

Unit-2 : Data Storage and Cloud Computing

[7Hrs]

Data Storage: Introduction to Enterprise Data Storage, Direct Attached Storage, Storage Area Network, Network Attached Storage, Data Storage Management, File System, Cloud Data Stores, Using Grids for Data Storage.

Cloud Storage: Data Management, Provisioning Cloud storage, Data Intensive Technologies for Cloud Computing. Cloud Storage from LANs to WANs: Cloud Characteristics, Distributed Data Storage.

Case Study: Online Book Marketing Service, Online Photo Editing Service



Google Cloud :File system

- Google Cloud Storage (GCS): Google Cloud Storage is an object storage service that offers scalable, durable, and highly available storage for large volumes of unstructured data. It is designed to store and serve a wide variety of data types, including images, videos, backups, and archives. Google Cloud Storage offers different storage classes, including Standard, Nearline, Coldline, and Archive, allowing users to optimize storage costs based on their data access requirements.
- Google Cloud Filestore: Google Cloud Filestore provides fully managed Network Attached Storage (NAS) for applications that require shared file storage.



Data Storage

- **Data storage:** Files and documents are recorded digitally and saved in a storage system for future use
- A huge amount of data is continuously generated, collected, stored, and analyzed through software.
- The most prevalent forms of data storage are **file storage, block storage, and object storage**, with each being ideal for different purposes.

8 FREE CLOUD STORAGE YOU SHOULD KNOW



MEGA
50 GB FREE



DRIVE
15 GB FREE



MEDIA FIRE
10 GB FREE



PCLOUD
10 GB FREE



ICE DRIVE
10 GB FREE



ONE DRIVE
5 GB FREE



ICOULD
5 GB FREE



SYNC.COM
5 GB FREE



Data Storage

- **File storage**
 - In file storage, **data is stored in files, the files are organized in folders**, and the folders are organized under a hierarchy of directories and subdirectories.
 - To locate a file, all you or your computer system need is the path—from directory to subdirectory to folder to file.
 - If you need to store very large or unstructured data volumes, you should consider block-based or object-based storage
 - **Example:** Harddrive, google drive etc.



Data Storage

- **Block Storage:**
- Block storage **breaks a file into equally-sized chunks** (or **blocks**) of data and stores each block separately under a unique address.
- Rather than conforming to a rigid directory/subdirectory/folder structure, blocks can be stored anywhere in the system.
- To access any file, the server's operating system **uses the unique address** to pull the blocks back together into the file, which **takes less time** than navigating through directories and file hierarchies to access a file.
- Example: Block Storage are SAN, iSCSI, and local disks.

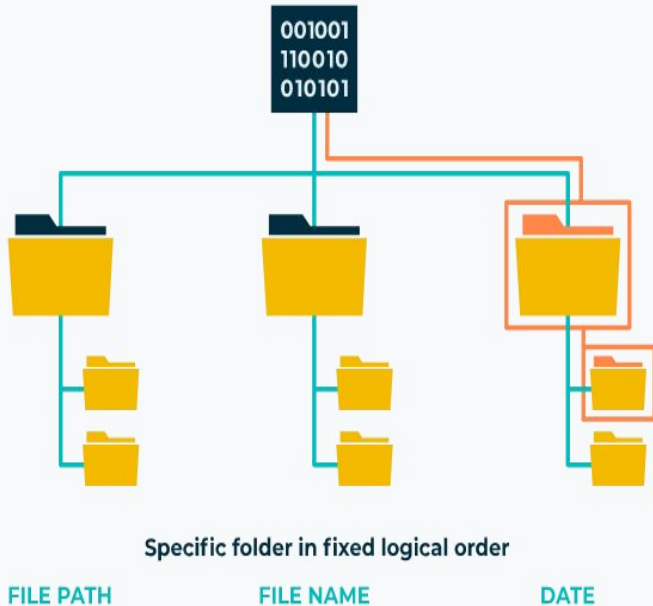


Data Storage

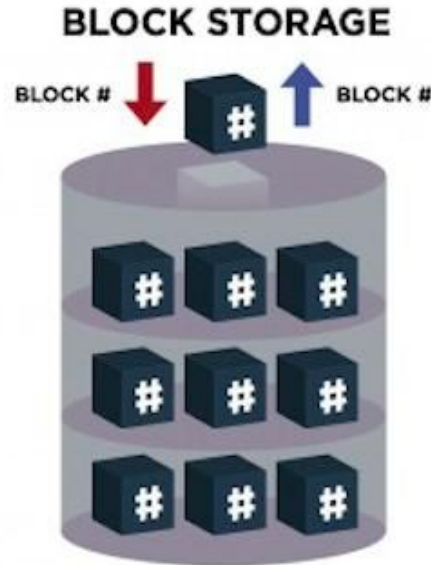
- **Object Storage:**
- unstructured media and web content like email, videos, image files, web pages, and sensor data produced by the Internet of Things (IoT).
- object is a simple, self-contained repository that includes the **data**, **metadata** (descriptive information associated with an object), and a **unique identifying ID number**.
- This information enables an application **to locate and access the object**.
- **Example:** storing objects like **videos and photos on Facebook**, songs on Spotify, or files in online collaboration services, such as Dropbox

Different Storage Types

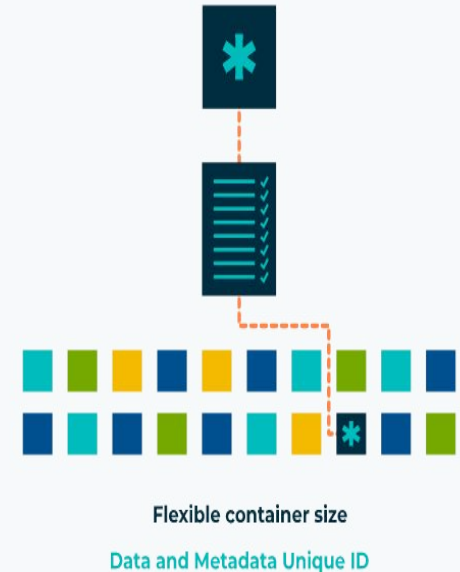
File Storage



Block Storage



Object Storage





Data Storage Challenges

- **Some challenges are :**
- massive data demand
- performance barrier
- power consumption and cost.



Data Storage Challenges

- **Massive Data Demand**
- An industry survey estimates the digital world to increase by 45 zettabytes by 2020, that is, one terabyte is equal to 10^3 gigabytes, one petabytes is equal to 10^3 terabytes, one exabytes is equal to 10^3 petabytes and one zettabytes is equal to 10^3 exabytes.



Data Storage Challenges

- **Performance Barrier**
- Rapid growth in data has caused a parallel increase in the size of databases.
- In the traditional storage method, the response time taken for queries is slow and it should be increased.
- Be it a social networking site, an enterprise database or a web application, all requires faster disk access to read and write data.



Data Storage Challenges

- **Power Consumption and Cost**
- Because of increase in storage demands, IT organizations and data centres **need larger storage with minimal cost.**
- Performance lags with minimal cost but has other expenses like licensing and maintenance.
- Apart from this, other factors such as power consumed by storage devices, cooling systems, man power for managing it and space for data centres are to be considered.



Introduction to Enterprise Data Storage,

- An **Enterprise Storage** System is a centralized repository for business information.
- Enterprise data storage allows businesses to store and access large volumes of company information.
- The size of data that businesses can store depends on the storage type they use.



Introduction to Enterprise Data Storage,

- For most large companies, a good data storage platform is essential to security and success
- Business **requires** that huge amounts of data be stored safely but also be **easily accessible**.
- *Enterprise storage is a **centralized repository** for business information that **provides** common data management, protection and data sharing functions through connections to computer systems.*



Introduction to Enterprise Data Storage,

- Main types of **Enterprise(business) Data storage**

1 Direct Attached Storage(DAS)

2 Storage Area Network(SAN)

3 Network Attached Storage(NAS)

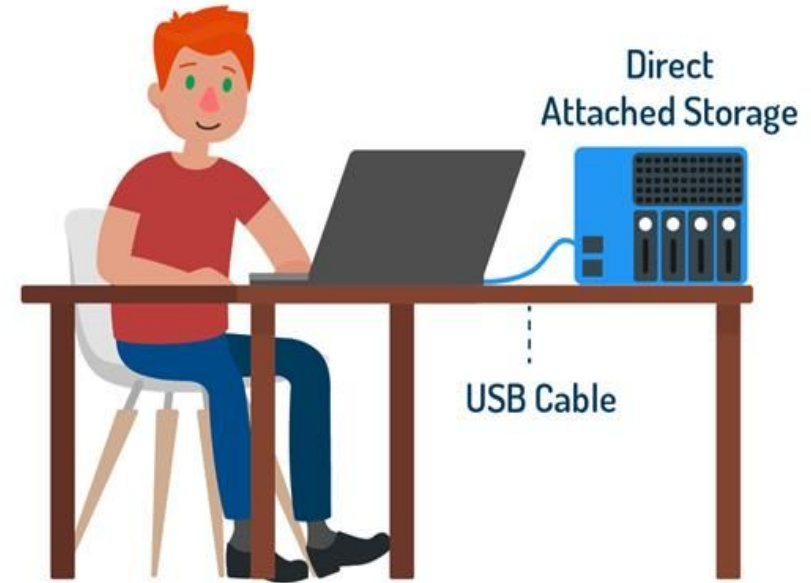


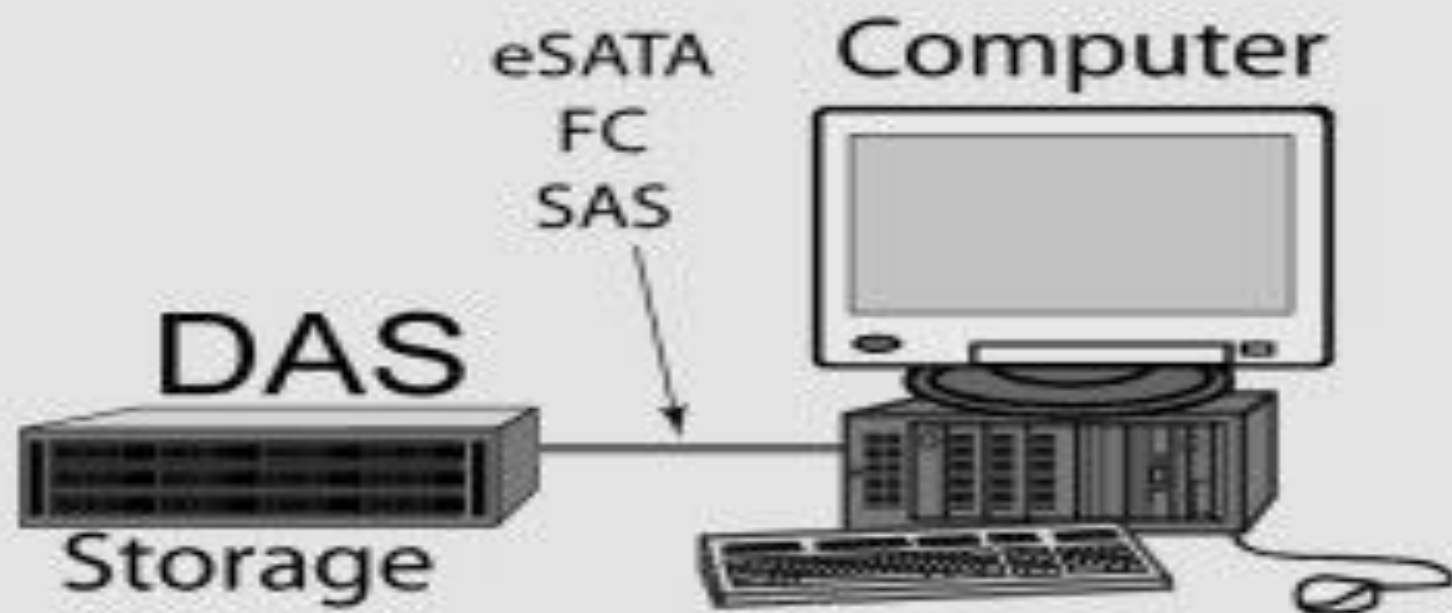
1 Direct Attached Storage(DAS)

- **Introduction of DAS**

- **Advantage of DAS**

- **Disadvantage of DAS**





Direct Attached Storage (DAS)



1 Direct Attached Storage(DAS)

● Introduction of DAS

- Direct-attached storage (DAS) is a type of storage that is attached directly to a computer without going through a network.
- The storage might be connected **internally or externally**.
- Only the **host computer can access the data directly**.
- Most servers, desktops and laptops contain an internal **hard disk drive (HDD)** or solid-state drive (SSD).
- Some computers also use external DAS devices.



1 Direct Attached Storage(DAS)

- **Introduction of DAS**

- In some cases, an enterprise server might **connect directly to drives** that are shared by other [servers](#).
- A direct-attached storage device is **not networked**.
- An **external DAS device** connects directly to a computer through an interface such as Small Computer System Interface (**SCSI**), Serial Advanced Technology Attachment (**SATA**), Serial-Attached SCSI (**SAS**), FC or Internet SCSI (**iSCSI**).



1 Direct Attached Storage(DAS)

- **Advantage of DAS**
- DAS can provide users with better performance than networked storage because the server does not have to traverse a network to read and write data, which is why many organizations turn to DAS for applications that require high performance.
- DAS is also less complex than network-based storage systems, making it easier to implement and maintain, and it is cheaper.



1 Direct Attached Storage(DAS)

- **Disadvantage of DAS**
- It has limited scalability
- lacks the type of centralized management
- backup capabilities available to other storage platforms.
- it can't be easily shared data

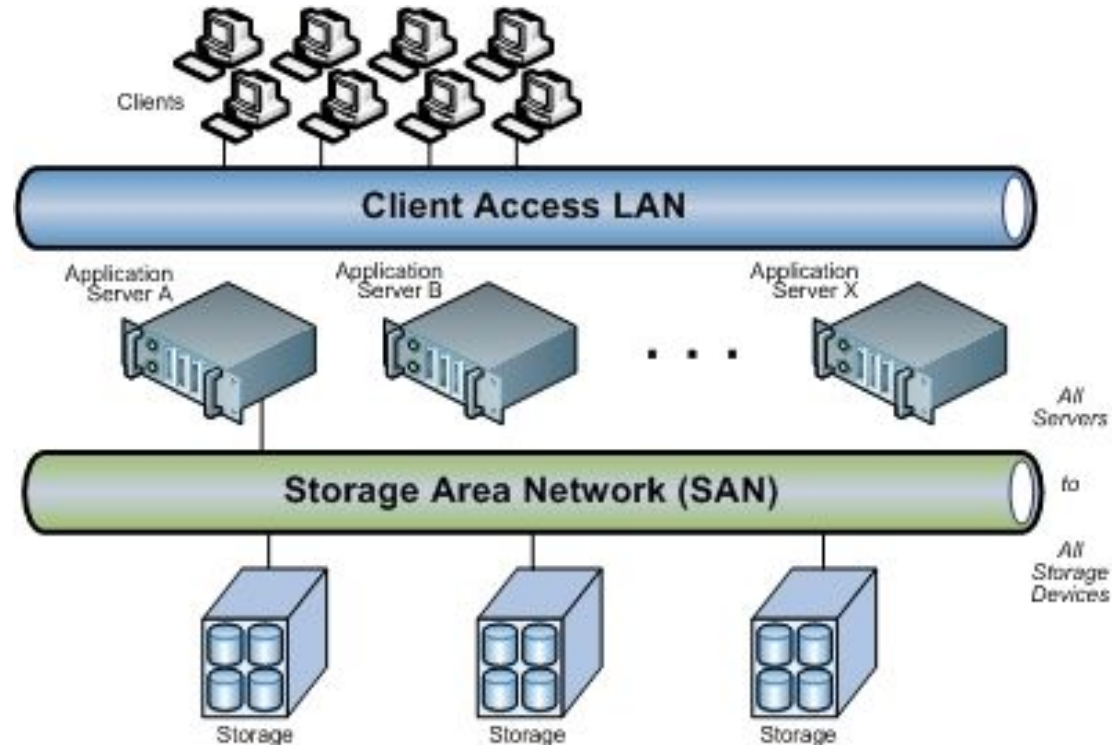


2 Storage Area Network(SAN)

● Introduction of SAN

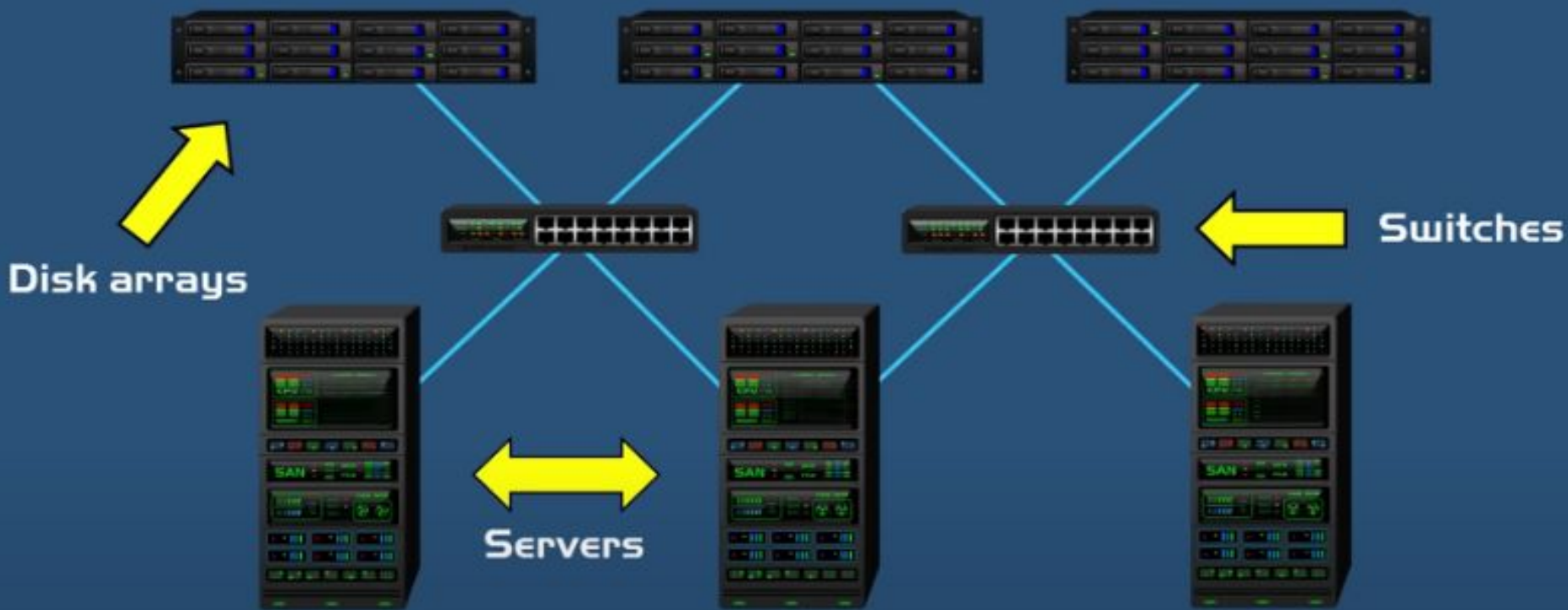
● Advantage of SAN

● Disadvantage of SAN



SAN

STORAGE AREA NETWORK





2 Storage Area Network(SAN)

- **Introduction of SAN**
- A **Storage Area Network** (SAN) is a specialized, high-speed network that **provides network access** to storage devices.
- SANs are typically **composed** of **hosts, switches, storage elements, and storage devices** that are interconnected using a variety of technologies, topologies, and protocols.



2 Storage Area Network(SAN)

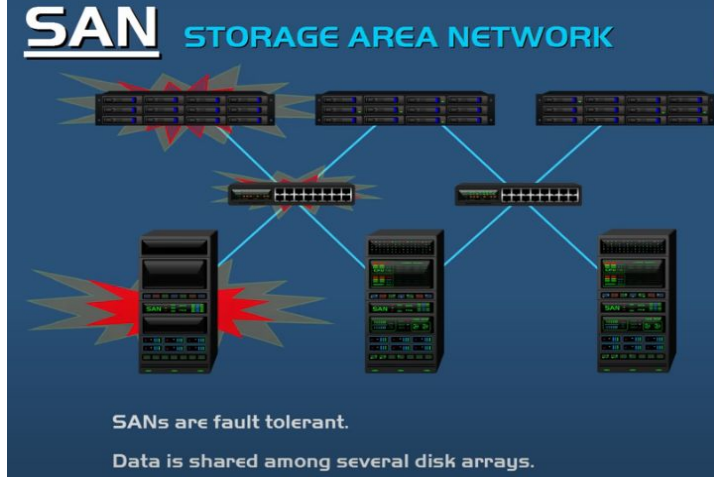
- **Introduction of SAN**
- **Traditionally**, only a limited number of storage devices could attach to a server, **limiting** a network's storage capacity.
- But a **SAN introduces** networking flexibility **enabling one server**, or **many heterogeneous servers** across multiple data centers, **to share a common storage utility**.



2 Storage Area Network(SAN)

- **Advantage of SAN**

- Simplified storage administration
- Disk mirroring
- Low cost of storage management
- Instant and real-time information
- Ability to boot itself and expand the storage capacity
- SAN is not directly attached to any particular server or network, SAN can be shared by all





2 Storage Area Network(SAN)

● Disadvantage of SAN

- If client computers need intensive data transfer then SAN is not the right choice. SAN is good for low data traffic
- More expensive
- It is very hard to maintain
- As all client computers share the same set of storage devices so **sensitive data can be leaked**. It is preferable not to store confidential information on this network.
- Poor implementation results in a performance bottleneck
- Not affordable for small business
- Require a high-level technical person

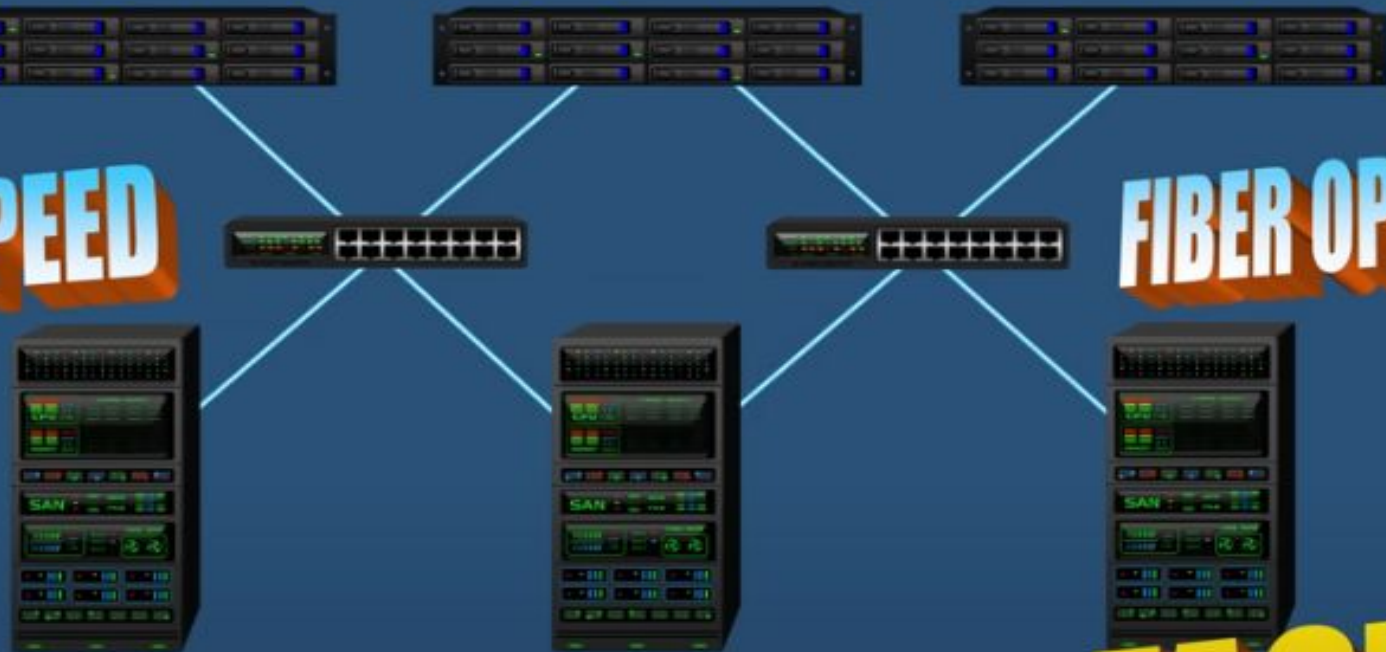
SAN

STORAGE AREA NETWORK



HIGH SPEED

FIBER OPTICS



Interconnected using **Fibre Channel**.

Speeds between **2 Gbit/s** – **128 Gbit/s**.

FAST!



3

Network Attached Storage(NAS)

Introduction of NAS

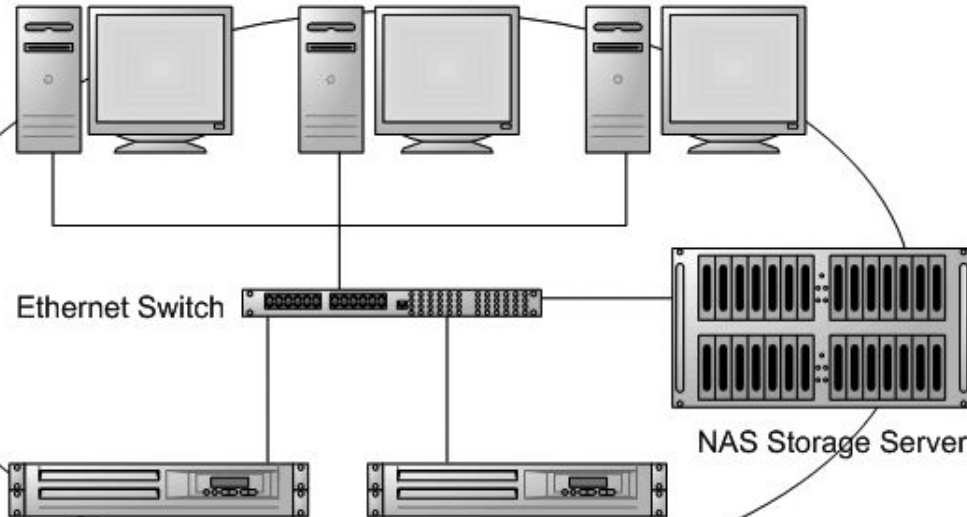
Advantage of NAS

Disadvantage of NAS

Network Attached Storage

Clients

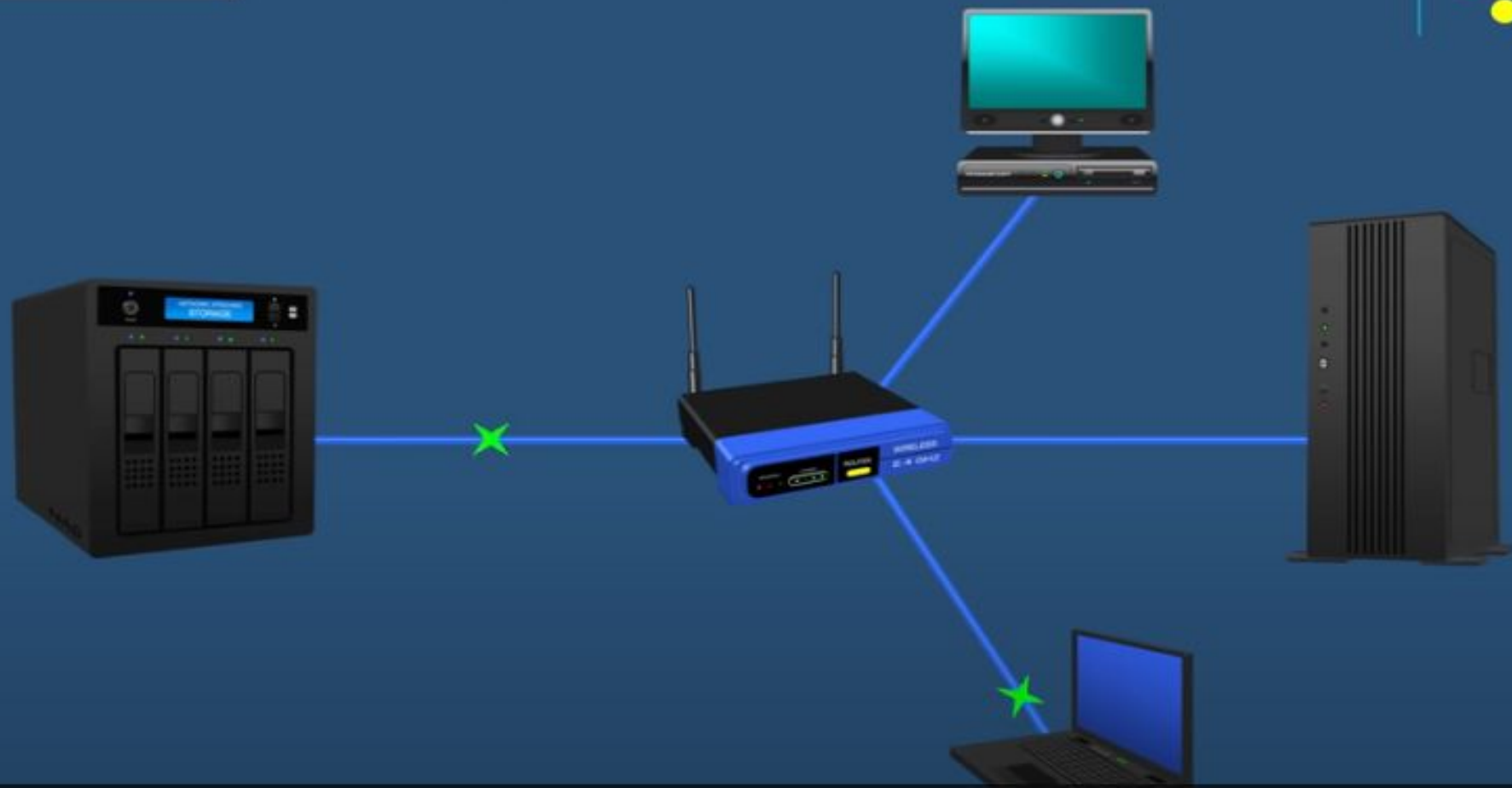
LAN



Servers

NAS

NETWORK ATTACHED STORAGE





3

Network Attached Storage(NAS)

Introduction of NAS

- An NAS device is a storage device **connected to a network** that allows **storage and retrieval of data** from a central location for authorised network users and varied clients.
- NAS is a centralized, file server, which **allows multiple users to store and share files over a TCP/IP network via Wifi or an Ethernet cable.**
- It is also commonly known as a **NAS box, NAS unit, NAS server, or NAS head.**



3

Network Attached Storage(NAS)

Introduction of NAS

- **Network Protocols:** TCP/IP protocols –i.e. Transmission Control Protocol (TCP) and Internet Protocol (IP)—are used for data transfer, but the network protocols for data sharing can vary based on the type of client.



Network Attached Storage(NAS)

● Advantage of NAS



Simple to operate, a dedicated IT professional is often not required

- Lower cost
- Easy data backup, so it's always accessible when you need it
- Good at centralising data storage in a safe, reliable way

● Disadvantage of NAS

- Out-of-sync data
- Reliability and accessibility issues if storage goes down

STORAGE TYPES COMPARISON

DAS

Servers



Storage Disks



NAS

Servers



Storage Disks



SAN

Servers



Storage Disks



	DAS	NAS	SAN
Protocols	SATA, SAS	SMB (CIFS), NFS	iSCSI, Fiber Channe
Type	File Storage	Block Storage	Block Storage
Speed	5-10ms	5-10ms	20-50ms
Data Transmission	IDE/SCSI	Ethernet, TCP/IP	FiberChannel
Complexity	Easy	Moderate	Complex
Management Costs	High	Moderate	Low
Storage Type	Sectors	Shared Files	Blocks
Supports Capacity Sharing?	No (possible manually)	Yes	Yes
Connected to Network or not?	No	Yes	Yes
Scalability	Low	Medium	High



Data Storage management

- Data storage Management tool must rely on policies which govern the usage of storage devices .
- Data Storage management refers to **the software and processes that improve the performance of data storage resources.**
- It may include network virtualization, replication, mirroring, security, compression, deduplication, traffic analysis, process automation, storage provisioning and memory management.



Data Storage management

- Storage management makes it possible to reassign storage capacity quickly as business needs change.
- Storage management techniques can be applied to **primary, backup or archived storage.**
- Primary storage** holds actively or frequently accessed data
- Backup storage** holds copies of primary storage data for use of disaster recovery
- Archived storage** holds outdated or rarely used data that must be retained for compliance or business continuity.



Data Storage management

- Storage provisioning is a management technique that assign storage capacity to servers,computer,virtual machines and other devices.
- It may **use** automation to allocate storage space in a networked environment.
- Intelligent storage management** uses software policies and algorithm to automate the provisioning (and de-provisioning) of storage resources,continuously monitoring data utilisation and re-balancing data placement without human intervention.



Cloud File System

- Introduction
- Ghost File System:Rokade
- Gluster File System:Kirti
- Hadoop File System:Sahil deshमुख
- XtreamFS: A Distributed and Replicated File System:Aboli
- Kosmos File System:Nishigandha
- CloudFS:Yash
- Google File system(GFS):Vashinavi



Cloud File System :Introduction

- A file system is a structure used in computer(o.s.) to store data on a hard disk.
- The file system is responsible for organizing files and directories, and keeping track of which areas of the media belong to which file and which are not being used.
- When we install a new hard disk, we need to partition and format it using a file system before storing data.
- Following file systems in use in **Windows OS**; **NTFS**(New Technology File System),**FAT32**(**F**ile **a**llocation **t**able),**EXT4**(most common **L**inux **f**ile **s**ystem) etc



Cloud File System :Introduction

FAT:

- FAT was planned for systems with very small RAM and small disks. It required much less system resources compared to other file systems like UNIX.

NTFS

- NTFS is much simpler than FAT.
- While files are used, the system areas can be customized, enlarged, or moved as required. NTFS has much more security incorporated.
- NTFS is not apt for small-sized disks.



Cloud File System :Introduction

- File system typically provide mechanism for **reading,writing ,modifying,deleting or organising files in folders and directories,**
- **Cloud file system** are specifically designed to be **distributed and operated in the cloud based environment.**



Cloud File System

- In cloud file systems, the **considerations** are:
- It must sustain basic file system functionality.
- It should be an open source.
- It should be grown-up enough that users will at least think about trusting their data to it.
- It should be shared, i.e., available over a network.
- It should be paralleling scalable.
- It should provide honest data protection, still on commodity hardware with only internal storage.



Cloud File System :Ghost File System

- **Ghost cloud file system** is used in [Amazon Web Services \(AWS\)](#).
- It gives high redundant elastic mountable, cost-effective and standards-based file system.
- A fully featured scalable and stable cloud file systems is provided by **ghost cloud file system**.
- GFS (Ghost File System) run over [Amazon's S3, EC2 and SimpleDB web services](#).
- When using GFS, [user can have complete control of the data](#) and [can be accessed as a standard network disk drive](#).



Cloud File System : Ghost File System

- **Features of Ghost CFS**

- Mature elastic file system in the cloud.
- All files and metadata duplicated across multiple AWS availability regions.
- FTP access.
- Web interface for user management and for file upload/download.
- File name search.
- Torrents are a method of distributing files over the internet.
- WebDav: WebDAV protocol provides a framework for users to create, change and move documents on a server
- Sideload describes the process of transferring files between two local devices,



Cloud File System :Ghost File System

- **Benefits of Ghost CFS**
- **Elastic and cost efficient:** Pay for what you use from 1 GB to hundreds of terabytes.
- **Multi-region redundancy:** Aiming to take advantage of AWS's 99.99% availability
- **Highly secure:** Uses your own AWS account (ghost cannot access your data).
- **No administration:** Scales elastically with built in redundancy—no provisioning or backup.
- **Anywhere:** Mount on a server or client or access files via a web page or from a mobile phone.



Cloud File System :Gluster File System



- GlusterFS is an open source, distributed file system capable of handling multiple clients and large data.
- **GlusterFS** clusters storage devices over network, aggregating disk and memory resources and managing data as a single unit.
- GlusterFS is **based** on a stackable user space design and delivers good performance for even heavier workloads.
- GlusterFS supports clients with valid IP address in network.



Cloud File System : Gluster File System



- Users no longer locked with legacy storage platforms which are costly and monolithic.
- GlusterFS gives users the ability to deploy scale-out, virtualized storage, centrally managed pool of storage.
- **Attributes of GlusterFS** include scalability and performance, high availability, global namespace, elastic hash algorithm, elastic volume manager, gluster console manager, and standards-based.

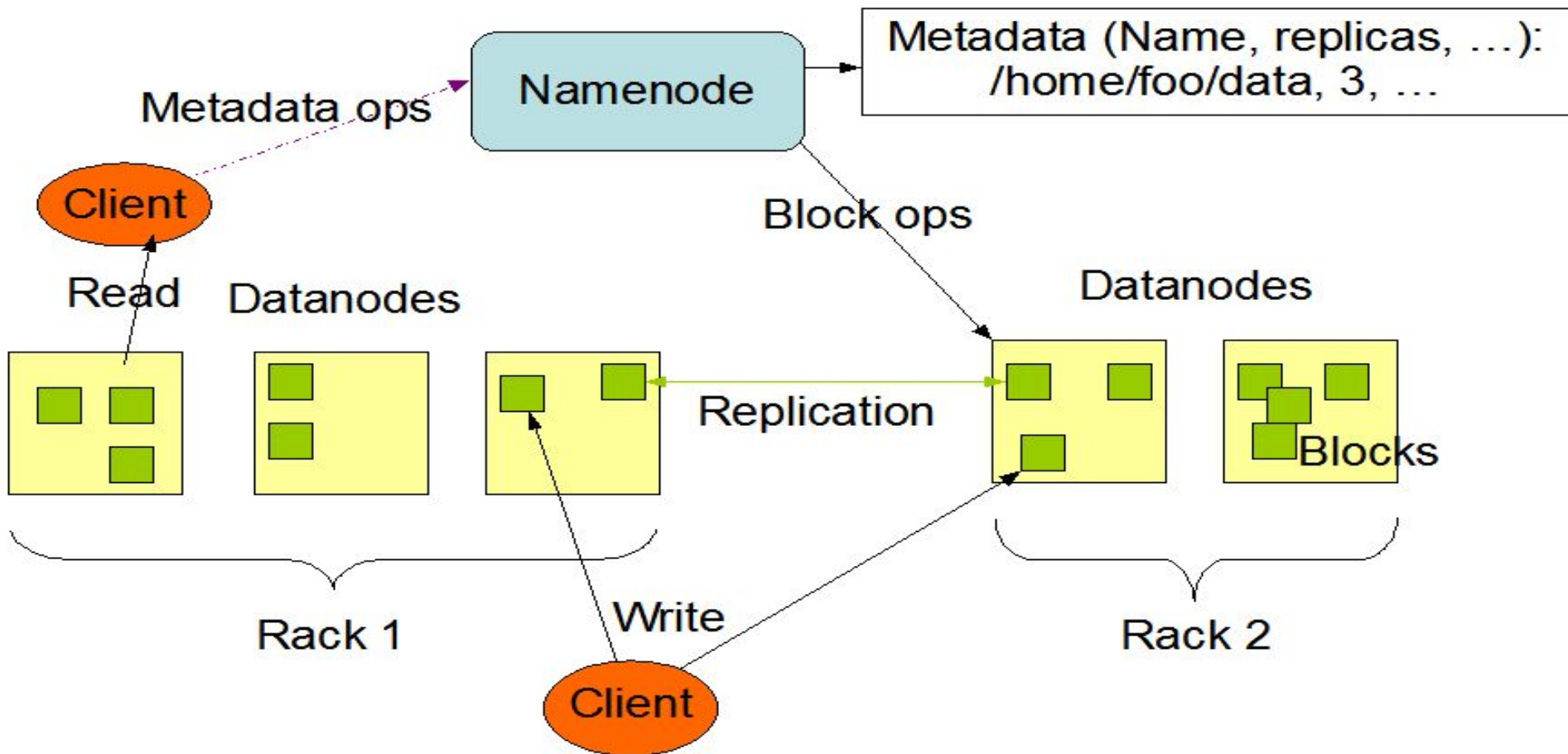


Cloud File System :Hadoop File Syst



- A distributed file system designed to run on commodity hardware is known as Hadoop Distributed File System (HDFS).
- In HDFS, files are stored in **blocks ranging** from **64 MB to 1024 MB**.
- The default size is 64 MB.
- The blocks will be distributed across the cluster and replicated for fault tolerance.

HDFS Architecture





Cloud File System :XtreemFS

- **XtreemFS:** A Distributed and Replicated File System
- **XtreemFS** is a distributed, replicated and open source.
- XtreemFS allows **users to mount and access files** via WWW.
- Engaging XtreemFS a user can **replicate the files across data centres** to reduce network congestion, latency and increase data availability.
- Installing XtreemFS is quite easy, but **replicating the files is bit difficult.**



Cloud File System :Kosmos File Syst



- Kosmos Distributed File System (KFS) gives **high performance with availability and reliability.**
- **For example**, search engines, data mining, grid computing, etc.
- It is deployed in C++ using standard system components such as STL, boost libraries, aio, log4cpp.
- KFS is incorporated with **Hadoop and Hypertable.**



Cloud File System : CloudFS

- CloudFS is a distributed file system to solve problems when **file system is itself provided as a service.**
- CloudFS is based on **GlusterFS**, a basic distributed file system, and supported by **Red Hat** and hosted by **Fedora**.
- There are really **three** production level distributed/parallel file systems that come close to the requirements for the cloud file systems: **Lustre**(Linux and cluster.), **PVFS2**(The Parallel Virtual File System (**PVFS**)) **and GlusterFS**.



Cloud File System: Google file System



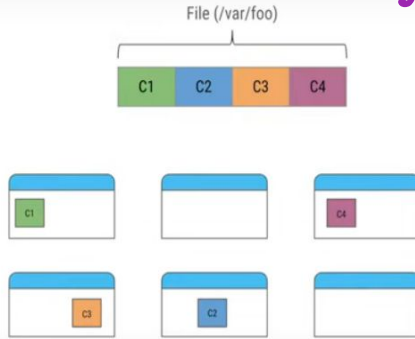
- It is **scalable distributed file system** for large distributed data-intensive applications.
- Developed by google
- Single file is not stored into single server, files are divided into multiple chunk
- GFS master only read the metadata of file
- Publish paper <https://static.googleusercontent.com/media/research.google.com/en/archive/gfs-sosp2003.pdf>

Cloud File System: Google file System

Chunks

Files split into chunks

- Each chunk of 64MB
- Identified by 64 bit ID
- Stored in Chunkservers

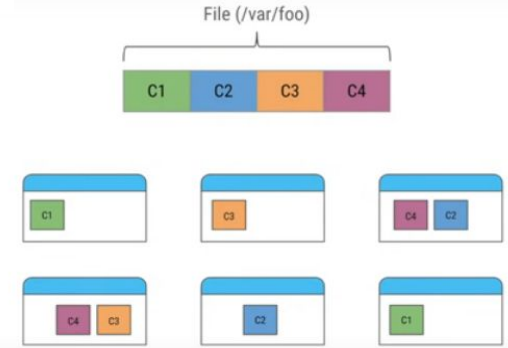


Chunkservers - Chunks of single file are distributed on multiple machines

Replicas

Files split into chunks

- Replica count by client
- commodity server failures



Replicas ensure durability of data if chunkserver goes down

GFS master

- Stores entire metadata of the cluster
- File names + chunk ids + chunk locations
- Access Control details



Chunkservers

GFS master

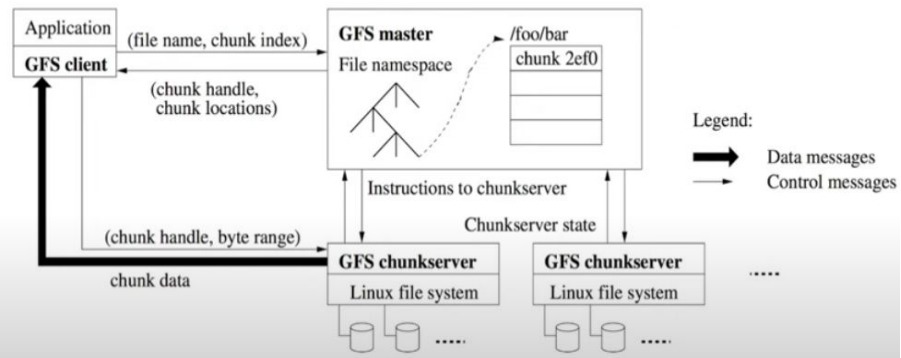
Client application

File	Chunks	Locations
/var/foo	ffe0	server1 (replica 1) server2 (replica 2)
	ff21	server 4 (replica 1)



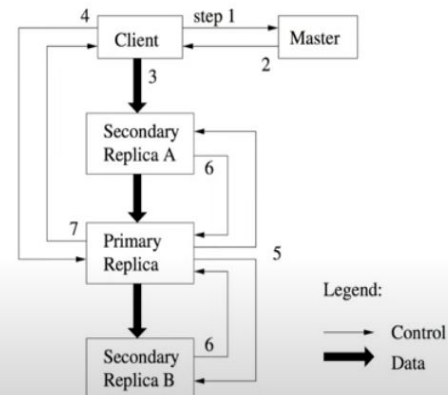
Chunkservers

Reads



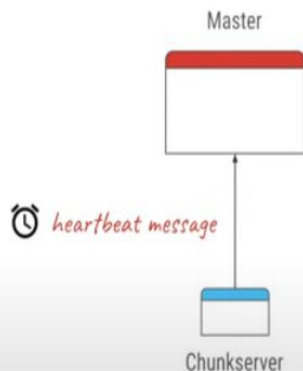
Writes

1. Ask for locations to write
2. Get replicate locations
3. Write data to closest replica.
4. Request commit to primary
5. Primary instructs order of writes to secondaries
6. Secondaries acknowledge
7. Primary ack to client



Heartbeats

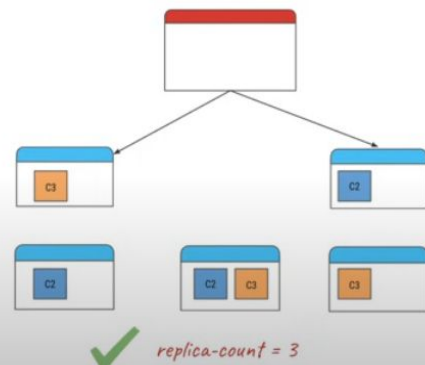
Regular heartbeats to ensure chunkservers are alive



Ensure chunk replica count

If chunkserver is down, master ensures all chunks that were on it are copied on other servers.

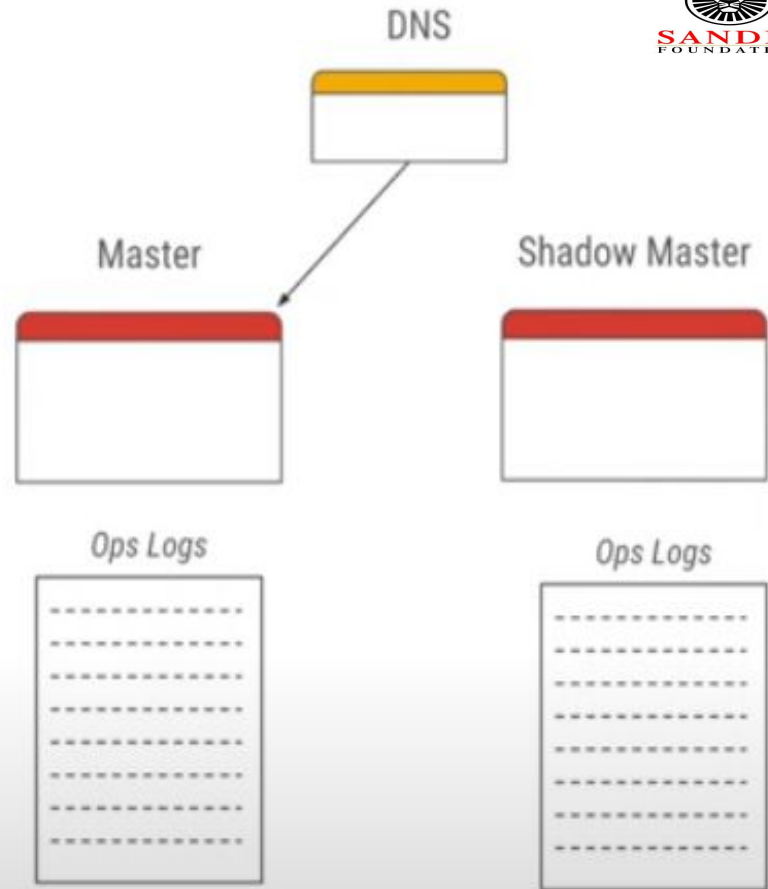
Ensures replica counts remains same.



Shadow Master

Single Point of Failure

- Ops log is replicated remotely
- Shadow master uses the logs
- DNS change can change master
- Shadow master may lag slightly





Cloud Data Stores

- What is Data stores?
- Distributed Data Store
- Types of Data Stores :BigTable,Dynamo:



Popular Cloud Data Stores

- **Amazon Web Services (AWS)** - Amazon DynamoDB, Amazon RDS (Relational Database Service), Amazon Redshift, Amazon Aurora, Amazon S3 (Simple Storage Service)
- **Microsoft Azure** - Azure Cosmos DB, Azure SQL Database, Azure Blob Storage, Azure Data Lake Storage
- **Google Cloud Platform (GCP)** - Google Cloud Bigtable, Google Cloud SQL, Google Cloud Storage, Google BigQuery
- **IBM Cloud** - IBM Db2 on Cloud, IBM Cloud Object Storage, IBM Cloudant
- **Oracle Cloud** - Oracle Database Cloud Service, Oracle NoSQL Database Cloud Service, Oracle Object Storage



Cloud Data Stores

- A data store is a data repository where data are stored as objects.
- Data store includes data repositories, flat files that can store data.
- **Data stores can be of different types:**
 - Relational databases (Examples: MySQL, PostgreSQL, Microsoft SQL Server, Oracle Database)
 - Object-oriented databases
 - Operational data stores
 - Schema-less data stores, e.g. Apache Cassandra or Dynamo
 - Paper files
 - Data files (spreadsheets, flat files, etc)



Cloud Data Stores: Distributed Data Store

- A Distributed Data Store is like a distributed database where users store information on multiple nodes.
- These kinds of data store are non-relational databases that searches data quickly over a large multiple nodes.
- **Examples** for this kind of data storage are **Google's BigTable, Amazon's Dynamo** and Windows Azure Storage.
- Some Distributed Data Stores use to recover the original file when parts of that file are damaged or unavailable by using forward error correction techniques.
- Others download that file from a diverse mirror.



Cloud Data Stores:Types of Data Stores:BigTable







- **Types of data stores:BigTable and Dynamo**
- **BigTable** is a compressed, high performance and proprietary data storage system construct on Google File System, Chubby Lock Service, SSTable and a small number of other Google technologies.
- **BigTable was developed in 2004** and is used in number of Google applications such as web indexing, Google Earth, Google Reader, Google Maps, Google Book Search, MapReduce, Blogger.com, Google Code hosting, Orkut, YouTube and Gmail.
- **Advantage** for developing BigTable includes scalability and better performance control.




Cloud Data Stores:Types of Data Stores:BigTable

- types of data stores:BigTable and Dynamo
- BigTable charts two random string values (row and column key) and timestamp into an **associated random byte array**.
- BigTable is designed **to scale into the petabyte** range across multiple machines and easy **to add more machines** and **automatically start using resources available without any configuration changes**.

Bigtable storage model

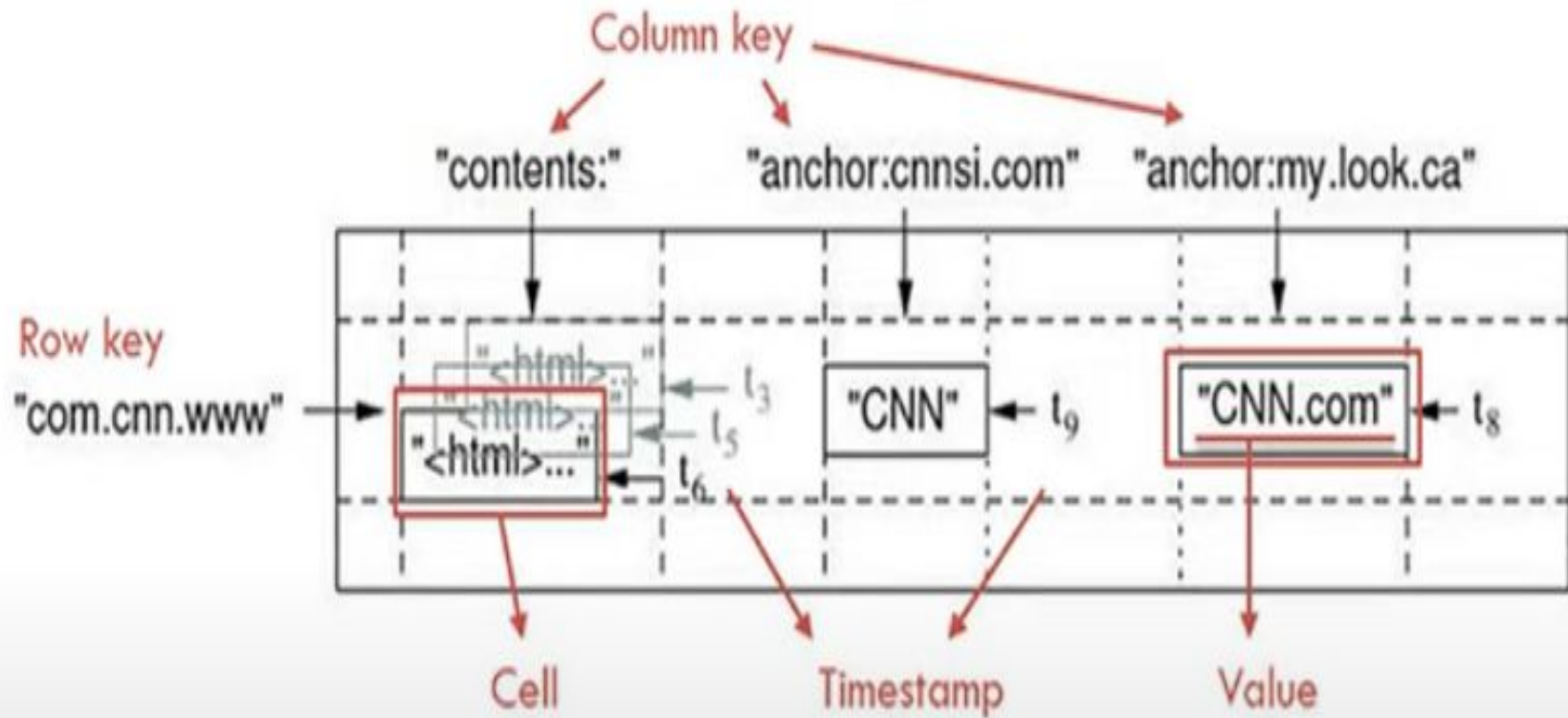
	Column family 1		Column family 2	
	<i>Column 1</i>	<i>Column 2</i>	<i>Column 1</i>	<i>Column 2</i>
Row key 1				
Row key 2				



t1

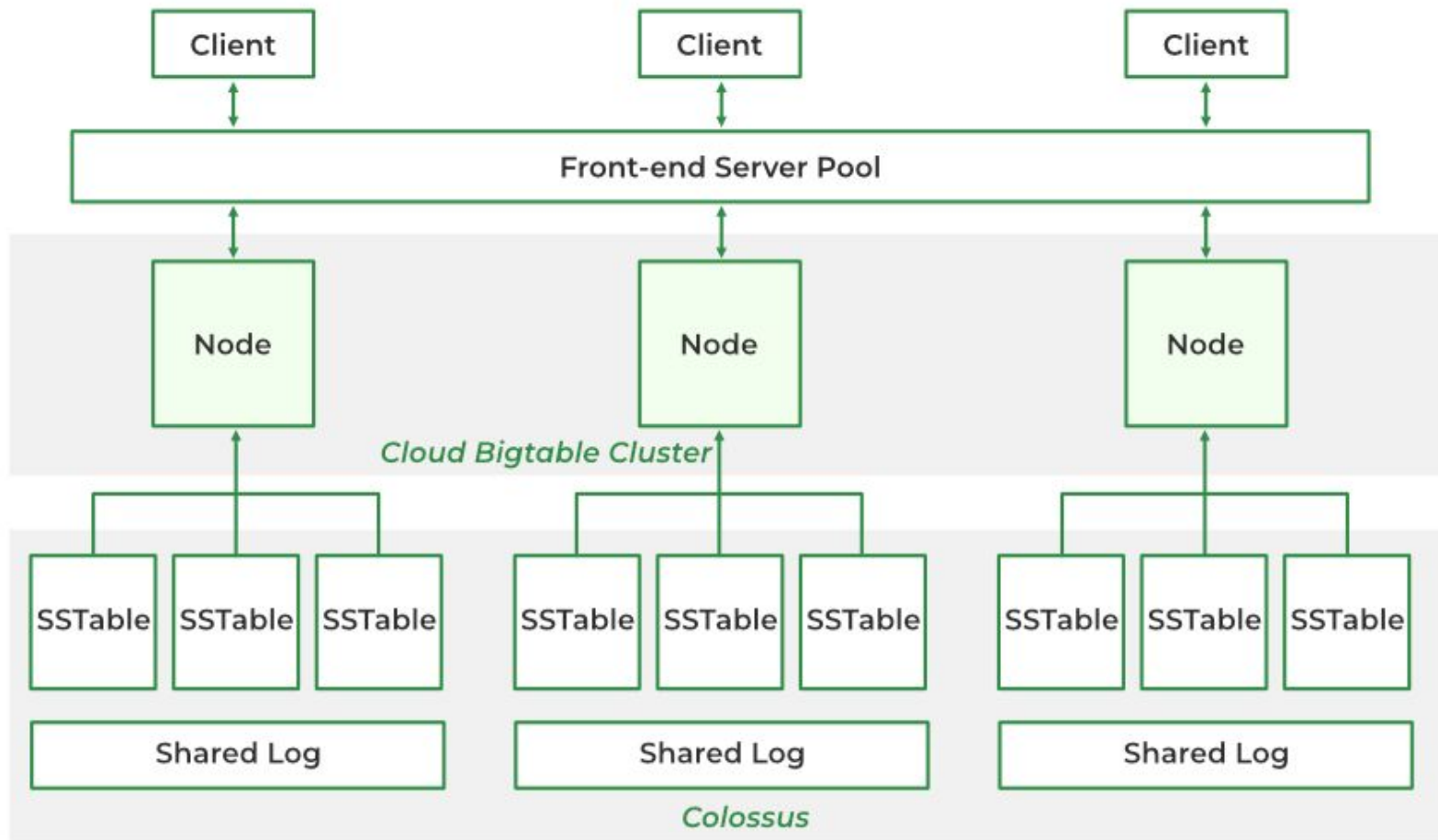
t2

t3



A Bigtable is a sparse, distributed, persistent multidimensional sorted map. The map is indexed by a row key, column key, and a timestamp; each value in the map is an uninterpreted array of bytes.

Bigtable architecture





Cloud Data Stores:Types of Data Stores:BigTable

- Other similar softwares are as follows:
- **Apache Accumulo:** Construct on top of Hadoop, ZooKeeper and economy. Server-side programming mechanism deployed in Java environment.
- **Apache Cassandra:** Dynamo's distributed design and BigTable's facts and numbers form adds simultaneously in Apache Cassandra, which uses Java.
- **Hbase:** Supports BigTable and Java programming language.
- **Hypertable:** Designed for cluster of servers especially for storage and processing.
- **KDI:** Kosmix stab to make a BigTable clone and is written in C++.



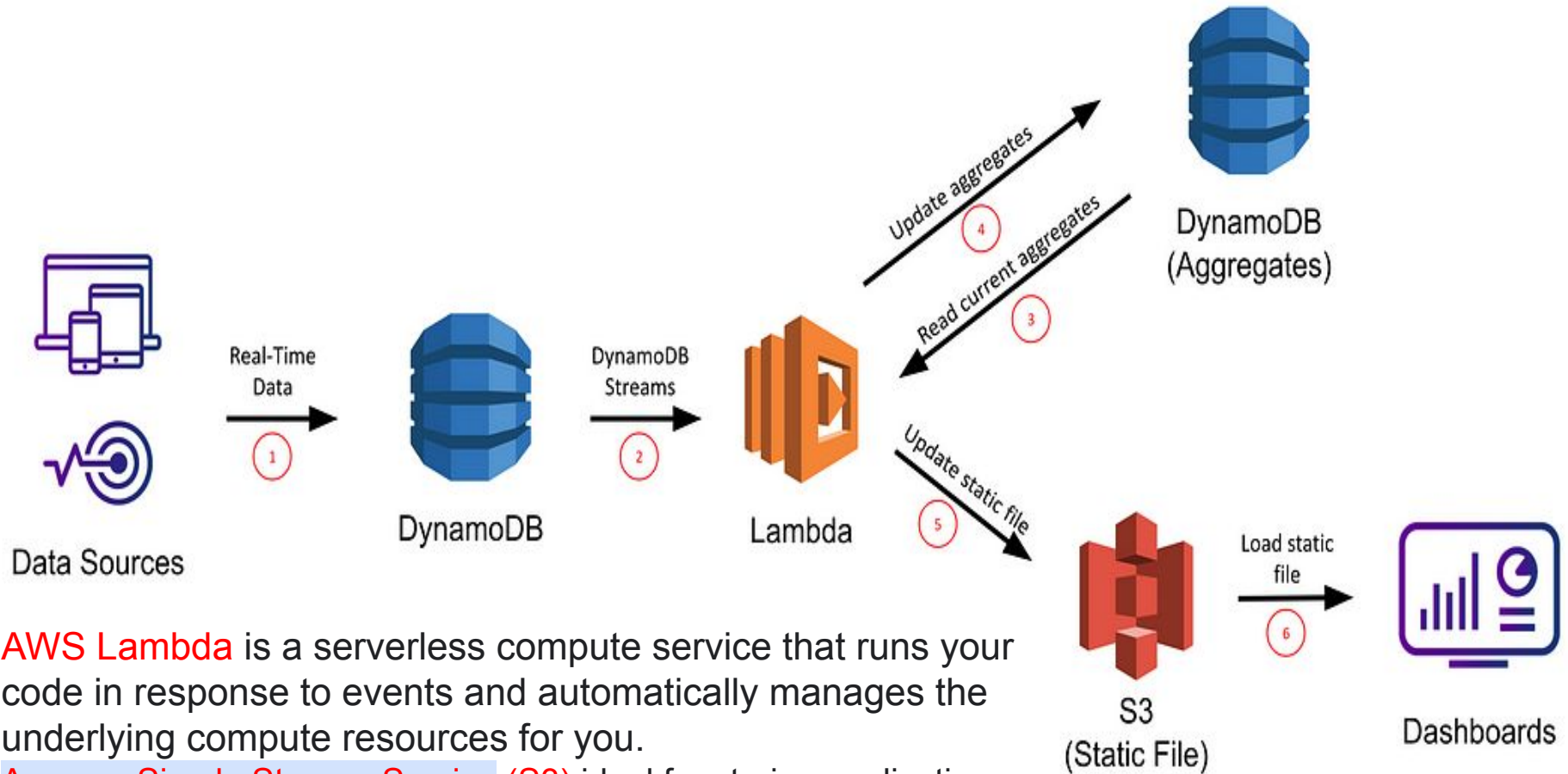
Cloud Data Stores:Types of Data Stores:Dynamo

- **Dynamo:** A Distributed Storage System
- Dynamo is a vastly offered, proprietary key-value structured storage system or a dispersed datastore. Dispersed storage systems are well-suited for storing unstructured data like digital media of all types
- It can act as databases and also **distributed hash tables** (DHTs).
- It is used with parts of Amazon web services such as Amazon S3.
- Dynamo is the most powerful **relational database available in World Wide Web**. (Relational databases have been used a lot in retail sites, to make visitors browse and search for products easily.)



Cloud Data Stores:Types of Data Stores:Dynamo

- It is difficult to create redundancy and parallelism with relational databases which is a single point failure.
- Replication is also not possible.
- Dynamo is a distributed storage system and not a relational database.
- Similar to a relational database it stores information to be retrieved; however, it stores the data as objects and not as tables.
- The **advantage** of using Dynamo is responsive and consistent in creating a distributed storage solution.



AWS Lambda is a serverless compute service that runs your code in response to events and automatically manages the underlying compute resources for you.

Amazon Simple Storage Service (S3): ideal for storing application content like media files, static assets, and user uploads.



Cloud Data Stores: Is Bigtable similar to DynamoDB?

- **Google Cloud Bigtable and AWS DynamoDB are both highly-available, scalable, globally distributed and fully-managed serverless NoSQL databases.**
- Both can function as a key-value store, however DynamoDB additionally supports a document model and Bigtable additionally supports a wide-column store.



Using Grids for Data Storage

- What is grids?
- Grid Storage for Grid Computing
- Grid Oriented Storage (GOS)



Example For Grids Computing

$$X = (5 \times 7) + (6 \times 3) + (4 \times 5)$$



Example of Grids Computing

In grid computing, each task is broken into small fragments and distributed across computing nodes for efficient execution. Each fragment is processed in parallel, and, as a result, a complex task is accomplished in less time. Let's consider this equation:

$$X = (5 \times 7) + (6 \times 3) + (4 \times 5)$$

Typically, on a desktop computer, the steps needed here to calculate the value of X may look like this:

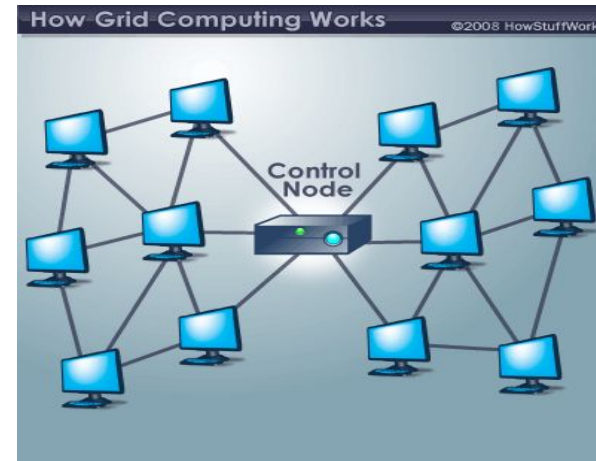
- Step 1: $X = 35 + (6 \times 3) + (4 \times 5)$
- Step 2: $X = 35 + 18 + (4 \times 5)$
- Step 3: $X = 35 + 18 + 20$
- Step 4: $X = 73$

However, the steps in a grid computing setup differ as three processors or computers calculate different pieces of the equation separately and combine them later. This implies fewer steps and shorter timeframes.



Using Grids for Data Storage

- What is grids Computing?
- Grid computing is a computing infrastructure that combines computer resources spread over different geographical locations to achieve a common goal.





Using Grids for Data Storage

- **Grid Storage for Grid Computing**
- Grid computing established its stand as an understood architecture, as it provides users and applications to use shared pool of resources.
- The compute grid connects computers both desktops and servers and storage across an organization.
- It virtualizes heterogeneous and remotely located components into a single system.
- Grid computing allows sharing of computing and data resources for multiple workloads and enables collaboration both within and across organizations.



Using Grids for Data Storage

- Grid Storage for Grid Computing
- Storage for grid computing requires a common file system to present as a single storage space to all workloads.
- Presently grid computing system uses NAS type of storage.
- NAS provides transparency but limits scale and storage management capabilities.
- To set the unique demands of the compute grid on its storage infrastructure, storage for the grid must be abnormally flexible.



Using Grids for Data Storage

- Grid Storage for Grid Computing
- DAS is basically not an option.
- Virtualization is a start, providing the single unit behaviour where the global filing system requires data compute grid.
- Due to this, SAN architectures are used.
- However, the scale of these SANs is beyond the capabilities of fibre channel.



Using Grids for Data Storage

- **Grid Oriented Storage (GOS)**
- GOS is a dedicated data storage architecture connected directly to a computational grid.
- It supports and acts as a **data bank and large supply for data if needed**, which can be shared among multiple grid clients.
- GOS is a **successor of Network-Attached Storage (NAS)** products in the grid computing era.
- GOS accelerates all kinds of applications in terms of **performance and transparency**.
- A GOS system contains multiple hard disks, arranged into logical, redundant storage containers like traditional file servers.



Using Grids for Data Storage

- **Grid Oriented Storage (GOS)**
- GOS deals with **long-distance, heterogeneous and single-image file operations.**
- GOS acts as a **file server** and uses **file-based GOS-FS protocol.**
- **Similar to GridFTP, GOS-FS integrates a parallel stream engine and Grid Security Infrastructure (GSI).**
- GOS-FS can be used as an underlying platform to utilize the available bandwidth and accelerate performance in grid-based applications.



Cloud Storage

- Cloud storage is a **part** of cloud computing.
- Cloud storage can be accessible through web-based applications maintained by the third party (service provider).
- Cloud storage is nothing but virtualized storage on demand called as **Data storage as a Service (DaaS)**.
- Cloud storage can be deployed in many ways. For example:
 - Local data (desktop/laptop) can be backed up to cloud storage.
 - A virtual disk can be 'sync' to the cloud and distributed.
 - The cloud can be used as a reservoir for storing data.



Cloud Storage

- Following are some additional cloud storage attributes:
- **Resource pooling and multi-tenancy:** Multiple consumers can use shared single storage device. Storage resources are pooled and consumers can be assigned and unassigned resources according to their needs.
- **Scalable and elastic:** Virtualized storage can be easily expanded on need basis.



Cloud Storage

- Following are some additional cloud storage attributes:
- Accessible standard protocols including HTTP, FTP, XML, SOAP and REST.
- Service-based: Consumers no need to invest, that is, no CAPEX (Capital Expenditure) and only pay for usage, that is, OPEX (Operational Expenditure).
- Pricing based on usage
- Shared and collaborative
- On-demand self-service



Data Management for Cloud Storage

- Introduction
- Cloud Data Management Interface (CDMI)
- Cloud Storage Requirements



Data Management for Cloud Storage

- **Introduction**
- Cloud storage should incorporate new services according to change of time.
- For cloud storage, a standard document is placed by SNIA(Storage Networking Industry Association), Storage Industry Resource Domain Model (SIRDM).
- **Figure** shows the SIRDM model which uses CDMI standards(Cloud Data Management Interfac)
- SIRDM model adopts **three metadata**:
- storage metadata, data metadata and user metadata.



Data Management for Cloud Storage

- **User metadata** is used by the cloud to find the data objects and containers.
- **Storage system metadata** is used by the cloud to offer basic storage functions like assigning, modifying and access control.
- **Data system metadata** is used by the cloud to offer data as a service based on user requirements and controls the operation based on that data.

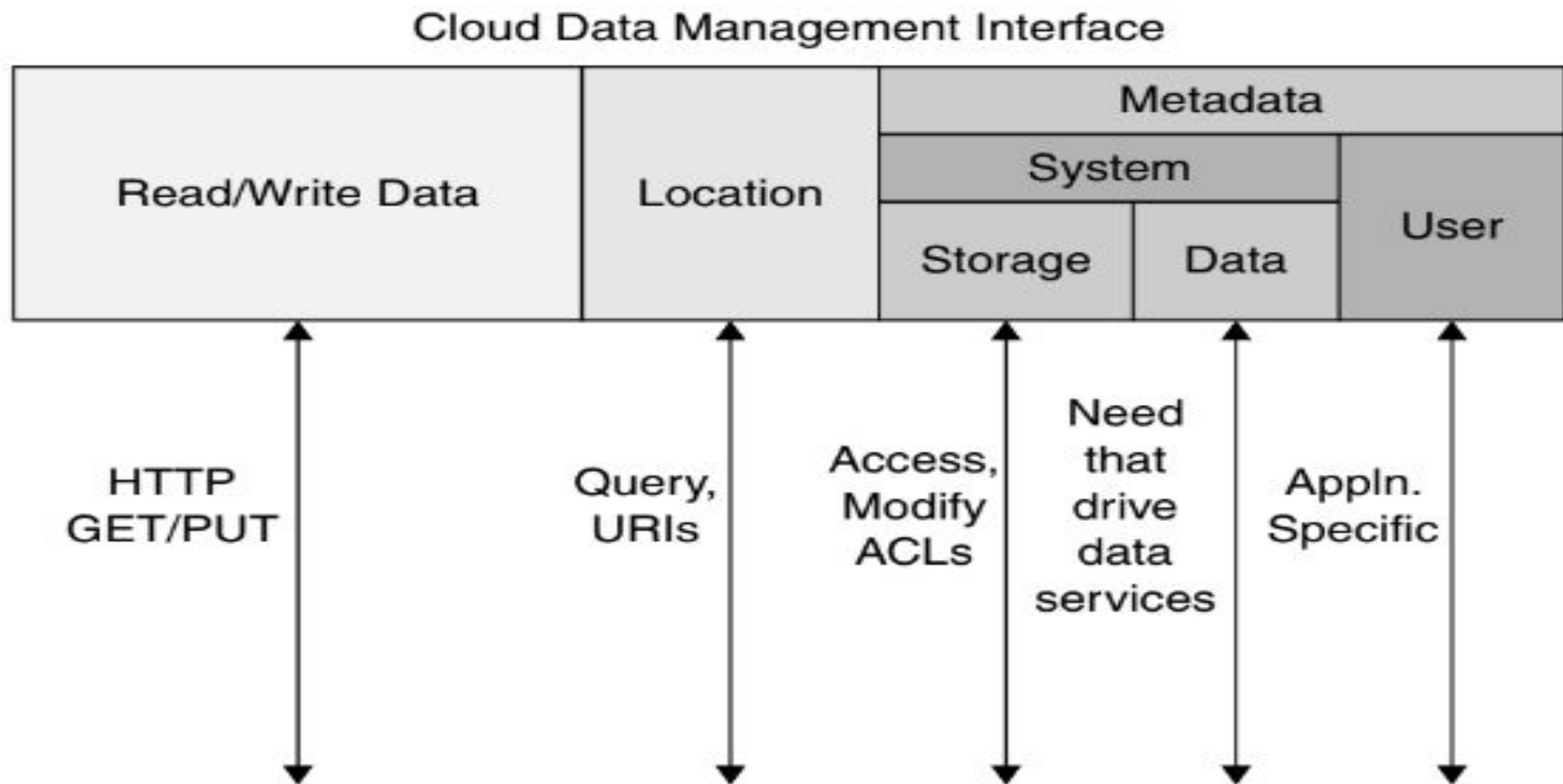


Figure 12.1 Cloud Storage Usage of SIRDm Model



Data Management for Cloud Storage

- Cloud Data Management Interface (CDMI)
- To create, retrieve, update and delete objects in a cloud ,the cloud data management interface (CDMI) is used.
- **The functions in CDMI are:**
 - Cloud storage offerings are discovered by clients
 - Management of containers and the data
 - Sync metadata with containers an objects



Data Management for Cloud Storage

- Cloud Data Management Interface (CDMI)
- CDMI is also used to manage containers, domains, security access and billing information.
- CDMI standard is also **used as protocols** for accessing storage.
- CDMI defines how to manage data and also ways of storing and retrieving it.
- **‘Data path’** means how data is stored and retrieved.
- **‘Control path’** means how data is managed.
- CDMI standard supports both data path and control path interface.



Provisioning Cloud Storage

- Cloud means **sharing** third party resources via the **Internet**.
- This **sharing can be done on need basis** and **there is no need to invest any infrastructure at consumers end**.
- Capacity of storage can be increased on need basis and can be done using multi-tenancy methods.



Provisioning Cloud Storage

- By adopting **Cloud Data Management Interface (CDMI)**, standard service providers can implement the method **for metering the storage and data usage of consumers.**
- This interface also helps the providers for billing to the IT organizations based on their usage.
- Advantage of this interface is that IT organizations **need not write/use different adapters used by the service providers.**



Data-intensive Technologies for Cloud Computing

- Introduction
- Processing Approach
- System Architecture



Data-intensive Technologies for Cloud Computing

● Introduction

- Data-intensive computing is a related type of computing which use **parallelism concept for processing large volumes of data**, called big data.
- Parallel processing approaches are divided into two types: compute-intensive and data intensive.



Data-intensive Technologies for Cloud Computing

- Introduction
- **Compute-intensive:** Applications which need more execution time for computational requirements
- **Data-intensive** : Applications which to try to find large volume of data and time in process.



Data-intensive Technologies for Cloud Computing

- . Processing Approach
- . Data-intensive computing platforms **use** a **parallel computing approach**.
- This approach **combines multiple processors and disks** as computing clusters connected via high-speed network.
- The **data that are needed to be processed** are independently done by computing resources available in the clusters.



Data-intensive Technologies for Cloud Computing

. Processing Approach

- There are many common **characteristics of data-intensive**
- The principle mechanism used for collection of the data and programs or algorithms to perform the computation
- Programming model used
- Reliability and availability
- Scalability of both hardware and software



Data-intensive Technologies for Cloud Computing

. System Architecture

- For data-intensive computing **an array of system architectures have been implemented.**
 - Architecture for data-intensive computing
 1. MapReduce
 2. HPCC

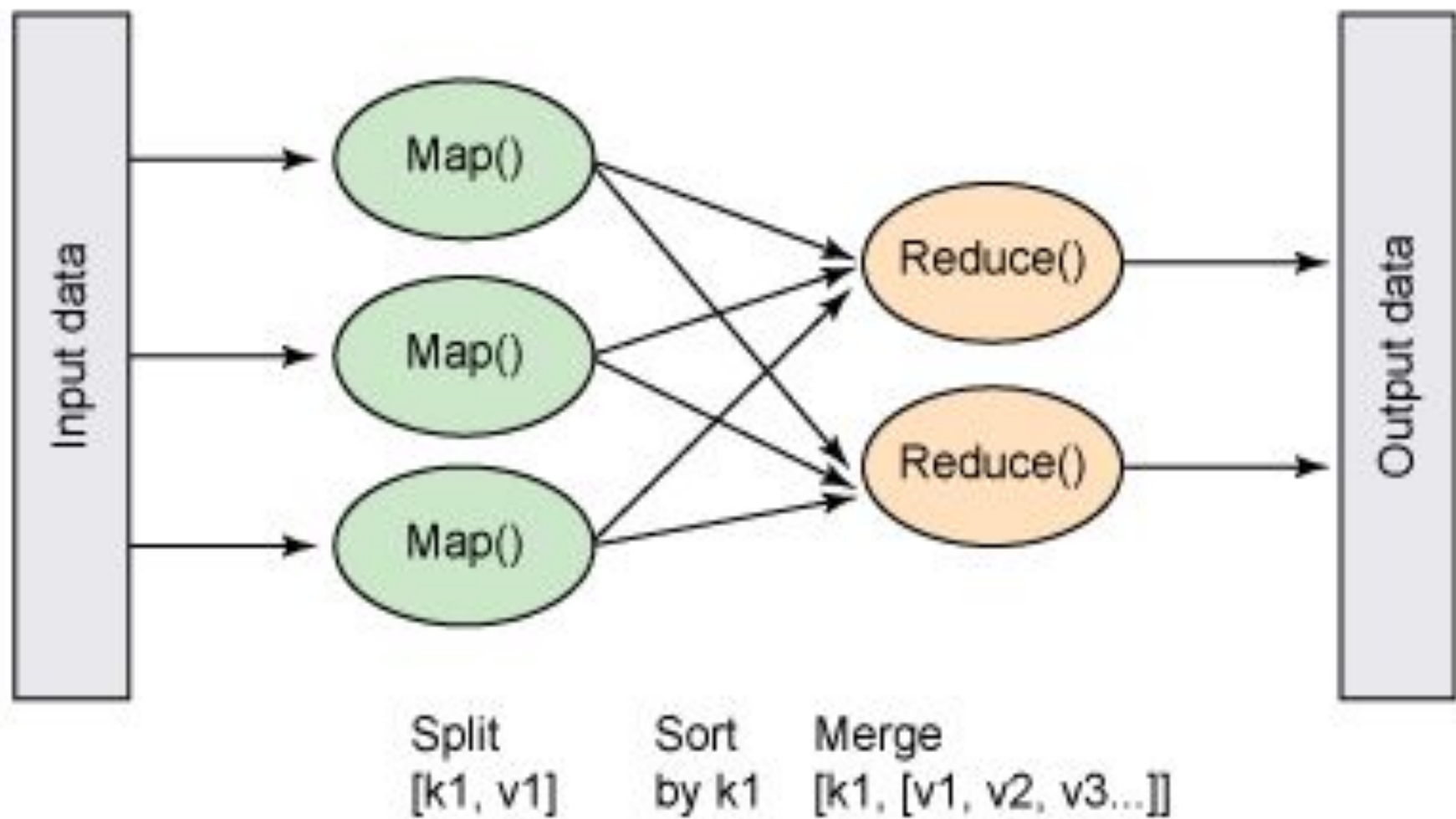


Data-intensive Technologies for Cloud Computing

. System Architecture

. 1. MapReduce

- MapReduce concept which is developed by **Google** and available as **open-source implementation known as Hadoop**.
- This project is used by **Yahoo, Facebook** and others.
- **To create a map function**, the MapReduce architecture uses a functional programming style using key-value pair.
- **Reduce function merges** all intermediate values using intermediate keys.
- Hence programmers who do not have experience in parallel programming can simply use a large distributed processing environment without any problem.





Data-intensive Technologies for Cloud Computing

. System Architecture

• **2. HPCC:**(High-Performance Computing Cluster).

- Developed by Lexis Nexis Risk Solutions called **LexisNexis**.
- LexisNexis Risk Solutions independently developed and implemented a solution for data intensive computing called the HPCC .
- The LexisNexis method structure clusters with commodity hardware that runs in Linux OS.



Data-intensive Technologies for Cloud Computing

. System Architecture

. 2. HPCC

- Custom **system software and middleware parts were created** and layered to provide the execution environment and distributed file system support that is essential for data-intensive computing on the base of Linux operating system.
- A new high-level language for data-intensive computing called ECL is also implemented by LexisNexis.



Cloud Storage from LANs to WANs

- ❑ Cloud Characteristics
- ❑ Distributed Data Storage.
- ❑ Application Utilizing Cloud storage



Cloud Storage from LANs to WANs

★ Cloud Characteristics

★ There are **three characteristics** of a cloud computing ,considered before choosing storage in cloud.

1.Computer power is elastic, when it can perform parallel operations.

e.g.Google's App Engine,

2.Data is retained at an unknown host server.

3.Data is duplicated often over distant locations.e.g.Amazon's S3



Cloud Storage from LANs to WANs

- Distributed Data Storage.
 - Data storage for the new generation of WWW applications through organizations like Google, Amazon and Yahoo.
 - new generation of applications require processing of data to a tune of terabytes and even peta bytes. This is accomplished by distributed services.
 - Following Database are used for distributed data storage
 - Amazon Dynamo
 - CouchDB
 - ThruDB



Cloud Storage from LANs to WANs

- Following Database are used for distributed data storage
- Amazon Dynamo
- It is a fully managed, serverless, key-value NoSQL database designed to run high-performance applications at any scale
- supports key-value and document data structures
- DynamoDB uses synchronous replication across multiple data centers^[4] for high durability and availability.



Cloud Storage from LANs to WANs

- Following Database are used for distributed data storage
- CouchDB

Apache CouchDB is an open-source document-oriented NoSQL database, implemented in Erlang

CouchDB uses multiple formats and protocols to store, transfer, and process its data. It uses JSON to store data, JavaScript as its query language using MapReduce, and HTTP for an API.[2]



Cloud Storage from LANs to WANs

- Following Database are used for distributed data storage
- **CouchDB**
- CouchDB aspires the Four Pillars of Data Management:
 - 1. Save: ACID compliant, save efficiently
 - 2. See: Easy retrieval, straightforward describing procedures, fulltext search
 - 3. Secure: Strong compartmentalization, ACL, connections over SSL
 - 4. Share: Distributed means



Cloud Storage from LANs to WANs

- Following Database are used for distributed data storage
- ThruDB
 - ThruDB aspires to be universal in simplifying the administration of the up-to-date WWW data level (indexing, caching, replication, backup) by supplying a reliable set of services:
 - Thrucene for indexing
 - Throxy for partitioning and burden balancing
 - Thrudoc for article storage



Cloud Storage from LANs to WANs

ThruDB is an open source database built on Apache's Thrift framework and is a set of simple services such as scaling, indexing and storage which is used for building and scaling websites.

ThruDB contains two services

ThruDoc – Document storage service

ThruIndex – Indexing and search service



Cloud Storage from LANs to WANs

❑ Application Utilizing Cloud storage

- ❑ **Online File Storage** :DropBox,Box.net,Live Mesh,Oosah,JungleDisk
- ❑ **Cloud Storage Companies**:Most of these service providers have a free test or offer some sort of free storage space.Box cloud storage,Amazon cloud,SugarSync online backup: SugarSync,Hubic online storage,Google cloud drive: Google
- ❑ **Online Book Marking Service**:Microsoft Labs lately launched Thumbtack, a new bookmarking application.
- ❑ **Online Photo Editing Service** :Online Photo Editors,**Photoshop Express Editor**:Picnik,Splashup,FotoFlexer,Pixier.us



*Thank
you*



Free Cloud Storage

Terabox:1024

Mpeg box

Telegram



AWS:File system

- Amazon S3 (Simple Storage Service): Amazon S3 is an object storage service that offers industry-leading scalability, data availability, security, and performance. It is suitable for a wide variety of use cases, including backup and restore, data archiving, data lakes, and big data analytics. S3 is not a traditional file system but rather an object storage system, where data is stored as objects within buckets.
- Amazon EFS (Elastic File System): Amazon EFS provides scalable file storage for use with Amazon EC2 instances in the AWS Cloud. It is designed to provide scalable, elastic, and shared file storage that is compatible with the NFSv4 protocol. Amazon EFS can be used to support a wide range of file-based workloads and applications, including content repositories, development environments, and data analytics workloads.

Both Amazon S3 and Amazon EFS have their own advantages and use cases. Amazon S3 is ideal for storing large amounts of unstructured data, while Amazon EFS is suitable for applications that require shared file storage and compatibility with the NFSv4 protocol. Depending on your specific requirements, you may choose one or both of these storage options for your cloud file system needs on AWS.



Azure:File system

Azure Blob Storage and Azure Files, which are commonly used for storing files in the cloud.

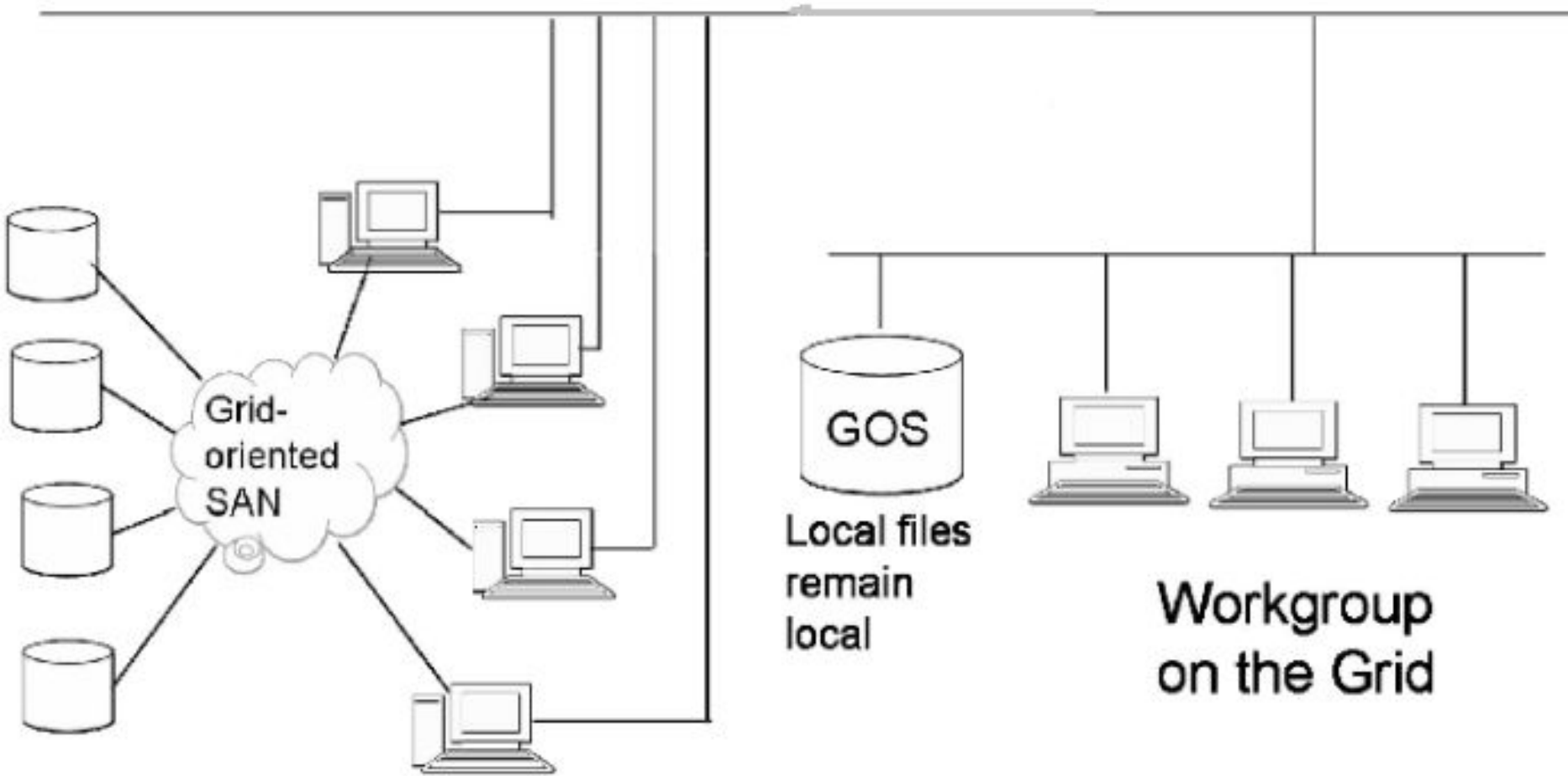
Azure Blob Storage: Azure Blob Storage is Microsoft's object storage solution for the cloud. It is designed to store and serve large amounts of unstructured data, such as text or binary data, such as documents, images, videos, and backups. Blob storage offers various tiers for different access patterns, including hot, cool, and archive tiers, allowing users to optimize storage costs based on their data access requirements.

Azure Files: Azure Files offers fully managed file shares in the cloud using the Server Message Block (SMB) protocol. It provides the ability to create file shares that can be accessed from multiple Azure virtual machines or on-premises systems over standard SMB protocols. Azure Files is suitable for scenarios requiring shared file storage, such as application data, user home directories, and configuration files.

- ACID stands for atomicity, consistency, isolation, and durability
- Atomicity is a feature of databases systems dictating where a transaction must be all-or-nothing
- consistency (or correctness) refers to the requirement that any given database transaction must change affected data only in allowed
- Isolation: defines how or when the changes made by one operation become visible to other
- durability: transactions are saved permanently and do not accidentally disappear or get erased,

Basic Concept

- SNIA:Storage Networking Industry Association
- SNIA is a not-for-profit global organization that leads the storage industry in developing and promoting vendor-neutral architectures, standards, and educational services that facilitate the efficient management, movement, and security of information.



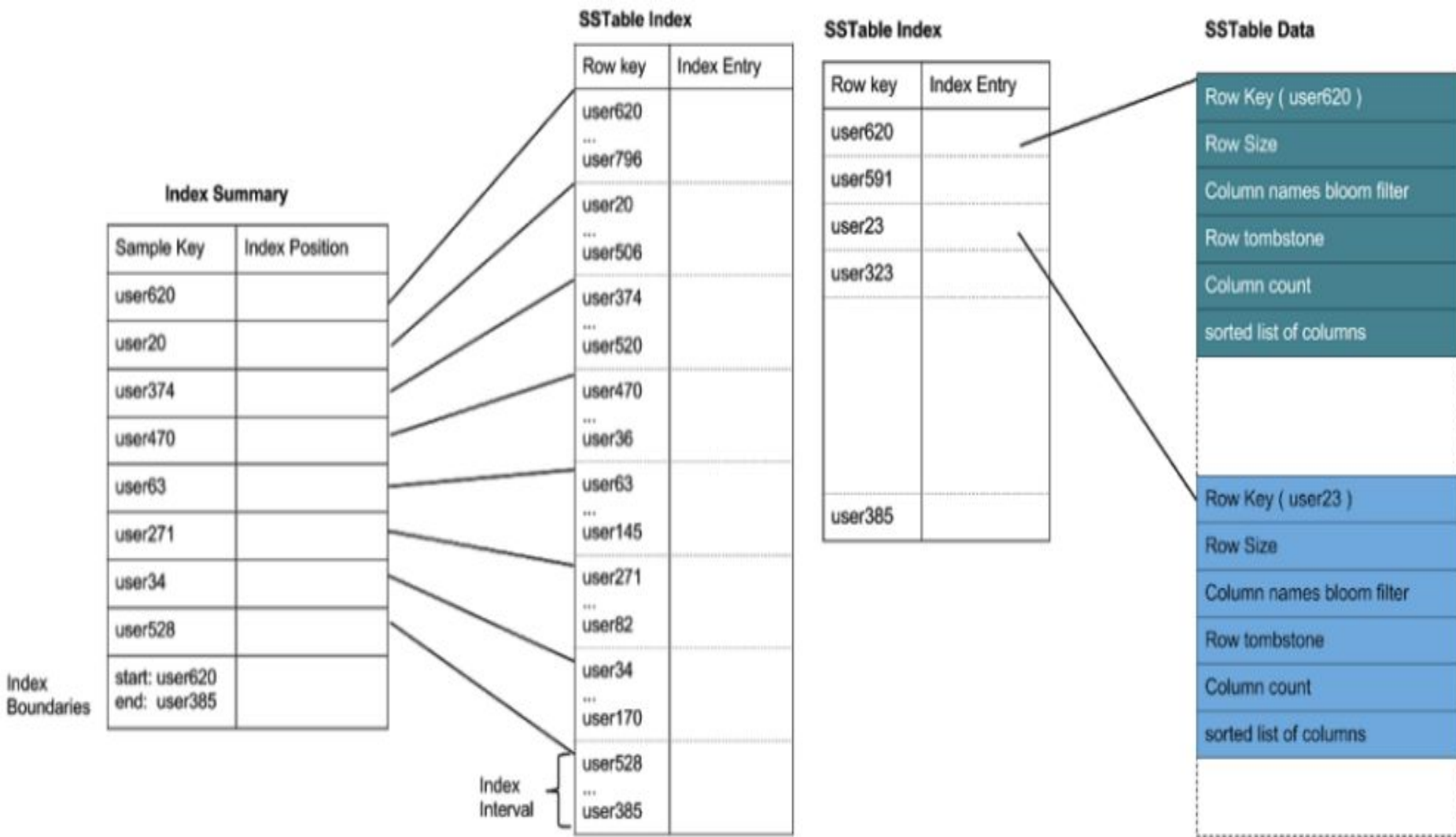
Grid-oriented Server Farm

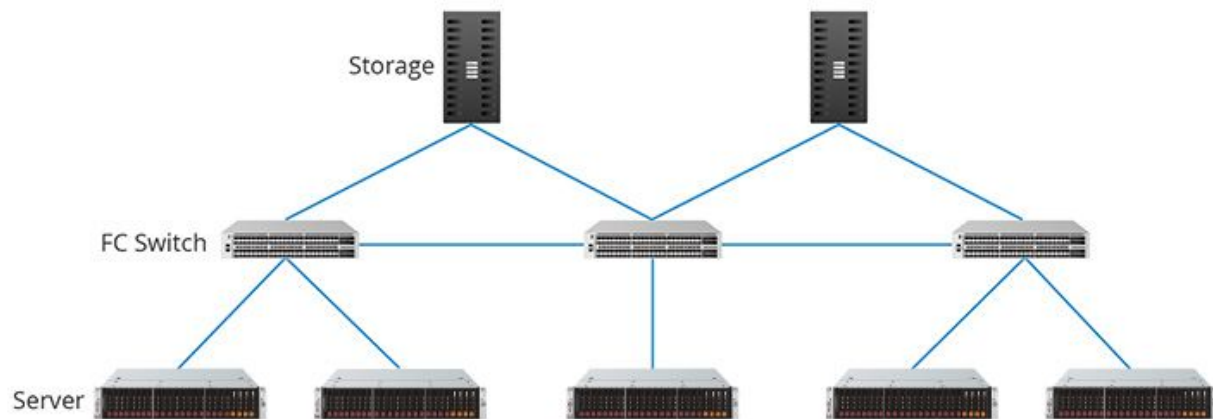
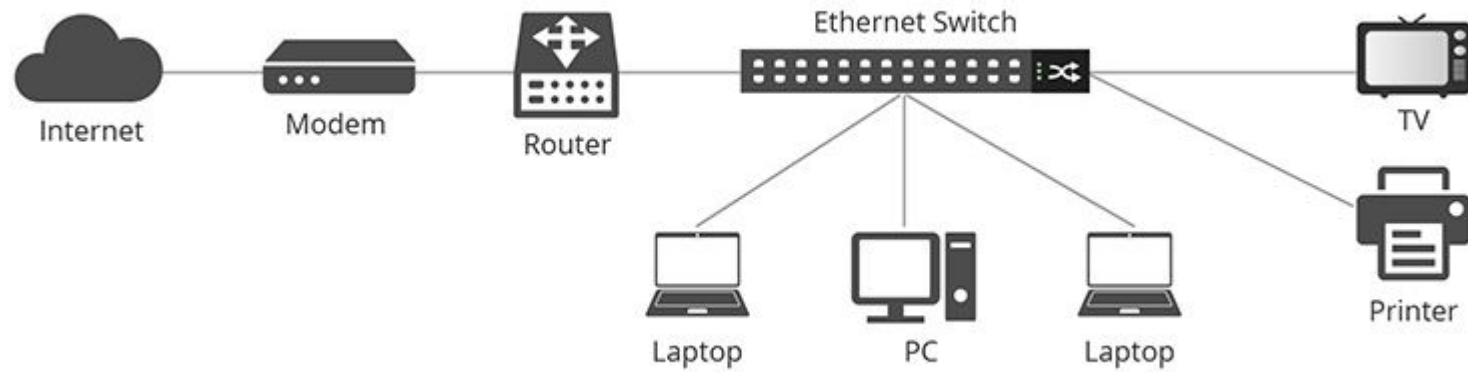
Extra:Chubby Lock Service

- Chubby is used extensively **inside Google in various systems such as GFS, BigTable.**
- Chubby is used **to elect a master, allow the master to discover the servers it controls, and allow clients to find the master.**
- It is also used to store metadata. Chubby is the root of its distributed data structures.
- **Google Chubby** is a highly available and persistent distributed **lock service** and configuration manager for large-scale distributed systems.
- It was first introduced in 2006 to **manage locks for resources** and store configuration information for various distributed services throughout the Google cluster environment.
- Since then, it has since **become a important component** of many Google services, including the Google File System, Bigtable, MapReduce etc

Basic Concept

- Sorted Strings Table (SSTable)
- is a persistent file format used by ScyllaDB, Apache Cassandra, and other NoSQL databases to take the in-memory data stored in memtables, order it for fast access, and store it on disk in a persistent, ordered, immutable set of files. Immutable means SSTables are never modified.
- An SSTable provides a persistent, ordered immutable map from keys to values, where both keys and values are arbitrary byte strings.
- **Colossus** is our cluster-level file system, successor to the Google File System (GFS).





Basic term

- **Ethernet** switches are the **basic building blocks** of networks, which bridge Ethernet devices together.
- Ethernet connects a wide range of computing devices to each other across a local or wide area network (LAN/WAN) using packet switching technology.
- **Fiber Channel(FC)** is a high-speed technology that transfers large quantities of raw block data between servers and data storage centers.
- Fibre Channel supports point-to-point connections in which a server physically access the attached storage directly.
- A **Fibre Channel switch** is a **networking device** that is compatible with the FC protocol and designed for use in a dedicated storage area network (SAN).
- FC switches work to bridge servers and storage.