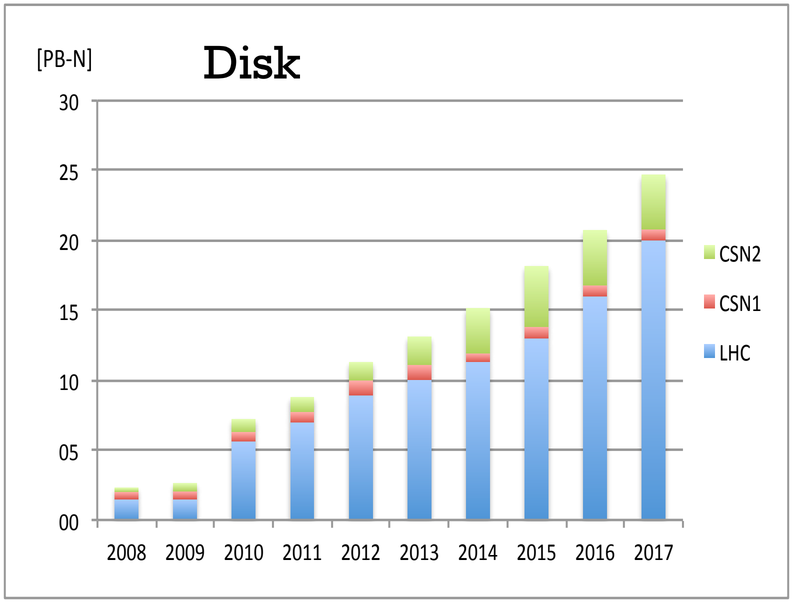
Storage evolution at CNAF

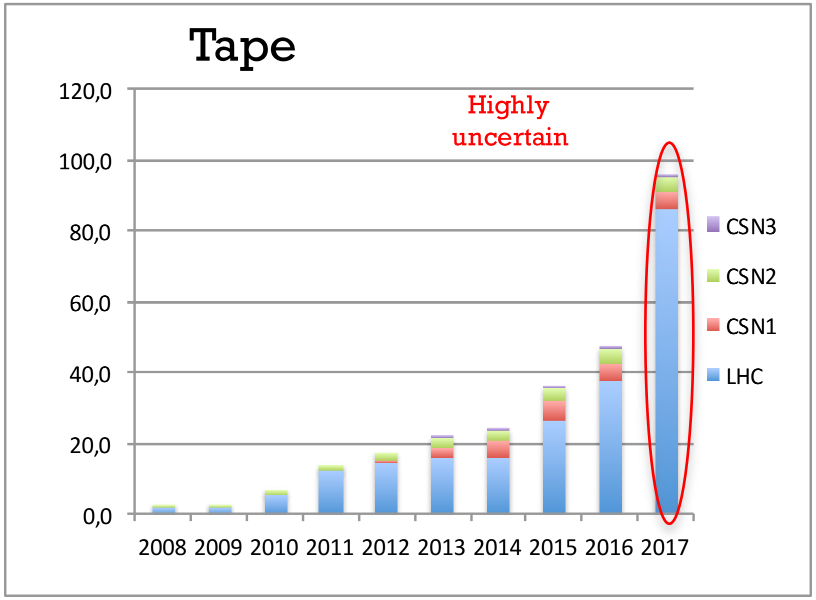
Vladimir Sapunenko,
CNAF

Introduction

- INFN Tier-1 is not only WLCG Tier-1
  - We host CPU and storage resources, both disk (29%) and tape (22%) for more than 25 experiments beside WLCG ones
  - Mainly Astro-particle collaborations
- Non definitive figures for LHC Run2 timescale
  - Taking into account not only LHC
  - Even more uncertain scenario for 2020+
- Change of mission for a Tier-1 center?
  - E.g. More focus on DM and fewer importance for computing?
- Does our current storage solution (GEMSS) fit with any requirements?
  - Several computing models (if any!) to cope with
    - Different storage usage, data access pattern, protocols….
  - Other constrains (budget, space, ….) can also condition the evolution of our storage system

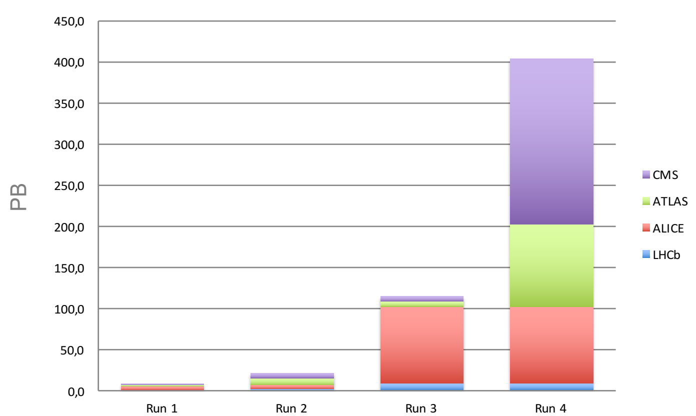
On-line and Near-line Storage grow trends at CNAF
(LHC Run2 time-scale)



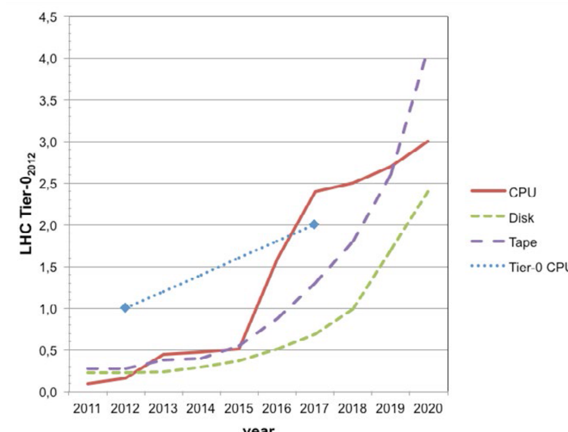


Long(er) term

- Difficult to extrapolate figures for CNAF
- Anyway, huge increase of resources foreseen and our Data Center will be unlikely able to support it (budget issues not considered)
  - Remote extension could have effect on storage model
  - First experience to be gained with HNSciCloud, Bari and Aruba



WLCG raw data volumes estimations per year of data taking (to be added: ESD, AOD)

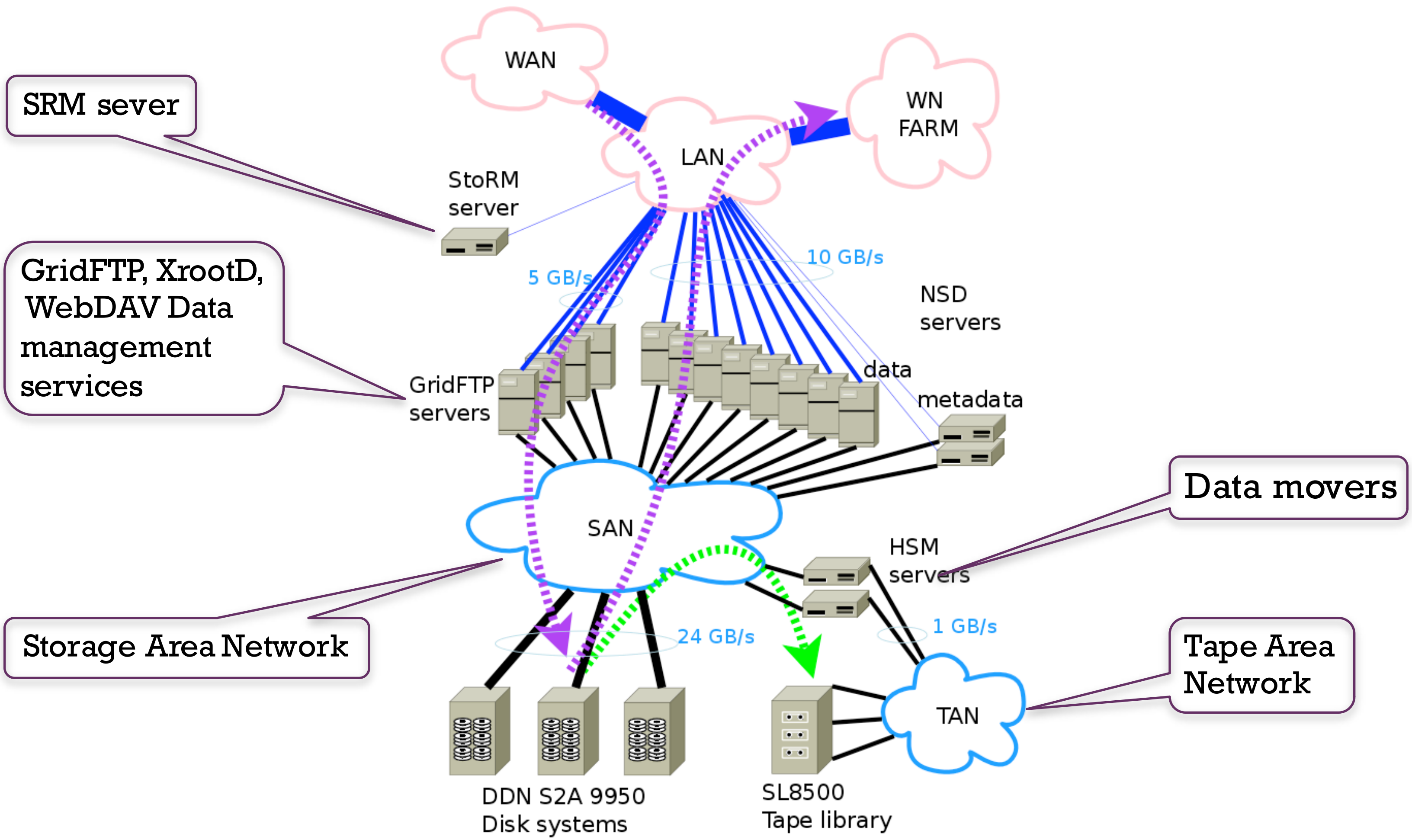


APPEC resource usage estimations compared to WLCG Tier-0

Operational conditions

- 4 LHC and more than 25 other HEP experiments
- ~18 PB of data online and 22 PB near-line (tapes)
- Accessed from ~15K concurrent processes
- Aggregated data bandwidth to storage ~ 90 GB/s
  - Actually observed:
    - on LAN ~ 20 GB/s (16 GB/s from 1 single experiment)
    - and WAN, ~ 2 GB/s (saturating 2x10Gbit uplinks)
- Continues configuration changes (new installations, data migrations)

Data flow in a single experiment cluster

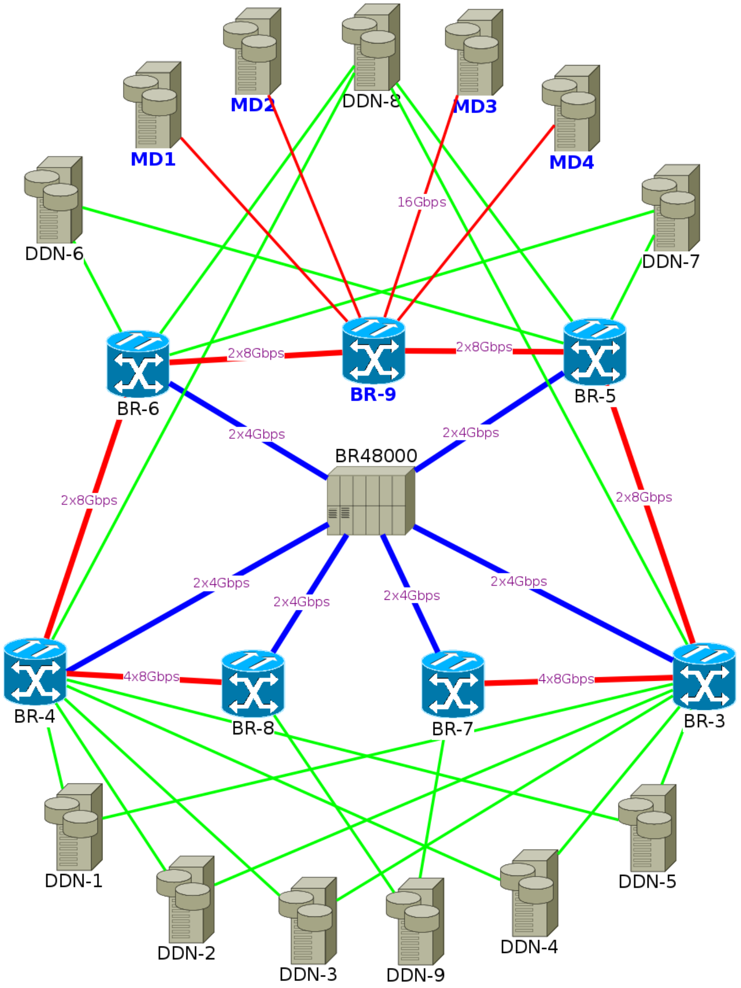


Data Center Storage model (current state)

- Fiber Channel SAN (and IB FDR is under deployment)
- Few but Big storage systems O(PB)
- Dedicated 10 Gbps I/O servers for big Exps
  - Moving to 4x10 Gbps servers
- Dedicated 10 Gbps GridFTP servers for big Exps
- Dedicated HSM nodes for big Exps
- Direct access via SAN from GridFTP, XrootD, WebDAV, servers to the storage
- Direct access via SAN from HSM nodes to the storage
- Targeted for high performance

Storage Area Network

- Single fabric
  - 1 director (core switch)
  - 7 edge switches
- Different technologies:
  - Core - FC4
  - Edge - FC8
  - Latest – FC16
  - *Soon - IB FDR*
- Total number FC of ports: 1360
- Dual link from every HBA to SAN disks (via separate edge switches)
- All tape drives connected to the core



Data Center Storage model (current state 2)

- GPFS as POSIX interface and back-end for all data management services provides
  - Flexibility in management
  - Performance
  - Failure resilience
- Dedicated clusters for big experiments
  - Management and Failure domain isolation
- Using dedicated disks (SAS) and servers to store and handle file system metadata
  - Data (I/O) servers can perform metadata handling in case of metadata servers failure

GPFS = Software Defined Storage

"Why you are using GPFS? It's boring, it just works…"

- GPFS is actively evolving
  - 3 Major releases in last three years
  - Incorporated New architectural approaches
    - Hadoop-like: SNC (Shared Nothing Cluster)
    - RAIN-like: Native RAID (usnig JBODs)
    - Geographically distributed: AFM (with local cache, AFS-like)
    - Local read-only cache on clients
    - Integration with OpenStack
- IBM GPFS and TSM operational costs
  - TSM ~50K€/year (including CNAF backup service)
  - GPFS ~50K€/year (for ~8 INFN sites)

Disk Performance offered and demanded

Exp	On-line usable (disk) storage, TB	Number of I/O servers (EA, Gbps)	TB(net)/ I/O server	Front-end (LAN) Bandwidth/TB, MB/s	Back-end (Storage) bandwidth /TB, MB/s	Max sustained bandwdth used/TB, MB/s
ATLAS	3500	8 (x10)	437	2.85	**4.7**	**2.0**
ALICE	1730	6 (x10)	288	4.33	**3.1**	**2.9**
CMS	3380	16 (x10)	211	5.91	**4.7**	**4.7**
LHCb	2520	12 (x10)	210	5.95	**4.7**	**2.0**
AMS	1540	8 (x10)	192	6.49	**6.8**	**6.8**
GR2	1250	6 (x10)	208	6.0	**4.4**	**1.3**
Virgo	428	16 (x1)	26	4.6	**6.2**	**1.8**
ARGO	320	12 (x1)	26	4.6	**6.2**	**2.5**

- TB (usable) per I/O server (LAN, WAN)
- LAN bandwidth (MB/s per TB)
- Storage bandwidth (MB/s per TB)

Technology caveats

- Disk capacity increasing much faster than performance
  - Sequential access rate is about 150MB/s for 4TB SATA disks
    - real sustained rate even lower (30-60MB/s)
  - Rebuild times for 4TB disks is about 50 hours
- More space per spindle + more CPU cores → IO congestion
  - to keep up with performance demand we need to deploy faster Disk Tier or Cache
    - Preliminary tests with SSD array demonstrated great improvements

Disk storage HW evolution (as we see it)

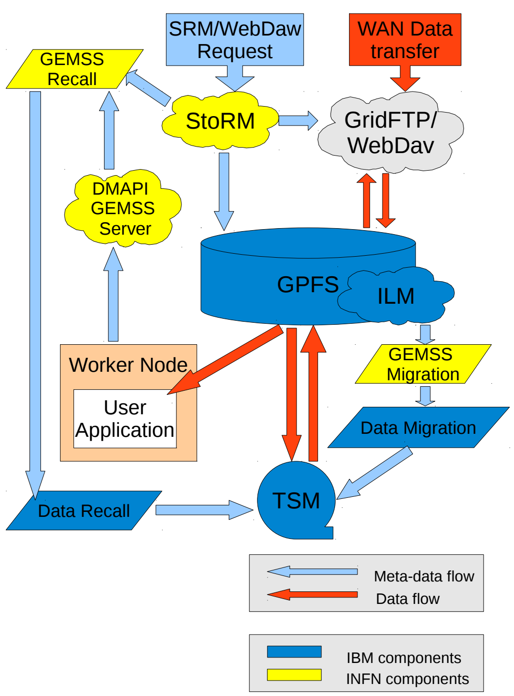
- There are no alternatives to enterprise-graded HDDs;
- 8 TB He-filled drives are being installed, expected higher performance in streaming I/O, up to 200 MB/s (as on data sheet) 10 TB He-filled disks already available in the market;
- No strong objections to use "small bricks", BUT
  - only data replication can provide acceptable level of protection from entire system failure
- Advantages of enterprise-graded storage systems:
  - lower efforts in management;
  - better support and problem resolution;
  - lower chance of entire system failure;

Disk Storage software evolution

- use of POSIX FS (as frequently requested and preferred by users);
- distributed RAIDs to minimize recovery time from hdd failure or use of mirroring (RAID10) with "Archival" hdd - continuous availability + serviceability
- use of Parallel FS to provide requested bandwidth;
- use of tiered storage requiring deployment of not negligible amount of expensive High Performance disks and implementation of HSM movers between slow and fast disks
  - can be done with TSM or directly with GPFS) and "pre-staging" of data to be processed ("transparent recall" will not work for I/O intensive jobs

Mass Storage System

- HSM: GEMSS
  - Integration of IBM GPFS and TSM + specific customization and the SRM interface StoRM
  - Very good performance and efficiency



- Disk-centric system with five building blocks
  1. **GPFS**: disk-storage software infrastructure
  2. **TSM**: tape management system
  3. **StoRM**: SRM service
  4. **TSM-GPFS** interface
  5. **Globus GridFTP**: WAN data transfers
- SRM is not essential – currently used only to "BringOnLine", could be replaced by direct WebDav/HTTP calls
- DMAPI Server Used to intercept READ events via GPFS DMAPI and re-order recalls according to the files position on tape
- Globus GridFTP or Storm WebDav service used for WAN data transfers

Near-line Storage evolution

- There is no any Open Source High Performance Storage solutions with HSM even on a horizon (apart from dCache)
- HPSS is targeted for performance, very expensive and VERY complicated in use
- GEMSS seems to be optimal solution as for expenses and for management efforts

Exploring new technologies: Dynamic Disk Pools (Dell)

- Recent storage systems from Dell
- Distributed RAID 8+2
- Pros:
  - Fast recovery (15 min respect to 50 hours)
- Cont:
  - Slightly lower performance (more computations needed)

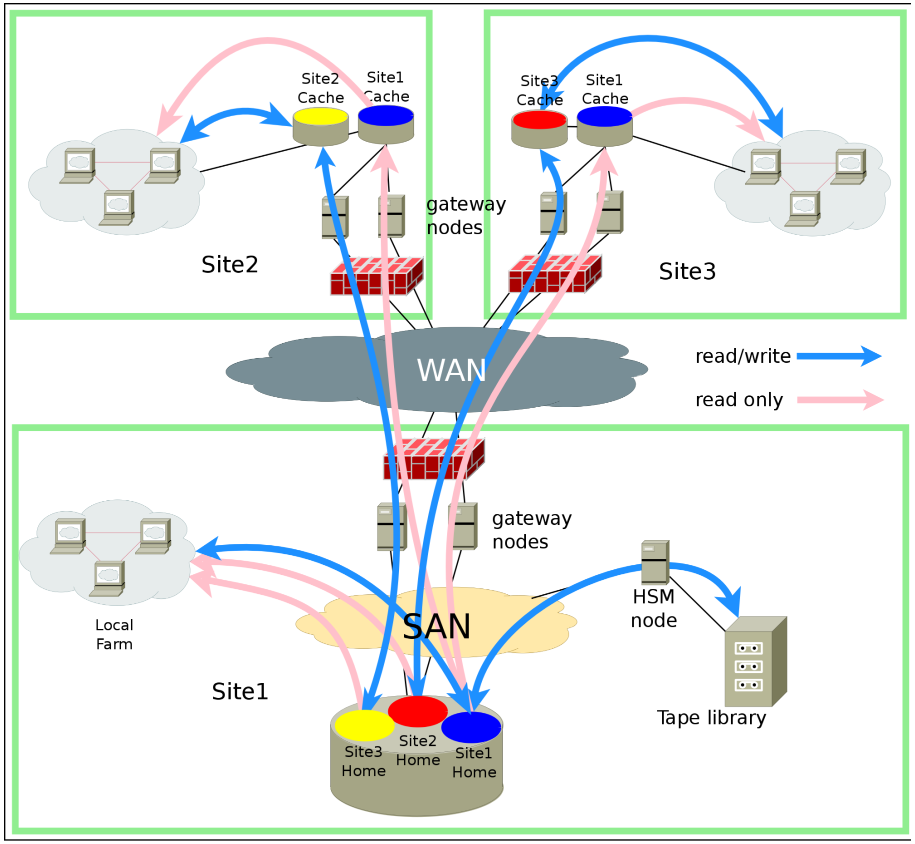
RAID configuration	RAID6 (8+2)	DDP (1x180)
N. of pools	18	1
N. disks per pool	10	180
% used for parity	20	20
N. of reserved capacity disks	0	6
N. of LUNs	18	18
Usable space, TB	576	556
Critical conditions (N. Of failed disks)	2 in 1 pool	2 in 1 pool
Recovery Time from critical, hours	50 (rebuild of 1 disk)	<0.3 (estimate)

Remote data access via GPFS AFM

Cache basics

- Asynchronous updates
- Writes can continue when the WAN is unavailable
- TCP/IP for communication between sites (NFS or GPFS protocol)

- Two sides
  - Home - where the information lives
  - Cache
    - Data written to the cache is copied back to home as quickly as possible
    - Data is copied to the cache when requested
- Communication is done using NFS (v3 and v4)
- GPFS has it's own NFSv3 client
  - Automatic recovery in case of a communication failure
  - Parallel data transfers (even for a single file)
  - Transfers extended attributes and ACL's



Use case of AMS: CNAF(Bologna)-ASI(Rome) remote data processing…

- Home site location: Bologna
- Remote site location: Rome
- Distance between sites: ~400km
- RTT: 23 ms
- Bandwidth: 100 Mbps
- Home FS size: 1.1 PB
- Cache size: 10 TB

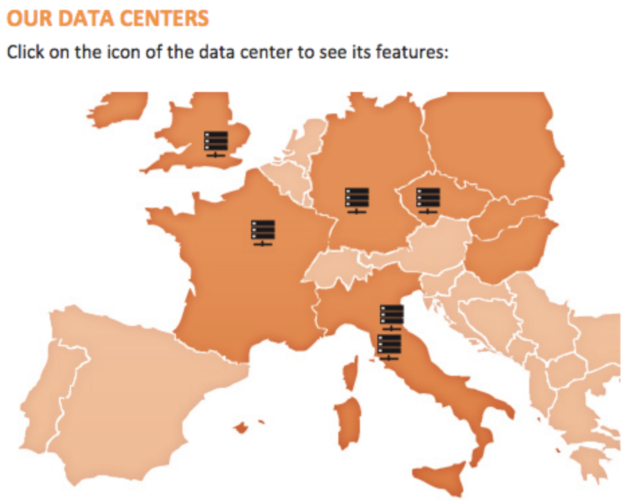
- A DB (based on ROOT TTree objects) with tags of events that have passed certain preselection requirements has been locally created.
- Each data processing job queries the preselection DB to look for the tags of interesting events, in order to access them (and only them) from a remote file.
- AFM Prefetch Threshold has been tuned to manage 10 GB files accessed randomly and sequentially.
- The final configuration allows us to process the same file remotely paying only a fraction of 15% in execution time.

… and Bari/RECAS

- Tests on going for remote extension of Tier-1 in Bari/RECAS
- 20 kHS06 available
- VPN (20 Gbps) configured
- AFM cache set-up completed
  - ~200 TB and 2 disk servers, 20 Gbps interconnection
- Transparent access to tape from Bari needs to be understood
  - First step: data on tape to be accessed directly from CNAF

The Aruba case

- Aruba is one of the biggest commercial cloud provider in Italy
- No cache present (yet) in Aruba excepting fo: LSF and Exp software
- Direct remote access to storage via xrootd
  - Viable only for few experiments (e.g. CMS, Alice…)
- Stageout to CNAF/Storm



Exoteric (non WLCG?) use cases

- http/WebDAV implemented as side service of Storm, can be used as independent service
- Easy to use interface, i.e. dropbox like or T1_as_a_usb_disk_attached_to_my_laptop, is a common request from smaller VOs.
  - We are experimenting with OwnCloud and GlobusOnline but if storm will be able to provide this kind of access if would be a great added value, maybe integrating OwnCloud itself.

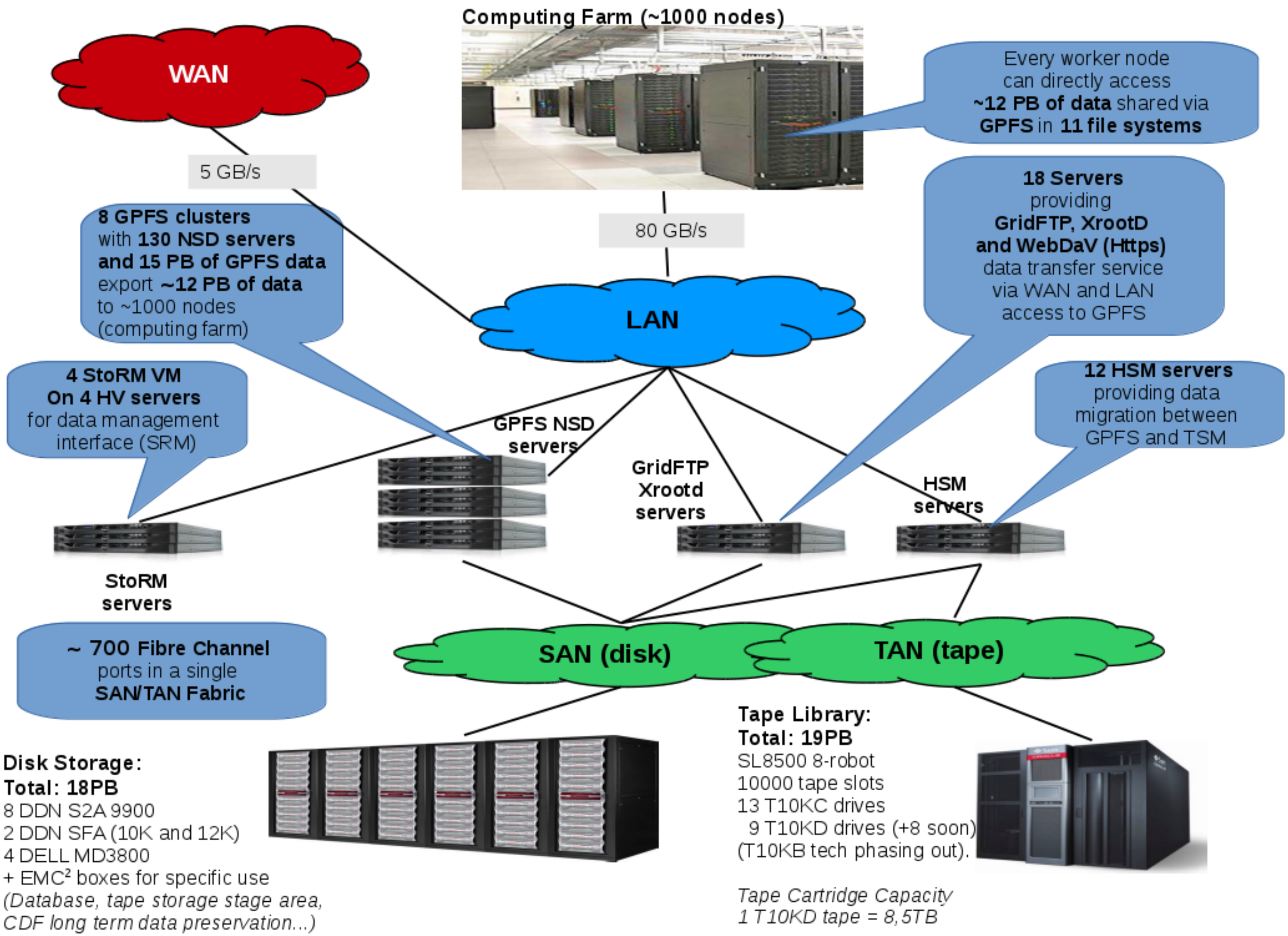
Conclusions

- SAN based solutions + clustered file system are still providing better performance and availability at lower costs
- DAS based solutions (EOS, dCache, Gluster) to ensure data availability still require data replication → doubling number of servers, raw storage space, footprint and power consumption
- Implementation of Erasure coding in RAIN system could overturn this situation
- Well defined metrics will help to make the choice

Backup slides

Metrics to confront PB-range storage solutions

- Capacity to Bandwidth ratio
  - TB (usable) per I/O server (LAN, WAN)
  - LAN bandwidth (MB/s per TB)
  - Storage bandwidth (MB/s per TB)
- Building block size
  - capacity, footprint (rack units), network ports
- Power consumption
  - KW/TB (including all components: disks, servers, network)
- Price (TCO)
  - KEuro/TB (including all components: disks, controllers, servers, network, software, man power)



Computing Farm (~1000 nodes)
WAN
5 GB/s
80 GB/s
LAN
Every worker node can directly access **~12 PB of data** shared via **GPFS** in **11 file systems**
18 Servers providing **GridFTP, XrootD and WebDaV (Https)** data transfer service via WAN and LAN access to GPFS
8 GPFS clusters with **130 NSD servers and 15 PB of GPFS data** export **~12 PB of data** to ~1000 nodes (computing farm)
4 StoRM VM On 4 HV servers for data management interface (SRM)
12 HSM servers providing data migration between GPFS and TSM
GPFS NSD servers
GridFTP Xrootd servers
HSM servers
StoRM servers
~ 700 Fibre Channel ports in a single **SAN/TAN Fabric**
SAN (disk)
TAN (tape)
Disk Storage:
Total: 18PB
8 DDN S2A 9900
2 DDN SFA (10K and 12K)
4 DELL MD3800
+ EMC² boxes for specific use
(Database, tape storage stage area, CDF long term data preservation...)
Tape Library:
Total: 19PB
SL8500 8-robot
10000 tape slots
13 T10KC drives
9 T10KD drives (+8 soon)
(T10KB tech phasing out).
Tape Cartridge Capacity
1 T10KD tape = 8,5TB

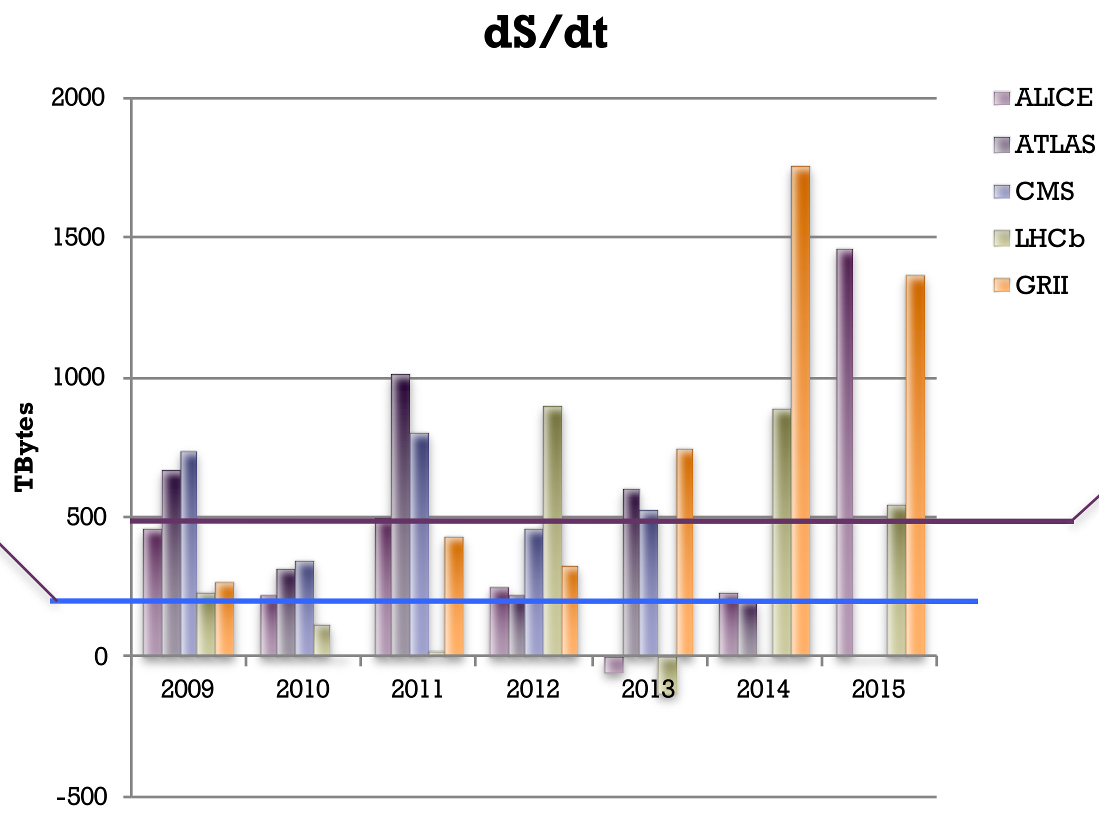
Flash back (to 2013 review)

- CERN EOS in 2013:
  - The concept: use of single disks (JBODs) without local RAIDs
  - Reed-Solomon error correction is ready and will be available in next release (in a month time) - still is not used in prod at CERN!
- in 2015 we have Reed-Solomon (erasure) coding in IBM's GSS (GPFS based) systems and in Dell MD3800 (last storage acquisition)
  - Dell MD3800 in production for 6 months
  - Working with IBM to verify GSS compatibility with our environment

MSS: Hardware

- TSM server – core of MSS system
  - Current version does not support redundant server configuration
    - Using "warm" spare server with shared storage
  - Observed HW limitation of current server during re-pack and data migration
    - Upgraded server HW to FC16 HBA pushing throughput to 1.6 GB/s
- Tape libray: Oracle-StorageTek SL8500
  - 21 PB total space used
  - 10000 slots
    - 4514 T10000D tapes used, 1622 free
  - 17 tape drives
  - Max capacity with tapes "D" ~8,5 PB
  - Expected demand by 2017 ~ 100 PB
    - Considering installation of second library.

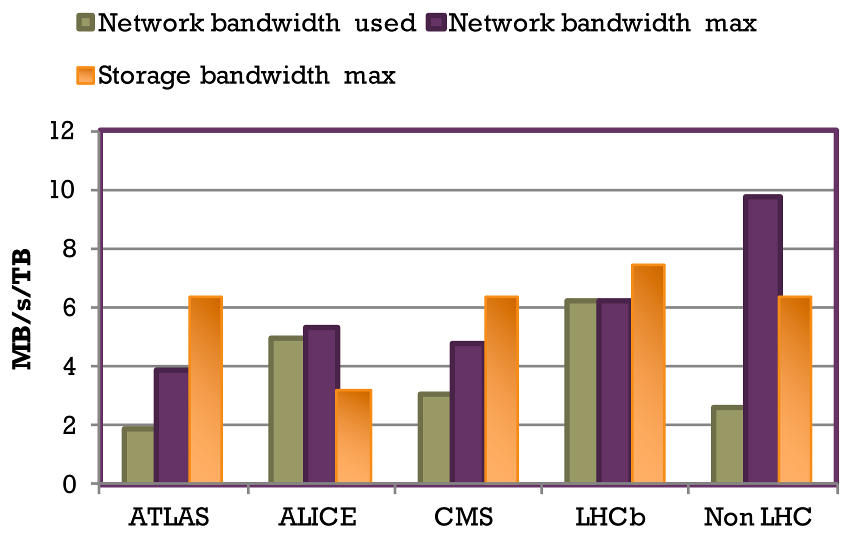
Delta in storage space allocated respect to a previous year



Optimal building block size

Current building block size

INFN Tier1

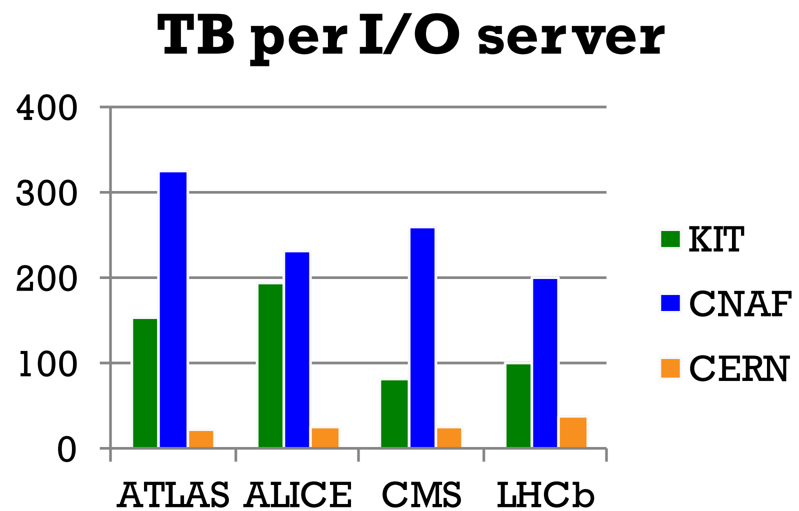


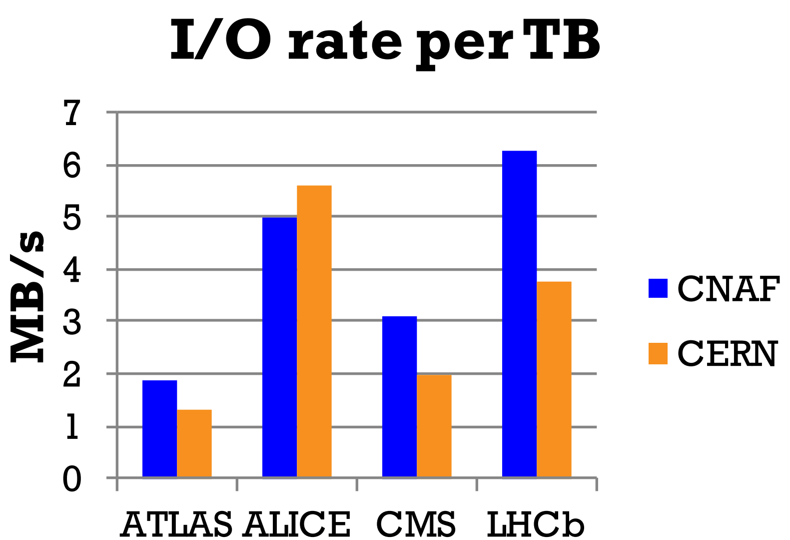
Network and storage bandwidth available and used by experiments for every TB of storage.

2013 Tender: 1.9 PB (usable) for ~306 €/TB
(included servers and FC switches, excluded VAT)

Confronting with other sites

- TB (usable) per server for CNAF, KIT and CERN (EOS only)
  - CERN: Raw/usable=2
  - KIT, CNAF: raw/usable=1.25





- I/O rate on LAN in MB/s normalized to storage volume (in TB), max sustained
  - Data from LeMon (CERN, CNAF), X. Mol talk at GDB 13/03/13 (KIT)

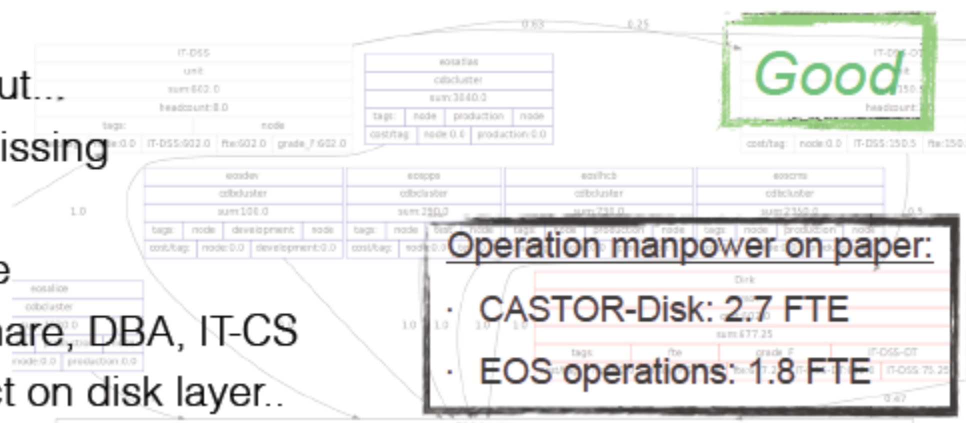
Storage operations costs at CERN

assumptions on:
- prices for HW @ 3years
- electricity cost
- disk operation manpower

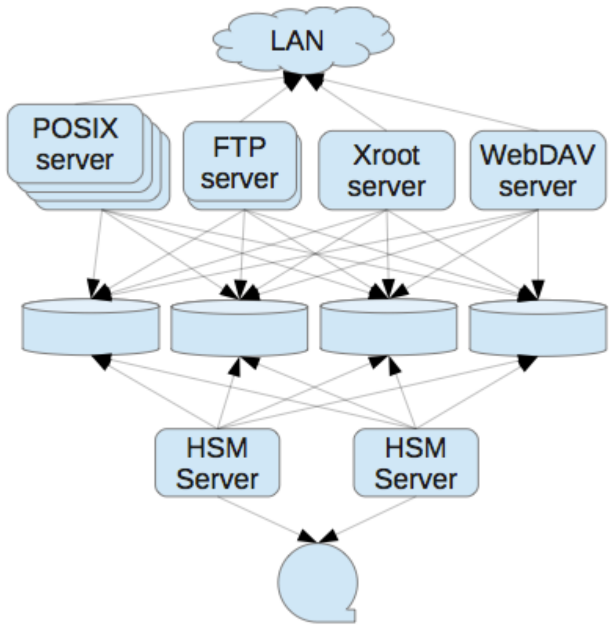
	CASTOR (CERN Advanced STORage manager)	EOS
HW+ electricity cost	16.5 CHF/1TBMonth	13.0 CHF/1TBMonth
operation manpower cost	2.7 CHF/1TBMonth	1.3 CHF/1TBMonth
partial "running" cost	**19.2** CHF/1TBMonth	**14.3** CHF/1TBMonth

Amazon S3: "reduced redundancy", Europe, 30PB: **42**US$ / 1TBmonth (no Network, I/O ops)

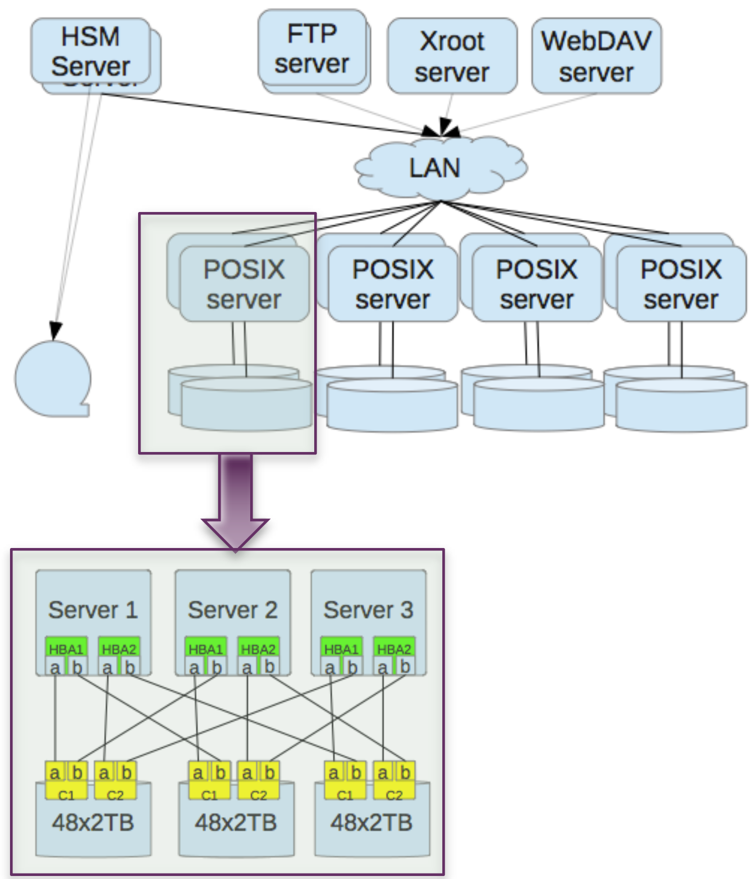
- Doing OK cost-wise, but...
  - Some manpower missing
    - development
    - sysadmins share
  - ORACLE license share, DBA, IT-CS
  - CASTOR tape effect on disk layer..



SAN vs. DAS



LAN
POSIX server
FTP server
Xroot server
WebDAV server
HSM Server
HSM Server



HSM Server
FTP server
Xroot server
WebDAV server
LAN
POSIX server
POSIX server
POSIX server
POSIX server
Server 1
Server 2
Server 3
HBA1
HBA2
a
b
C1
C2
48x2TB
48x2TB
48x2TB

Advantages of SAN solutions

- Redundancy
  - With a Shared disk file system more than one server can access one storage device → protection against server failure, possibility to take server off-line for maintenance without compromising access to data
- Scalability
  - Adding more storage to a server does not require HW modification on server side
- Dedicated network for server-storage communication
  - Servers with different roles (I/O servers, data movers, HSM nodes) can work independently
- Centralized management

Drawbacks of SAN solutions

- Scalability
  - Building blocks are of order of PB is huge respect to requested (yearly) increment for single experiments
  - If fully loaded, expansion of a few % require big expenses
  - If not fully loaded, being highly optimized, expansion is very challenging
  - Performance problem in a single component can affects the whole system ("slow disk" problem)
- Flexibility
  - Considerable efforts to preserve performance after small configuration changes
  - To move some TB from one exp to another require some intensive data re-balancing which can affect performance of both systems

General architecture view