The Abdus Salam
International Centre for Theoretical Physics

Introduction to Storage for Climate Data

Stefano Cozzini(CNR/IOM Democritos), Gilberto Díaz(ULA)

May 16, 2011

Introduction

Many applications perform relatively simple operations on vast amounts of data. In such cases, the performance of a computer's data storage devices impact overall application performance more than processor performance.

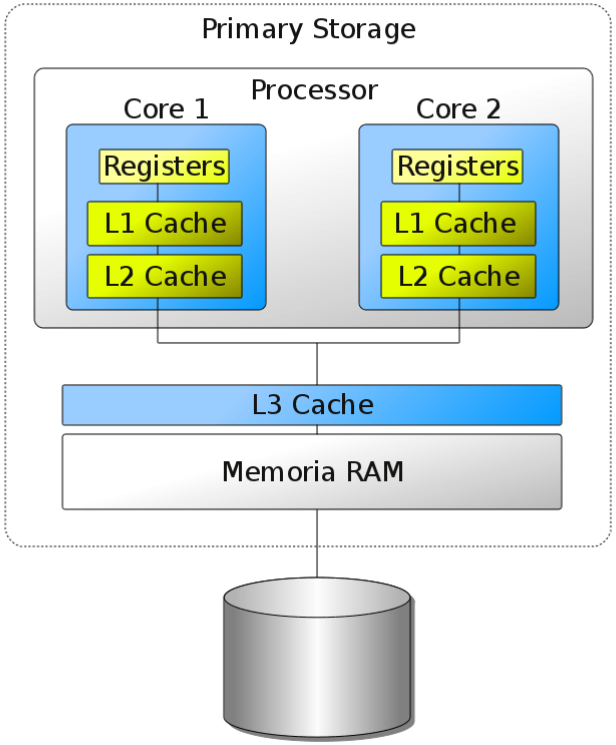
Introduction

Data storage devices include, but are not limited to:

- processor registers
- caches memories
- main memory
- disk (hard, compact disk, etc.)
- and magnetic tape.

Memory Hierarchy

- Computers architectures try to keep data close to the processors in order to feed them continuosly.
- However, while the capacity of storage devices increases, the distance to the processors also increases.



Current Mass Storage Devices

- We are interested particularly in low cost storage devices with big capacity and high performance.
- Nowaday, Magnetic hard Disk Drives still are the technology which include all these features.

Current Hard Disk Drives Technologies

We can find several magnetic hard disk tecnologies today

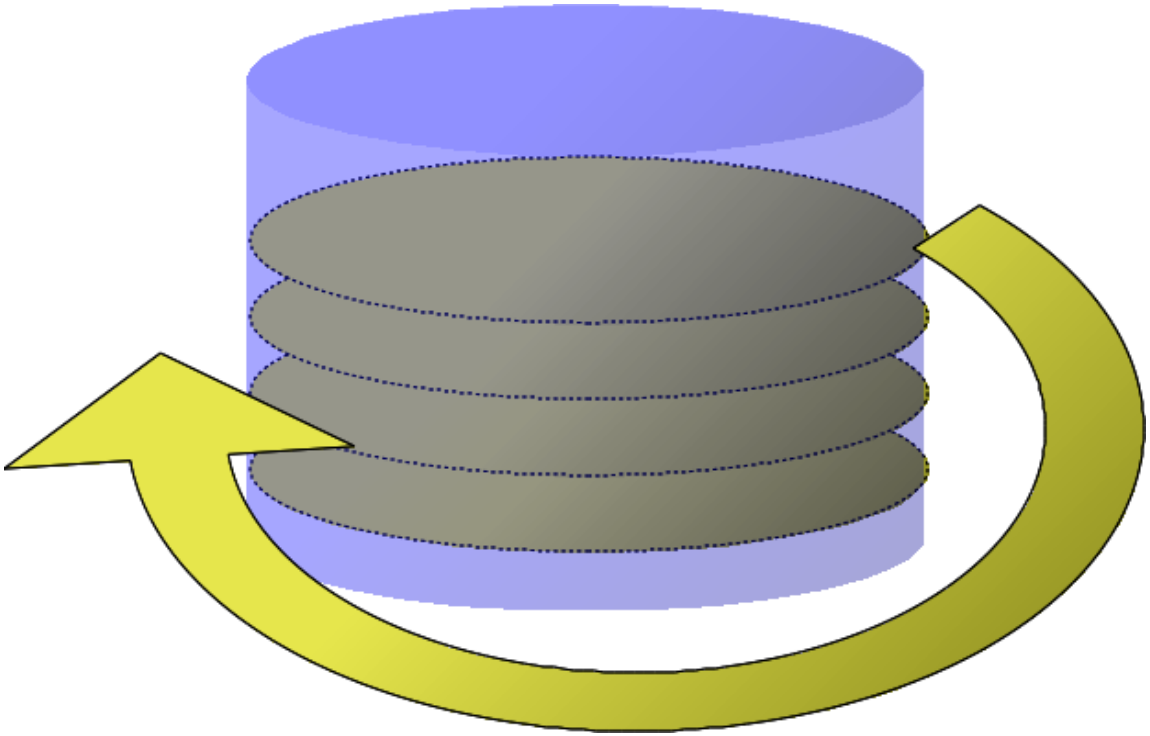
- Serial Advanced Technology Attachment (SATA)
- Small Computer System Interface (SCSI)
- Serial Attached SCSI (SAS)

Current Hard Disk Technologies

- Today disk space is cheap, a single (SATA) disk drives provides up to 3TB.
- However, performance is another story. Fastest hard disk drive (SAS) bandwith is around 3Gbps

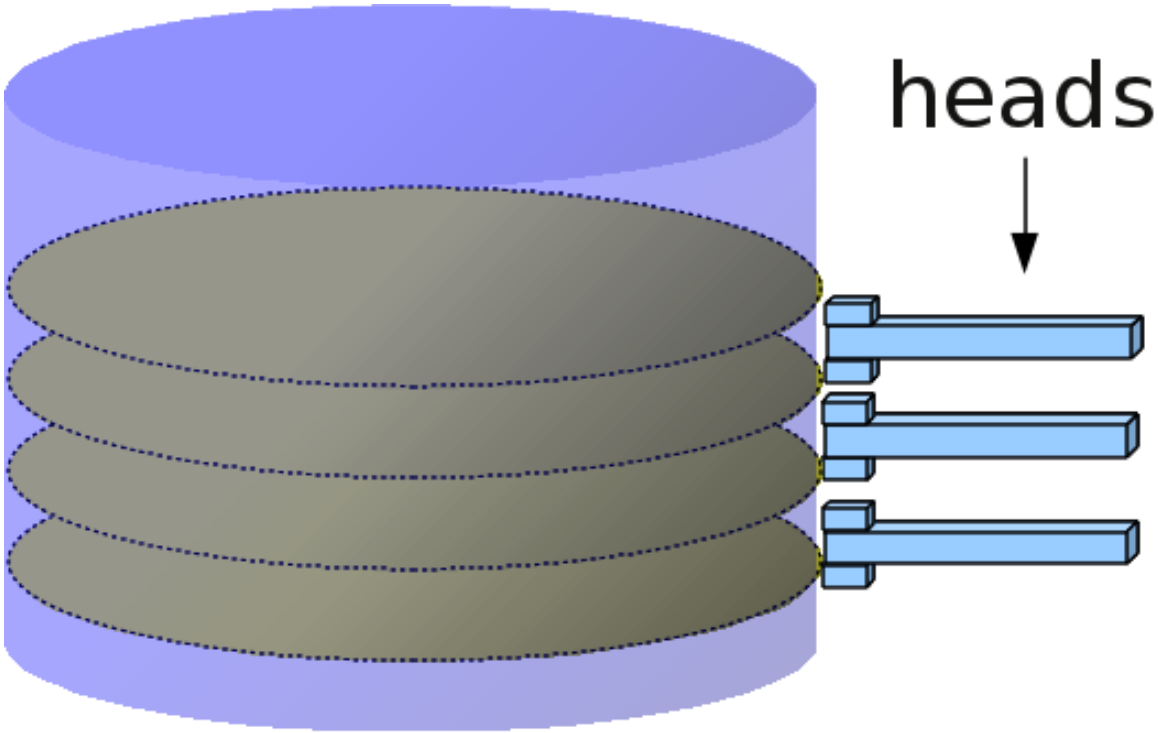
HDD components

A typical HDD includes a plurality of magnetics disks spun by a spindle motor.



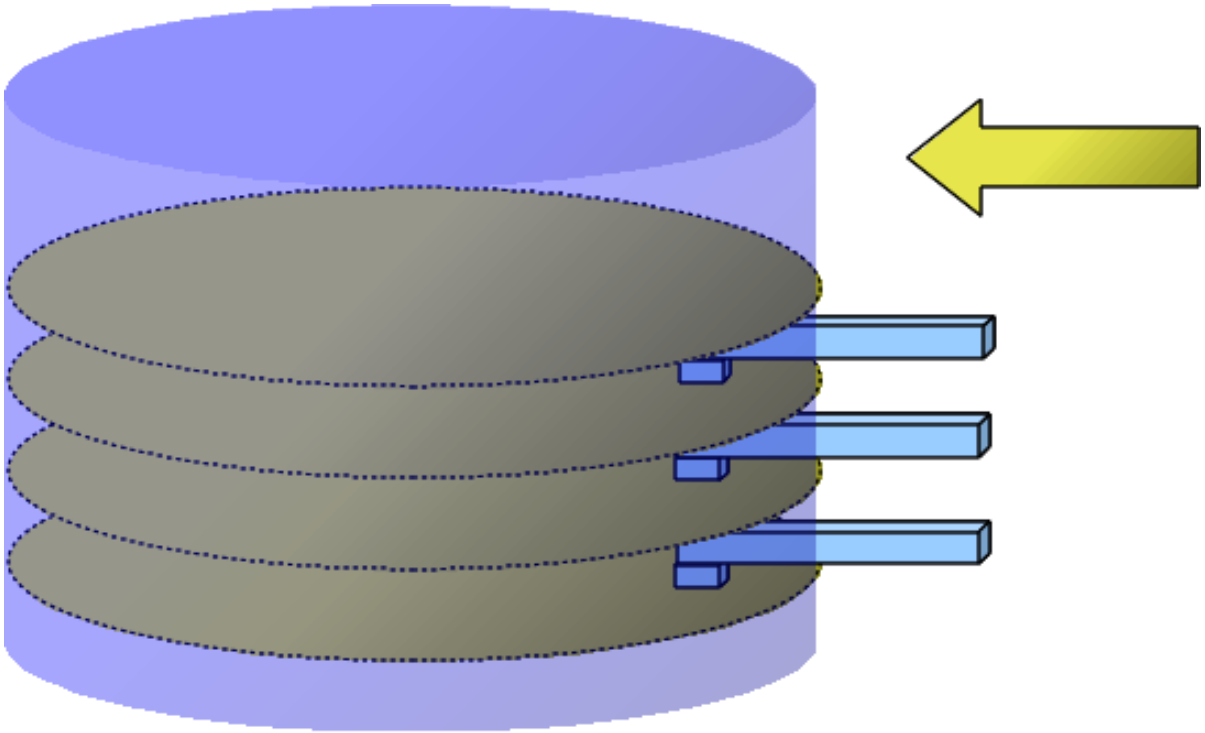
HDD components

Read/Write heads supported bye slider suspension assembly which are moved by some actuators in radial direction.



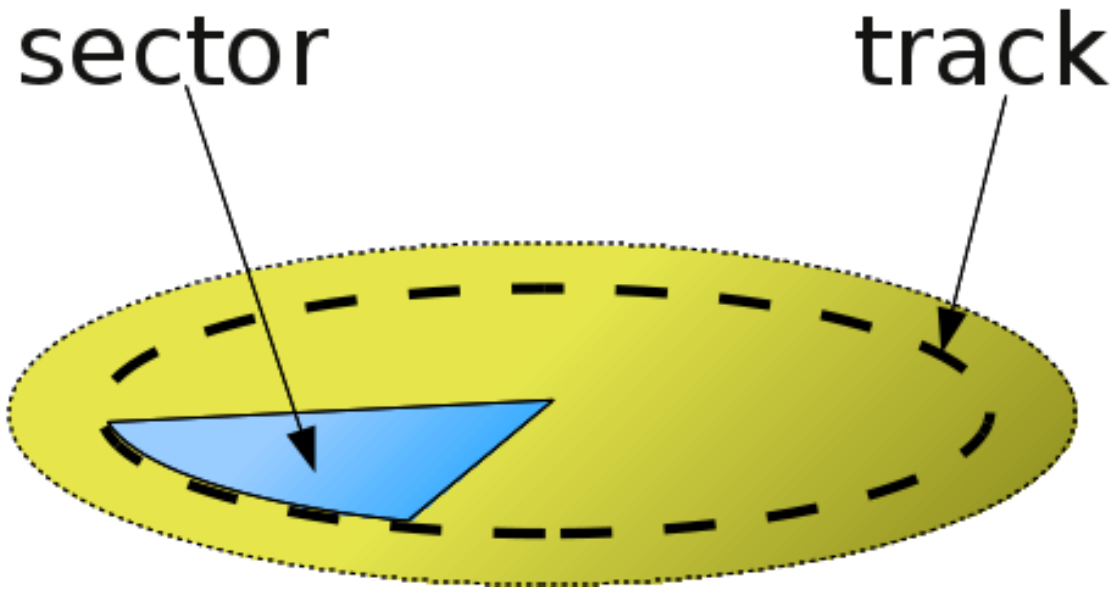
HDD components

Read/Write heads supported bye slider suspension assembly which are moved by some actuators in radial direction.



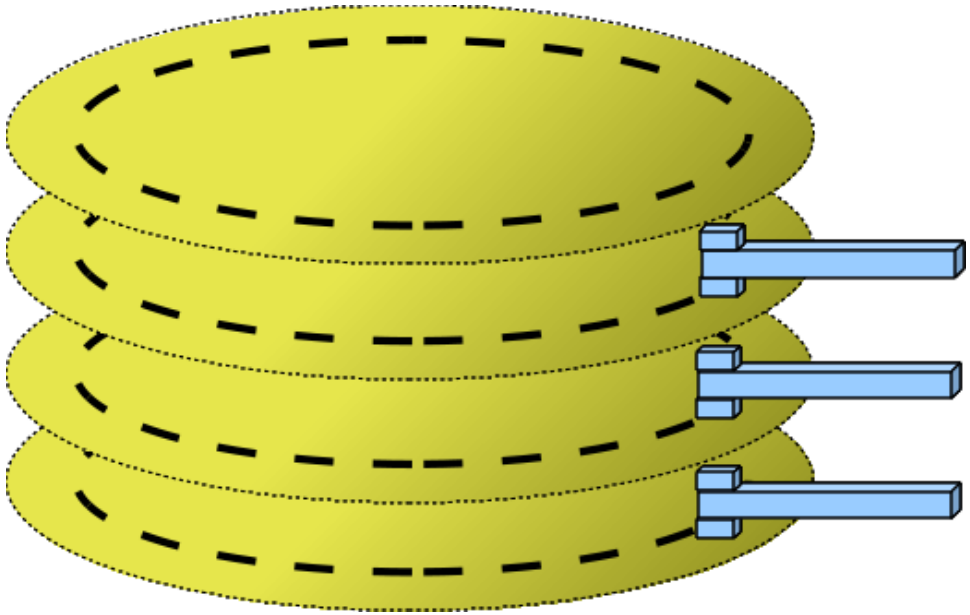
HDD components

We can identify, on each plate, specific zones.



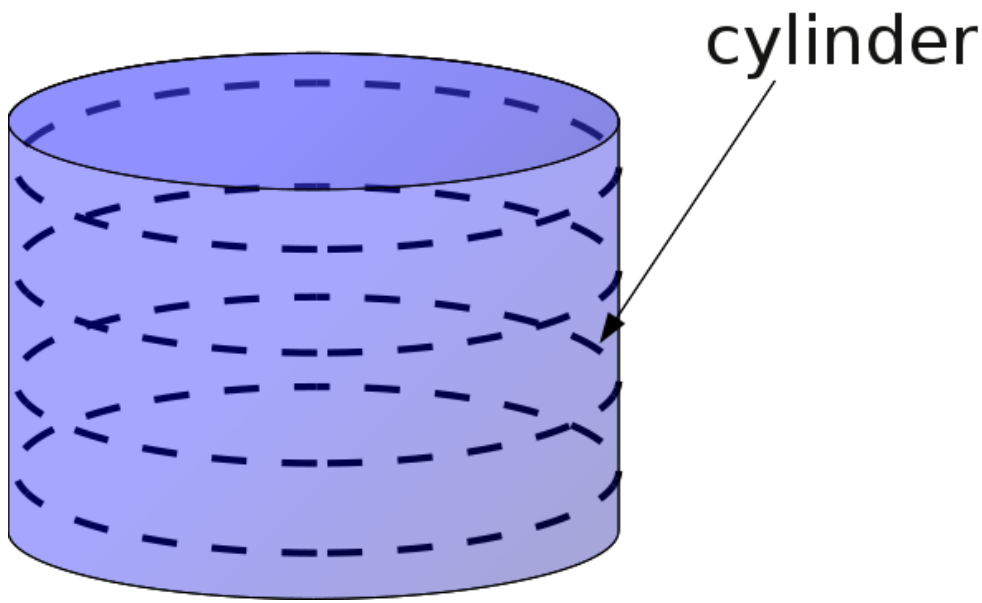
HDD components

If we see one specific track on all plates, we can see a cylinder.



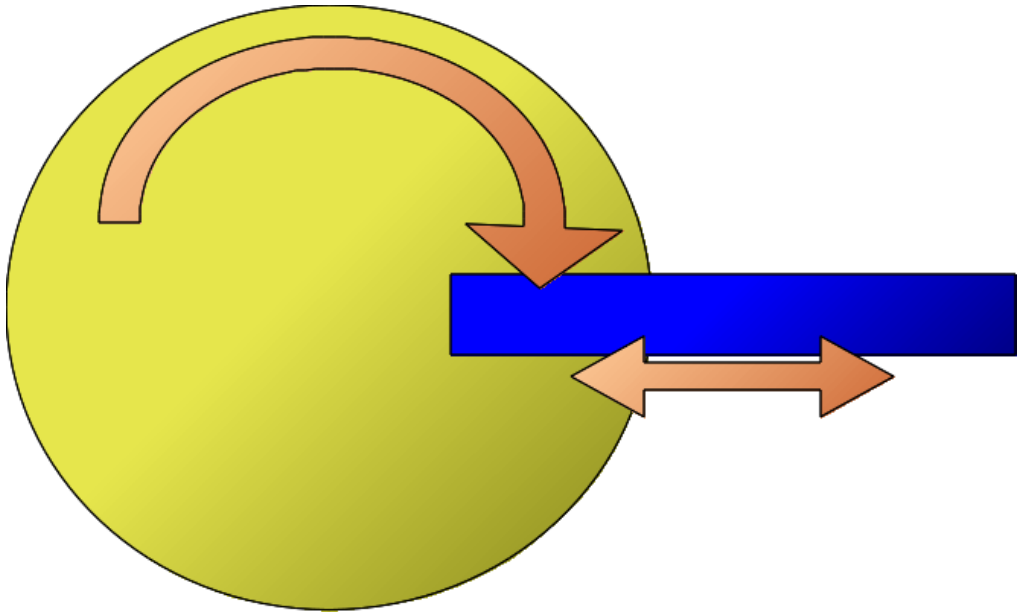
HDD components

If we see one specific track on all plates, we can see a cylinder.



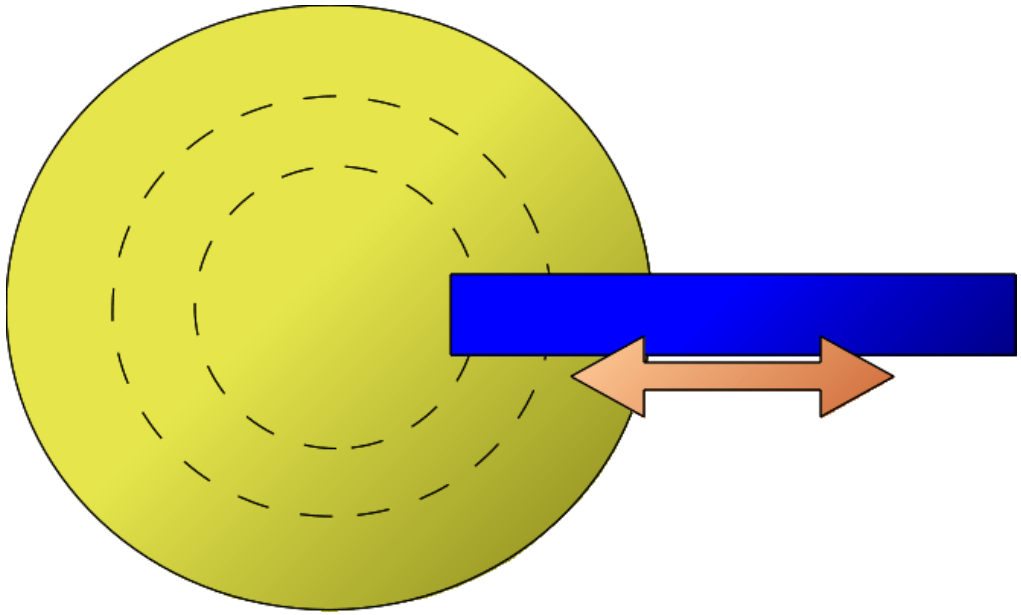
HDD components

The policy used by the heads, to respond every read/write operation is similar to the policy used by a lifter.



HDD components

In this way, the most visited zone of the plate is the zone in the middle.



Redundant Array of Independent Disks

- One way to improve bandwidth is to define a logical device which consists of multiple disks.
- With this sort of approach a single I/O transaction can simultaneously move blocks of data to multiple disks.
- For example, if a logical device is created from eight disks, each of which is capable of sustaining 10 MB/sec, then this logical device is capable of delivering up to 80 MB/sec of I/O bandwidth.

Redundant Array of Independent Disks

We can get some better performance using RAID.

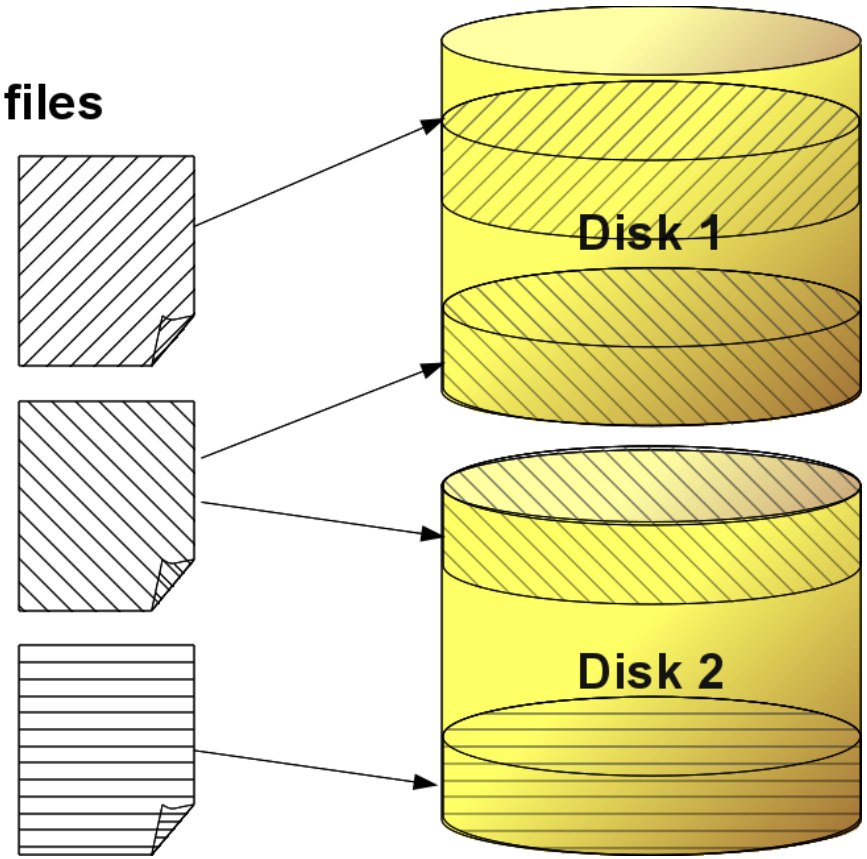
- S: Hard disk drive size.
- N: Number of hard disk drives in the array.
- P: Average performance of a single hard disk drive (MB/seg).

Linear RAID

$Performance = P$

No redundancy

$Capacity = N * S$

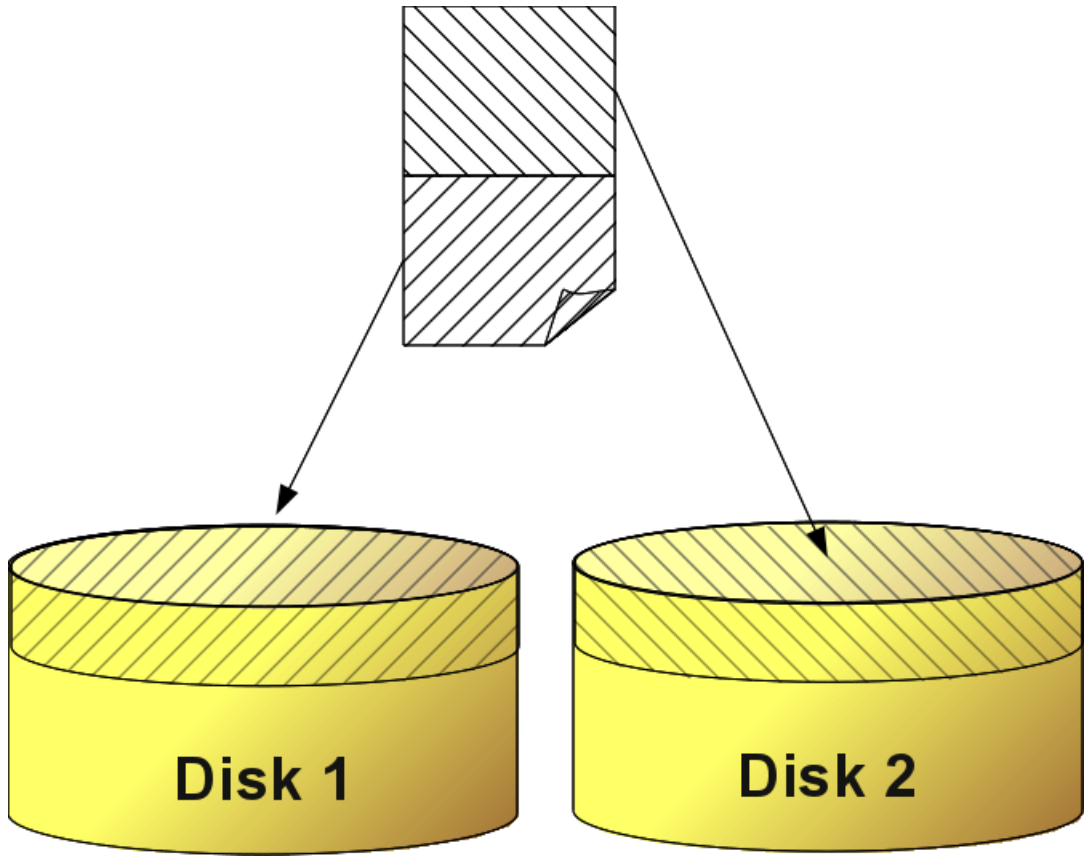


RAID 0

$Performance = P * N$

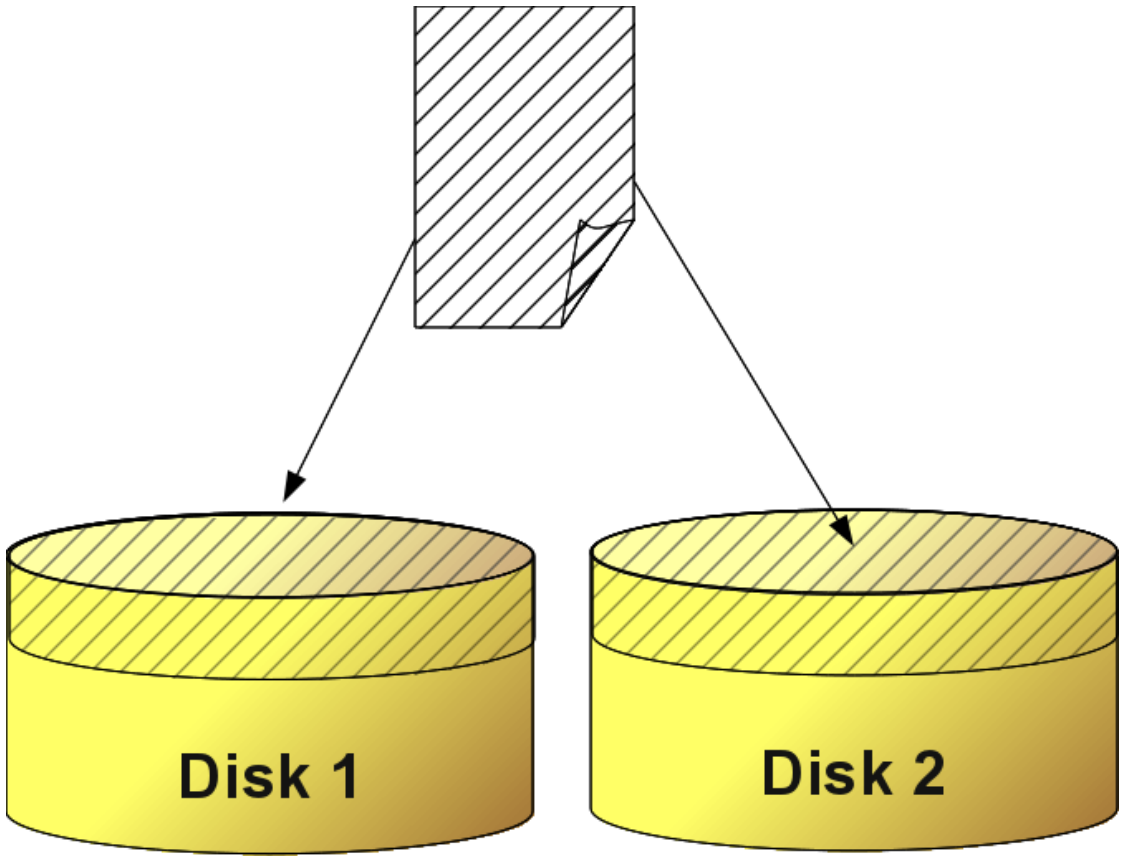
Striping

$Capacity = N * S$



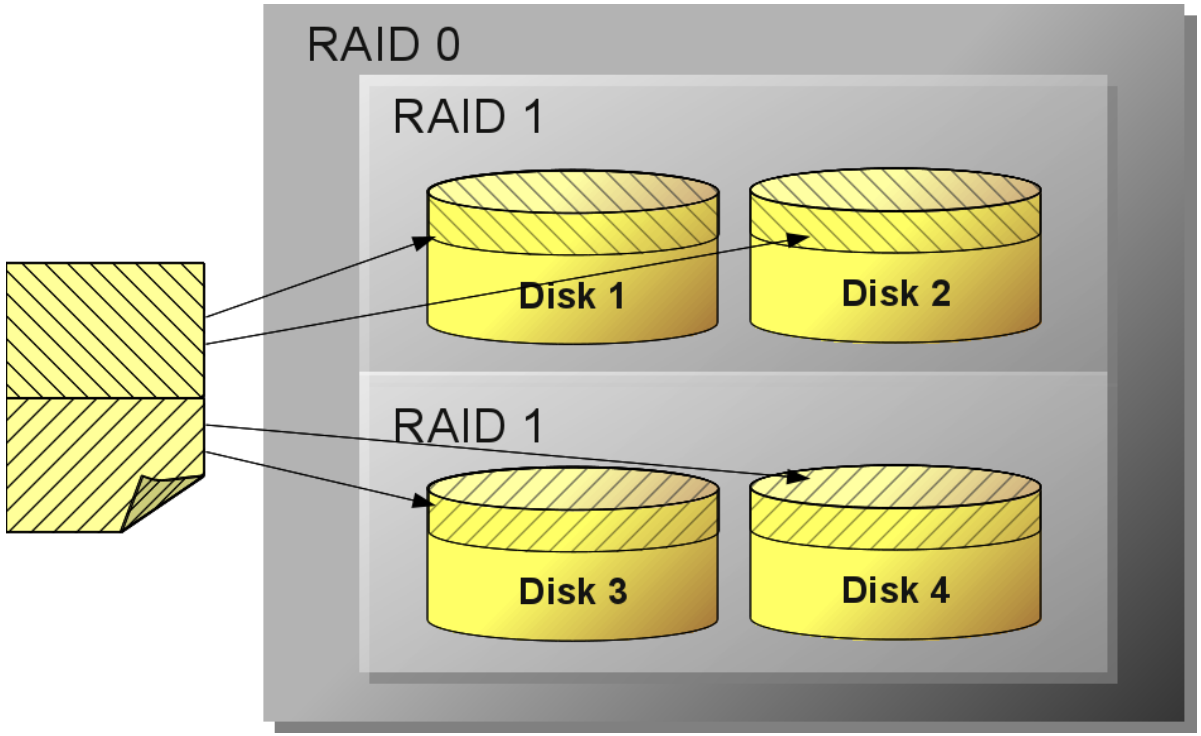
RAID 1

$WritePerf. = P$
$ReadPerf. = P * N$
Redundancy
$Capacity = S$



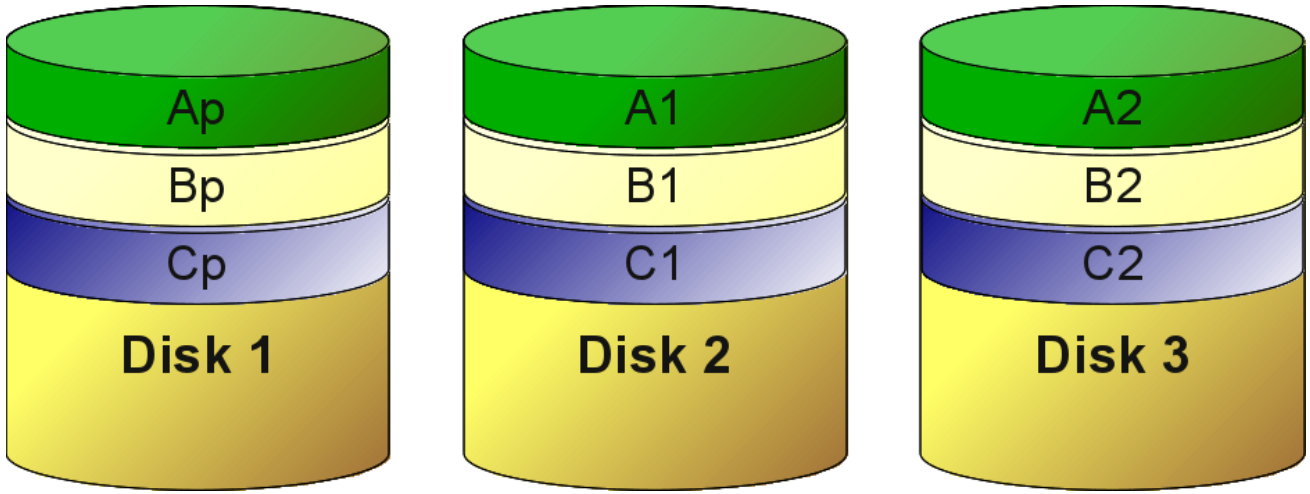
RAID 10

Redundancy
Striping



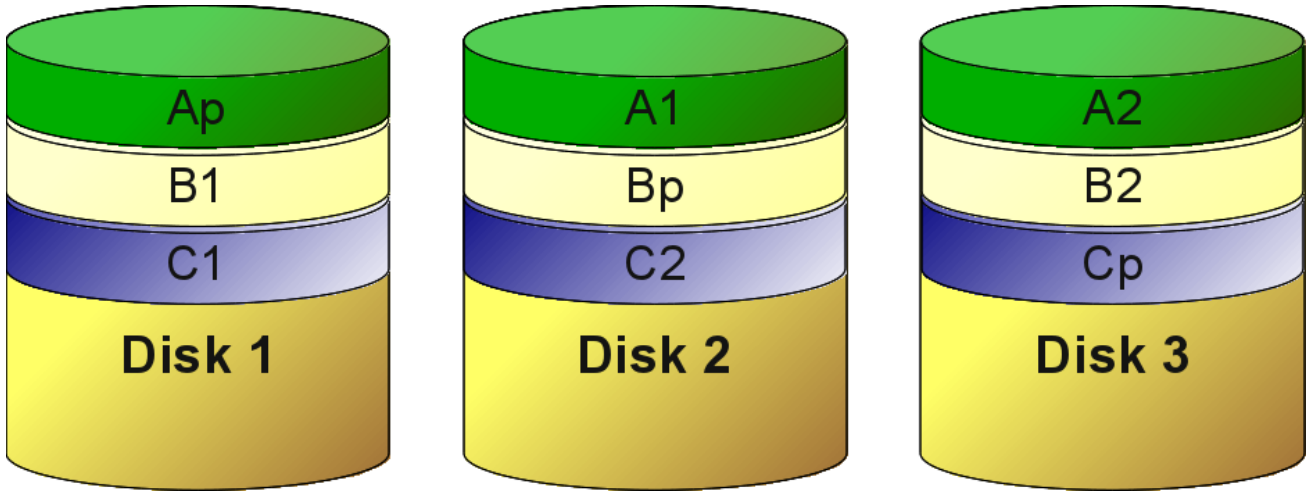
RAID 4

Parity Disk Bottleneck



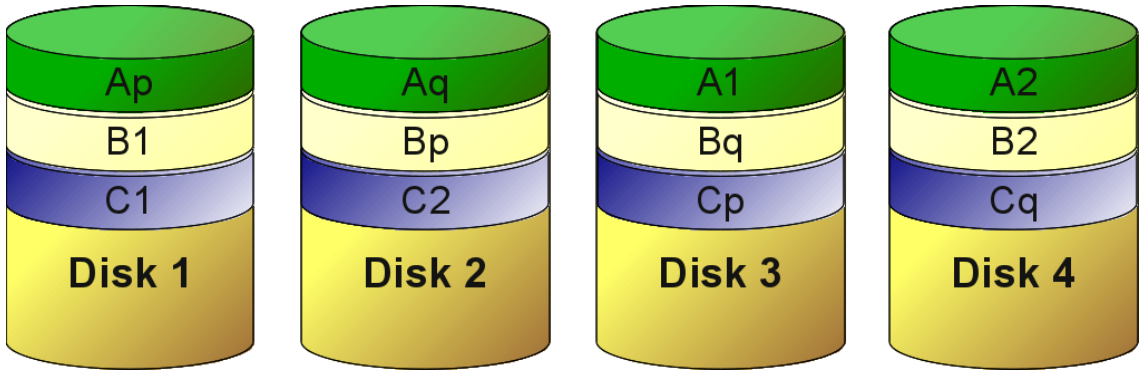
RAID 5

One disk can fail
Distributed parity



RAID 6

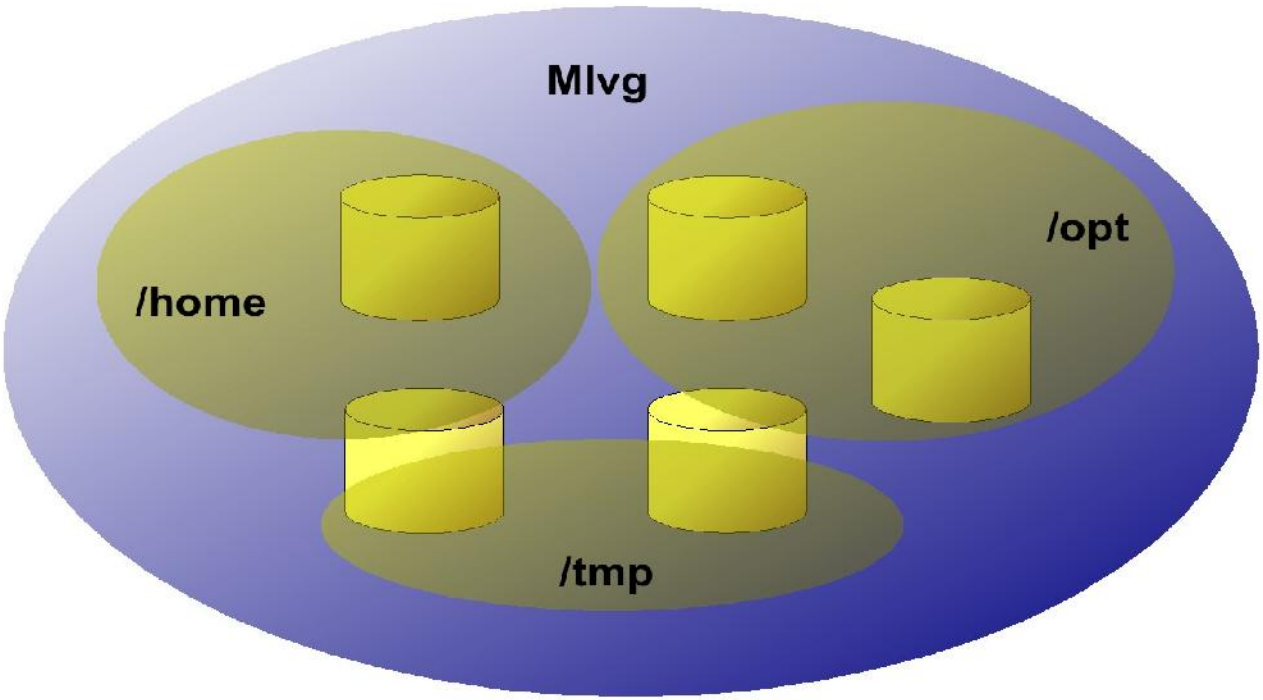
Two disk can fail
Double parity code



Logical Volume Manager

From phisical devices we can create:

- Volume groups (Mlvg)
- Logical volumes (logical partitions): /home, /opt, /tmp



Software RAID

Several Operating Systems provide software RAID. For instance, Linux provides

- Linear RAID
- RAID 0
- RAID 1
- RAID 1+0
- RAID 5
- RAID 6

File System

A file system is a set of methods and data structures used to organize, store, retrieve and manage information in a permanent storage medium, such as a hard disk. Its main purpose is to represent and organize resources storage.

File System

The elements of a file system are:

- Name space: It is a way to assign names to the items stored and organize them hierarchically.
- API: Is a set of calls that allow the manipulation of stored items.
- Security Model: Is a scheme to protect, hide and share data.
- Implementation: Is the code that couples the logical model to the storage medium.

Basic Concepts

- Disk: A permanent storage medium of a certain size.
- Block: The smallest unit writable by a disk or file system. Everything a file system does is composed of operations done on blocks.
- Partition: A subset of all the blocks on a disk.
- Volume: The term is used to refer to a disk or partition that has been initialized with a file system.
- Superblock: The area of a volume where a file system stores its critical data.

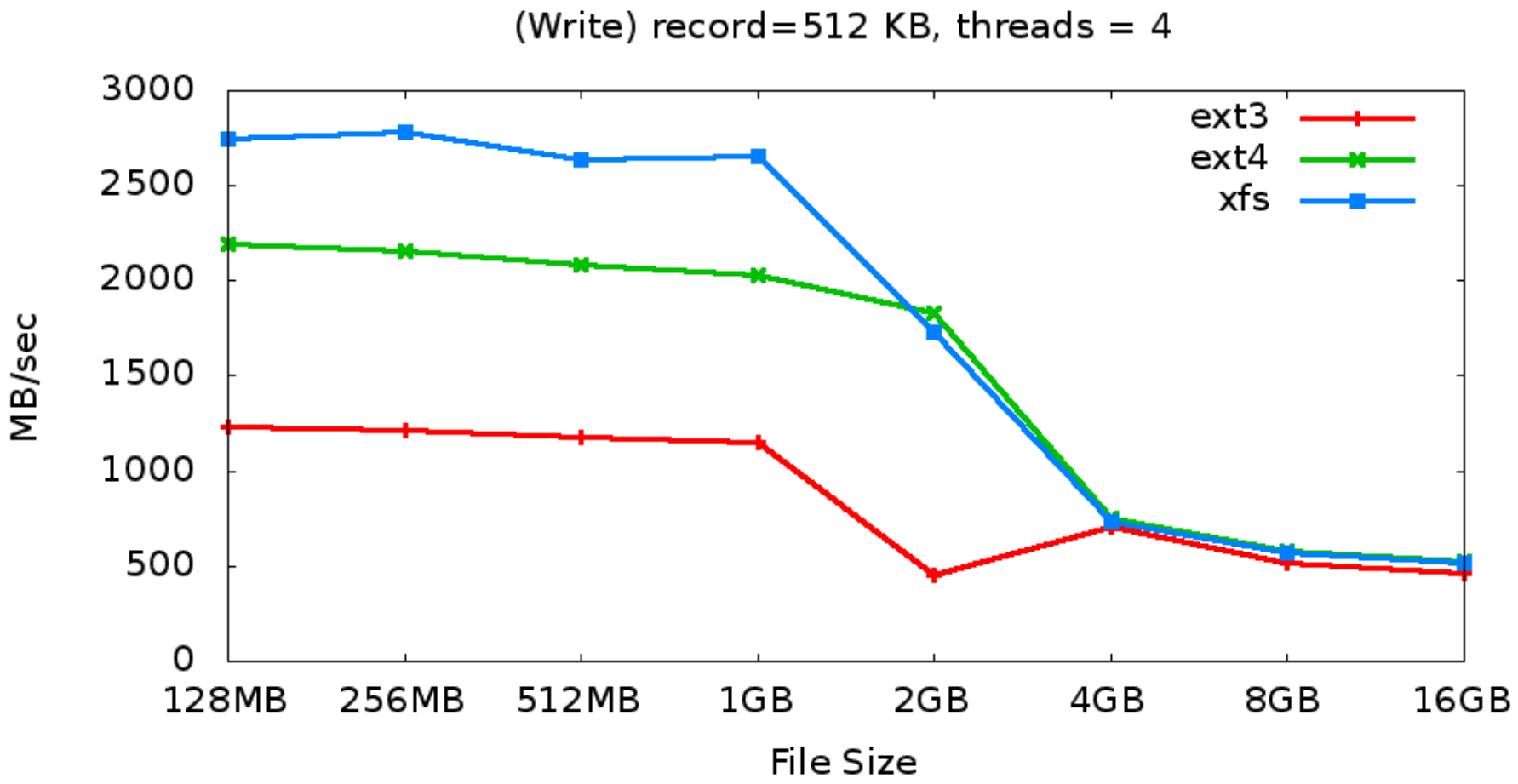
Basic Concepts

- Metadata: A general term referring to information that is about something but not directly part of it. For example, the size of a file is very important information about a file, but it is not part of the data in the file.
- Journaling: A method of insuring the correctness of file system metadata even in the presence of power failures or unexpected reboots.
- Attribute: A name and value associated with the name. The value may have a defined type (string, integer, etc.).

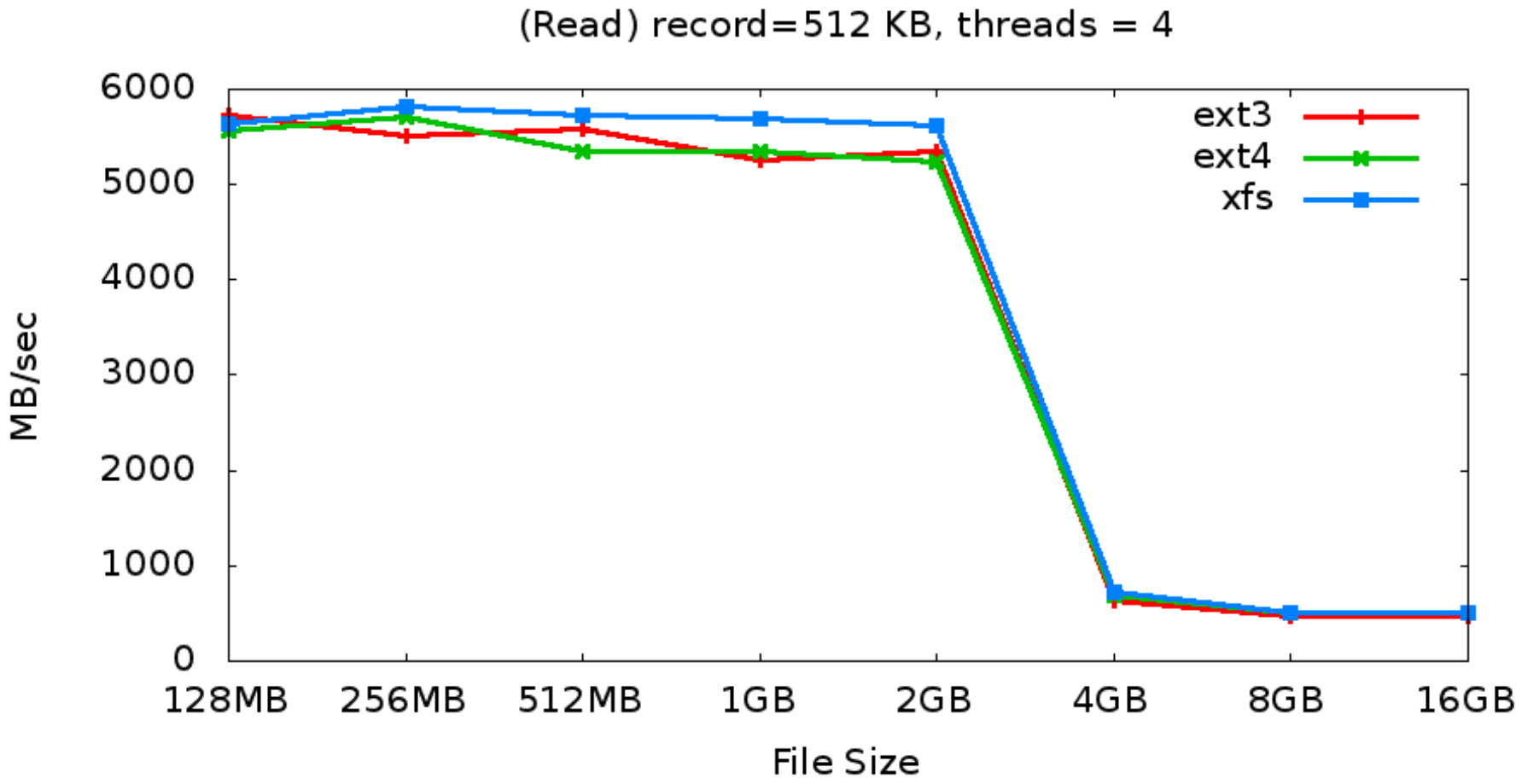
File Systems

- FAT
- NTFS
- Ext2, Ext3, Ext4
- Reiserfs, xfs, jfs

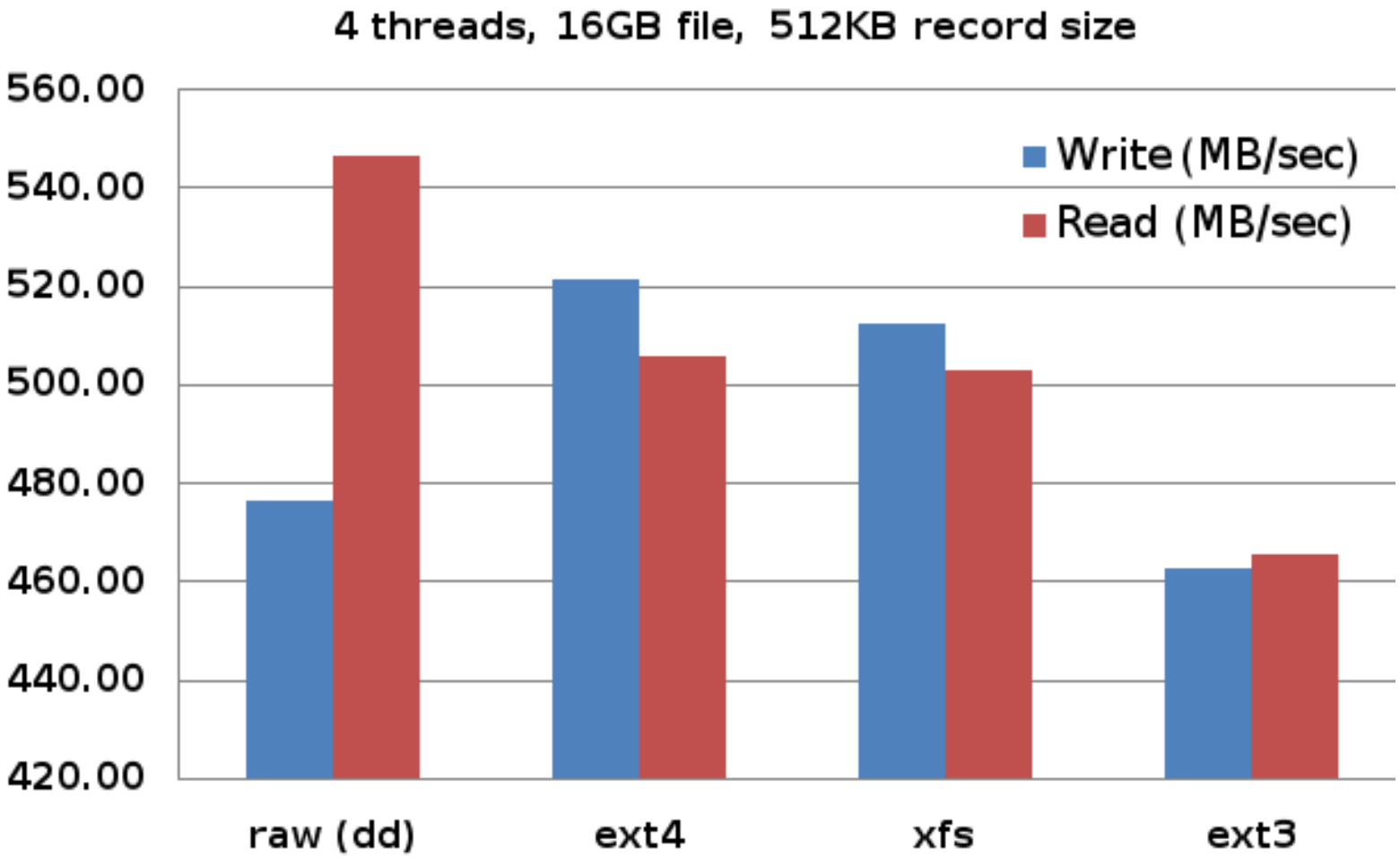
Some Performance Tests



Some Performance Tests

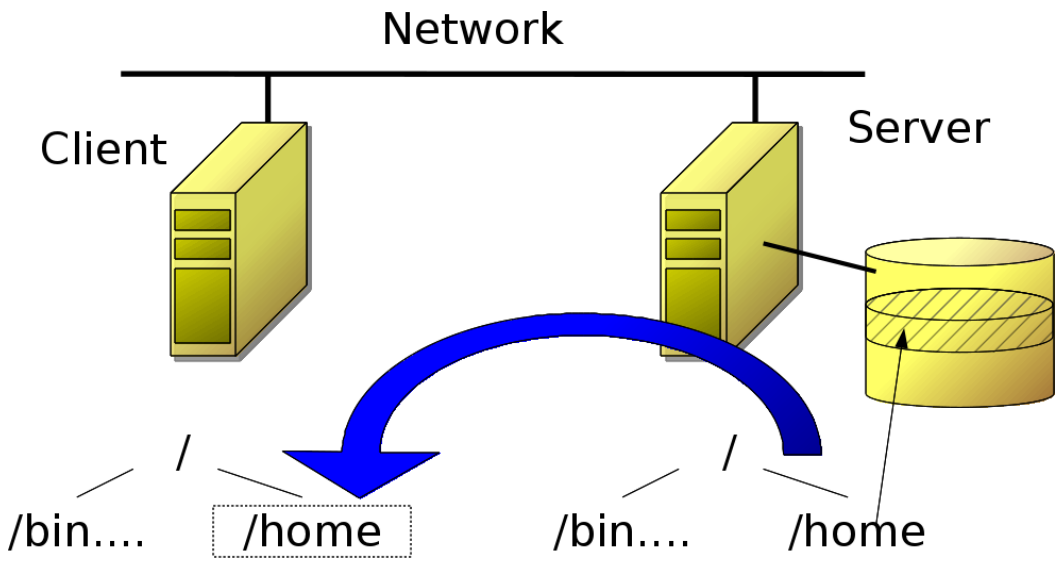


Some Performance Tests



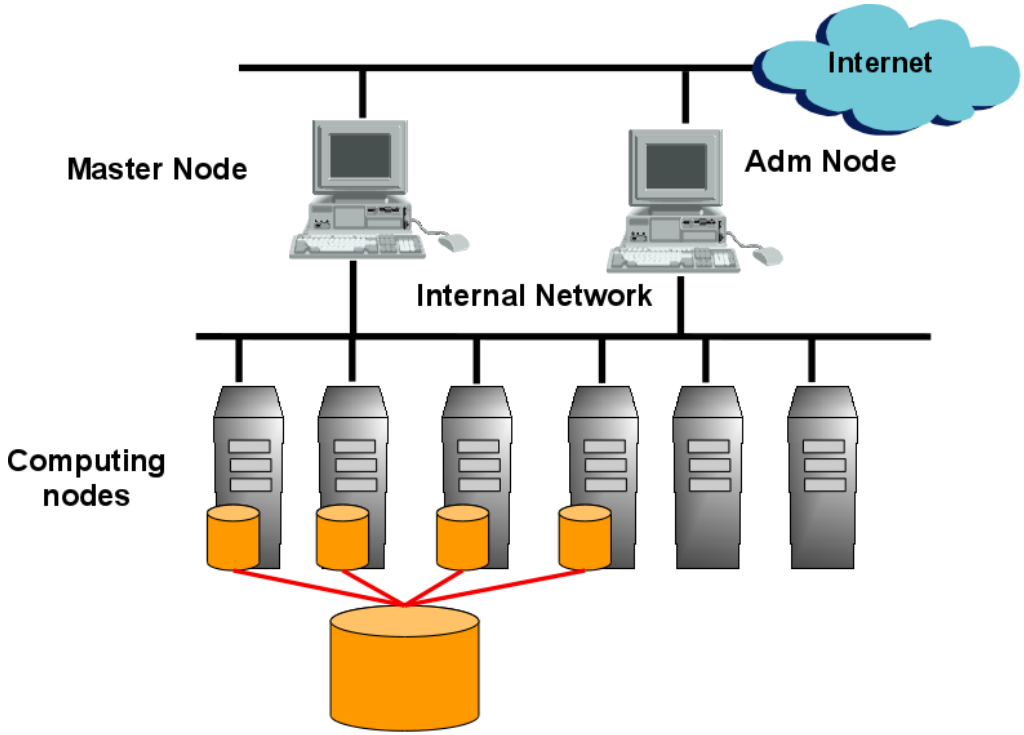
Distributed File Systems

A distributed file system makes available the data across the network:
NFS, coda, AFS, etc.



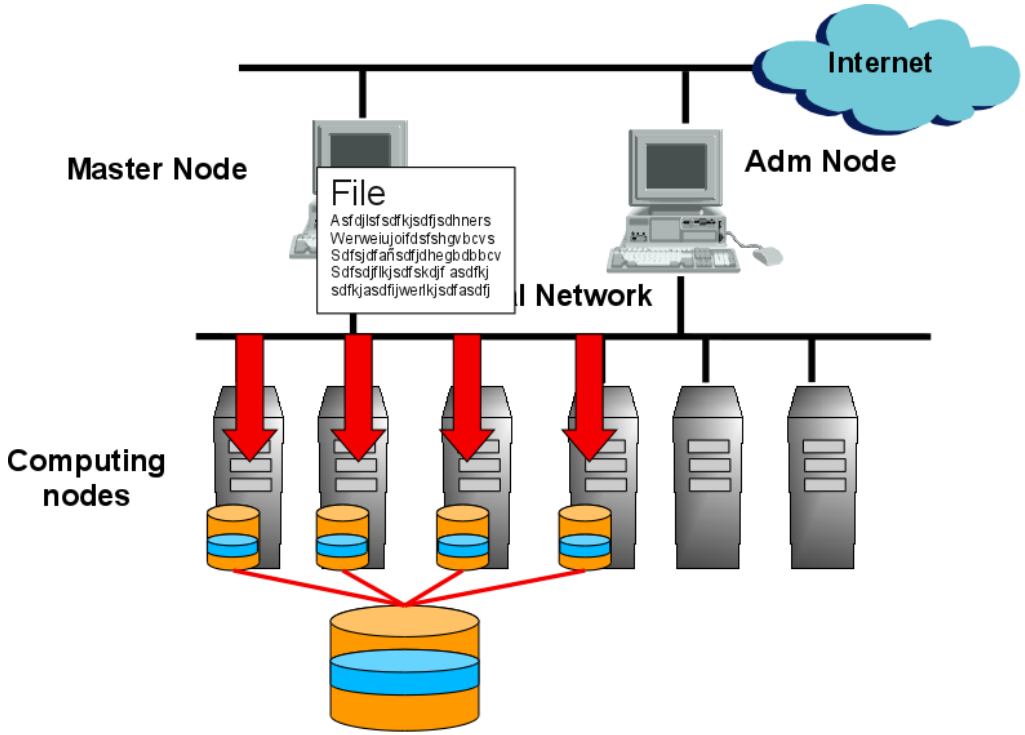
Parallel File Systems

A parallel file system leverage all disks available across the network



Parallel File Systems

A parallel file uses all IO controllers at the same time.



Parallel File Systems

We can find several implementations of Parallel File Systems:

- Lustre from Cluster File System, then from Sun, now from Oracle (GPL)
- GPFS from IBM (Private)
- PVFS from ANL (GPL)
- GlusterFS from Gluster Inc. (GPL)
- PanFS from Panasas (Private)
- IBRIX from Ibrix (Private).
- Etc.

Parallel File Systems

In general, a Parallel File Systems has the following components

- Metadata Server
- IO Servers
- Clients