Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

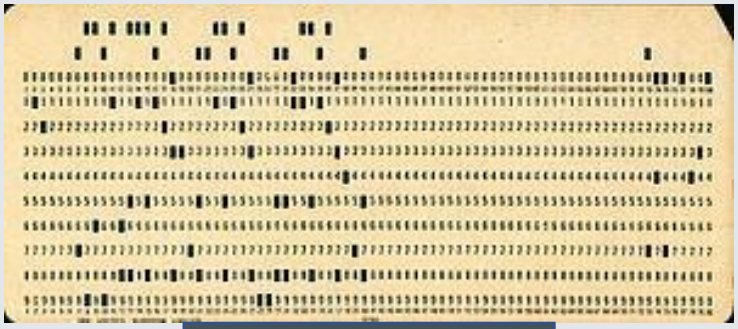
Associate Professor

Learning Outcomes

After this lecture, you will be able to

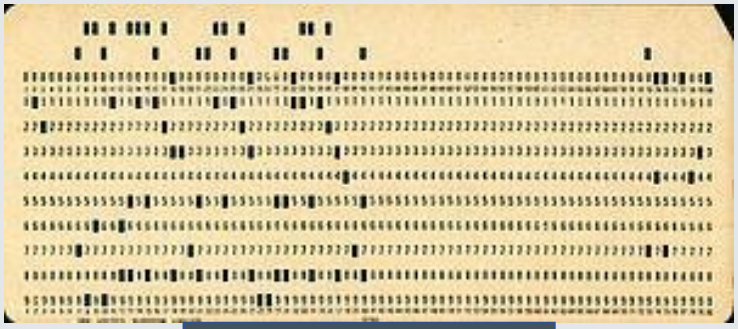
- differentiate between file system (FS) and distributed file system (DFS)
- understand how DFS works?
- explore the advantages of DFS?
- understand scalable computing over the internet.

What is File System (FS)?



Punch Cards

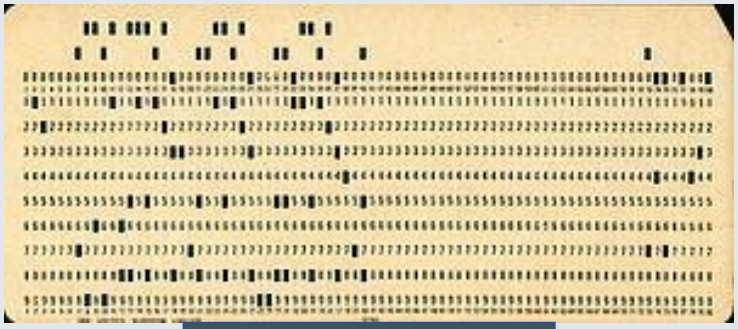
What is File System (FS)?



Punch Cards

File Cabinet

What is File System (FS)?

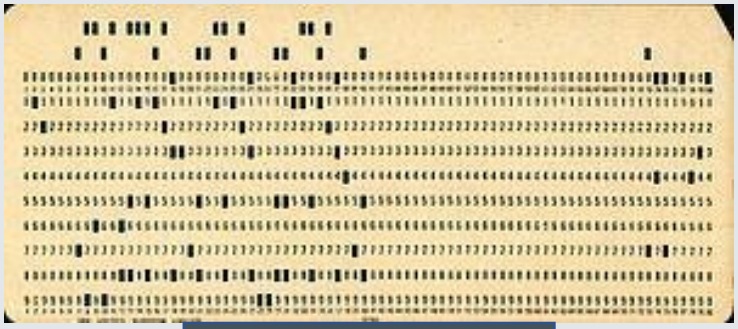


Punch Cards

File Cabinet

- This is very similar to what we do nowadays to archive papers in government intuitions who still use paper work on daily basis

What is File System (FS)?



Punch Cards

File Cabinet

- This is very similar to what we do nowadays to archive papers in government intuitions who still use paper work on daily basis
- Instead of storing information on punch cards; we can now

Example

Hard Disk

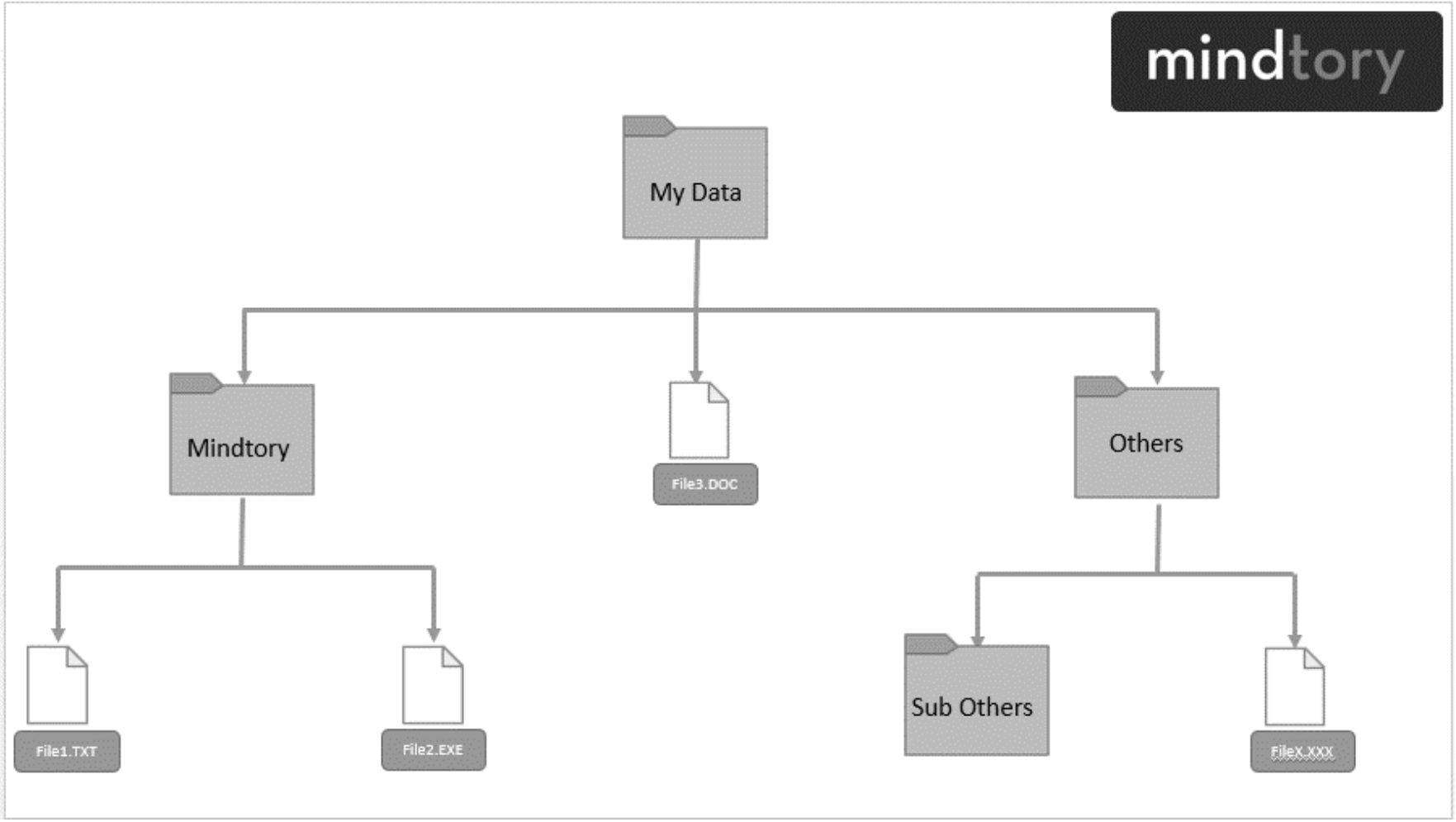
Example

Hard Disk

Flash Drive

- Related data are still categorized as files; related groups of files are stored in folders.
- Each file has a name, extension and icon. The file name gives an indication about the content it has while file extension indicates the type of information stored in that file.
  for example; EXE extension refers to executable files, TXT refers to text files…etc.

Example to file system – FS



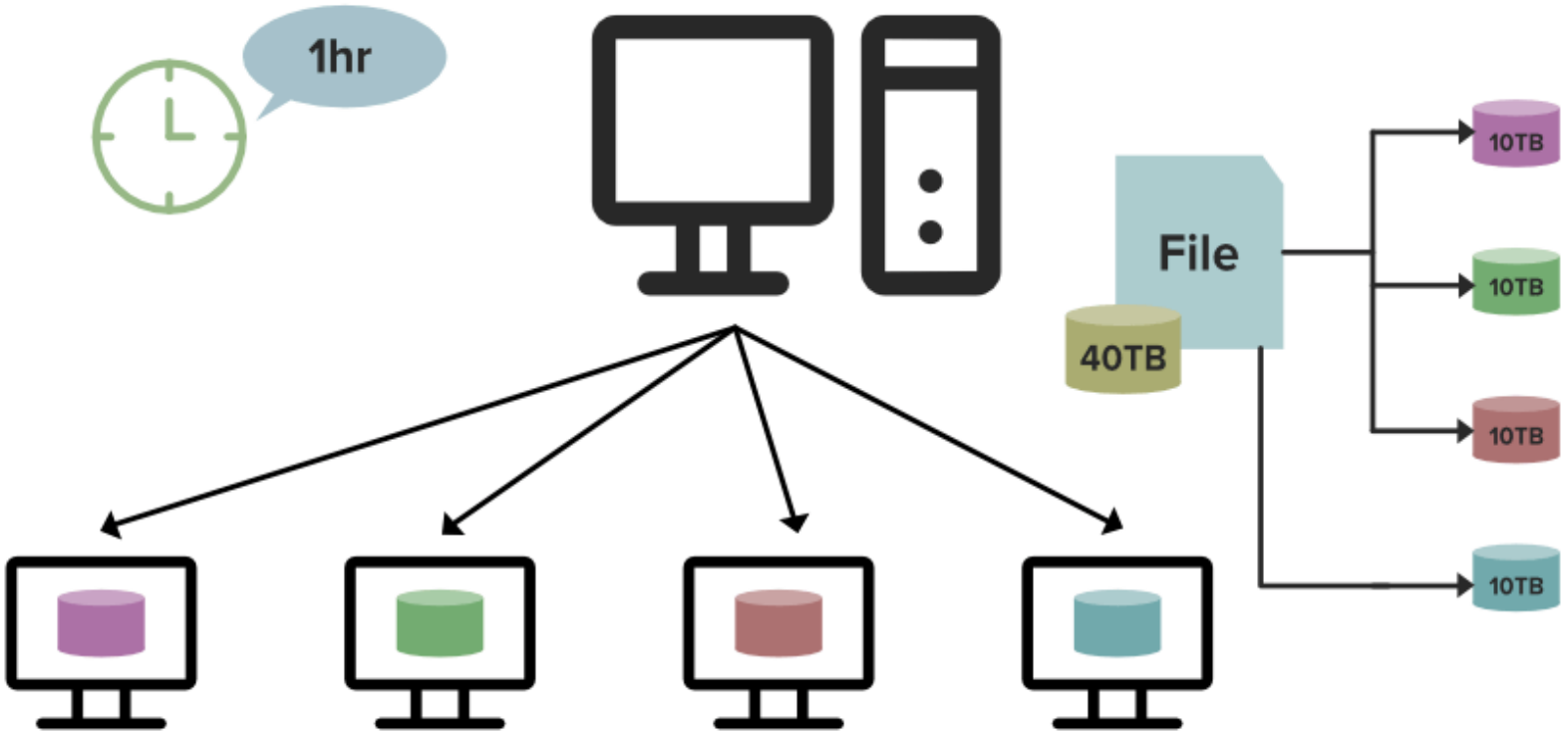
Example to file system – FS

- File management system is used by the operating system to access the files and folders stored in a computer or any external storage devices.
- Imagine file management system as a big dictionary that contains information about file names, locations and types. File management system is capable of handling files within one computer or a cluster. But what if we have many? So here comes DFS

What is Distributed file system (DFS)?

- In Big Data, we deal with multiple clusters (computers) often. One of the main advantages of Big Data which is that it goes beyond the capabilities of one single super powerful server with extremely high computing power.
- The whole idea of Big Data is to distribute data across multiple clusters and to make use of computing power of each cluster (node) to

Distribution Concept



DFS has two components

- Location Transparency: Location Transparency achieves through the namespace component.
- Redundancy: Redundancy is done through a file replication component.

Features of DFS

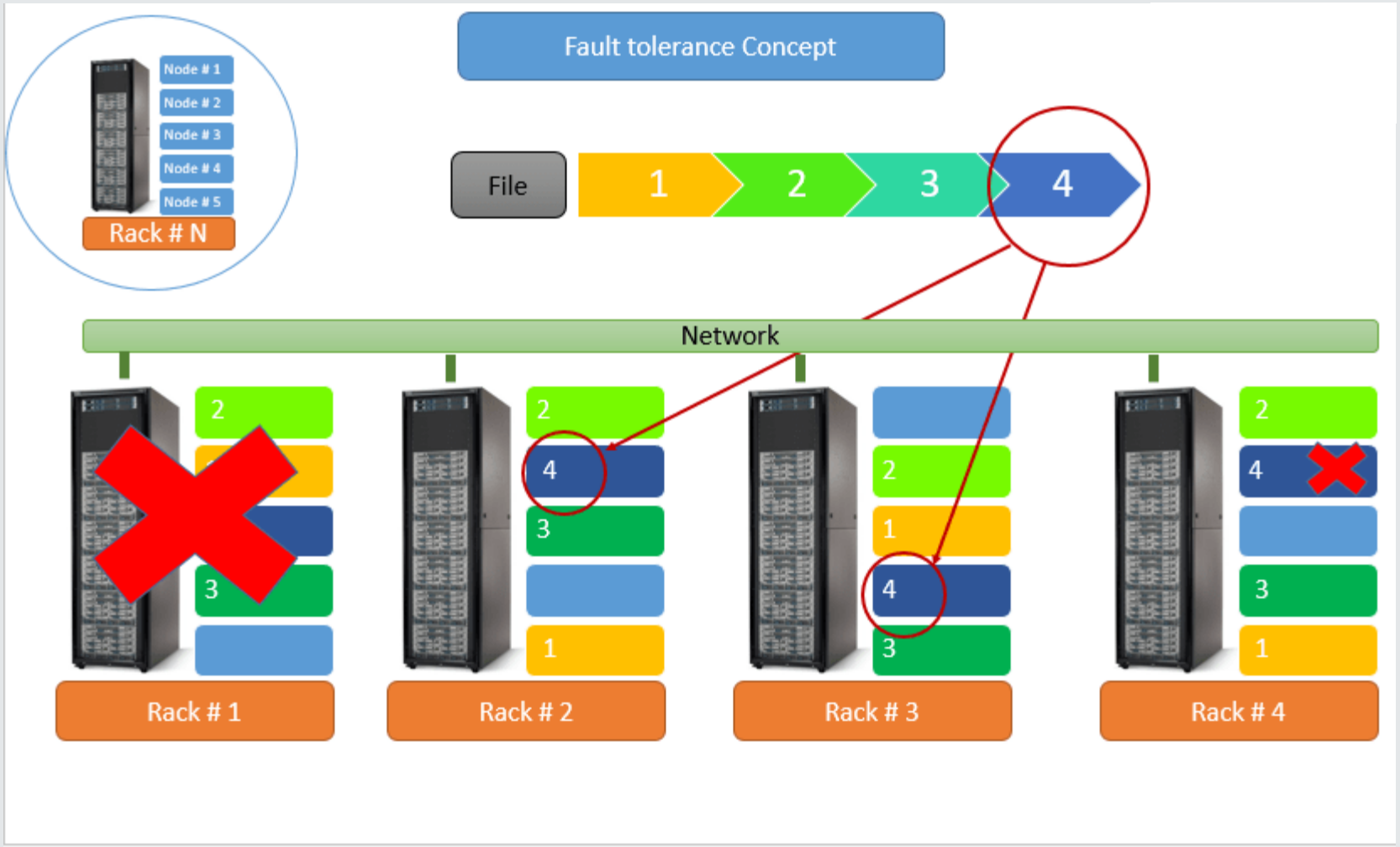
- Transparency
  - Structure transparency:
  - Access transparency
  - Naming transparency
  - Replication transparency
- User mobility
- Performance
- Simplicity and ease of use
- High availability

How Distributed file system (DFS) works?

Distributed file system works as follows:

- Distribution
- Replication
  - *Fault Tolerance*
  - *High Concurrency*

Fault Tolerance Concept

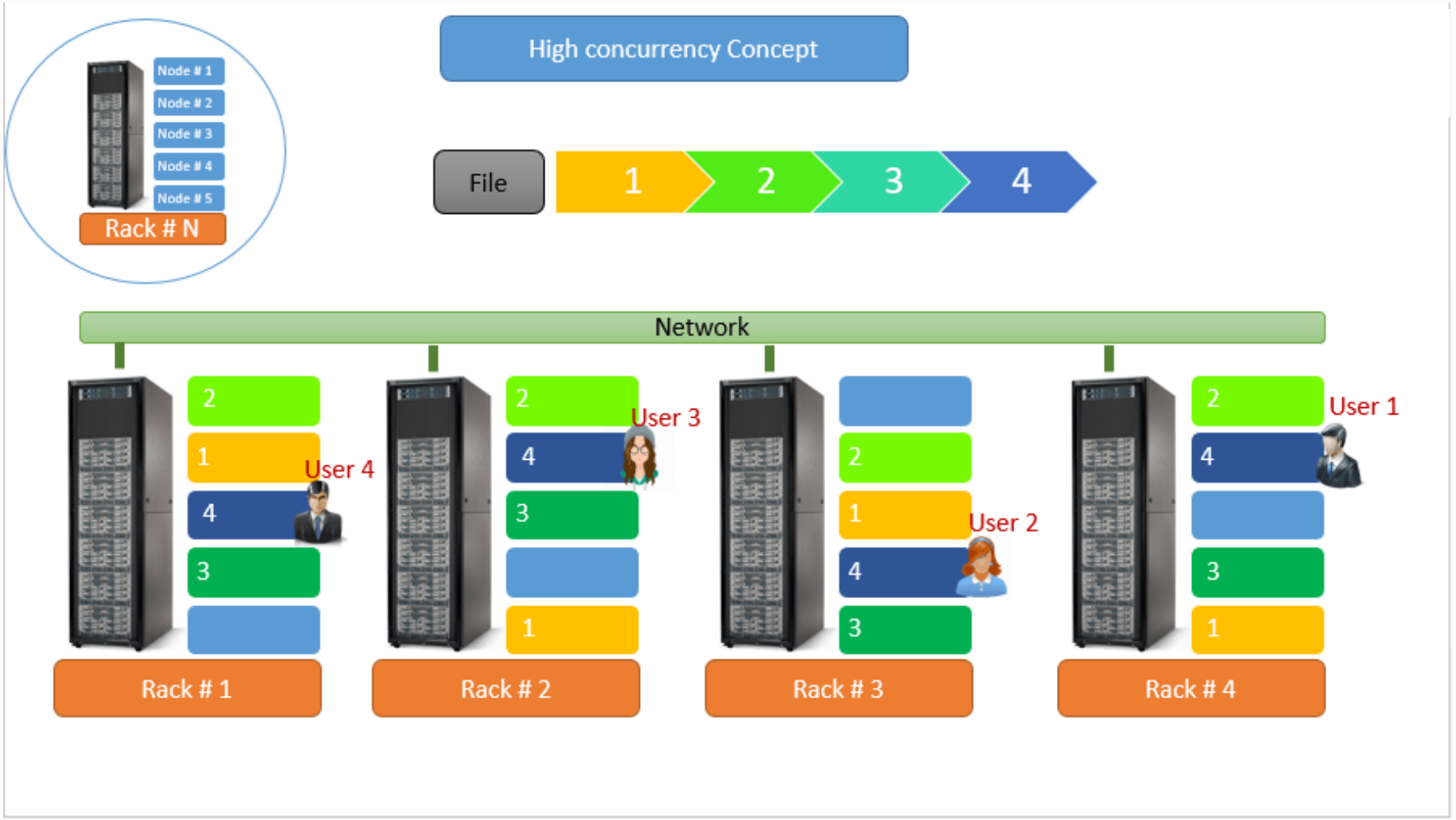


Data replication is a good way to achieve fault tolerance and high concurrency; but its very hard to maintain frequent changes. Assume that someone changed a data block on one cluster; these changes need to be updated on all data replica of this block

What are the Advantages of Distributed File System (DFS)?

- Scalability
- Fault Tolerance
- High Concurrency

High Concurrency Concept



Advantages

- DFS allows multiple user to access or store the data.
- It allows the data to be share remotely.
- It improved the availability of file, access time and network efficiency.
- Improved the capacity to change the size of the data and also improves the ability to exchange the data.
- Distributed File System provides transparency of data even if server or disk fails.

Disadvantages

- In Distributed File System nodes and connections needs to be secured therefore we can say that security is at stake.
- There is a possibility of lose of messages and data in the network while movement from one node to another.

Disadvantages

- Database connection in case of Distributed File System is complicated.
- Also handling of the database is not easy in Distributed File System as compared to a single user system.
- There are chances that overloading will take place if all nodes tries to send data at once.

Scalable Computing Over the Internet

What is scalability?

With Cloud hosting, it is easy to grow and shrink the number and size of servers based on the need. This is done by either increasing or decreasing the resources in the cloud. This ability to alter plans due to fluctuation in business size and needs is a superb benefit of cloud computing especially when experiencing a sudden growth in demand.

Scalability computing over internet

- The Age of Internet Computing
- High-Performance Computing
- High-Throughput Computing
- Three New Computing Paradigms
- Computing Paradigm Distinctions
- Distributed System Families
- Degrees of Parallelism

The Age of Internet Computing

Supercomputer Sites

The Age of Internet Computing

Supercomputer Sites

Large Data Centres

The Age of Internet Computing

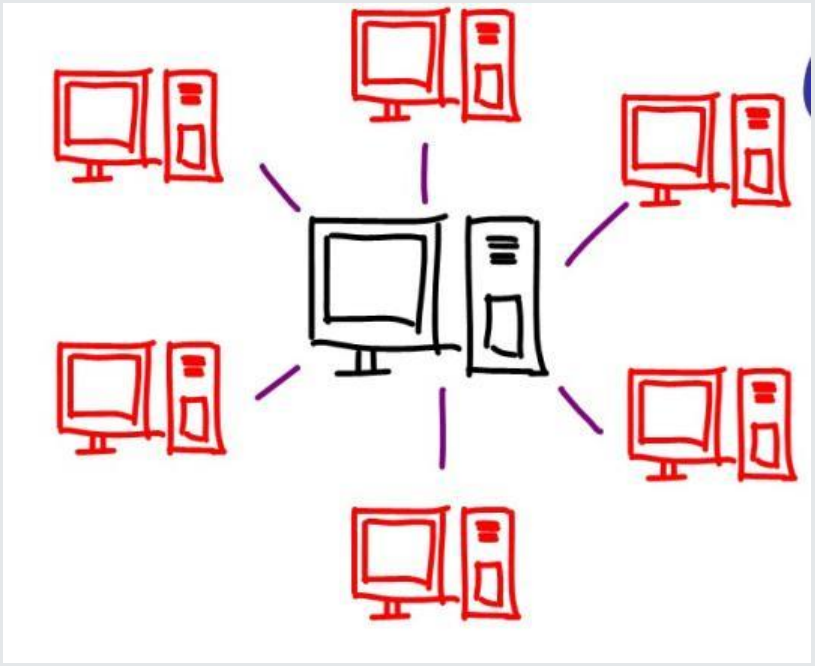
Supercomputer Sites

Large Data Centres

Linpack Benchmark

The Age of Internet Computing

Parallel Computing



Distributed Computing

The Age of Internet Computing

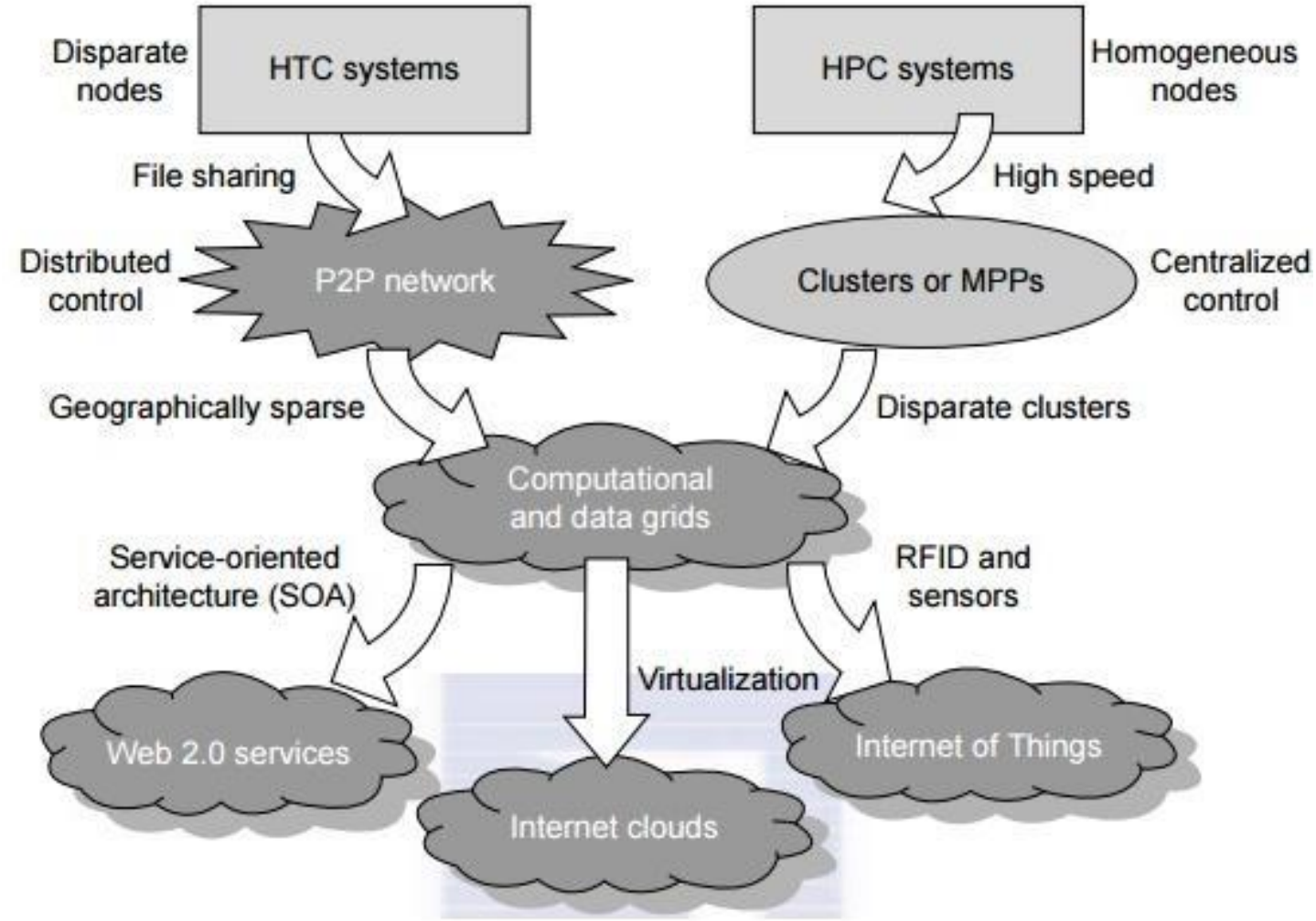
Data canters

Fast Server

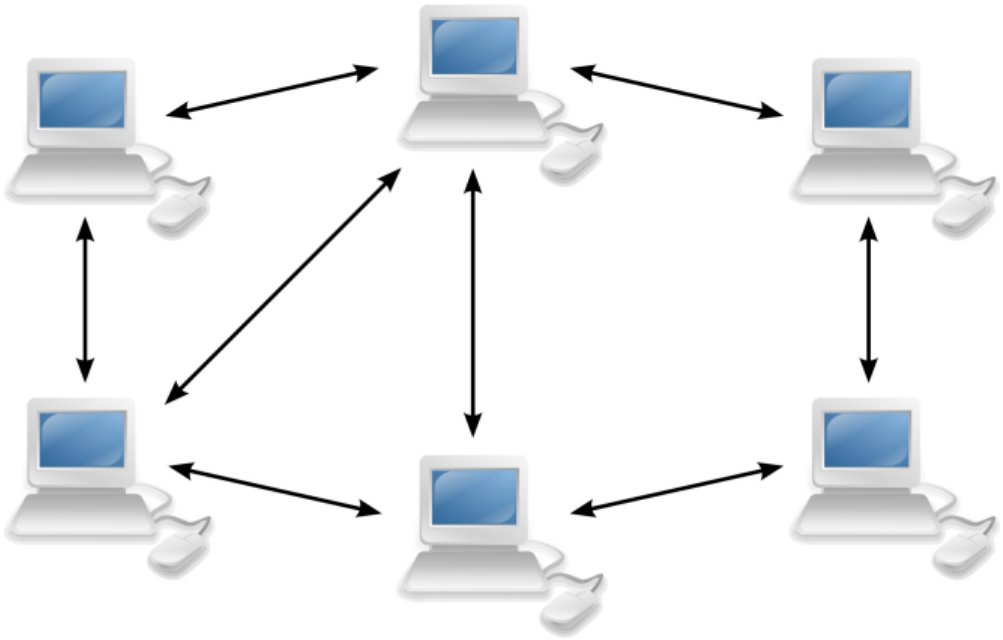
Storage systems

High-bandwidth networks.

The Platform Evolution



The Platform Evolution



Peer-to-Peer Network

High-Performance Computing

- For many years, HPC systems emphasize the raw speed performance
- The speed of HPC systems has increased from Gflops in the early 1990s to now Pflops in 2010.
- This improvement was driven mainly by the demands from scientific, engineering, and manufacturing communities.

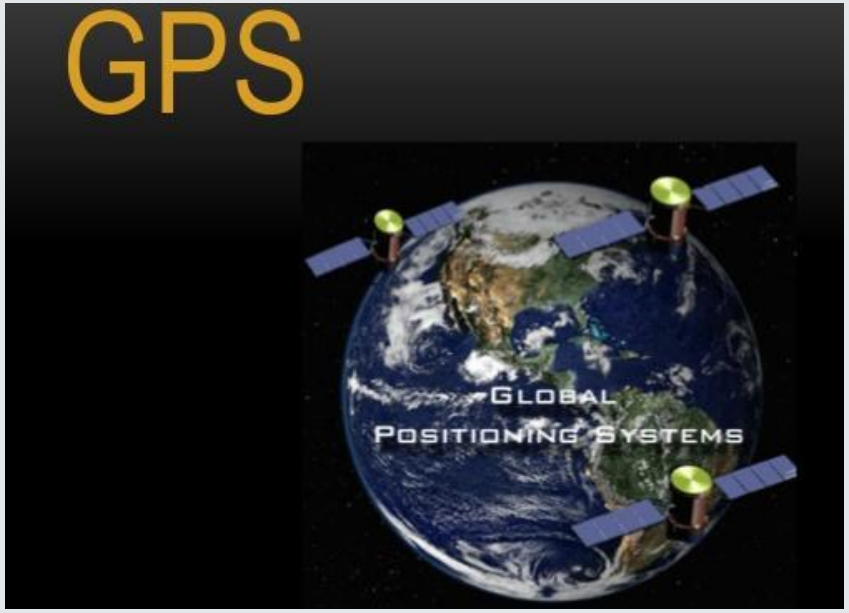
High-Throughput Computing

- The development of market-oriented high-end computing systems is undergoing a strategic change from an HPC paradigm to an HTC paradigm.
- This HTC paradigm pays more attention to high-flux computing
- The main application for high-flux computing is in Internet searches and web services by millions or

High-Throughput Computing

- The performance goal thus shifts to measure high throughput or the number of tasks completed per unit of time.
- HTC technology needs to not only improve in terms of batch processing speed, but also address the acute problems of cost, energy savings, security, and reliability at many data and enterprise computing centers.

Three New Computing Paradigms



Sensor technologies

Three New Computing Paradigms

Computing Paradigm Distinctions

- The high-technology community has argued for many years about the precise definitions of centralized computing, parallel computing, distributed computing, and cloud computing.
- Distributed computing is the opposite of centralized computing.
- The field of parallel computing overlaps with distributed computing to a great extent, and cloud

Distributed System Families

- Since the mid-1990s, technologies for building P2P networks and networks of clusters have been consolidated into many national projects designed to establish wide area computing infrastructures, known as computational grids or data grids.
- Recently, we have witnessed a surge in interest in exploring Internet cloud resources for data-intensive applications.
- Internet clouds are the result of moving desktop computing to service-oriented computing using server

Degrees of Parallelism

- Fifty years ago, when hardware was bulky and expensive, most computers were designed in a bit-serial fashion.
- Over the years, users graduated from 4-bit microprocessors to 8-,16-, 32-, and 64-bit CPUs. This led us to the next wave of improvement, known as instruction-level parallelism (ILP), in which the processor executes multiple instructions simultaneously rather than only one instruction at a time.

Degrees of Parallelism

- In this scenario, bit-level parallelism (BLP) converts bit-serial processing to word-level processing gradually.
- For the past 30 years, we have practiced ILP through pipelining, superscalar computing, VLIW (very long instruction word) architectures, and multithreading.
- ILP requires branch prediction, dynamic scheduling, speculation, and compiler support to work efficiently


That's all for now...