University of Windsor

04 Introduction to Distributed Database Management System

Kalyani Selvarajah
School of Computer Science
University of Windsor

Advanced Database Topics
COMP 8157 01-02-03
Fall 2023

Today’s Agenda

System Architectures

Introduction to Distributed Databases

Advantages and Disadvantages of DDBMSs

Architectures of a DDBMS

https://domains.upperlink.ng/elementor-947/

Procrastination! Is it healthy?

Introductory Questions

Why Do You Want Distributed Databases?

What are the different System Architectures?

What are the challenges with DDBMS?

How DDBMS's architecture differ from Centralized DBMS's?

Distributed vs Centralized Databases

Multiple databases are distributed across different computers/nodes/servers .

vs

Entire database run on a single computer/node/server.

Google Spanner,
Azure Cosmos DB, BigQuery,
Data warehouses, etc.

vs

MySQL, SQL Server,
PostgreSQL, etc.

Why You Want Distributed Databases?

1. Scalability:

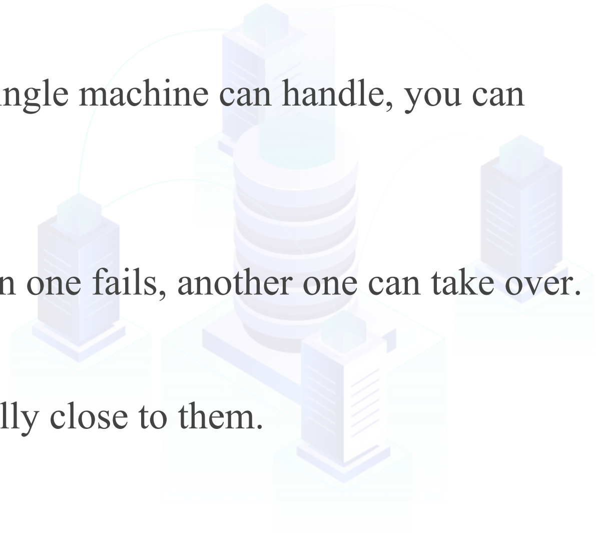
   If data volume, read load, or write load grows bigger than a single machine can handle, you can potentially spread the load across multiple machines.

2. Fault tolerance/high availability:

   You can use multiple machines to give you redundancy. When one fails, another one can take over.

3. Latency:

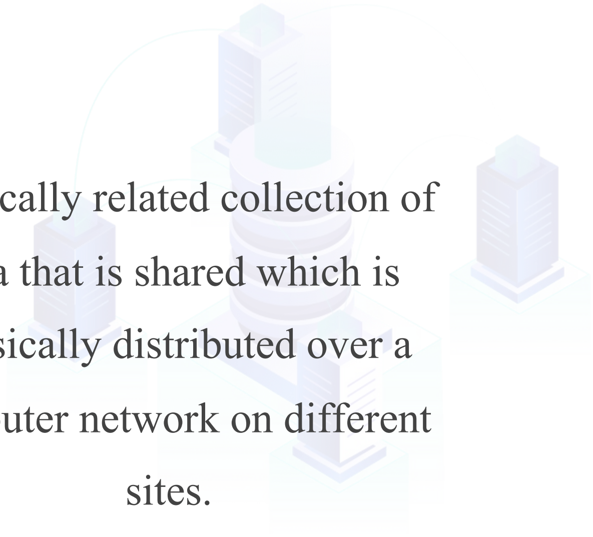
   Each user can be served from a datacenter that is geographically close to them.



Parallel vs Distributed Databases

a DBMS that runs across multiple processors and is designed to execute operations in parallel, whenever possible.

vs



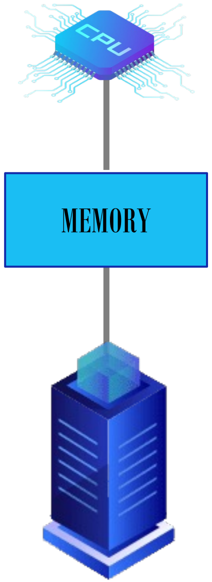
a logically related collection of data that is shared which is physically distributed over a computer network on different sites.

System Architectures

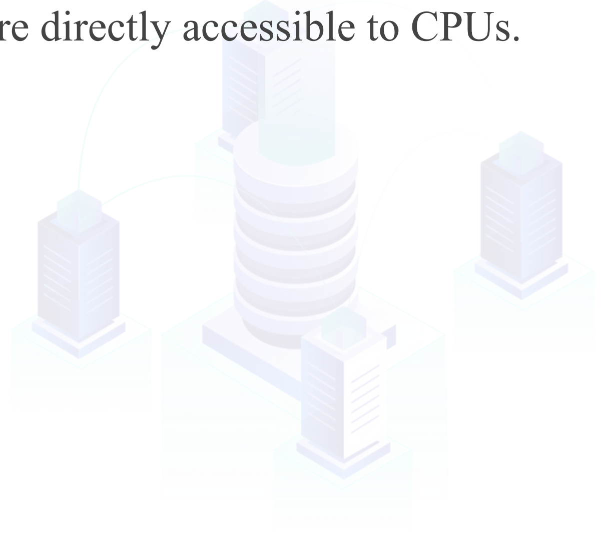
A DBMS's system architecture specifies what shared resources are directly accessible to CPUs.

System Architectures

A DBMS's system architecture specifies what shared resources are directly accessible to CPUs.

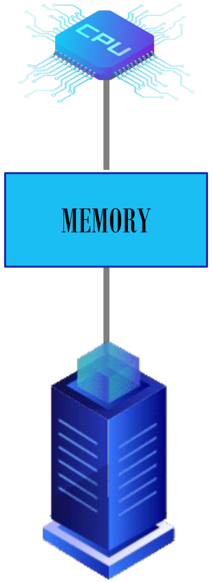


Shared Everything

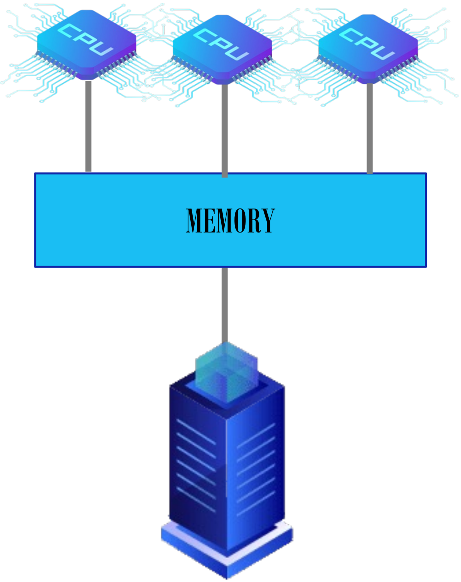


System Architectures: Scaling to Higher Load

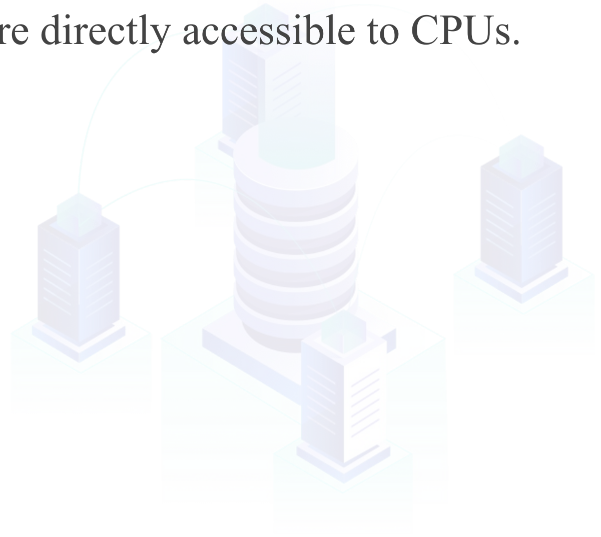
A DBMS's system architecture specifies what shared resources are directly accessible to CPUs.



Shared Everything

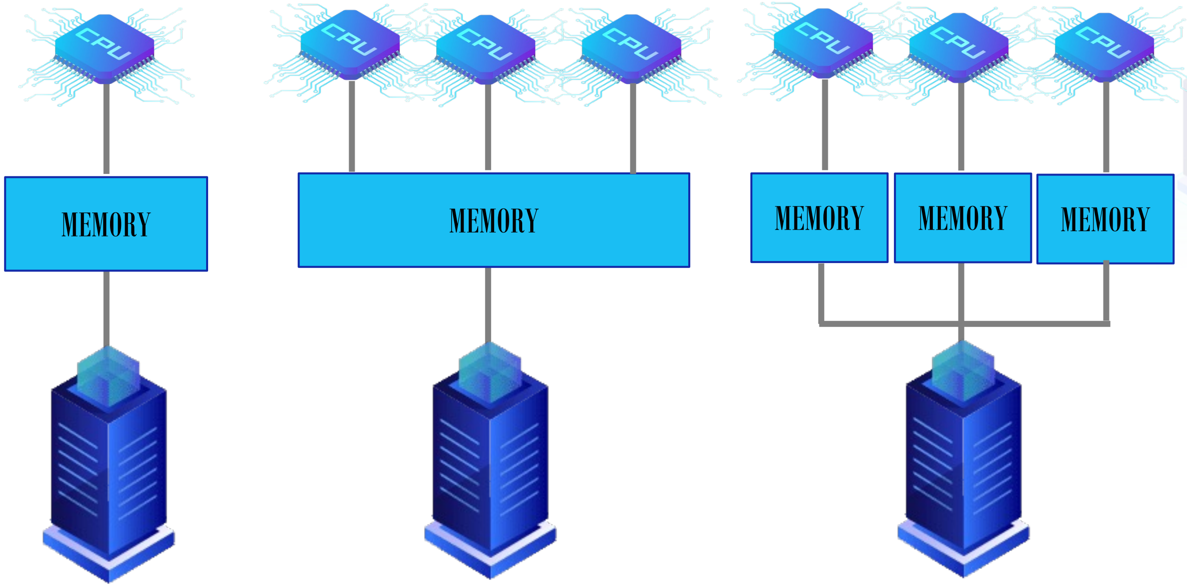


Shared Memory



System Architectures: Scaling to Higher Load

A DBMS's system architecture specifies what shared resources are directly accessible to CPUs.



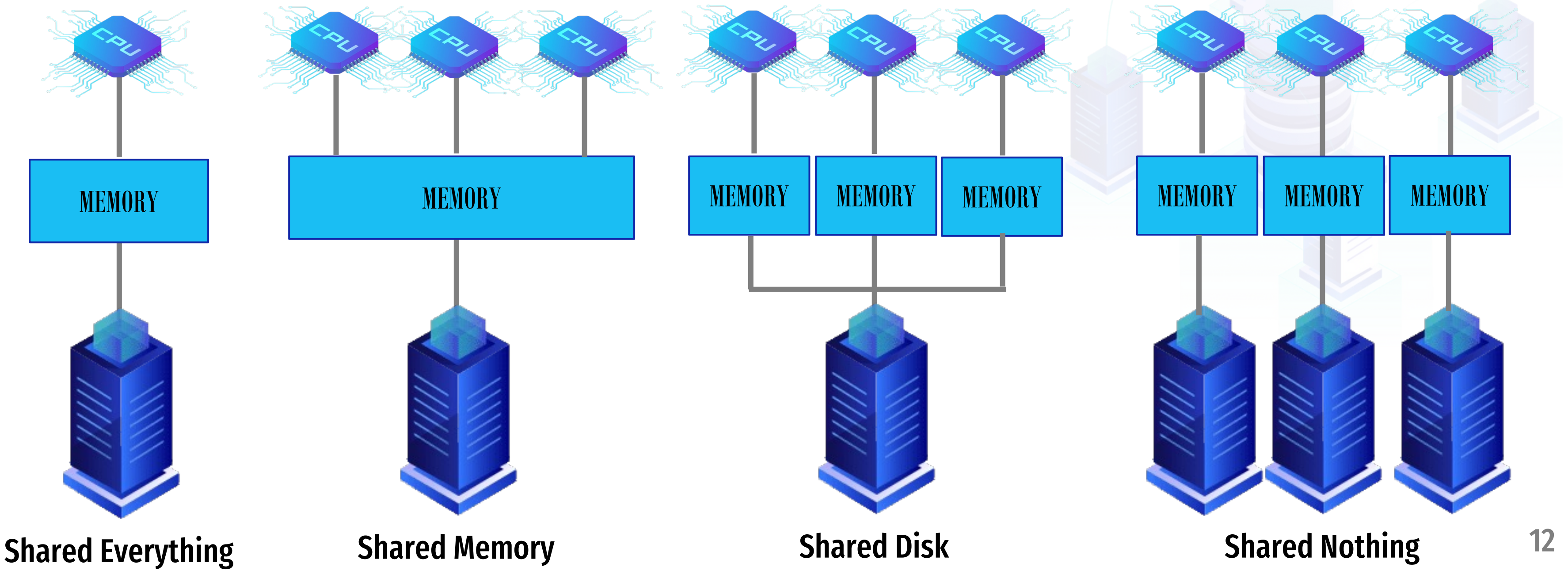
Shared Everything

Shared Memory

Shared Disk

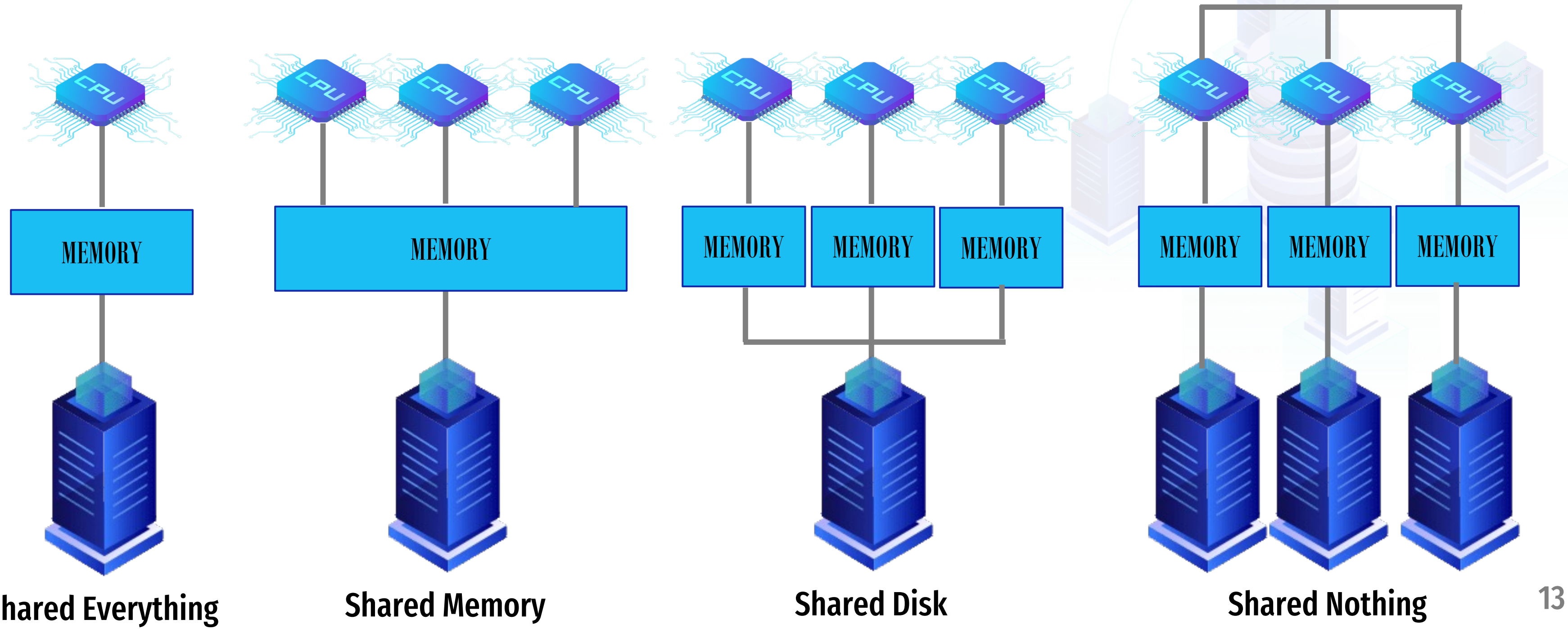
System Architectures: Scaling to Higher Load

A DBMS's system architecture specifies what shared resources are directly accessible to CPUs.

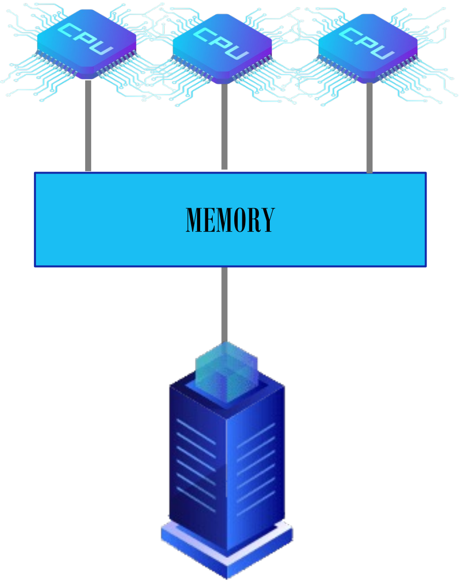


System Architectures: Scaling to Higher Load

A DBMS's system architecture specifies what shared resources are directly accessible to CPUs.



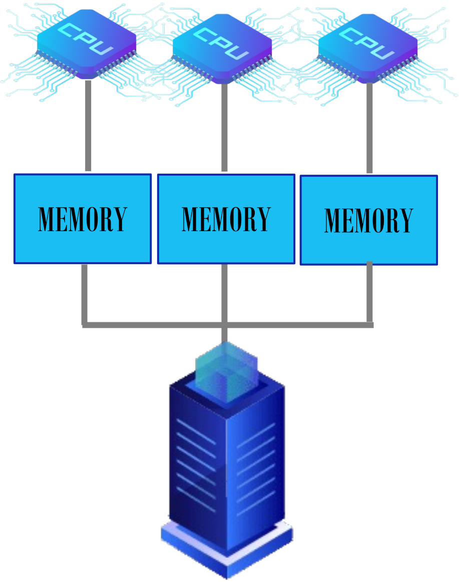
Shared Memory



Shared Memory

- ✓ All the CPUs can access the same memory and are all controlled by a single operating system.
- ✓ A fast interconnection network allows any processor to access any part of the memory in parallel.
- ✓ Advantages: simplicity and load balancing.
- ✓ Problems: cost, limited extensibility and low availability.

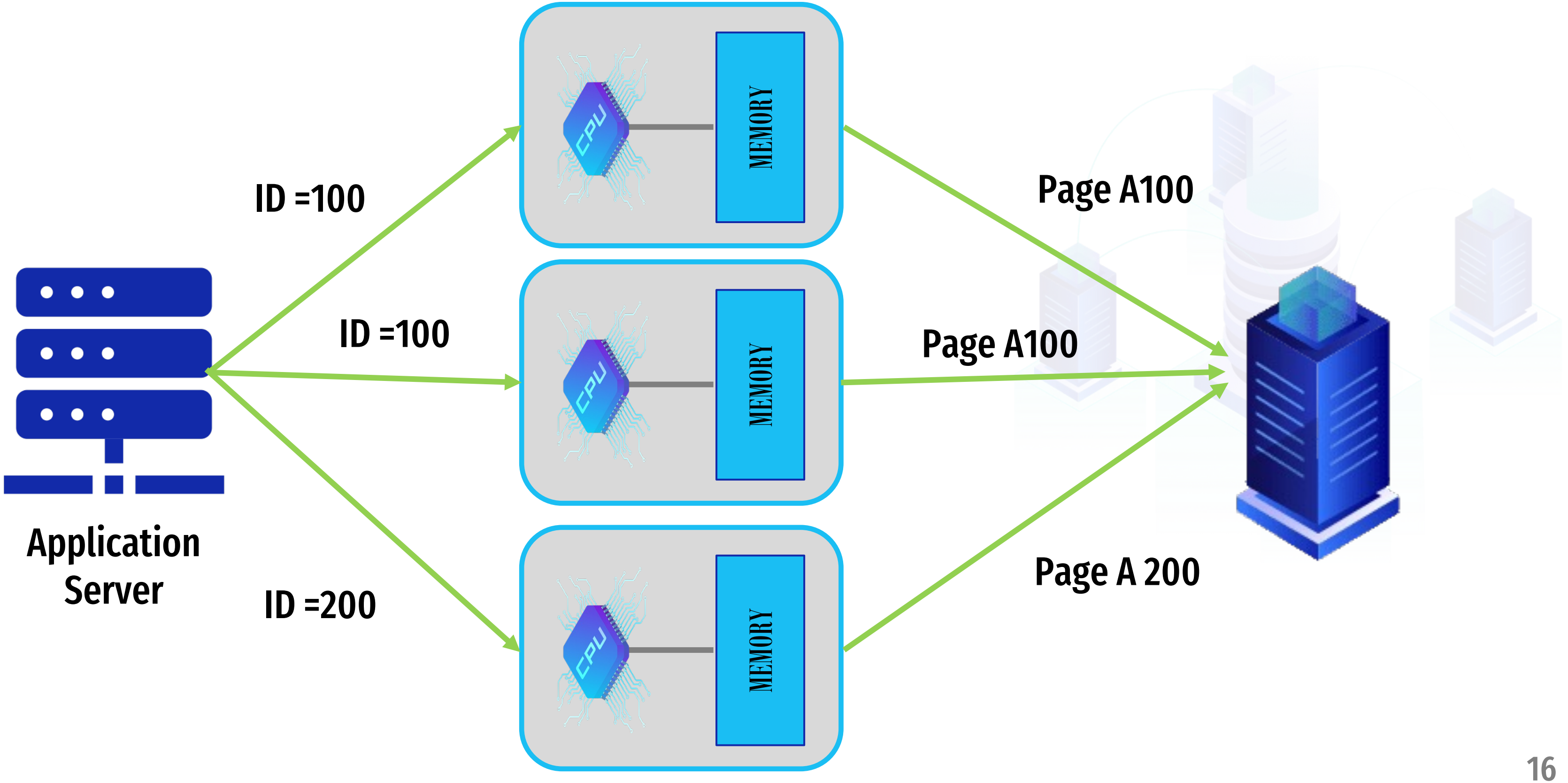
Shared Disk



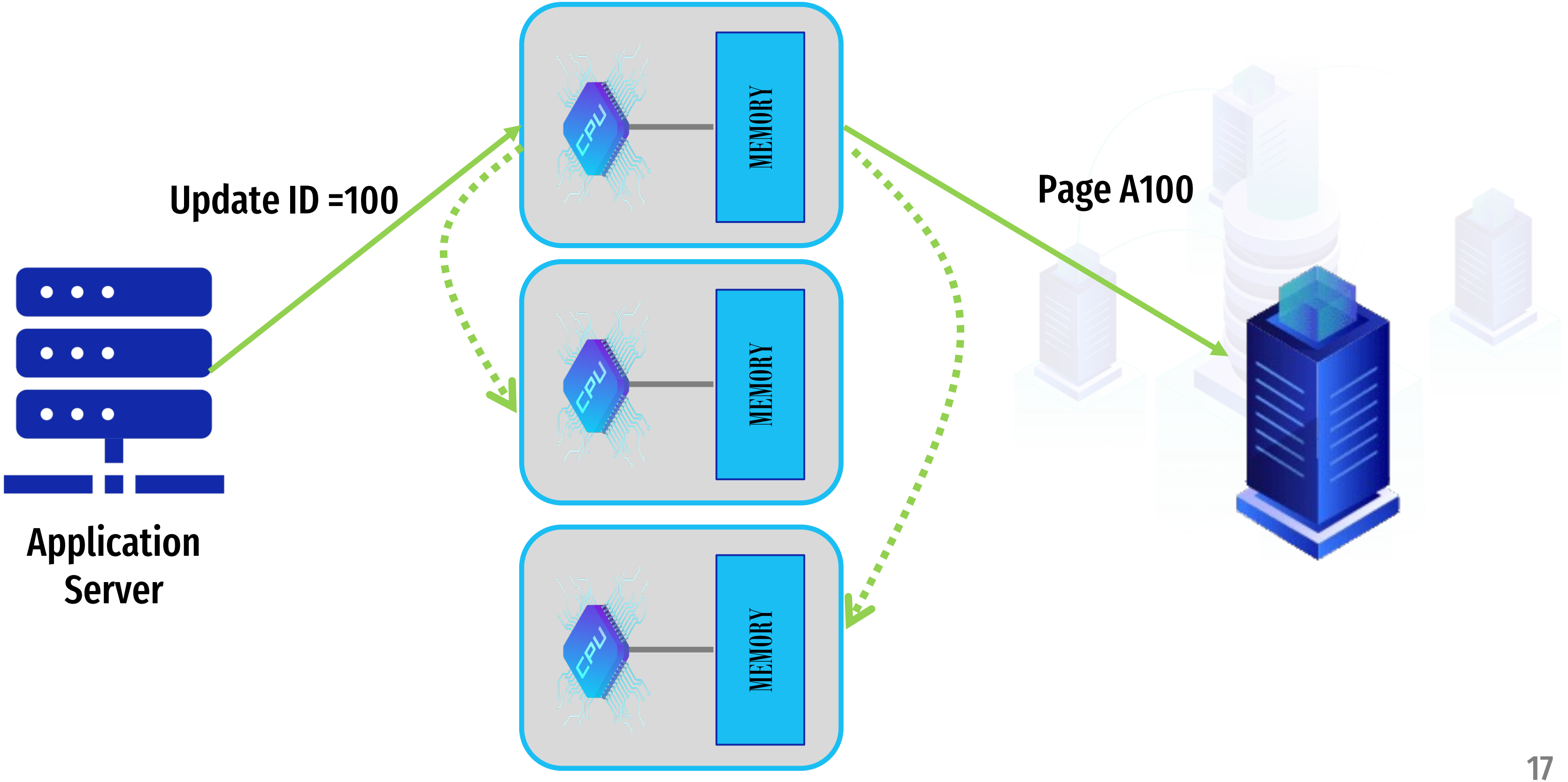
Shared Disk

- ✓ All CPUs can access a single logical disk directly via an interconnect network, but each have their own private memories.
- ✓ Disk sharing architecture requires suitable lock management techniques to control the update concurrency control.
- ✓ This is the present architecture in today's cloud environment.- The most notable parallel database system which uses shared-disk is Oracle.
- ✓ Advantages: lower cost, good extensibility, availability, load balancing, and easy migration from centralized systems.
- ✓ Problems: complexity and potential performance problems

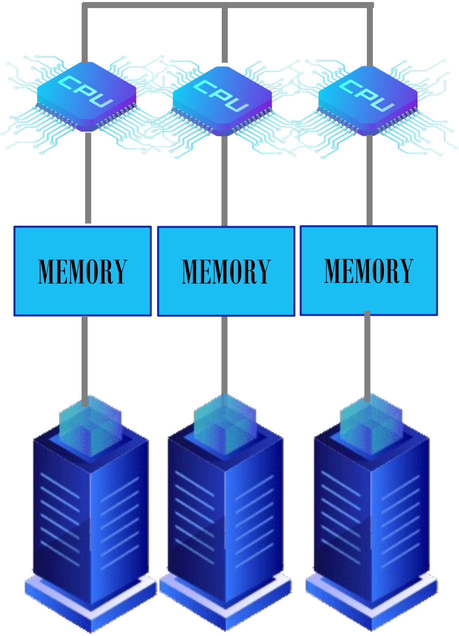
Shared Disk Example



Shared Disk Example



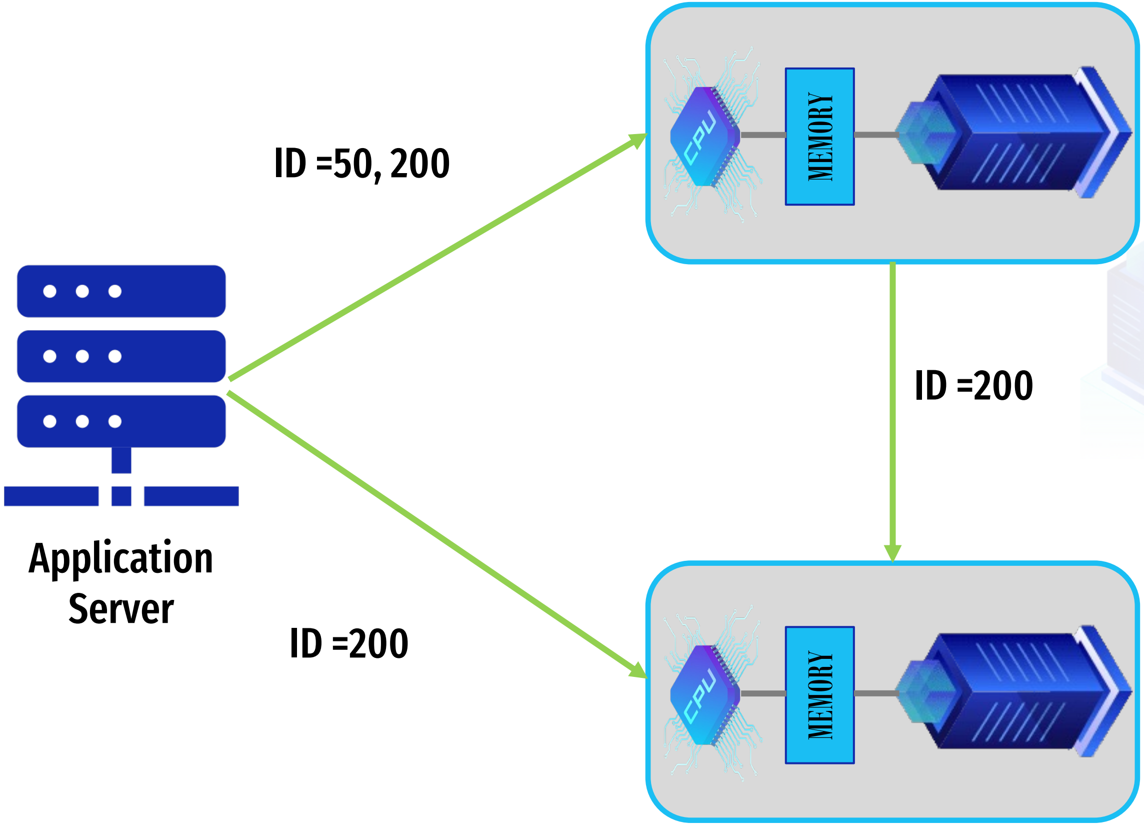
Shared Nothing



Shared Nothing

- ✓ Each node uses its CPUs, RAM, and disks independently.
- ✓ Any coordination between nodes is done at the software level, using a conventional network.
- ✓ No special hardware is required by a shared-nothing system.
- ✓ Advantages: low cost, high extensibility, and high availability.
- ✓ Problem: Higher complexity

Shared Nothing Example



ID =50, 200

ID =200

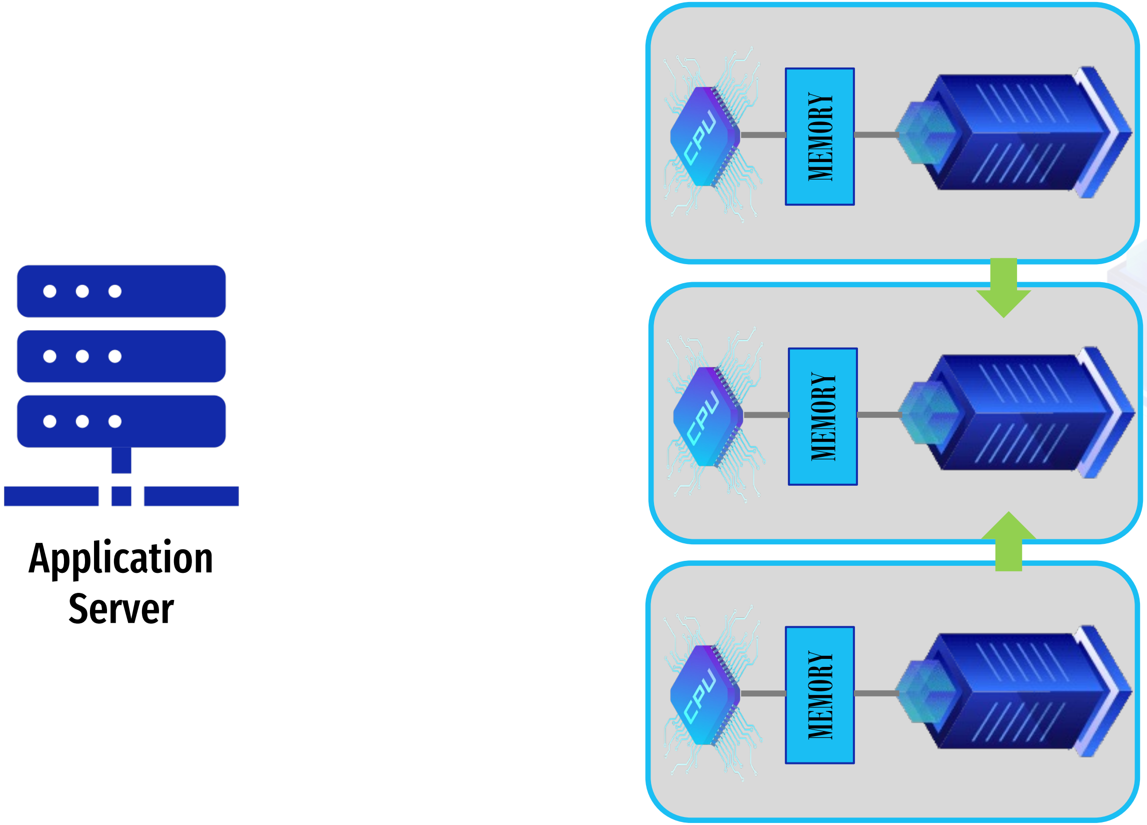
ID =200

Application Server

ID	…..
1	……
………	………
199	……

ID	…..
200	……
………	………
400	……

Shared Nothing Example



ID	…..
1	……
………	………
199	……

ID	…..
151	……
………	………
250	……

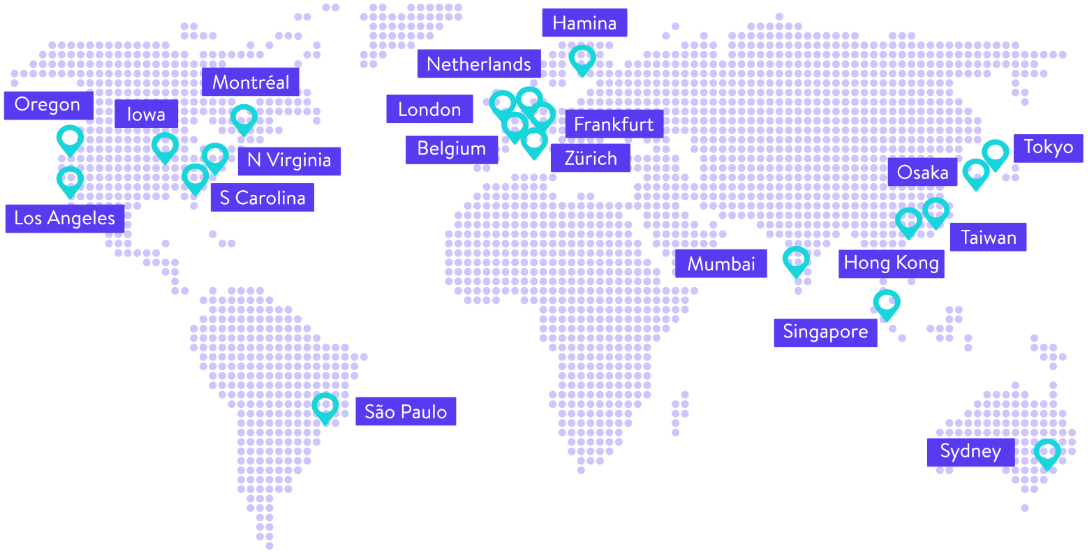
ID	…..
200	……
………	………
400	……

Definition of Distributed Databases

A distributed database management system (DDBMS) is the software that manages a collection of multiple, logically interrelated databases located at the nodes of a distributed system and provides an access mechanism that makes this distribution transparent to the users. A distributed DBMS is logically integrated but physically distributed.

Example of a Distributed DBMSs

“**Google** runs on hundreds of thousands of servers—by one estimate, in excess of 450,000—racked up in thousands of clusters in dozens of data centers around the world. It has data centers in Dublin, Ireland; in Virginia; and in California, where it just acquired the million-square-foot headquarters it had been leasing. Google runs more than 2,000,000 servers in



History of Distributed DBMS

Peer-to-peer systems (P2P)

Client/server

Cloud computing

- Infrastructure-as-a-service (IaaS)
- Platform-as-a-service (PaaS)
- Software as- a-service (SaaS)
- * Database-as-a-service (DBaaS)

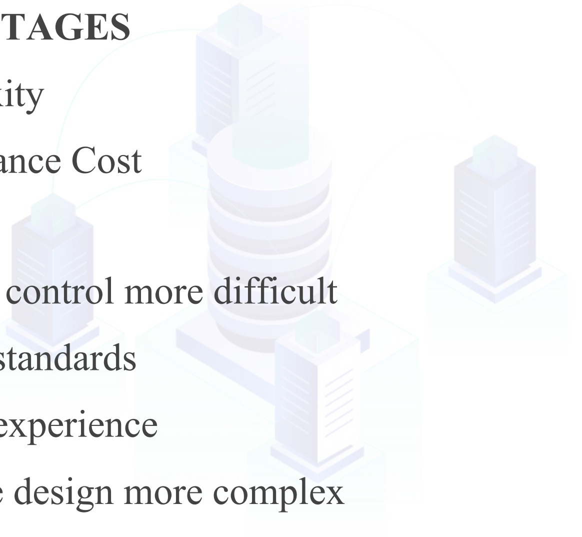
Advantages and Disadvantages of DDBMSs

ADVANTAGES

- ✓ Reflects organizational structure
- ✓ Improved shareability and local autonomy
- ✓ Improved availability
- ✓ Improved reliability
- ✓ Improved performance
- ✓ Economics
- ✓ Modular growth
- ✓ Integration
- ✓ Remaining competitive

DISADVANTAGES

- ✓ Complexity
- ✓ Maintenance Cost
- ✓ Security
- ✓ Integrity control more difficult
- ✓ Lack of standards
- ✓ Lack of experience
- ✓ Database design more complex



Homogeneous and Heterogeneous DDBMSs

Homogenous Nodes

All sites use the same DBMS product

Much easier to design and manage.

Heterogenous Nodes

Sites may run different DBMS products.

Translations are required to allow communication between different DBMSs.

Data may be required from another site that may have:

✓ different hardware;

✓ different DBMS products;

✓ Both different hardware and different DBMS products.

Functions of a DDBMS

Use the same concept in single node DBMSs to support transaction processing and query execution in distributed environments.

Query optimization & Execution, Concurrency Control, Logging & Recovery, Communication services & Security control

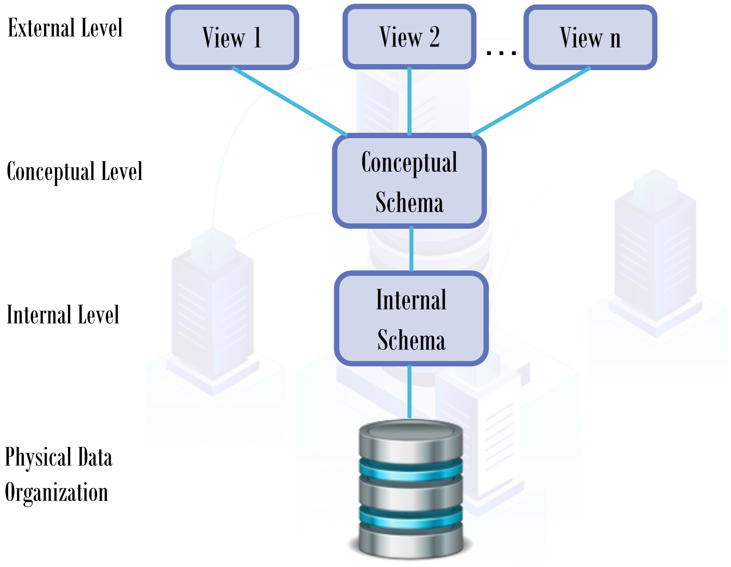
In addition, we expect a DDBMS to have the following functionality:

- ✓ Extend communication services to provide access to remote sites and allow to transfer of queries and data among the sites using a network;
- ✓ Extend system catalog to store data distribution details;
- ✓ Distribute query processing, including query optimization and remote data access;
- ✓ Extend security control to maintain appropriate authorization/access privileges to the distributed data;
- ✓ Extend concurrency control to maintain consistency of distributed and possibly replicated data;
- ✓ Extend recovery services to take account of failures of individual sites and the failures of

Architectures of a DDBMS

Due to diversity of distributed DBMSs, there is no accepted architecture equivalent to ANSI/SPARC 3-level architecture.

However, it may be useful to present one possible reference architecture that addresses data distribution.



ANSI/SPARC 3-level architecture

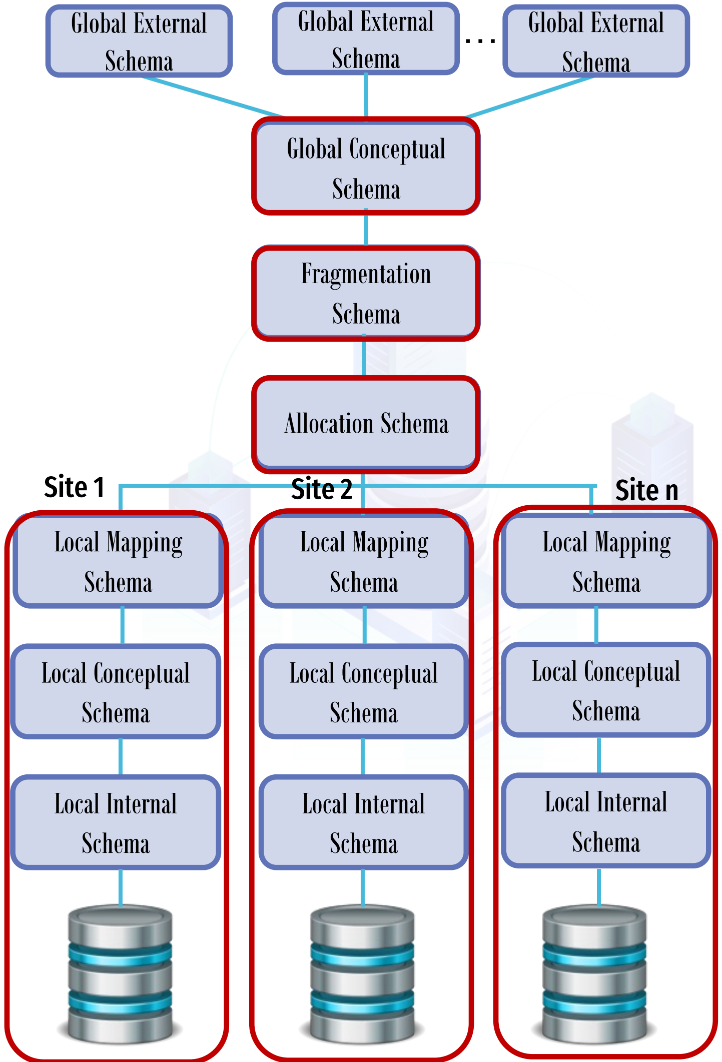
Architectures of a DDBMS

Due to diversity of distributed DBMSs, there is no accepted architecture equivalent to ANSI/SPARC 3-level architecture.

However, it may be useful to present one possible reference architecture that addresses data distribution.

Reference Architecture for a DDBMS consists of the following schemas:

- ✓ Set of global external schemas.
- ✓ Global conceptual schema (GCS).
- ✓ Fragmentation schema and allocation schema.
- ✓ Set of schemas for each local DBMS conforming to 3-level ANSI/SPARC.



Summary

Major different between DDBMS and Centralized DBMS.

Various type of system architectures: Shared Everything, Shared Memory, Shared Disk and Shared Nothing

Define DDBMS: the software that manages a collection of multiple, logically interrelated databases located at the nodes of a distributed system.

Major Advantages and Disadvantages of DDBMSs.

Homogeneous and Heterogeneous DDBMSs.

Owing to the diversity of distributed DBMSs, it is much more difficult to present an equivalent architecture that is generally applicable.