DISTRIBUTED FILE SYSTEM

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Chennai Mathematical Institute

The ever-growing imbalance between computation and I/O is one of the fundamental challenges for current petascale and future exascale systems. – Zhao and Raicu, Illinois Institute of Technology, 2013.

What Comes Next?

byte

kilobyte

megabyte

gigabyte

??

???

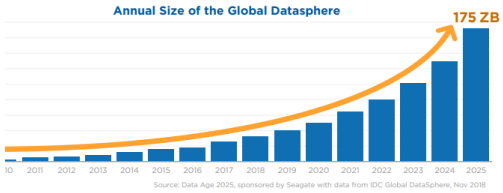
????

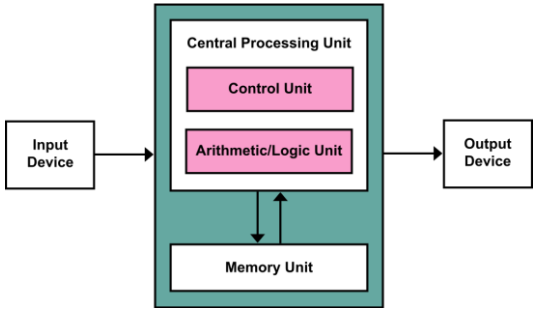
?????

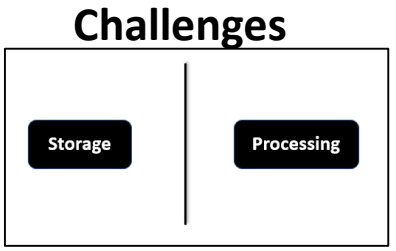
Sizes

Name	Size
Byte	8 bits
Kilobyte	1024 bytes
Megabyte	1024 kilobytes
Gigabyte	1024 megabytes
Terabyte	1024 gigabytes
Petabyte	1024 terabytes
Exabyte	1024 petabytes
Zettabyte	1024 exabytes
Yottabyte	1024 zettabytes

Recap

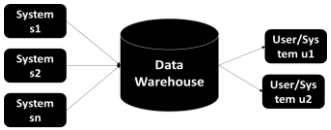

Annual Size of the Global Datasphere
175 ZB
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025

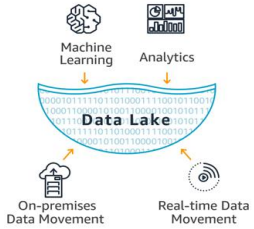

Central Processing Unit
Control Unit
Arithmetic/Logic Unit
Memory Unit
Input Device
Output Device


Challenges
Storage
Processing

Recap

Data Storage

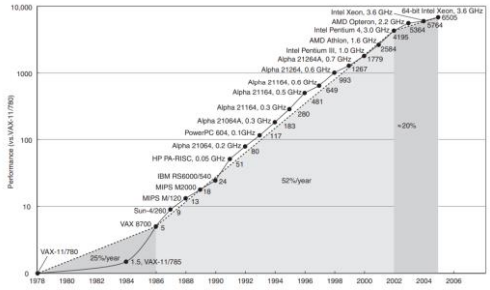




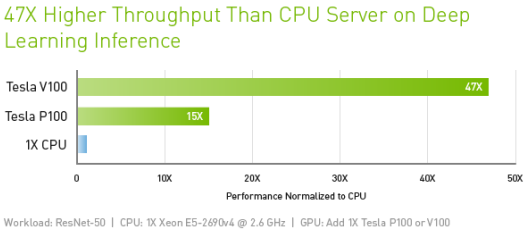
Amazon S3

STaaS

Data Processing



CPU Performance



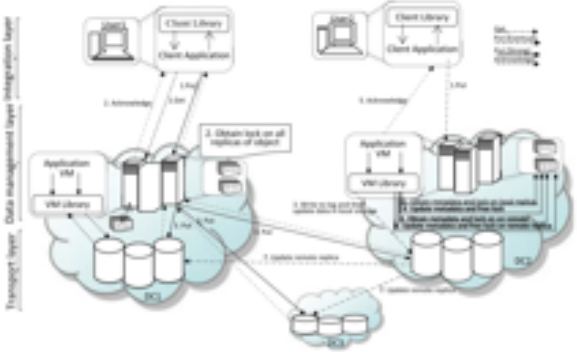
GPU Performance

SuperComputers

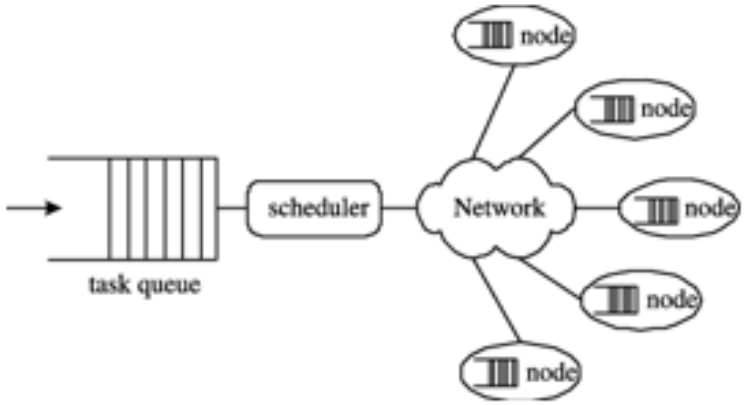
Cloud Computing

Two kinds of Big Data Opportunities

Storage



Processing



So, we have the cloud. But, how to store and retrieve data? How to process jobs?

What is an operating system?

Yarn is now the Apache Hadoop Operating System

Apache Hadoop

Open source platform for reliable, scalable, distributed processing of large data sets, built on clusters of commodity computers.

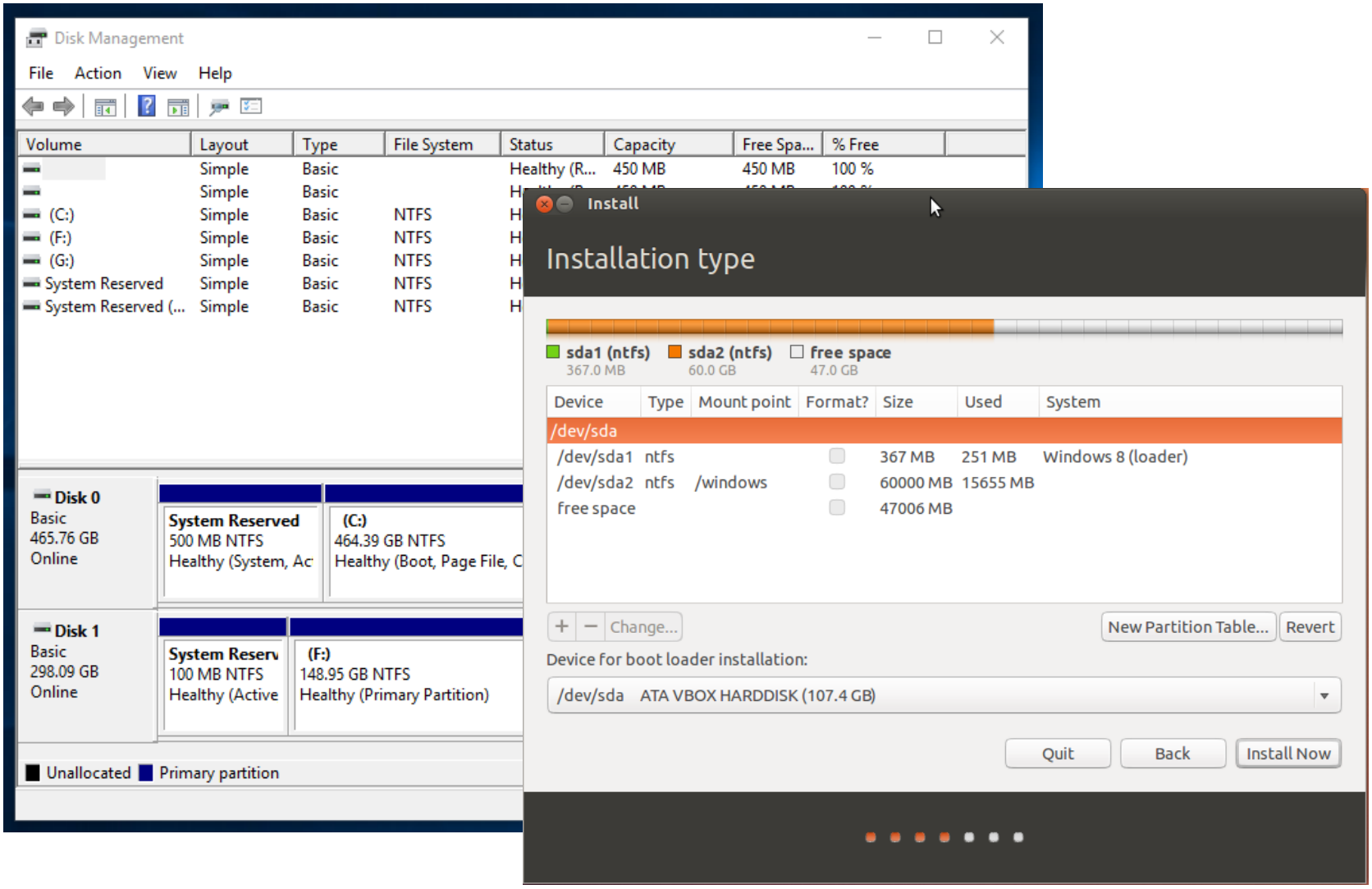
Agenda

- File Systems
  - Introduction
  - File and Folders – How are they stored?
  - Windows/Unix/Miscellaneous File Systems
  - File Allocation Methods
  - Free Space Management
  - Compression
- Distributed File System
  - Hadoop Distributed File System (HDFS)

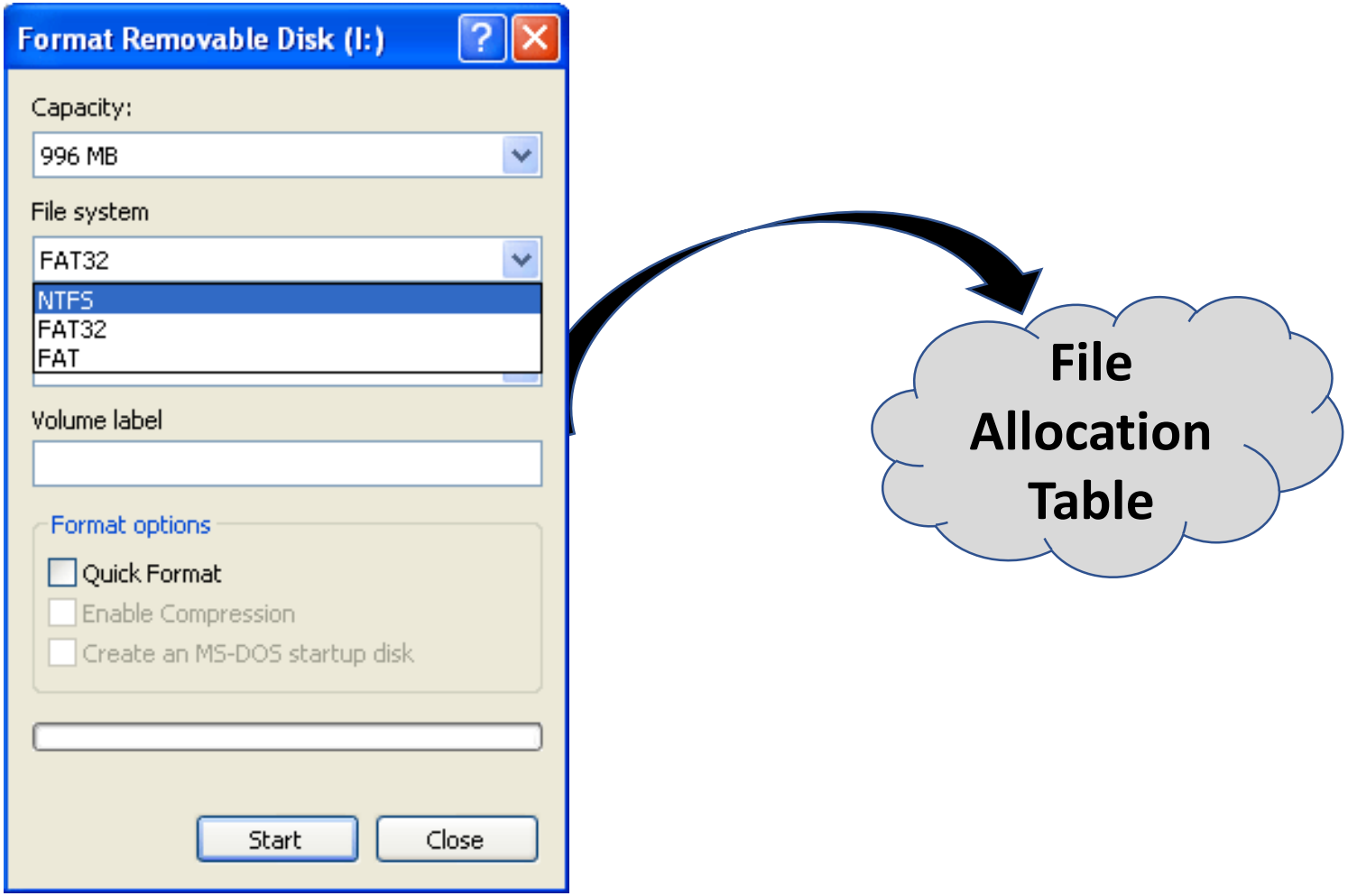
File System

How to store and retrieve files?

Disk Partitioning

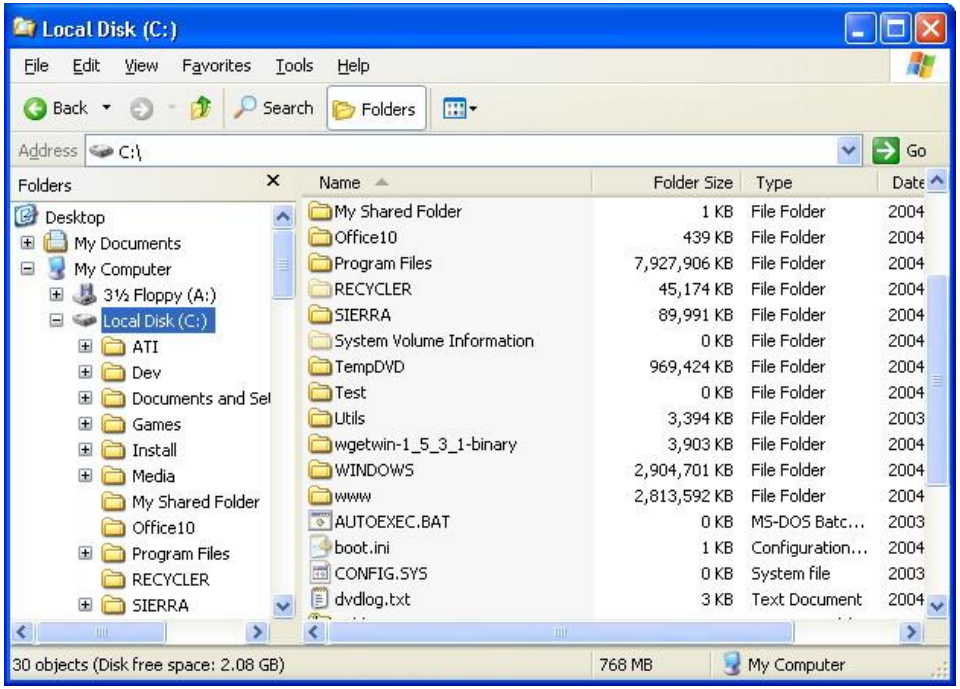


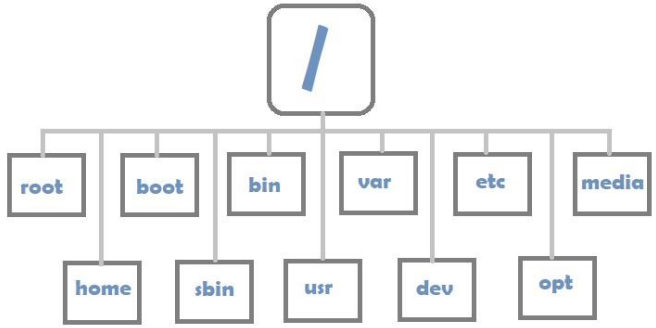
Formatting

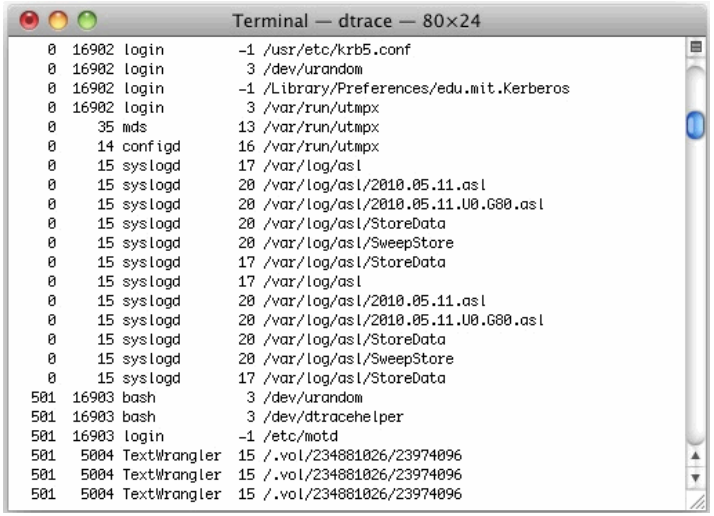

Format Removable Disk (I:)
Capacity:
996 MB
File system
FAT32
NTFS
FAT32
FAT
Volume label
Format options
Quick Format
Enable Compression
Create an MS-DOS startup disk
Start
Close
File Allocation Table

Files and Folders

- An operating system interface to storage media.

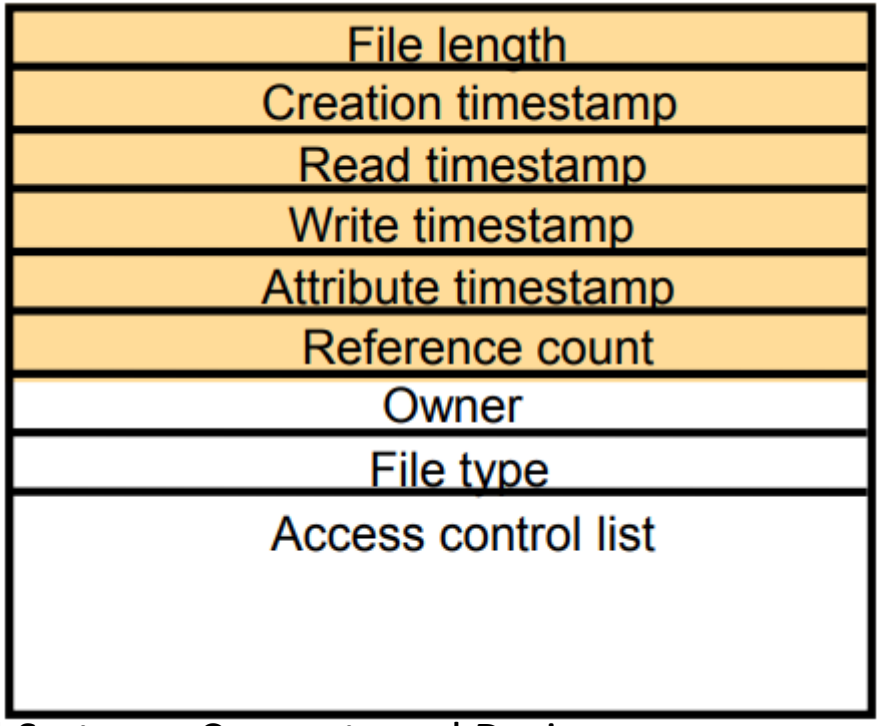






File

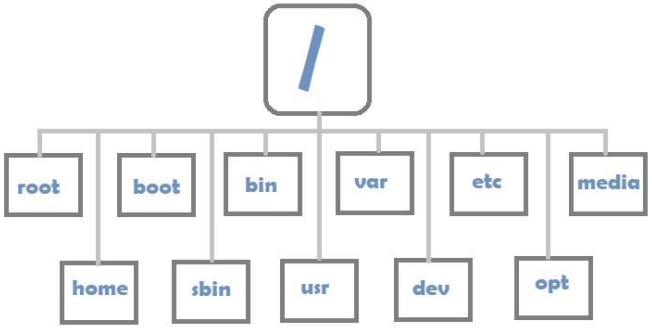
- A Central Object of a File System
- Made of Header and Content



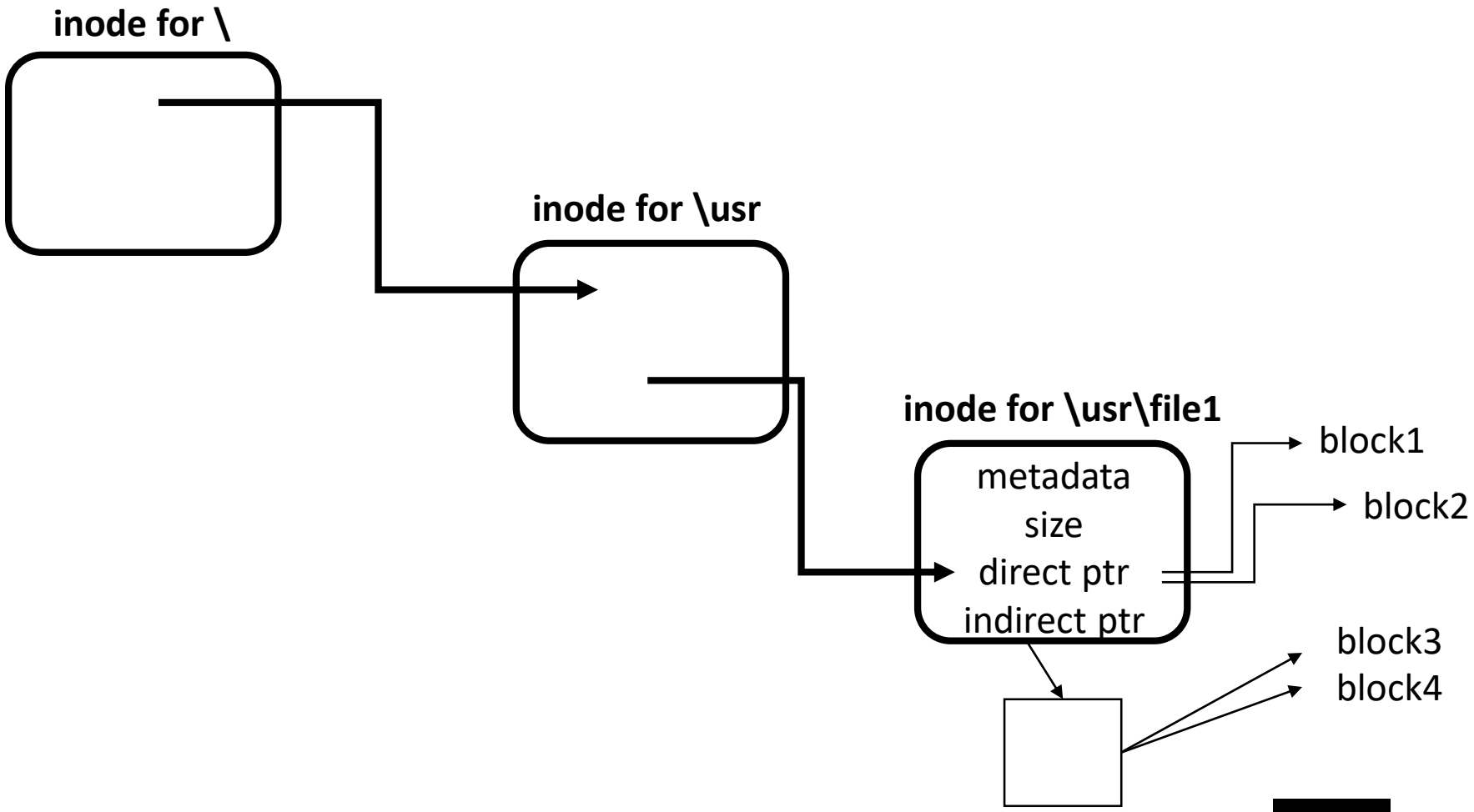
Source: Distributed Systems: Concepts and Design

Unix/Linux File System

- Everything is a file!
  - CD/DVD, USB, …
- Hierarchical
  - / (root) is the top level element
- Accessed through commands
  - cat, cd, cp, mkdir, ls, rmdir, …



inodes (in linux)



Inodes

- Every file has an inode number

```
himanshu@ansh:~$ stat test.txt
  File: 'test.txt'
  Size: 22              Blocks: 8          IO Block: 4096   regular file
Device: 807h/2055d      Inode: 3673414     Links: 1
Access: (0664/-rw-rw-r--)  Uid: ( 1000/himanshu)   Gid: ( 1000/himanshu)
Access: 2018-02-01 16:49:49.256422217 +0530
Modify: 2018-02-01 16:46:59.628037156 +0530
Change: 2018-02-01 16:46:59.708035450 +0530
 Birth: -
himanshu@ansh:~$
```

Hardlinks

- Two filenames for the same file.
- Both the names are mapped to same inode number.

```
root@tryit-right:~# touch f1
root@tryit-right:~# touch f2
root@tryit-right:~# ls
f1  f2
root@tryit-right:~# stat f1
  File: f1
  Size: 0               Blocks: 0          IO Block: 4096   regular empty file
Device: 68h/104d        Inode: 19497       Links: 1
Access: (0644/-rw-r--r--)  Uid: (    0/    root)   Gid: (    0/    root)
Access: 2019-12-21 05:58:56.820000000 +0000
Modify: 2019-12-21 05:58:56.820000
Change: 2019-12-21 05:58:56.820000
 Birth: -
```

```
root@tryit-right:~# ln f1 f3
root@tryit-right:~# ls
f1  f2  f3
root@tryit-right:~# stat f3
  File: f3
  Size: 0               Blocks: 0          IO Block: 4096   regular empty file
Device: 68h/104d        Inode: 19497       Links: 2
```

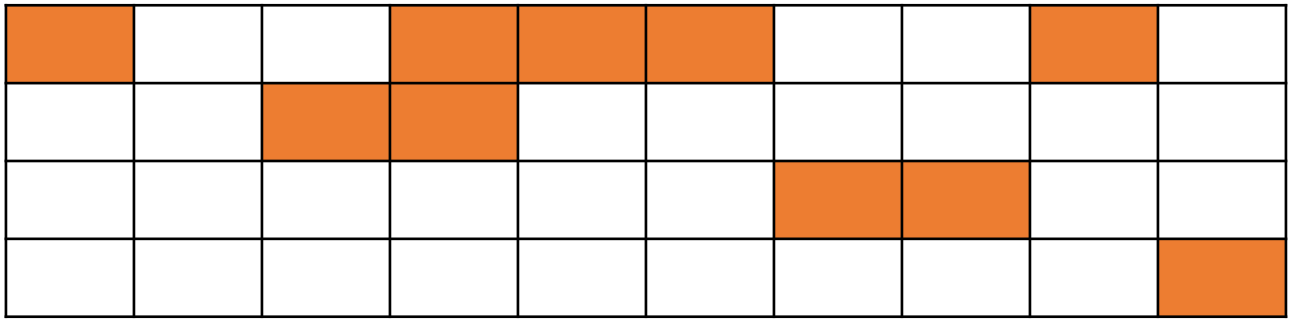
softlinks are just paths to file.

File Permissions

```
dave@howtogeek:~/work$ ls -l
total 80
drwxr-xr-x 2 dave dave  4096 Aug 23 08:02 archive
-rw-rw-r-- 1 dave dave   780 Aug 20 11:11 command_cls.page
-rw-rw-r-- 1 dave dave   828 Aug 20 11:11 command_exit.page
-rw-rw-r-- 1 dave dave   819 Aug 20 11:11 command_gc.page
-rw-rw-r-- 1 dave dave   799 Aug 20 11:11 command_osm.page
-rw-rw-r-- 1 dave dave   829 Aug 20 11:11 command_quit.page
-rw-rw-r-- 1 dave dave   832 Aug 20 11:11 command_satellite.page
-rw-rw-r-- 1 dave dave   811 Aug 20 11:11 command_street.page
-rw-rw-r-- 1 dave dave 28127 Aug 20 11:11 GC Help.mm
-rwxrwxr-x 1 dave dave    46 Aug 20 11:11 mh.sh
-rw-rw-r-- 1 dave dave 16149 Aug 20 11:11 window_tool.page
dave@howtogeek:~/work$
```

File Allocation Methods

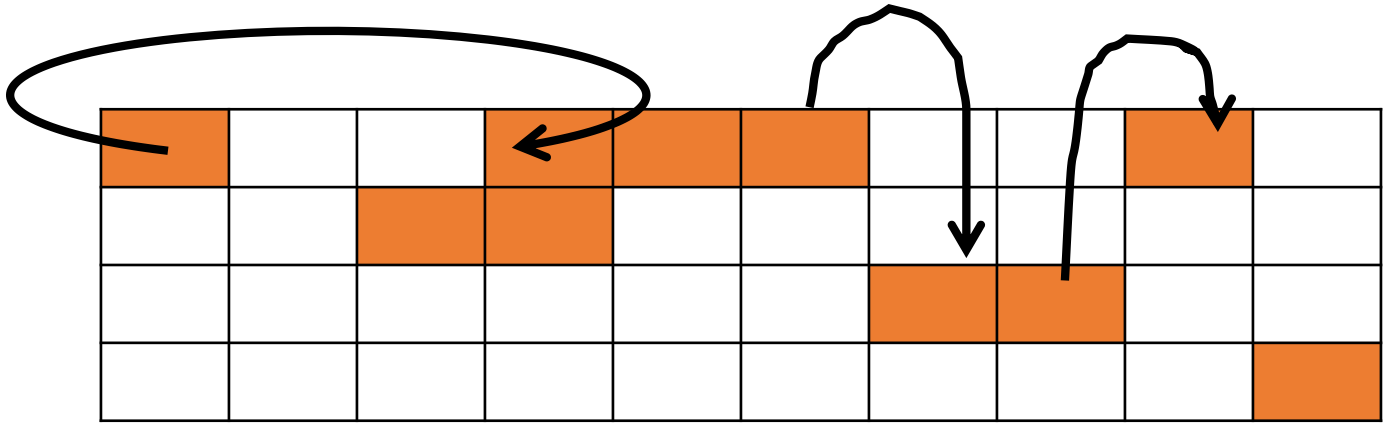
How would you like it if we contiguously write blocks to disk?



Data stored in blocks but need not be in contiguous blocks.

File Allocation Methods

Linked File Allocation



Each file is a linked list of disk blocks

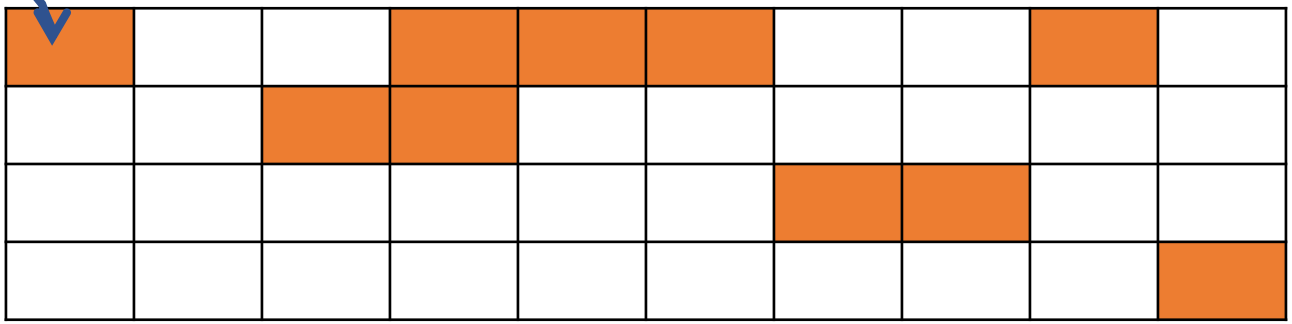
File Allocation Methods

Indexed Allocation

Each file has an index block that stores array of block addresses.

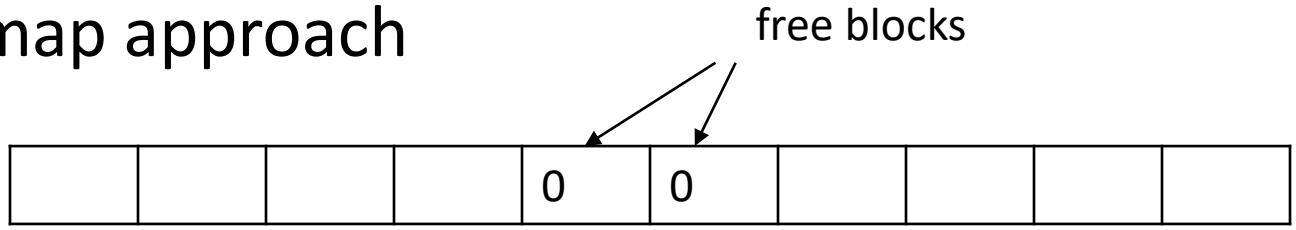
File	Index Block Address
cmi.txt	20

20: Index
1
4
5
6
9



Free Space Management

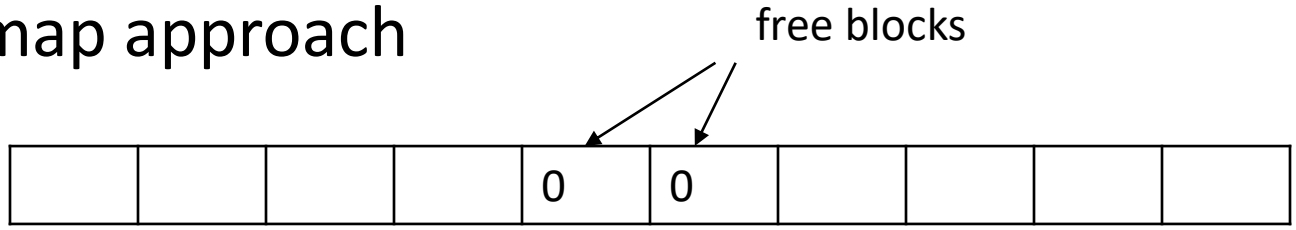
- Bitmap approach



  - Assume disk size = 1 Terabyte, block size = 4 KB. How much space will we need to store the free space bitmap?

Free Space Management

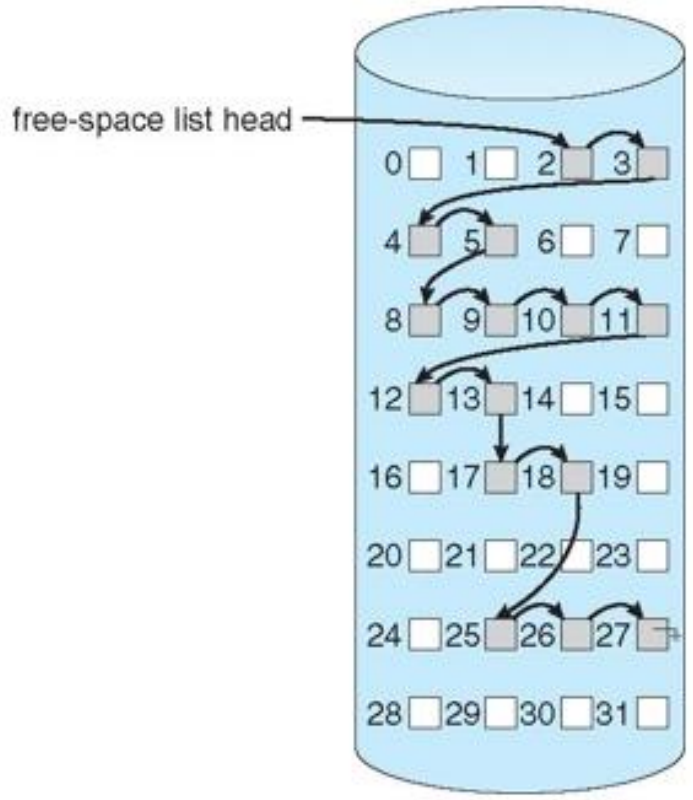
- Bitmap approach



  - Assume disk size = 1 Terabyte, block size = 4 KB. How much space will we need to store the free space bitmap?
    - 1 TB / 4 KB = $2^{40}/2^{12} = 2^{28}$ = 32 MB.

Free Space Management

- Free-list approach



Source: OS Concepts – 9th Edition. Silberschatz, Galvin and Gagne

Windows File Systems

- CDFS
  - CD ROM File System: ISO 9660-compliant standard.
  - Directory/File names shorter than 32 characters, with max depth of 8 levels!
- UDF (Universal Data Format)
  - created primarily for DVD
  - ISO 13346-compliant
- FAT (File Allocation Table) File System
  - Used in DOS and Win 9x.
  - Serious restrictions on file size, filename length, etc.
- NTFS (Native FS for Windows)
  - Windows 10 uses NTFS!

Criteria	NTFS5	NTFS	exFAT	FAT32	FAT16	FAT12
Max Volume Size	2 ^ 64 clusters – 1 cluster	2 ^ 32 clusters – 1 cluster	128PB	32GB	2GB	16MB
Max Files on Volume	2 ^ 32 -1	2 ^ 32 -1	Nearly Unlimited	4194304	65536	
Max File Size	2 ^ 64 bytes	2 ^ 44 bytes	16EB	4GB minus 2 Bytes	2GB	16MB
Max Clusters Number	2 ^ 64 clusters – 1 cluster	2 ^ 32 clusters – 1 cluster	4294967295	4177918	65520	4080
Max File Name Length	Up to 255	Up to 255	Up to 255	Up to 255	8.3	Up to 254

http://www.ntfs.com/ntfs_vs_fat.htm

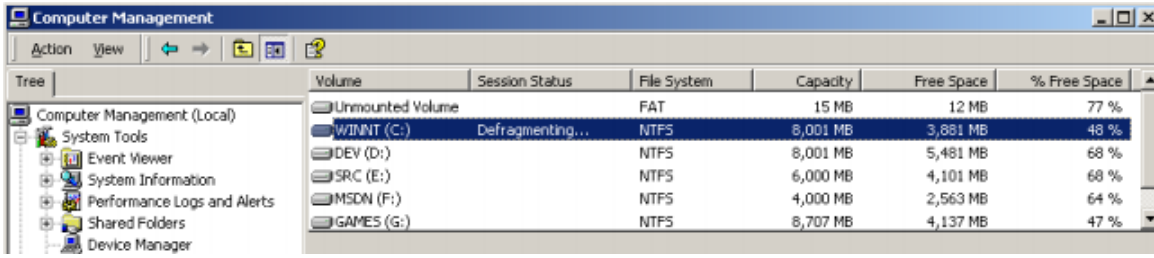
Compression

- Why compress while storage and retrieval?

Compression

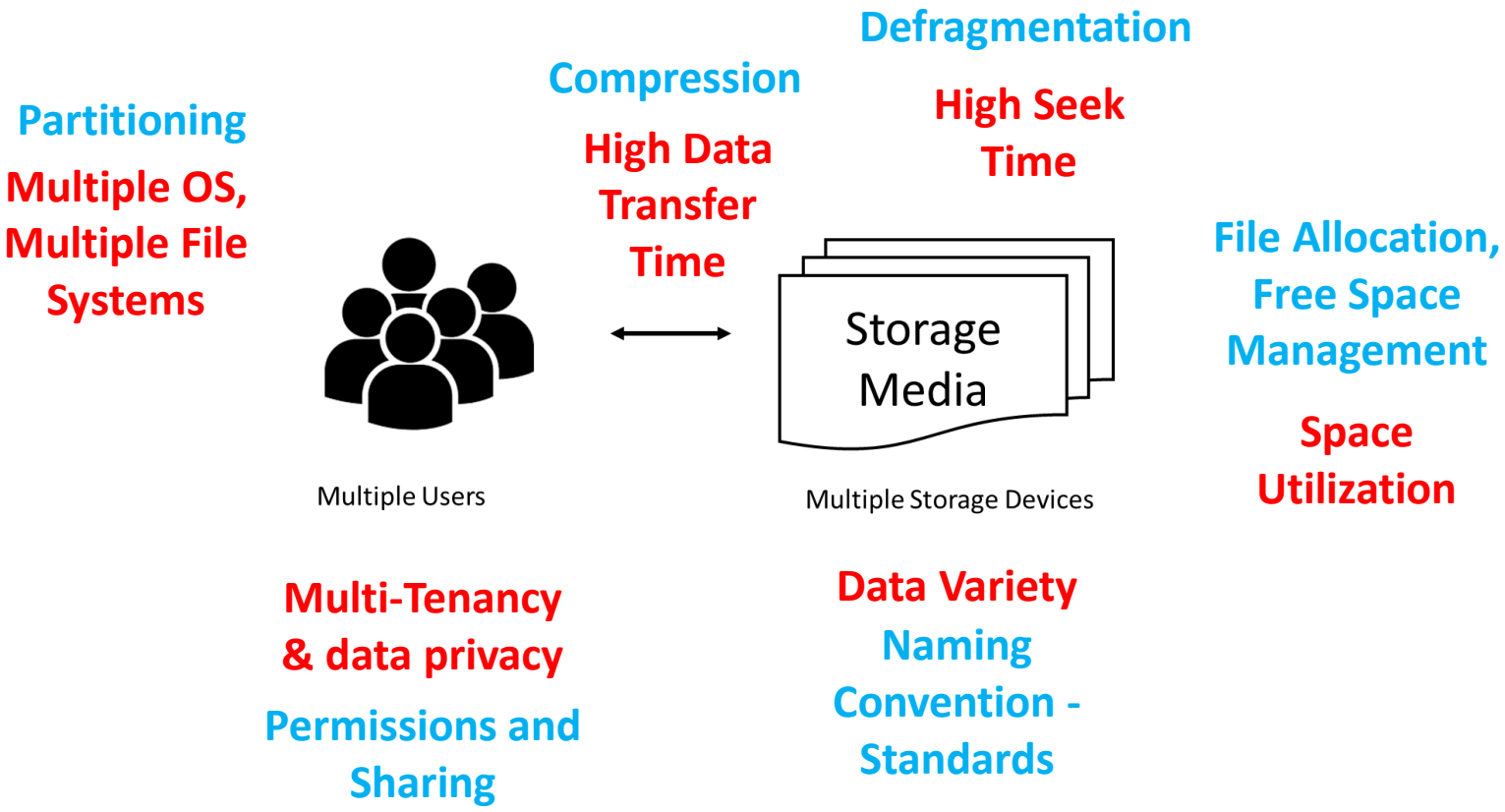
- Why compress while storage and retrieval?
  - To narrow the gap between computation and I/O
  - Usually computation power is much higher, I/O speed is too low.

The Complex World of File Systems



- Defragmentation
- Partitioning
- Compression
- Sharing and Permissions
- Naming Convention
- File Allocation and Free Space Management
- Multiple users and multiple storage media
- …

The Complex World of File Systems



BUSINESS INSIDER INDIA

TECH INSIDER BUSINESS POLICY STRATEGY ADVERTISING SCIENCE

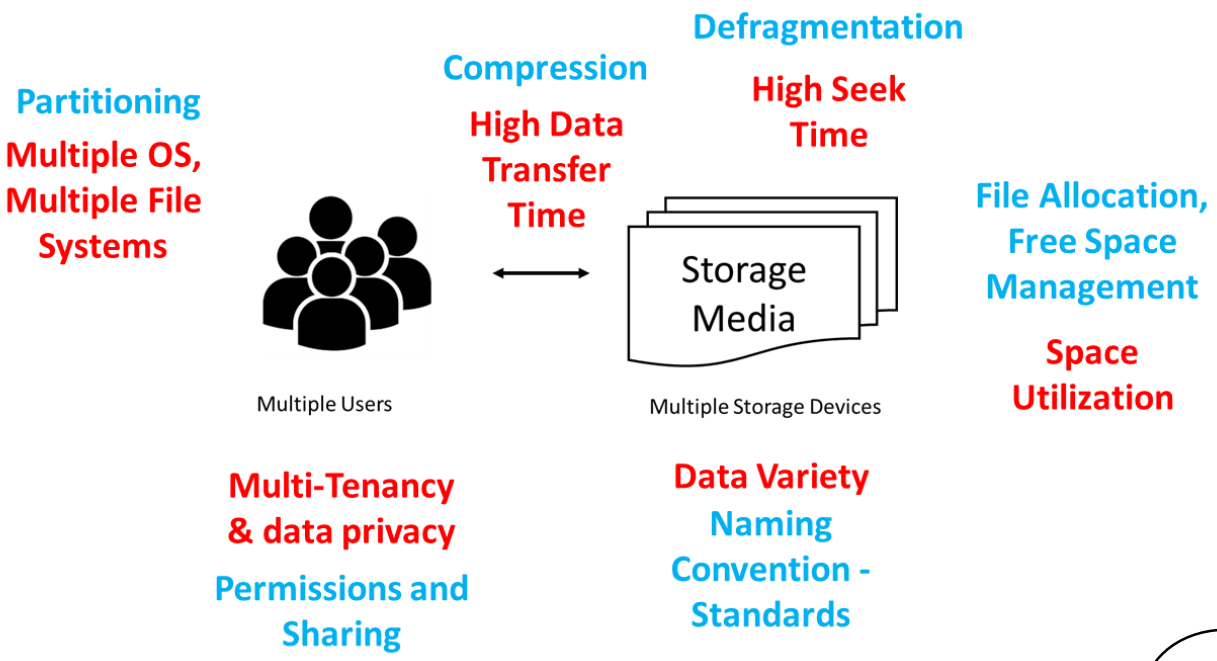
ALL

Home › Tech › News » Linus Torvalds, Creator Of The Linux Operating System, Warned Developers Not To Use An Oracle-Owned File System

Linus Torvalds, creator of the Linux operating system, warned developers not to use an Oracle-owned file system because of the company's 'litigious nature'

ROSALIE CHAN | JAN 13, 2020, 23:36 IST

Summary



File systems are key to handling data.

Variety of FS exist
NTFS, FAT, DOS, CDFS, NFS, …