CMS: T2 dynamic data placement and plans for T2 disk space handling

2013 WLCG Collaboration Workshop
11. November 2013

Oliver Gutsche
for
CMS Computing

Introduction

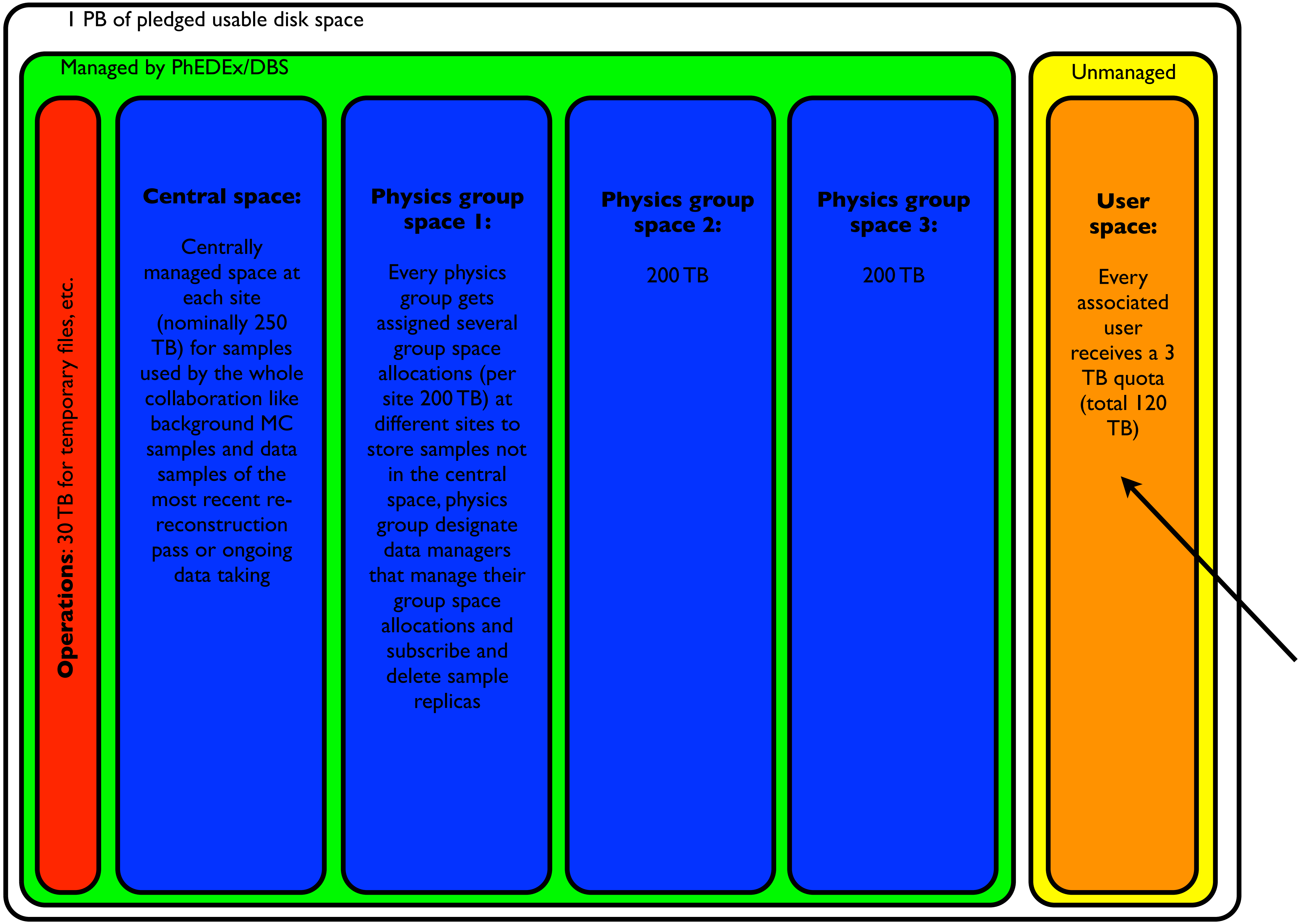
- CMS plans to optimize its disk usage on the Tier-1 and Tier-2 level
  - Analysis has shown that sizable fractions of samples stored on the Tier-2 level are not accessed at all or only accessed once
- CMS plans to implement dynamic data placement and automatic cache release following Atlas' example
  - At the same time, CMS will propose to the CMS physics groups to simplify the current setup which is manpower intensive and does not reward efficient usage of disk space

Current setup

- T1 sites disk is integrated into MSS, no direct control
- T2 disk is used by both the CMS data management system (PhEDEx, managed space) and local users storing their files (not tracked, unmanaged space):
  - Users are associated to sites (site A provides user space for user X). Unmanaged space is under the control of local site data managers who work with their associated users to stay within the storage possibilities of the site.
  - Managed disk space is separated logically in space allocation for different functions: central, physics groups, operations

T2 managed space allocations

- Example site: 1 PB of pledged usable disk space, 40 associated users



1 PB of pledged usable disk space

Managed by PhEDEx/DBS

Operations: 30 TB for temporary files, etc.

Central space:

Centrally managed space at each site (nominally 250 TB) for samples used by the whole collaboration like background MC samples and data samples of the most recent re-reconstruction pass or ongoing data taking

Physics group space 1:

Every physics group gets assigned several group space allocations (per site 200 TB) at different sites to store samples not in the central space, physics group designate data managers that manage their group space allocations and subscribe and delete sample replicas

Physics group space 2:

200 TB

Physics group space 3:

200 TB

Unmanaged

User space:

Every associated user receives a 3 TB quota (total 120 TB)

The different geographical regions are providing user space (recommendation is 2-4 TB per user) for the regional user community (usually organized in countries, if a country with CMS collaborators does not have sufficient T2 capacity, needs get satisfied by close-by T2 sites)

Disadvantages of current setup

- T1 disk is inaccessible for analysis
- T2 physics group space is manpower intensive to manage and leads to inefficient data placement with samples stored at T2 sites never or rarely accessed
- Experience from LHC run 1 also showed that there is increased demand for user space

New setup

- T1 disk is separated from tape, see previous talk
  - CMS can open Tier-1 resources for analysis when not used by central workflows

- T2 disk:
  - Resolve physics groups and other allocation and simplify space management:
  - Operations allocation of 30 TB will remain
  - 60% of remaining disk at every T2 site is centrally managed
  - 40% of remaining disk is regionally managed providing user space

- Use automated systems to initiate distribution of samples via PhEDEx and automated systems to release caches again
  - Based on popularity information of samples

Popularity

CMS POPULARITY

VICTOR LE NETTOYEUR

CERN IT ES

- CMS' popularity service and cache release was developed by CERN IT-ES based on ideas from Atlas
- Currently tracks file usage from
  - CMS analysis system CRAB
  - Xrootd monitoring (currently only for CERN)
- **Popularity is the center piece of automation!**
- To be added:
  - Global Xrootd monitoring tracking
  - Recent CMSSW releases are also instrumented to feed information into the popularity system while accessing files (covers also interactive access and access outside the CMS analysis system)

Automatic cache release

CMS

CMS POPULARITY

VICTOR LE NETTOYEUR

CERN IT ES

- When the popularity of a sample decreases, caches are released in the following order
  - First reduce the number of replicas on Tier-2 level
  - Followed by removing all replicas on Tier-2 level
  - Followed by removing the replica on Tier-1 level (sample is still on tape)

- Status:
  - Currently evolving and optimizing the cache release algorithm, first successfully used for EOS at CERN (T2_CH_CERN) followed by manual deletion requests
- Plans:
  - Allow for multiple cache release algorithms

Dynamic data placement

- When a new analysis format (AOD) sample is produced, it is pre-placed by the operations teams
  - Current proposal: at least at one Tier-1 on disk, and at one or two different Tier-2 sites
  - Evolution: enable clever pre-placement using for example past experience, seasonal conferences and other special events, hot analyses topics, etc.

- When the demand for a sample (popularity) increases (measured through popularity service and analysis job queue depth)
  - Additional replicas will be made automatically
  - Sites are intelligently chosen taking into account current usage of disk and CPU at the sites

- Dynamic data placement will be also tightly coupled to AAA and the CMS data federation
  - Algorithm has to be advanced to detect if a sample is better accessed over the WAN rather than replicated

Technicalities

- Recovery of samples that have been released from disk completely
  - Check automatically analysis queues and user requests (analysis system includes data discovery step which would come back negative if a requested sample in not on disk anymore)
  - Re-stage sample to disk at Tier-1 site and let dynamic data placement take over
  - Allow also for user requests to re-stage sample to disk at Tier-1 site
- Distribution of samples will be tightly coupled to site readiness information
  - A working site is prerequisite to allow for dynamic data placement
  - If analysis access and transfers to site do not work reliably, dynamic data placement cannot allow for samples to be placed at the site.
  - This goes hand in hand with CMS push to invest more into site readiness and evolve the site readiness tests and definition.
- Auto-approval of PhEDEx requests made through automatic cache release and dynamic data placement
  - To reduce latency of execution of placement and cache release decisions by the systems, we ask the sites to be allowed to auto approve the PhEDEx requests made by the systems

Moving forward

- Automatic cache release has to work first
  - Will start from the already existing implementation of Victor for CMS
  - Work started refining the cache release algorithm, first tests on EOS at CERN to keep the physics groups honest and obey their quotas was performed in manual mode
  - Need work to couple Victor to PhEDEx for automatic requests
- Next step: transform the already existing central space to be handled by dynamic data placement
- Lots of work will need to go into the development of the algorithms
- Next step: In coordination with physics groups, we will resolve the physics group allocations and have dynamic data placement use the full centrally managed disk space allocation

Conclusion

- This is not a presentation of the final setup of automatic cache release and dynamic data placement
  - It more defines the boundary conditions of an automated system to more efficiently use the disk space at the Tier-2 sites
  - We hope to get feedback from the sites while implementing and optimizing the system
- We hope that automatic cache release and dynamic data placement will be a big improvement to the current system
  - We need it fully functioning and tested at high scale by the begin of LHC run 2 in 2015