Storage Technologies

Bernd Panzer-Steindel

CERN IT

CERN School of Computing 2009

How to build a storage system

Basic storage components

Interconnects

File systems

Mass storage

Cloud storage

Complexity, energy and costs as boundary conditions

What we want :

The whole is bigger than the sum of the individual parts

What we usually get:

The whole is much smaller than the sum of the individual parts

**Need to understand the hardware and software aspects,
but the real tool to solve this problem is BRAINWARE**

Storage system properties, order of importance

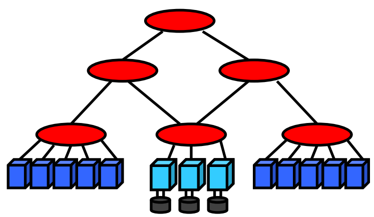
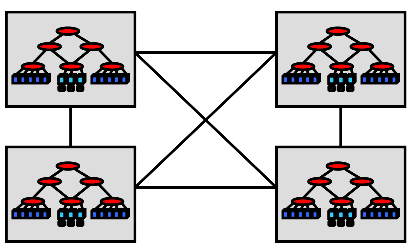
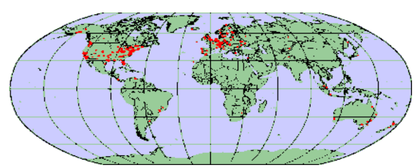
1. **Reliability**
2. **Basic functionality**
3. **Performance**
4. **Fancy functionality**

All 4 items should be considered from the beginning in the design and cross-checked regularly during prototyping

CERN School of Computing

Physical and logical connectivity

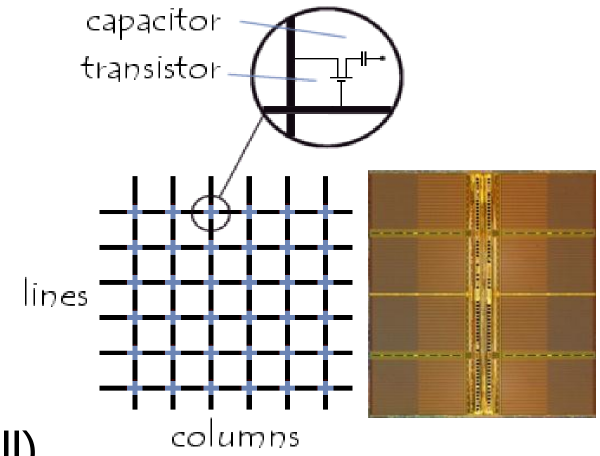
Complexity

		Hardware	Software
Components		CPU, disk, memory, motherboard	
PC, disk server			Operating system Device drivers
Cluster, Local fabric	 	Network, Interconnects	Resource Management software
World Wide Cluster	 	Wide area network	Grid and Cloud Management software

Chapter 1

Basic storage devices

Components: Memory I

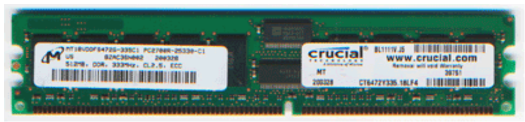


SDRAM : Synchronous Dynamic Random Access Memory

Characteristics :

- Volatile storage, one cell = one transistor plus one capacitor, needs constant refresh cycles (every ~64 ms per cell)

- Registered or buffered DIMMs (Dual Inline Memory Modules),
  Market : 1, 2, 4, 8 Gbyte DIMMS, → 16 Gbytes in 2010



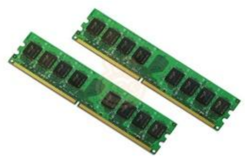
- Dominating market share:
  DDR2, DDR3 (Double Data Rate SDRAM, generation 3)

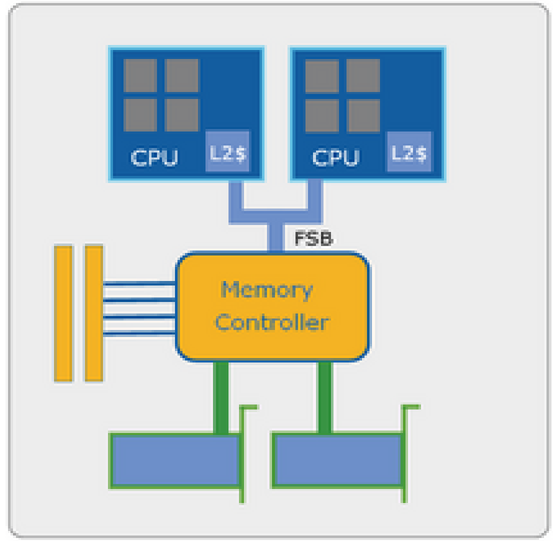
- ECC integrated, (Error Correction Code), 8 data bytes + 1 parity byte,
  can correct single bit errors 10^14 Bit Error Rate == 1 week at 100 MB/s

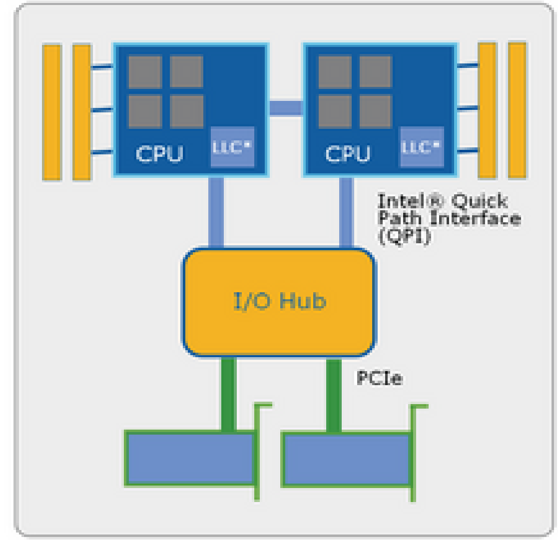
- Current production lines use Structure sizes of 50-60 nm,
  → 30-40 nm in Q4 2009

Components: Memory II

Performance characteristics







- DDR2 200 – 400 MHz I/O bus frequency DDR3 400 – 800 MHz
  e.g. DDR3-1600
  → 800 MHz memory controller * 2 (double) * 64 bit (width) = 12.8 Gbytes/s

- Memory banks can be combined on the motherboard via dual/quad channels
  → another speed increase by a factor 2 or 4

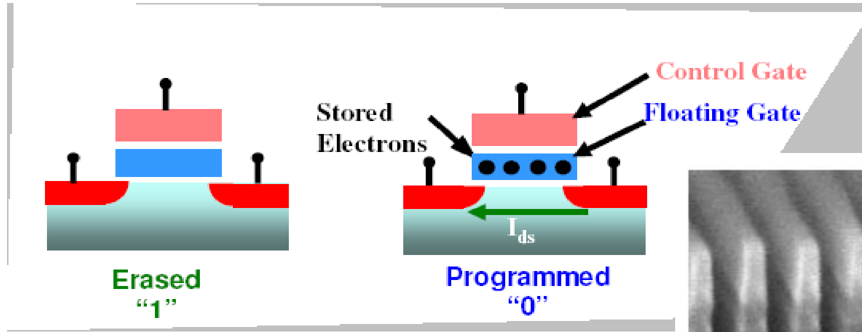
- The memory access latency is about 10 ns

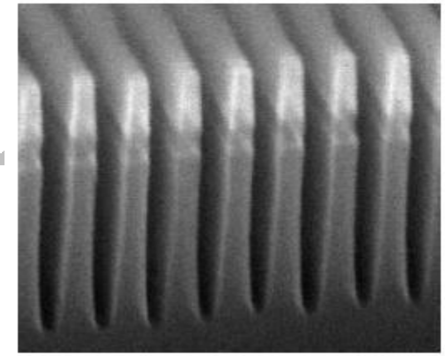
- DDR uses 1.8 V and DDR3 1.5 V
  electrical power usage scales with frequency and voltage
  $P \sim V^2$ $P \sim f$

CERN School of Computing

Components: Memory III

Flash memory





Sub-40mm NAND Flash gate

- Non-volatile
- High density, only single floating-gate-transistor per cell
- currently 40-50 nm structures
- cost effective, uses same techniques as the processor industry
- Two types of flash implementations
  NOR → addressable per cell, slow read and write speed
  NAND → block addressable (no random access, all access is sequential), fast read and write, dominating the market (11 B$ revenues)

- SLC (Single Level Cell) versus MLC (Multi Level Cell)
  SLC → faster performance, more expensive, higher endurance (100000 erase cycles versus 10000), less redundancy (ECC) needed

Capacity is doubling every 12 month since 1994.

Components: Memory IV

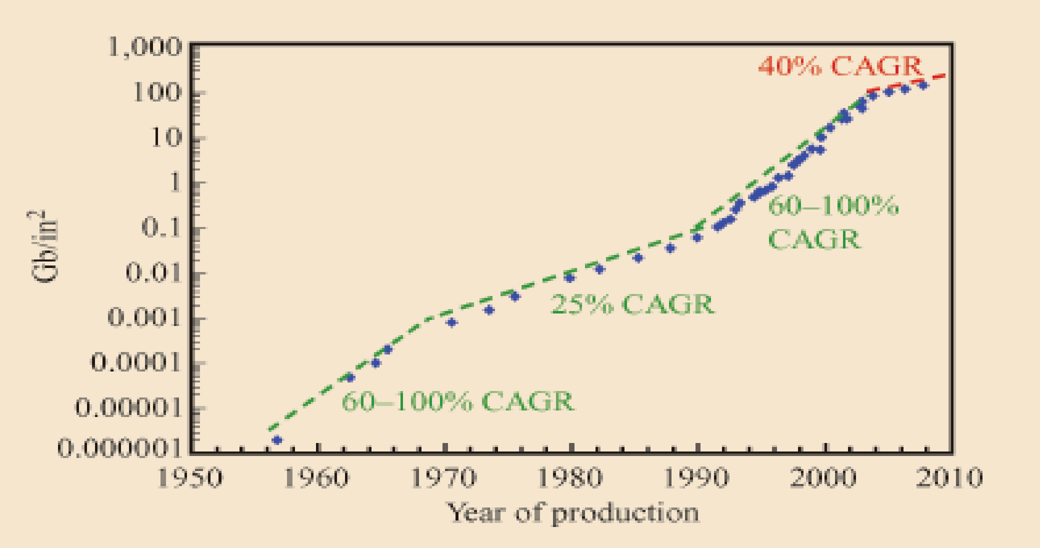
New memory technologies

Effort to produce non-volatile, fast switching, high endurance, low cost memory

FeRAM Ferroelectric RAM
MRAM Magnetoresistive RAM
PRAM Phase-Change RAM
……………………

- Major investments and activities since ~1990
- Difficulties with storage density, long-term stability, reliable production
- MRAM prediction from 2005 → 2 B$ market share in 2008 Reality = 25 M$
- Overall memory market value is about 60 B$ / year

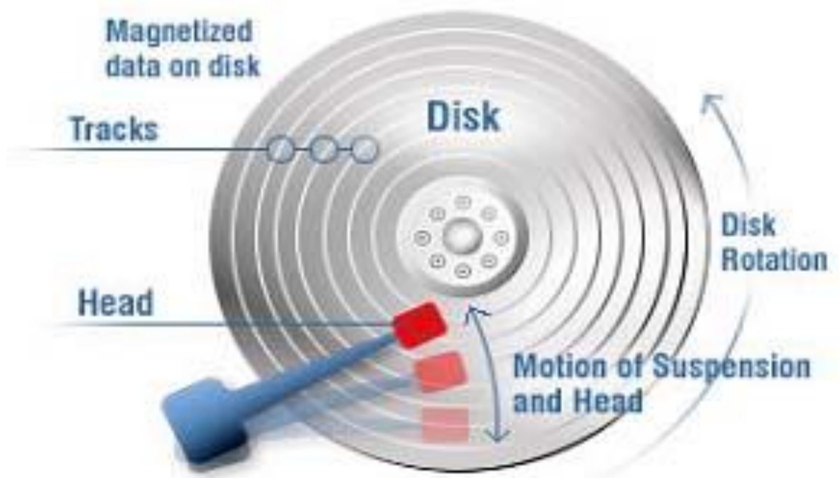
Components: Hard Disk I



Defining properties

1. Magnetic recording density,
   areal density = recording density plus track density
   BPSI : bits per square inch, Gbits/in^2

   the controller uses Reed-Solomon ECC encoding before writing to disk, 20% of the data on disk are for error correction
   (a CD has 66% redundancy data)

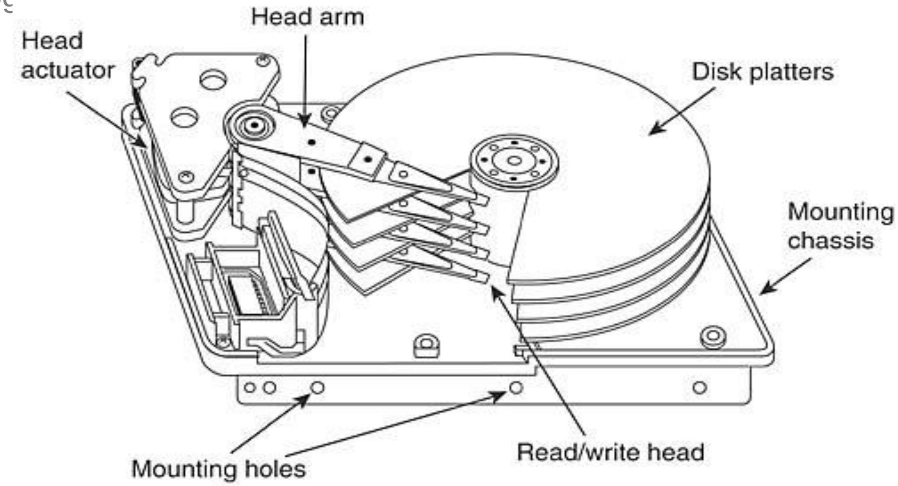


2. Form factor 3.5“ 2.5“ 1.8“ 1“

3. Number of platters, single sided double sided

4. Internal cache size (SDRAM), 8, 16, 32 MBytes

Components: Hard Disk II



Defining properties

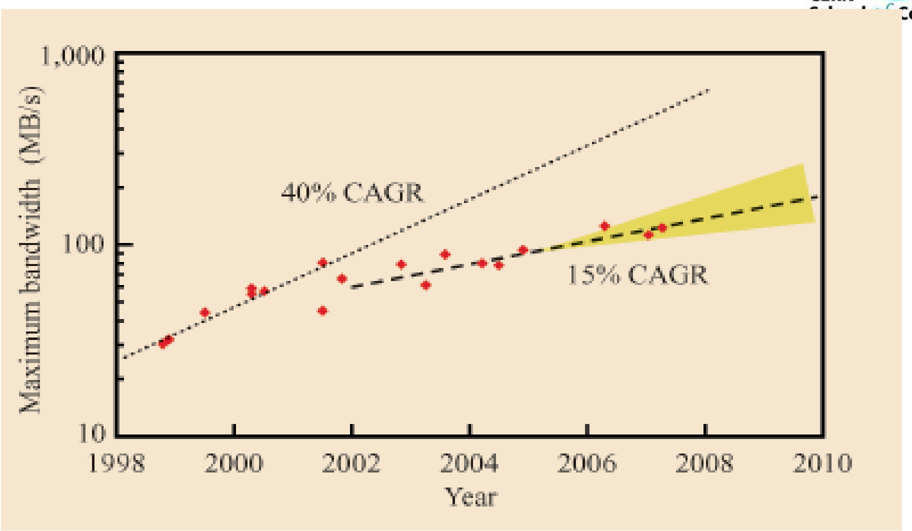
4. Spindle rotional speed
   → Notebooks 5400 rpm , server 7200 rpm , 10000 rpm, high end 15000 rpm
   → green drives change their rotational speed on the fly
   → some work on 20000 ongoing → power problems, mechanical stability
   spindle motor primary source of power consumption

6. Quality
   → MTBF = Mean Time Between Failure
   low end drives : 300000 h MTBF, high end drives : 2000000 h MTBF
   definition of duty cycle ! 24h * 7 d or 8h per day
   → error rates, one un-recoverable bit error per n Bits read/written
   e.g. One bit in 10^14 bits == < 3 days at 50 MB/s

7. Electronic interface (SATA, SAS, SCSI, FC, etc.) (see next pages)

Components: Hard Disk III

Derived properties



1. **Sequential I/O performance**
   → areal density + rotational speed
   + interface up to 150 MB/s

2. **Capacity**
   → areal density + form factor + #platters

3. **Access Time = Command Overhead Time + Seek Time + Settle Time + Latency**
   → rotational speed + form factor + actuator quality
   as low as 3 ms

4. **Power consumption**
   → rotational speed + form factor
   in the range of 5-12 W

Large growth rate (40% per year) for the disk capacities, while the seq. performance has a very low improvement rate and the access time is essentially constant

CERN School of Computing

Components: Hard Disk IV

More parameters :

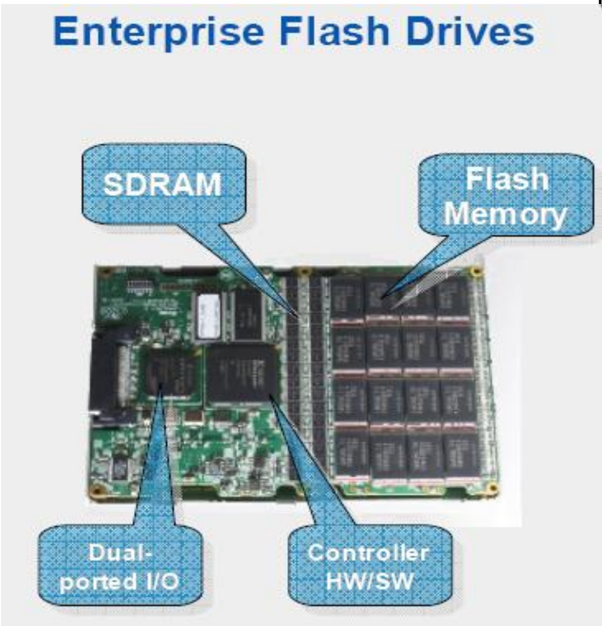
- Size of the read-ahead buffers in the controller cache (64 Kbytes)

- Write caching

- Support for NCQ or TCQ
  Native command queuing (SATA), tagged commend queuing (SCSI)
  Must be supported by the RAID controller and the disk
  Allows the disk to order the requests, optimize performance,~ 32 commands deep
  Kernel IO driver available , multiple commands to be issued at a time

CERN School of Computing

Components: SSDisk



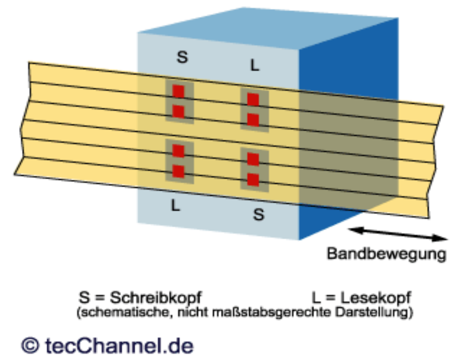
Solid State Disk

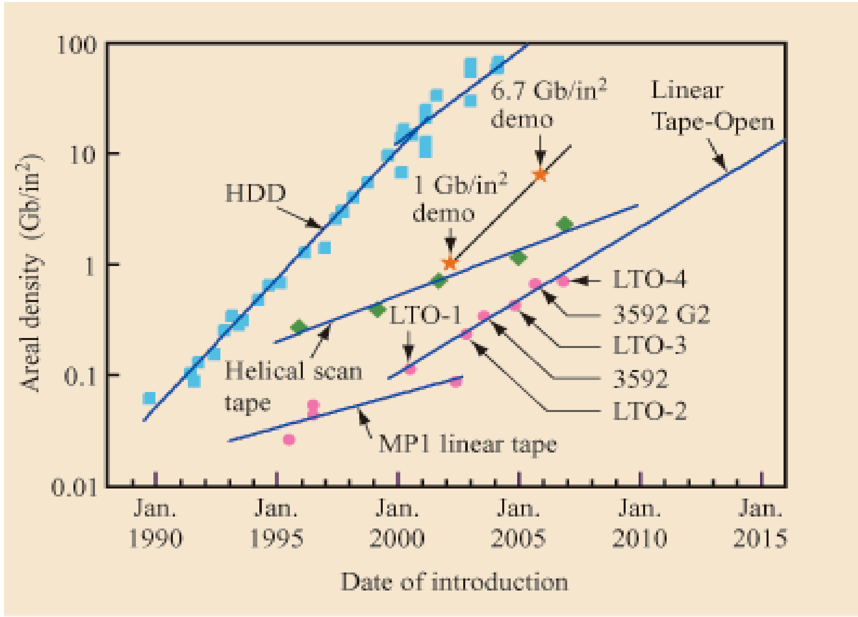
- Replacing the magnet recording platters with flash memory
- More complicated controller needed (cost, performance)
  → endurance, wear-leveling access algorithms
- Low power consumption < 2W
- High performance possible,
  → > 300 MB/s sequential, 10000 IOPs (Input Output Operations per Second)
- ~100 suppliers in the is market (HDD market: 5)
  Lot's of consolidation and competition, large variations in price/performance
- In 2009 density equality was reached between HDD and SSD for 2.5" (1 TB disks)
- Need new file system design for SSD → block size, wear-leveling, etc.
- Problems with benchmarks, SSD controller differences to HDD controller

CERN School of Computing

Components: Tape I

- LTO (Linear Tape-Open) format dominates the market, ~90 % share
- HP, IBM , Quantum consortium
  >20 Exabyte of tape space sold since 2000
- Linear track technology, 70 tracks/mm
- Density is less than hard disks, but with a similar growth rate
- LTO-4 : 800 GB cassettes, 120 MB/s maximum transfer speed
  LTO-5 : early 2010, 16 TB cassettes, 180 MB/s
- Quite active developments, long term roadmap, technology improvements every 2 years





CERN School of Computing

Components: Tape II

- Industry quoting include a compression ratio of 2:1 always maximum speed of the device

- Cost per GB storage , include drives and silos
  e.g. 20 KCHF for a drive plus server
  400 KCHF for a 10000 slot silo

- Drive MTBF is a bout 250000 hours
  Read bit error rate is about 10^{-17}

- Random access times: 2-3 minutes

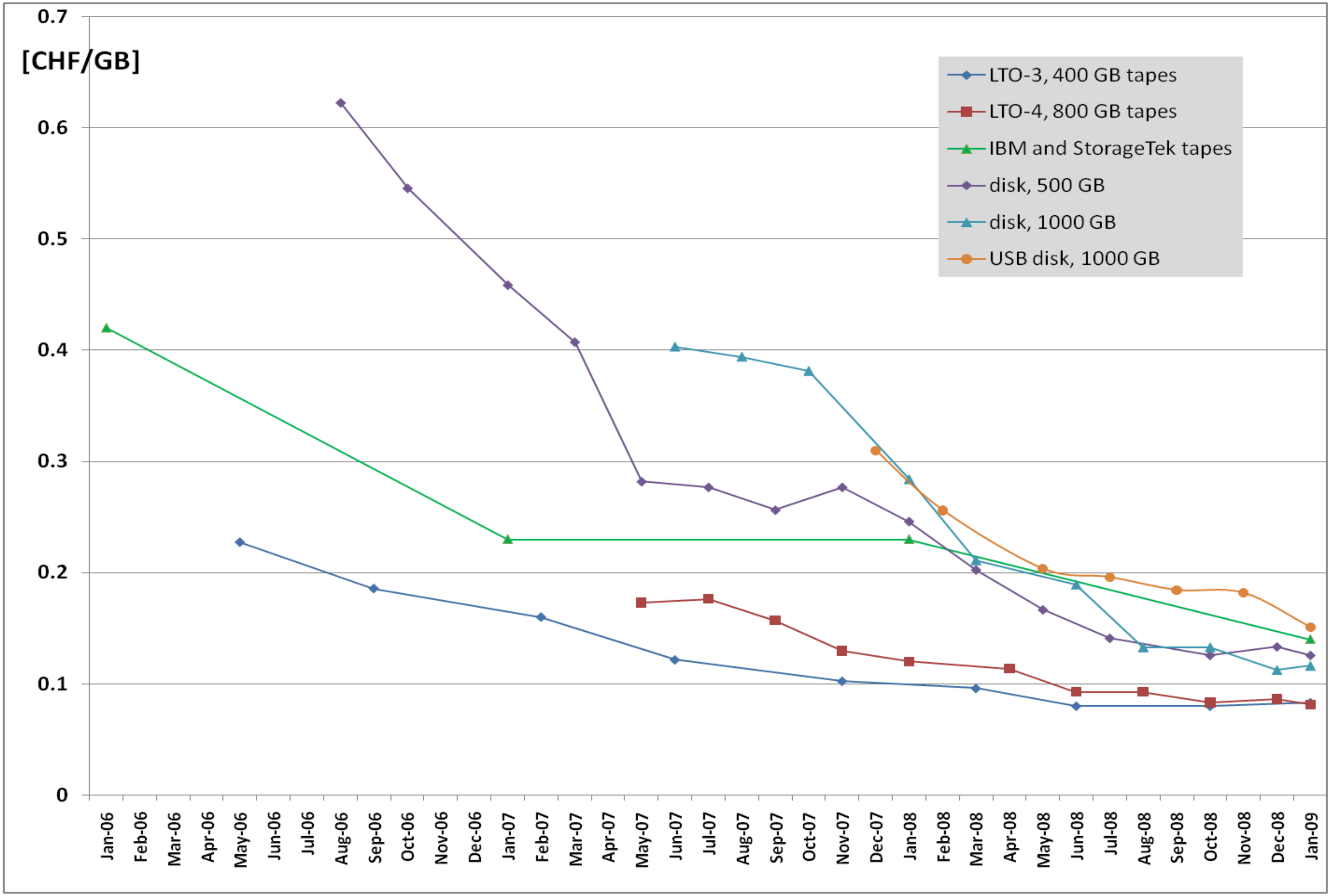
- Data streaming necessary to achieve reasonable performance !

- LTO covers 90 % of the automated library market
  Tape total revenues per year :4 B$
  compared to 26 B$ for disk storage systems (integrated, not bare disks)

- Long principle lifetime of the media, but technology change every 2 years
  → Major implications for the operation and maintenance

CERN School of Computing

Components: Cost comparisons and evolution I



CERN School of Computing

Components: Cost comparisons and evolution I

1 PB of space

		cost	power	infra*	seq r/w	IOPs
Memory	→	25 MCHF,	1.3 MW	(x3)	3.0 PB/s	10^{13}
SSD, high end	→	18 MCHF,	0.03 MW	(x4)	3.0 TB/s	10^{8}
SSD, low end	→	5 MCHF,	0.06 MW	(x4)	1.0 TB/s	10^{7}
Hard Disk, high end	→	1.05 MCHF	0.03 MW	(x3)	0.4 TB/s	10^{6}
Hard Disk, low end	→	0.15 MCHF	0.01 MW	(x2)	0.1 TB/s	10^{5}
Tape	→	0.08 MCHF	0.01 MW	(x1.2)	0.5 GB/s	1

Error bar certainly 20 % and things are changing fast

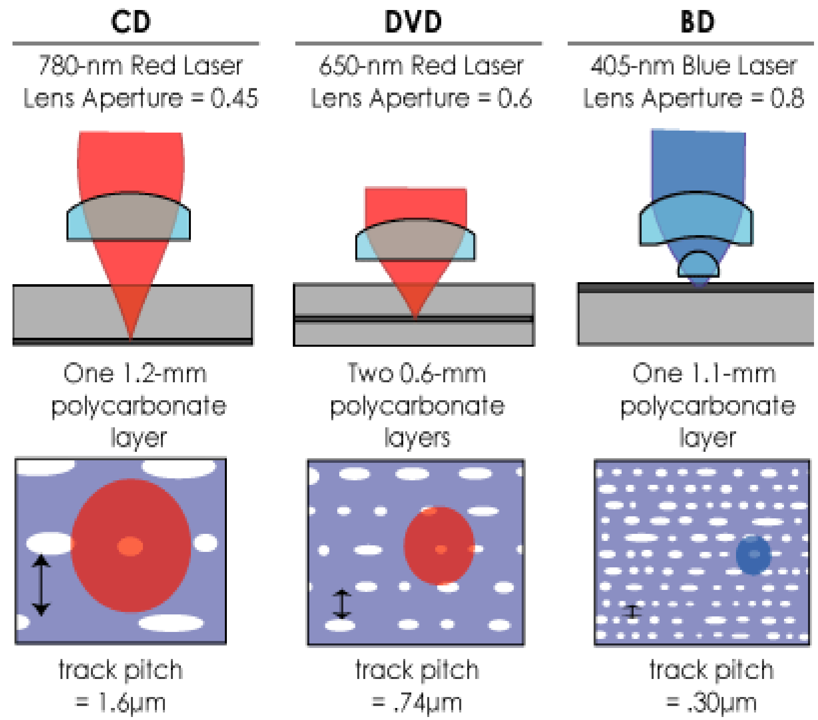
(4 GB DIMM, 64 GB SSDisk, 0.3/1 TB disk, 1 TB tape)

* Infrastructure overhead, multiplication factor for costs and power

Components: Optical devices

Laser frequency, disk size, max. performance, costs

- **CD**
  
  780 nm, 0.7 GB, 10 Mbytes/s, 0.7 €/GB
- **DVD**
  
  650 nm, 4.7 GB, dual layer, 30 MB/s, 0.24 € /GB
- **Blue-Ray**
  
  405 nm, 25 GB, dual layer, 36 MB/s, 0.11 € /GB
- Optical juke-boxes
  up to 2000 slots, multiple drives



Long media lifetime, but regular format changes → availability of drives

CERN School of Computing

Components: Holographic Storage

Developments since the 1960s
First tests in the 60s (TCRA Laboratories)
First prototype by Bell Labs in 1998

- InPhase promise in 2005 : holographic disk with 300 GB, 60 x DVD
  First product shipment in 2008, 180 $ per 300 GB disk,
  20 Mbytes/s read speed, 18000$ for the drive

- In 2009, General Electric introduced a 500 GB prototype holographic disk

In the labs

- Quantum holographic storage, 2 * 35 bit, around a single electron

- Five dimensional holographic storage
  (3, plus color plus phase) 1.1 terabytes / cm^3

- MEMS (Micro-Electro-Mechanical Systems) based memory
  → millipede (prototype 2005, no production yet)

Large discrepancy between expectations and market availability !!

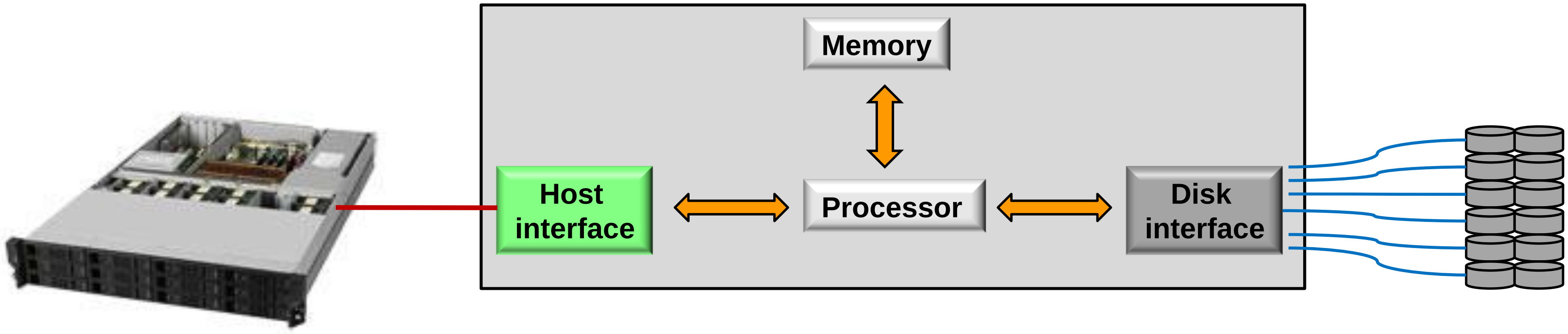
Chapter 2

Hardware interconnects

Components: RAID controller I

Managing one or multiple disks and interfacing to the host processor

e.g. 3Ware, Adaptec, Areca,…



DAS Direct Attached Storage
- → controller directly attached to the motherboard via PCI-E
- → from one to 48 disks in an enclosure, market ‘sweet-spot’ is 24 bay

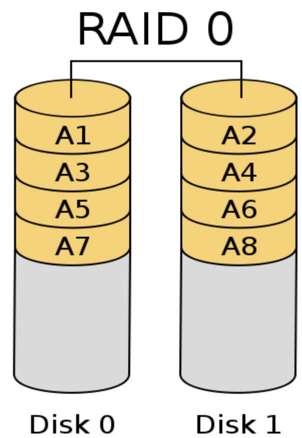
NAS Network Attached Storage
- → HBA (Host Based Adaptor) on the motherboard connects to the host interface on the controller (longer physical distances possible)
- → controller and disks in external enclosure, 8 – 48

SAN Storage Area Network
- → specific network attachment via Fiber Channel (FC)

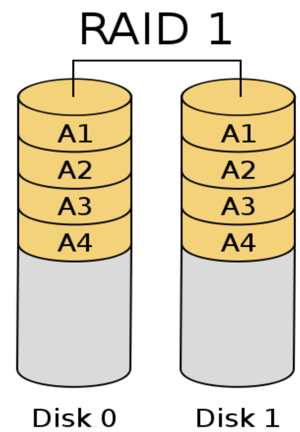
CERN School of Computing

Components: RAID controller II Redundancy

- Combining multiple disks
- Performance reasons, but mainly reliability
  → MTBF and intrinsic bit error rate
  → performance penalties during disk recovery (RAID rebuild)

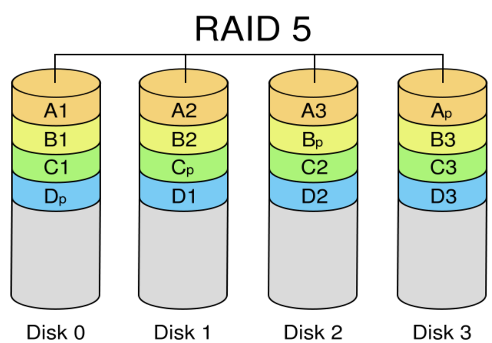


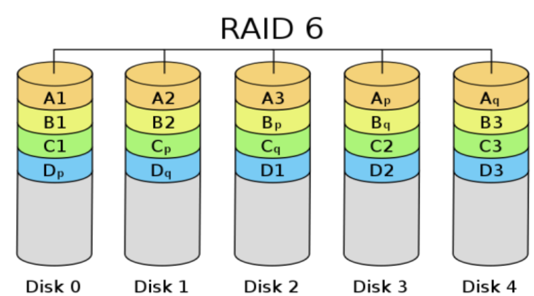
Striped



Mirror

- Striping can be combined with redundancy
  raid0 + raid5 = raid50
- Multi TB disks and many devices require RAID6

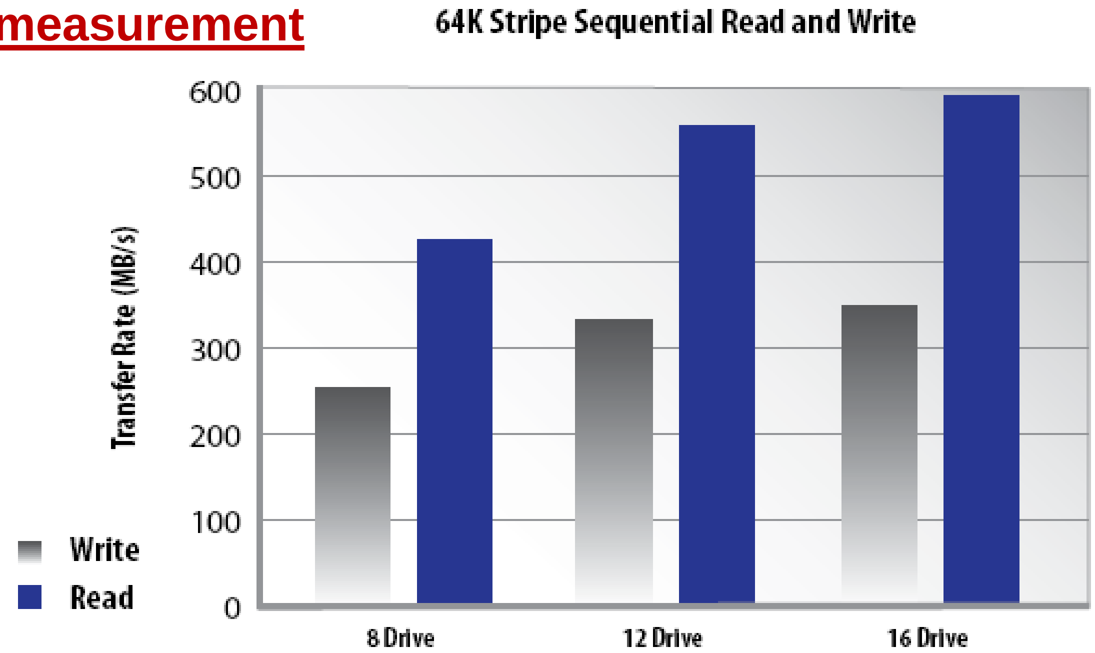




CERN School of Computing

Components: RAID controller III Performance

Example measurement



- ➢ Writing speed penalty due to parity calculations, worse for RAID6
- ➢ Need powerful controller
- ➢ PCI-E interface to the motherboard → 1 Gbyte/s max
- ➢ 16 disks can do 17 Gbytes/s r/w
- ➢ Performance increase NOT proportional to the number of disks

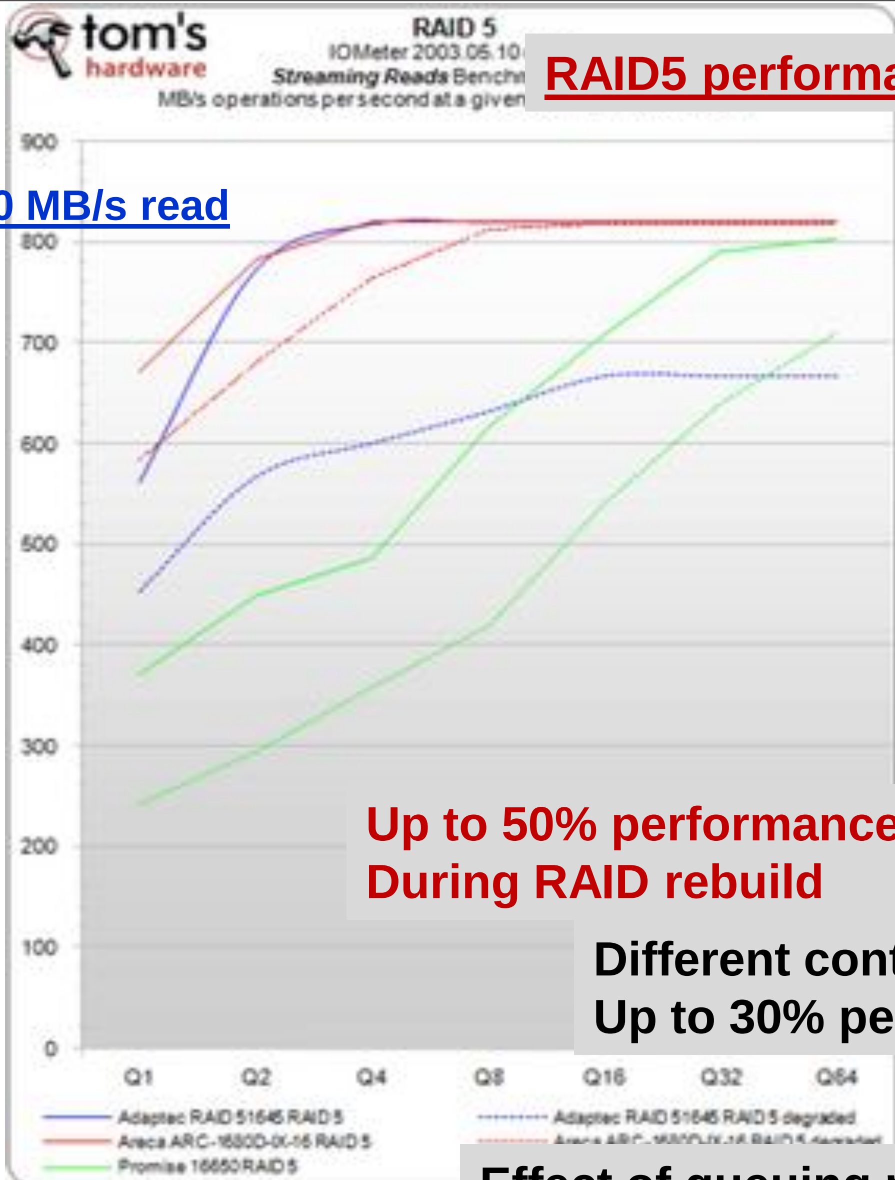
- **RAID5 configuration with 8/12/16 disks**
- **3Ware controller, 16-way**
- **10k RPM Western Digital disks**
- **XFS file system**

CERN School of Computing

Components: RAID controller IV Performance

RAID5 performance example

800 MB/s read



700 MB/s write

Up to 50% performance degradation
During RAID rebuild

Different controllers
Up to 30% performance differences

Effect of queuing parameters (NCQ)

CERN School of Computing

Device interfaces

SCSI Small Computer System Interface
→parallel 16bit, 320 Mbytes/s, 16 devices, 12m cable length

SAS Serial Attached SCSI
→ serial, 750 Mbytes/s, 4 devices (16k with expanders), 8m cable length

FC Fiber Channel
→serial, 1200 Mbytes/s, point-to-point (switches, hubs), 50 km optical fibre, channel and network

SATA Serial ATA
→serial, 300 Mbytes/s, point-to-point, 1m cable length

SCSI essentially stopped the development of new faster versions, moves to SAS, new standards this year SATA 600, FC20

Often these protocols are also used to describe the quality of the disk, e.g. SCSI disks have a better MTBF than SATA disks
That is wrong !

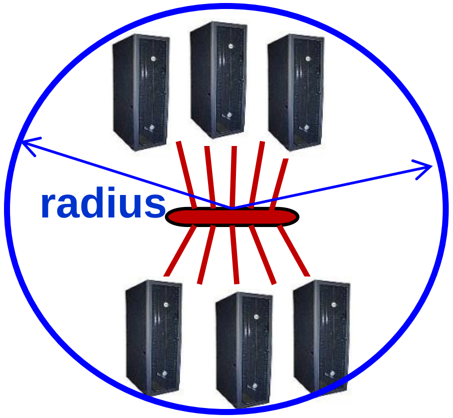
Network interfaces

- **Ethernet** packet based, tree network, 10 Gbit NICs (Network Interface Cards) max,
  40-100 Gbit standard in 2010,
  copper and fiber e.g. 10-15 m for 10 Gbit copper
  possibility of packet drops and retries, high and unpredictable latency (ms),
  wide-spread, large market, many management tools, very cost effective

- **Infiniband** point-to-point serial link, switched fabric, low latency (3-5 µs),
  uses RDMA (Remote Direct Memory Addressing), copper and fiber,
  15m for 4x SDR = 10 Gbit copper cable, can scale to 120 Gbit/s

- **Fiber Cannel** point-to-point, arbitrated loop, switched fabric, 2/4/8/10 Gbits/s,
  SAN Storage Area Network, fiber cables dominant (km range),
  copper possible < 3m, well established technology for storage,
  SCSI commands via FC, failsafe connection of SCSI
  devices via FC

Cable length limitations
→Power density of servers
→Cooling limits KW/m²



CERN School of Computing

Network transport

Combining the physical network layer with the device interfaces

Transfer of the interface protocol on top of the network layer

- iSCSI   SCSI over Ethernet
- FCoE   Fiber-Channel over Ethernet
- TCP over Fibre-Channel
- TCP over Infiniband
- …. More permutations are on the market

Combination of controller hardware implementation and software
Kernel drivers

Keywords:
Cost factor, interoperation, taking into account existing infrastructure, performance versus overheads

Chapter 3

Low level software interconnects

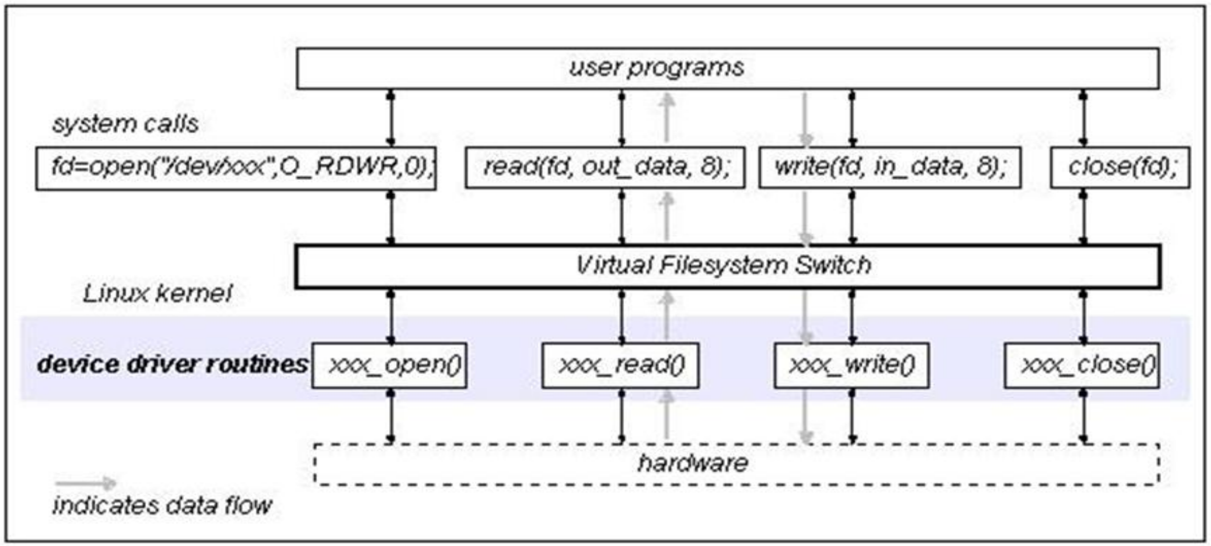
Device Diver I

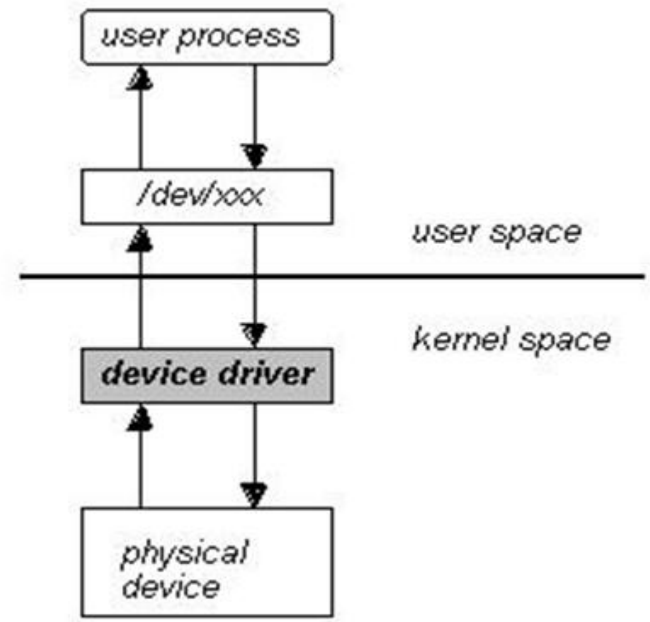
Need to make the physical devices visible to the OS

Disk device driver → SCSI, SATA, FC

RAID controller combine several disk drives into one single device

Mapping the user I/O commands onto the corresponding device commands



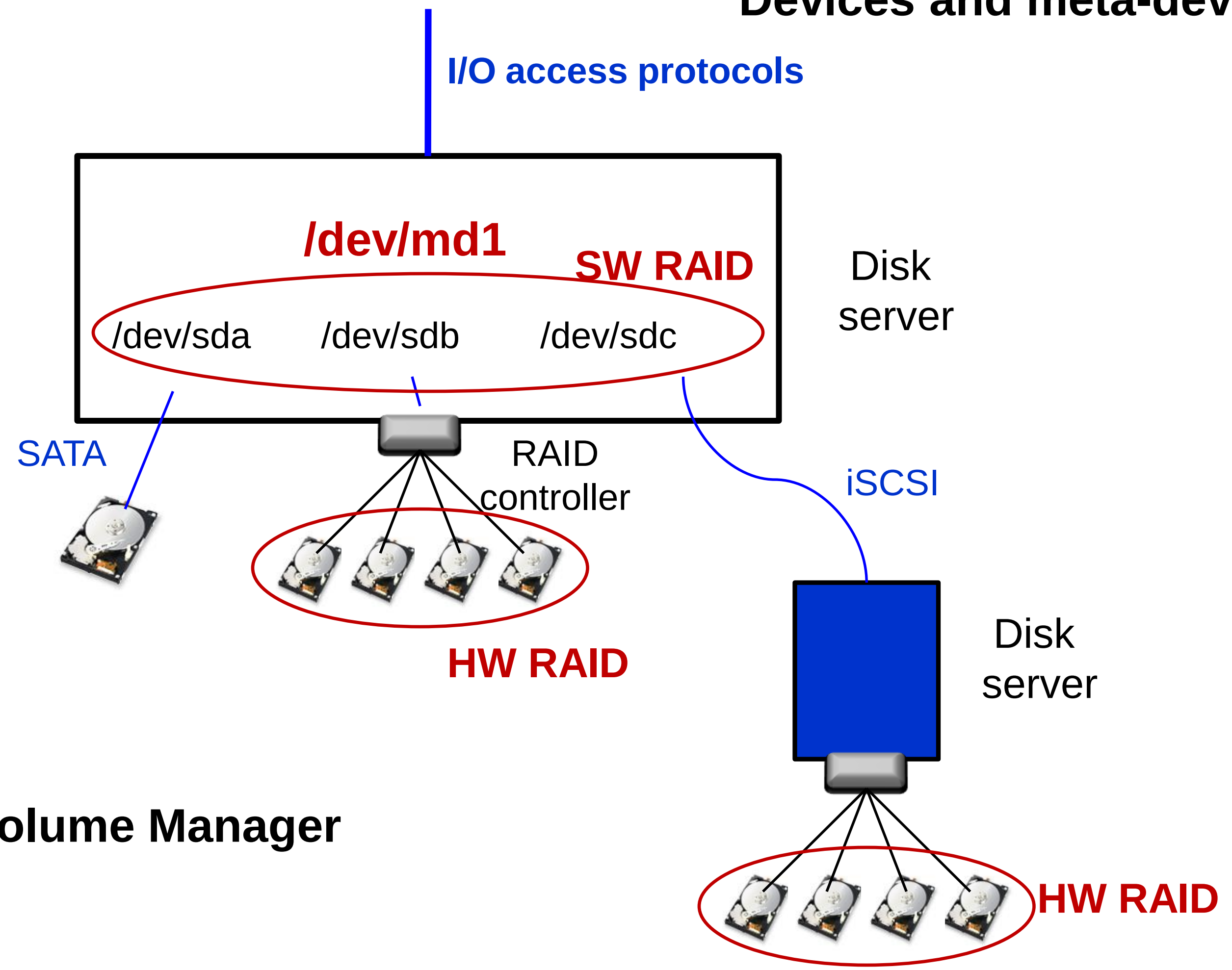


Device Diver III

Creating RAID aggregates based on various combination of devices

Hardware RAID ← → software RAID

Devices and meta-devices



Tools :
LVM Logical Volume Manager
raidtools

System I/O scheduler

Management of the I/O layer in the Linux kernel

→ Linux I/O scheduler   Linus Elavator

Read/write blocks are mappings of disk cylinder/head/sector

Block requests are put into a queue and sorted sequentially
→First order disk IO optimization

Time ordered FIFO queue in addition with request expiration times
→ Avoids ‘starving’ of small requests

Look ahead algorithms for further tuning

Read is synchronous and write is asynchronous
Favors the writing of data

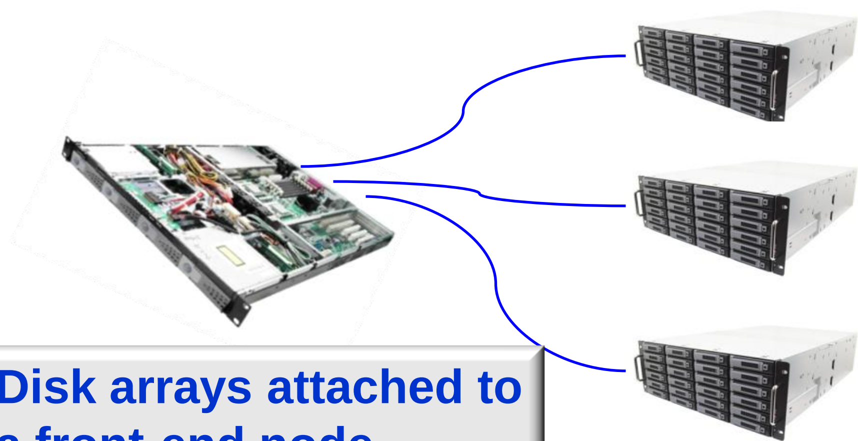
Some tuning parameters which effects the throughput and the balancing of read and write streams

Disk Server

How much Processing capacity is needed for the storage servers ?!

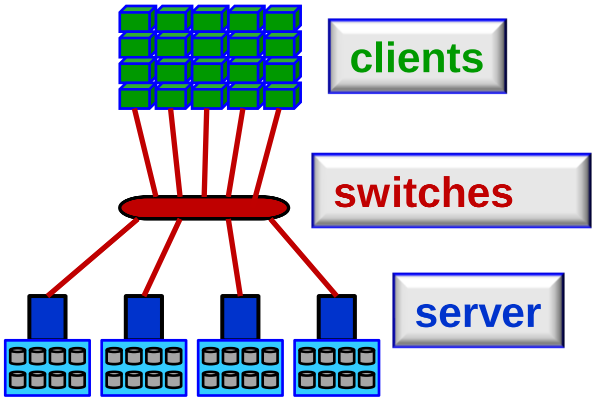
- The actual data transfer, disk ← → network
- Compression and decompression
- Encryption and decryption
- File system operations (list, find, move, delete, etc.)
- Daemons for the higher layer data management software
- Monitoring of the system
- Data integrity checks
- File transfer daemons and protocol

Integrated disk server

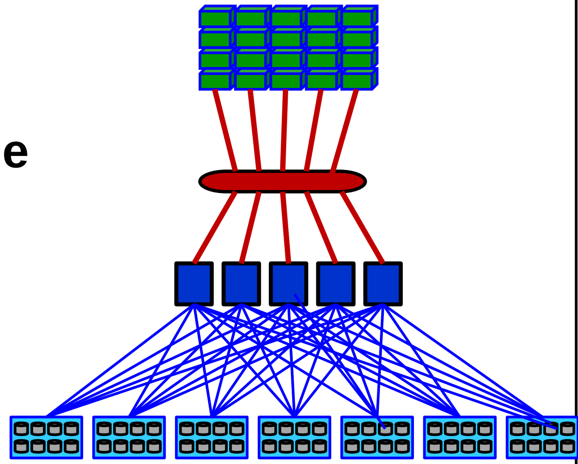


Disk arrays attached to a front-end node

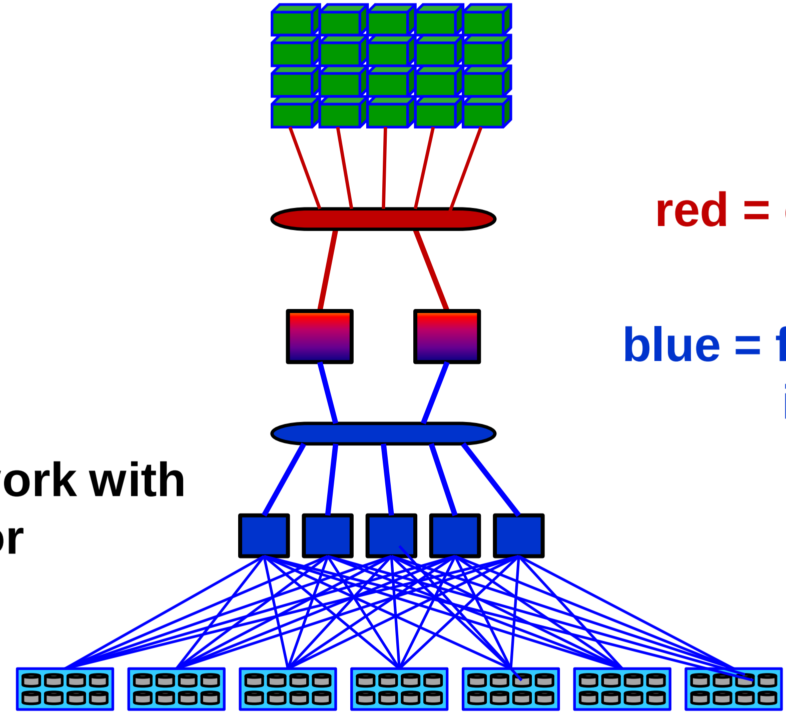
Storage network topologies



Homogeneous network between the storage nodes and the clients



Different network type between the Disk server and the disks



Storage Area network with front-end nodes for storage export

red = ethernet

blue = fibre channel
infiniband

Chapter 4

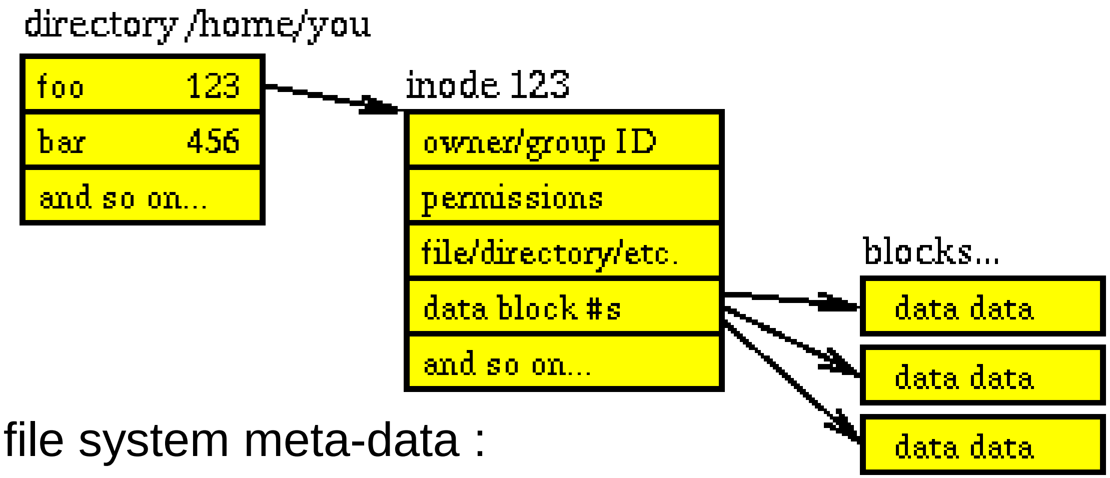
File Systems

CERN School of Computing

Local file systems I

Make the storage devices available to the user applications

Physical :
Mapping of disk blocks to files
Logical:
Hierarchical arrangement of directories



Stores the actual file data and structural file system meta-data :

- **Superblock** → file system type and size, mount status; several copies mixed with the file data
- **Inodes** → file type (executable, block ,etc.), access times, file size, owner and group, ACLs Access Control Lists, number of links, etc.
- **Directories**
- **Journals** → transaction logs, separated from the data, allows easy and fast recovery from crashes, enables data consistency

CERN School of Computing

Local file systems II

Most common file systems under Linux are :
EXT3 , EXT4 (lately integrated and released) , XFS, ReiserFS, JFS, OCFS

Selection criteria :
Performance and functionality differences versus support quality and wide spread experience

Developments :

ZFS developed for SOLARIS, Linux port via FUSE, license issues
key features are

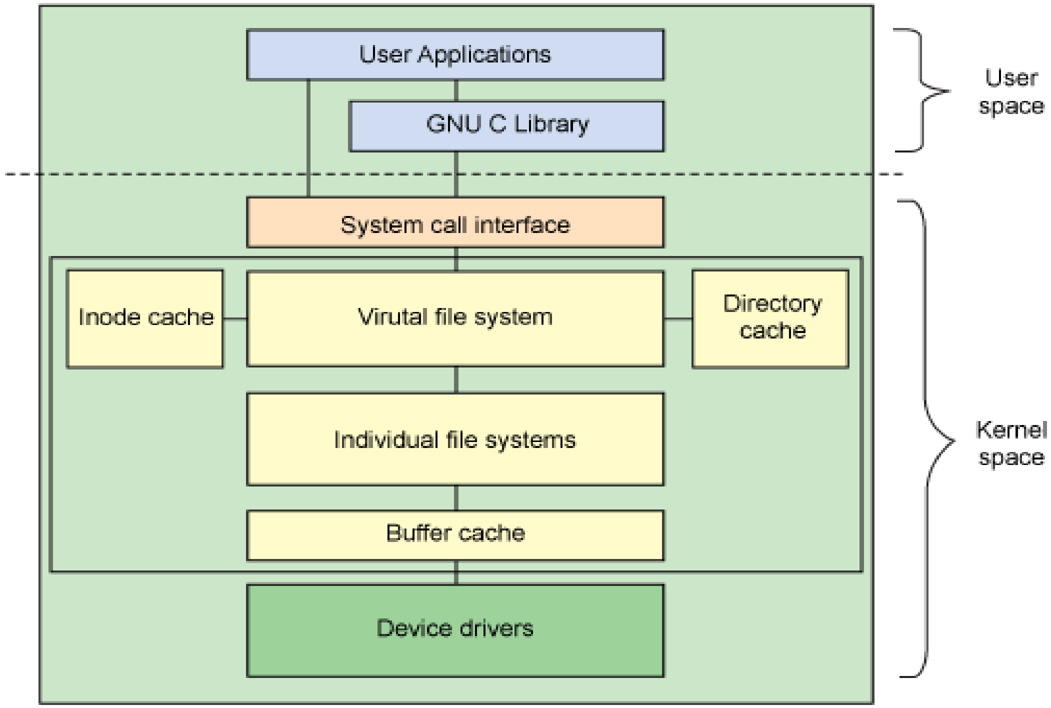
BTRFS features are writable snapshots, pooling of multiple devices, efficient small file support, and optimization for SSDs

Benchmarks and comparisons can easily be found on the web, but not easy to interpret

http://www.phoronix.com/scan.php?page=article&item=ext4_btrfs_nilfs2&num=1

CERN School of Computing

Local file systems III



Some tuning parameters

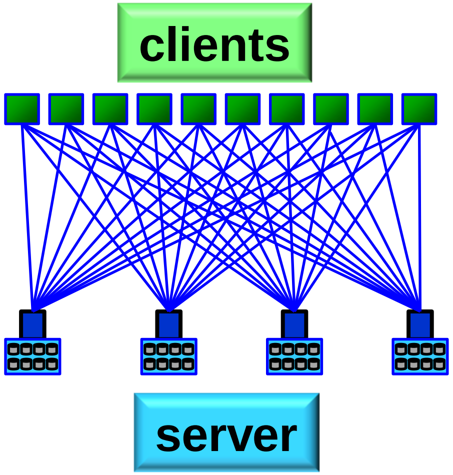
- The number of inodes defines the number of files in a file system
- Intrinsic block size (1-4kB) (multiple of disk block size0
- Application level direct IO, avoiding the buffer cache
- Buffer cache flushing algorithms
- Journaling options → information level versus safety and speed, extra disk

Limits :
Maximum length of file names, size of files, size of file system
File system dependent performance penalty for large number of files in a directory

Network file systems I

Most simple variant is NFS, Network File System (version 3)

- Remote mount a server disk partition on a client node
- Not very scalable, becomes quickly unmanageable
  When trying to mount many servers on many clients
- No redundancy, server failures critical
- Simple security implementation
- Wide spread, plenty of experience and tuning guides



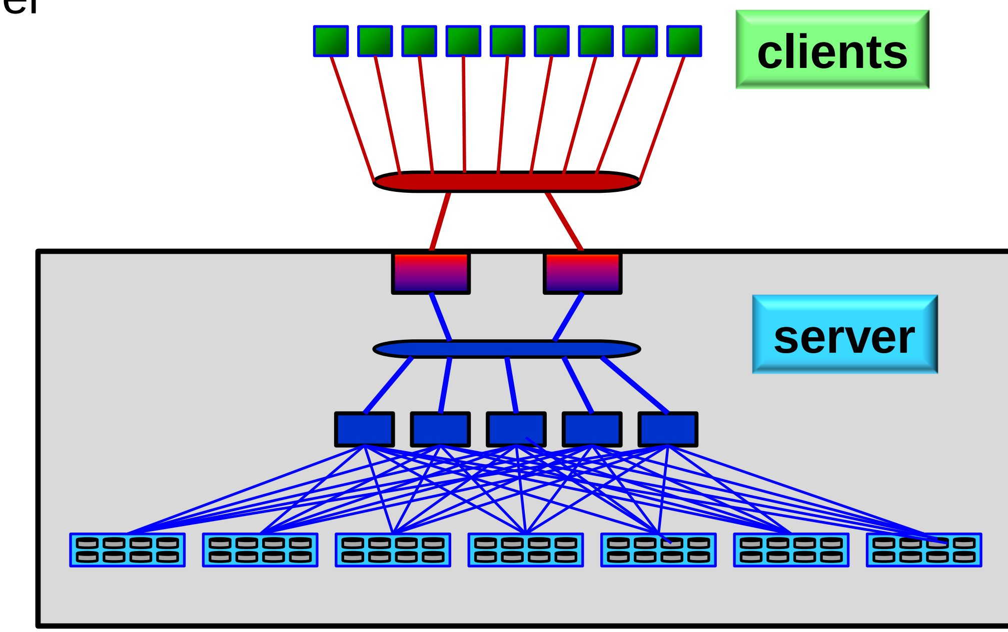
http://nfs.sourceforge.net/nfs-howto/

CERN School of Computing

Network file systems II

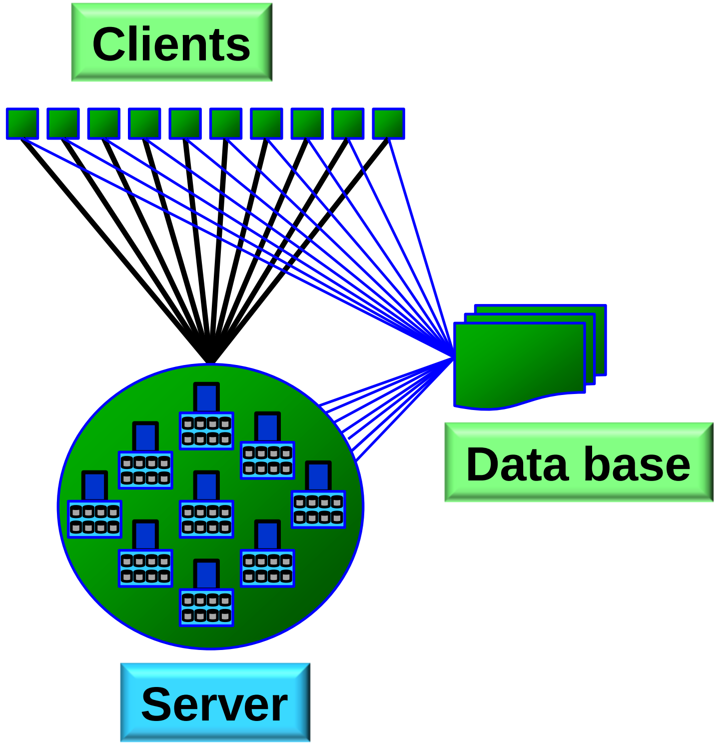
Commercial hardware based NFS solutions

- At least 60 different vendors (EMC, NetApps, Isilon, Blue-Arc, …)
- NFS compliant clients, but proprietary server
- 'Looks' like a simple NFS server,
  but has much better :
  - Scalability in size and performance
  - Security
  - Redundancy and fault tolerance



Cluster file systems I

Aggregation of local file systems and Server nodes



Meta-data server is the new important component
→ Mapping of files to locations
→ Data base implementation (Oracle, MySQL, ….)

Control data flow between the clients and the Meta-data server

Data flow directly between clients and disk server

Two types of implementations :

1. Device driver implementation via the virtual file system
   the application accesses the data via a File system syntax
   mount point, looks like a local file system, same commands (ls, rm, mkdir, etc.)

2. Translation of application IO commands (open, read, write, seek, close) via
   special IO library linked into the executable. Special commands for ls/rm/mkdir …

CERN School of Computing

Cluster file systems II

AFS Andrew File system: open-source, home directory usage, small files storage, not tuned to high performance, long term experience, problematic load balancing

DPM Disk Pool Manager: open-source, CERN development, used in > 200 T2 and T3 sites, large file storage, no mount point – IO libraries in the application

GFS Global File System: open-source, commercial support available, ~100 nodes optimal

GPFS General parallel File System: commercial (IBM), origin in the US ASCI supercomputer initiatives, large scale parallel IO

Lustre 'Linux Clustre': open-source, commercial support, large scale clusters, origin also in the supercomputer initiatives

pNFS Parallel Network File System: open-source, extension NFS4, standard soon, first prototypes available

Cluster file systems III

Performance depends on :

- ➢ The underlying hardware
- ➢ The implementation of the transfer protocols
- ➢ The meta-data server
- ➢ The implementation and configuration of server caching and client caching
- ➢ Configuration of block sizes, striping factors, read-ahead values
- ➢ Load-balancing mechanism → requirement for homogeneous hardware

Redundancy in case of server failures is a weak point

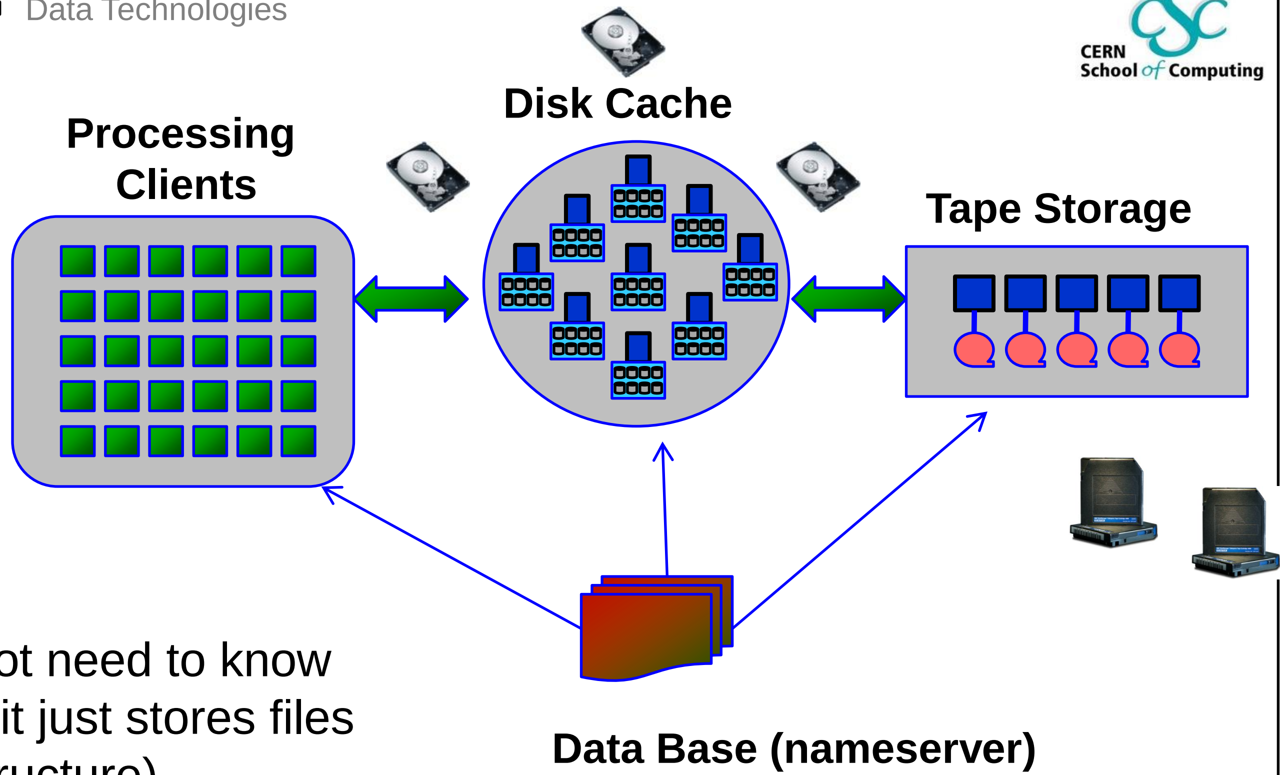
- Mostly not or weakly covered in the software design
- Configuration of time-outs and retries
- Relies heavily on a redundant hardware setup :
  → dual controller, dual network links, RAID, SAN, homogeneous and high quality hardware, etc.

HEP cluster file system evaluations :

http://hepix.caspur.it/storage/hep_pdf/2008/Spring/Maslennikov-FSWG-Final-Report.pdf

Mass Storage Systems I

Adding a tape storage system to a cluster file system



The disk layer is just a cache for the tape storage, assuming that   tape space >> disk space

The application in principle does not need to know About the details of this hierarchy, it just stores files Under a unique name (directory structure)

<u>Hierarchical storage</u> : transparent internal movements of data between different storage pools (aggregations of disk servers)
e.g.
Fast SSD disks → SATA disks → low access disks (can be powered down)
→Fast tape pool with many drives → low access tape pool

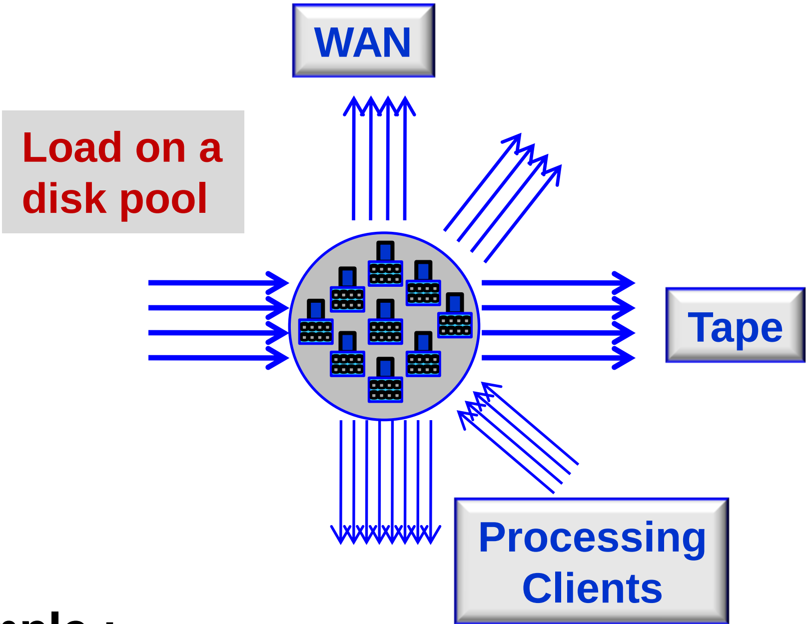
In High Energy Physics application with varying IO requirements and large storage demands even the “simple, two layer hierarchy” does not work well

CERN School of Computing

Mass Storage Systems II

Optimization of performance with quite different concurrent application requirements:

- Tape and WAN are critical components
- Small files for tape don't work
- Write priority for tape
- Write on disk has priority over read
- Low number of high speed streams versus large number of low speed streams
- Streams per disk spindle is key
  → Space will be 'for free'
  but with a need for a guaranteed minimum disk space, SSD disks are not yet an alternative for large pools



Example :

Disk pool = collection of disk server

- 1 GB/s input, 20 streams
- 1 GB/s output to tape, 30 streams
- 1 GB/s output, 2000 streams
- 0.2 GB/s input, 2000 streams
- 0.3 GB/s output to WAN , 100 streams

CERN School of Computing

Mass Storage Systems III

HPSS High Performance Storage System: commercial (IBM), large scale supercomputer installations
http://www.hpss-collaboration.org/hpss/index.jsp

CASTOR CERN Advanced Storage Manager: open-source, CERN development of an integrated mass storage system
http://castor.web.cern.ch/castor/

GPFS + TSM Cluster file system with hooks into a backup system
Tivoli Storage manager: commercial (IBM)
http://www-01.ibm.com/software/tivoli/products/storage-mgr/

dCache + ENSTORE Disk pool manager developed at DESY and a mass storage interface developed at Fermilab
open-source
http://www.dcache.org/
http://www-ccf.fnal.gov/enstore/

CERN School of Computing

Mass Storage Systems IV

Industry term : Data lifecycle management (DLM)

- Combining hierarchical storage, mass storage and backup
- Products available from EMC, HP, IBM, Symantec,……

"….appropriate combination of storage devices, media types, and network infrastructure to create a proper balance of performance, data accessibility, easy retrieval cost, and data reliability…."

- Difficulty to define the right strategies for data storage and movement

Chapter 5

Storage characteristics

Storage system requirements

What are the requirements from the application(s) for our storage system ?

→ Define the I/O characteristics

1. Sequential I/O performance ← → random access I/O
2. Relative and absolute read and write operations
3. Management I/O operations
4. Number of concurrent I/O streams
5. Amount of space needed
6. Expected growth rate, space and performance
7. Expected availability (24*7, max downtime of x hours per year)
   → backup, redundancy level

CERN School of Computing

Storage system parameters I

- Raw data transfer speed versus Control data flow speed
- Time to open a file (authentication, filename resolution, database search, server contact, protocol and daemon response times, etc.)
  1. Abstract name : Muon data from 10th June 2011
  2. Logical name : /experiment A/raw/muon/date/data12345
  3. Site mass storage : Goettingen:/mass storage/exp A/raw/muon/…
  4. Site specific : disk server 33:/filesystem/data/…../bla12345
  5. Node specific : device 17, raid5 controller 2, blocks 147464-148000
  6. Disk specific : cycinder 57, track 45, sector 120-138
- Time to delete a file
- Time to register a file (close)
- Time to list files
- Dependencies on number of files and file size distribution
  → many small files versus few large files

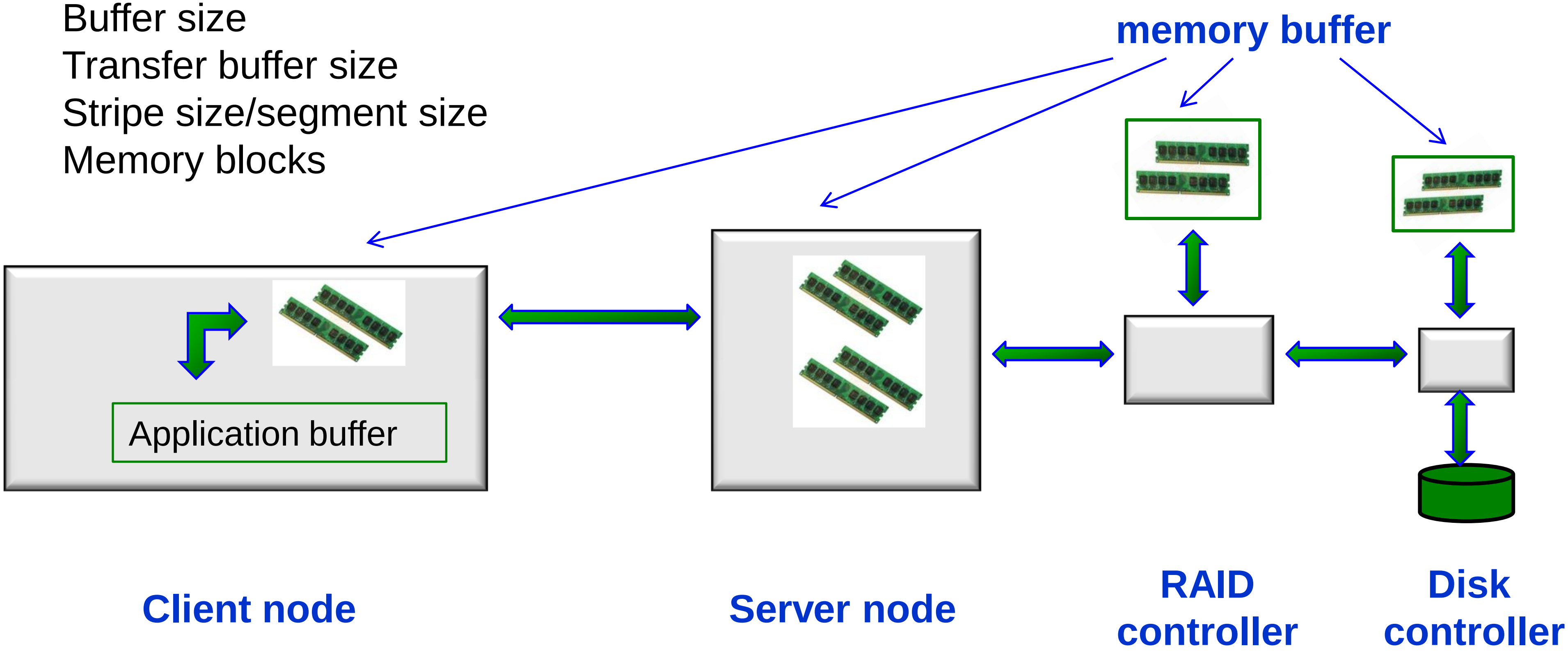
CERN School of Computing

Storage system parameters II

Caches all over the place

Tuning of parameters:

Read-ahead
Write-back/Write-through
Block size
Buffer size
Transfer buffer size
Stripe size/segment size
Memory blocks



Storage protection

Need to protect against power interruption
- Voltage spikes
- Micro-cut
- Short interruptions <5min
- Long term power break-down

Availability of the services and data loss protection
Write caching improves considerably the performance, but
In case of power loss, data is lost

Software protection → transaction logs, syncing of data writes

Battery backup for caches (e.g. RAD controller)
UPS (uninterruptable power supply) for the servers = large scale battery backup to cover interruptions of up to 5 min

Diesel generators for long term interrupts and critical service protections

CERN School of Computing

Debugging Tools

Linux , low level system analysis tools

- top atop
- vmstat mpstat
- netstat
- iostat
- strace
- lsof
- nfsstat

(sysstat package)

Prototyping and scaling tests !!

Disk, file system

- hdparms
- tune2fs

Benchmarks

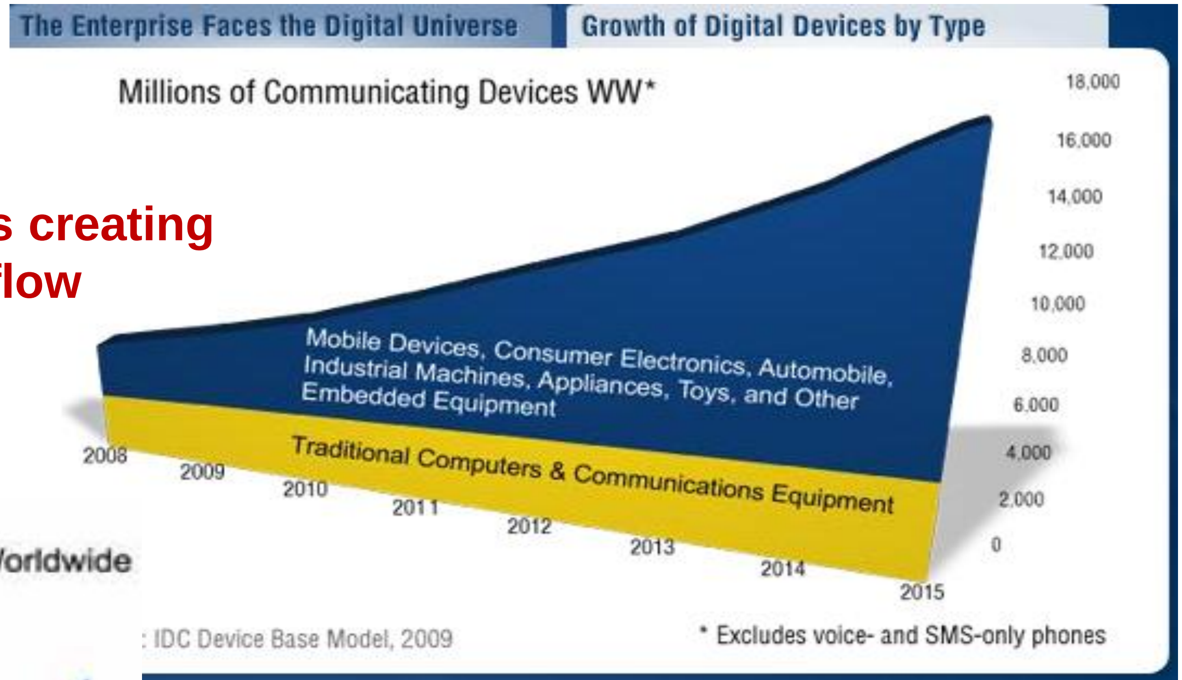
- Bonnie
- Netperf
- Protocol specific read/write client program

Chapter 6

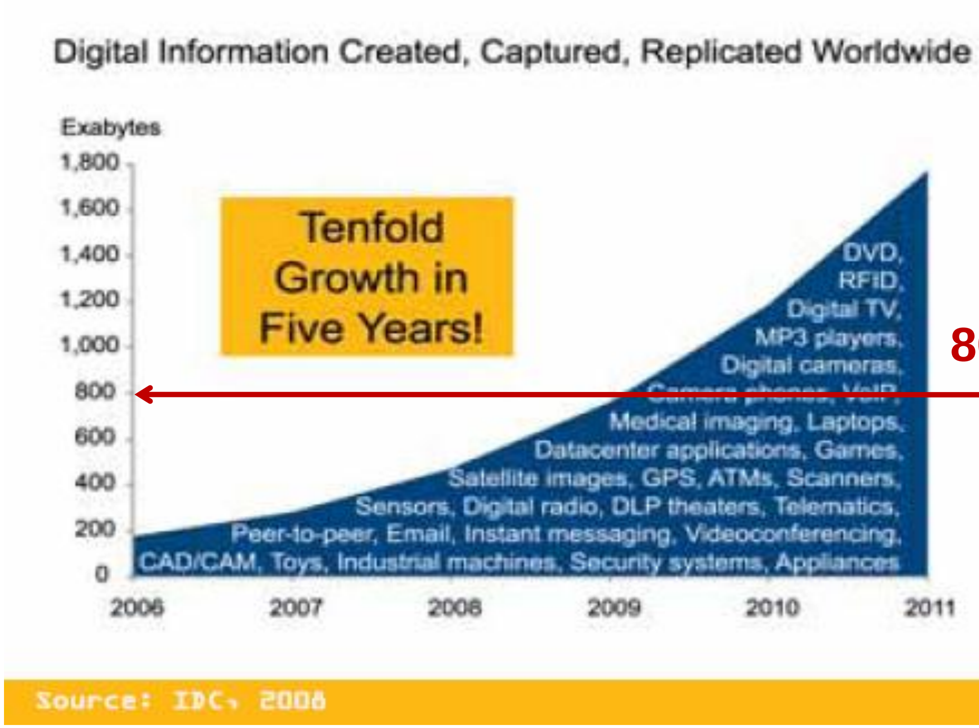
Information and the power problem

CERN School of Computing

Information Growth



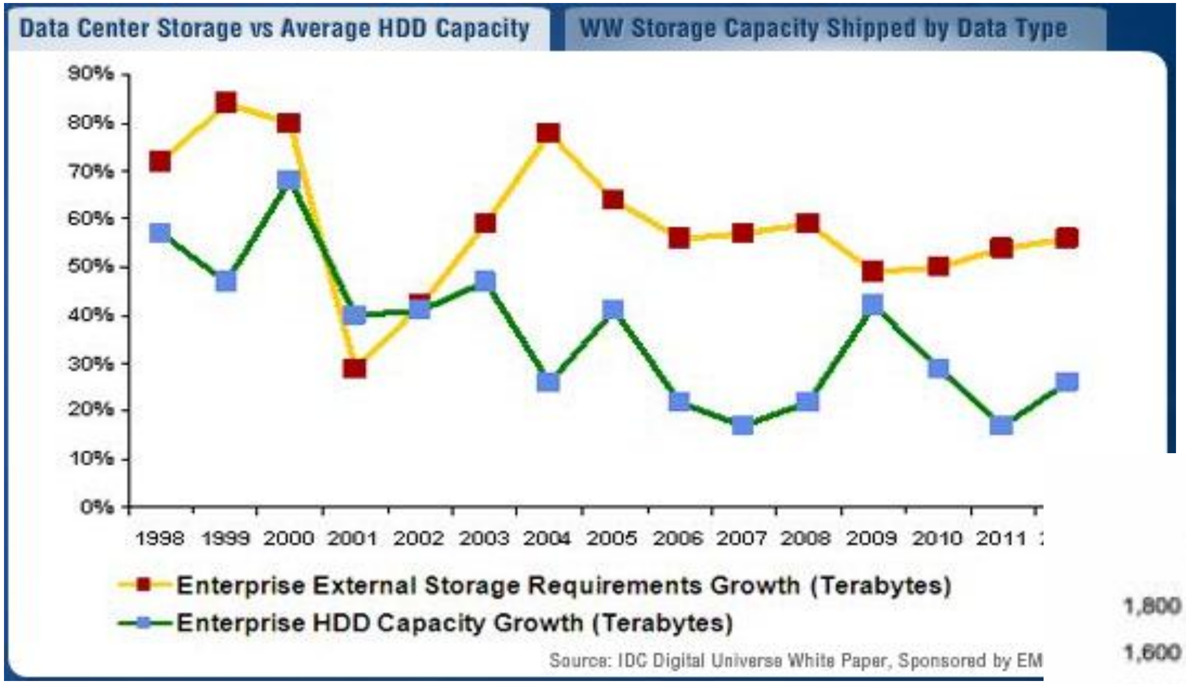
6 Billion devices creating an information flow



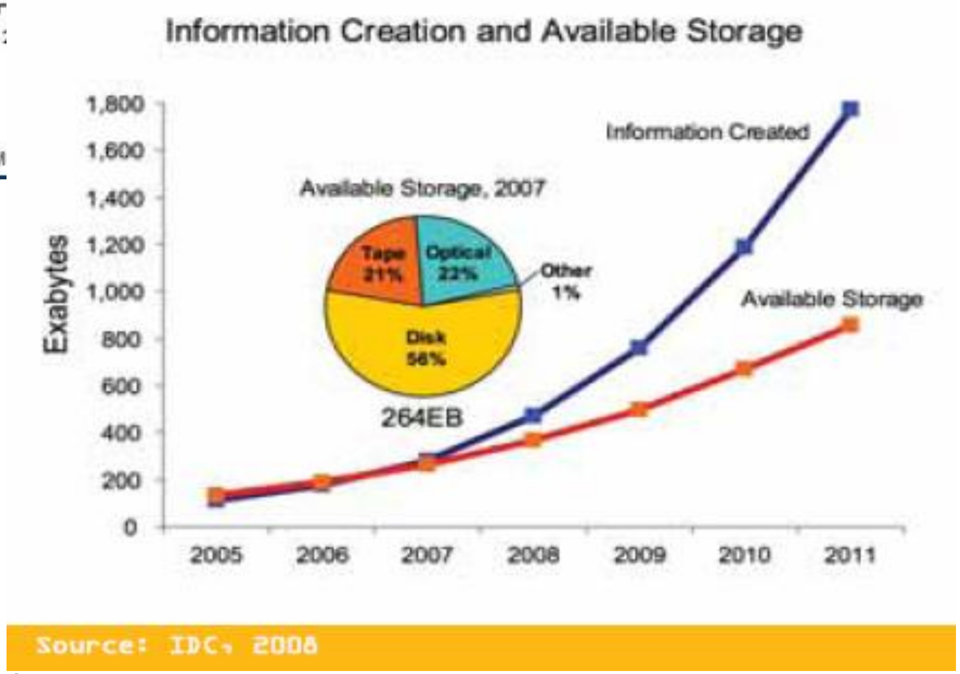
800 ExaBytes, 10^{18} Bytes

CERN School of Computing

Storage Problem



The information growth is exceeding the available storage capacity in the market



The Power Problem I

Electrical power usage scales with frequency and voltage

$P \sim V^2 \quad P \sim f$

Power crisis in the processor industry (2005/2006)

→ Deviate from performance through frequency increase to performance through more processing units   multi-core

Side effects on storage

→ More cores = more programs = more streams
Good parallel programming is the exception

Many concurrent sequential read/write streams

→ = random access performance = bad performance

At the same time on the storage side

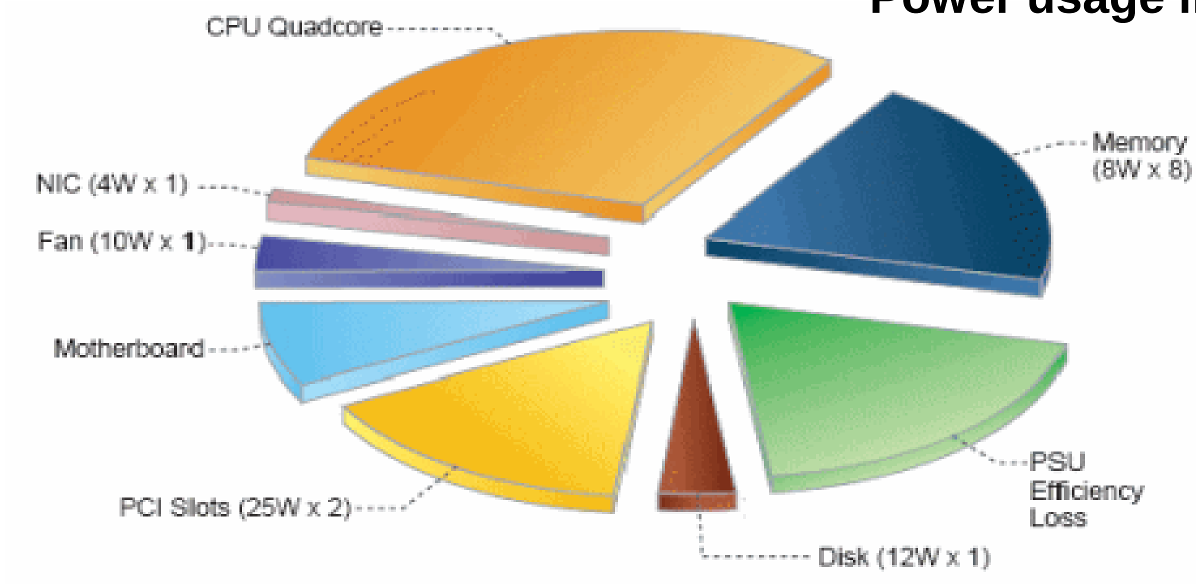
→ higher capacity disks with stagnating performance
Green drives with varying RPM

SSD disks 'easing' the problem, but cost issues

CERN School of Computing

The Power Problem II

Power usage in a PC node



The “Google-Way” :
Specially designmotherboards,
Local Battery,
High efficient power-supply, 12V-only, efficient voltage-converters

CERN School of Computing

Ambiente

Ambient conditions

→ **Leakage currents,** the processor chip leakage currents increase with temperature, ambient temperature dependent, negative feed-back loop

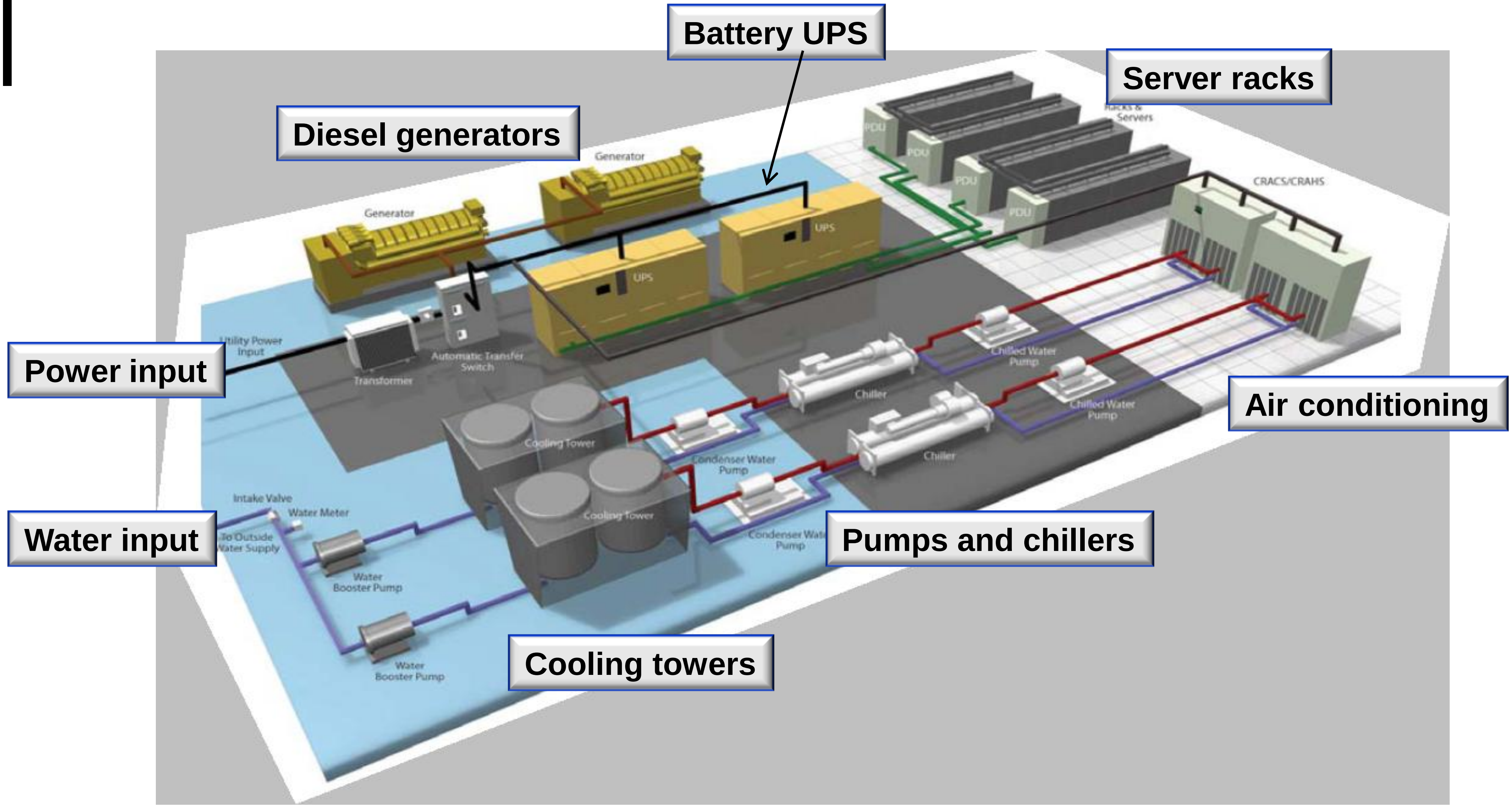
→ **Vibrations,** server quality disks have acceleration sensors integrated to protect against strong vibrations, possible reason for lower MTBF values

video

http://www.youtube.com/watch?v=tDacjrSCeq4

CERN School of Computing

Computer Center Layout

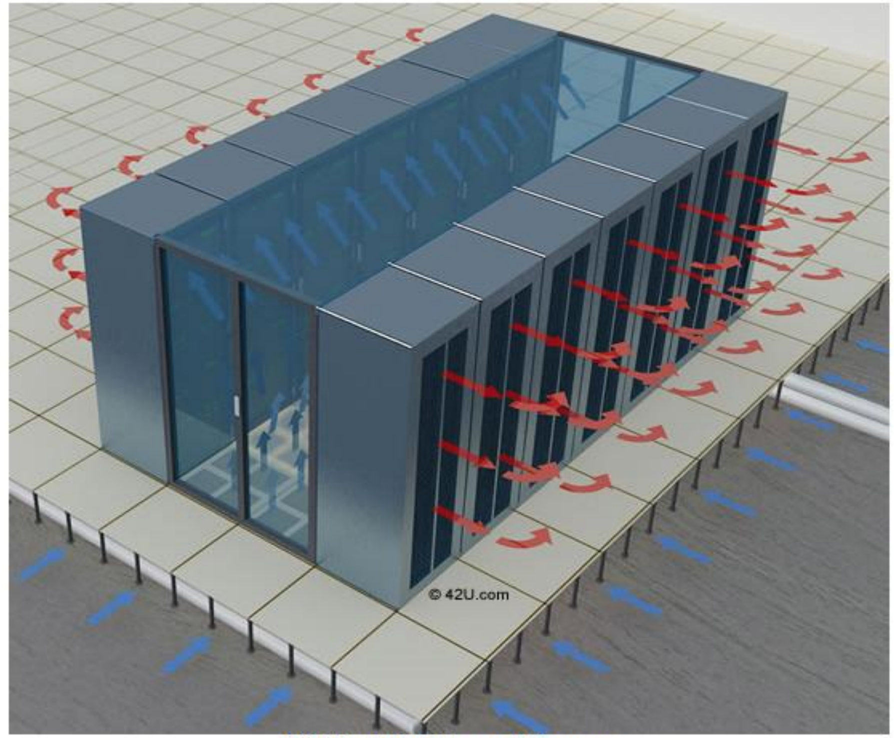


CERN School of Computing

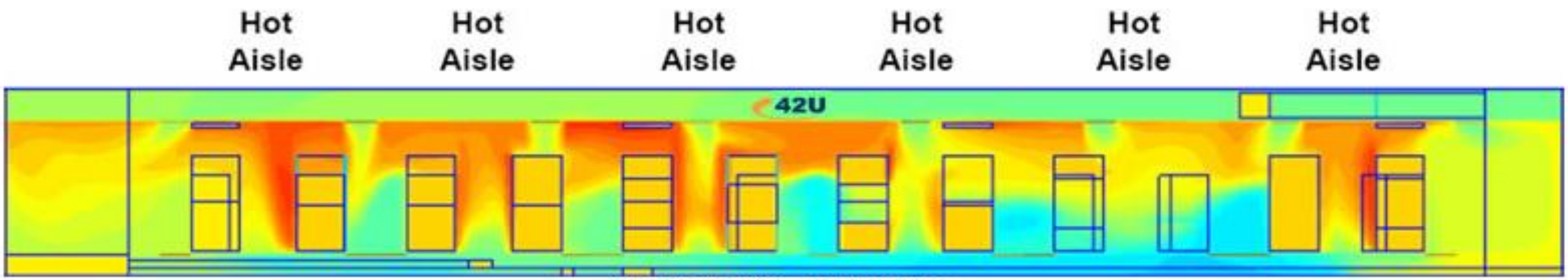
Cooling techniques I

Air cooling with hot and cold aisle on a raised floor

2-5 KW/m^2 density wide spread in data centers

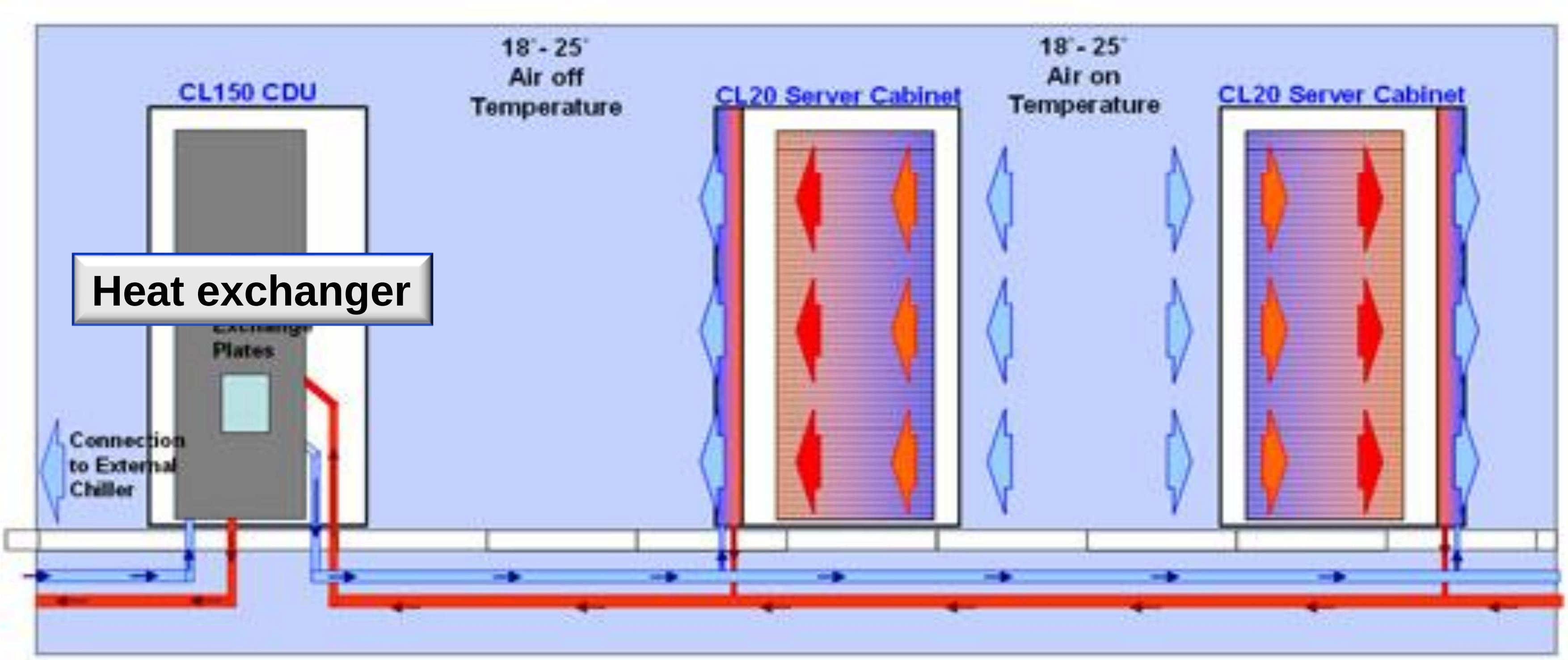


Cold Aisle Containment Diagram



SynapSoft™ 4.0 - Thermal Map

Cooling techniques II



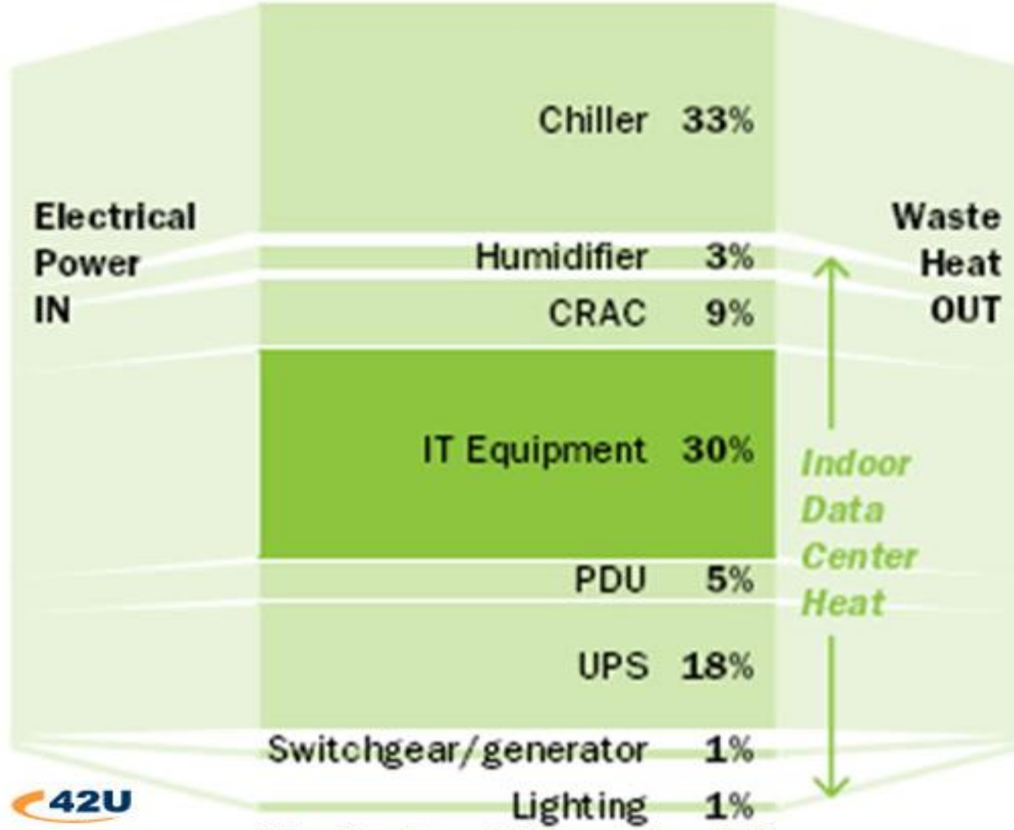
Water cooled racks, IRC In-Row Cooling

Still using the hot and cold aisle concept

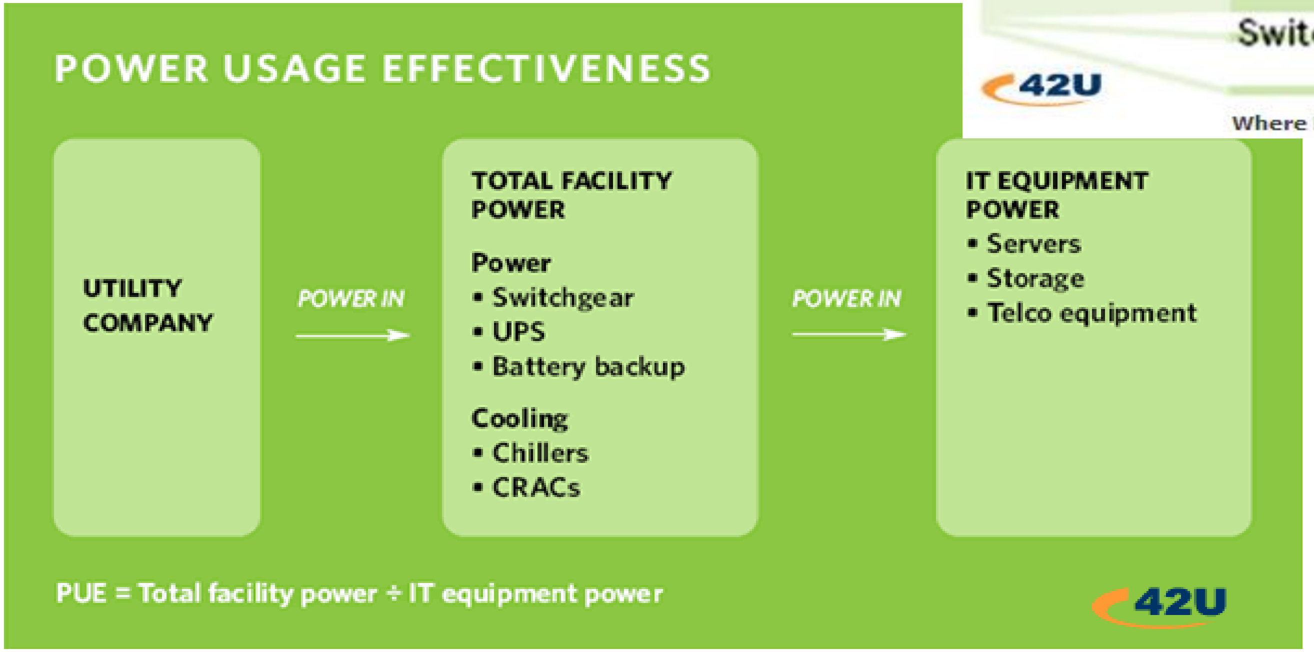
With densities above 10 KW/m^2

Power and cooling efficency

PUE as a measurement for Efficiency and green computing



Where Does Energy Go? (source: Green Grid)



Power Usage Effectiveness - PUE (source: Green Grid)

CRAC computer room air conditioning

CERN School of Computing

Computer mega-center

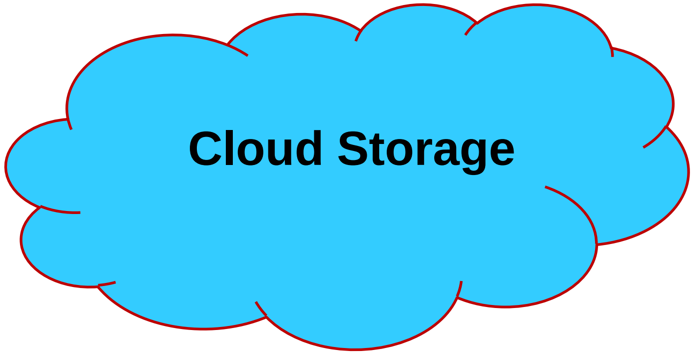
- Google computer center (project 02)
  Area of 2 football fields, estimate > 1000000 cores
  Columbia river (US) , cheap hydro electric power available
  Amazon started a building there too

- In total Google has probably > one million servers world-wide
  → 200 MW electricity needs plus cooling

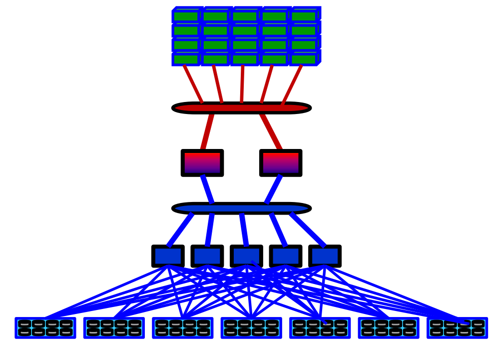
- Each server with 2-3 disks → 1000 PB ?!

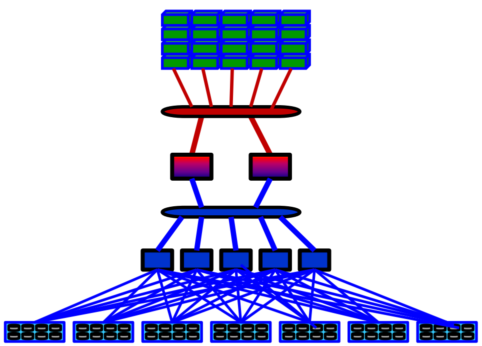
video

http://www.youtube.com/watch?v=zRwPSFpLX8I



Chapter 7
Clouds





CERN School of Computing

Cloud/Grid computing and storage I

Cloud/grid storage implementations vary:

- Coupled cluster file systems plus another layer of software/middleware and an additional data base management system

  example:

  The different Tiers (T0, T1, T2, T3) in the LCG project have different file system and mass storage implementations (Castor, dCache+Enstore, HPSS, GPFS, DPM, etc.). They are ‚linked' by a common software layer called SRM, which hides all the different storage implementation and access details from the user (srmput, srmcopy, srmget, etc.). The additional data base setup is provided by the experiments

- Extended cluster file systems which span over several geographic locations

  example:

  Hadoop
  the Google file system
  BitTorrent is in principle a world-wide file system implementation
  AFS, Andrew File System, is not an explicit one, but used in HEP for some world-wide access to user home directory data
  some enhanced Lustre and GPFS implementations

CERN School of Computing

Cloud/Grid computing and storage II

Originally companies have build large data centers to cope with demands for their core business and only later started to 'sell' free capacity to everybody.
→ This created new business models Amazon, Google, etc.

Another Driving force in the market is a strong commodity trend :
More than 50% of currently sold PC's are notebooks
and the Netbook share has the highest growth rates

→ **Netbook + Cloud Computing** !!

This goes along with various cloud interface and programming models:

CloudStore, Hadoop and Hypertable, Amazon EC2, Backup S3 ElasticDrive, Nirvanix GoGrid, Vmware vCloud,…..

Cloud/Grid computing and storage III

Quite a few companies are offering ‘low cost’ online backup space :
box.net, Live Mesh, DropBox, Oosah, JungleDisk, Mozy, Nirvanix, ………
But
e.g. The HP ‘experiment’ in this area failed last year on a large scale

Interesting site which monitors performance and availability of sites providing Cloud computing services :

http://www.cloudclimate.com/

But what if :
-- network interrupts
-- service disruption
-- company goes bust
-- deletes your data (human errors)
-- changes the price strategies and you have to move the data

All that has already happened with a variety of companies
→ TCO Total Cost of Ownership

Summary (sort of…)

There is no generic recipe on how to build a storage system.
Fast product evolution and changing commodity equipment trends requires a constant evaluation and adaptation of a storage system.

The focus must be on the overall behavior of the total system, don't get lost in technical details of the single components

Performance is of course a major parameter, but the driving costs will come from the operational and support efforts

Keep it simple, you will get complexity 'for free'

Don't fall into the 'free-software' trap,
→ Total Cost of Ownership must be considered
→ home-grown versus commercial products

CERN School of Computing

Links

General hardware information :
http://www.tomshardware.com

Storage industry:
http://searchstorage.techtarget.com/
http://www.byteandswitch.com/

Clouds

http://www.usenix.org/events/fast/

http://cloudslam09.com/

HEP storage talks and information
http://www.hpc2n.umu.se/events/workshops/09/hepix/conferenceTimeTable.py.html