MHPC
Master in High Performance Computing

Introduction to data management and storage for HPC (part 1)

Stefano Cozzini

CNR/IOM and eXact-lab srl

SISSA
Scuola Internazionale Superiore di Studi Avanzati

ICTP The Abdus Salam International Centre for Theoretical Physics

MHPC
Master in High Performance Computing

Agenda

- Introducing the big-data problem
- Data intensive science
- Data management considerations
- Data infrastructures and tools
- Basic concepts on storage

SISSA ICTP

MHPC
Master in High Performance Computing

Aim of this introduction

- Frame the problem and the discussion around DATA:
  - What are Big Data ?
  - Which kind of challenges ahead of us ?
- Highlight different topics on data management
- Give you a short recap/introduction on basic concepts about storage

SISSA ICTP

MHPC
Master in High Performance Computing

INTRODUCING BIG DATA PROBLEM

SISSA
ICTP

MHPC
Master in High Performance Computing

The 3 original V's of big data..

- Velocity

Data are produced at speed higher than the speed you are able to move/analyze and understand them..

- Variety
  - Data range from simulation to remote sensing information, from instruments to market analysis etc..
  - datasets come in a variety of data formats and span a variety of metadata standards
- Volume

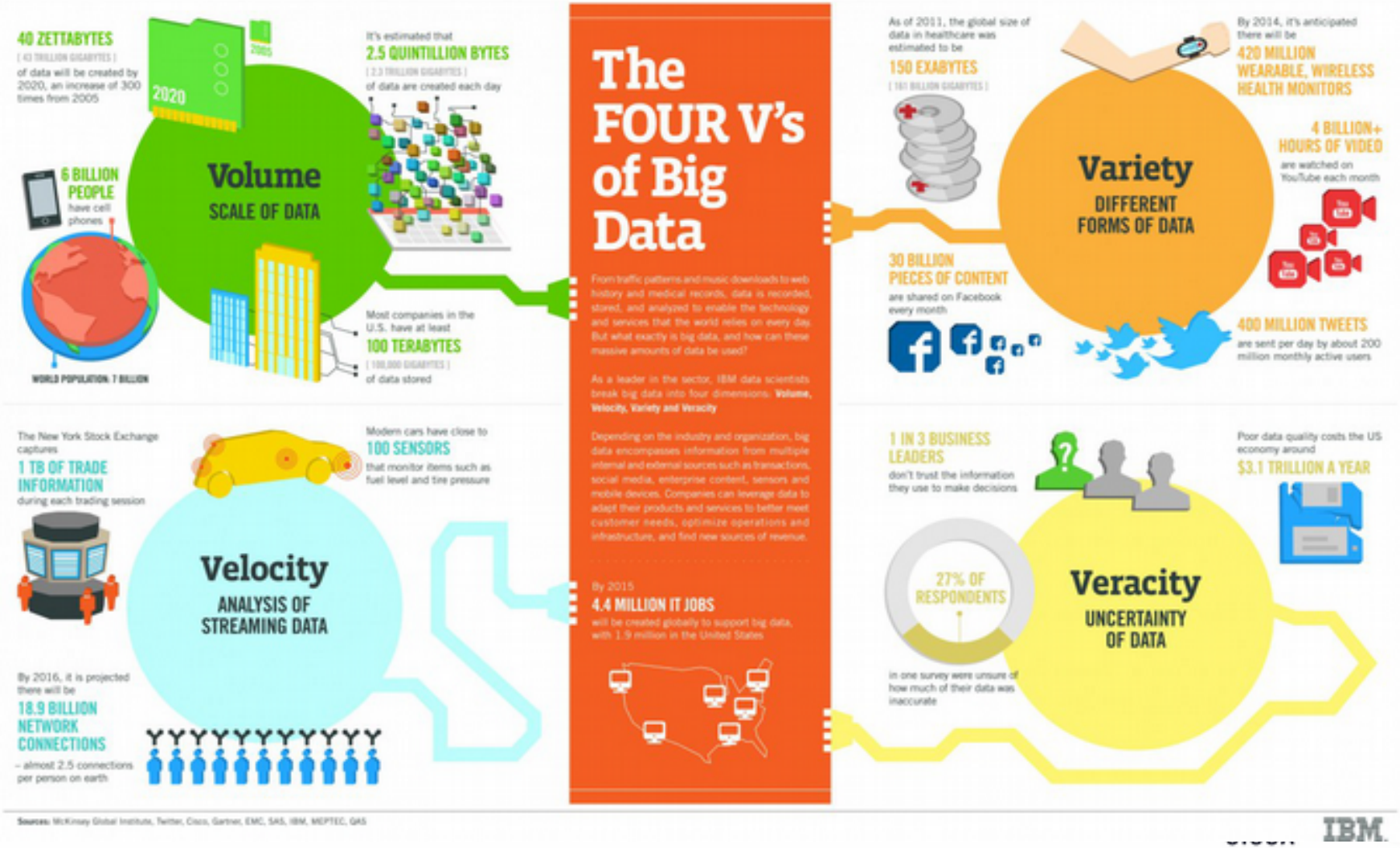
"From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days… and the pace is accelerating"

Veracity: refers to the trustworthiness of the data. Can we rely on the fact that the data is representative? There are inherent discrepancies in all the data collected and these should be understood and/or removed.

SISSA ICTP

MHPC
Master in High Performance Computing

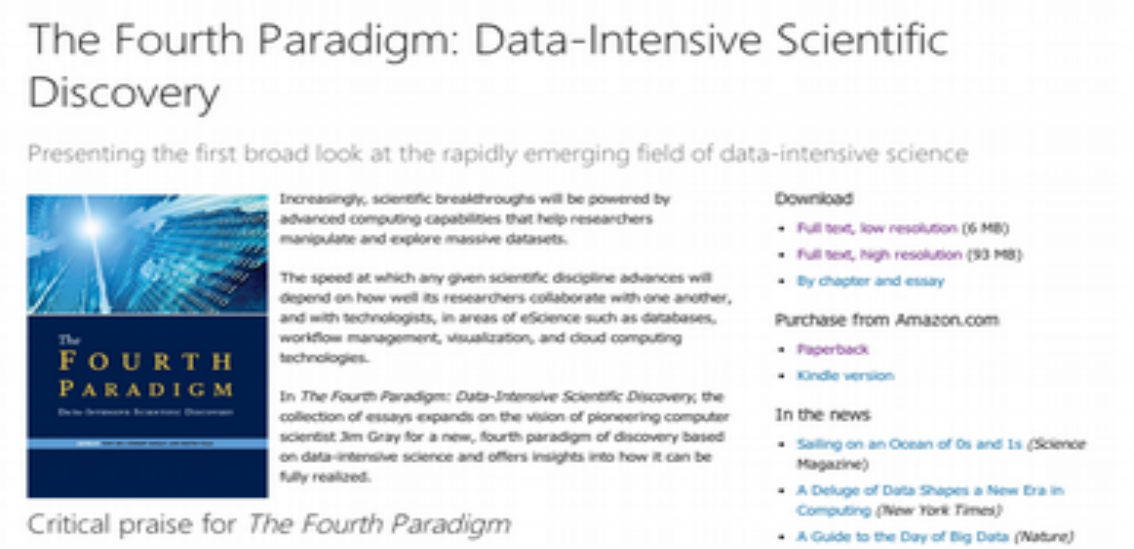
IBM'S infographics



MHPC
Master in High Performance Computing

Data-intensive science

- A "fourth paradigm" after experiment, theory, and computation..



It involves collecting, exploring, visualizing, combining, subsetting, analyzing, and using huge data collections

SISSA ICTP

MHPC
Master in High Performance Computing

Importance of data management in science

'Editorial: Data's Shameful Neglect' (10 September 2009) in Nature 461, p. 145, doi:10.1038/461145a.

"Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly".

"Data management should be woven into every course in science, as one of the foundations of knowledge"

SISSA ICTP

MHPC
Master in High Performance Computing

Challenges & Requirements

Challenges:

- Deluge of observational data, “exaflood” of simulation model outputs
- Need for collaboration among groups, disciplines, communities
- Finding insights and discoveries in a “Sea of Data/Data lake”

Requirements:

- New tools, techniques, and infrastructure
- Standards for interoperability
- Institutional support for data stewardship, curation

SISSA ICTP

MHPC
Master in High Performance Computing

Which kind of data ?

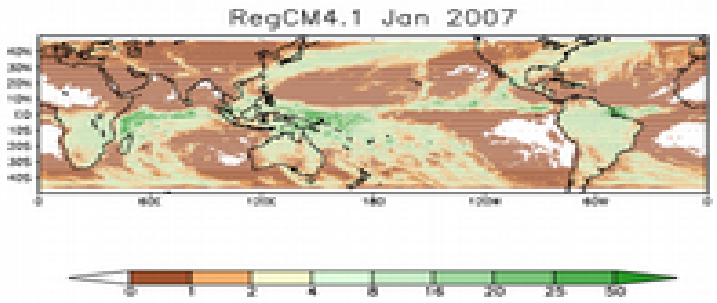
- Observational: data captured in real time, usually unique and irreplaceable e.g. brain images, survey data
- Experimental: data from experimental results, e.g. from lab equipment, often reproducible, but can be expensive e.g. chromatograms
- Simulation: data generated from test models where model and metadata may be more important than output data from the model e.g. economic or climate models
- Derived or compiled: resulting from processing or combining 'raw' data, often reproducible but expensive e.g. compiled databases, text mining, aggregate census data
- Reference or canonical: a (static or organic) conglomeration or collection of smaller (peer reviewed) datasets, most probably published and curated e.g. gene databanks, crystallographic databases

SISSA ICTP

MHPC
Master in High Performance Computing

Data Deluge in science

- A local example: climate change with RegCM4
  - Output generated is:
    - X 32bit variables for X*Z*Y point of the domain every T hours of simulation for N years of simulation
  - One large example currently studied at ICTP Equator Belt
    - Number of variables: ~ 50
    - domain: 832x250x18 points
    - Frequency: 3 hours
    - Length : 150 years



$50 * 832 * 250 * 18 * 6 * 365 * 150 * 4 = 2.459808 \times 10^{14}$!!! ~250 TB !!!

SISSA ICTP

MHPC
Master in High Performance Computing

A new sector: High-Performance Data Analysis

- simulation-based and analytics-based data analysis are complex enough to require the use of high-performance computing (HPC) methods and resources (on-premise or in the cloud).

SISSA ICTP

MHPC
Master in High Performance Computing

Data vs Metadata..

- "Data" implies accompanying metadata (e.g. precise definitions of quantities, equations of interrelationships, scientific units of measurement, error analysis, etc.)
- In experimental sciences the data is all the information required to repeat the experiment and the resulting data reported from that experiment.
- In data-driven sciences the data is the methodology of data collection and the contents of the database at a given time.
- In computational science the data is the program used to compute the results, the parameterisation of the program and the results of the calculation.

SISSA ICTP

MHPC
Master in High Performance Computing

The FAIR principles

Data should be:

- Findable
- Accessible
- Interoperable
- Reusable
- From: http://www.nature.com/articles/sdata201618

SISSA ICTP

MHPC
Master in High Performance Computing

FAIR principles

Findable

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

SISSA ICTP

MHPC
Master in High Performance Computing

FAIR principles

Accessible

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

SISSA ICTP

MHPC
Master in High Performance Computing

FAIR principles

Interoperable

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

SISSA ICTP

MHPC
Master in High Performance Computing

FAIR principles

Reusable

- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
- R1.1. (meta)data are released with a clear and accessible data usage license
- R1.2. (meta)data are associated with detailed provenance
- R1.3. (meta)data meet domain-relevant community standards

SISSA ICTP

MHPC
Master in High Performance Computing

Scalability and "Big Data"

- What's the big deal about big data?
  - aren't more and faster computers and larger disks the solution?
- The missing parts:
  - Network
    - What about data transfer ?
  - Software
    - What about the software ?
  - People
    - Who is maintaining the infrastructure ?

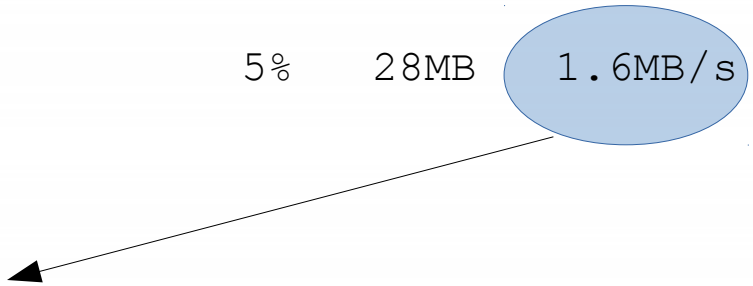
SISSA ICTP

MHPC
Master in High Performance Computing

Network issues..

- Moving 13 TB of data from HPC center in Gorjansko to ICTP Trieste..

```
[exact@arctur1 2008]$ scp air.2008.18.nc cozzini@democritos.sissa.it:
Password:
air.2008.18.nc                                5%   28MB   1.6MB/s   04:53
ETA
```



13,000,000 MB / 1.6 MB ~ 94 days
!!!

SISSA ICTP

MHPC
Master in High Performance Computing

Data infrastructure crucial aspects

Large datasets:

Deal with large datasets and large data rates from experiments

Reliability :

Increase the level of QoS and SLA, e.g. increasing reliability by replicating data sources and increasing accessibility by copying source to several places

Accounting

Allow monitoring and checking of resource usage

Integration

Provide the same set of services that are understandable (compatible) between domains

Interoperation

Interoperation through common standard schemes

Access

Broadband data access Allow transparent and secure remote access to data

Source: e-irg blue paper on data management 2012
http://www.e-irg.eu/images/stories/dissemination/e-irg-blue_paper_on_data_management_v_final.pdf

SISSA ICTP

MHPC
Master in High Performance Computing

Data infrastructure crucial aspects

Data preservation

Allow long-term availability of data

High quality

Quality of data to enable advanced and cross-disciplinary access and enrichment operations

Economic justification

As the scientific community is operating on increasingly larger datasets and want to preserve the information concerned, the infrastructure provided should have a clear roadmap of technology exchange and backwards compatibility.

Access control

Provide the infrastructure to allow for fine-grained access control

SISSA ICTP

MHPC
Master in High Performance Computing

Plenty of nice projects on data management and data repository

SISSA ICTP

MHPC
Master in High Performance Computing

Some specific purpose data repository

- NFFA- IDRP (by CNR/IOM)
- EUSMI project (infrastructure lead by CNR/IOM)

Thesis Projects available

please contact me if interested

SISSA ICTP

MHPC
Master in High Performance Computing

HPC AND I/O

SISSA
ICTP

MHPC
Master in High Performance Computing

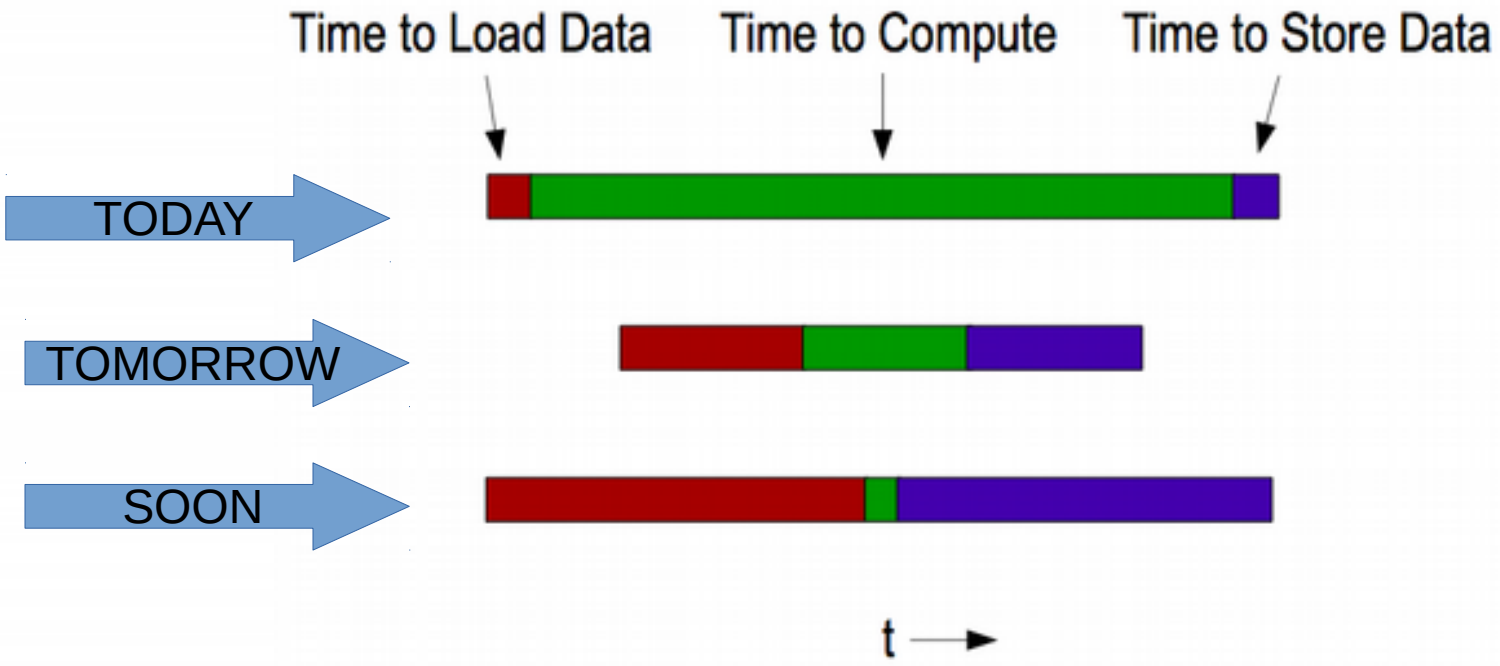
IOPS vs FLOPS

- HPC is today compute-centric
- But:
  - You can only compute as fast as you can move data
- Then:
  - scientific computing needs data accessibility rather than computing speed
  - HPC workflow will be soon be bounded by the speed of the storage system..

Source: IDC Direction 2013

SISSA ICTP

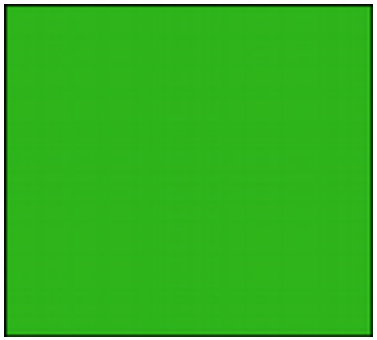
"A supercomputer is a device for converting a CPU-bound problem into an I/O bound problem." [Ken Batcher]



Energy consideration

computing 1 calculation
≈ 1 picojoule

moving 1 calculation
≈ 100 picojoule



Source: IDC Direction 2013

MHPC
Master in High Performance Computing

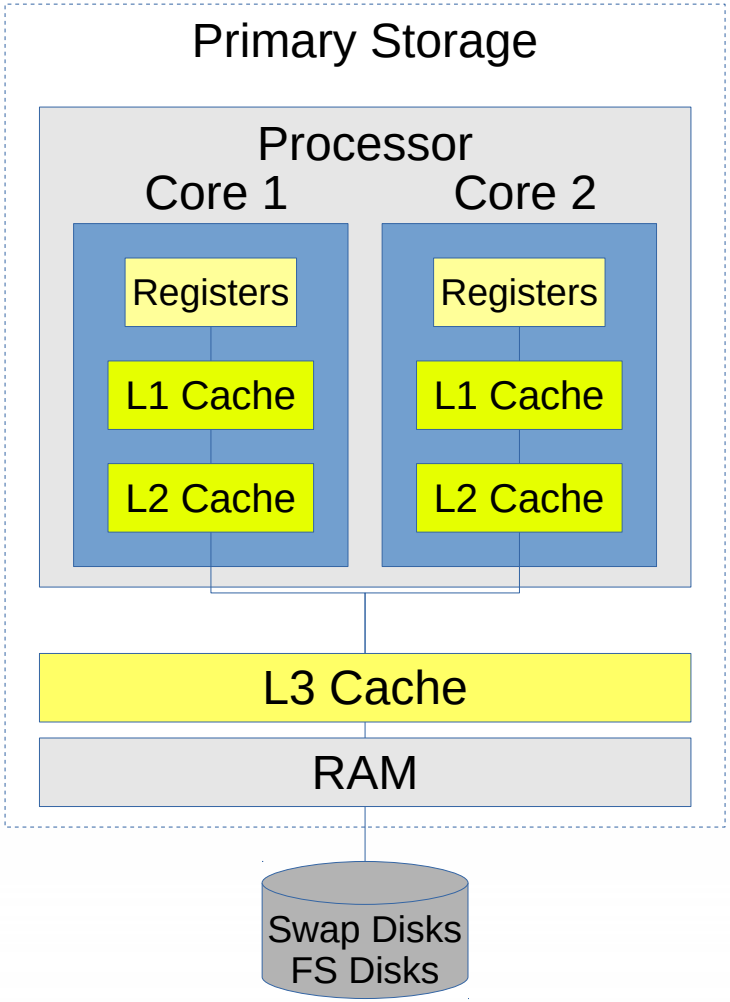
storage 101

SISSA
ICTP

Memory Hierarchy

Computers architectures try to keep data close to the processors in order to feed them continuously.

However, while the capacity of storage devices increases, the distance to the processors also increases.



Internal Memory - processor registers and cache
Main Memory - system RAM and controller cards
On-line mass storage - secondary storage
Off-line bulk storage - tertiary and off-line storage

Storage Hierarchy

Same as with the memory hierarchy of Register -> Cache (L1->L2->L3) -> RAM storage follows a hierarchy with multiple levels:

- RAM disk, I/O buffers or file system cache
- Local disk (flash based, spinning disk) (SATA, SAS, RAID, SSD, JBOD, ...)
- Local network attached device or file system server (NAS, SAN, NFS, CIFS, Lustre, GPFS, ...)
- Tape based archival system (often with disk cache)
- External, distributed file systems (Cloud storage)

Key metrics

- Bandwidth: volume of data read/written in a second

  → throughput metric

- IOPs: number of IO request processed by second

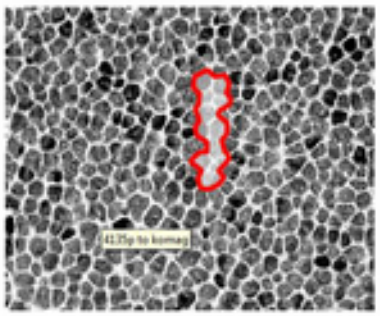
  → Is it a latency or a throughput metric ?

- Order of magnitudes
  - Intel v2/v3 CPU-DRAM: 60 GB/s
  - IB link: 5-10 GB/s
  - Hard Drive: ~100- 150 MB/s

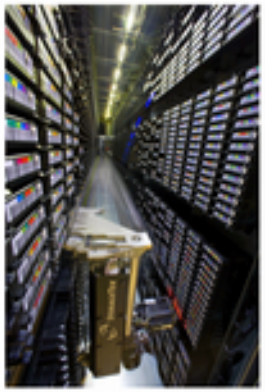
Cache/Swap

- Disk I/O is much slower than main memory I/O, typically about a 100x (varied with hardware):
  - typically applications use buffers (libc/stdio)
- In typical workloads certain data is accessed repeatedly **beyond** an application lifetimes:
  - OS maintains buffer of recently used data
  - buffer competes with applications for RAM
  - OS can substitute swap disk for RAM
- Memory management unit (MMU) organizes address space in pages

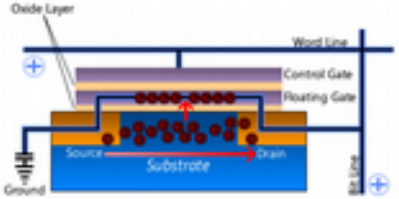
Building storage system from bits



Magnetic or Solid State storage bits



Storage Devices



Software to aggregate many devices for performance

Software to handle device failures (erasure codes on blocks, across devices)

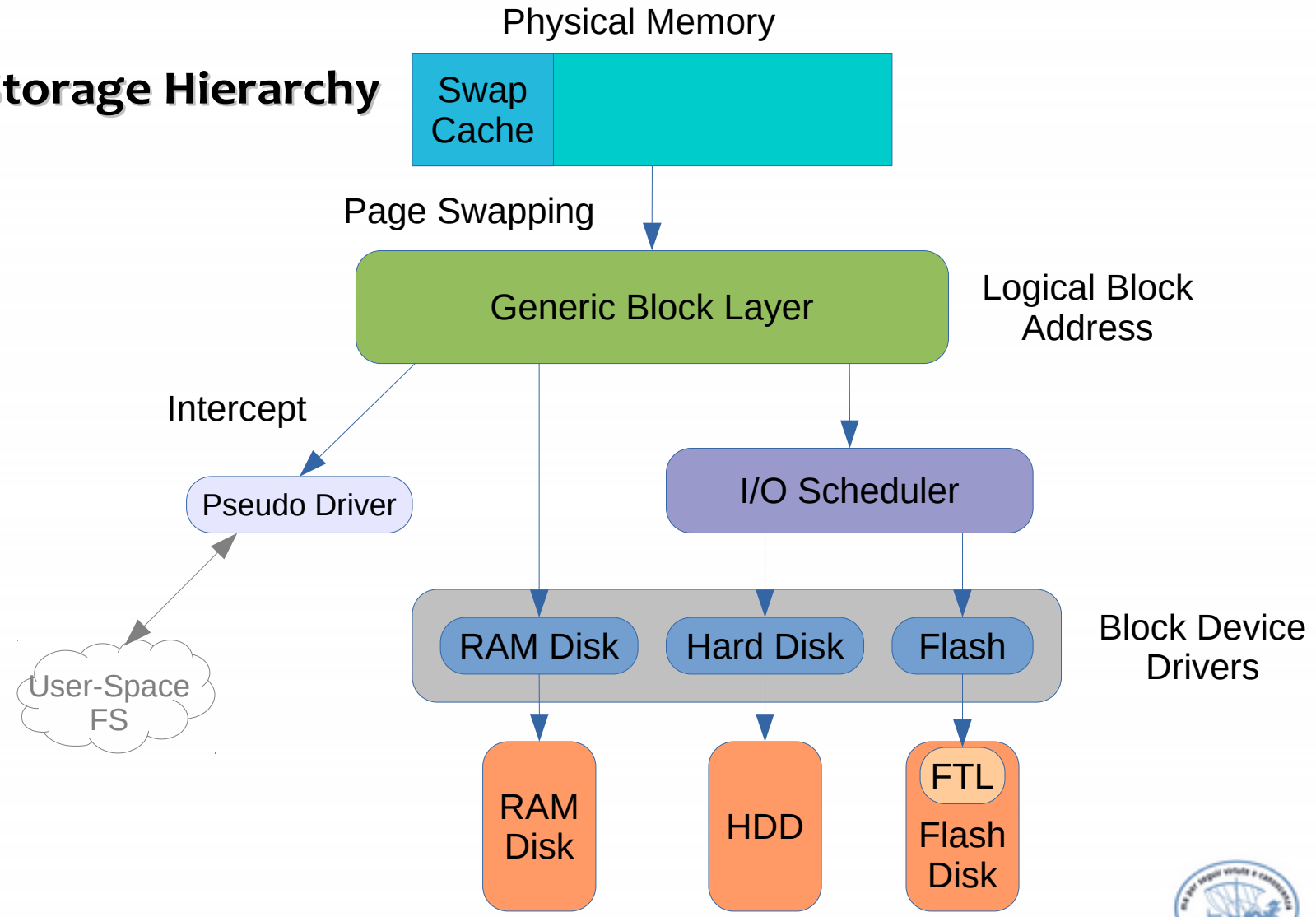
Software to handle software failures (server failover, write-ahead logging)

RAM Disk / Solid State Drive

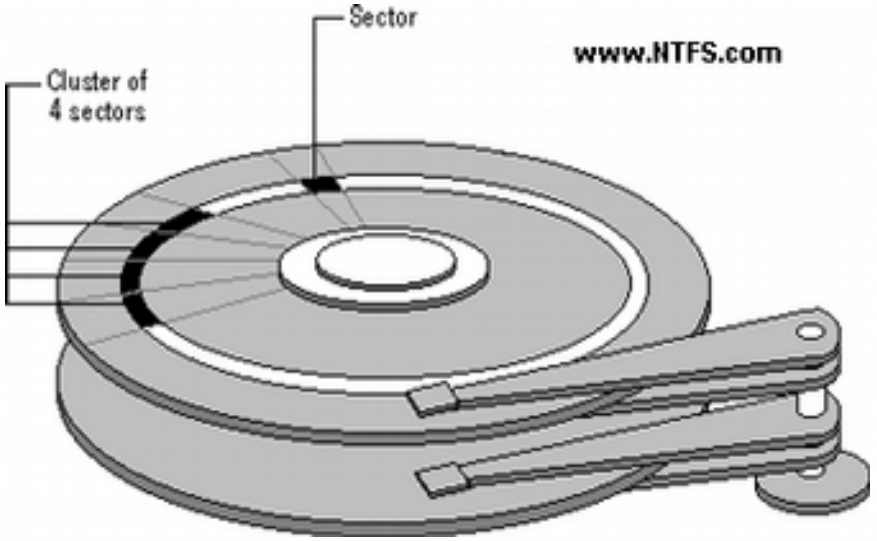
- Unix-like OS environments very frequently create (small) temporary files in /tmp, etc.
  - faster access and less wear with RAM disk
- Linux provides "dynamic RAM disk" (*tmpfs*)
  - only existing files consume RAM
  - automatically cleared on reboot (-> volatile)
- Solid state drive is a **non**-volatile RAM disk
  - uses same interface as (spinning) hard drive
    - Battery buffered DRAM (fast, no wear, expensive)
    - Flash based (varied speed, wears out, varied cost)

MHPC
Master in High Performance Computing

Storage Hierarchy



SISSA ICTP



HDD

- Rotating mechanical device
  - 7200, 1000, 15000 rpm.
- Head on the right track (seek time) 4 ms
- Head on the right sector (latency) 2ms
- Capacity: 4-12 TB
- Bandwidth: Read / Write ~ 150/200 MB/s
- At constant rotating speed, where should I put my data to get max bandwidth ?

Current HDD technology

We can find several magnetic hard disk technologies today:

- Serial Advanced Technology Attachment (SATA)
- Serial Attached SCSI (SAS)
- Advanced Technology Attachment ([P]ATA/[E]IDE) (obsoleted by SATA)
- Small Computer System Interface (SCSI) (obsoleted by SAS)

Solid-State Drive (SSD)

pros:

- lower access time and latency
- no moving parts (silent, less susceptible to physical shock, low power consumption and heat production)
- available over SATA, SAS, PCIe, FC buses

cons:

- expensive, low capacity; usage limited to special purposes only (hardly used for data-servers)
- limited write-cycle durability (depending on technology and ... price)
  - SLC NAND flash ~ 100K erases per cell
  - MLC NAND flash ~ 5K-30K erases per cell
  - TLC NAND flash ~ 300-500 erases per cell

SSD technology

- SLC – Single Level Cell
  - One threshold, one bit
  - 10^5 to 10^6 write cycles per page
- MLC – Multi Level Cell
  - Multiple thresholds, multiple bits (2 bits)
  - 10^4 write cycles per page
  - Denser and cheaper, but slower and less reliable
- TLC – Triple Level Cell
  - Cheapest, slowest writes
  - 500 write cycles per page!

SSD vs HDD

Latency ÷40: 0.1 ms vs 4 ms

Bandwith x3: 450 MB/s vs 150 MB/s

Capacity ÷8: 1 TB X SSD vs 8 TB (2016)

Price / Byte x10: $0.4 / GB vs $0.05 / GB

MHPC
Master in High Performance Computing

NVME

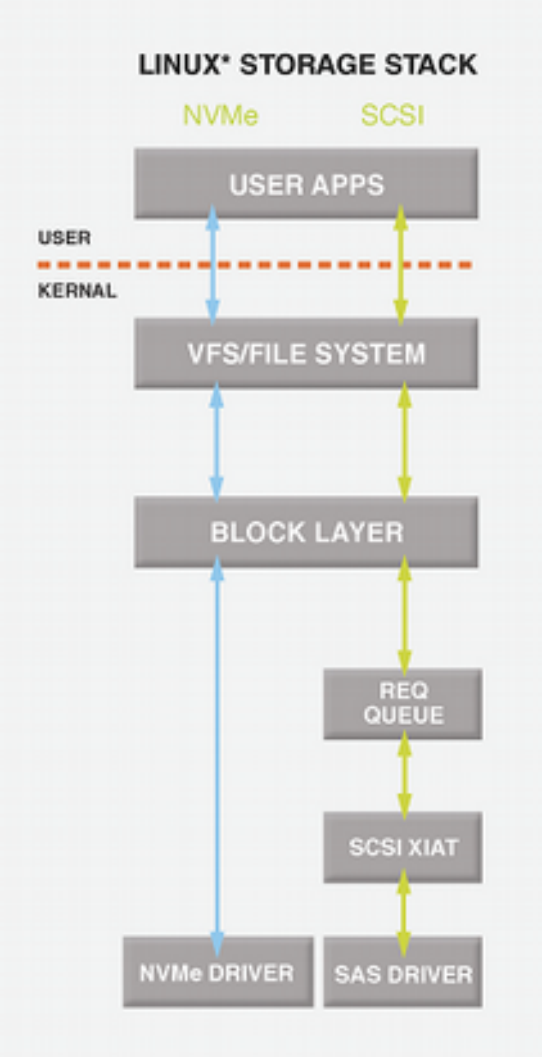
None-volatile Memory Express (NVMe)

SSD can be connected on PCIe bus

SATA /SAS protocols designed for mechanical drive and are now the bottleneck

SAS SSD: 100 µsec - 450 MB/s

NVMe SSD: 20 µsec - 2500 MB/s



SISSA ICTP

Memory/Disk prices

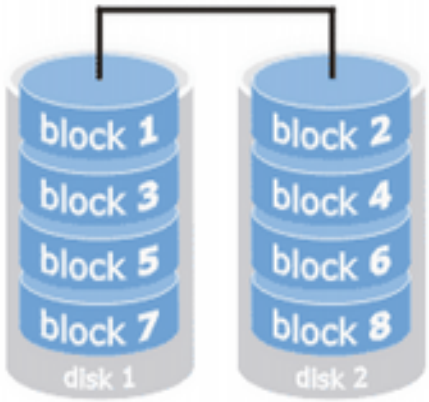
Description	$Price	1yr % chg	$/GB	
Samsung 16 GB 1600 MHz DDR3	76	-4%	$4.75	4.7
Samsung 32 GB 2133 MHz DDR4	306	+61%	$9.56	
Intel 3D Xpoint 32 GB SSD M.2 2280	80		$2.50	
Kingston 32 GB microSD U1/Class 10	13	-48%	$0.41	
SanDisk 32 GB microSD U3/Class 10	29	-41%	$1.10	
Samsung 1 TB SSD MZ-75E1T0	327	+6%	$0.33	0.33
Samsung 1 TB SSD MZ-7KE1T0BW	421	-2%	$0.42	
Seagate 2TB HDD ST2000DM006	67	-1%	$0.033	
Seagate 4TB HDD ST4000DM000	98	-16%	$0.025	0.025
Seagate 8TB 5900 RPM HDD ST8000AS0002	227	0%	$0.028	
HGST 8TB 7200 RPM HDD HUH728080ALE604	275	-42%	$0.034	
Seagate 10TB HDD ST100000VN004	340		$0.034	
HGST 10TB HDD HUH721010ALE604	350		$0.035	

Data from web search, August 22 2017

The Disk Bandwidth/Reliability Problem

Disks are slow: use lots of them in a parallel file system

However, disks are unreliable, and lot's of disks are even more unreliable



This simple two-disk system is twice as fast, but half as reliable, as a single-disk system

RAID

- RAID is a way to aggregate multiple physical devices into a larger virtual device
  - Redundant Array of Inexpensive Disks
  - Redundant Array of Independent Devices
- Invented by Patterson, Gibson, Katz, et al
  - http://www.cs.cmu.edu/~garth/RAIDpaper/Patterson88.pdf
- Redundant data is computed and stored so the system can recover from disk failures
  - RAID was invented for bandwidth
  - RAID was successful because of its reliability

RAID reliabilty and performance..

Reliability or performance (or both) can be increased using different RAID “levels”.

Let us examine some of the most important:

Definitions:

S: Hard disk drive size.

N: Number of hard disk drives in the array.

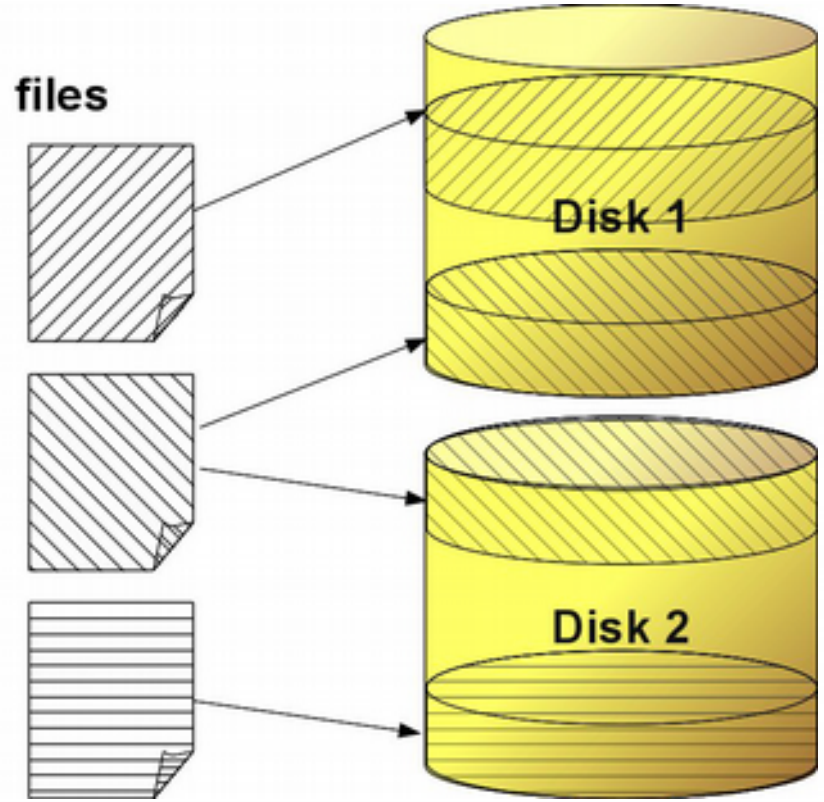
P: Average performance of a single hard disk drive (MB/sec).

LINEAR RAID

Performance = P

NO REDUNDANCY

Capacity = N * S

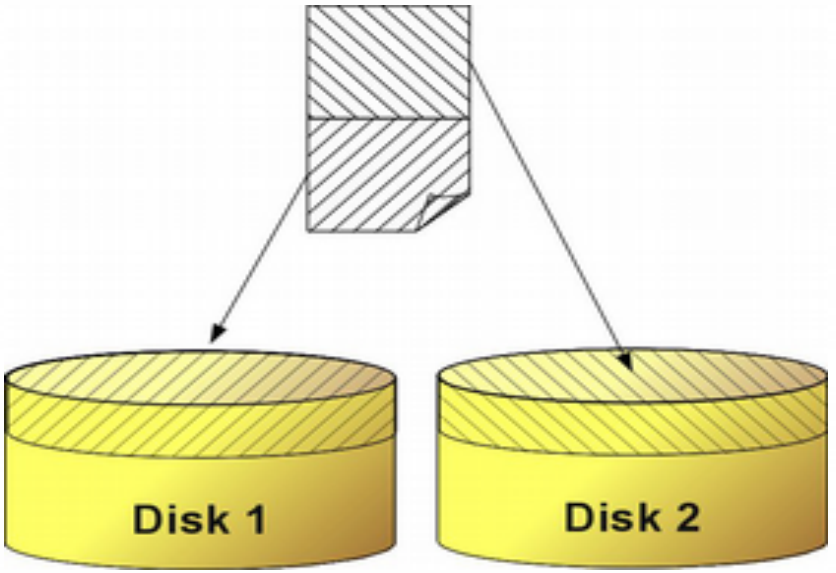


RAID 0

Performance = P * N

STRIPING

Capacity = N * S



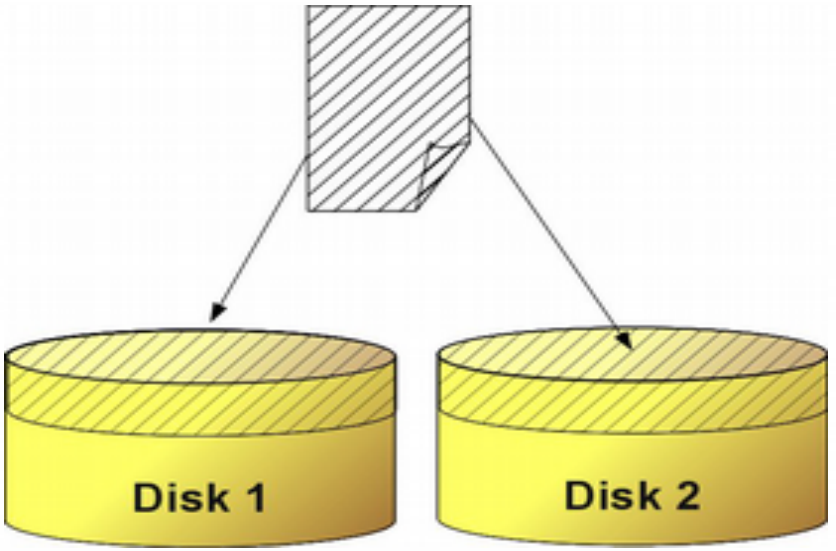
RAID 1

Write Perf. = P

Read Perf. = P * N

REDUNDANCY

Capacity = S

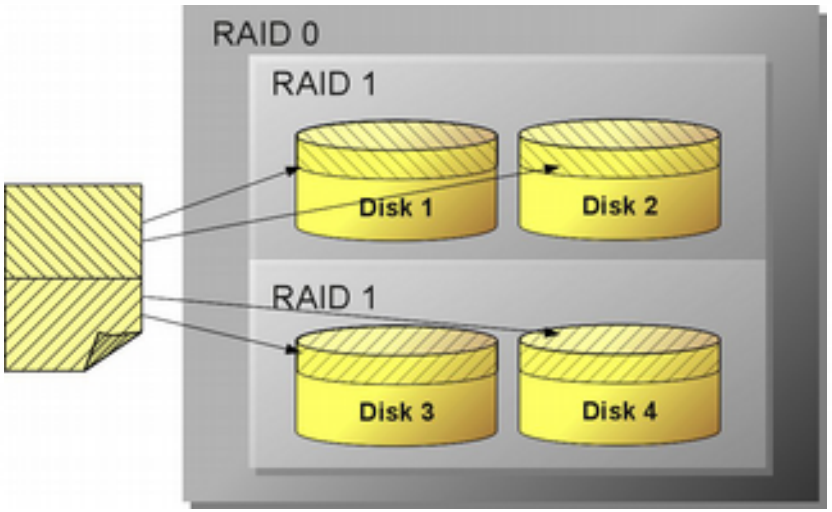


Nested RAID levels

RAID 10 / RAID 1+0 and RAID 0+1

REDUNDANCY

STRIPING



Raid 1+0 / 10: mirrored sets in a striped set

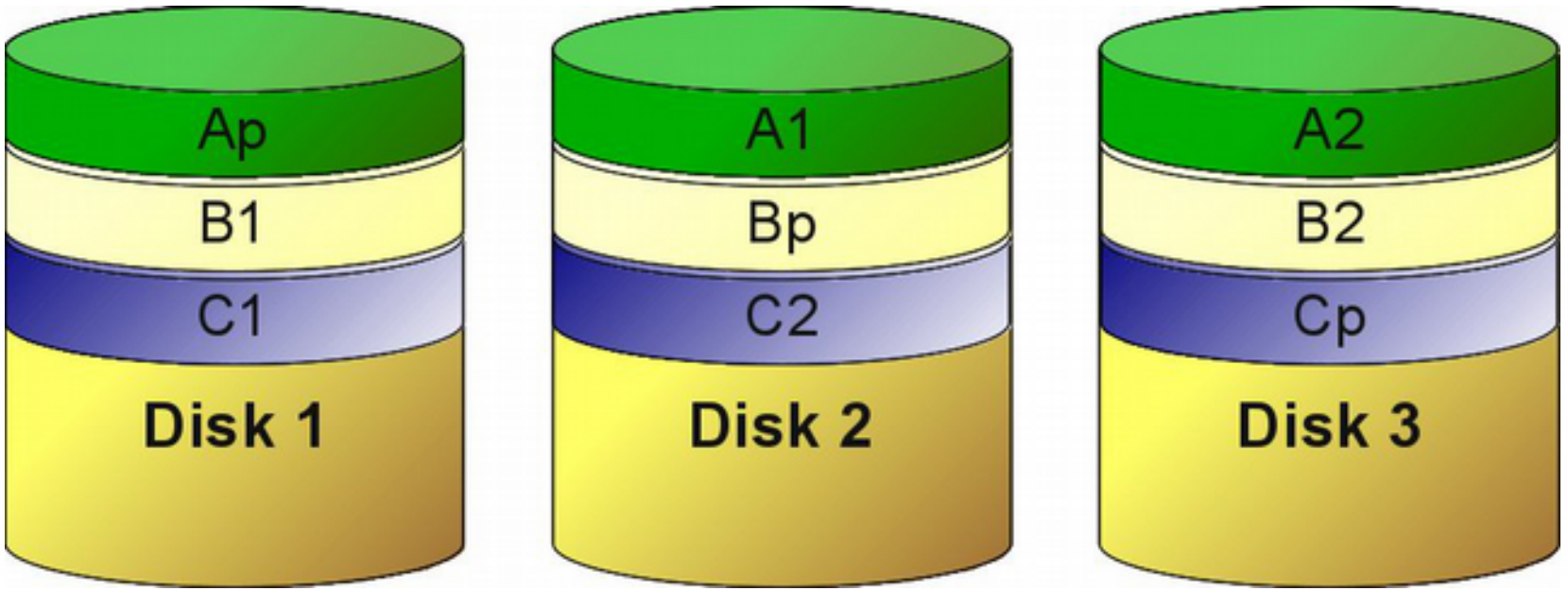
the array can sustain multiple drive losses so long as no mirror loses all its drives

Raid 0+1: striped sets in a mirrored set

if drives fail on both sides of the mirror the data on the RAID system is lost

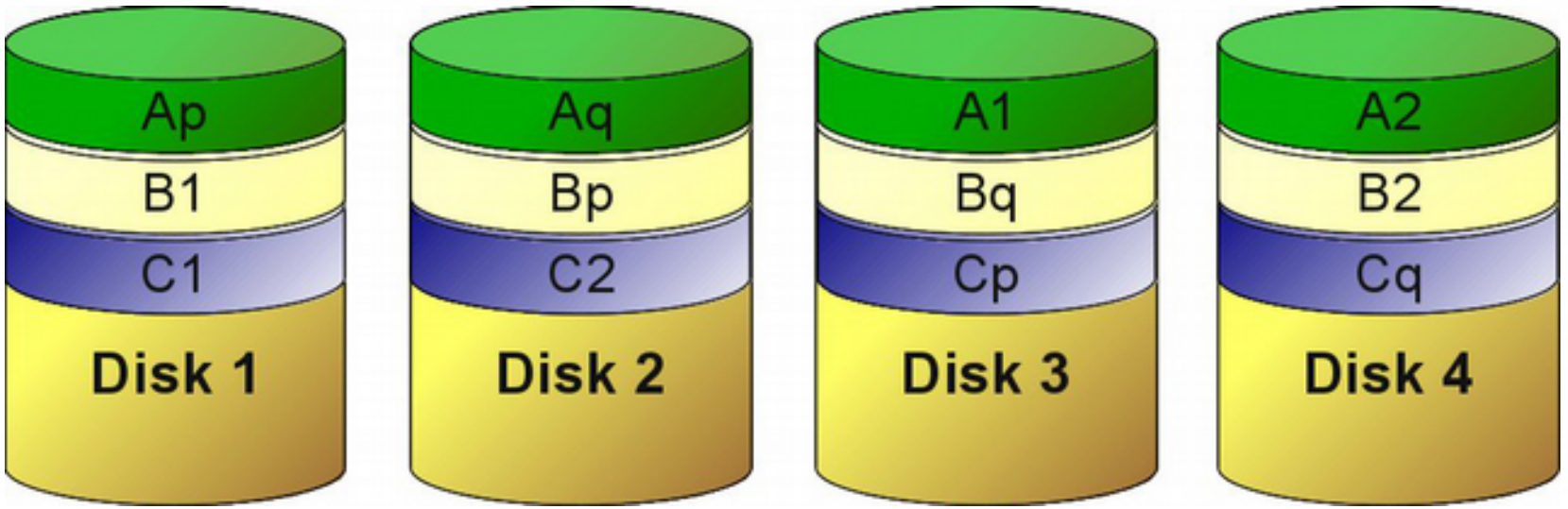
RAID 5

One disk can fail
Distributed parity



RAID 6

Two disks can fail
Double distributed parity code



Notes on redundancy

Computing and updating parity negatively impact the performance. Upon drive failure, though, lost data can be reconstructed, and any subsequent read can be calculated from the distributed parity such that the drive failure is masked to the end user.

However, a single drive failure results in reduced performance of the entire array until the failed drive has been replaced and the associated data rebuilt.

The larger the drive, the longer the rebuild takes (up to several hours on busy systems or large disks/arrays).

Hot spare

Both hardware and software RAIDs with redundancy may support the use of a hot spare drive, a drive physically installed in the array which is inactive until an active drive fails, when the system automatically replaces the failed drive with the spare, rebuilding the array with the spare drive included. A hot spare can be shared by multiple RAID sets.

Subsequent additional failure(s) in the same RAID redundancy group before the array is fully rebuilt can cause data loss.

RAID 6 without a spare uses the same number of drives as RAID 5 with a hot spare and protects data against failure of up to two drives, but requires a more advanced RAID controller and may not perform as well.

Level	Description	Minimum # of drives	Space Efficiency	Fault Tolerance	Read Benefit	Write Benefit
RAID 0	Block-level striping without parity or mirroring.	2	1	0 (none)	nX	nX
RAID 1	Mirroring without parity or striping.	2	1/n	n-1 drives	nX	1X
RAID 5	Block-level striping with distributed parity.	3	1-1/n	1 drive	(n-1)X	(n-1)X
RAID 6	Block-level striping with double distributed parity.	4	1-2/n	2 drives	(n-2)X	(n-2)X
RAID 1+0/10	Striped set of mirrored sets.	4	*	needs 1 drive on each mirror set	*	*
RAID 0+1	Mirrored set of striped sets.	4	*	needs 1 working striped set	*	*

http://en.wikipedia.org/wiki/RAID

Measure performance

dd: to measure storage system performance

```
# dd if=/dev/zero of=/tmp/toto bs=1k count=1k oflag=sync
1024+0 records in
1024+0 records out
1048576 bytes (1.0 MB, 1.0 MiB) copied, 1.17594 s, 892 kB/s
# dd if=/dev/zero of=/tmp/toto bs=1k count=1k
1024+0 records in
1024+0 records out
1048576 bytes (1.0 MB, 1.0 MiB) copied, 0.00291734 s, 359 MB/s
```

Questions:

1- Why such a difference between the two runs?

2 - How to measure Read bandwidth?