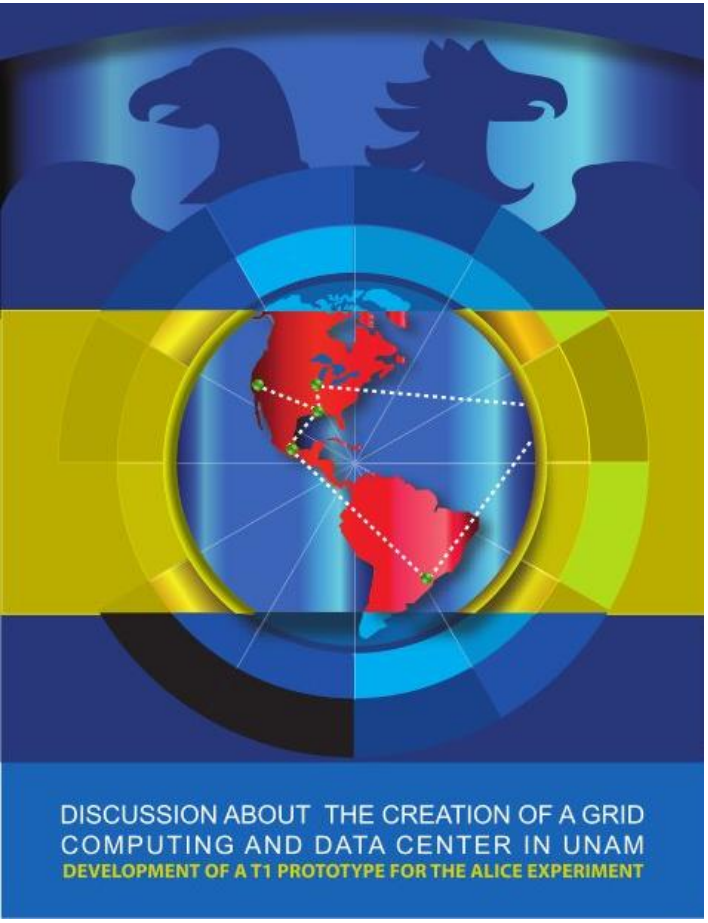
Grid of the Americas Workshop



Storage Trends

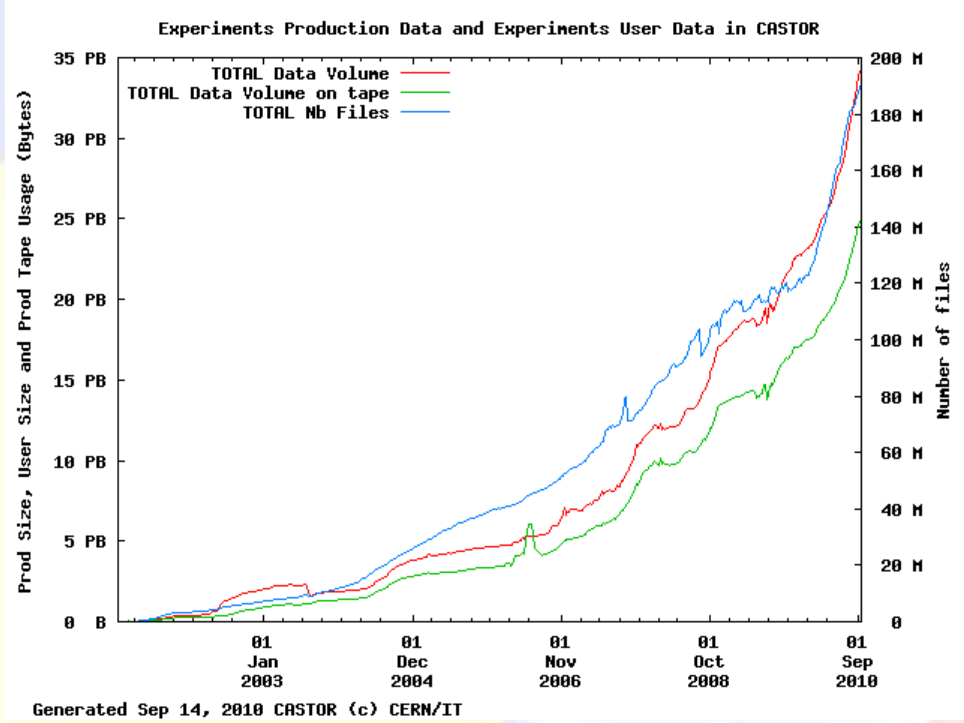
P. Vande Vyvre – CERN/PH
ALICE experiment

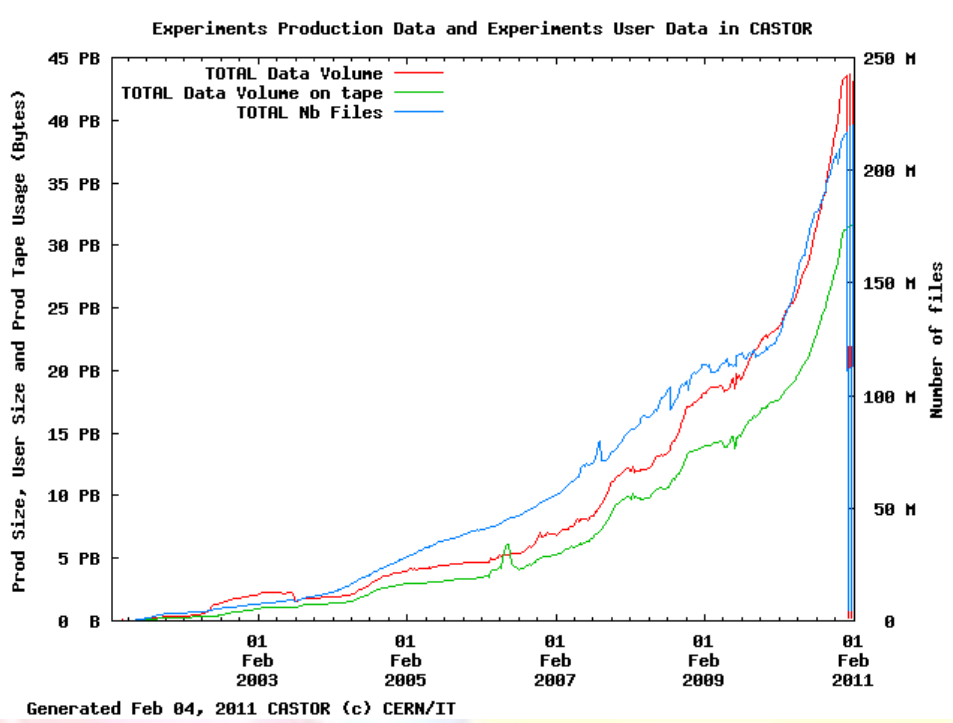
Outline

- LHC Data Storage Challenges
- Storage technologies (Optical, Magnetic, Solid State)
- Storage attachment
- Magnetic Disks
- Magnetic Tapes & Tape libraries
- Global File Systems
- Storage architecture
- A case study
- Conclusion

LHC Data Storage Challenges (1)

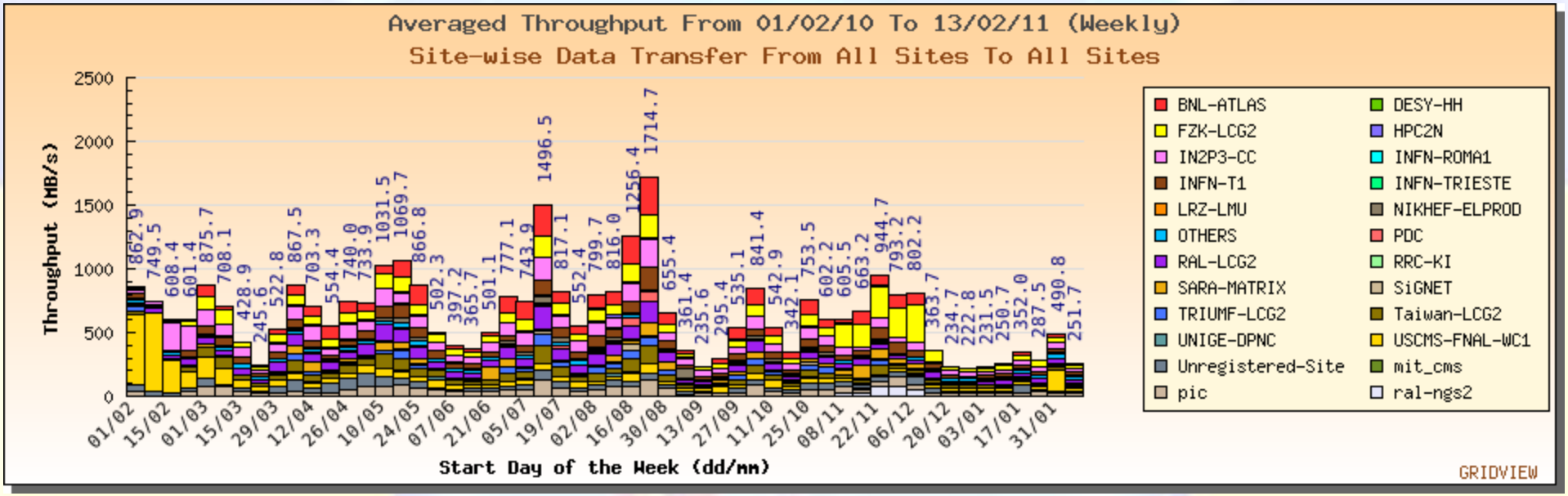
- Challenge 1: Recording the data
  - Lots of data… (due to inexpensive sensors and cheap computing)
    Total amount of data in CASTOR at CERN mid-Sept 2010:
    35 PB of data – 25 on tape
  - … with a challenging slope
    Amount of data added to CASTOR at CERN within the last 150 days:
    +5 PB of data – +5 to tape
    Data taking (pp, PbPb), frenetic data reconstruction and analysis, internal CASTOR housekeeping





LHC Data Storage Challenges (2)

- Challenge 2: Sharing the data
  - Make data available to end users
  - Huge network and storage load: read data at T0 and write them at T1's



LHC Data Storage Challenges (3)

- Challenge 3: Accessing and processing the data
  - Some tasks are purely sequential (e.g. reconstruction)
  - Some tasks have a high random access fraction (e.g. analysis)
  - Identify and quantify the workload
  - The challenge of designing the storage of a Tier 1 is to obtain the best possible performances, at an affordable cost, for all types of tasks.

Data Storage

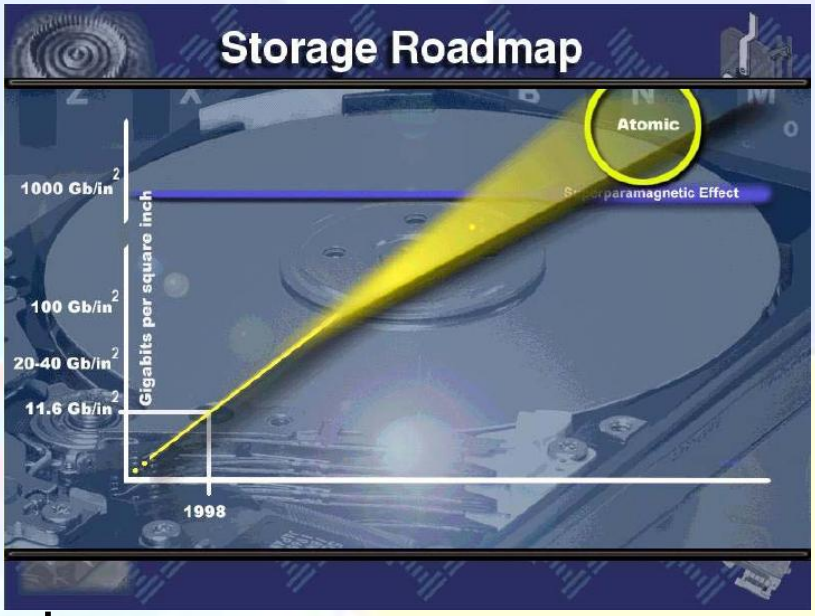
- How can the storage technologies and their current trends help us to address these challenges ?
- Dominating factors:
  - Performance
  - Purchase cost
  - Cost of ownership including manpower, power, etc
  - Carbon footprint
  - Scalability, openness to new technologies
  - Ease of duplication, operation
- Data Storage Technologies
  - Optical
  - Magnetic
  - Solid state

Optical Storage

- Either commodity (CD, DVD) overtaken by magnetic data storage
- Or futuristic R&D project (e.g. holographic data storage) which regularly sets higher target compared to the mass market magnetic storage devices.
  Holographic storage was already THE technology of the future 15 years ago when the ALICE TDR was written.
- Regularly considered by HEP (LEP or LHC time) but does never sustain the comparison with magnetic data storage.
- However, the fact that this technology was not used so far in HEP doesn't preclude its use in the future.
- A good storage architecture should accommodate new technologies and devices.

Magnetic disks

- Highly competitive resulting into huge concentration:
  - 5 large vendors: Hitachi, Samsung, Seagate, Toshiba, WD
  - IBM => Hitachi (2002), Maxtor => Seagate (2006), Fujitsu => Toshiba (2009)
- Performance of commodity disk has exploded with PC market
- Several reliability grades often associated with the attachment technology:
  - SATA (Serial ATA):
  - SAS (Serial SCSI)
  - FCS (Fibre Channel)



- Limited margin for commodity products.
  Some margin preserved by creating several product lines.
- Reliability is an issue: assess it and address it

Hard Disk Reliability

- Reliability - MTBF (millions hours):
  - 1-2 Million hours MTBF often announced by manufacturers
  - 0.4 measured by customers with a large installed base (Google, CERN)
  - The MTBF announced is not measured by manufactures but often calculated on the basis of the failure rate of a set of disks when used in burn-in temperature conditions
  - PBs => tens of 1000 disks => 1 or a few disk to be replaced every day
  - Addressed with Redundant Array of Inexpensive Disks (RAID)

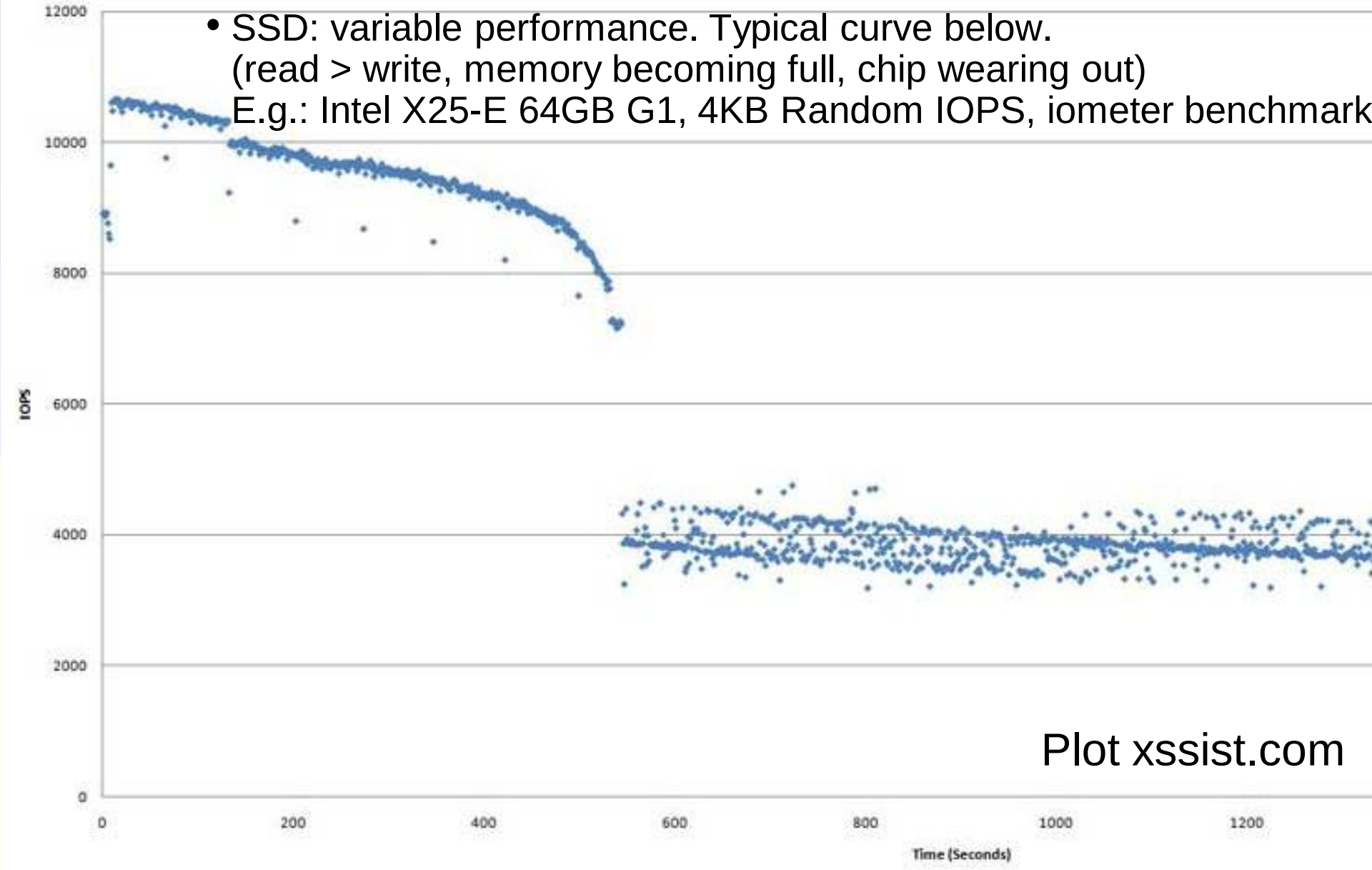
- "Seagate is no longer using the industry standard Mean Time Between Failures (MTBF) to quantify disk drive average failure rates. MTBF has proven useful in the past, but it is flawed.
  To address issues of reliability, Seagate is changing to another standard: Annualized Failure Rate (AFR)."
  Typical value for a server disk: 0.5 %

SSD (Solid State Disk)

- THE new “must have” device
- Solid State only, no mechanical part
- Two technologies:
  - DRAM
    - Volatile
  - NAND-Flash based:
    - Non-volatile
    - Bits are erased and programmed in blocks (EEPROM)
    - Endurance (Flash memory can only be programmed and erased a limited number of times)
      - MLC (Multi-Level Cell) typical life 10,000 or fewer cycles
      - SLC (Single-Level Cell) typical life of 100,000 write cycles
    - Moreover: date cannot be overwritten in situ; only complete pages can be erased. An erase page includes several write pages (Write Amplification factor).
- Excellent MTBF but limited number of P/E cycles
- IOPS: much better than magnetic disk but with variations

SSD IOPS Performance

- SSD: variable performance. Typical curve below.
  (read > write, memory becoming full, chip wearing out)
  E.g.: Intel X25-E 64GB G1, 4KB Random IOPS, iometer benchmark

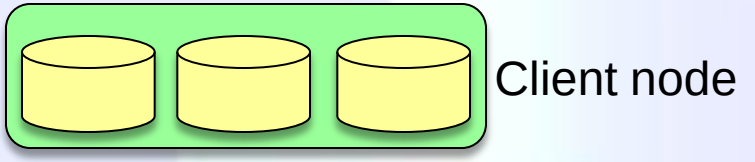


Storage Technologies Characteristics

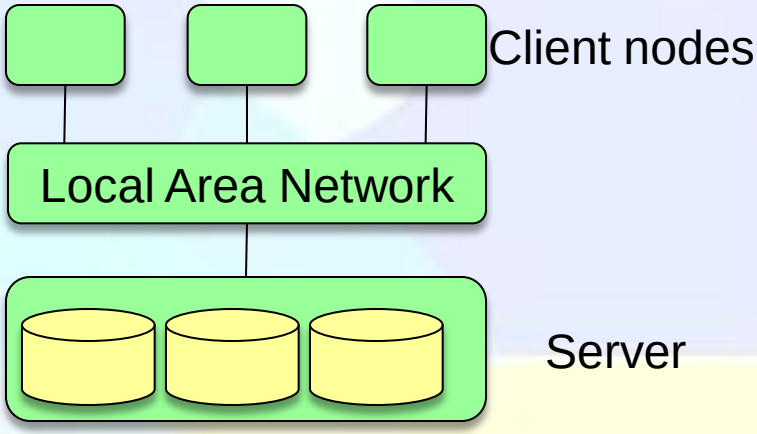
	Optical CD/DVD	Magnetic Disk 3.5" SATA	SLC-Flash SSD	DRAM SSD
Density (Gb/in^2)	0.3 / 2.2	300	30	30
Rotation speed (RPM) Avrg access time (ms)	200-500	7200 4	NA 0.1 - 1	NA 10^{-3}
External Transfer Speed (MB/s)	0.15–10 / 1	130	Read 265 Write 215	Read 410 Write 260
IOPS		300	3k – 30k	70k – 5M
Power Idle (W) Power Operation (W)		7 - 13 3 - 9	0.05 - 0.1 1 - 3	0.1 3

Storage attachment

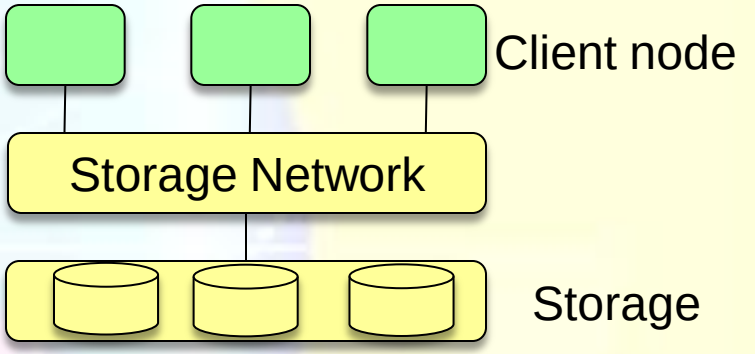
- Direct-Attached Storage
  - SATA: Serial ATA (1.5G, 3G, 6G)
  - SAS: Serial Attached SCSI (3G, 6G)



- Network-Attached Storage (NAS)
  - Ethernet, Infiniband



- Storage Area Network (SAN)
  - FC: Fibre Channel (2G, 4G, 8G)
  - IB: Infiniband
  - iSCSI: SCSi commands over Eth.



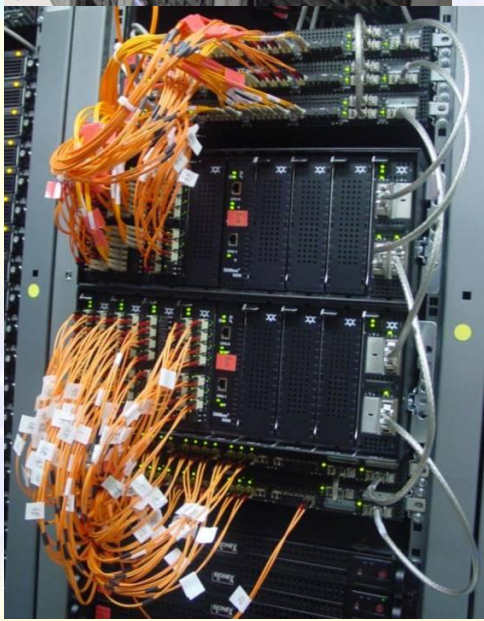
Hard Disks Characteristics

Typical characteristics of hard disks

	Magnetic	Magnetic	Magnetic	SSD
Attachment Std	SATA	SAS	FC	SATA
Attachment Gen.	1.5G & 3G	3G & 6G	FC2G, FC4G	1.5G
Form factor (Inches)	2.5 - 3.5	2.5 - 3.5	2.5 - 3.5	2.5 - 3.5
Capacity (GB)	500-2000	300-1000	150-600	32-512
Rot. Speed (kRPM)	7.2-10	7.2-15	10-15	NA
Media Cost (USD/GB)	0.065 - 0.100 for 3.5 “	0.2 for 3.5”	0.7 for 3.5”	3.5 - 4.5 Flash SSD

Storage Arrays and Racks

- Disk Enclosure
  - Group 12 - 24 disks in 1-4U
  - Interfaces: Inside SATA, SAS, FC
    Outside: FC or Ethernet
  - Inside controller:
    Group the disks into file systems
    RAID 0,1,3,5,6,10,30,50,60
- Rack
  - 15 - 20 enclosures
  - 200 – 400 disks
  - Home-made or integrated (EMC, SGI, etc)
- Storage Area Network
  - Connection between the storage arrays and the servers



Magnetic Tape

- To tape or not to tape ?
  Do we still need tape ?
  A dinosaur device in a brave new world ?

- Market segments:
  - Low-range
    - 8mm, DAT, etc
  - Mid-range for computer centre:
    - LTO (Linear Tape-Open) consortium (HP, IBM, Quantum and many others)
      Presently the 5th generation
    - IBM 3592 linear tape
    - STK 10000 linear tape
  - High-end
    - Sony DIR1000, Redwood etc

Magnetic Tapes

Typical characteristics of magnetic tapes for computing centers

	HP-IBM-Quantum LTO5	IBM TS1130	Oracle/Sun/STK T10000C
Connect. Std	SAS	Fibre Channel FICON	Fibre Channel FICON
Connect. Gen.	6G	2G, 4G, 8G	FC4G
Transfer speed (MB/s)	140	160	240
Capacity (GB)	1500	1000	5000
Media Cost (CHF/GB)	0.1	0.1	0.3

Tape Library

- Automate tape manipulation
- “Infinite” scaleability
- Tens to thousand of cartridges “almost” online
- Still the latency to load a cartridge is in (tens) of seconds

Tapes Library

Typical characteristics of tape libraries for computing centers

	HP-IBM-Quantum LTO5	IBM TS1130	HP-IBM-Quantum LTO5	Oracle/Sun/STK T10000C
Library	HP ESL	IBM TS3500	STK SL8500	STK SL8500
Capacity (Cartridges)	712	16’000	100’000	100’000
Capacity (PB)	0.7	16	150	500

Global File Systems

- Benefits of a global view of the whole storage system
  - Introduce coherence
  - Ease management
- A cluster or global file system is the ideal tool
- Some handle the archiving to tertiary storage
- Performance is possible
- Prefer a hardware agnostic file system
- Must be usable by the data access package: XROOTD in ALICE.

GPFS

dCache

GFS

CASTOR
CERN Advanced STORage manager

lustre

StorNext

System Design (1)

1. Data volume now and in the future (10 PB now)

- Possible to have PBs of data on disk
- Tape storage is still cost effective for very large data volumes and is “infinitely” scaleable…
- …but it has some drawbacks: more complex architecture, media checking and evolution to new generations etc.
- If tape needed, no need to have it online:
  at last HEP is like industry and using tape for what it is good at: archive !
- Be agnostic: review the design criteria and make the maths: cost of 10 PB on tape vs 10 PB on disk
  - 10 PB on tape: 2000 cartridges of 5 TB + 20 drives
  - 10 PB on disk: 300 servers of 33 TB

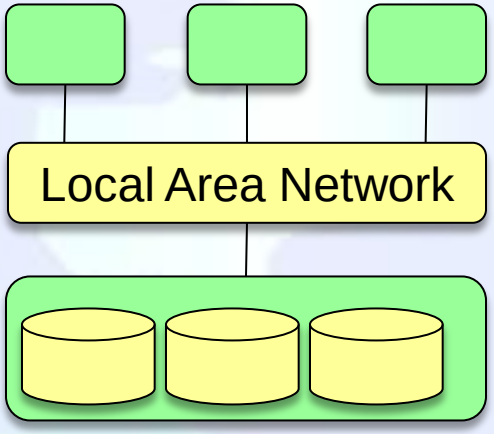
System Design (2)

2. Data throughput now and in the future

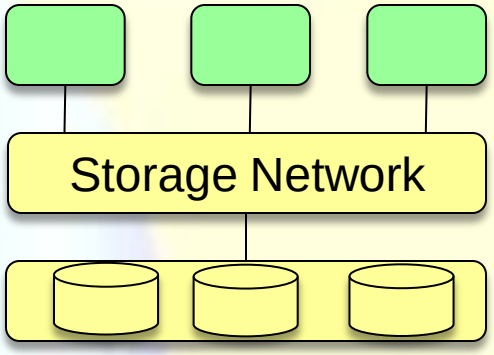
- Input/Output data traffic
- Local data traffic
- Previous decision will have a big impact
  - 10 PB on tape: 20 drives with 200 MB/s => 4 GB/s
  - 10 PB on disk: 300 servers at 100 MB/s => 30 GB/s
- Part of the disk storage can be implemented in SSD
  - A storage array or the server can then saturate a FC8G or 10Gb Eth => 1 GB/s per box
- Depending on the needs, the optimum can be found by mixing technologies, again home-made or integrated.

System Design (3)

3. Design of the basic box

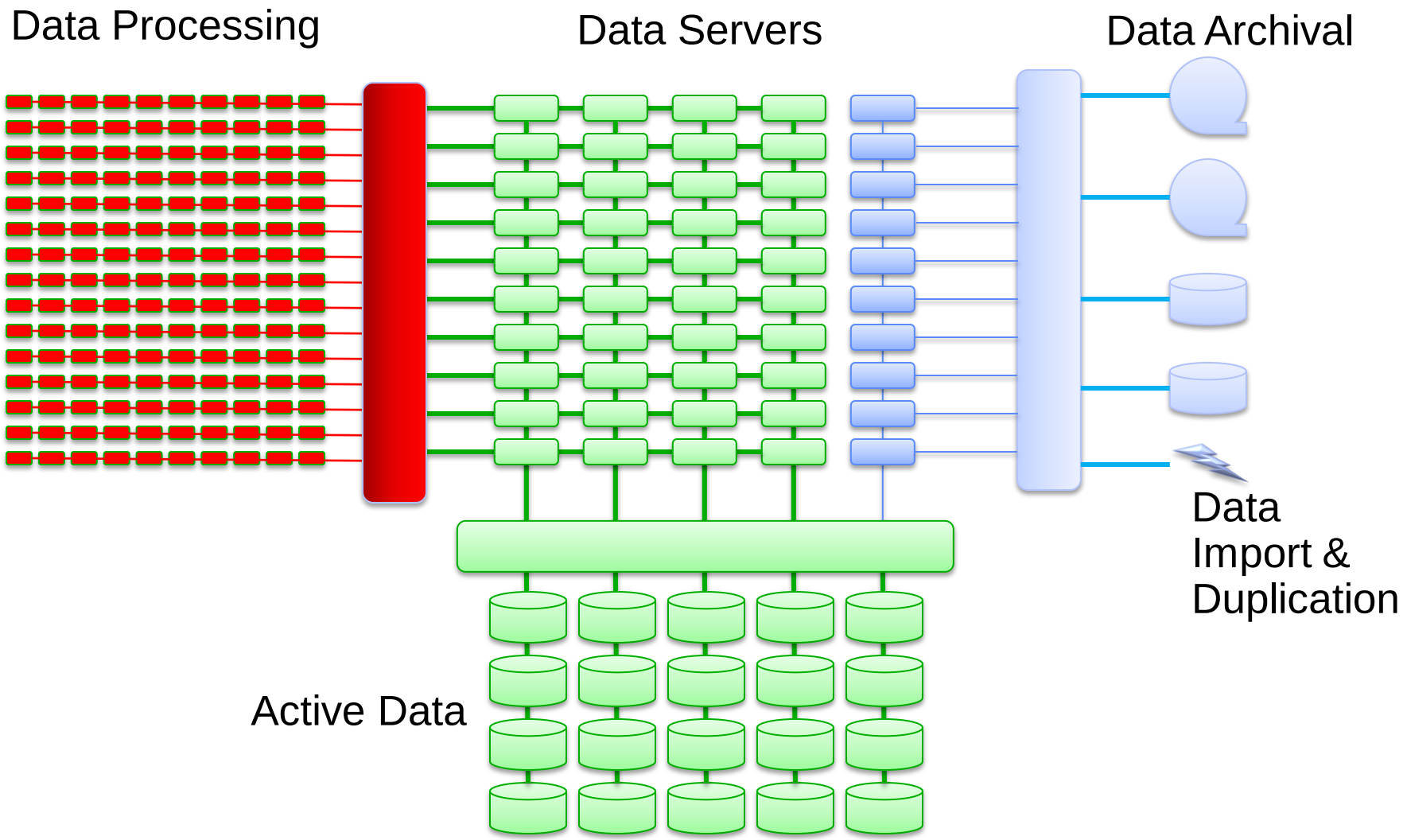


- Basic box: efficiency of all steps from procurement to operation
- Disk server on LAN
  - Rack-mountable 4U box
  - 1 mother board with 2 CPUs, 1 x 10 Gb Eth
  - No controller => direct attachment via SAS and sw RAID
  - 36 SATA disks of 3 TB configured as 17 file systems of 2 mirrored HDs Total usable capacity 34 TB, 1 GB/s
  - 6 file systems of 4+2 RAID6 HDs: 72 TB



- Storage array on SAN
  - Rack-mountable 4U box
  - 1 controller : 2 x FC4G, hw RAID
  - 24 SATA disks of 3 TB configured as 3 file systems of 6+2 RAID6 HDs Total usable capacity 54 TB, 800 MB/s

System Design (4)



R&D and Training

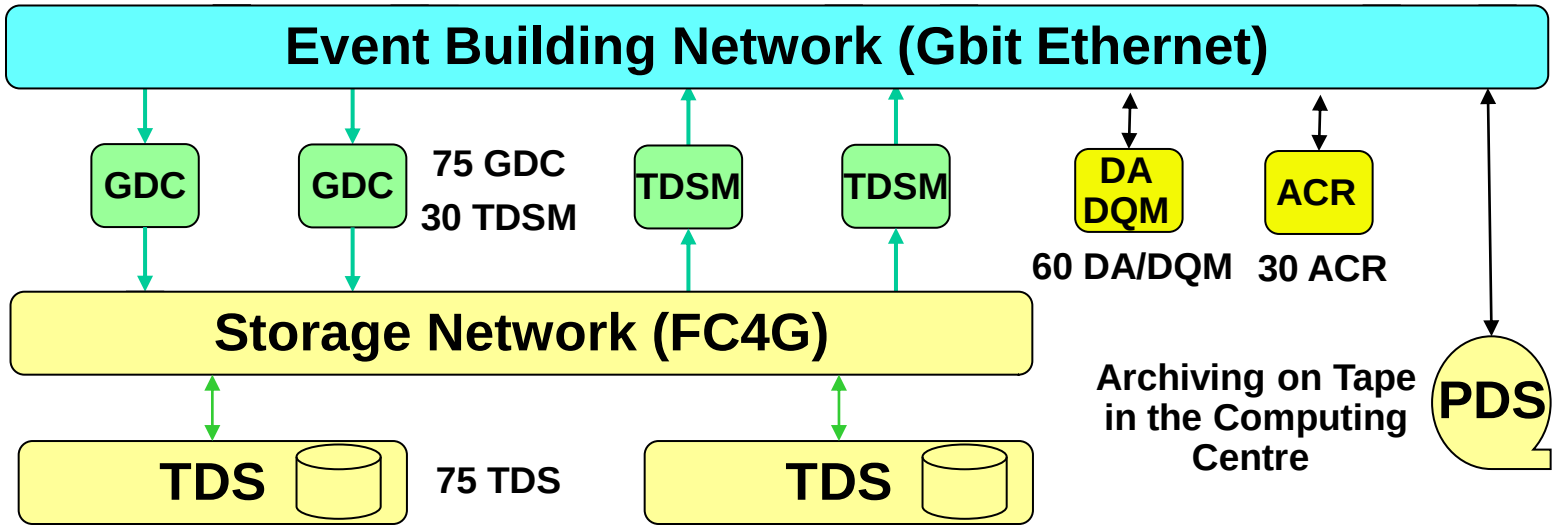
- A new Tier-1 also opens new opportunities
- R&D
  - Hardware
  - System and infrastructure software
  - Application software
- Training and education
  - Computing experts to design, build, and operate the Tier 1
  - Scientists to use it
- Cooperation and partnership
  - Academic institutions.
    Ready to join forces with you to support this project.
  - Commercial actors
- Manpower is essential

Case study: ALICE DAQ Storage

- Large project with a long duration: long term support, evolution and interoperability are key design factors
- Requirements: 100 TB disk buffer with an I/O bandwidth of 3 GB/s
- R&D and tests:
  - Conferences: technology in action, contacts with vendors (e.g. SNW)
  - Validation of technology in the lab (Fibre Channel FC2G)
  - Selection of hw components. Loan of components is essential !
- Architecture coherence
  - Selection of the cluster file system
- Staged Deployment and commissioning
  - Adoption of FC4G
  - Deployment in several stages (10, 20, 40, 100 % of the performance). The SAN has nicely followed this evolution
- Be ready for the unforeseen
  - Data quality monitoring: larger processing power than anticipated together with a higher throughput data access
  - Use of IP-based clients of the cluster file system

ALICE Data Storage Architecture

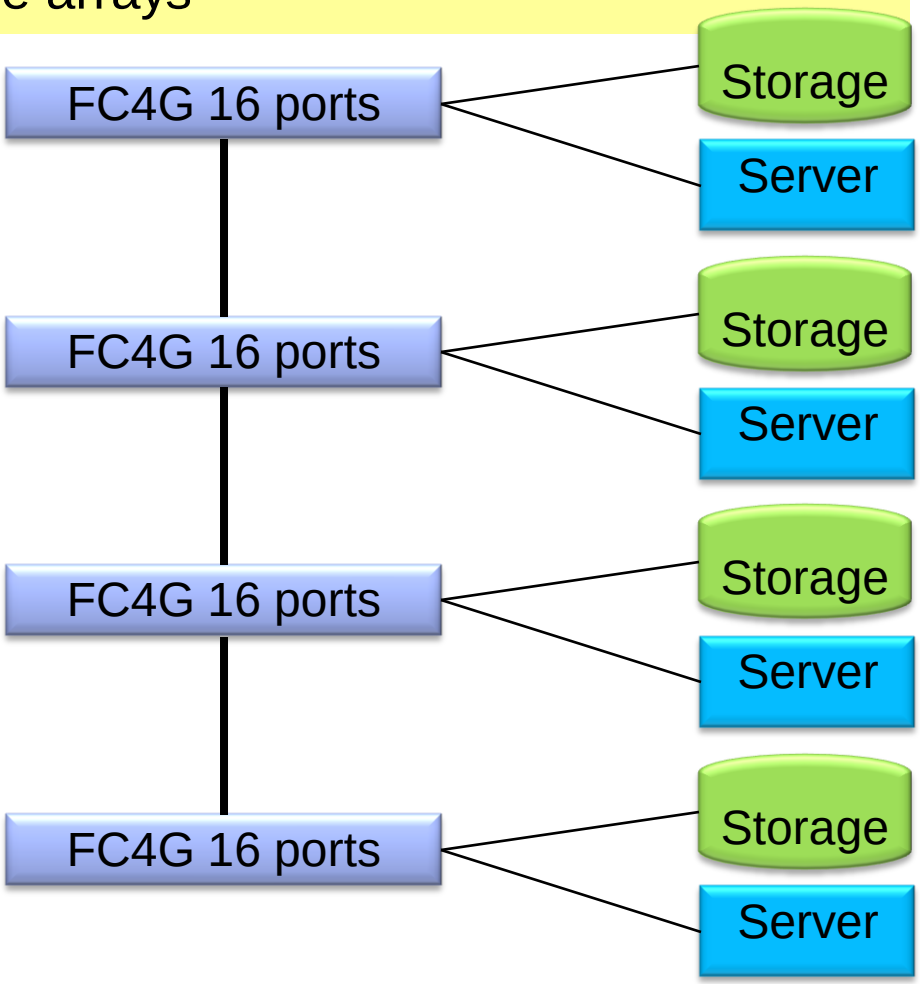
- 180 ports FC4G
- 75 Transient Data Storage (TDS) storage arrays each divided in 3 RAID6 volumes
- Nodes accessing data over FC4G:
  - 75 data formatting and writing-only nodes (GDC)
  - 30 reading-only nodes (TDSM) exporting data files to Permanent Data Storage (PDS)
- Nodes accessing data over IP
  - 90 nodes for data quality monitoring
- Unique logical view by cluster file sharing!



SAN Evolution over Time

2003-2008 (hw deployment 10-40% performance):

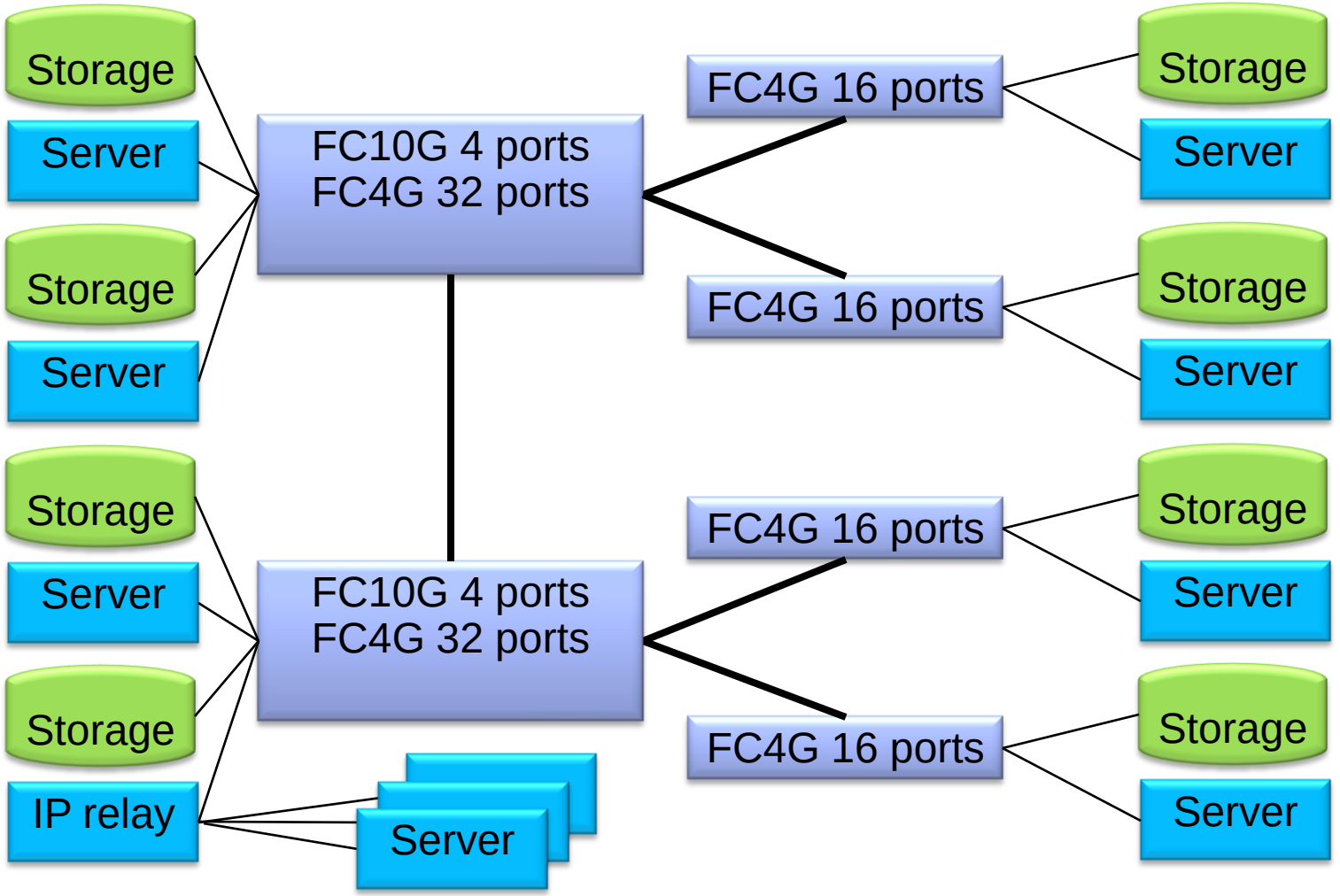
- 1-4 FC switches (16 FC4G, 4 FC10G)
- Up to 30 nodes
- Up 30 storage arrays



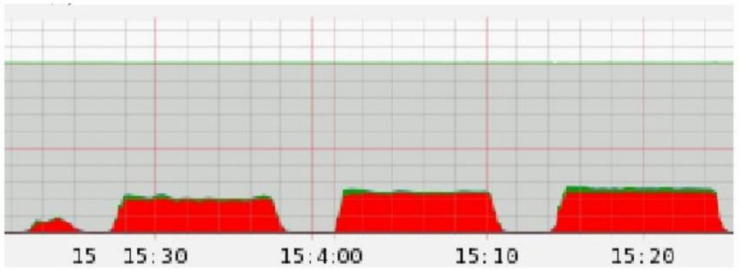
SAN Evolution over Time

2009: (hw deployment 100% performance)

- 2 enterprises FC switches (8 slots, 16 FC4G, 4 FC10G)
- 100+ nodes over FC4G, 90 nodes over IP
- 75 storage arrays



Cluster File System Performance

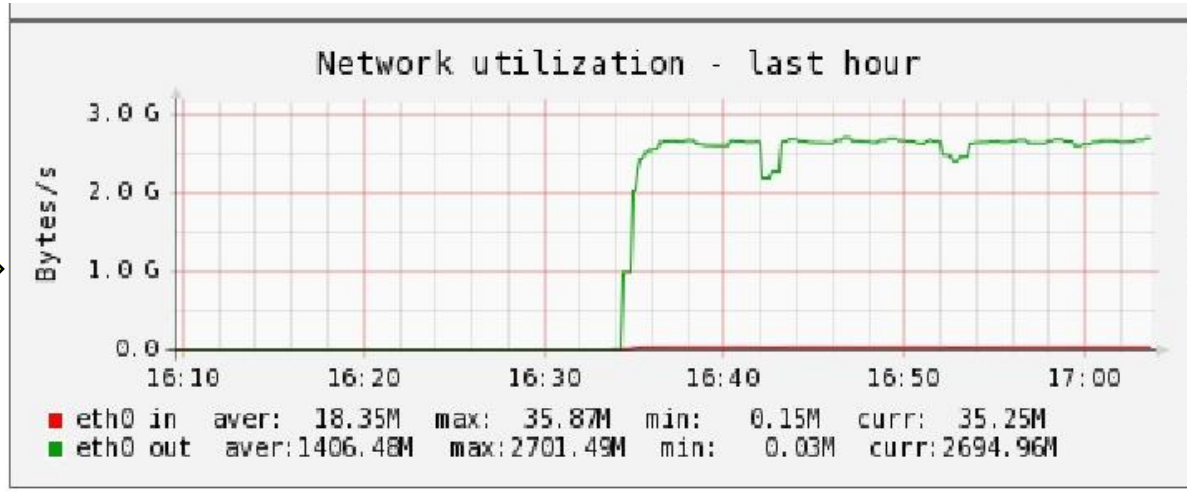


StorNext®

Writing to file system
Up to 4.5 GB/s

Reading from file system
Up to 2.5 GB/s

Aggregate data traffic: 7 GB/s



Trends and buzzwords

- Data deduplication
  - Target: e.g. the same mail with attachment sent to many people
  - HEP: explicit duplication over the GRID
- Virtualization
  - Go away from the concept of my local C: disk
  - HEP: going to this direction with the GRID, Cloud etc
- Storage clouds
  - Combination of commodity hardware infrastructure and value-added software is the best way to deliver a good service

Conclusion

- HEP produces lots of data and requires a solid storage system
- The technology is available to build it
- Define criteria to select technology & equipment. Test, demonstrate and deploy
- Keep a critical eye: all technologies want to claim primacy and be essential
- Designing a new Tier 1 is also a fantastic opportunity to build new partnerships and initiate new R&D and training activities
- Ready to collaborate

References

- *Early history & a 50 years perspective on magnetic disk storage,* A. S. Hoagland, Magnetic Disk Heritage Center, 2005.
- *CERN Storage Update*, Dirk Duellmann, Pres. HEPiX Fall 2008, Taipei, Taiwan, Oct. 2008.
- *Storage Management*, Ian Fisk, Pres. CHEP'10, Taipei, Taiwan, Oct. 2010.
- *A Comparison of Data-Access Platforms for BaBar and ALICE analysis Computing Model at the Italian Tier1*, A. Fella et al., Proc. CHEP'09, Prague, Czech Republic, Mar. 2009.
- *More than an interface — SCSI vs. ATA*, D. Anderson et al., Proc. 2nd Annual Conf. on File and Storage Technology (FAST), Mar. 2003.
- *Deep Dive on Solid State Storage - The Technologies & Architectures*, D. Martin, Pres. SNW Fall 2010, Dallas, USA, Oct 2010.
- *Intel X25 SSD Performance test*, xssist.com, Mar. 2010.
- *Tape Operations*, Vladimír Bahyl, Pres. CASTOR Review, CERN, Sep. 2010.

Thank You !

Questions ?