INTRODUCING TECHNOLOGIES FOR HANDLING BIG DATA

CONTENTS

- Distributed and parallel Computing for Big Data
- Introducing Hadoop
- Cloud Computing and Big Data
- In-Memory Computing Technology for Big Data

- Among the technologies that are used to handle, process and analyse big data the most popular and effective innovations have been in the field of distributed and parallel processing, Hadoop, in memory computations, big data cloud.
- Most popular:- *Hadoop*
- Organisations use it to extract maximum output from normal data usage practices at a rapid pace.
- Cloud computing helps companies to save cost and better manage resources.

BIG DATA

- Big Data can’t be handled by traditional data storage and processing systems.
- For handling such type of data,

Distributed and Parallel Technologies are more suitable.

DISTRIBUTED AND PARALLEL COMPUTING FOR BIG DATA

- Distributed Computing

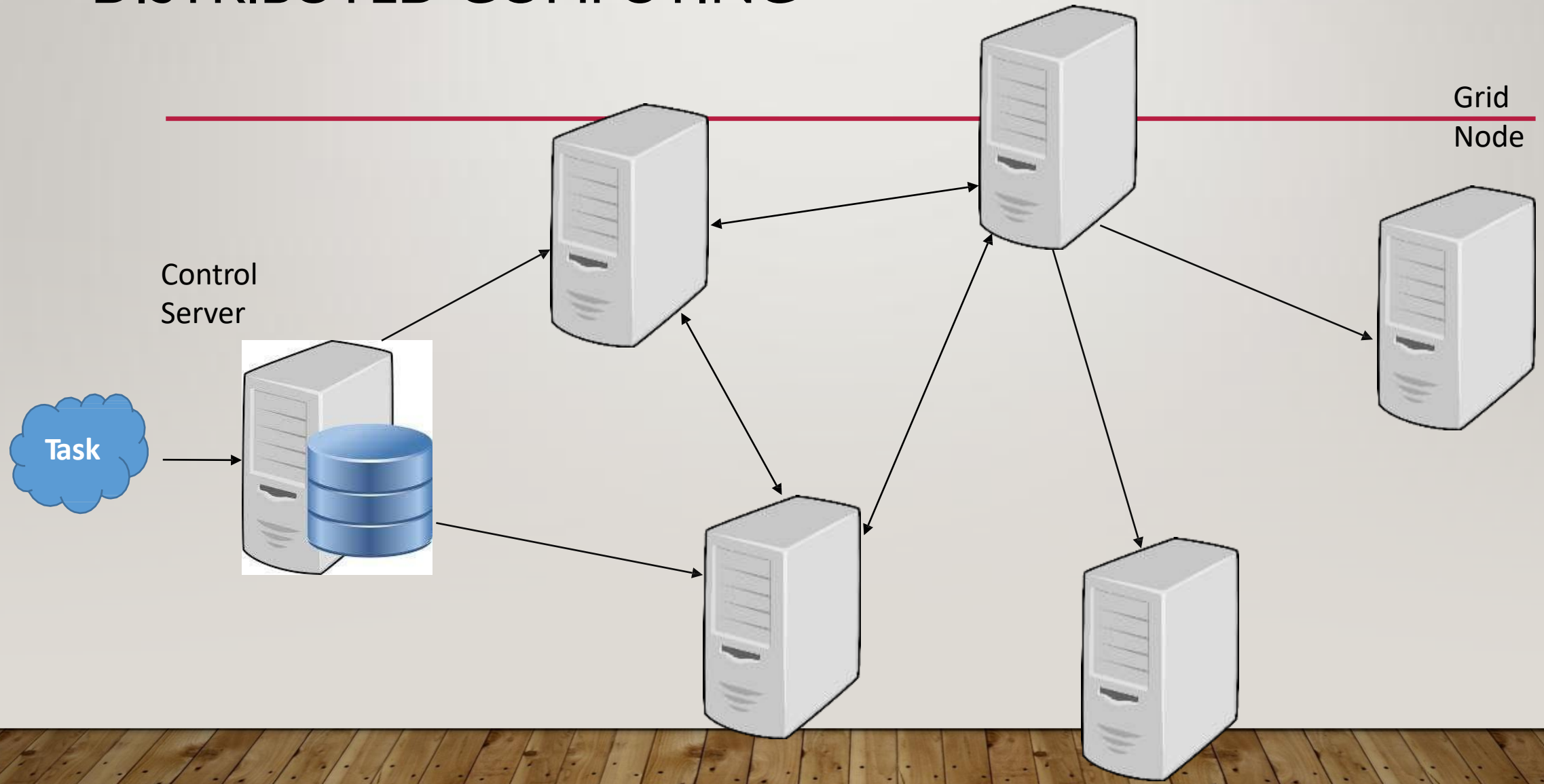
Multiple computing resources are connected in a network and computing tasks are distributed across these resources.

☺ Increases the Speed

☺ Increases the Efficiency

☺ more suitable to process huge amount of data in a limited time

DISTRIBUTED COMPUTING



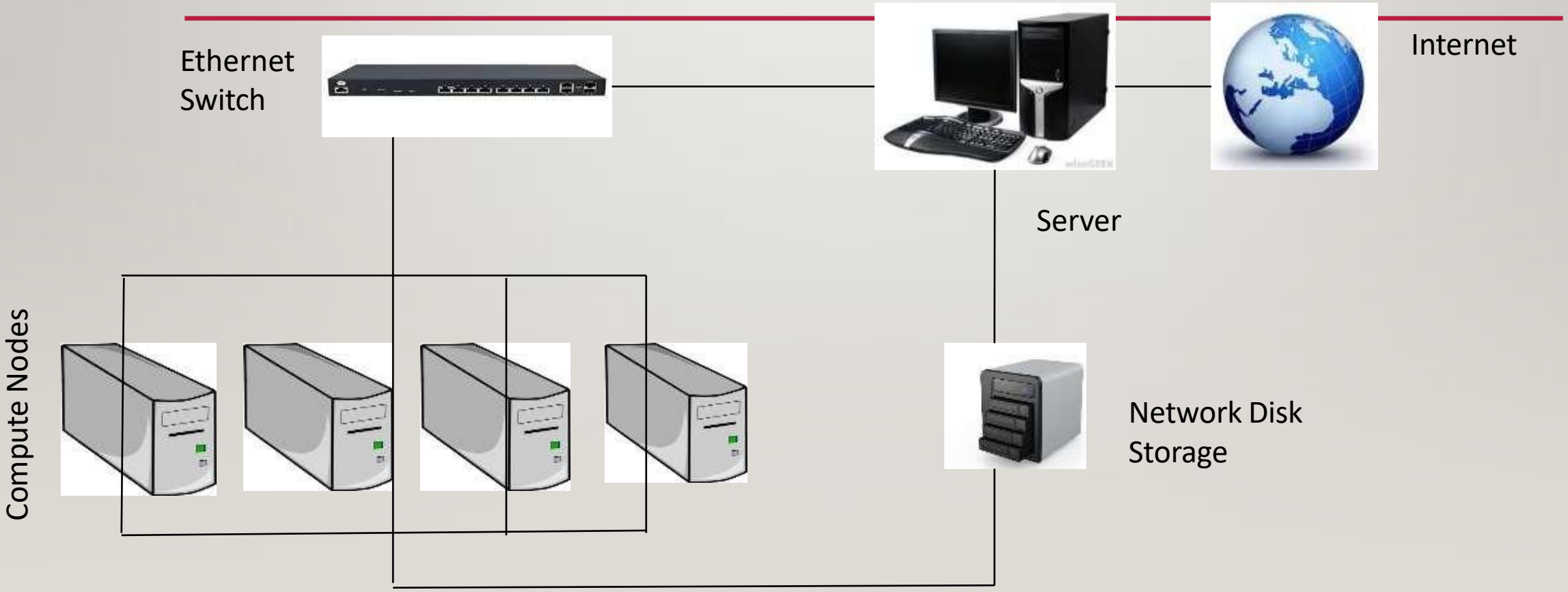
DISTRIBUTED AND PARALLEL COMPUTING FOR BIG DATA

Parallel Computing

- Also improves the processing capability of a computer system by adding additional computational resources to it.
- Divide complex computations into subtasks, handled individually by processing units, running in parallel.

Concept – processing capability will increase with the increase in the level of parallelism.

PARALLEL COMPUTING



BIG DATA PROCESSING TECHNIQUES

- With the increase in data, forcing organizations to adopt a data analysis strategy that can be used for analysing the entire data in a very short time.

Done by Powerful h/w components and new s/w programs.

The procedure followed by the s/w applications are:

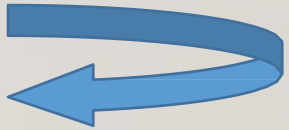
1) Break up the given task
2) Surveying the available resources
3) Assigning the subtask to the nodes

ISSUES IN THE SYSTEM

- Resources develop some technical problems and fail to respond
- virtualization.

Some processing and analytical tasks are delegated to other resources.

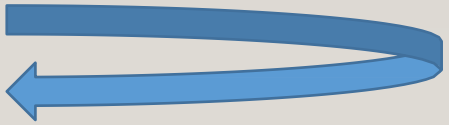
✓Latency : can be defined as the aggregate delay in the s/m bcoz of delays in the completion of individual tasks.



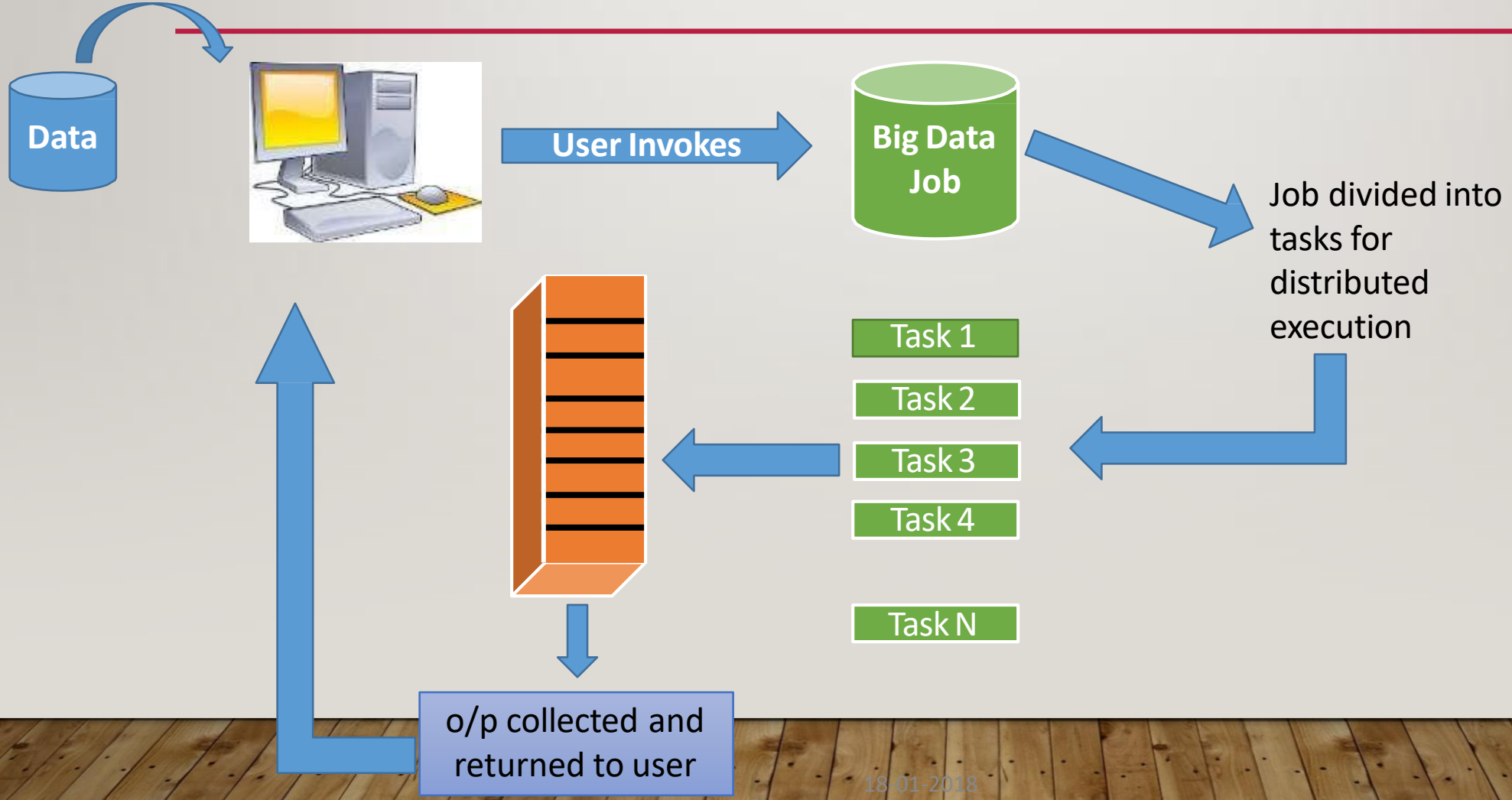
✓System delay

Also affects data management and communication

Affecting the productivity & profitability of an organization.



DISTRIBUTED COMPUTING TECHNIQUE FOR PROCESSING LARGE DATA



MERITS OF THE SYSTEM

☺ Scalability

The system with added scalability, can accommodate the growing amounts of data more efficiently and flexibly.

☺ Virtualization and Load Balancing Features

Load Balancing – The sharing of workload across various systems.

Virtualization – creates a virtual environment

h/w platform, storage device and OS

PARALLEL COMPUTING TECHNIQUES

1) Cluster or Grid Computing
- primarily used in Hadoop.
- based on a connection of multiple servers in a network (clusters)
- servers share the workload among them.
- overall cost may be very high.

2) Massively Parallel Processing (MPP)
- used in data warehouses.

PARALLEL COMPUTING TECHNIQUES

- Single machine working as a grid is used in the MPP platform.
- Capable of handling the storage, memory and computing activities.
- Software written specifically for MPP platform is used for optimization.
- MPP platforms, EMC Greenplum, ParAccel , suited for high-value use cases.

3) High Performance Computing (HPC)

- Offer high performance and scalability by using IMC.

PARALLEL COMPUTING TECHNIQUES

- Suitable for processing floating point data at high speeds.
- Used in research and business organization where the result is more valuable than the cost or where strategic importance of project is of high priority.

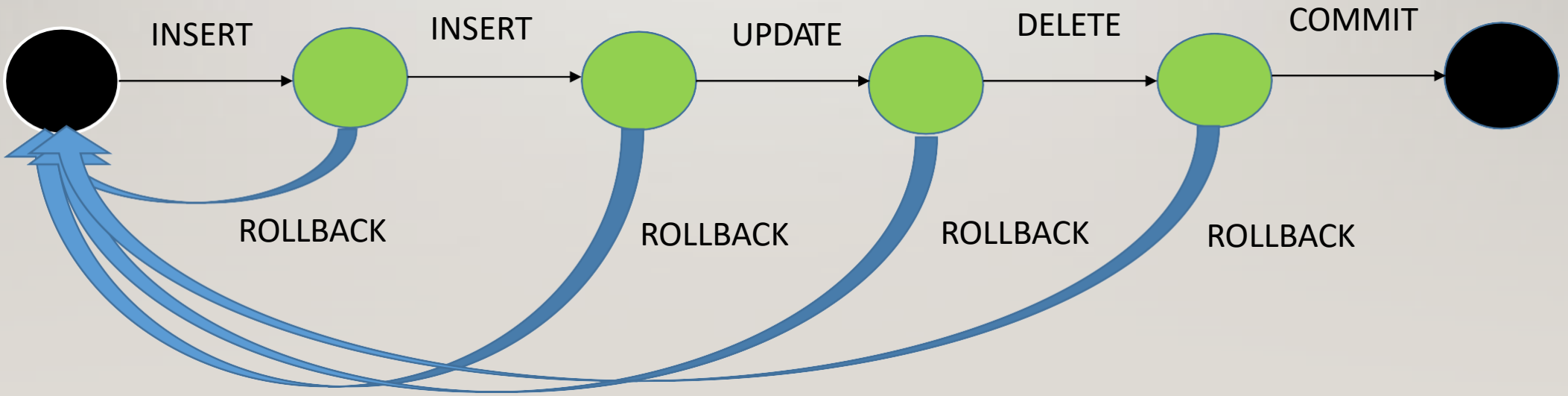
DIFFERENCE B/W DISTRIBUTED AND PARALLEL SYSTEMS

Distributed System	Parallel System
• Independent autonomous system connected in a n/w for accomplishing specific task.	• Computer s/m with several processing units attached to it.
• Coordination is possible b/w connected computers that have their own m/y and CPU	• Common shared m/y can be directly accessed by every processing unit in a n/w.
• Loose coupling of computers connected in a n/w, providing access to data and remotely located resources.	• Tight coupling of processing resources that are used for solving a single, complex problem.

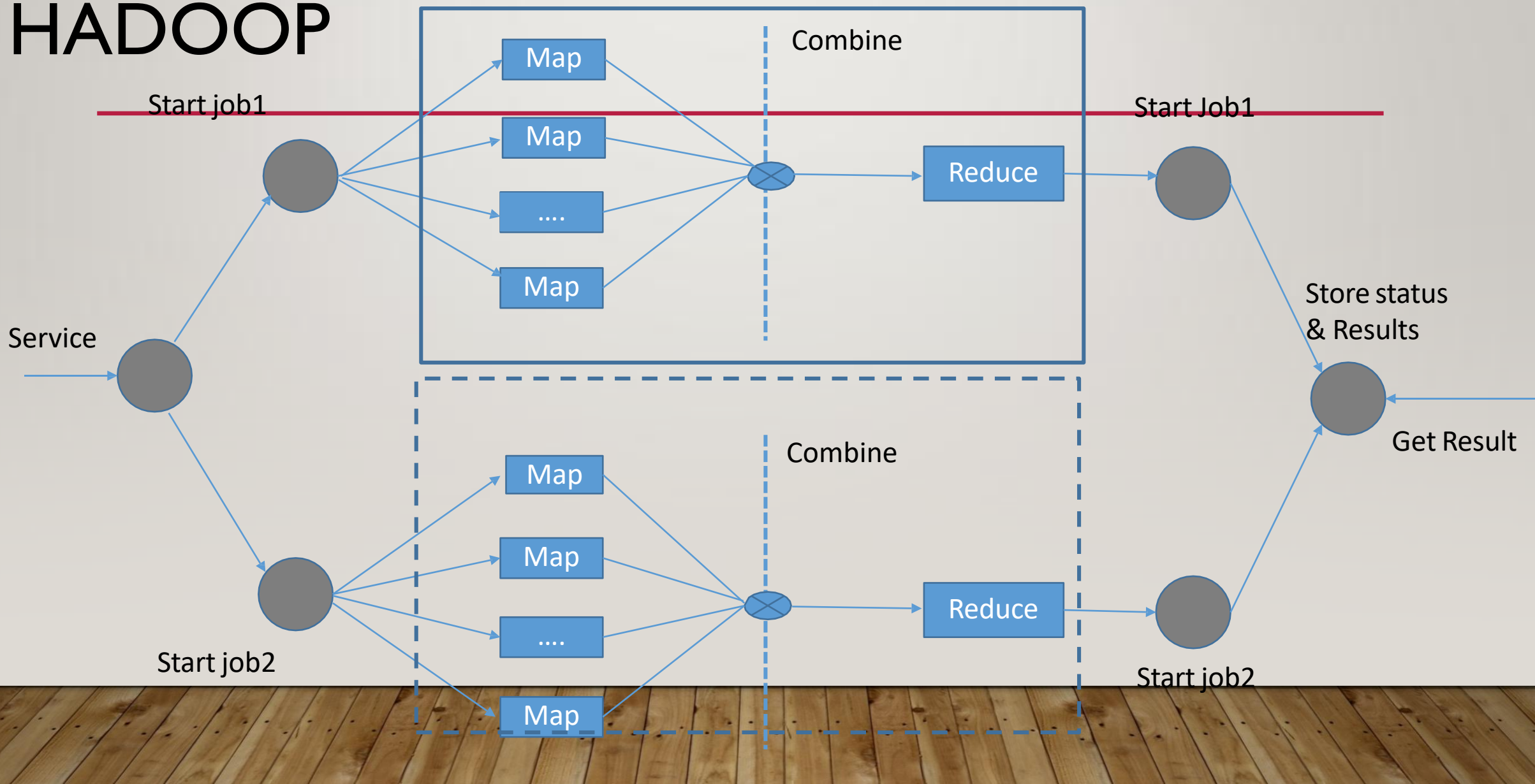
INTRODUCING HADOOP

- Hadoop is a distributed system like distributed database.
- Hadoop is a ‘software library’ that allows its users to process large datasets across distributed clusters of computers, thereby enabling them to gather, store and analyse huge sets of data.
- It provides various tools and technologies, collectively termed as the Hadoop Ecosystem.

COMPUTING MODEL OF DISTRIBUTED DATABASE

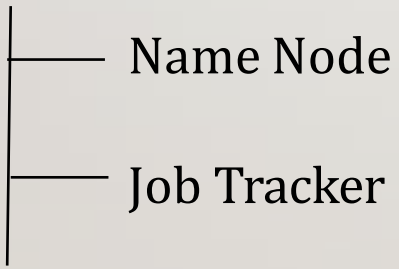


COMPUTING MODEL OF HADOOP

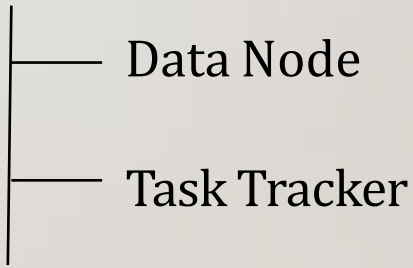


HADOOP MULTINODE CLUSTER ARCHITECTURE

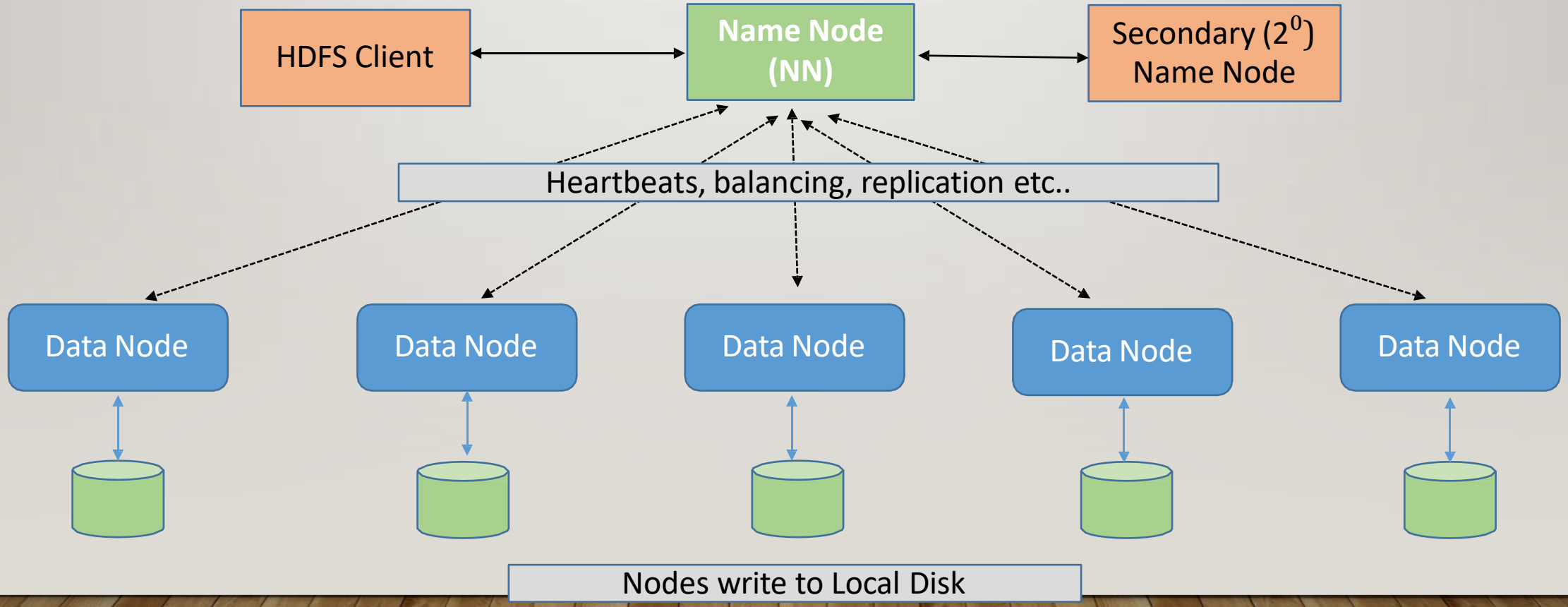
- Hadoop cluster consist of single Master Node and a multiple Worker Nodes.
- Master Node



Worker Node



HADOOP MULTINODE CLUSTER ARCHITECTURE



- To process the data, Job Tracker assigns tasks to the Task Tracker.
- If a data node cluster goes down while processing is going on, then the NN should know that some data node is down in the cluster, otherwise it can't continue processing.
- Each DN sends a "Heart Beat Signal" after every few minutes to make NN aware of active/inactive status of DNs. – Heartbeat Mechanism.

HDFS AND MAPREDUCE

- HDFS – used for storage.

-MapReduce – used for processing.

Hadoop Distributed File System (HDFS)

- fault tolerant storage s/m.
- Large size files from terabytes to petabytes.
- attains reliability by replicating the data over multiple hosts.
- The default replication value is 3.

HDFS AND MAPREDUCE

- File in HDFS is split into large block size of 64 MB.
- Each block is independently replicated at multiple data nodes.

MapReduce

- Parallel processing frame work.
- Helps developers to write programs to process large volumes of unstructured data.

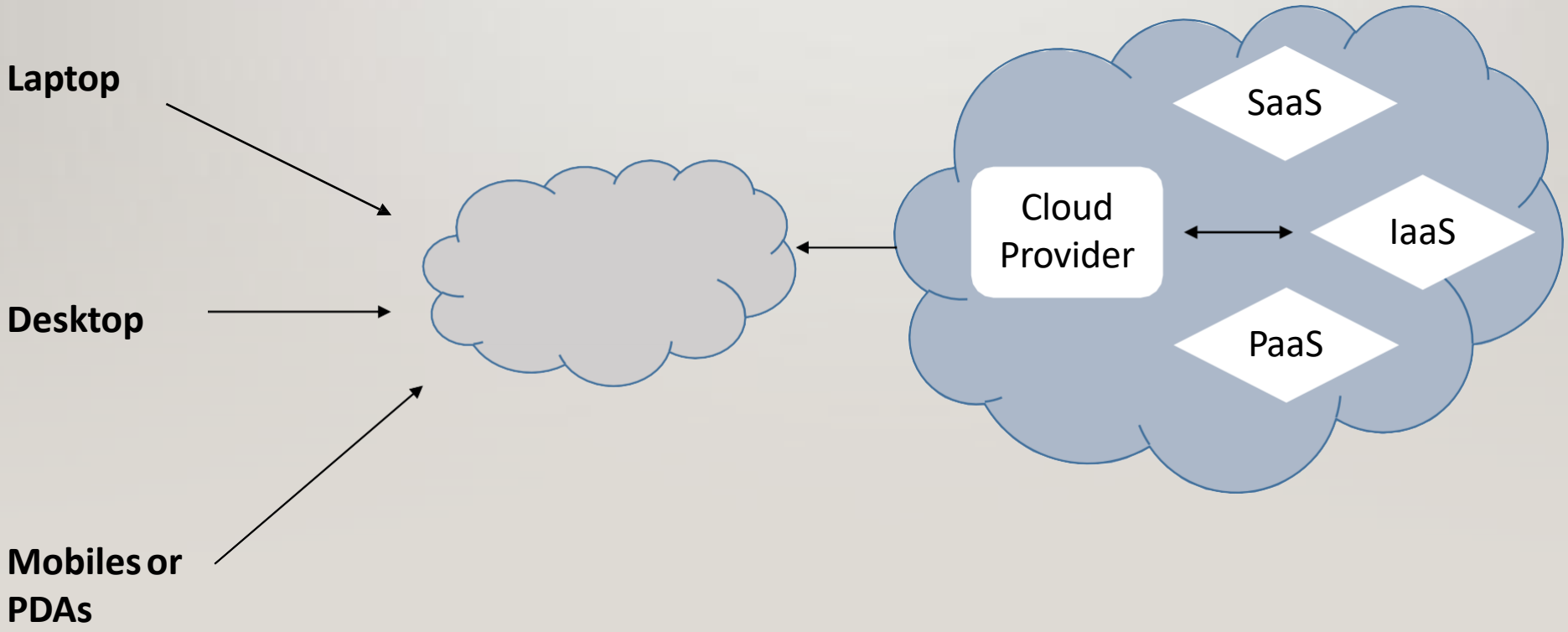
- Mapper function
- Reducer function

CLOUD COMPUTING AND BIG DATA

Cloud Computing is the delivery of computing services—servers, storage, databases, networking, software, analytics and more—over the Internet ("the cloud").

Companies offering these computing services are called **cloud providers** and typically charge for cloud computing services based on usage, similar to how you are billed for water or electricity at home.

CLOUD COMPUTING AND BIG DATA



FEATURES OF CLOUD COMPUTING

- **Scalability** – addition of new resources to an existing infrastructure.

- increase in the amount of data , requires organization to improve h/w components.
- The new h/w may not provide complete support to the s/w, that used to run properly on the earlier set of h/w.
- Solution to this problem is using cloud services - that employ the distributed computing technique to provide scalability.

- **Elasticity** – Hiring certain resources, as and when required, and paying for those resources.
- no extra payment is required for acquiring specific cloud services.
- A cloud does not require customers to declare their resource requirements in advance.

- **Resource Pooling -** multiple organizations, which use similar kinds of resources to carry out computing practices, have no need to individually hire all the resources.

- **Self Service –** cloud computing involves a simple user interface that helps customers to directly access the cloud services they want.

- **Low Cost** – cloud offers customized solutions, especially to organizations that cannot afford too much initial investment.

- cloud provides pay-us-you-use option, in which organizations need to sign for those resources only that are essential.

- **Fault Tolerance** – offering uninterrupted services to customers

CLOUD DEPLOYMENT MODELS

- Depending upon the architecture used in forming the n/w, services and applications used, and the target consumers, cloud services form various deployment models. They are,

  - Public Cloud
  - Private Cloud
  - Community Cloud
  - Hybrid Cloud

- **Public Cloud (End-User Level Cloud)**

- Owned and managed by a company than the one using it.
- Third party administrator.
- Eg : Verizon, Amazon Web Services, and Rackspace.
- The workload is categorized on the basis of service category, h/w customization is possible to provide optimized performance.
- The process of computing becomes very flexible and scalable through customized h/w resources.
- The primary concern with a public cloud include security and latency.

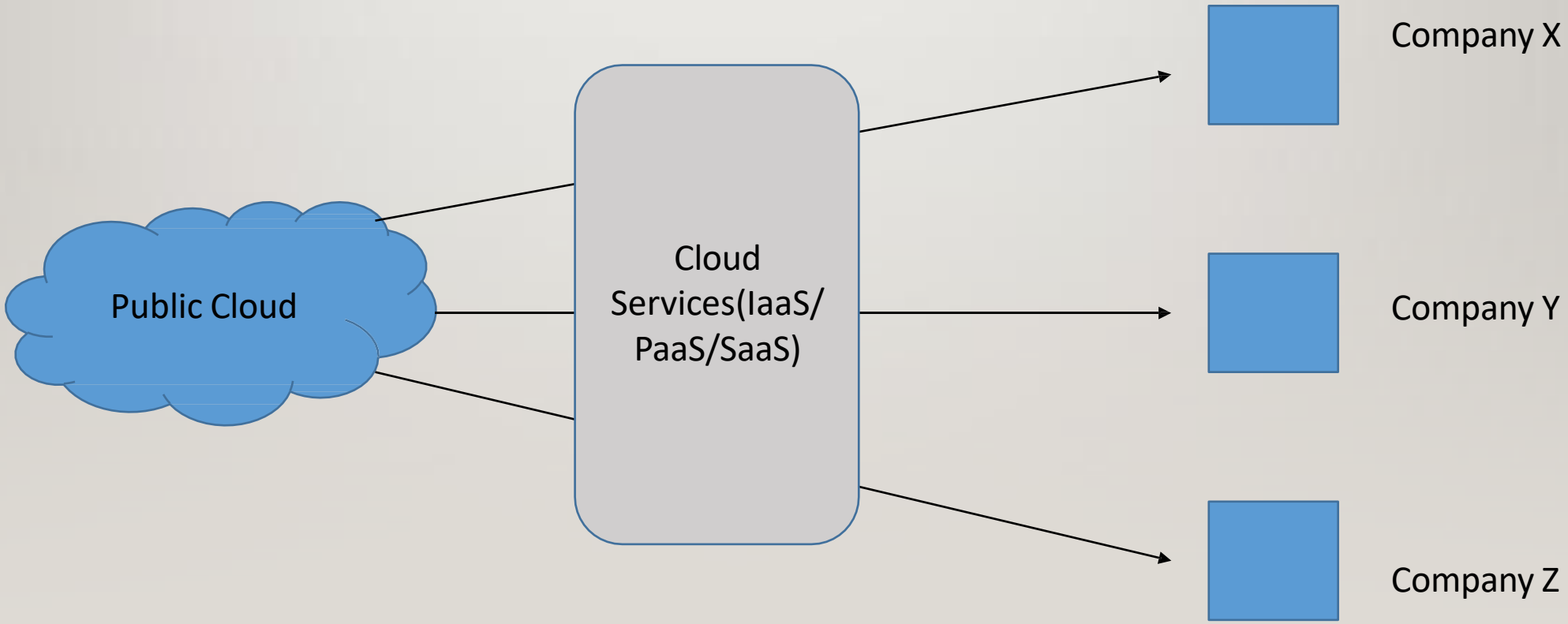


Fig : Level of Accessibility in a Public Cloud

- **Private Cloud (Enterprise Level Cloud)**

- Remains entirely in the ownership of the organization using it.
- Infrastructure is solely designed for a single organization.
- Can automate several processes and operations that require manual handling in a public cloud.
- Can also provide firewall protection to the cloud, solving latency and security concerns.
- A private cloud can be either on-premises or hosted externally.

on premises : service is exclusively used and hosted by a single organization.

hosted externally : used by a single organization and are not shared with other organizations.

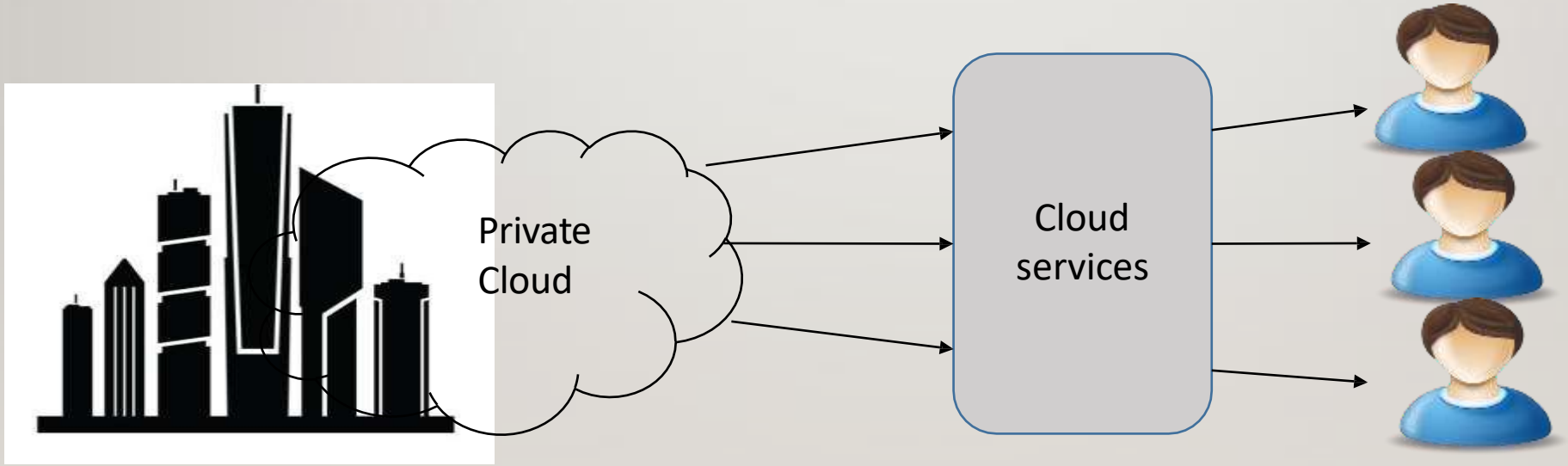


Fig: Level of Accessibility in a Private Cloud

- **Community Cloud**

-Type of cloud that is shared among various organizations with a common tie.

-Managed by third party cloud services.

-Available on or off premises.

Eg. In any state, the community cloud can provided so that almost all govt. organizations of that state can share the resources available on the cloud. Because of the sharing of resources on community cloud, the data of all citizens of that state can be easily managed by the govt. organizations.

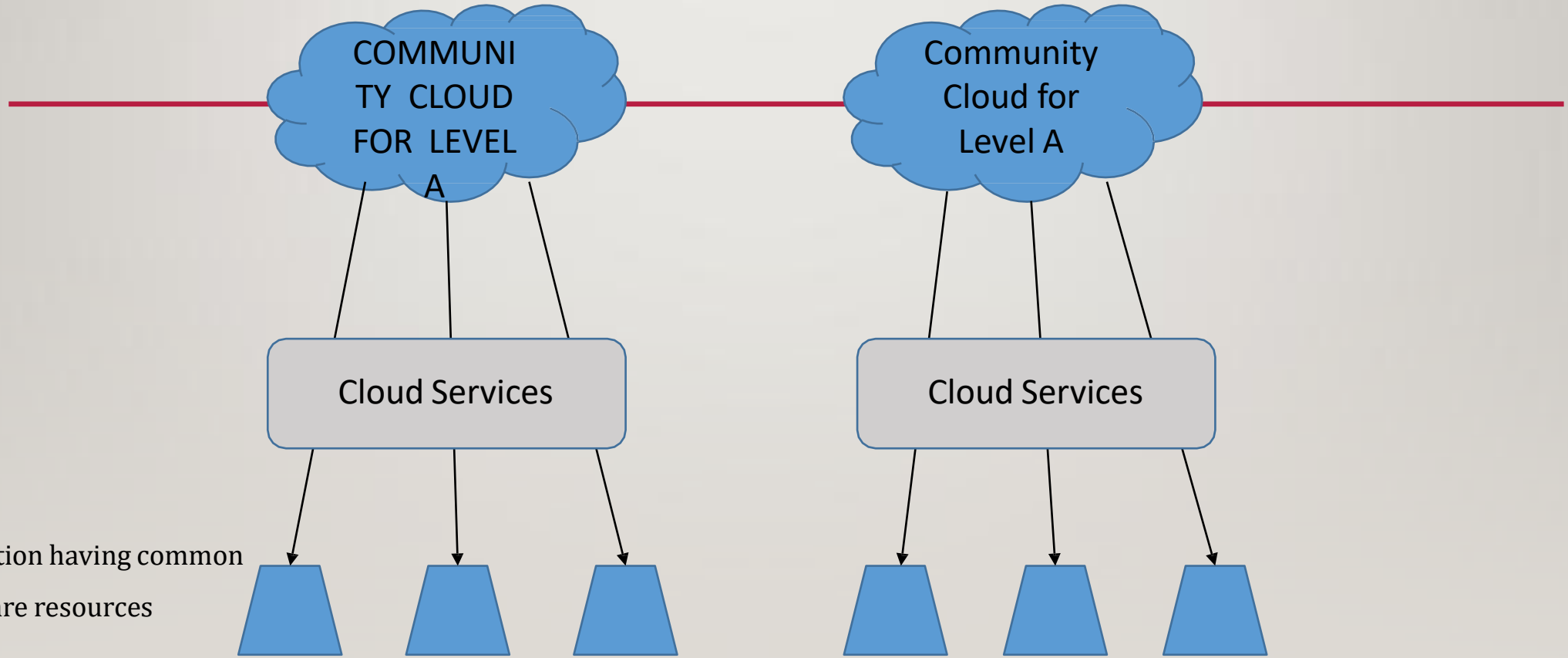


Fig : Level of Accessibility in Community Clouds

- **Hybrid Cloud**

-various internal or external service providers offer services to many organizations.

-In hybrid clouds, an organization can use both types of cloud, i.e. public and private together – situations such as **cloud bursting**.

- Organization uses its own computing infrastructure, high load requirement, access clouds.

The organization using the hybrid cloud can manage an internal private cloud for general use and migrate the entire or part of an application to the public cloud during the peak periods.

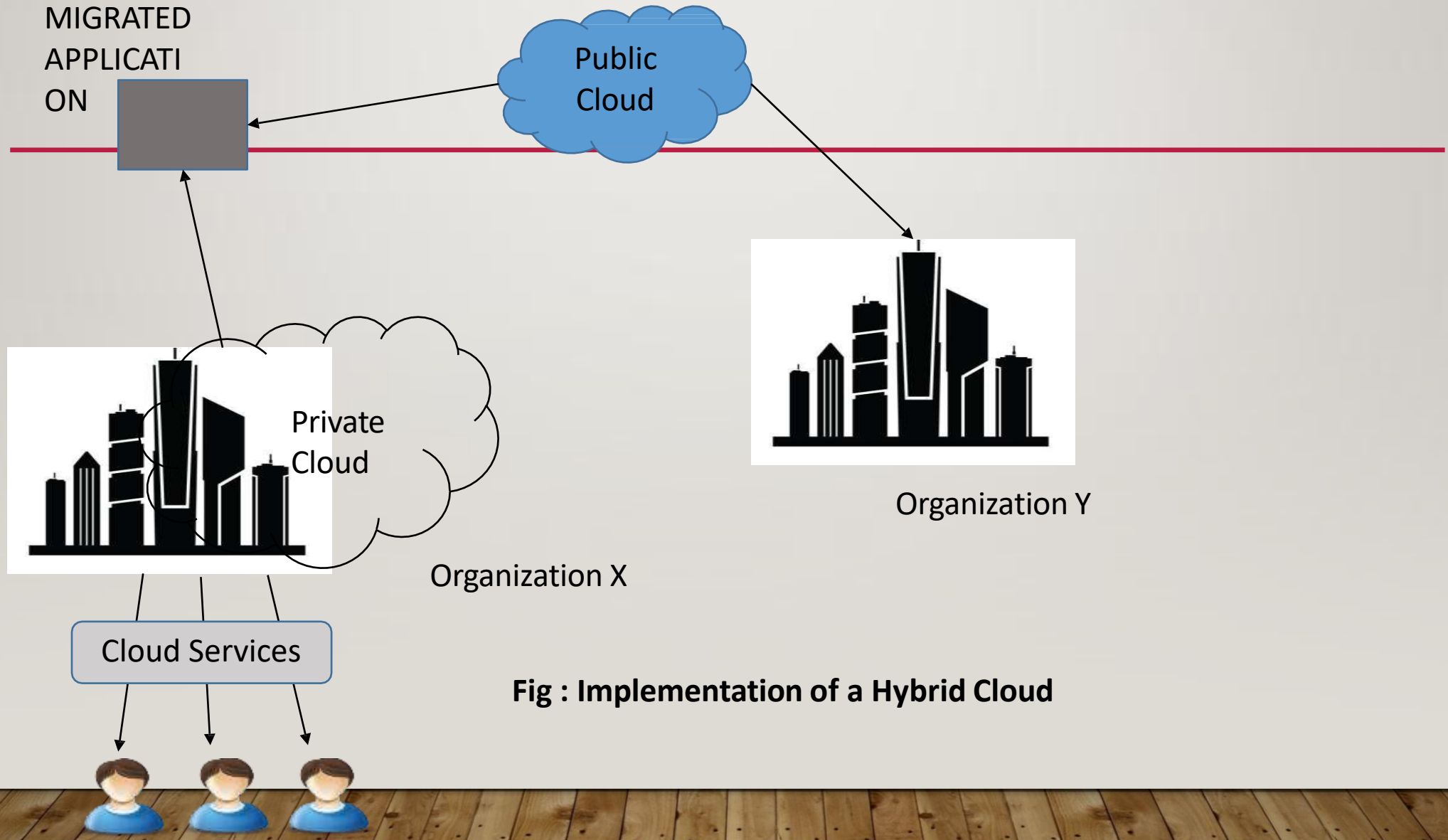


Fig : Implementation of a Hybrid Cloud

CLOUD SERVICES FOR BIG DATA

- In big data Iaas, Paas and Saas clouds are used in following manner.
- Iaas:- Huge storage and computational power requirement for big data are fulfilled by limitless storage space and computing ability obtained by iaas cloud.
- Paas:- offerings of various vendors have stared adding various popular big data platforms that include mapreduce, Hadoop. These offerings save organisations from a lot of hassels which occur in managing individual hardware components and software applications.
- Saas:- Various organisation require identifying and analysing the voice of customers particularly on social media. Social media data and platform are provided by SAAS vendors.In addition, private cloud facilitates access to enterprise data which enable these analyses.

IN MEMORY COMPUTING TECHNOLOGY

- Another way to improve speed and processing power of data.
- IMC is used to facilitate high speed data processing e.g. IMC can help in tracking and monitoring the consumers activities and behaviours which allow organizations to take timely actions for improving customer services and hence customer satisfaction.
- Data stored on external devices known as secondary storage space. This data had to be accessed from external source.
- In the IMC technology the RAM or Primary storage space is used for analysing data. Ram helps helps to increase computing speed.
- Also reduction in cost of primary memory has helped to store data in primary memory.

Thank You