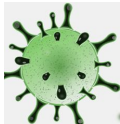
# Evoluzione dello Storage

Vladimir Sapunenko

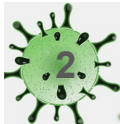
Workshop “CNAF Reloaded”
10 Novembre 2020



# Come dobbiamo gestire l’Evoluzione?

Secondo me bisogna seguire questo percorso

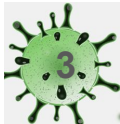
1. Raccolta e definizione dei requisiti da parte degli utenti
2. Definizione dei servizi necessarie per realizzare richieste di punto 1.
3. Analisi della soluzione attuale
4. Selezione delle Soluzioni tecnologici più adatti
5. Valutazioni dei costi e degli risorse umane necessarie
6. Implementare le modifiche alla soluzione attuale



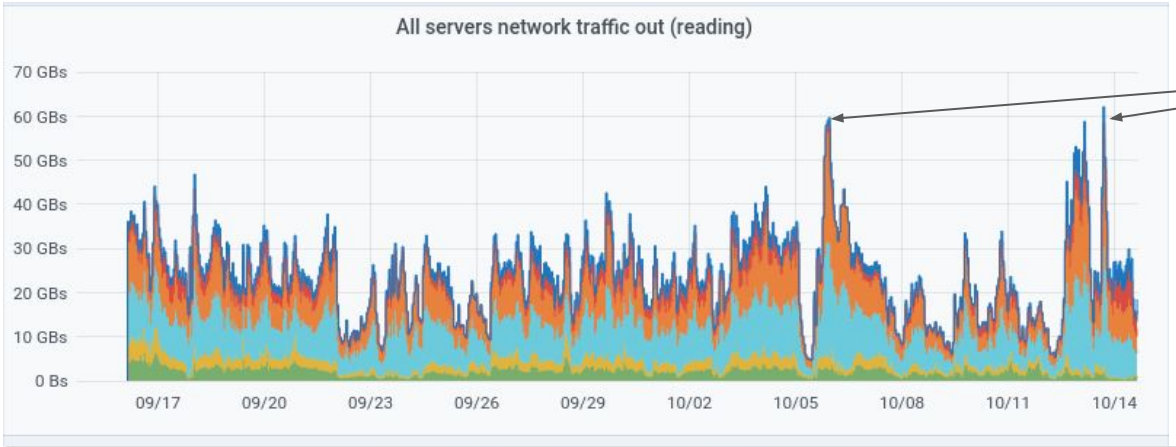
# Disk storage in produzione (41PB)

| Sistema | modello | Capacita', TB | esperimenti | scadenza |
|---|---|---|---|---|
| ddn-10, ddn-11 | DDN SFA12k | 10752 | Atlas, Alice, AMS | 03/2021 |
| 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 |
| os6k8 | Huawei OS6800v3 | 3400 | ALICE, GR2 | 2022 |
| 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 |

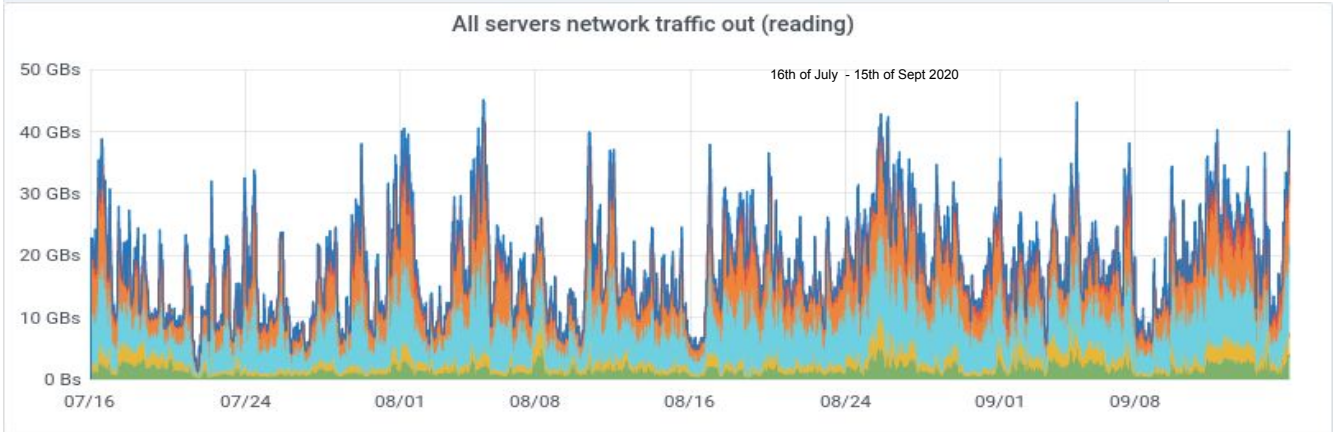
In verde - sistemi da traslocare, in rosa - da dismettere, in giallo - da decidere



# Network: l’uso in aumento

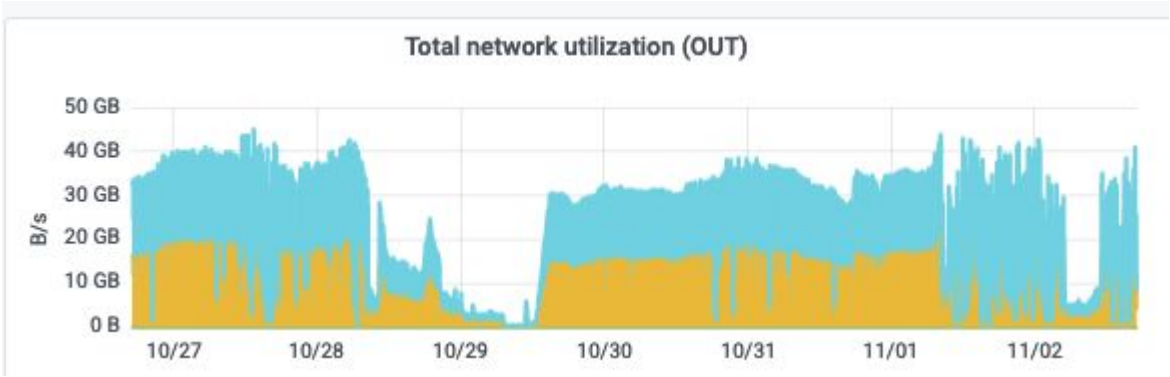



Saturazione (quasi) link
400Gb CINECA-CNAF



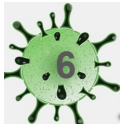
# Requirements per lo storage sul disco

- Accesso via
  - POSIX
  - XrootD
  - SRM (almeno per lo buffer d’avanti all tape)
  - WebDAV/HTTP
- I/O rate 5 MB/s per ogni TB (non e’ un numero a caso)
  - Ad. es. CMS ha 8193TB di spazio disco e in grado tirare fino a 40GB/s



# Considerazioni tecnologici

- Con la crescita prevista delle pledge per disco possiamo starci dietro con aumento dimensione del singolo disco
  - numero dei dischi può rimanere invariato (7-8k)
- Aumento delle prestazioni - progresso tecnologico (PCI3->PCI4, EDR IB, 200GbE, dual-actuator HDD) ma sempre indietro dalla crescita di capacità
  - Rate di I/O relativo alla capacità va in diminuzione (MB/s/TB)



# Modello attuale (GPFS + Enterprise storage)

- Numero di server in grosso modo invariato
- Crescita di capacità definita dall'aumento capacità del disco
- Efficienza dell'uso dello spazio ~99% dello spazio utilizzabile (80% del RAW)
- Resilienza a fallimento dei server (fino a N-N/2)
- Ricostruzione di un disco gestito da FW sulla base di RAID distribuito (o EC), trasparente per i server.

Numero di server: 10 serv x exp LHC +10 serv per non_LHC = 50

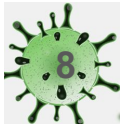
- NSD server e Data export server possono essere condivisi
- Da aggiungere costi manutenzione delle licenze SW

# Studio delle Soluzioni alternativi

Abbiamo valutato

- dCache
- EOS
- Lustre
- Object Storage
- CephFS

Negli prossimi slides riporto alcuni problematiche di varie soluzioni che hanno portato alla scelta di CephFS come soluzione più adatta alle nostri esigenze



# dCache

- Prodotto di DESY
- Difficilmente utilizzabile da esperimenti non LHC
  - Specialmente per chi vuole accesso POSIX
- Singoli storage nodes con la protezione locale via RAID controller
- HA (protezione contro spegnimento di un server) ed eliminazione dei HostSpot fatta via replicazione dati su altri server - overhead + del doppio sullo spazio RAW

# EOS

- Open Source del CERN
- Poco diffuso anche nel modo scientifico
- Difficilmente utilizzabile da esperimenti non LHC
- Manpower dedicato per la gestione (almeno 1FTE diviso tra 2-3 persone)
  - ~60KE/anno
- Bassa densita' - Best Practice: 25 HDD per storage node
- limite dell'occupazione ~80-90% (Best Practice dice 60-80%)
- Ricostruzione dati usa risorse dei server
- Works better with many "small" storage servers

# Lustre

- OpenSource, proprieta’ di DDN (free download per la versione -1), supporto a pagamento
- Prestazioni alti (tunabile per un tipo di data flow)
  - Problematico supporto per data flow misto
- Works better with a few “large” storage servers or Enterprise storage systems

# Object Storage

- No POSIX (ovviamente)
- Basso interesse da parte degli utenti

# CephFS

- OpenSource, proprieta’ di RedHat, supporto a pagamento
- Mancano alcuni funzionalita’ che usiamo in PROD con GPFS
- Manca integrazione con StoRM
- Per XrootD c'è stato un tentativo di interfacciamento fatto al CERN e RAL
  - Da verificare lo stato attuale
- Prestazioni generalmente più basse
- Works better with many “small” storage servers

# Piano di lavoro

- CEPH e CephFS
  - Studio di varie configurazioni con lo Testbed esistente
    - 2 mesi, 2 persone, 0.5 FTE
  - Studio di fattibilità di interfacciamento CephFS con StoRM
    - 2 mesi, 2 persone, 0.2 FTE con supporto da parte SDDS
  - Studio Interfacciamento con XrootD
    - 2 mesi, 1 persona, 0.2 FTE
  - Studio di possibilità di integrazione CEPH (al livello di Block device) con EOS services
    - 2 mesi, 2 persone, 0.2 FTE (Progetto collaborazione con CERN)
  - Installazione dello cluster con lo HW mirato alla architettura CEPH (gara in corso)
    - 1 settimana partendo dalla consegna HW
  - Studio di fattibilità di utilizzo CephFS in un cluster di O(1000) nodi e I/O di alta intensità
    - 2 mesi, 2 persone, 0.2 FTE (con supporto da parte farming)
  - Confronto con GPFS delle prestazioni e stabilità

# Tape Libraries

- 7.3 PB liberi (complessivamente sulle 2 librerie). Usati 80.8 PB
  - Scritture ultimi 6 mesi: 1.1 PB al mese

| Library | Tape drives | Max data rate/drive, MB/s | Max slots | Max tape capacity, TB | Used slots | Used capacity, PB |
|---|---|---|---|---|---|---|
| SL8500 (Oracle) | 17*T10KD | 250 | 10000 | 7.4 | 10000 | 80 |
| TS4500 (IBM) | 19*TS1160 | 400 | 6198 | 20(30) | 355(+400) | 5 |

# Considerazioni su Tape Repack (copia dati da SL8500 a TS4500)

- Procedura nativa
  - tutti dati passano via unico TSM server -> limite supremo e' il canale FC
  - Max 2x16Gb=4GB/s
  - La copia viene fatta da un drive T10KD a un TS1160 (N+N)
  - Con 8 tape drive T10KD possiamo raggiungere (realisticamente) 1.6 GB/s
    - Per copiare 80TB ci vuole 590gg (~2 anni)
  - Con 16 tape drive (e stop per la PROD) - 3.2GB/s (~1 anno)
- Facendo recall sul disco e migrazione sulla nuova libreria
  - Si puo usare piu server (nodi HSM)
  - Il limite e' il numero di tape drive T10KD (e solo meta' dei TS1160)
  - Altra meta dei TS1160 può rimanere in PROD
  - Stima del tempo ~ un anno

# Repack (cont.)

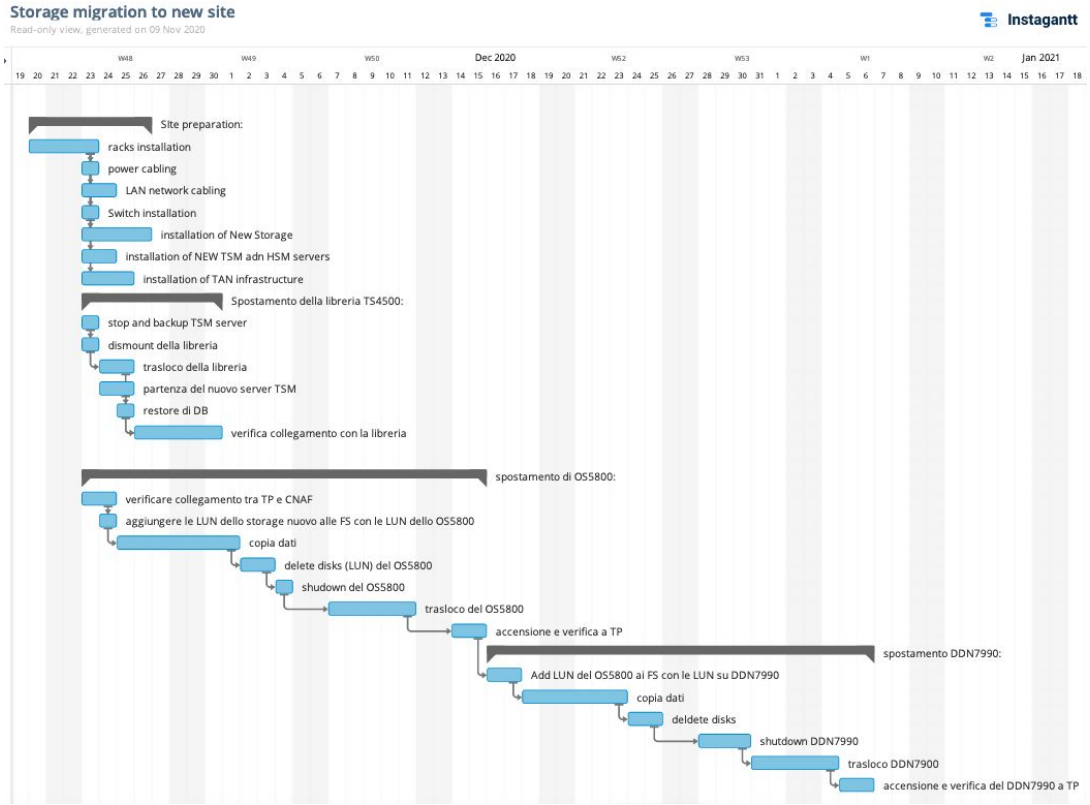
- Prima o poi va fatto
- Se lo facciamo prima Non dovremo traslocare la libreria vecchia (SL8500)
  - Costi di manutenzione della libreria in crescita (ora 50K/anno, +15% dopo 10 anni)
  - Tape drive T10KD non trovano piu in vendita
  - Non dovremo pagare il servizio di trasloco ~50 KEuro
- Siamo ancora in tempo per farlo prima del trasloco (se partiamo subito)
- Occorre comprare le media per 80PB
  - JE (attuale, 20TB) ~ 4000 cassette => spesa immediata ~800 KEuro
  - 30TB per cassetta con nuovi tape drive previsti a meta' del 2021
- Trasloco della TS4500 può essere ritardato rispetto lo spostamento del disco
  - Va bene con schedule del LHC?
  - Va fatta la separazione dello cache dallo disco (in ogni caso)

# Backup slides

# Ceph Testbed

- Hardware
  - 8 nodes
  - 4 jbods (SAS) 60 X 8TB disks→ total  1920TB RAW space
- Focus of the second round of tests:
  - Installation procedure → based on a new developed puppet modules
  - EC 6+2 , Failure domain host
  - Next steps  → Performance tuning
  - Multiple service object storage and block storage in addition to POSIX FS
  - Finding optimal EC setup

# Piano di spostamento



# 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 and GPFS, tests are foreseen

# Costi del disco

| Year | Vendor | Tender price without VAT (-22%), Euro/TB real capacity | price/TB (-10% maintenance charge), Euro/TB | CERN price, (Replica 2, no VAT, no maintenance, EUR/TB |
|---|---|---|---|---|
| 2014 | DELL | 151.26 | 136.1 | 172 |
| 2015 | E4/DDN (2nd choice) | 162 | 145.8 | 155 |
| 2016 | SIELTE/Huawei | 155.1 | 139.5 | 98 |
| 2017 | TIM/Huawei | 88.85 | 88.66 | 93 |
| 2018 | TIM/Huawei | 102.5 | 100.5 | 102,3 |
| 2019 | E4/DDN | 113.9 | 103.5 | 69.75 |
| 2020 | Sielte/Huawei | 79 | 70.87 | 60 |

# Componenti del prezzo disco

- HW
  - HDD
  - RAID controllers
  - Servers
  - Infrastruttura LAN e SAN
- Protection method
  - RAID - risparmio sul numero dei dischi, electric power, spazio
  - Replicazione -> aumento di spesa di corrente elettrica
- Software
  - Costi del supporto:
    - in casa -> manpower (personale dedicato)
    - Out source -> pagamento manutenzione