INFN-T1 storage infrastructure

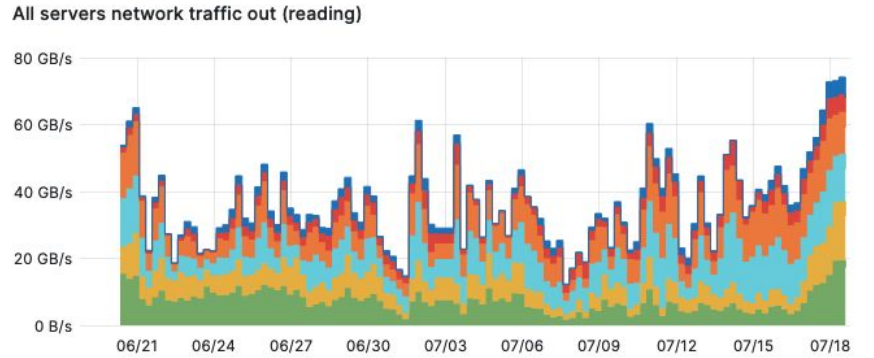
Vladimir Sapunenko

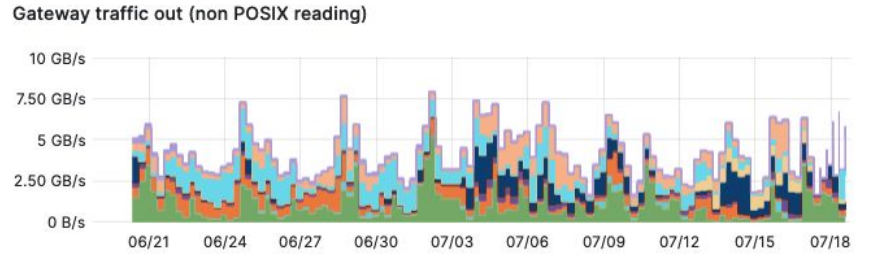
CNAF overview

- Storage provided by CNAF (total)
  - 70 PB usable on disks, 80% overall efficiency compared to raw disk space
  - 130 PB on tapes
- CNAF provide
  - Tier-1 services for all 4 LHC experiments
  - Tier-0/1 services for non-LHC experiments
- General data and computing services for ~40 HEP and Astrophysics experiments

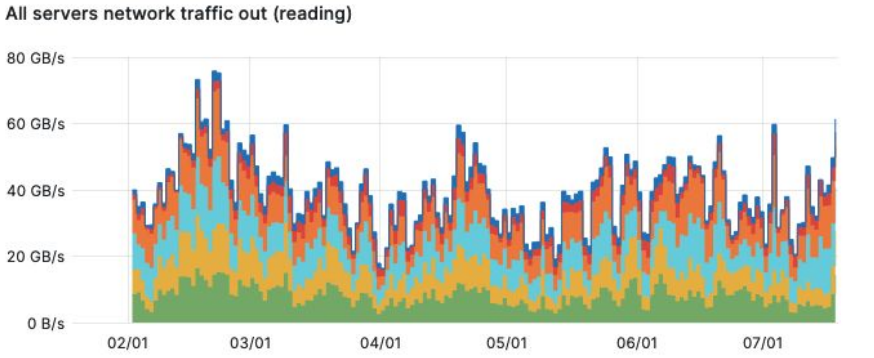
Some performance numbers

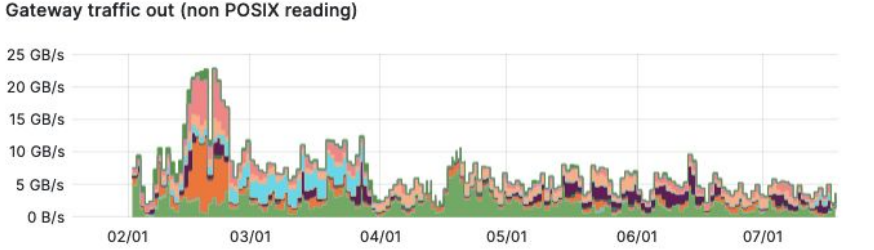
Last month





Last 6 months





Disk storage in prod

Installed: 113PB - 33PB (in dismissione)=80.6PB Pledge 2024: 82.08PB, Used: 48.8PB

Storage system	Model	Net capacity, TB	Experiment	End of support
ddn-10, ddn-11	DDN SFA12k	10120	ALICE, AMS	12/2022
os6k8	Huawei OS6800v3	3400	GR2, Virgo	07/2024
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	12/2024
md-5, md-6 e md-7	Dell MD3820f	50	metadati, home, SW	11/2023 e 12/2024
os18k1, os18k2	Huawei OS18000v5	7800	LHCb	7/2024
os18k3, os18k5, os18k5	Huawei OS18000v5	11700	CMS	6/2024
ddn-12, ddn-13	DDN SFA 7990	5840	GR2,GR3	2025
ddn-14, ddn-15	DDN SFA 2000NV	24	metadati	2025
os5k8-1,os5k8-2	Huawei OS5800v5	8999	ATLAS	2027
Cluster CEPH	12xSupermicro SS6029	3400	ALICE, cloud, etc.	2027
od1k6-1,2,3,4,5,6	Huawei OD1600	60000	ALICE,ATLAS,LHCb, CMS	2031

Storage overview (I/O servers)

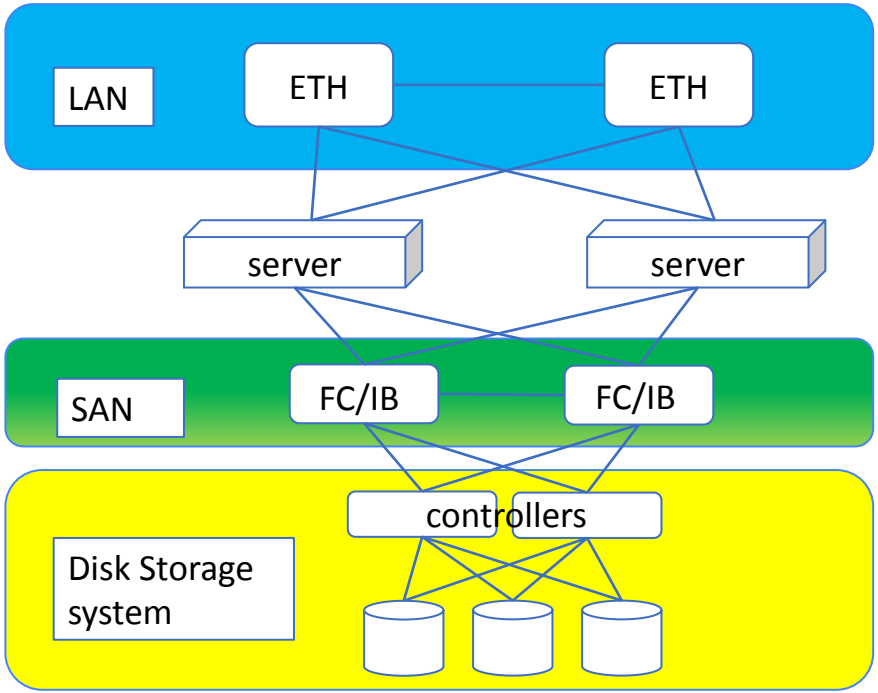
- 54 I/O servers in total (NSD, WebDAV, XrootD, HSM)
- Dual power supply
- RAM: 32 GB (2x10 GbE) - 128 GB (4x10 GbE) - 512 GB (2x100GbE)
- Local System disk protected by mirroring (RAID1)
- LAN: 2x**10**GbE - 2x**25**GbE 2x**100**GbE
  - All servers connected to both core switches (Ethernet bonding, active-active)
- SAN: All servers connected to two FC or IB switches (active-active)
- Every Storage device accessible from several servers (shared storage model)
  - Any server can be taken off-line at any time without compromising data access.

More recent Storage hardware

- Huawei OceanStore Micro 1500/1600 - 10PB of net space
  - 4 controllers x 4 FC32
  - 55 x 14TB disks in each Storage pool with Dynamic RAID6 protection
- 8 systems x 10PB
- 8 FC switches
- 40 server
  - 512GB of RAM
  - 4 x FC32 ports
  - 2 x 100 Gpbs ethernet
- IBM TS4300 Tape library
  - 18 x TS1170 tape drives
  - 50TB tape cartriges

Storage interconnect infrastructure

- Fiber Channel
  - FC16, FC32 (16-32 Gb/s)
  - I/O Server to storage
  - HSM nodes to storage
  - HSM nodes to tape drives
- InfiniBand
  - FDR (56 Gb/s) EDR(100Gbps)
  - I/O Server to DDN and Huawei storage
- Ethernet
  - Nx10 GbE, 2x25GbE or 2x100 GbE
  - I/O Server to LAN
  - Ethernet bonding (N=1-4)



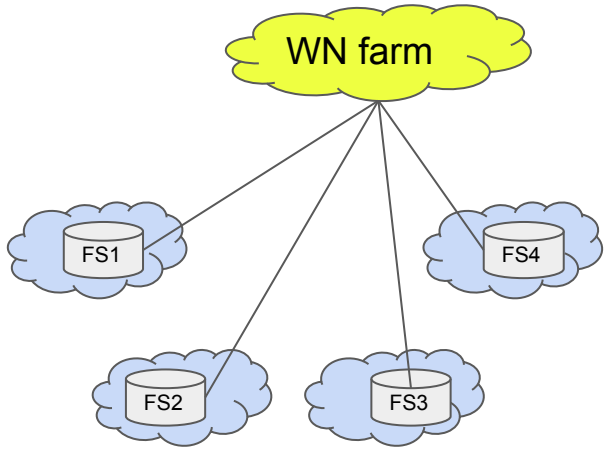
Storage Services

- Remote and local access
  - SRM (StoRM)
  - GridFTP
  - XrootD
  - WebDAV
- Local access only
  - POSIX via GPFS (~1000 clients, mainly WNs)

Software

- Proprietary
  - GPFS (IBM, site license)
  - TSM (IBM, per socket license)
- Open Source
  - GridFTP
  - XrootD
  - StoRM (locally developed)
  - GEMSS (locally developed)

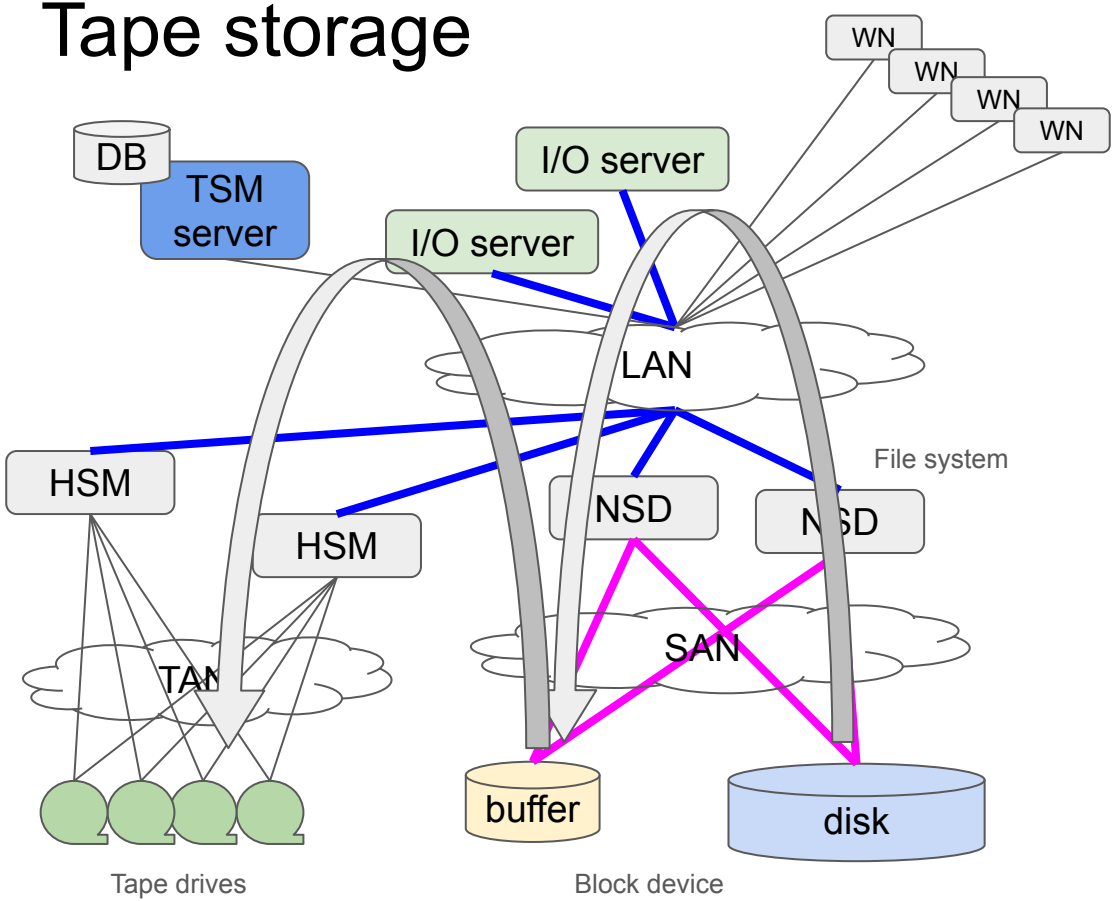
Multi-cluster architecture, remote FS mount



Storage HW overview (tape)

- Oracle(StorageTek) SL8500 library (being decommissioned)
  - 16 T10kD (scientific data) tape drives (250 MB/s I/O data rate)
  - 7 T10kC (backup and recovery service) tape drives
- IBM tape library TS4300
  - 19 TS1160 tape drives (20TB/tape, 400 MB/s I/O data rate)
- IBM tape library TS4300 (not yet in production)
  - 18 TS1170 tape drives (50TB/tape, 400 MB/s I/O data rate)
- TSM servers (2xFC16 to TAN)
  - Production: 1 active, 1 stand-by, 1 devel
  - Backup service: 1 active
- HSM servers (2xFC16 to TAN, FC or IB to disks) → (FC to tape, 2x25Gbps to disks)
  - Production: 5 active, 1 stand-by
  - Testbed: 2 active

Tape storage



WN - worker (compute) nodes

NSD - Network Shared Device

SAN - Storage Area Network (IB)

TAN - Tape Area Network (FC)

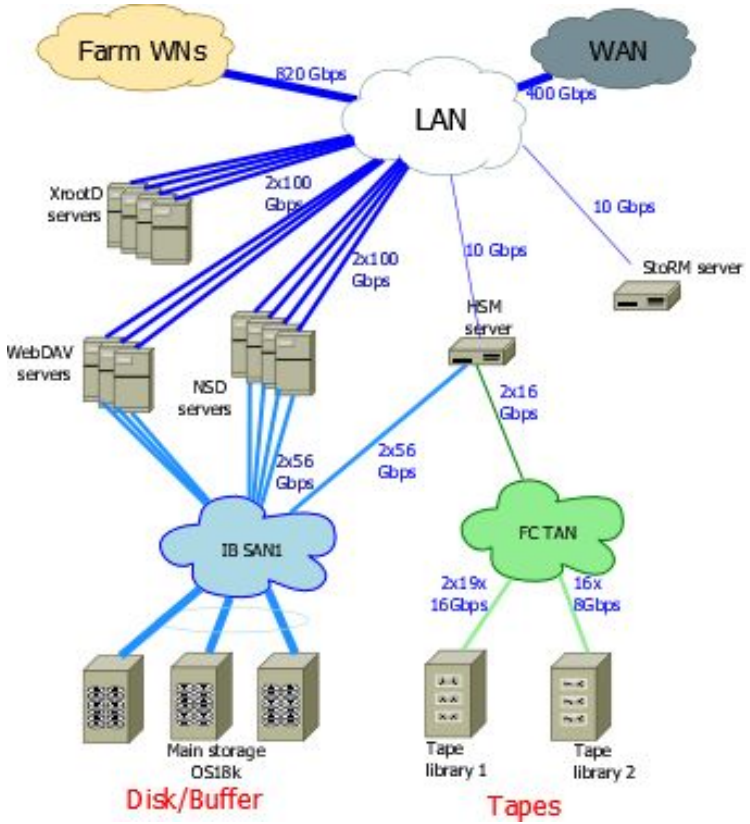
HSM - Hierarchical Storage Management

Architectural choice

- Solution well consolidated over the years
- Storage servers:
  - SAN-based solution
  - Backend: Infiniband 56Gbps (FDR) and 100Gbps (EDR) and FiberChannel 16, 32 Gbps
  - Frontend: 2x100 GbE, 2x25 GbE and 4x10 GbE
- Software:
  - Parallel file system IBM Spectrum Scale (aka GPFS) as POSIX interface and backend for all data management and data transfer services
  - Interface to tape: IBM Spectrum Protect (aka TSM) + in-house optimization layer
  - Advantages:
    - performance
    - relying on stable and well supported sw
    - minimizing support effort

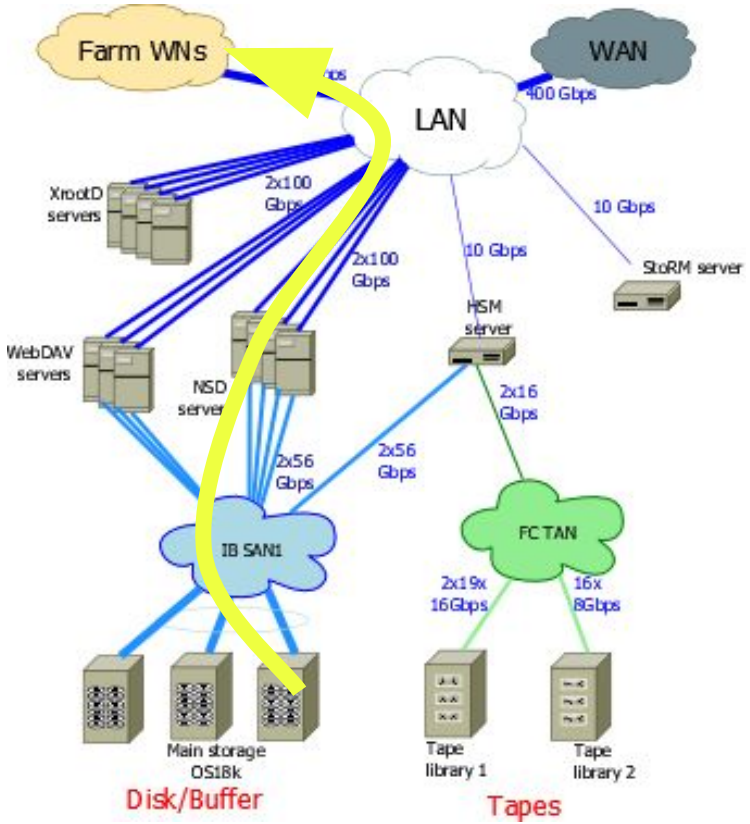
Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



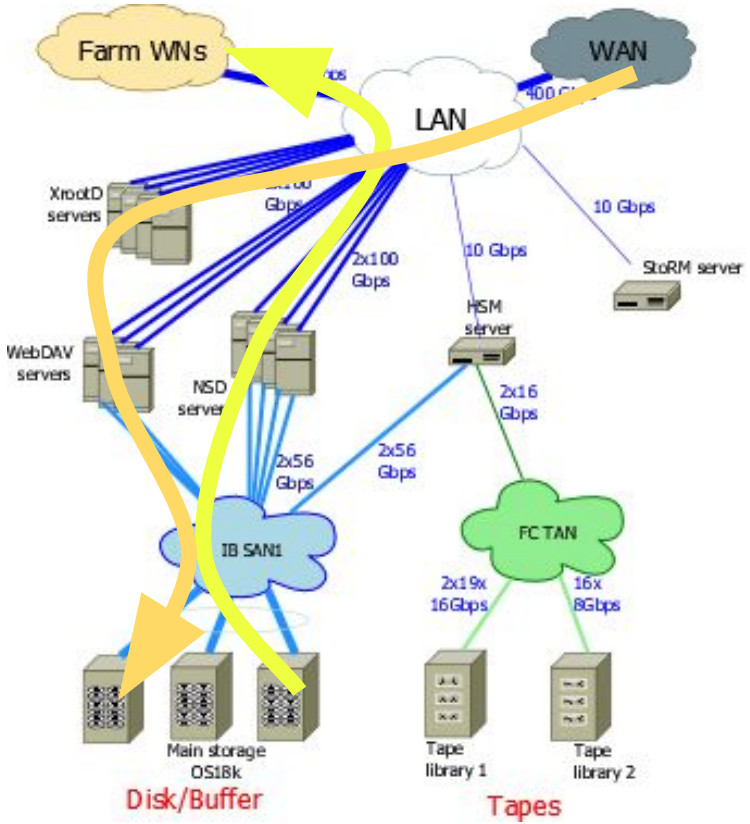
Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



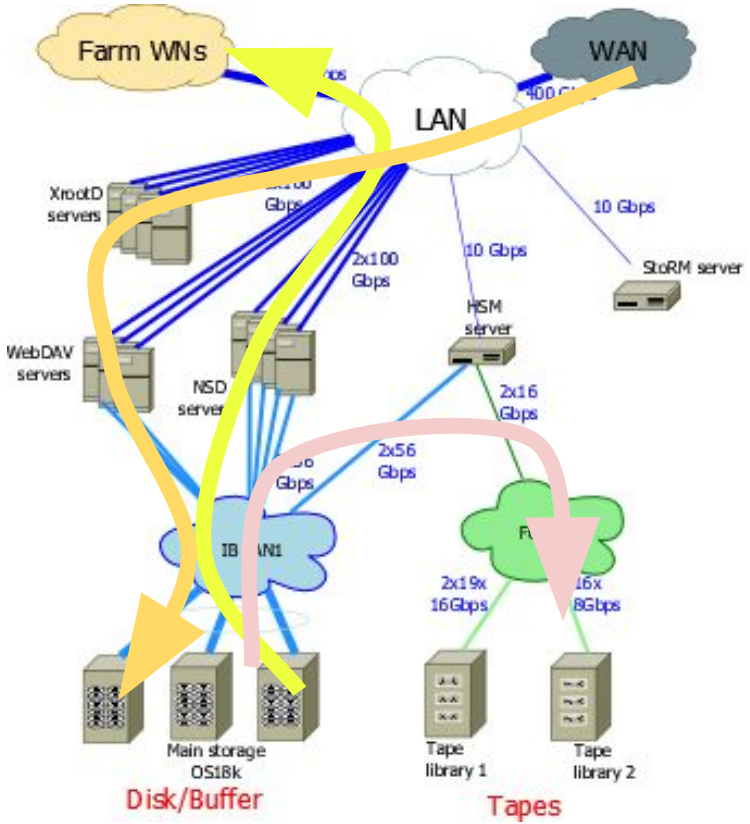
Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



Typical data flow (CMS)

- A single big experiment has a dedicated cluster
- Dedicated servers:
  - 4 NSD servers
  - 3 StoRM WebDAV servers
  - 4 XrootD servers
  - 1 StoRM frontend/backend server (VM)
  - 1 HSM server
- > 1000 clients mounting filesystem



Distributed RAID vs. RAID6

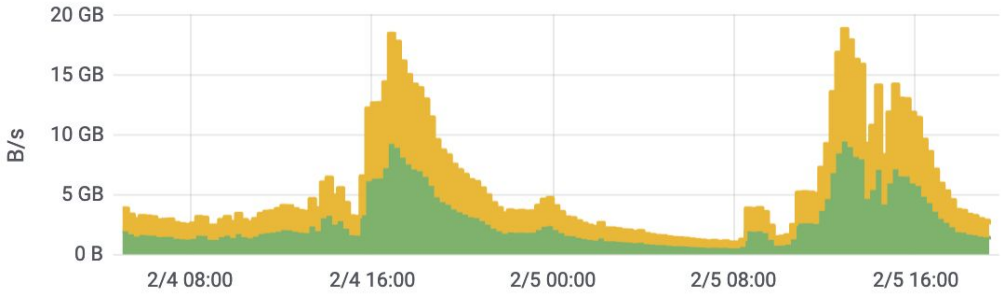
- Usually implemented as “floating” RAID6 (8+2) over bigger (>10) disk pool
- Using “reserved capacity” to restore missing blocks in case of disk failure
- With a disk pool big enough recovery time becomes significantly reduced
  - Failure of 4TB disk: Disk pool of 180 disks – 3.5 hours to restore redundancy (under heavy I/O load) against 20-22 hours in traditional RAID6
  - Failure of 6TB disk: Disk pool of 95 disks – 3 hours to restore redundancy
  - Failure of 8TB disk in RAID6 (8+2) takes ~50 hours for reconstruction
- Drawback of Distributed RAID
  - I/O performance may be affected by ~20%

Procedures and human power

- All Storage resources and services managed by the **Data Management and Storage group** of 6 FTE
  - 6 technologists + 2 technicians
- Intensive pre-production testing of new HW
- Installation and configuration of servers and services via Foreman/Puppet
- Extensive monitoring and automated actions via SENSU
- Redundancy everywhere (no single point of failure)
  - Active-standby (with manual failover) if no redundancy possible
- All the work done during working hours
- HW support contracts – 4h on site (but only during working hours)

Backup slides

LHCb example



Servers (dedicated):

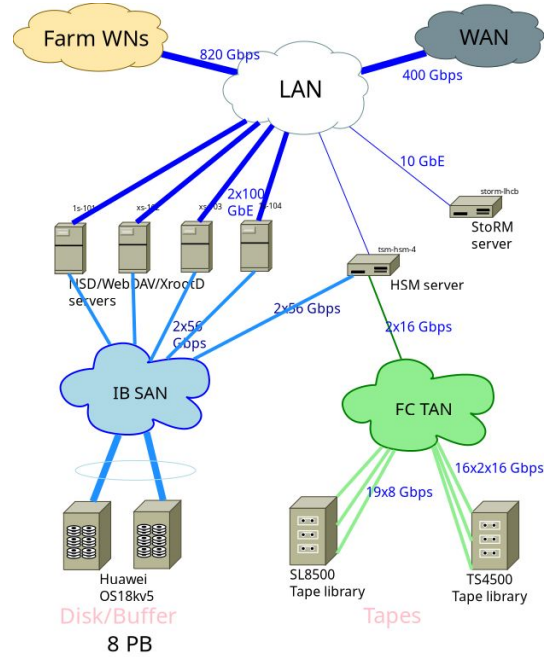
- 4 as WebDAV, XrootD and NSD
- 1 (VM) as StoRM FE/BE
- 1 as HSM

Storage (shared with other exps)

- 2 Huawei OS18Kv5 (main storage)
- 1 SL8500 Tape Library
- 1 TS4500 Tape Library

So, only 4 I/O servers and

2 service nodes for 8PB of data!



ATLAS example

Servers (dedicated):

- 3 as GridFTP and NSD
- 6 as NSD
- 2 as WebDAV and XrootD
- 2 as metadata servers
- 1 (VM) as StoRM FE/BE
- 1 as HSM

Storage (shared with other exps)

- 2 Dell MD3820f as metadata storage
- 2 DDN SFA12k as main storage
- 3 Huawei OS18Kv5 (main storage)
- 1 SL8500 Tape library