

Introduction to Cluster Computing

Dr. Hrachya Astsatryan,
Institute for Informatics and Automation Problems,
National Academy of Sciences of Armenia,
E-mail: hrach@sci.am



1

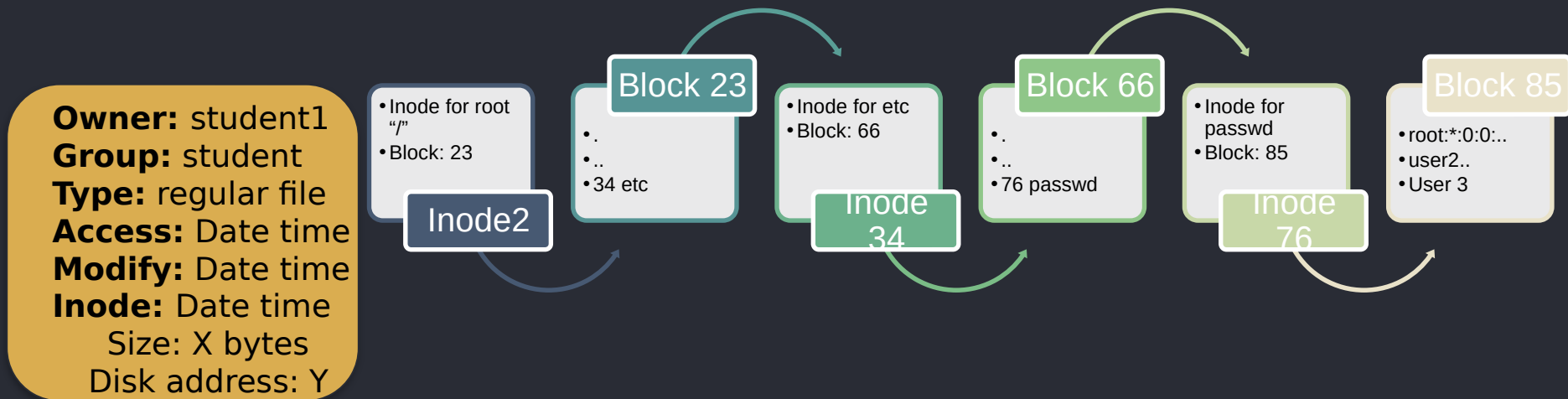


INTRO TO
DISTRIBUTED
FILE SYSTEMS

File systems

A file system is a way of organizing, storing, and naming data on storage media in computers.

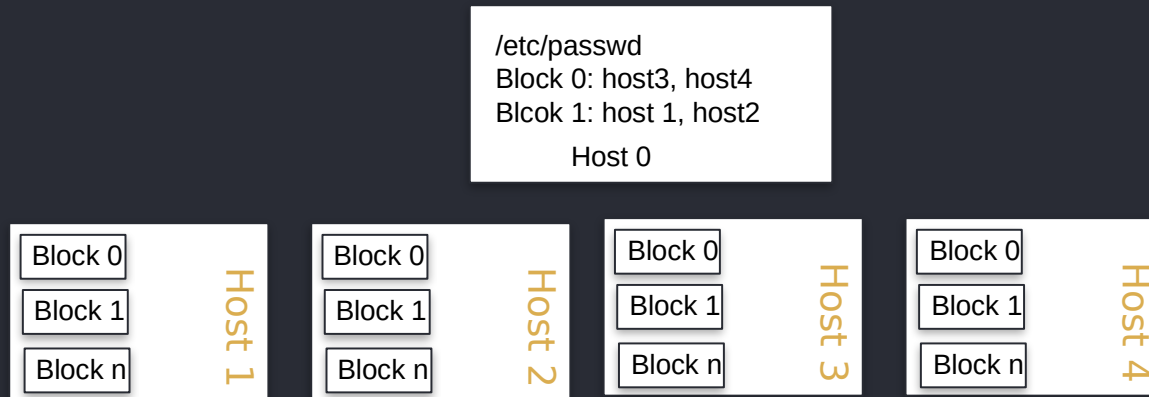
File systems are also used in other electronic equipment: digital cameras and voice recorders, mobile phones, etc.



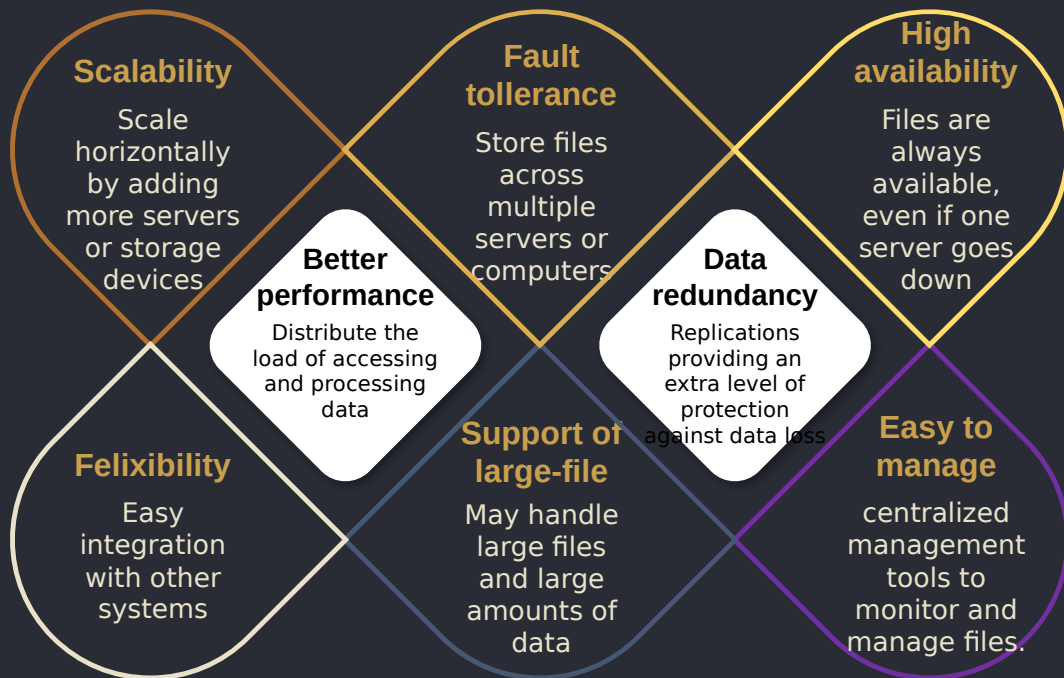
DFS

DFS is a type of file system storing files across multiple servers or computers, rather than being stored on a single machine.

- More storage
- More fault tolerance
- Distributed access



DFS key features



DFS use cases

Media streaming

Store and serve large media files, such as videos and music, to multiple clients

Backup & archiving

Backup and archive large amounts of data for disaster recovery and long-term retention

Cloud storage

Amazon S3
Microsoft Azure

Data analytics

Big data analytics to store and process large amounts of data in parallel

High-availability

Always available even if one or more servers fail

Large files

Store large files, such as videos, images and scientific data

DFS main types

01

Centralized

Centralized

Centralized server that manages access to the files and controls how they are distributed among the nodes.

Examples

- NFS - Network File System
- AFS - Andrew File System

02

Distributed

Distributed

Client-server architecture where the metadata servers manage the file system namespace and metadata while the object storage servers store the data

Examples

- Lustre, Ceph, Google File System, Hadoop Distributed File System.

03

Peer-to-peer

Peer-to-peer

Each node acts as both a client and a server, allowing files to be shared and accessed directly between nodes

Examples

- BitTorrent, Gnutella, Distributed Hash Table

DFS pros & cons

✗ CONS

- Complexity
- Cost
- Administrative overhead



✓ PROS

- Reliability
- Scalability
- Availability

DFS unit of transfer

✓ BLOCK-BASED

Meaning: If the writes lie in different blocks, the final file may represent the changes of both users, as only the block containing each write is written, not the whole file.

Benefits: Convenient in the sense that once the file arrives, the user will not encounter any delays.

Example

- Ceph
- GlusterFS
- OpenEBS
- HDFS

✓ FILE-BASED

Meaning: If two users write to two different parts of the file, the final result will be the file from the perspective of one user, or the other, but not both.

Convenient because the user can "get started" after the first block arrives and does not need to wait for the whole (and perhaps very large) file to arrive.

Example

- NFS
- AFS

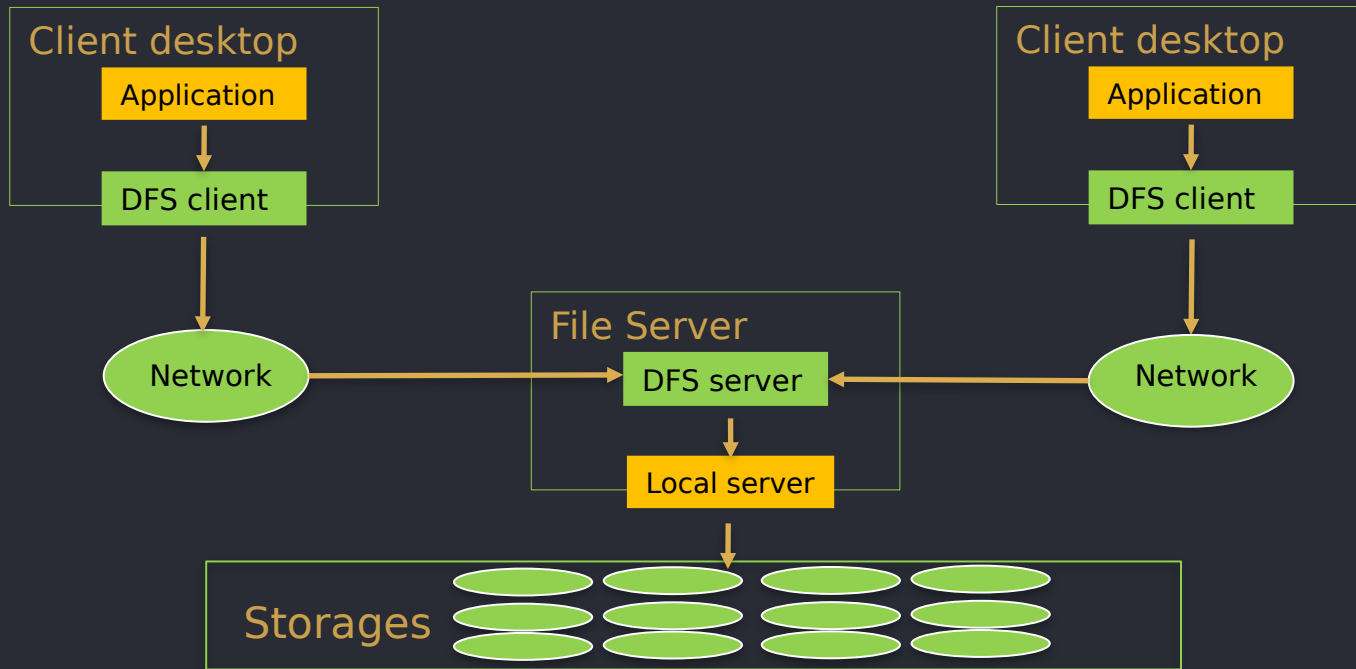


2



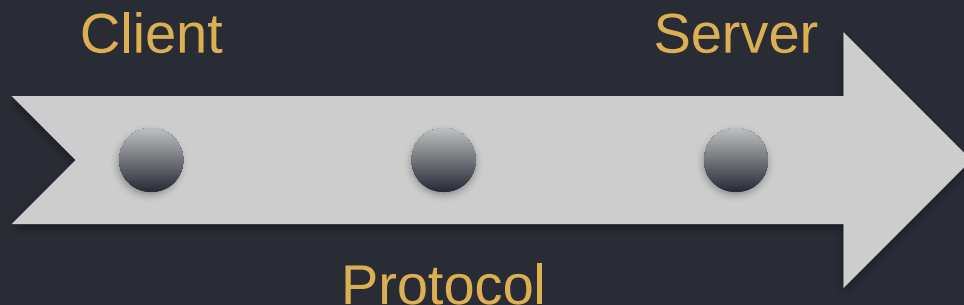
CENTRALIZED DFS

Centralized DFS architecture



Network file systems (NFS)

Developed by SUN microsystems in 1985. Motivated by extending a Unix file system to a distributed environment. But, further extended to other OS as well.



XDR (eXternal Data Representation) provides a way for programmers to pass data among heterogeneous machines without writing procedures to convert among the hardware data representations.

NFS server installation

Check Linux OS: `cat /etc/os-release`

Install NFS Kernel Server in Ubuntu

- `sudo apt install nfs-kernel-server`

Create an NFS Export Directory

- `sudo mkdir -p /mnt/ds`

Remove any restrictions in the directory permissions:

- `sudo chown -R nobody:nogroup /mnt/ds`

Read, write and execute privileges to all the contents inside the directory:

- `sudo chmod 777 /mnt/ds`

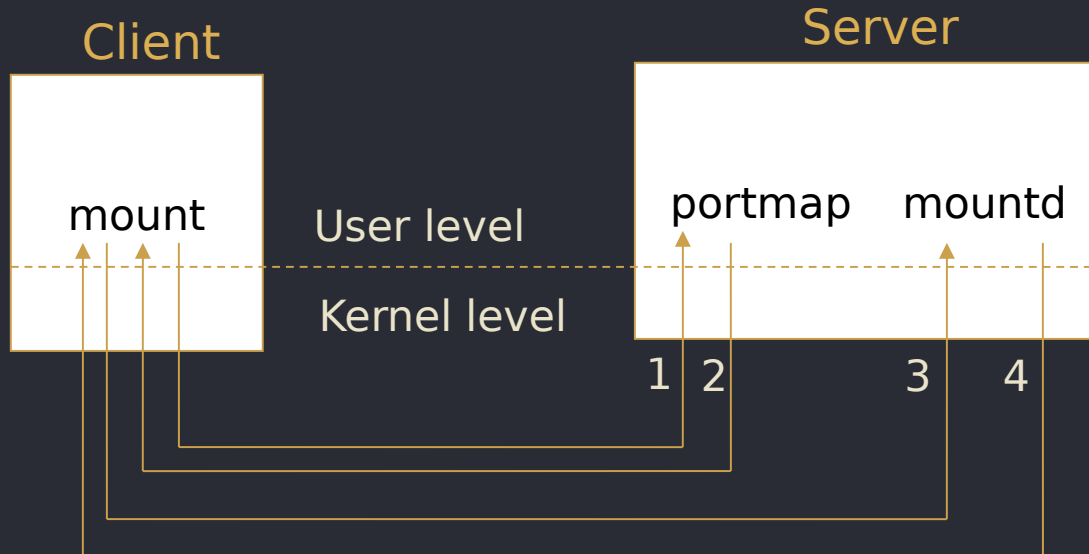
Grant NFS share access

- `/etc/exports /mnt/ds 10.0.0.0/24(rw,sync,root_squash, subtree_check)`
(rw: Stands for Read/Write, sync - Requires changes to be written to the disk before they are applied.)

Export the NFS Share Directory

- `sudo exportfs -a`
- `sudo systemctl restart nfs-kernel-server`

NFS client implementation



- `sudo apt install nfs-common`
- `mount -t ds:/mnt/ds /tmp/ds`

NFS implementation

`/etc/init.d/nfs {start|stop|restart|`

- `portmap`
- `rpc.statd`
- `nfsd`
- `rpc.mountd`
- `rpc.rquotad`

Turns NFS on over reboot

- `$ chkconfig nfs on`

• `/etc/init.d/portmap start`

- `portmap`
- `rpc.statd`
- Manual mount the NFS partitions
- `$ mount 10.0.0.1:/home /home`
- `$ mount 10.0.0.1:/import /import`
- Mount the NFS partitions at boot time
- `$ cat /etc/fstab`
- `10.0.0.1:/home /home nfs defaults 0 0`
- `10.0.0.1:/import /import nfs defaults 0 0`
- `$ mount -a`

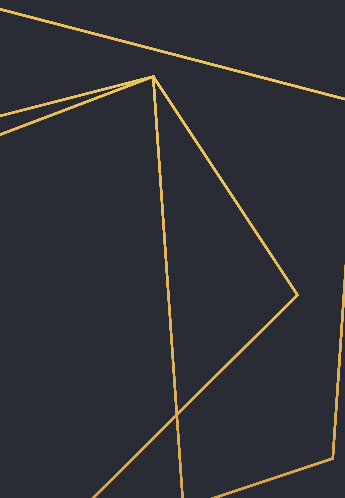
Andrew File System (AFS): overview

Developed - joint effort of Carnegie Mellon University and IBM

- based on client server architecture
- scale better in a large, distributed environment compare to NFS.
- support for secure authentication and access control, while NFS has weaker security features
- faster for large file transfers, compare to NFS which is generally faster for small file transfers, while AFS
- clients download and cache file
- server keeps track of clients that cache the file
- simple, effective
- keep track of clients that have cached files



3



PARALLEL
DISTRIBUTED
SYSTEMS

Parallel DFS architecture

- A parallel distributed file system is typically used in high-performance computing environments, while a centralized file system is more common in smaller businesses or personal use.
- Ceph provides object, block and file storage with a focus on scalability, availability and fault tolerance.
- Lustre handles large amounts of data and high concurrency, and provides high throughput, low latency, and scalability.

Google file system - criteria

- Built from a large amount of ordinary inexpensive equipment, which often fails.
- The system must store many large files. As a rule, several million files, each from 100 MB and more.
- Ceph provides object, block and file storage with a focus on scalability, availability and fault tolerance.
- Lustre handles large amounts of data and high concurrency, and provides high throughput, low latency, and scalability.

Menti 2: 8795 1265

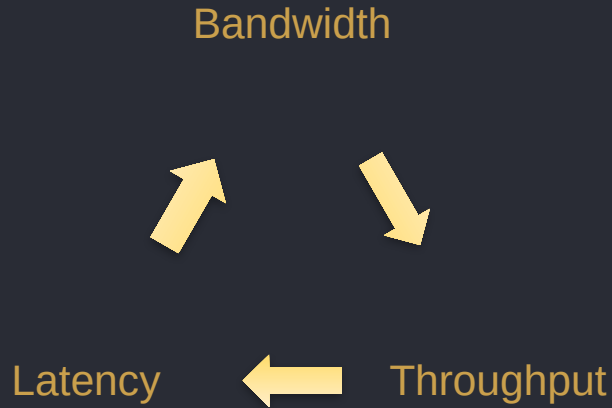


4



NETWORK PARAMETERS

Network parameters



- Bandwidth - capacity of the system
- Throughput - no. of bits that can be pushed through
- Latency (Delay) - delay incurred by a bit from start to finish

Bandwidth

- Number of bits that can be transmitted over a certain time.
- It is usually expressed in terms of bits per second (bps), or sometimes in bytes per second (Bps).
- If bandwidth is B , transmission time is $1/B$.

Pizza delivery example: The bandwidth here represents the size of the pizza delivery vehicle. A larger vehicle can carry more pizzas simultaneously, allowing faster and more efficient delivery to your hungry friends. Similarly, higher bandwidth in a network enables faster data transfer, as more data can be sent in a single communication.

Latency

- The time taken to send a unit of data between two points in a network
- A low latency network is a network in which the design of the hardware, systems and protocols are geared towards minimizing the time taken to move units of data between any two points on that network

Pizza delivery example: The latency is the time it takes for the pizza to arrive at your doorstep after you place the order. A quick delivery represents low latency, while a delay in pizza arrival signifies higher latency.

Latency

Propagation delay = distance/speed of signal.

- Depends on the speed with which the electromagnetic signal (light) travels in the medium -- 2×10^8 m/s in fiber.

Queuing Delay

- At each intermediate node or router, a packet is queued. How long does it have to wait? Dependent on the load on the network -- how many packets are traversing that router?

Latency = Propagation delay + Queue delay + Transmission delay

Throughput

- Throughput is a measure of how many units of information a system can process in a given amount of time.

Pizza delivery example: Throughput is the rate at which pizzas are delivered: A driver with good throughput efficiently serves multiple customers, similar to a network with high throughput delivering data packets effectively.

Throughput= latency?

Top500: Network interconnections

- Infiniband - 237
- 10 Gigabit EtherNet - 119
- Custom - 74
- Gigabit EtherNet - 62
- Proprietary - 8

Network interconnections: Infiniband

InfiniBand is a powerful new architecture designed to support I/O connectivity for the Internet infrastructure.

Data Rate	Theoretical Bandwidth (unidirectional)	End-to-End Latency	Technology
Gigabit Ethernet	125 MB/s	25 ~ 65 microseconds	

Data Rate	Theoretical Bandwidth (unidirectional)	End-to-End Latency	Generation
10Gb/s SDR	1 GB/s	2.6 microseconds	Mellanox InfiniHost III
20Gb/s DDR	2 GB/s	2.6 microseconds	Mellanox InfiniHost III
40Gb/s QDR	4 GB/s	1.07 microseconds	Mellanox ConnectX-3
40Gb/s FDR-10	5.16 GB/s	1.07 microseconds	Mellanox ConnectX-3
56Gb/s FDR	6.82 GB/s	1.07 microseconds	Mellanox ConnectX-3