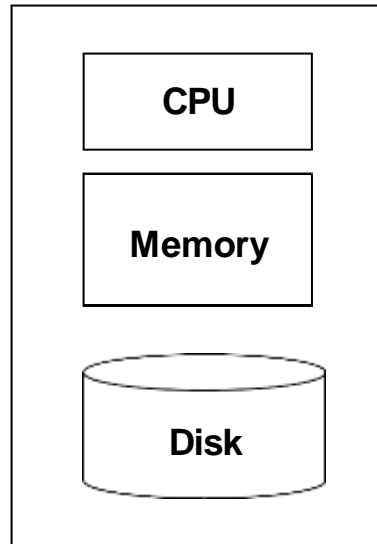

INTRODUCING TECHNOLOGIES FOR HANDLING BIG DATA

Single Node Architecture

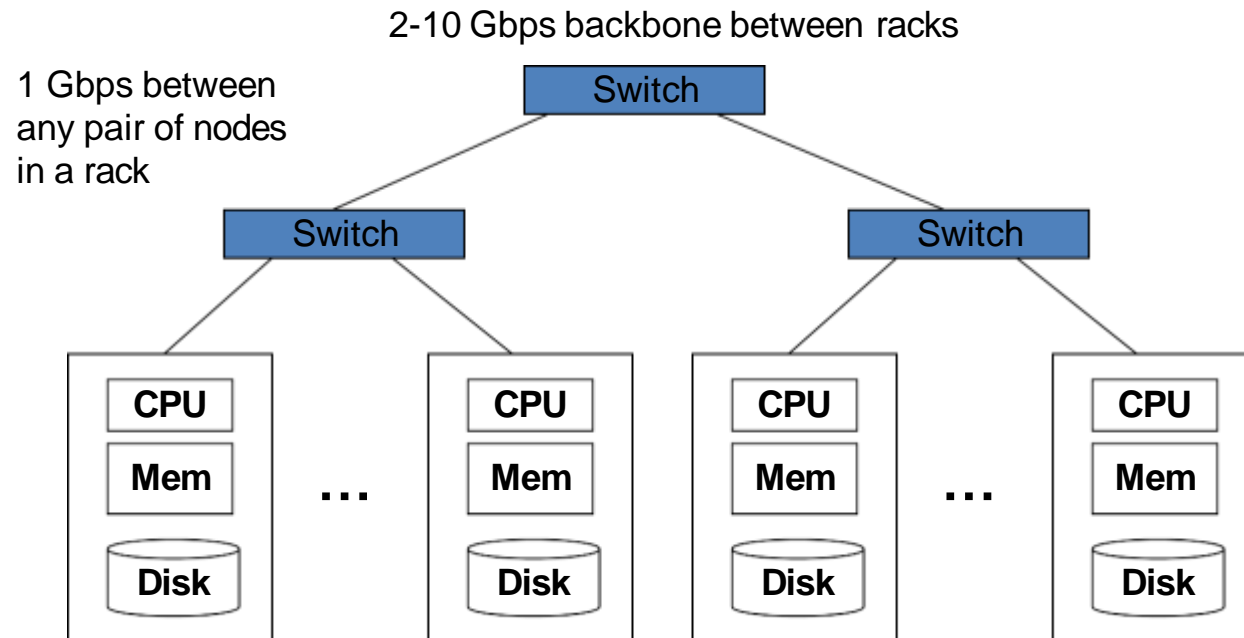


“Classical Architecture”

Motivation: Google Example

- 20+ billion web pages x 20KB = 400+ TB
- 1 computer reads 30-35 MB/sec from disk
 - ~4 to 8 months to read the web
- ~1,000 hard drives to store the web
- Takes even more to **do** something useful with the data!

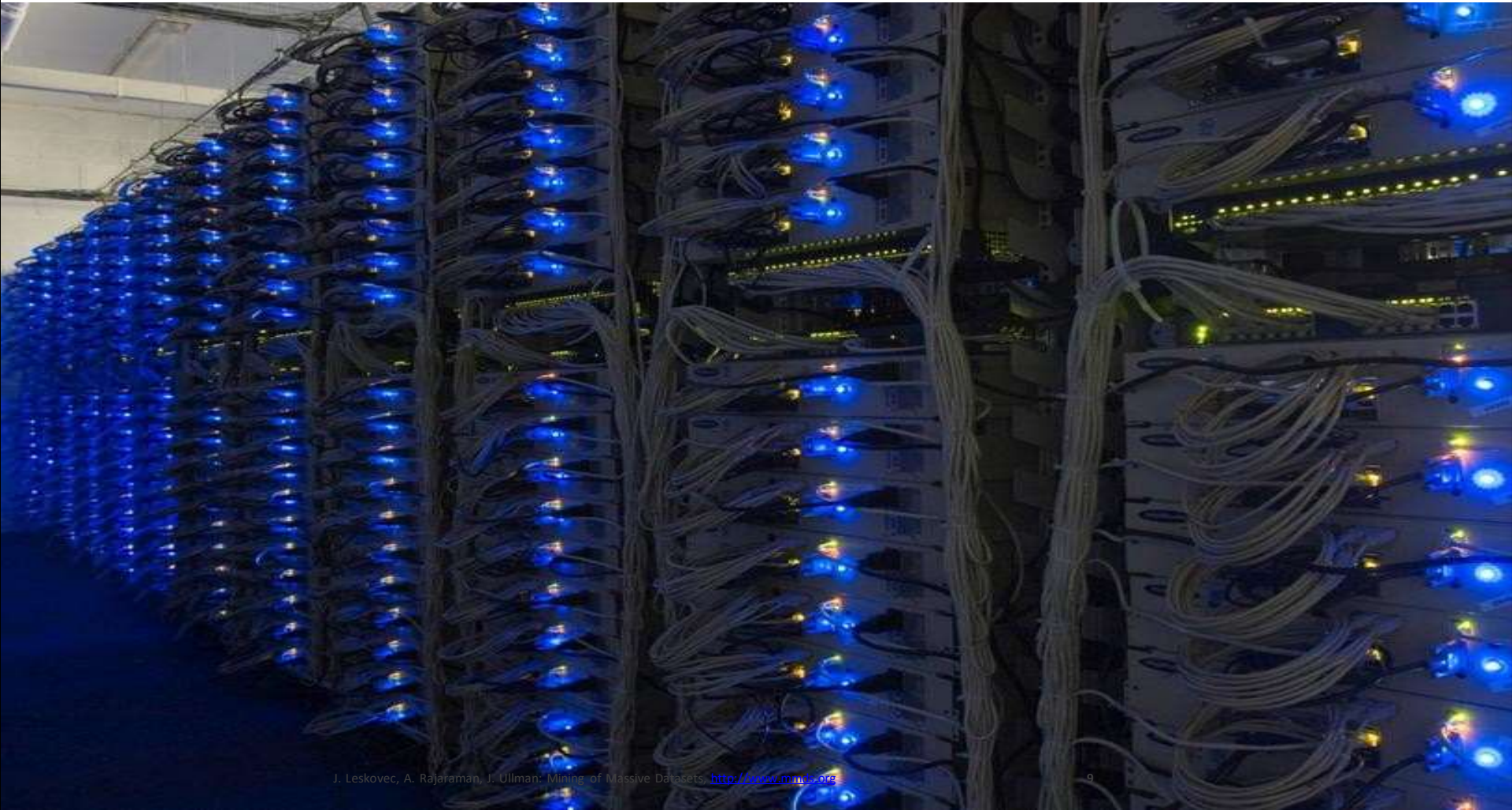
Commodity Clusters Architecture



Commodity Clusters Challenges

- Massive datasets
 - Tens to hundreds of terabytes
 - Cannot mine on a single server
- Node Failures
- Persistent Datastore
- Network Bottleneck
- Distributed Programming is Hard
 - In order to solve this problem we have MAP REDUCE and other Big Data framework

In 2011 it was guest estimated that Google had 1M machines, <http://bit.ly/Shh0RO>



History Of High Performance Computing

- 1980's Parallel Computing.
- 1990's DCS.

PARALLEL COMPUTING

- These were **shared memory** multiprocessors, with multiple processors working side-by-side on shared data.
- In the mid 1980's, a new kind of parallel computing was launched when the Caltech Concurrent Computation project built a supercomputer for scientific applications from 64 Intel 8086/8087 processors

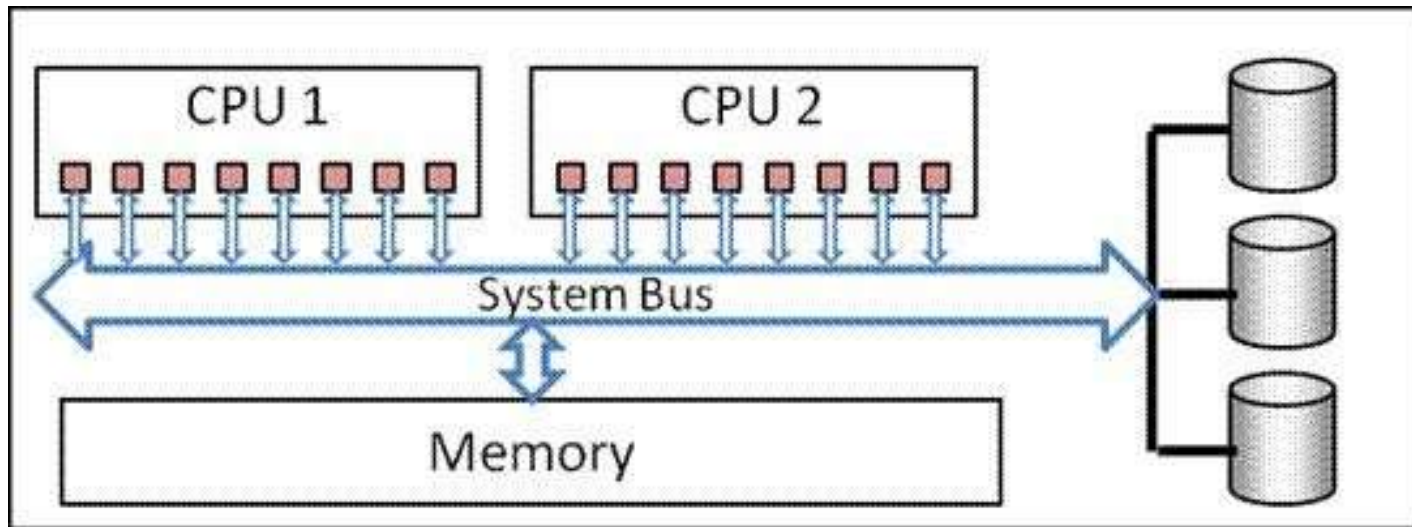
PARALLEL COMPUTING FOR BIG DATA

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 Architecture



Distributed Computing

- Distributed Computing (Cluster/Grid).
- Starting in the late 80's, **clusters** came into market.
- A cluster is a type of parallel computer built from large numbers of off-the-shelf computers connected by an off-the-shelf network.
- Today, clusters are the workhorse of scientific computing and are the dominant architecture in the data centers that power the modern information age.

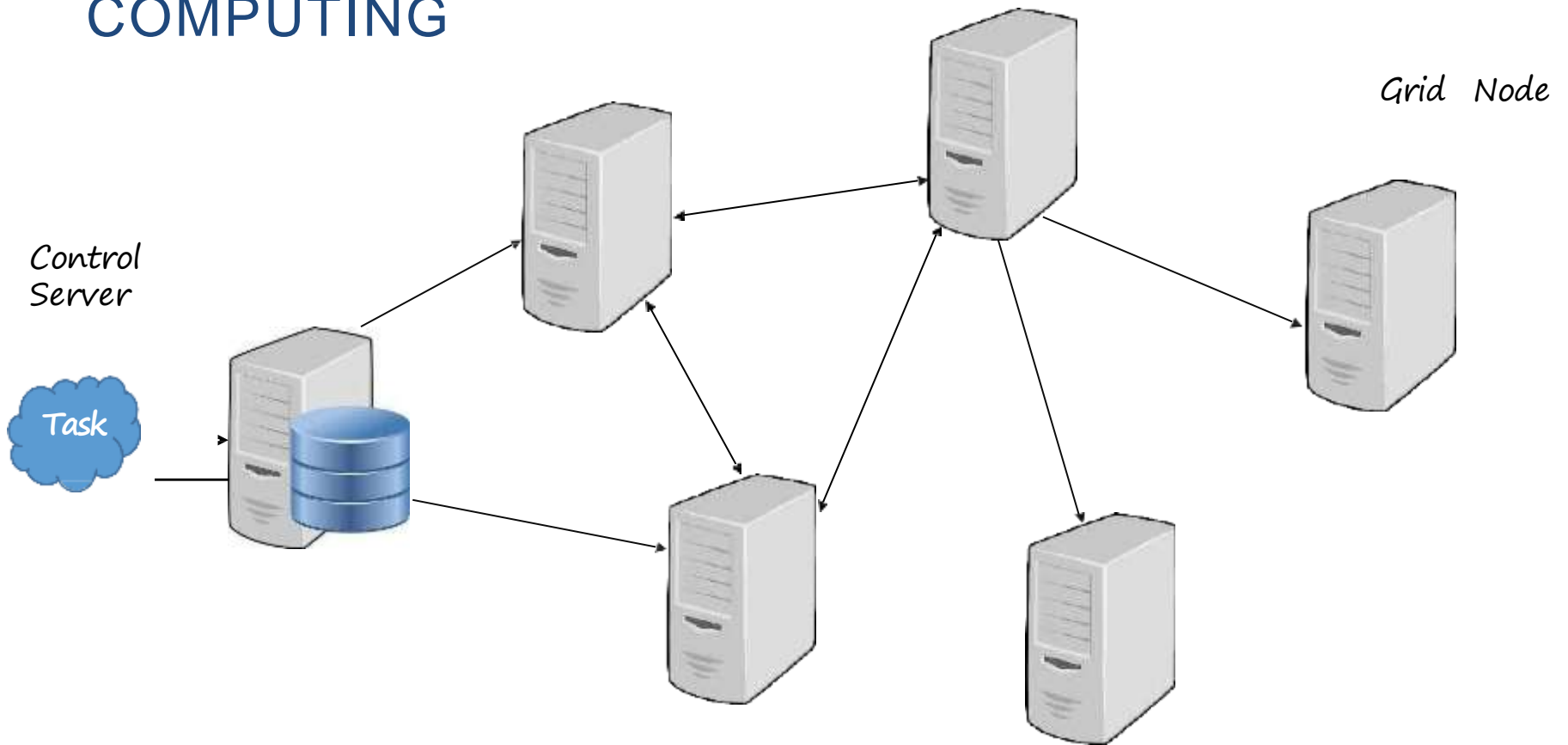
DISTRIBUTED COMPUTING FOR BIG DATA

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

- Increases the Speed
- Increases the Efficiency

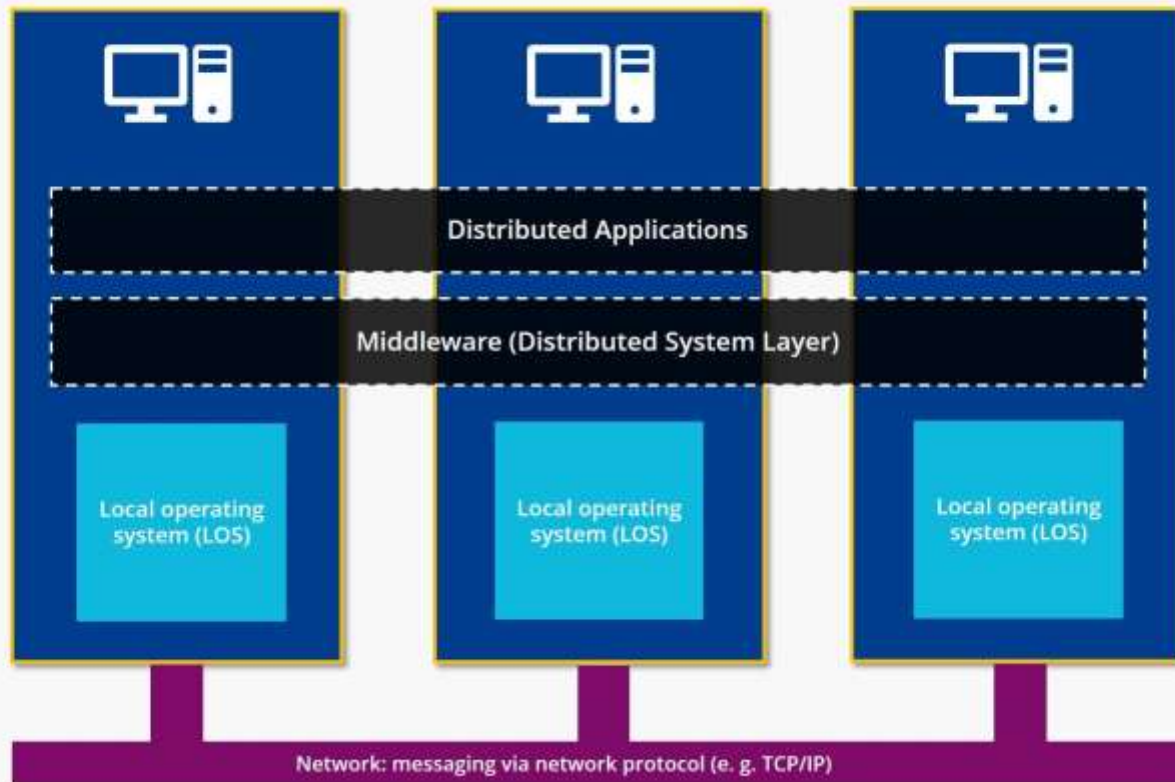
more suitable to process huge amount of data in a limited time

DISTRIBUTED COMPUTING

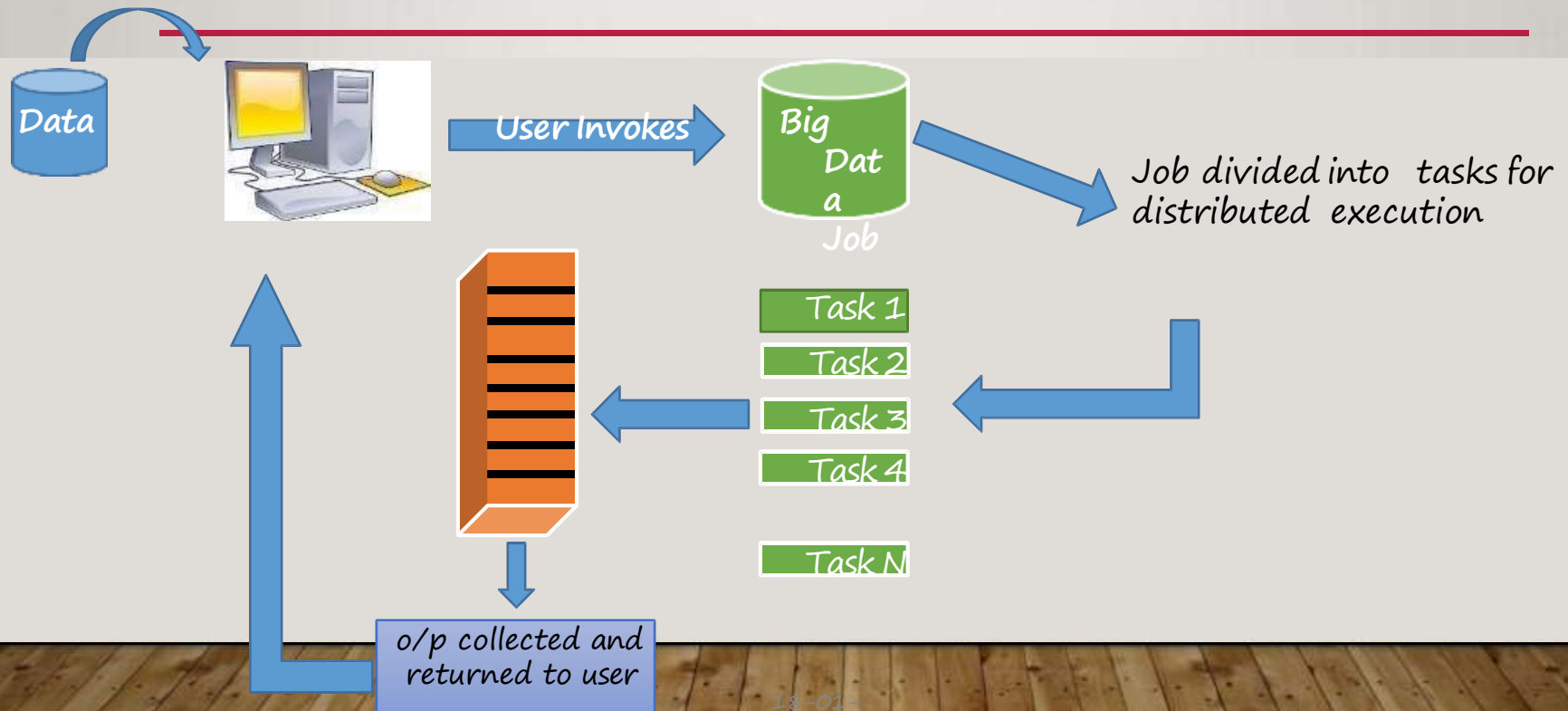


Architecture

Distributed Computing



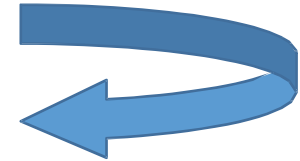
DISTRIBUTED COMPUTING TECHNIQUE FOR PROCESSING LARGE DATA



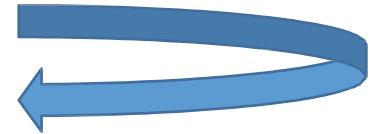
18-01-
2018

ISSUES IN THE SYSTEM

✓ 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

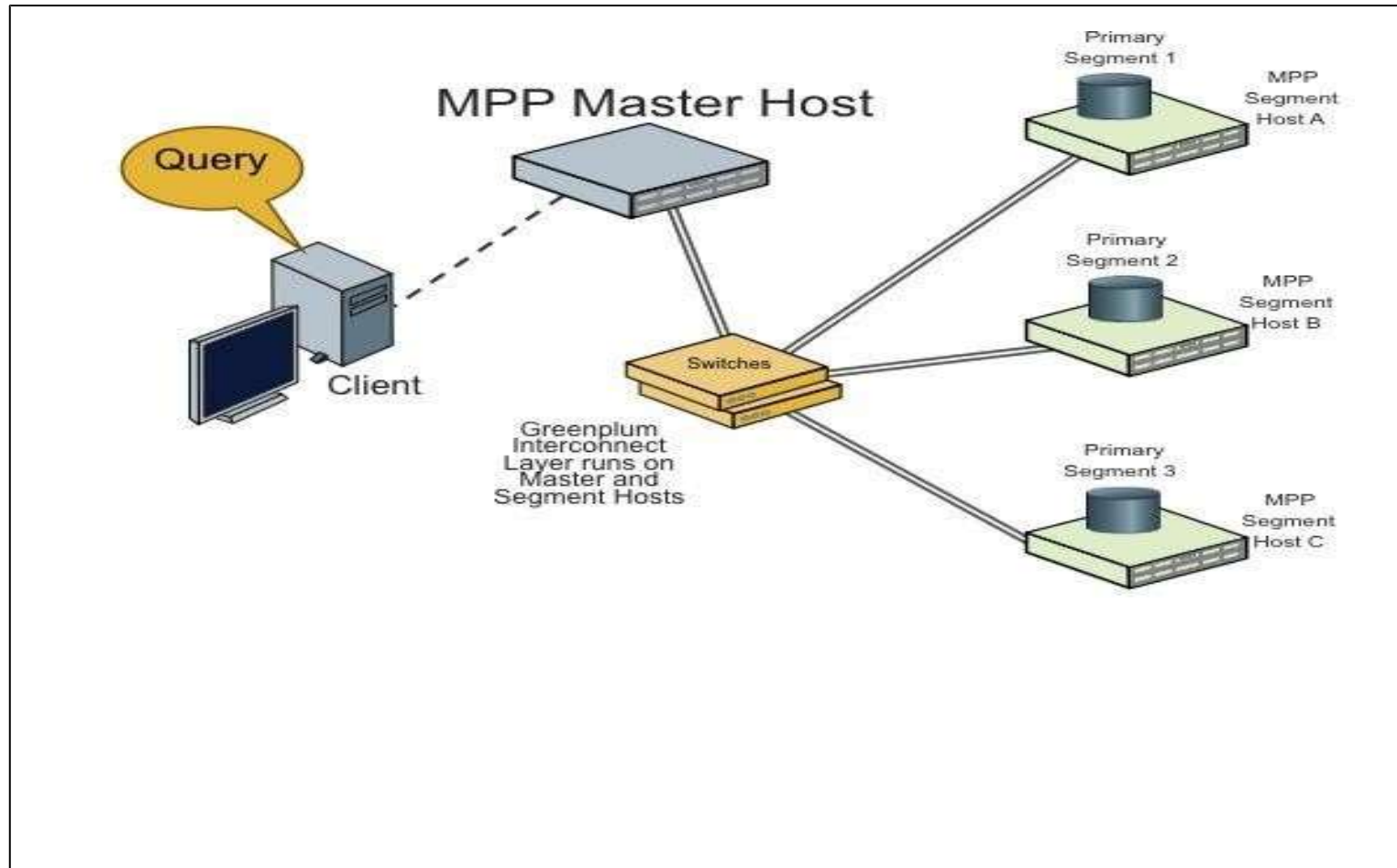


Affecting the productivity & profitability of an organization.

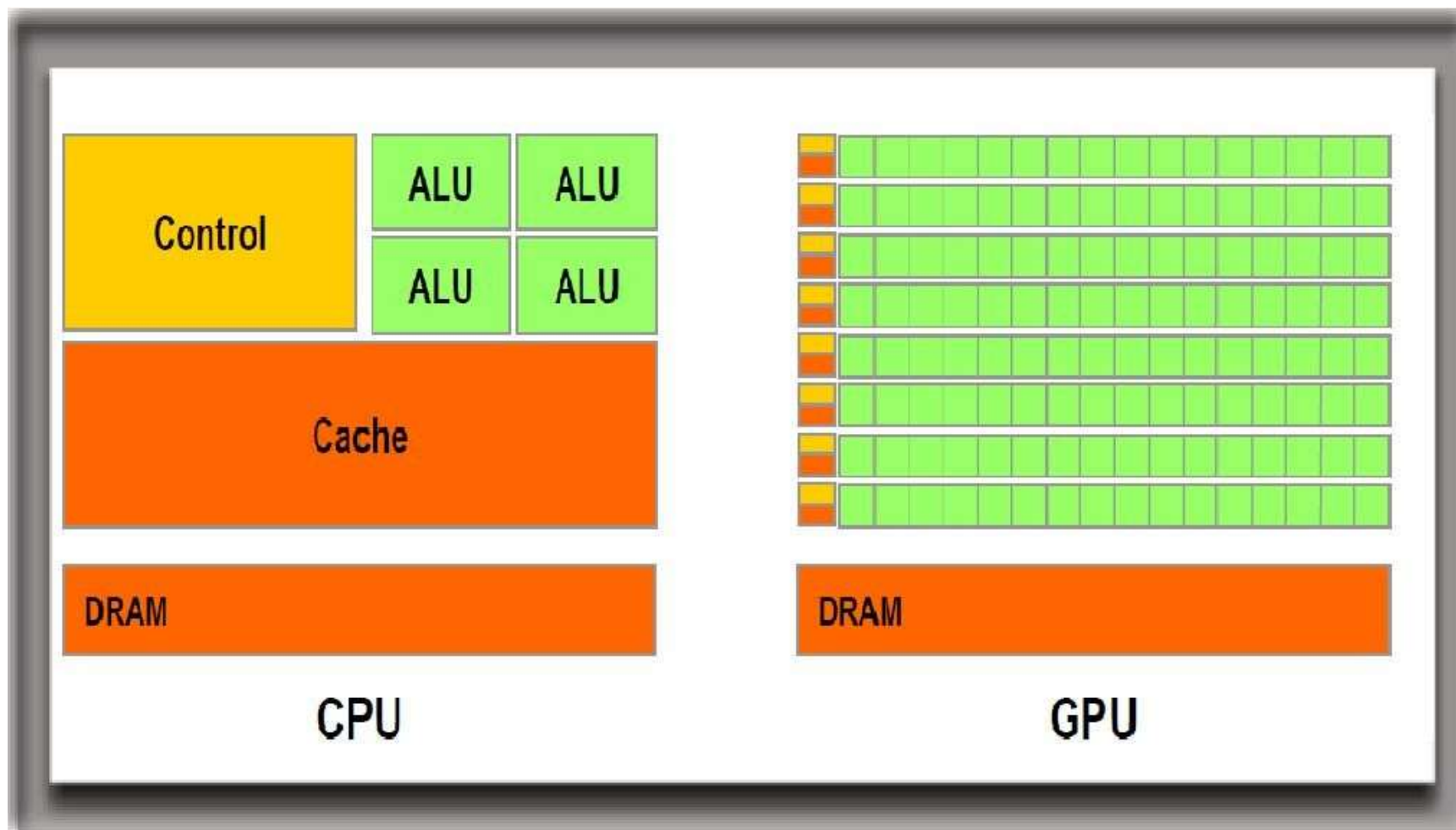
Massively Parallel Processing

- MPP (massively parallel processing) is the coordinated processing of a program by multiple processor that work on different parts of the program, with each processor using its own operating system and memory .
- Typically, MPP processors communicate using some messaging interface. In some implementations, up to 200 or more processors can work on the same application.
- An "interconnect" arrangement of data paths allows messages to be sent between processors.
- Typically, the setup for MPP is more complicated, requiring thought about how to partition a common database among processors and how to assign work among the processors.
- An MPP system is also known as a "loosely coupled" or "shared nothing" system.








Architecture



CPU vs GPU



GPU's BENCHMARKS

GPU	Cuda Cores/ Stream Processors	3D Mark Graphics Score
Nvidia GeForce RTX 3090	10496	19970 
AMD Radeon 6900XT	5120	19167 
Nvidia GeForce RTX 3080	8704	17699 
AMD Radeon 6800	3840	15217 
Nvidia GeForce RTX 3070	5888	13722 
AMD Radeon 6700XT	2560	11986 
Nvidia GeForce RTX 3060 Ti	4864	11833 

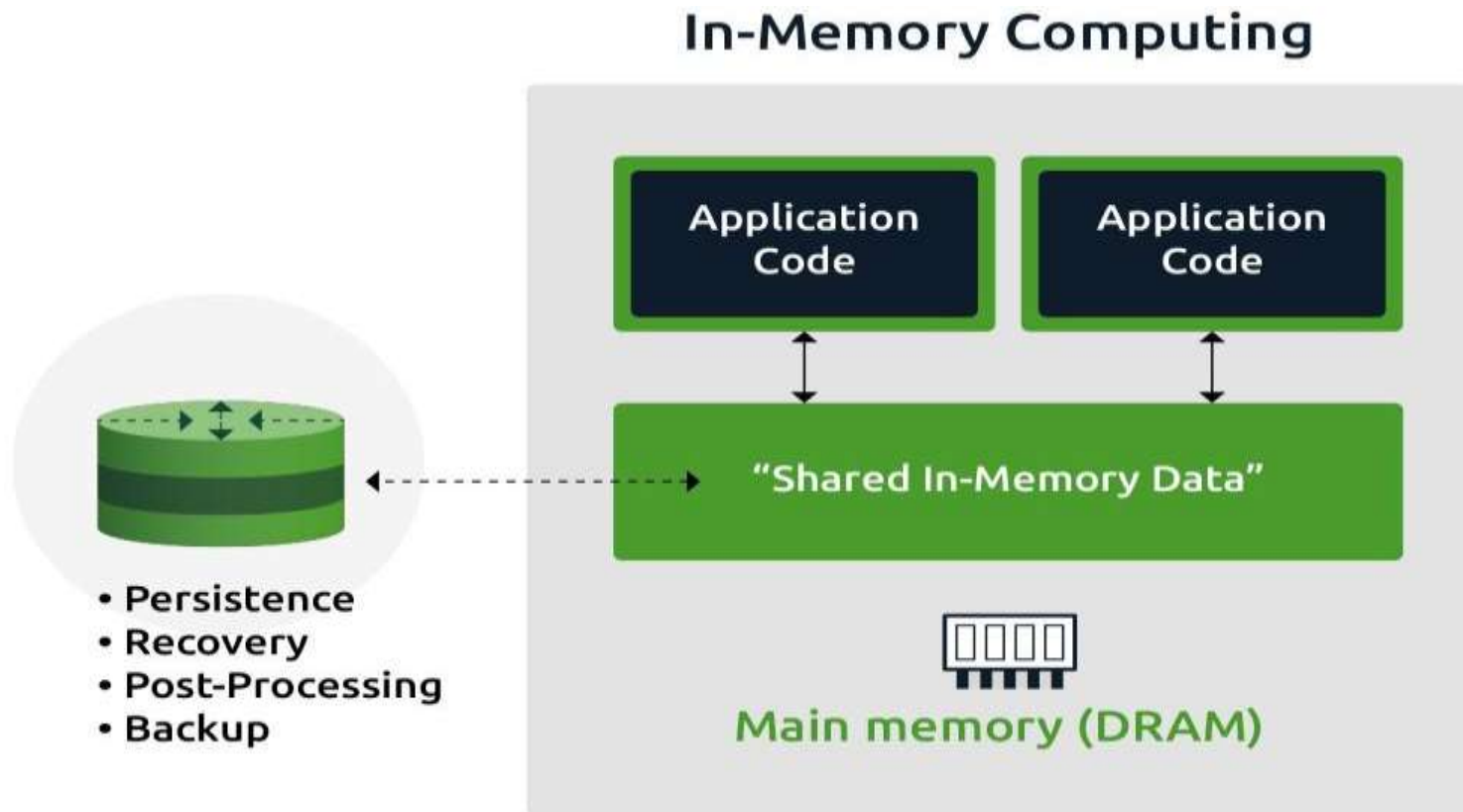
In-memory Computation

- In-memory computation works by eliminating all slow data accesses and relying exclusively on data stored in RAM.
- Overall computation performance is greatly improved by removing the latency commonly seen when accessing hard disk drives or SSDs.
- Software running on one or more computers manages the computation as well as the data in memory, and in the case of multiple computers, the software divides the computation into smaller tasks which are distributed out to each computer to run in parallel. In-memory computation is often done in the technology known as **in-memory data grids** (IMDG).

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.

Architecture



MERITS OF THE HIGH PERFORMANCE COMPUTING ARCHITECTURE

- **Scalability**
- **Virtualization**
- **Load Balancing Features**
- **Optimization**

PROBLEMS OF THE HIGH PERFORMANCE COMPUTING SYSTEM

Large-scale Computing

- **Large-scale computing** problems on **commodity hardware**
- **Challenges:**
 - **How do you distribute computation?**
 - **How can we make it easy to write distributed programs?**
 - **Machines fail:**
 - One server may stay up 3 years (1,000 days)
 - If you have 1,000 servers, expect to loose 1/day
 - People estimated Google had ~1M machines in 2011
 - 1,000 machines fail every day!

Storage Infrastructure

- **Problem:**

- If nodes fail, how to store data persistently?

- **Answer:**

- **Distributed File System:**

- Provides global file namespace
- Google GFS; Hadoop HDFS;

- **Typical usage pattern**

- Huge files (100s of GB to TB)
- Data is rarely updated in place
- Reads and appends are common

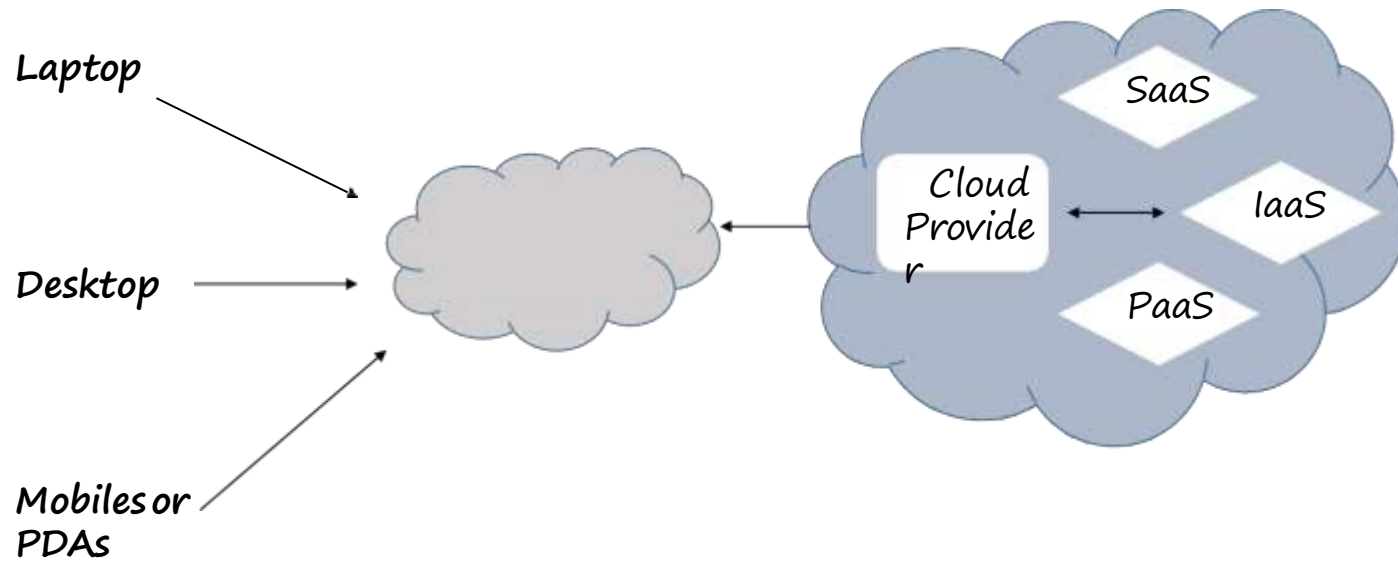
J. Leskoveri, "Data Science in the Cloud: Managing Massive Datasets", <http://www.mmds.org>

CLOUD COMPUTING AND BIG DATA

Cloud Computing is the delivery of computing services, 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.

- **Problem** : 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.

FEATURES OF CLOUD COMPUTING

- **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.

FEATURES OF CLOUD COMPUTING

- **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.

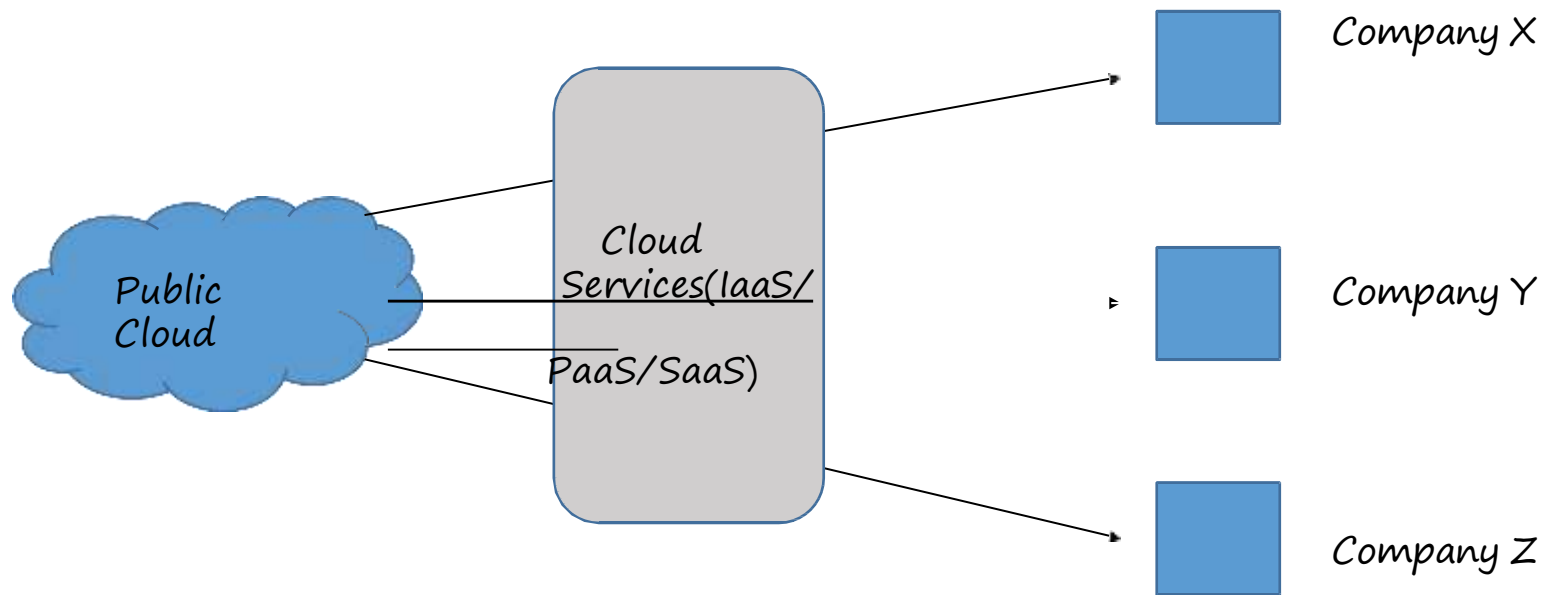


Fig : Level of Accessibility in a Public Cloud

• **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.

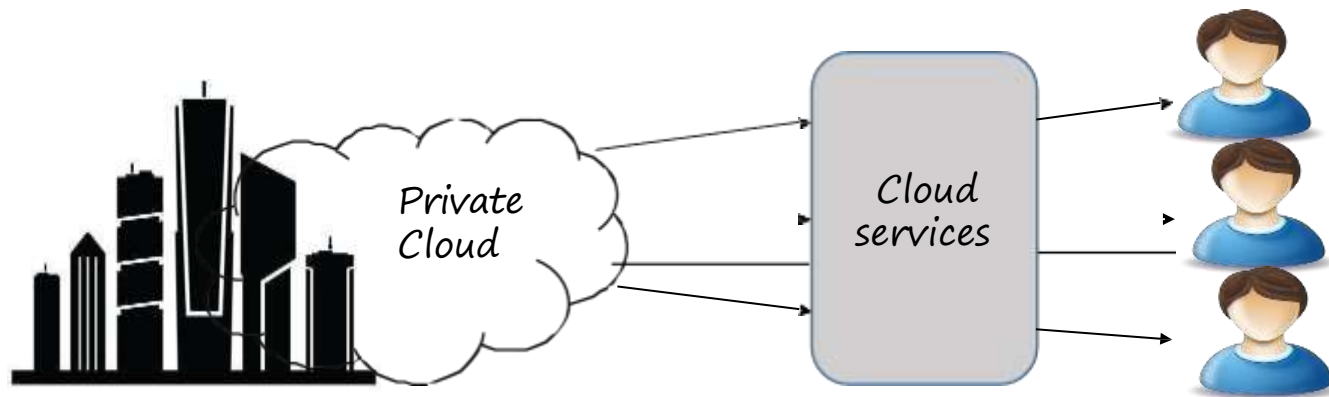


Fig: Level of Accessibility in a Private Cloud

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 be 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.

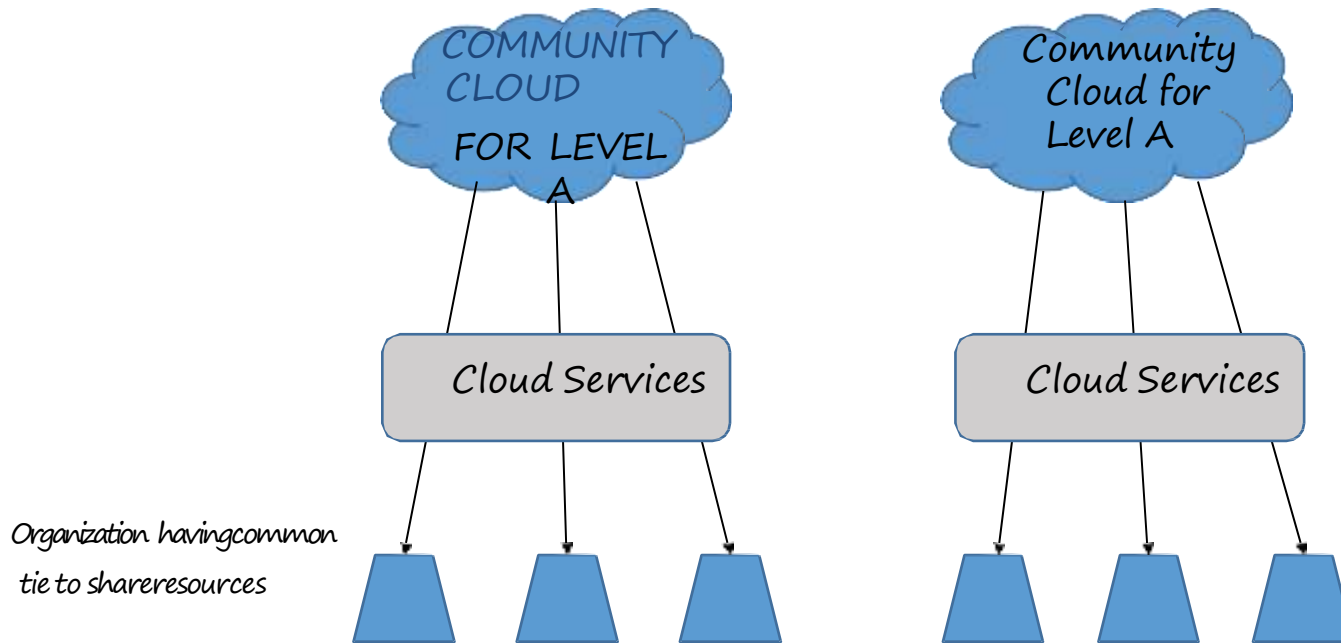


Fig : Level of Accessibility in Community Clouds

Hybrid Cloud

- Various internal or external service providers offer services to many organizations.
- In hybrid clouds, an organization can use both types of cloud, 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.

MIGRATED
APPLICATION

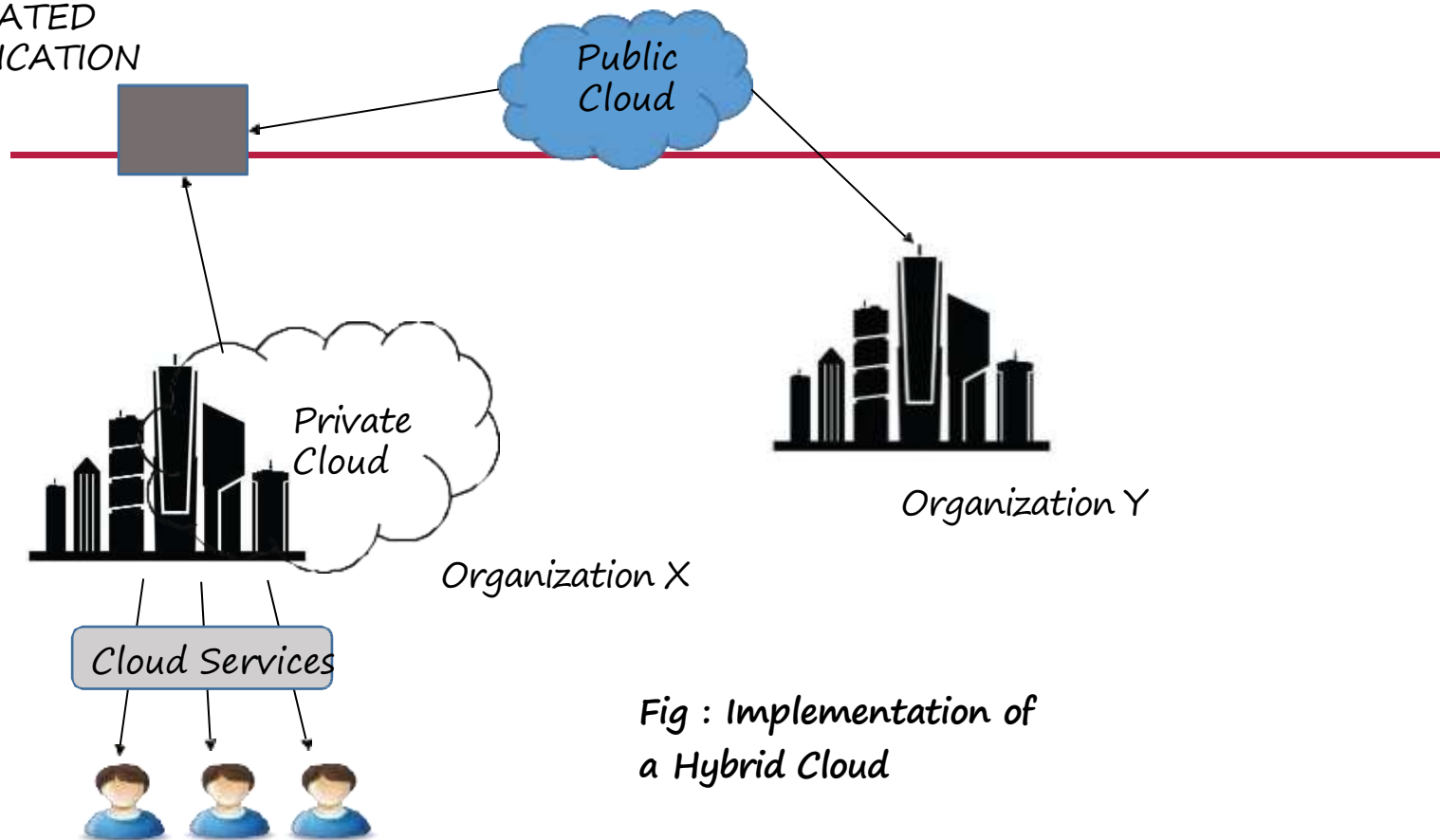


Fig : Implementation of
a Hybrid Cloud

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 started adding various popular big data platforms that include mapreduce, Hadoop. These offerings save organizations from a lot of hassels which occur in managing individual hardware components and software applications.
- SaaS :- Various organisation require identifying and analyzing 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.

CLOUD

Data Centers

Thousands

FOG

Nodes

Millions

EDGE

Devices

Billions



Why Edge & Fog Computing

- The Internet of Things (IoT) is driving business transformation by connecting everyday objects and devices to one another and to cloud-hosted services.
- Current deployment models emphasize mandatory cloud connectivity; These are two of the primary issues with connecting edge devices to the cloud for all services:
- Connected devices are creating data at an exponentially growing rate, which will drive performance and network congestion challenges at the edge of infrastructure.
- There are performance, security, bandwidth, reliability, and many other concerns that make cloud-only solutions impractical for many use cases.

Edge Computing

- Edge computing is a modern computing paradigm that functions at the edge of the network. It allows client data to be processed closer to the data source instead of far-off centralized locations such as huge cloud data centers.
- Edge computing takes data storage, enterprise applications, and computing resources closer to where the user physically consumes the information.
- This allows for efficient data processing as large amounts of data can be processed near the source itself, thereby reducing internet bandwidth, eliminating costs and allowing applications to be used in remote locations. This also helps in security as there is not interaction with any public cloud.

Edge Computing

- Edge devices can be laptops, sensors, smartphones, gateways, etc.
- Autonomous vehicles, Internet of Things, Software as Service, Voice Assistants, Predictive maintenance machines or services, Traffic Management, etc. All of these need real time data processing and reduced latency.
- In any use cases that have to do with latency-sensitive processing of information, Edge computing would be the best solution because data does not have to traverse over a network to a data center or cloud for processing.

“As the Internet of Things evolves, the rise of edge computing becomes inevitable. We need to start thinking how does edge computing fit our digital strategy and start building the implementation roadmap. Its time for a strategic consultation! “

Fog Computing

- Fog computing is between the Edge and the cloud, it can take data from the Edge before it reaches the cloud. The data taken in at the fog is then classified as to what is relevant and what is not.
- After this, the relevant data remains in the cloud for storage, and the rest of the unimportant data gets deleted or remains in a fog node for remote access.
- Fog computing allows backup infrastructure to be located closer to the data source, minimizing data transfer latency and improving backup and restore speeds. This is particularly beneficial for organizations with large amounts of data or time-sensitive applications.
- By performing data deduplication, compression, and pre-processing at the fog layer, only relevant and unique data needs to be transmitted to the central backup repository or cloud. This reduces the amount of data transferred over the network, optimizing bandwidth usage and minimizing costs.

Fog Computing

Fog computing is a computing architecture in which a series of nodes receives data from IoT devices in real time. These nodes perform real-time processing of the data that they receive, with millisecond response time. The nodes periodically send analytical summary information to the cloud.

“Fog computing has emerged as a technology trend that is complementary to cloud computing, and for the IoT, both fog and cloud are necessary. Machines generate tremendous amounts of data and (the) vast majority of this data is extremely ephemeral, which can be examined and dropped. The fog is necessary as the first layer of quick response to the very fast-moving information. The cloud is the repository to which the fog layer pushes data for post-facto analysis.”

