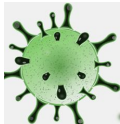
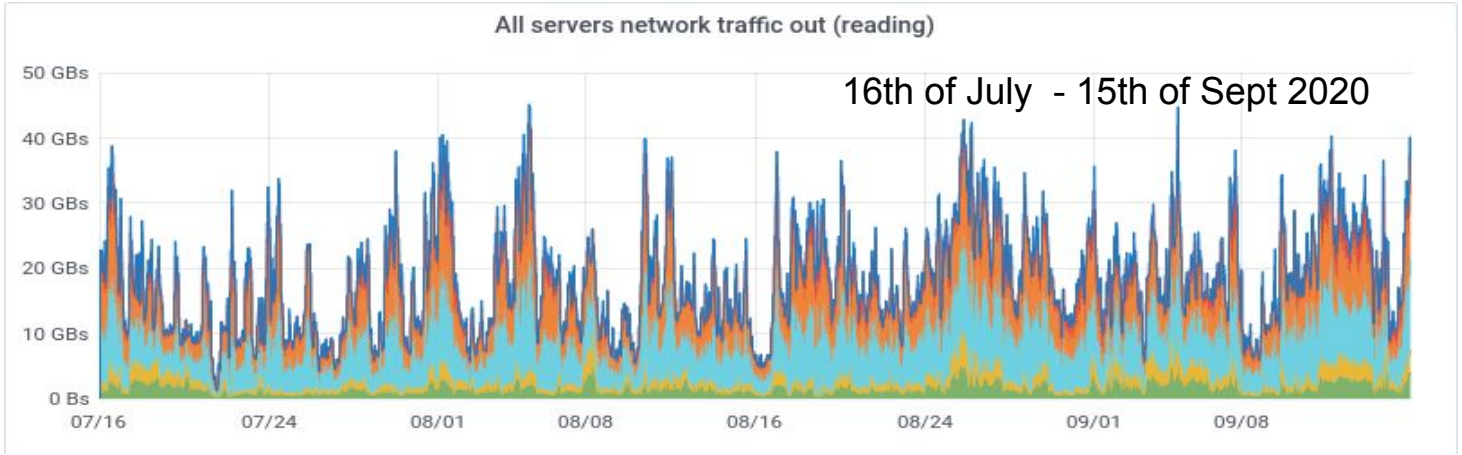
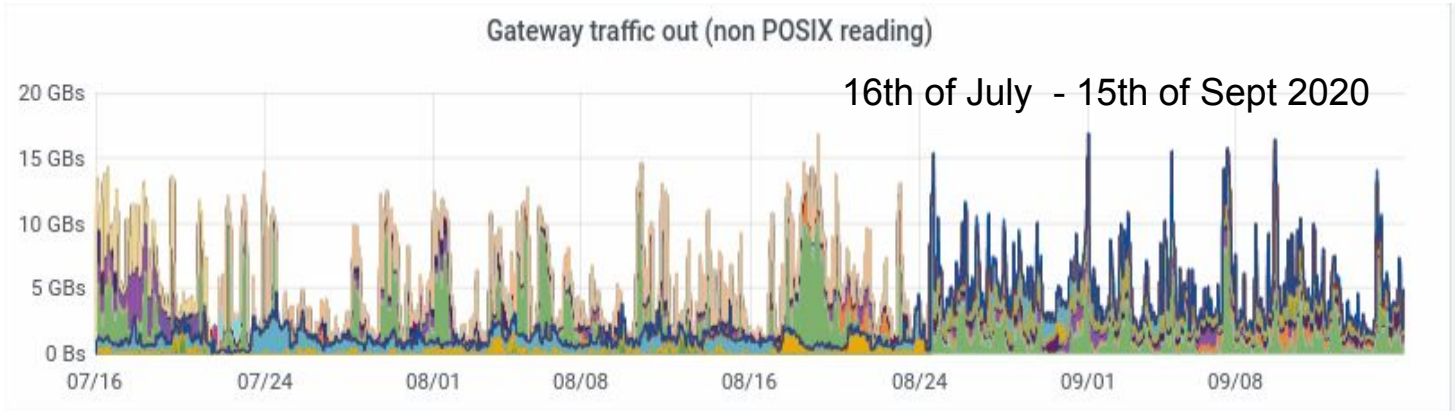
State of Storage

CdG 18th of September 2020



Business as usual





Disk storage in produzione

Installed: 41.07 PB Pledge 2020: 45 PB Used: 32.6 PB

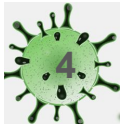
Sistema	modello	Capacita', TB	esperimenti	scadenza
ddn-10, ddn-11	DDN SFA12k	10752	Atlas, Alice, AMS	03/2021
os6k8	Huawei OS6800v3	3400	ALICE, GR2	2022
md-1,md-2,md-3,md-4	Dell MD3860f	2308	DS, Virgo, Archive	11/2021
md-7	Dell MD3820f	20	Metadati, home, SW	04/2021
md-5, md-6	Dell MD3820f	8	metadati	06/2021
os18k1, os18k2	Huawei OS18000v5	7800	LHCb, ALICE	2023
os18k3, os18k5, os18k5	Huawei OS18000v5	11700	ATLAS, CMS	2024
ddn-12, ddn-13	DDN SFA 7990	5060	GR2,GR3	2025
ddn-14, ddn-15	DDN SFA 2000NV	24	metadati	2025

Recent movements

Added +5.06 PB (ddn12, ddn13)
Removed - 1.91 PB (ddn04, 4 JBOD Supermicro)
Net increase = 3.15 PB

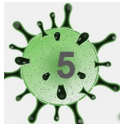
- Gpfs_tsm_alice: DDN04 -> OS18K (- 960TB over-pledge)
- Gpfs_data: OS6K8 -> DDN12,DDN13 (+550TB)
- Gpfs_alice: OS18K -> OS6K8 (+2100-800=1300TB)
- Gpfs_ds50: +450TB
- 800TB can be assigned to LHCb and/or ALICE
- 1050TB can be assigned to GR2/GR3

~4.5PB of data has been moved



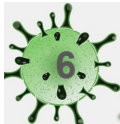
Delay in entering into PROD 2019 storage

- End of Service for GPFS 4.2 on Sept 2020
- In July GPFS on worker nodes updated to 5.0.5-1 (latest available)
- After installing GPFS ver. 5.0.5 we have discovered serious bug which provocated shutdown of entire cluster while adding new storage to existing filesystems
- There was no immediate solution, the bugfix have been released only on Aug.20 with new version of the GPFS package (5.0.5-2)
- Update of GPFS on the farm completed this week
- Added 450TB to gpfs_ds50 on Sept.17
- ALICE was not affected (gpfs_alice not mounted on worker nodes)



Recent problems

- ALICE
  - 85,417,346 “dark data” files (out of 241,735,401) - confirmed by exp, working
- CMS
  - Use of WebDAV in PROD: problems with Phedex configuration (GGUS 148379)
- ATLAS
  - Zero-size files in buffer area: Atlas is informed and currently checking
    - This has triggered general check for zero-sized files on all FS (see next slide)
- VIRGO
  - Reordering data on tape: not clear how needed it is; need to free disk space
- ARGO
  - 77TB in 3.2M files: to be archived and moved to tape
- LHCb
  - Very small amount of failures in upload/download led to the discovery of a bug in the puppet module for StoRM gridftp CentOS7 -> fixed



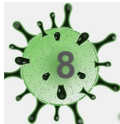
File di dimensione 0

	Tot files	Size 0 files	% size 0 files
ALICE disk	241735401	**41169**	0,02
ALICE tape	6535492	**75805**	1,16
ATLAS disk	19321176	**58486**	0,30
ATLAS tape	15009688	**22897**	0,15
CMS disk	2919436	**3189**	0,11
CMS tape	7761359	**192**	0,00
LHCb disk	42524169	**912470**	2,15
LHCb tape	4467837	**488**	0,01
gpfs_data	128254660	**2137662**	1,67
gpfs_archive	22521838	**140909**	0,63
TOTAL	491051056	**3393267**	0,69

- Per ATLAS tape controlli a campione
  - PrepareToPut che ha successo e poi nessun trasferimento
- Su data soprattutto file di borexino (650k) e auger (780k)
- Su archive soprattutto file di borexino (84k) e opera (50k)
- Agli esperimenti verranno inviate le liste dei file vuoti
  - Liste dai FS tape inviate ad ATLAS, Borexino e Opera
    - Borexino file cancellati da referente
    - ATLAS e Opera attesa risposta

Ceph Testbed Deployment

- Ceph test cluster ready for further studies
- Hardware
  - 8 nodes
  - 4 jbods (SAS) 60 X 8TB disks→ total 1920TB RAW space
- Focus of the second round of tests:
  - Installation procedure
  - Performance tuning
  - Network topology
  - Multiple service object storage and block storage in addition to POSIX FS
  - Finding optimal EC setup



Object Storage S3 (Test) instance @CNAF

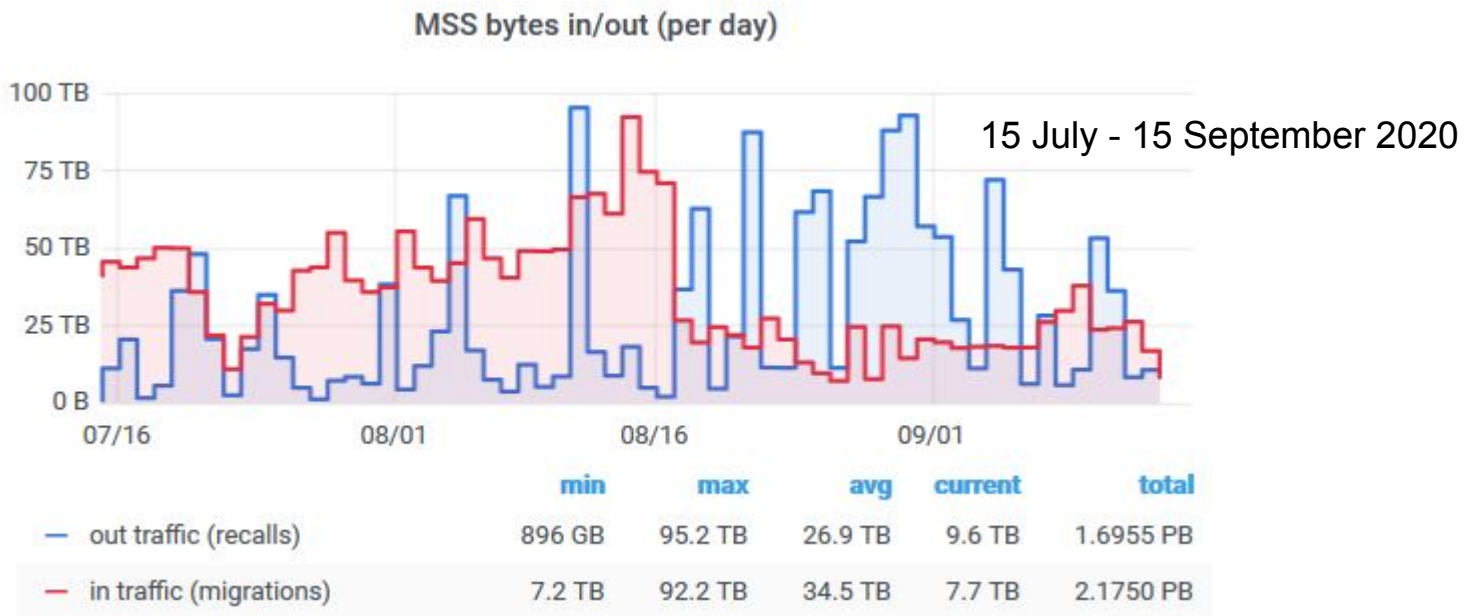
- GPFS v5 Shared Nothing cluster of 5 nodes
- 2 protocol nodes in Cluster Export Services (CES) to export Object Service:
  - 8TB of disk space (easily expandable), triple replica
  - DNS load balancer
  - Integration with Openstack Cloud@CNAF Cluster under test (main goal: configure IBM Openstack Swift with Cloud@CNAF remote Keystone, to avoid Openstack Controller bottleneck for data throughput)
- In parallel with the GPFS v5 cluster, a CEPH cluster has been setup by SDDS as Cloud@CNAF Swift Object Storage backend:
  - High data throughput easily supported: data transfer flow from HAProxy to CEPH Monitors
- Performances and stability to be compared between CEPH (open source) and GPFS (license protected), tests are foreseen

Migrazione a CentOS 7 e GPFS 5

- GPFS 4 è in end-of-support
  - Migrazione a GPFS 5 prevede upgrade del SO a CentOS 7
- StoRM e HSM:
  - storm-ams (ams, darkside, dampe, juno): martedì 29/9 (ore 8-19)
  - storm-lhcb + HSM (solo GPFS 5): martedì 6/10 (ore 8-19)
  - storm-cms + HSM (SO e GPFS): martedì 13/10 (ore 8-14)
  - storm-atlas + HSM (SO e GPFS): martedì 20/10 (ore 8-14)
  - storm-archive + HSM (solo GPFS): martedì 27/10 (ore 8-14)
  - HSM alice (solo GPFS): da definire (pochi minuti)
- GridFTP/xrootd/webdav
  - Da migrare a GPFS 5 senza down (già a CentOS 7)
- Server GPFS
  - Da migrare a CentOS 7 e GPFS 5 senza down

Stato tape

- 7.3 PB liberi (complessivamente sulle 2 librerie). Usati 80.8 PB
  - Scritture ultimi 6 mesi: 1.1 PB al mese
  - Repack per liberare 500 TB in corso (spostando 60 TB)
  - Avviata procedura di acquisto di ulteriori 6.6 PB per arrivare a pledge 2020 (95 PB)



Recall in GEMSS

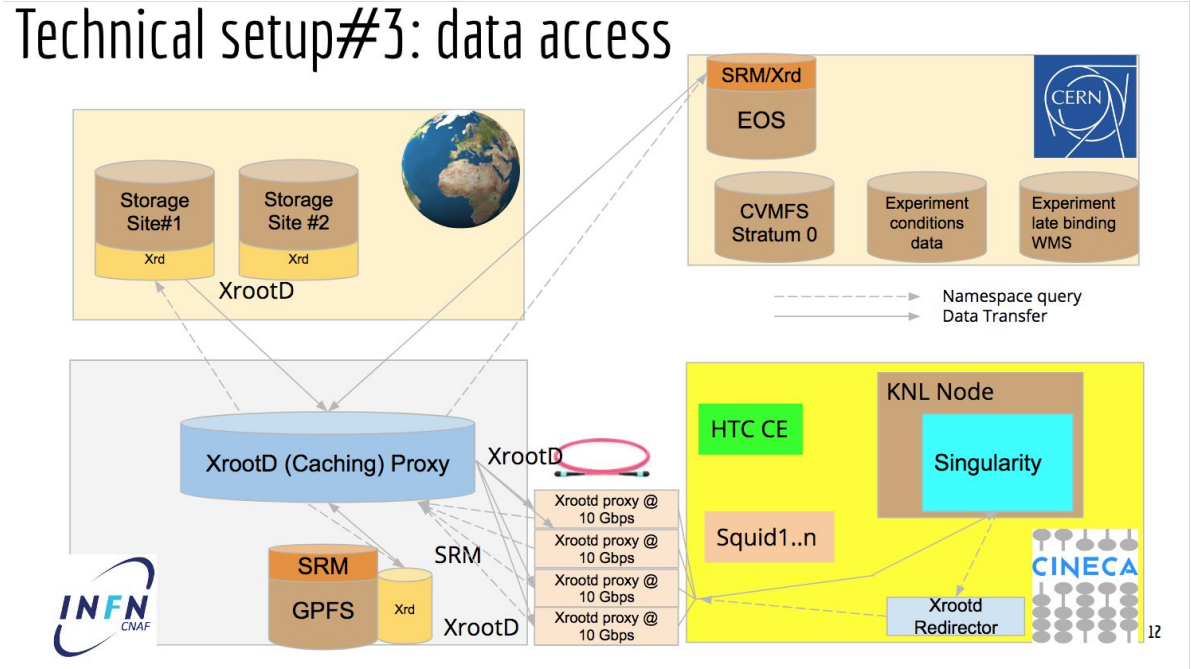
- Modifica per ottimizzare recall da più librerie
  - Prima, per ogni FS, unica coda di tape da leggere
    - e un certo numero di tape leggibili in contemporanea
  - Ora, per ogni FS, una coda per libreria
    - e un certo numero di tape leggibili in contemporanea per ciascuna libreria
  - Primi test sono ok
  - Da adeguare l’orchestratore, che gestisce dinamicamente i numeri di tape leggibili in contemporanea

Oracle Database: Atlas Muon Calibration

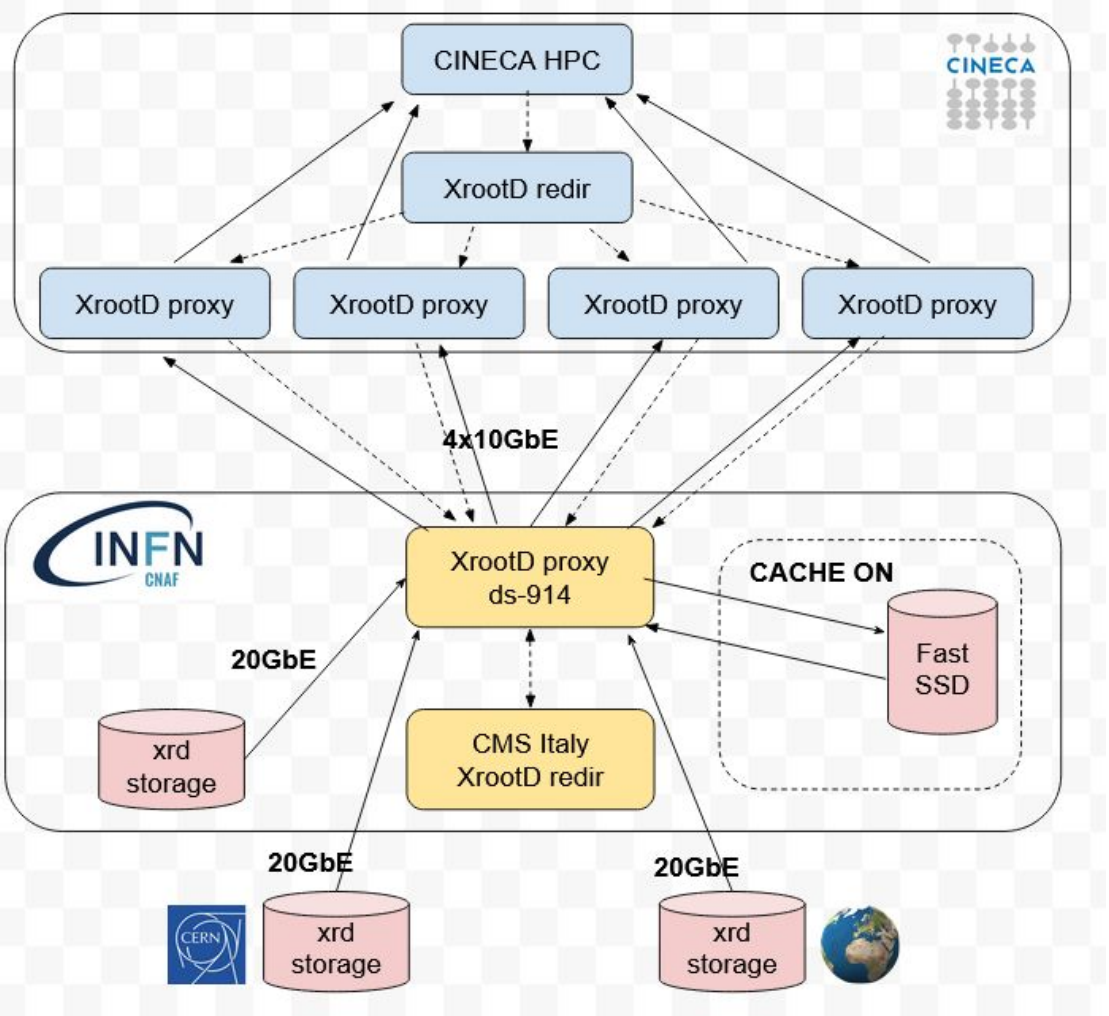
- Iniziata attività per aggiornare ad Oracle 19c e migrare su nuovo hardware
  - DB Atlas è Oracle 11g, come anche le istanze CERN ed Umich
  - CERN e Univ. Michigan migreranno ad Oracle 19c (CERN entro 2020, Umich?)
  - DB risiede su server Dell (2009) e storage in via di dismissione
  - nuova installazione: VM su nuovi HV e nuovo storage di servizio
  - non più Oracle RAC cluster, ma Single Instance DB su VM
- Previsione: entro 2020
- Coordinamento - contatti con CERN DBA (Atlas) e DBA Umich
  - atlas-db-cross-coordination@cern.ch, atlas-dba@cern.ch, smckee@umich.edu
- Man power - effort
  - Tier-1: alessandro.cavalli@cnaf.infn.it (~ 5 - 10%)
  - possibilità scambio informazioni tecniche Oracle con Servizio Sistema Informativo

Data management per job su partizione KNL Cineca

- Estensione del Tier1 su sistema HPC, “Prace Project Access” alla comunità LHC Italy per utilizzare la partizione A2 di Marconi (già usato al 240%!)



Courtesy of
T. Boccali



- Stiamo investigando la possibilità che l'XrootD proxy @CNAF faccia anche caching su fs dedicato.
- Partners in crime: T. Boccali, D.Ciangottini, D.Spiga (+ CNAF and CINECA people of course)