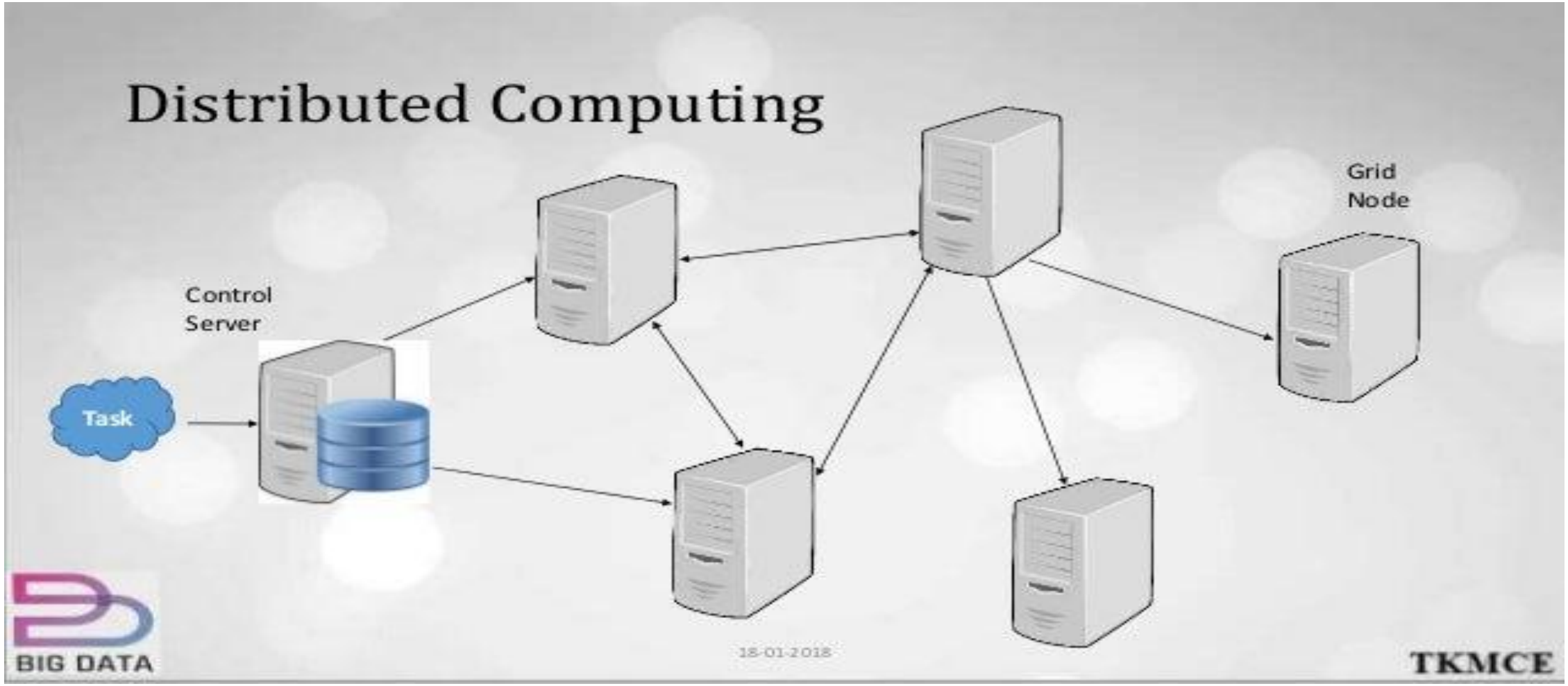
UNIT-3
Technologies for Handling Big Data

Prof. Vipul Gamit
PIET-MCA
Parul University

Distributed and Parallel Computing for 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 Computing
  - Multiple computing resources are connected in a network and computing tasks are distributed across these resources.
    - Increase the speed
    - Increase the Efficiency
    - More suitable to process huge amount of data in a limited time.

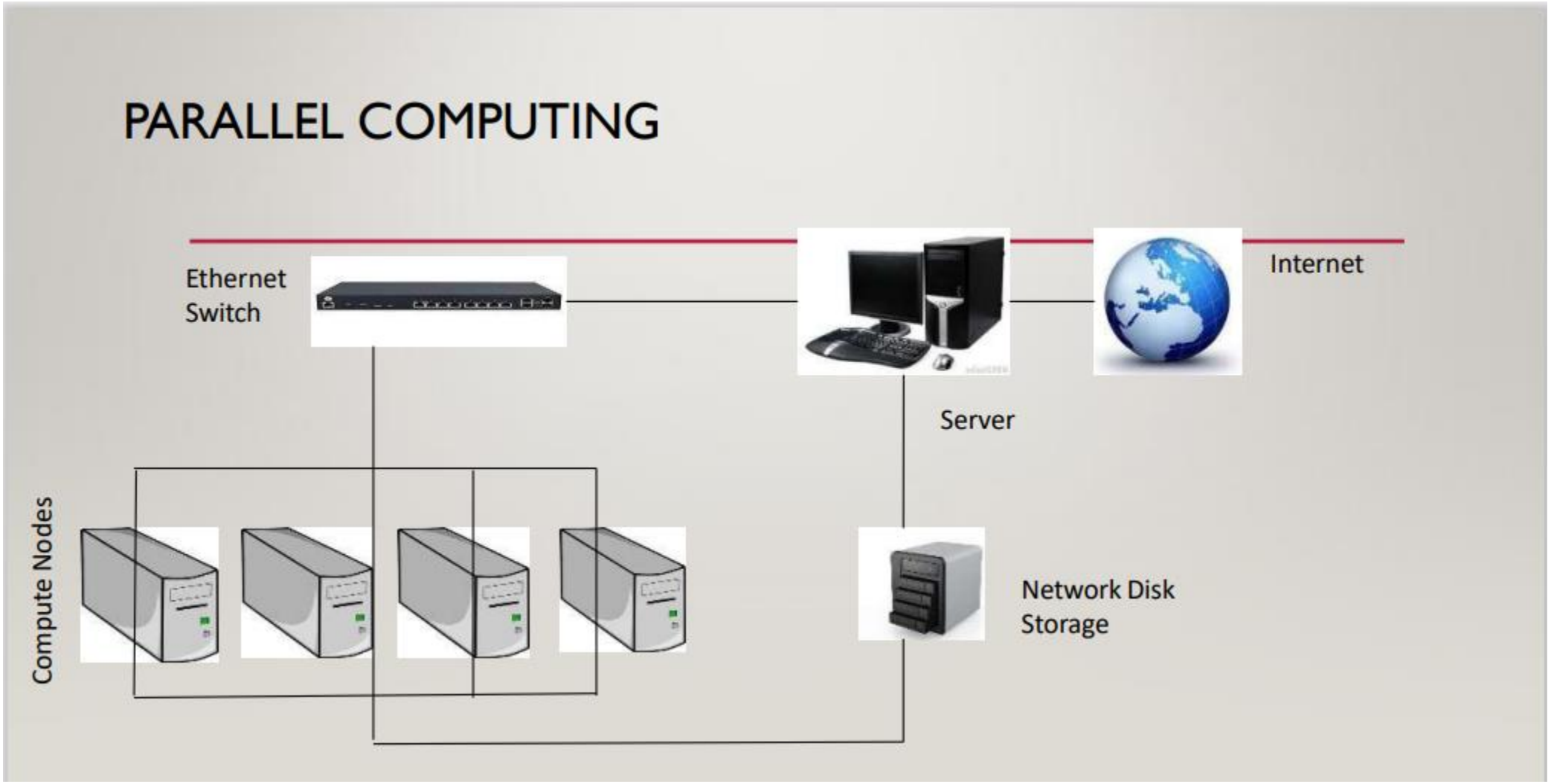
Distributed Computing



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



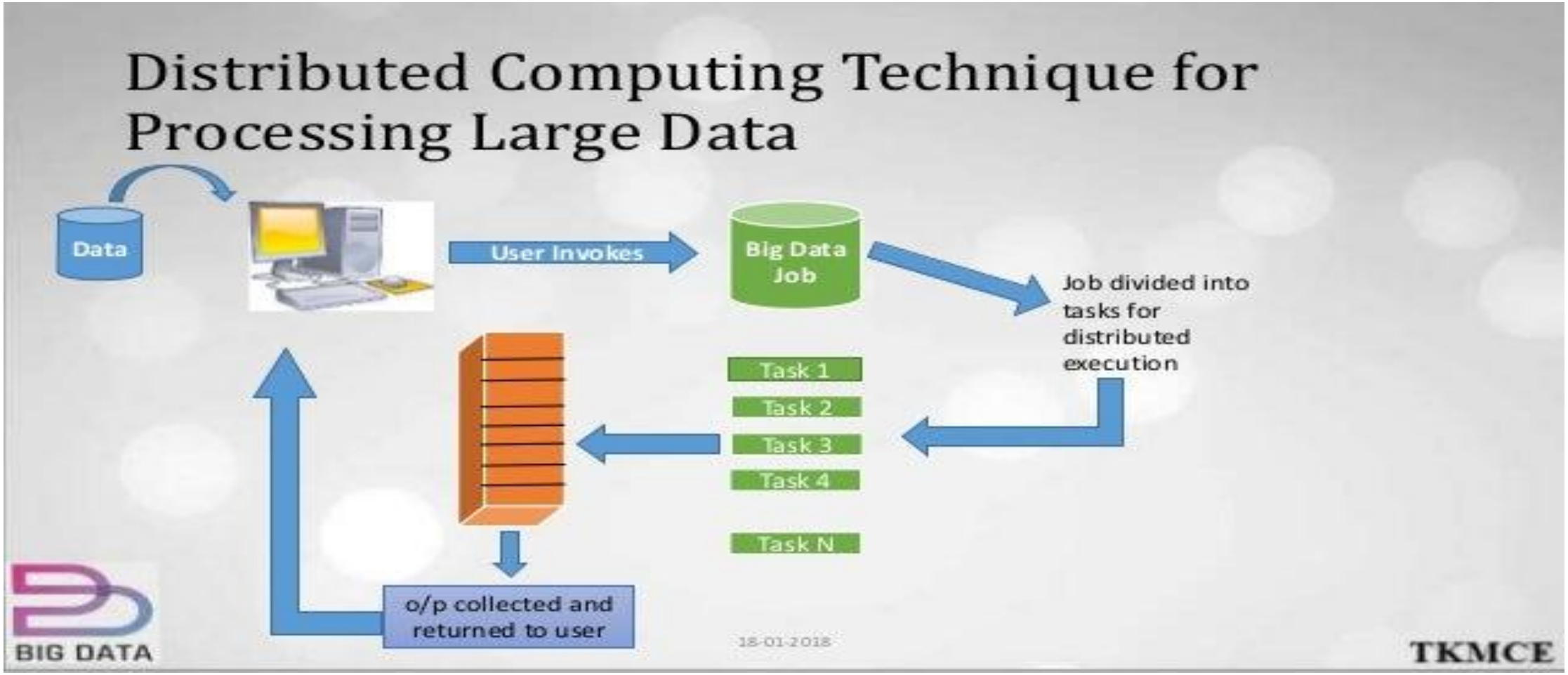
BIG DATA PROCESSING TECHNIQUES

- With the increase in data, forcing organizations to adopt a data analysis strategy that can be used for analyzing 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 because 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



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
   - 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
   - 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 between Distributed and Parallel Computing for Big Data

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

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.

In-Memory Computing Technology for Big Data

- Another way to improve speed and processing power of data.
- IMC is used to facilitate high speed data processing e.g. IMC can hel;p in tracking and monitoring the consumers activities and behaviors 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 analyzing data. Ram helps to increase computing speed.
- Also reduction in cost of primary memory has helped to store data in primary memory.