INFN and DOMA

Luca dell’Agnello, Daniele Cesini
INFN Tier-1 storage system status

• CNAF is currently offering disk and tape resources to more than 30 experiments

• Up to now, aiming to consolidate resources in few boxes
  • ~30 PB-N on disk (➔ ~34 PB-N in ~2 months)
    • 11 GPFS file-systems
    • 12 (➔ 13) storage devices
    • 70 (➔ 62) disk-servers
  • Moving to 100 GbE on LAN
    • Fewer disk-servers (with new tenders: ~1 server per net PB)
  • ~50 PB on tape (5 active + 5 stand-by servers)
    • TSM for tape management

• Usual protocol zoo
  • SRM, gridftp, XrootD, http/WebDaV
Miscellaneous issues (1)

• Storage systems
  • RAID6 has operational limitations on large volumes (i.e. rebuilds)
    • Also distributed RAID is (probably) not the solution (at least with the implementations we have tested)
    • Replication of data or Erasure Code?
  • Is the (enterprise level) huge brick architecture still viable with larger and larger volumes?
    • E.g. move to replica of data with “cheap” hw?
  • Also (very timidly) considering alternatives to Spectrum Scale (GPFS)
    • Not cheap but operationally cost saving

• Tape library
  • In 2016-2017 (before the flood): 17 tapes with read errors found
    • Most likely because of dust
  • 150 tapes damaged in the flood
    • Only 6 out of the 70 sent to the lab for recovery have shown errors
    • But not clear long term effect of humidity
  • Up to now, all tapes have been regularly repacked (Oracle drive A ➔ B ➔ C ➔ D)
    • As by-product, systematic check of tapes
  • Now preparing a tender for a new library
  • We need a new strategy for checking tape status on the present library
    • Continuous low rate read of all tapes?
    • How much will be accessed data acquired up to this year?
Miscellaneous issues (2)

• Tape drives are shared among experiments
  • But the allocation is static in our system (GEMSS)
  • Configuration can be manually preemptively changed by administrators

• Some worries for the plan of using “tape as a disk”
  • Avoid non bulk access to tape library
  • Limiting as much as possible to write data that will be removed, since intense repack is a resource-consuming activity that could limit the overall performances.
  • To be discussed and tested
Miscellaneous issues (3)

• Support
  • Resources and users/experiments increasing at “flat staff”
    • This consideration has driven the choice of our storage model!
  • Largest operational burden from non-WLCG experiments
  • Effort to fit all experiments in the same model (i.e. WLCG)
    • Not an easy task!
      • “Why cannot we use simple tools like rsync?” “No certificates please! Could we use username and password?” “Should we really use SRM?”
      • “Experienced” users invoke the Holy Graal: “We need cloud” 😊

• Access
  • “no SRM, no tape”
    • Small collaboration w/o SRM send us list of files to be recalled…..
    • Standard alternative?
Farm remote extensions (1)

• Some functional tests on cloud providers (Aruba, Azure)
  • No cache, xrootd access

• In 2017 ~13% of CPU pledged resources to WLCG experiments located in Bari-RECAS data center
  • Transparent access for WLCG experiments
    • CNAF CEs and LSF as entry-point
    • Auxiliary services (i.e. squids) in Bari
  • Similar to CERN/Wigner extension
  • 20 Gbps VPN provided by GARR
    • All traffic with farm in Bari routed via CNAF
  • Disk cache provided via GPFS-AFM
    • “Transparent” extension of CNAF GPFS
Data access in Bari-RECAS

- **GPFS AFM**
  - A cache providing geographic replica of a file system
  - Manages RW access to cache
- **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
- **AFM configured as RO for Bari-ReCaS**
  - ~400 TB of cache vs. ~11 PB of data
- **Several tunings and reconfigurations required!**
- **In any case decided to avoid submission of high throughput jobs in Bari (possible for Atlas)**
- **Alice jobs access data directly through XrootD**
Farm remote extensions (2)

- In 2018 ~180 kHS06 provided by CINECA
  - CINECA, located in Bologna too, is the Italian supercomputing center (~15 Km far from CNAF)
  - 216 WNs (10 Gbit connection to rack switch and then 4x40 to router aggregator)

- Dedicated fiber directly connecting Tier-1 core switches to our aggregation router at CINECA
  - 500 Gbps (upgradable to 1.2 Tbps) on a single fiber couple via Infinera DCI

- No disk cache, direct access to CNAF storage
  - Quasi-LAN situation (RTT: 0.48 ms vs. 0.28 ms on LAN)

- In production since March
  - Need to disentangle effects from migration to CentOs7, singularity etc… to have a definitive assessment on efficiency

Jun 5, 2018
With this new type of hybrid devices (Packet/DWDM) it is possible to define on demand true high bandwidth circuits.
Possible use case for SD-WAN: Data Lake

According to our understanding, with data lake model:

- Fewer data replicas
- Fewer costs for storage hw and support (CPU only sites, opportunistic sites)
  - We need (probably) some smart caching system (XDC?)
  - WAN connectivity!
A very synthetic summary of INFN computing

- 29 data centers
  - 1 very large center (INFN Tier-1) and ~10 ones (mainly Tier-2s)
  - Excepting dedicated systems (e.g. experiments acquisition systems, online and trigger)
- ~90 FTEs on computing (~60 FTEs for technical support)
  - But ~30% of the staff is not INFN!
- CPU: ~70,000 cores (~800 kHS06)
  - Some HPC farms too (~400 Tflops)
- Disk: ~60 PB storage disk on line
  - Several flavors (StoRM, DPM, dCache, vanilla GPFS, Lustre, etc....)
- Tape: ~100 PB storage tape available (~70 PB used)
Plans and tests at INFN sites

• National caching infrastructure based on XroootD (XDC, CMS)
  • Pilot test in preparation among CNAF, Bari and Pisa

• HTTP-Based Caches (XDC)
  • To support site extension to remote locations or to use cloud/HPC diskless sites
  • Test-bed at CNAF
  • Plan to extend to a regional level

• Plans for a distributed storage system (+ cache) based on DPM
  • Involved (Atlas) sites: Napoli, Frascati, Roma
INFN Tier-1 short term evolution

• We need a simplified data management interface
  • Some attempt to mask X.509 (e.g. dataclient, an home-made wrapper to globus-url-copy)
  • Also OneData under evaluation (XDC and HSN projects) to ease access from users

• Add token based authz to storage services (i.e. by-pass the bias against certificates)
  • For the moment only a pre-production service with webdav + IAM for small community (Cultural Heritage)
  • Working on StoRM2

• Standard interfaces
  • Storm+nextcloud integration
    • Smart (i.e. easy) interface (also) for recall for non SRM users

• Tape recall optimization
  • Dynamic drives allocation
INFN Tier-1 “vision” plans

- Strong preference to use standard protocols for Data Access, transfer, AAI etc..
  - Replacement of SRM?
  - http/webdav?
  - Token based Authz
- Exploitation of remote extensions of our data center
  - Opportunistic too
  - CNAF could be part and offer storage to a future INFN cloud
  - Understand which type of cache could help
  - Possibly have at infrastructure level some functionalities now at application level
    - i.e. data replication, QoS?, self-healing?
- CNAF is definitely interested in participating to tests for the data lake
- CNAF will be probably part of several data lakes
  - Not only WLCG!
Backup slides
Tests with XrootD caches

- Working on creating a national caching infrastructure
  - Effort common to CMS and XDC

- **Objective: to deploy a national level cache**
  - geographically distributed cache servers
  - heterogeneous resources and providers
  - Leverage national networking to optimize the total maintained storage resources

- Collection of important data for evaluating the benefits on a realistic scenario

*Credits: D. Ciangottini, D.Spiga, T.Boccali, A.Falabella, G.Donvito – CMS and XDC*

Starting with sites with homogeneous resources (gpf-storm). Then extending to other sites (e.g. Legnaro) on a second step.
Local site scenario with XROOTD proxy cache

Credits: D. Ciangottini, D.Spiga, T.Boccali – CMS and XDC

- Create a **cache layer near cpu resources**
- Bring it up **on demand**
- **Scale horizontally**
- **Federate caches in a content-aware manner**
  - redirect client to the cache that currently have file on disk

/stash/myarea/file.root
Distributed scenario with XROOTD CACHE redirector

Credits: D. Ciangottini, D.Spiga, T.Boccali – CMS and XDC

**Geographically distributed cache**

- The very same technology used on local scenario can be **geo-distributed**
- Use **ephemeral storages** to enhance jobs efficiency
- Leverage high speed links to **reduce the total amount of allocated space**
HTTP-Based Caches

- In collaboration with XDC
Our scenario

● Explore distributed storage evolution to improve overall costs (storage and ops) taking in account:
  ○ Single common namespace and interoperability.
  ○ User analysis is often based on clusters hosted on medium sites (Tier2) and small sites (Tier3).

● In order to reconcile these two trends, the target of my activity is to study a distributed storage system featuring a single access point to large permanent storage and capable to provide efficient and dynamic access to the data. In this view, medium sites like Tier2 and small sites like Tier3 will not necessarily require large storage systems, simplifying local management.

  This can be achieved by the adoption of a distributed storage and caching technologies.

● This activity takes place in the same context of the Data Lake project having very similar motivations.
Our implementation

- The Disk Pool Manager (DPM) is a data management solution widely used within ATLAS, in particular in three Italian Tier2.
- The latest versions of DPM are used in our implementation, that offer the possibility to manage volatiles pools to be used as caches.

By exploiting the fast connections between sites, we are deploying a first testbed among Naples, Frascati and Roma-1 using DPM. The aim is to study and develop a configuration in which a primary site represents a single entry point for the entire archiving system and each site can use its storage as permanent storage or as local cache.

Using a cache system the local site administrators can be dispensed from managing a complete storage system. The site became transparent for the central operations of the experiment.
Conclusions

- A first testbed using DPM among Naples, Frascati and Roma-1 is almost ready.
- Study of the best caching policy for the volatile pools.
- Evaluation of the performance of the developed prototype.
- System integration in the current ATLAS data management infrastructures.
- Synergies:
  - collaborations with the Naples BELLEII computing group (Silvio Pardi (INFN-NA), Davide Michelino (GARR))
  - collaborations with the DPM development group.
- Create conditions for easy replication of the system on other sites or in other contexts.