Data Engineering

Lecture 2

Lecture Agenda

- What is big-data?
- History of cloud computing
- Types of clouds
- Data-engineering
- Python libraries for data science and cloud computing

What is Big Data?

Big data is a collection of large datasets that cannot be processed using traditional computing techniques. **It is not a single technique or a tool**.

Benefits

- Using the information kept in the social network like Facebook, twitter, etc.
- Using the information in the social media like preferences and product perception of their consumers.

Sources of Big Data

Social networking sites: Facebook, Google, LinkedIn

E-commerce site: Amazon, Flipkart, Alibaba

Weather Station: All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather

Telecom company: Airtel, Vodafone

Share Market: Stock exchange across the world

3V's of Big Data

Velocity: The data is increasing at a very fast rate.

It is estimated that the volume of data will double in every 2 years.

Variety: Data is structured as well as unstructured.

structured – predefined schema

unstructured – each document has different properties

Volume: The amount of data which we deal with is of very large size of **Peta** bytes.

(1 Peta = 1024 TB)

Handling Big Data

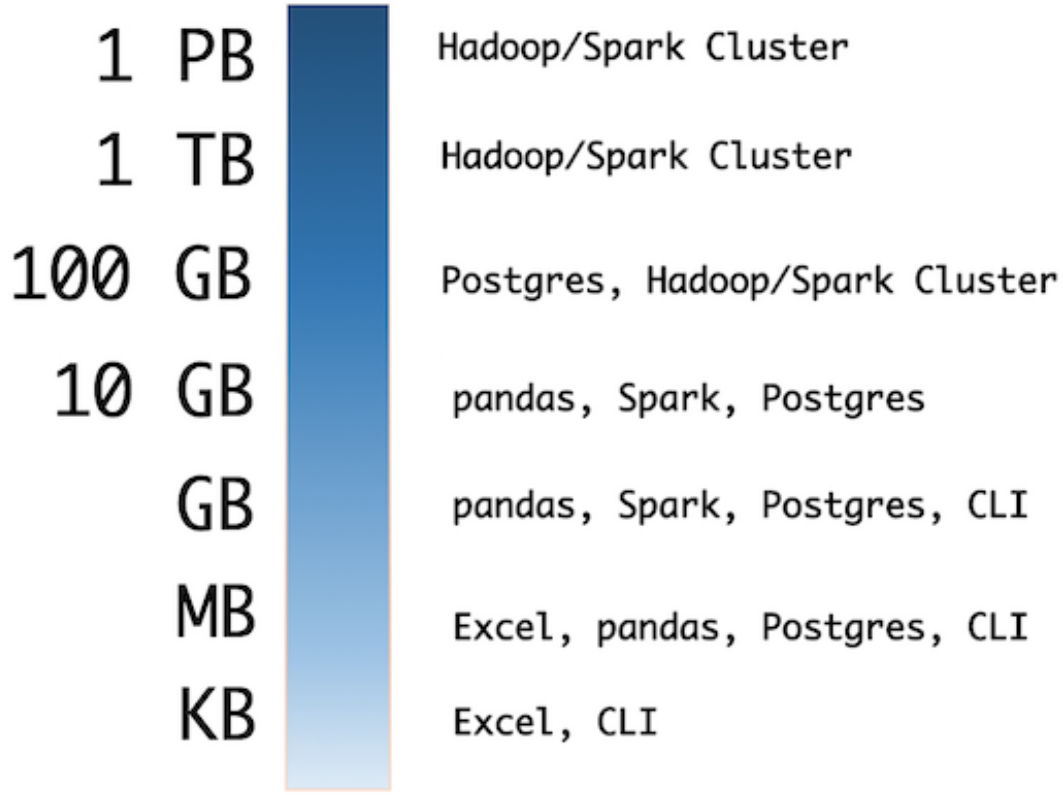
Excel

Pandas – Lecture 3

Hadoop – Lectures 4,5

Spark – Lectures 6,7

Streaming – Lectures 10,11



What is Cloud?

Datacenter hardware and software that the vendors use to offer the computing resources and services.

Cloud Computing

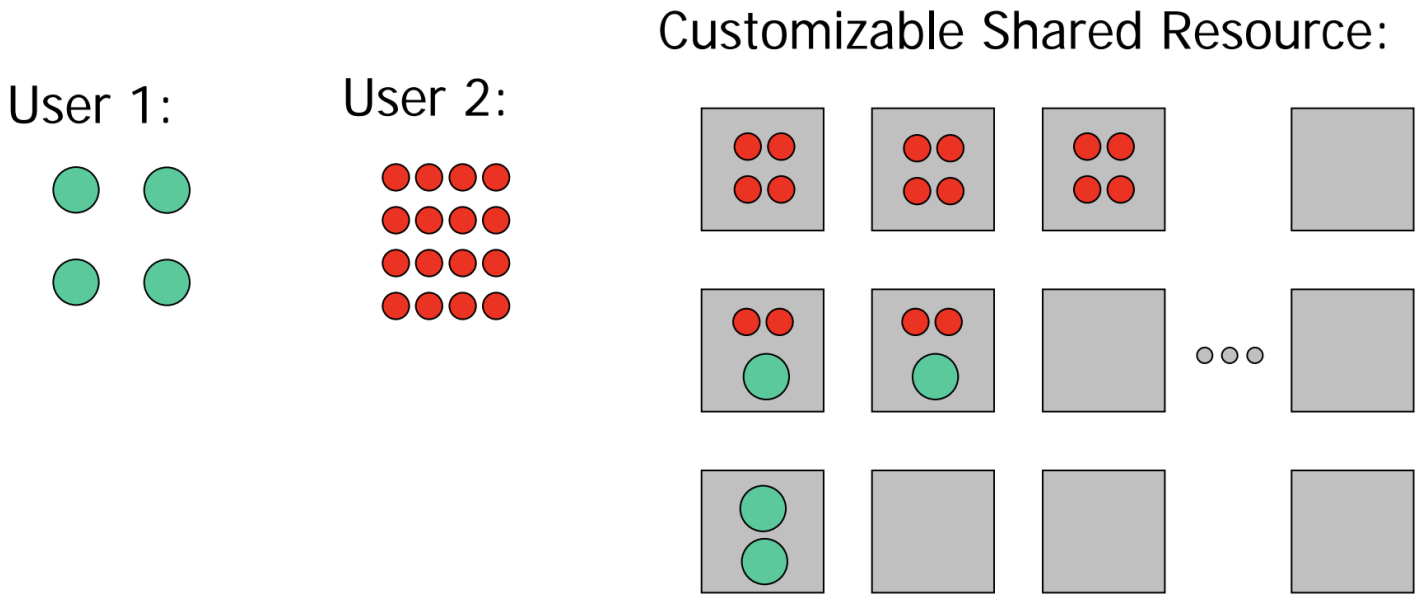
- Represents both the cloud & the provided services
- Why call it “cloud computing”?
  - Some say because the computing happens out there "in the clouds” ☺
  - Wikipedia: "the term derives from the fact that most technology diagrams depict the Internet or IP availability by using a drawing of a cloud."

Cloud Computing Services

- **S**oftware **a**s **a** **S**ervice (SaaS) – applications though the browser
- **P**latform **a**s **a** **S**ervice (PaaS) - Delivery of a computing platform for custom software development as a service
- **I**nfrastructure **a**s **a** **S**ervice (IaaS) - Deliver of computer infrastructure as a service
- And more ..

Resource Sharing

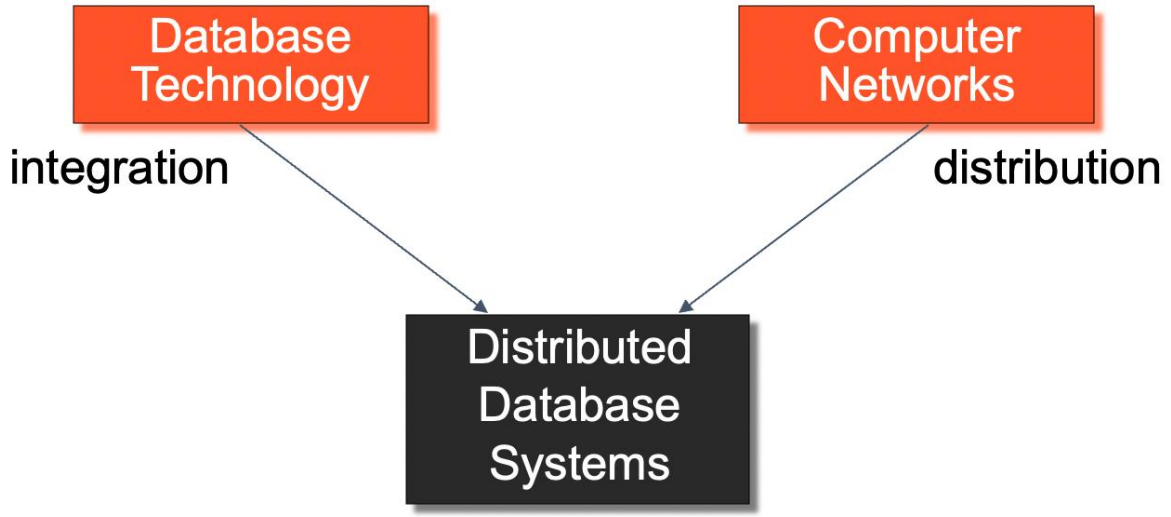
Offering computing resources as a service or utility through **virtualization**



Distributed Computing

Multiple processing elements that are interconnected by a computer network and that cooperate in performing their assigned tasks.

What is being distributed? Processing logic, Function, Data, Control

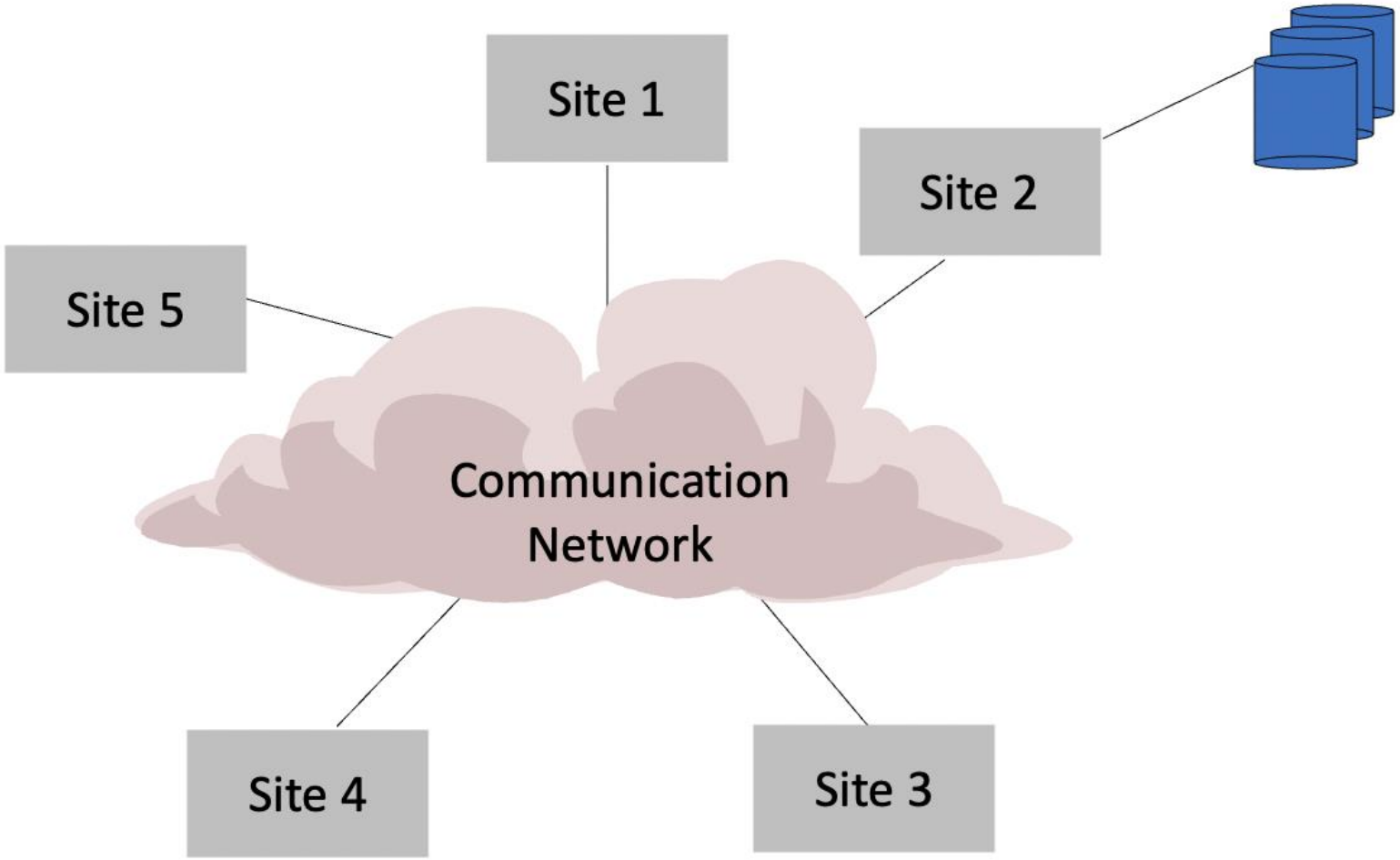


Distributed Database System

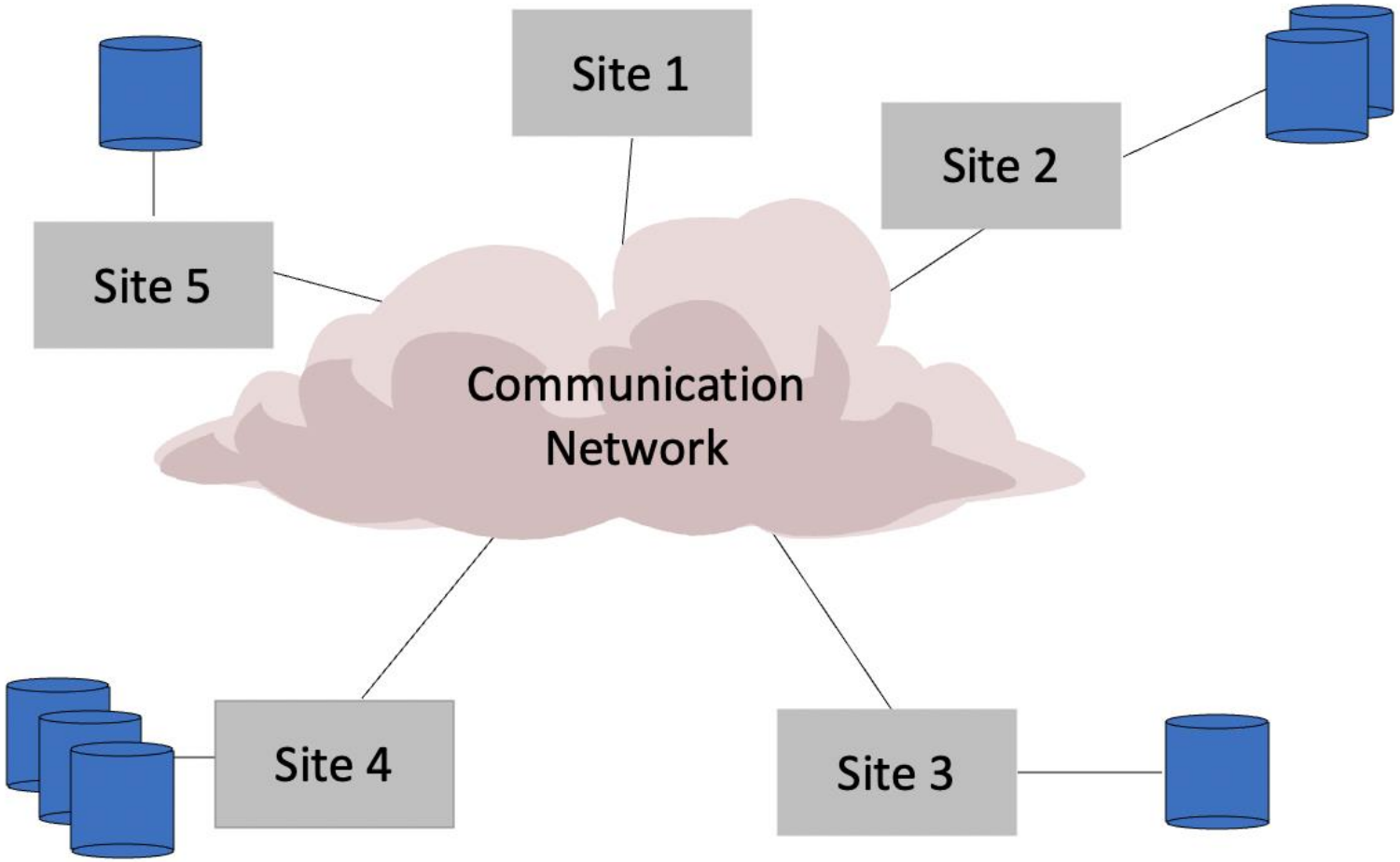
A **distributed database** (DDB) is a collection of multiple, logically interrelated databases distributed over a computer network.

A distributed database management system (D–DBMS) is the software that manages the DDB and provides an access mechanism that makes this distribution transparent to the users.

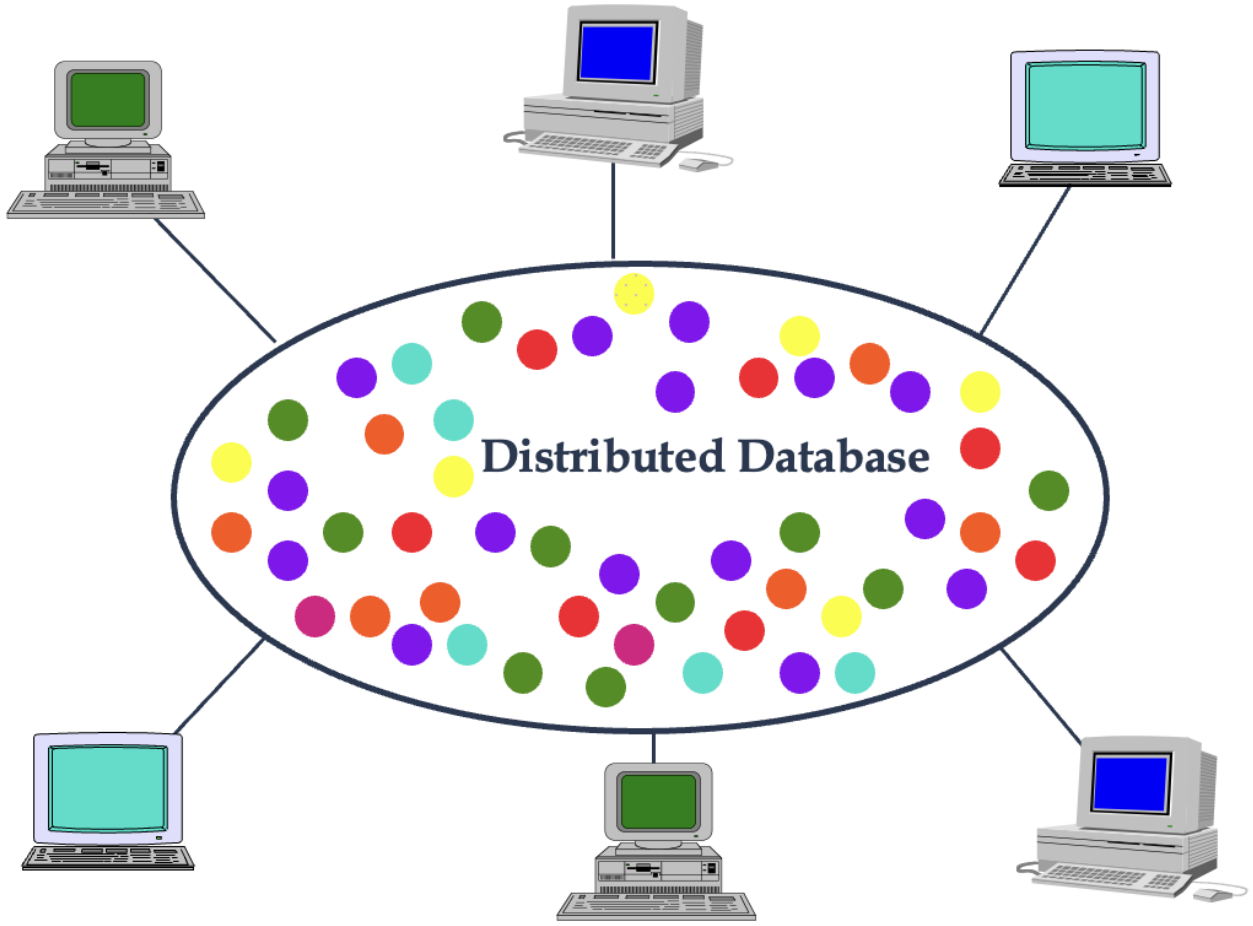
Centralized DBMS



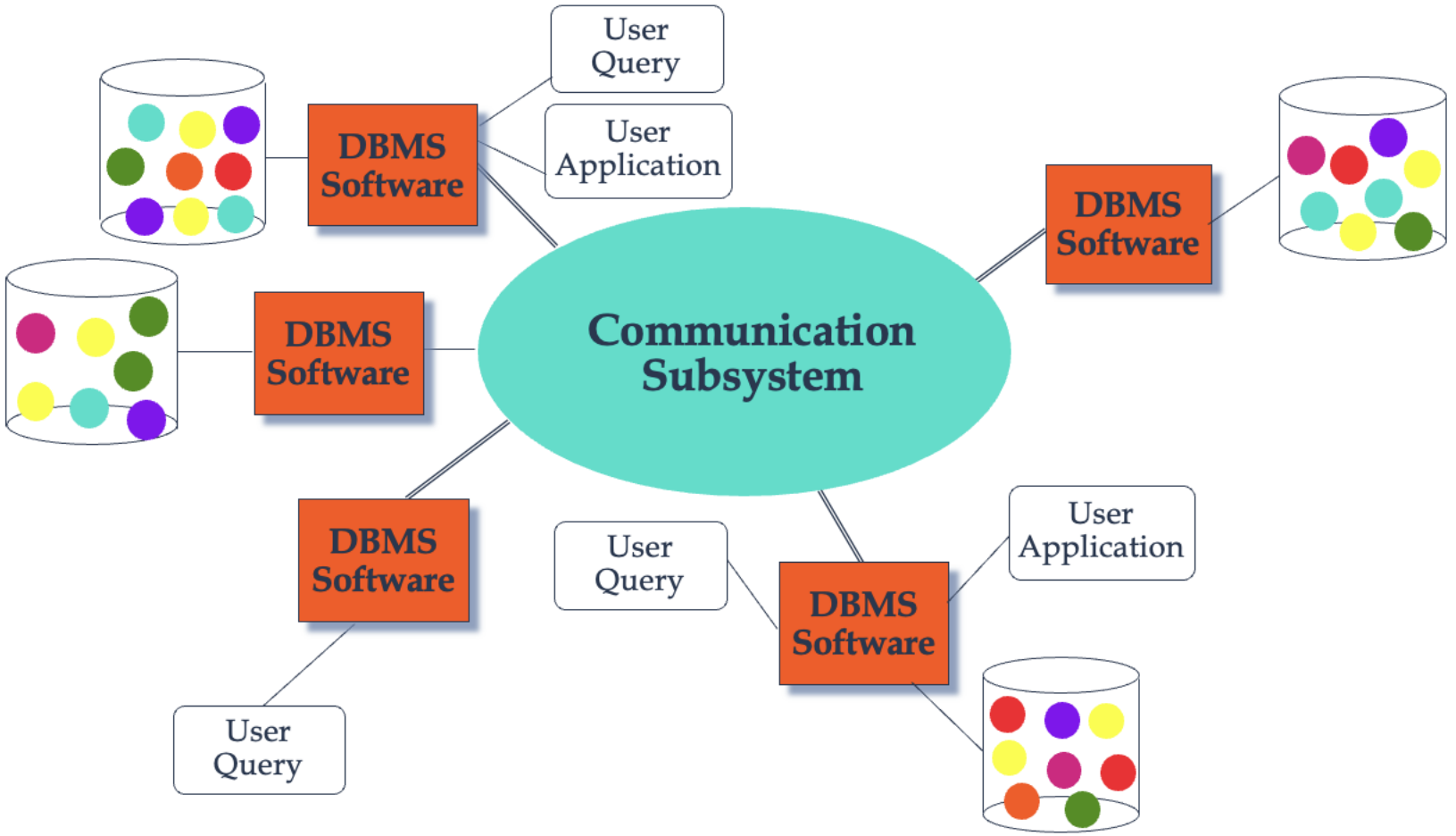
Distributed DBMS



Distributed Database – User View



Distributed Database – Reality



Distributed DBMS Issues

- **Distributed Database Design**
  - How to distribute the database
  - Replicated & non-replicated database distribution
  - A related problem in directory management
- **Query Processing**
  - Convert user transactions to data manipulation instructions
  - Optimization problem
    - min{cost = data transmission + local processing}

Distributed DBMS Issues

- **Concurrency Control**
  - Synchronization of concurrent accesses
  - Consistency and isolation of transactions' effects
  - Deadlock management
- **Reliability**
  - How to make the system resilient to failures
  - Atomicity and durability

Parallel Computation

Asynchronous Programming

Process vs Threads

A **process** is an instance of a running program.

Process may contain one or more **threads**, but a **thread** cannot contain a **process**.

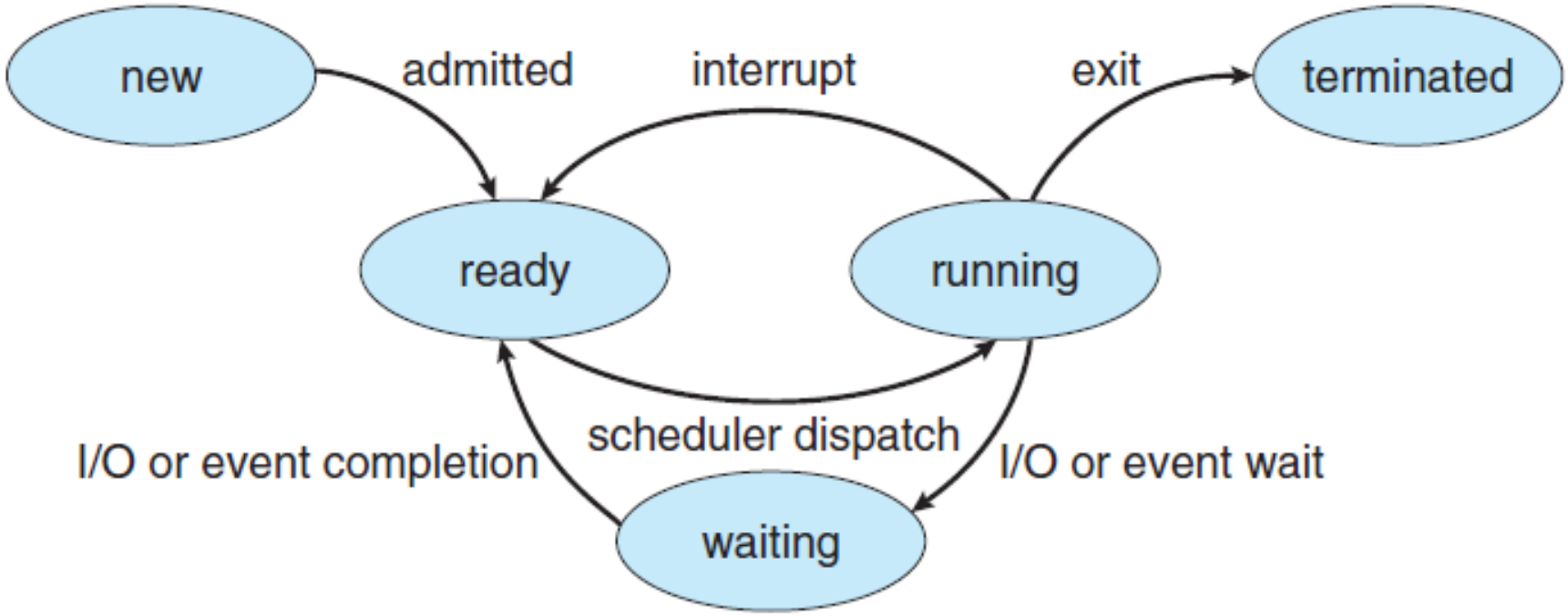
Process has a self-contained execution environment. It has its own memory space.

Process don't share its memory

A **thread** is made of and exist within a **process**, every **process** has at least one **thread**.

Multiple **threads** in a **process** share resources.

Process Life Cycle - PLC

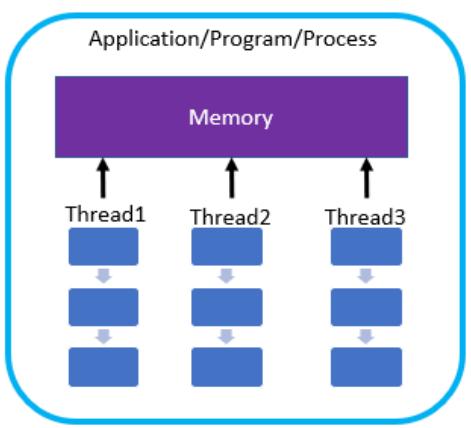


Parallel Computers

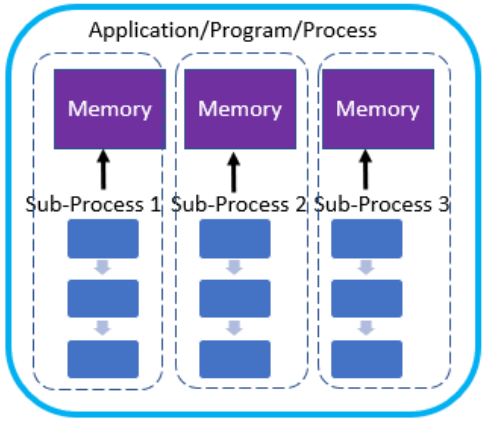
Multiprocessor/multicore: several processors work on data stored in shared memory

Cluster: several processor/memory units work together by exchanging data over a network

Co-processor: a general-purpose processor delegates specific tasks to a special-purpose processor (GPU)



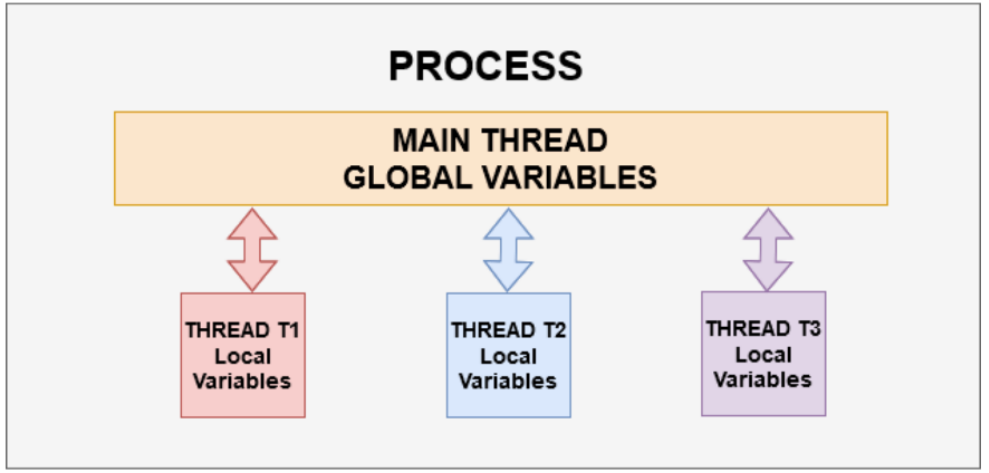
Multi-Threading



Multi-processing

Parallel Computation

Threading allows you to have different parts of your process run **concurrently**. For example, A web-browser could be a process, an application running multiple cameras simultaneously could be a process.

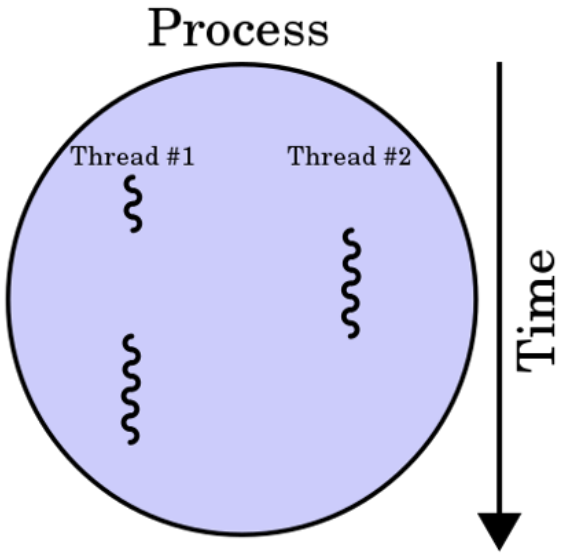


multiple threads work together to achieve a common goal

Parallel Computation

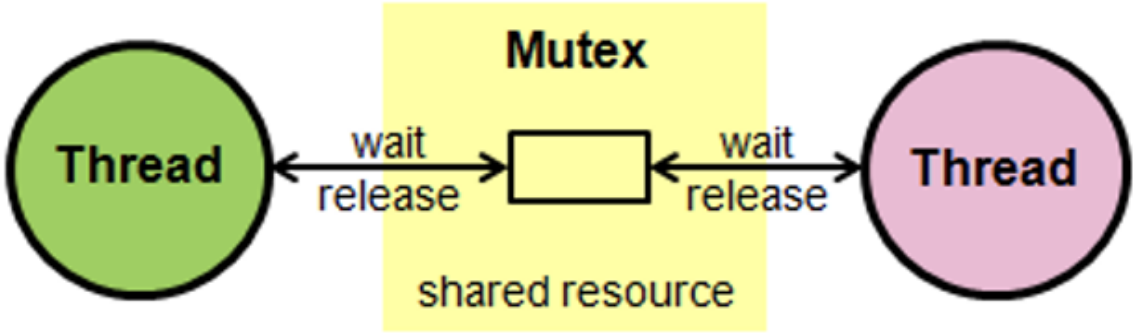
Multi-threading allows the program to speed up the execution provided that it has multiple CPUs. It also lets you perform other tasks while the I/O operations are being performed.

Threads within the same process can share the memory and resources of the main thread.



Synchronizing Threads

The threading module provided with Python includes a simple-to-implement **locking mechanism** that allows you to synchronize threads. A new lock is created by calling the ***Lock()*** method, which returns the new lock. It is also called “Mutex”.



Synchronizing Threads

The *acquire(blocking)* method of the new lock object is used to force the threads to run **synchronously**.

If *blocking* is set to 0, the thread returns immediately with a 0 value if the lock cannot be acquired and with a 1 if the lock was acquired. If blocking is set to 1, the thread blocks and waits for the lock to be released.

The ***release()*** method of the new lock object is used to release the lock when it is no longer required.

Process Hierarchies

- Parent creates a child process, child processes can create its own process
- Forms a hierarchy
  - UNIX calls this a "process group"
- Windows has no concept of process hierarchy
  - all processes are created equal