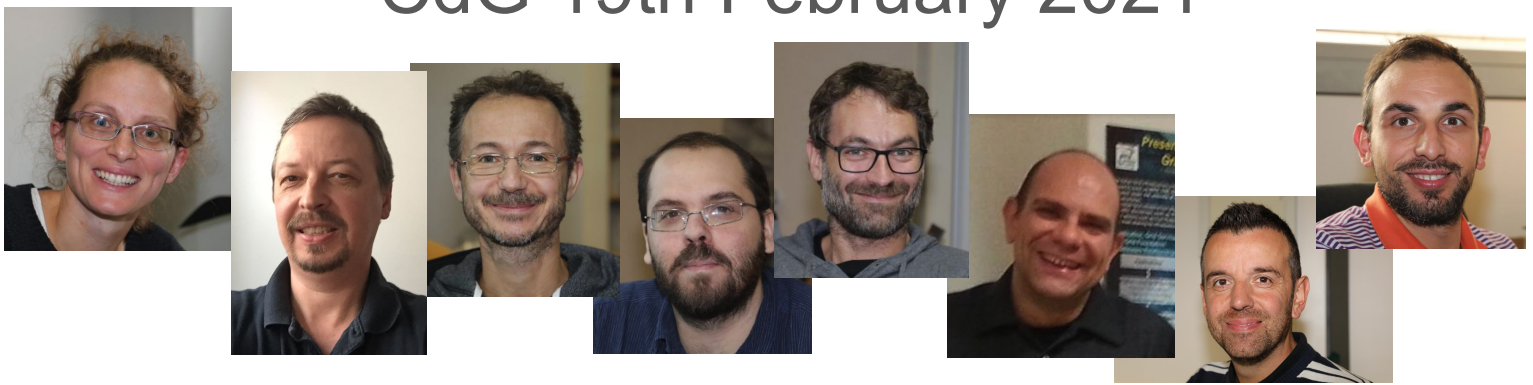
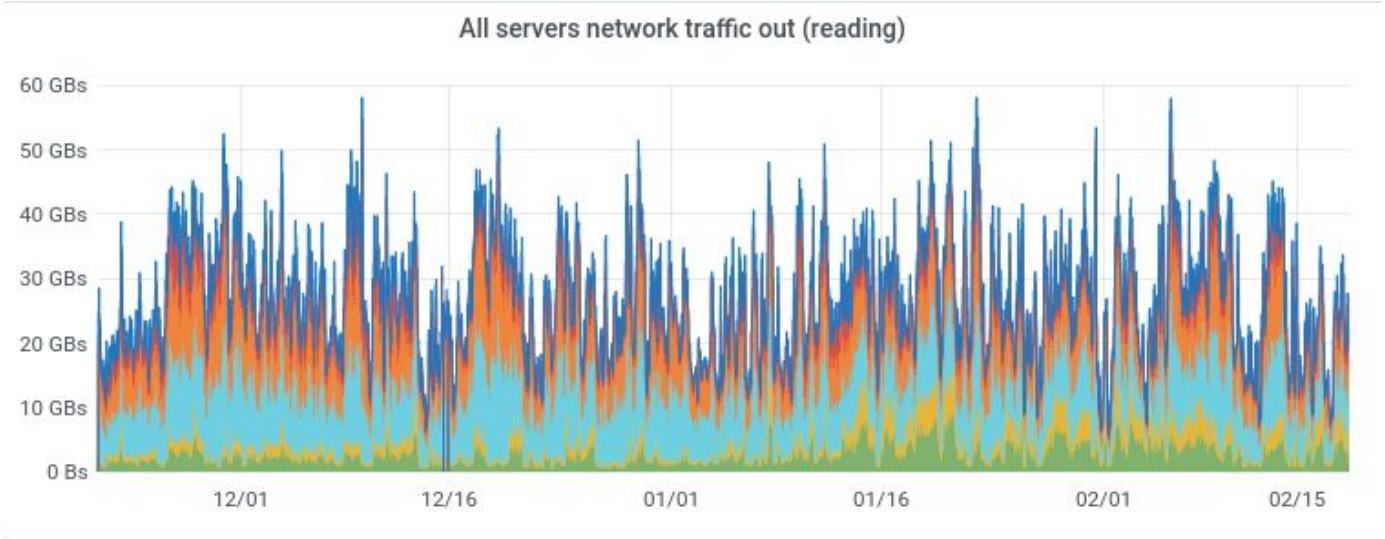
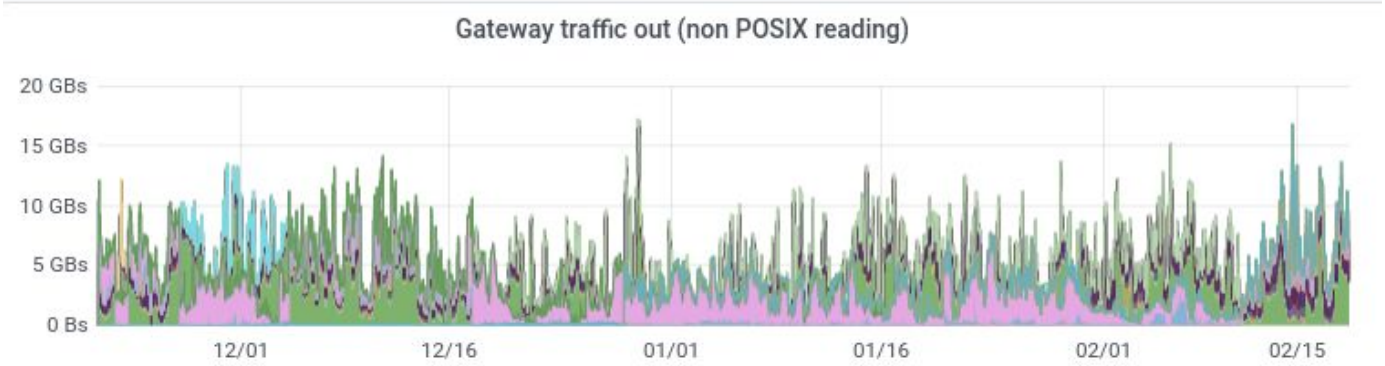
State of Storage

CdG 19th February 2021



Business as usual





Disk storage in produzione

Installed: 41.07 PB Pledge 2020: 45 PB Used: 36.1 PB

Sistema	modello	Capacita', TB	esperimenti	scadenza
ddn-10, ddn-11	DDN SFA12k	10752	Atlas, Alice, AMS	03/2021→ 06/2023
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

Interventi di manutenzione

- Mellanox: Eccessiva usura dei dischi SSD - aggiornamento FW - FATTO
- Aggiornamento FW sugli HBA dei HSM servers - FATTO
- Inserimento dei nuovi switch FC nella TAN di produzione - in corso oggi

Recent problems

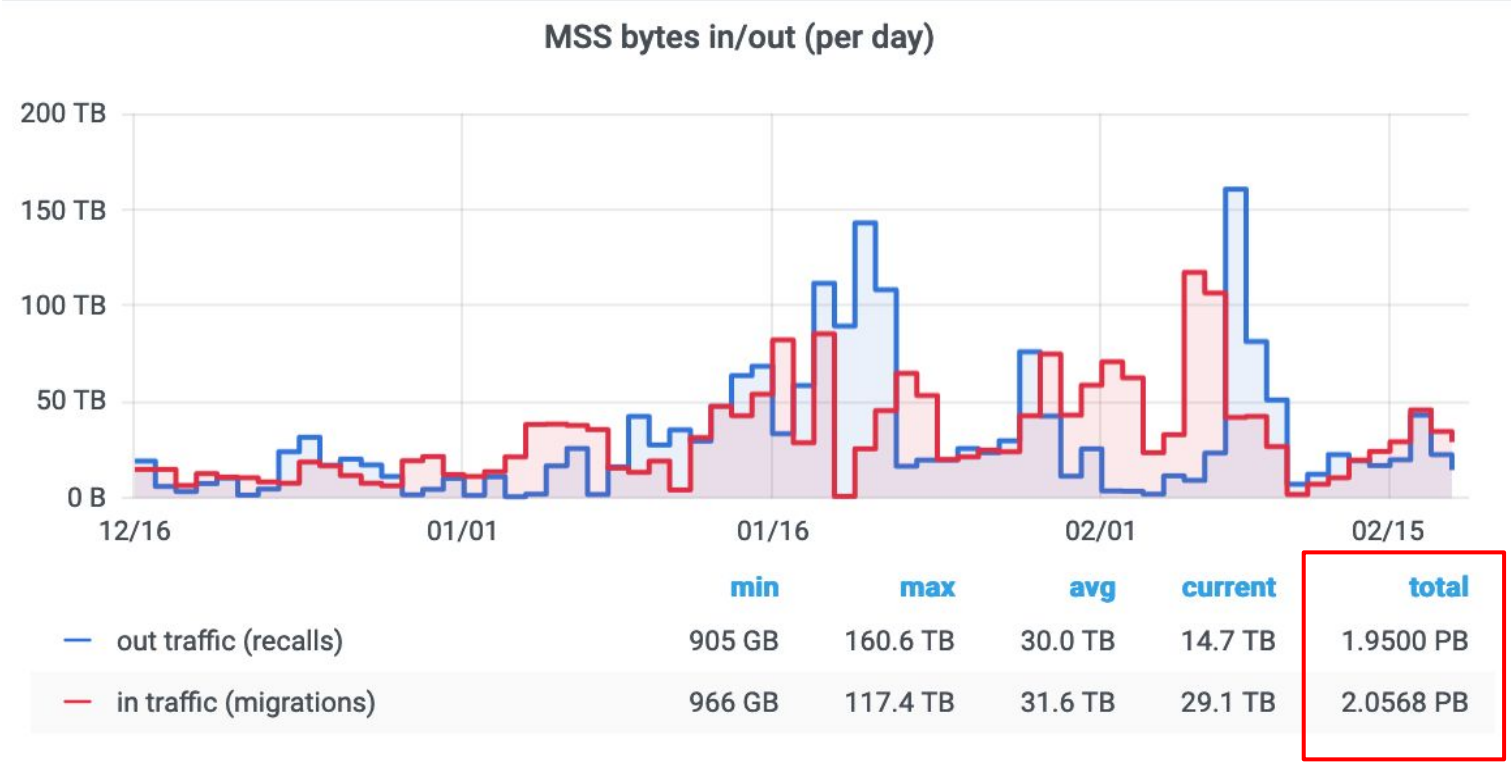
- ATLAS
  - Trasferimenti molto lenti con AGLT2 (GGUS ticket 150642)
    - Network problems? Under investigation
  - Performance StoRM WebDAV versus gridftp (GGUS ticket 150288)
    - Under investigation by StoRM developers (GGUS ticket 150181)
  - Falliment TPC Pull (GGUS ticket 150288) per errata configurazione di un endpoint StoRM WebDAV
- CMS
  - Fallimenti TPC Push (GGUS ticket 150299)
    - Connessione chiusa da UNL per TPC connection contention di XRootD (bug XRootd5)
  - Token BoL scaduti (GGUS ticket 150339), bug già trovato in passato da fissare in FTS
- LHCb
  - Fallimenti TPC (GGUS ticket 150223), per errata configurazione endpoint StoRM WebDAV
- VIRGO
  - Fallimenti stress test per Stashcache, dovuti a un bug CVMFS
  - Da concludere migrazione del filesystem, necessario togliere path hardcoded dagli eseguibili

Novità sui servizi

- Dopo un periodo di test, configurato StoRM WebDAV in produzione per LHCb, sostituendo uno dei 4 gridftp/xrootd dell'esperimento
  - Verificato con l'esperimento che non c'è necessità di fs montato R/W sui WN; R/O è sufficiente;
  - L'esperimento si organizza per scrivere con https e leggere posix, senza poter più sfruttare gfal-xattr per conoscere il path sul fs (funzionava con srm://)
- Proposta di convertire anche un gridftp di Atlas a StoRM WebDAV
- CTA-LST: scrittura limitata al gruppo (IAM) lst-data-manager, sia per StoRM WebDAV che per StoRM backend
- Fs dedicato gpfs_escape per il datalake ESCAPE (non più gpfs_atlas)
- Per ESCAPE, fornito un server origin XRootD (X509 voms CMS) che fornisce dati dal datalake a un CRAB test CMS girato al CINECA.

Stato tape

15 December - 15 February 2021



	min	max	avg	current	total
out traffic (recalls)	905 GB	160.6 TB	30.0 TB	14.7 TB	1.9500 PB
in traffic (migrations)	966 GB	117.4 TB	31.6 TB	29.1 TB	2.0568 PB

Stato tape

- 11 PB liberi (complessivamente sulle 2 librerie). Usati 85 PB
  - Gran parte delle scritture su nuova libreria
    - Tutti LHC
    - Xenon, CTA, Virgo
- Valutazioni in corso su futuro della vecchia libreria
  - In vista dello spostamento al tecnopolo

Library	Tape drives	Max data rate/drive, MB/s	Max slots	Max tape capacity, TB	Installed cartridges	Used capacity, PB
SL8500 (Oracle)	16*T10KD	250	10000	8.4	~10000	79
TS4500 (IBM)	19*TS1160	400	6198	20(30)	750	6

Requisiti per lo storage near-line (tapes) ?

- Abbiamo notato un aumento significativo in rate di scritture > 5 GB/s (ATLAS)
- Il dimensionamento del sistema attuale (buffer ATLAS di 570TB) permette di accettare in modo continuo
  - Solo scrittura o lettura: ~2.8 GB/s
  - Scrittura e lettura contemporanee: ~1.5 GB/s
- Se viene richiesto supportare injection rate più alto dobbiamo ridimensionare il buffer
  - Ad es. per supportare 5 GB/s (2.5 GB/s scrittura su buffer e 2.5 GB/s lettura per migrare su tape) dobbiamo predisporre un buffer ~2 PB con dischi rotanti
    Oppure 25-30 TB di dischi NVMe + un buffer di overflow ~1 PB su dischi rotanti

Tape drives performance vs. file size

Recente recall da archive di 86000 file per un totale di 340 GB

- File da 4 MB

Sistema tape funziona bene con file di dimensioni >= 1 GB

Test di lettura con 1 tape drive TS1160 da 1 tape (HSM) server:

- rate nativo 400MB/s
- 100 GB in file di diverse dimensioni (scritti uno dopo l'altro su tape):
  - 1 file da 100 GB
  - 10 file da 10 GB
  - 100 file da 1 GB
  - 1000 file da 100 MB
  - 10000 file da 10 MB

Traditional vs dynamic allocation

Sample comparison: real CMS bulk recalls. Similar number of files and TB read

Traditional

Recall period: 18-23 Apr 2019

Duration: 138 hours

Number of files: 98k

Data read: 319.5 TB

Avg drives used: 3.7

Avg throughput: 650 MB/s

Dynamic

Recall period: 17-19 Jan 2021

Duration: 72 hours

Number of files: 92k

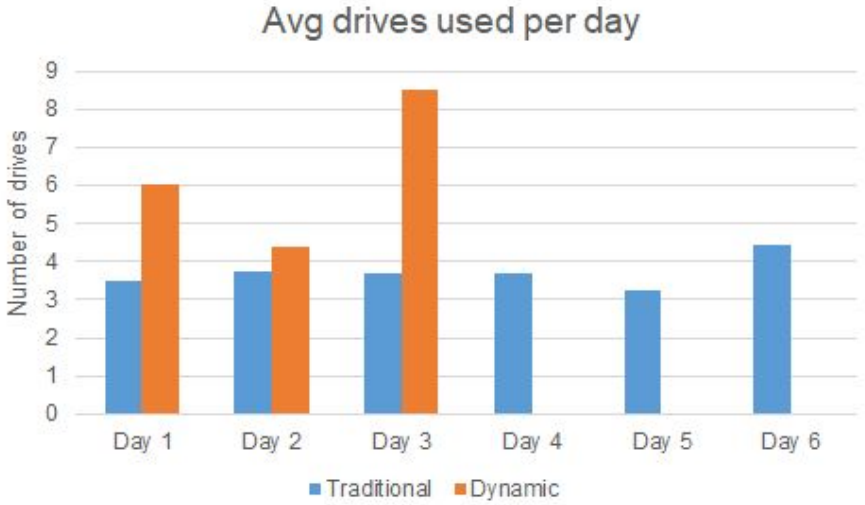
Data read: 313.5 TB

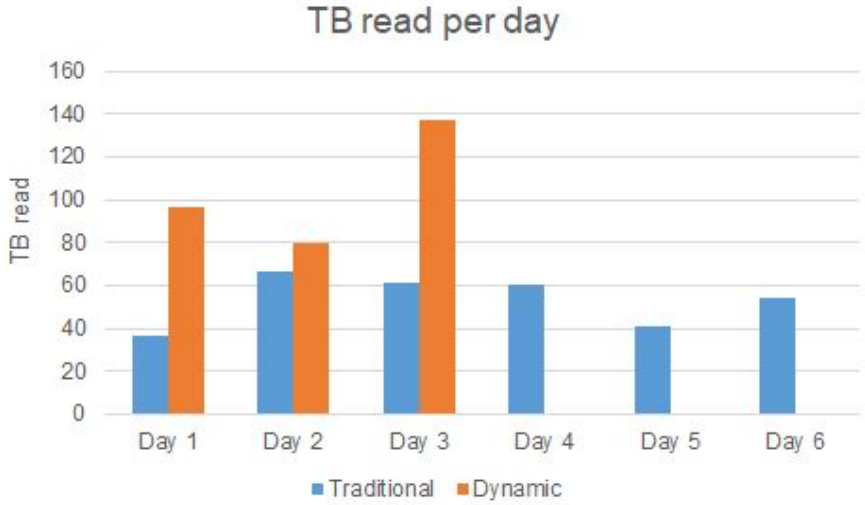
Avg drives used: 6.3

Avg throughput: 1.2 GB/s (+85

Traditional vs dynamic allocation

Sample comparison: real CMS bulk recalls. Similar number of files and TB read





Max drives used per day:

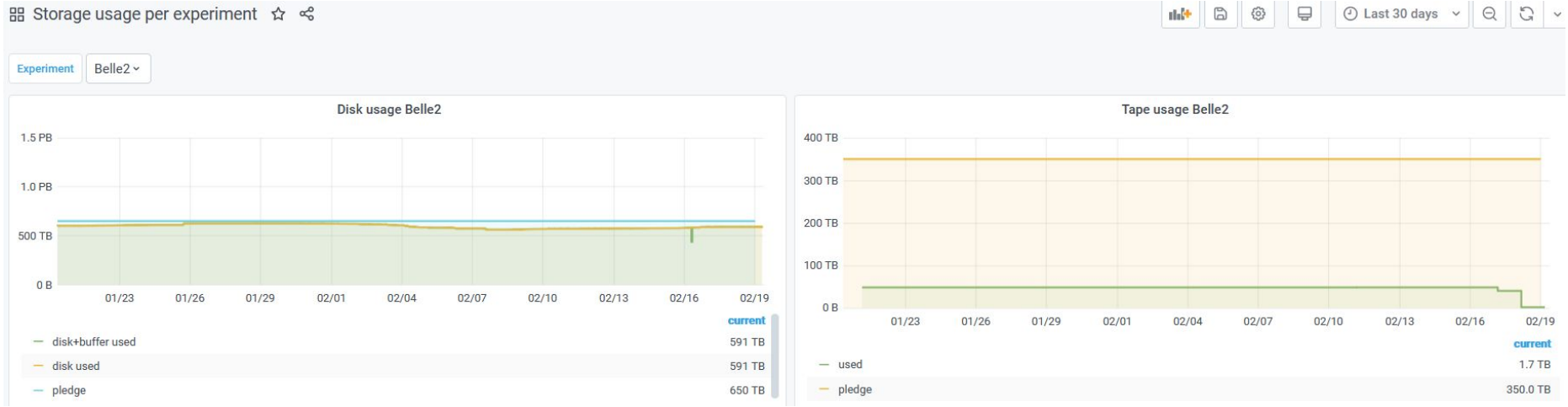
4.4 traditional vs 8.5 dynamic

Max TB read per day:

66.5 traditional vs 137 dynamic

Accounting storage

Nuova dashboard pubblica su Grafana con occupazione disco e tape per esperimento



https://t1metria.cr.cnaf.infn.it/d/ZArHZvEMz/storage-usage-per-experiment