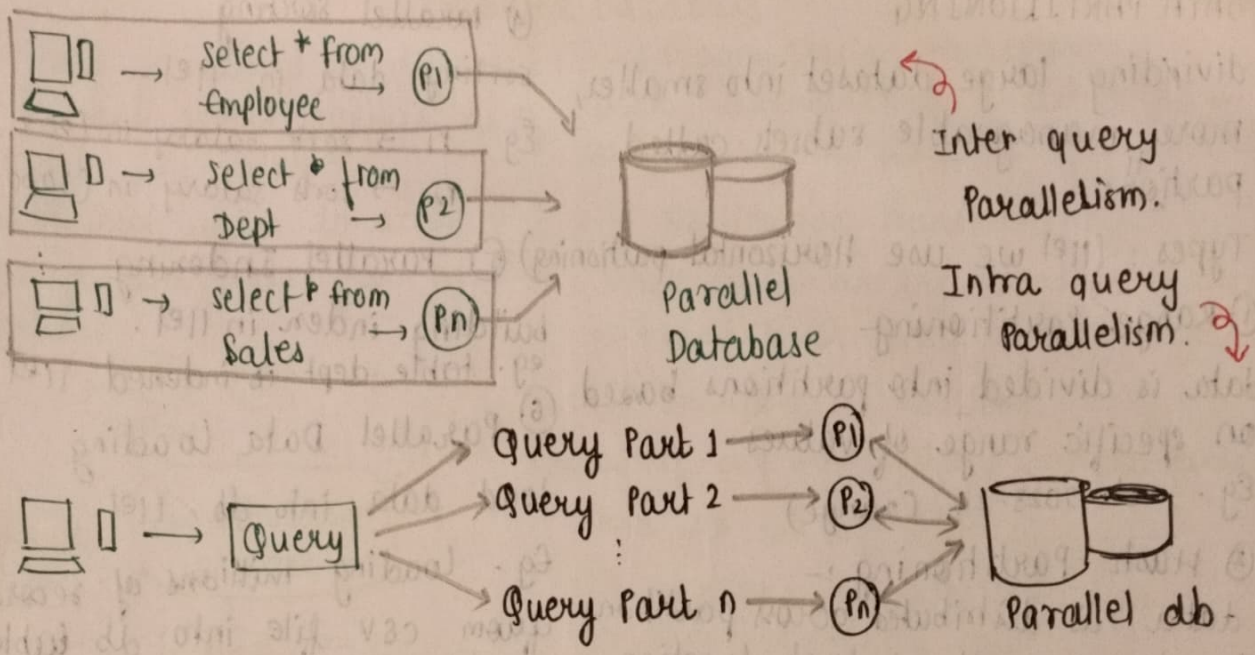
④ Hybrid
Memory - mixed
Disk - mixed
Scalability - variable
Complexity - high

PARALLEL QUERY EVALUATION

① INTER QUERY PARALLELISM
- This technique allows to run multiple queries on different processors simultaneously.
- Execution of multiple queries in ||el how?? dividing the workload
- hard to achieve
- Eg 6 queries, each take 3sec
  total time = 18 seconds
  But inter query ||el take 3sec only
- It is about " how would we execute all the above queries simultaneously by using ||el servers, so that each transaction need not wait for the other to complete".

② INTRA QUERY PARALLISM.
- In this, query is divided into subqueries which can run simultaneously on diff. processors, so minimized query evalution time.
- This improves response time of system.
- Execution of single query in ||el by dividing workload among various processors.

JOINS



PARALLEL QUERY OPTIMIZATION

Query optimization? → Most effective way to execute a given query.

Goal → minimize resources required. (efficient)
- → increase speed of returned results.
- → efficient use of available resources.

Approaches
- → use index
- → Aggregate table
- → Denormalization (combine multiple table into single table)
- → vertical partioning
- → horizontal partioning

DATA PARTITIONING

dividing large dataset into smaller, more manageable subsets called partitions.

Types : (IIel we use Horizontal partioning)

① Range Partitioning
data is divided into partitions based on specific range of values
Eg. cet 2023-24 (range)

② Hash Partitioning :-
data is distributed across partitions based on result of hash function applied on one/more columns.

③ Round-Robin Partitioning -
each row is placed in next available partition in sequence
$D = i \bmod n$ D → disk no.
i → record no.
n → total no. of disk

④ Parallel Sorting.
sorting data in IIel
Eg P_1 → sorts salary in (2000, 3000)
P_2 → sorts salary in (3000, 4000) ...

⑤ Parallel Indexing
building index in IIel.
eg. table dept is indexed IIel

⑥ Parallel Data Loading
load data into db IIel
Eg. loading millions of records from csv file into db table, where file is split into chucks & loaded in parallel.

PARALLELIZING INDIVIDUAL OPERATIONS

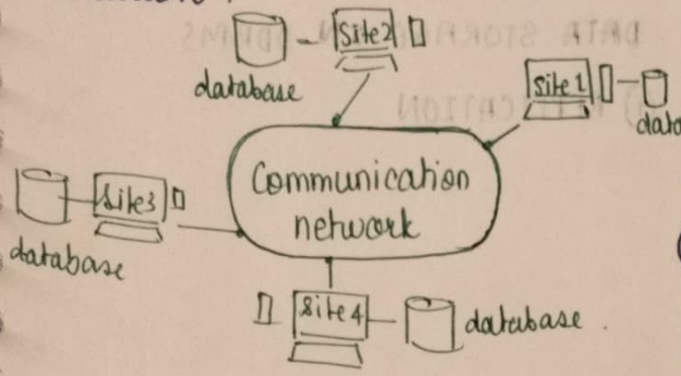
① Parallel Scanning
process of reading data from tables IIel
eg. Scanning large table, where each partition (eg. year, name, etc) is scanned by different processor.

② Parallel Joins
performing joins using IIel processing.
Eg. Joining large facts table with multiple dimension tables, where join is processed in IIel.

③ Parallel Aggregations
process of aggregate func in IIel

Distributed Database.

The data is stored in different places but works together as if it's in one place.
This allows users to access & manage the data from multiple location.



Features of DDBMS -

① Complex problems can be solved efficiently

② Scalability - adding more nodes or servers to the system.

③ High Availability - distributed databases enhance availability.

④ More computing power is generated at comparatively lower cost.

⑤ Individual processing elements can be managed independently.

⑥ User in this system gets feel that he is ~~knowing~~ working in a single centralized db system.

Advantages (refer to)

① Management of distributed data with different levels ~~flow~~ of transparency

→ Network transparency
user unaware about data comes from which server.

→ Fragmentation transparency.
user unaware of weather he is using original data or fragmented

→ Replication transparency.
user is unaware weather data accessed is original or replicated.

② Increase reliability.
if one system fails, other takes over its all function.

③ Increase availability

④ Less operating cost.

⑤ Ease for expansion

Disadvantages

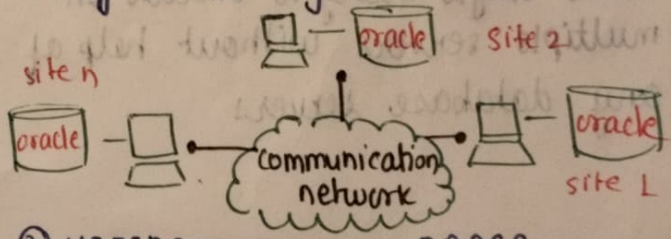
① System becomes complex to manage and control.

② Security issues must be carefully managed.

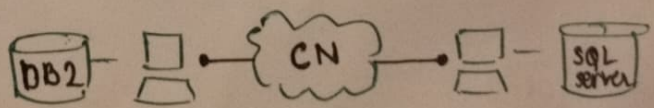
③ Requires deadlock handling.

Types of DDBMS :—

① HOMOGENEOUS DATABASE
All different sites ⇒ same DBMS software
- easy to manage.



② HETEROGENEOUS DATABASE
All different sites ⇒ different DBMS software
- a particular site might be completely unaware of other sites.
Hence, translations are required for different sites to communicate



REST IS
INCOMPLETE