Distributed caching system for a multi-site DPM storage

G. Carlino[2], A. De Salvo[1], A. Doria[2], B. Spisso[2], E. Vilucchi[3]
[1]INFN - Roma1 Unit - Italy
[2]INFN - Napoli Unit -Italy
[3]INFN - Laboratori Nazionali di Frascati - Italy

Motivations

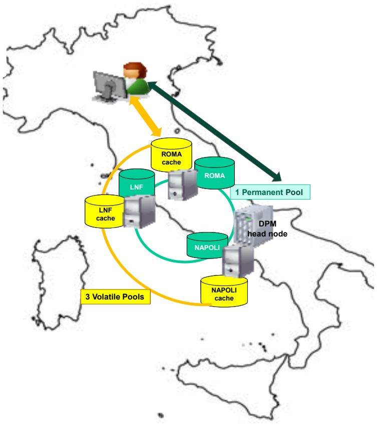
The LHC experiments, WLCG and funding agencies have started a process of optimization of the storage hardware and human resources needed for storage operations. The main keywords for this process are:

- **Common namespaces - Distributed storage and redundancy - Different QoS (storage media) - Geo-awareness - Caches**

- The DPM storage system is used since 2006 in 3 out of 4 ATLAS Tier2s in Italy. The DPM/DOME latest release has been used to verify how it fulfills some/all of the optimization requirements, making it compatible with future data model evolutions.

DPM volatile pools as local caches

- When a file is requested to a volatile pool for the first time, the pool retrieves it with a customized script and keeps it locally for the next usages.
- Complex retrival algorithms can be implemented and the file source can be another storage or any kind of Data Federation.
- **In our setup the cache interacts with Rucio Data Management to get any ATLAS file locally.**
- Different file sources could be used for other VOs



DPM multi-site setup

- The DPM head node is located in Naples .
- It's the only system front-end and manages 3 disk nodes located in Naples, in Rome and in Frascati.
- A common permanent pool includes storage areas from each disk node.
- 3 volatile pools, one per site.
- All system nodes have a 10Gbps link to the WAN.
- The file systems in Rome are built on CEPH

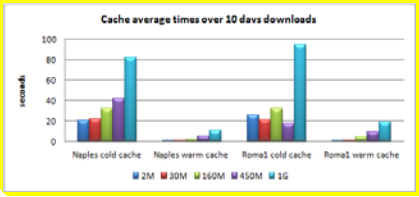
Use cases for our setup:

- Small diskless sites with a local cache, diskless Tier3s
  - Simplified local storage management.
  - Local users can benefit from cached data for thier analysis.
- Distributed storage with a single end-point
  - A single common namespace for remote and different storage media.
  - Simplified operations from the experiment point of view.

Cache tests

Each cache is accessible through the same front-end but with a specific path, that refers to the site. For example, to get a file from Roma cache:
davs://**t2-dpm-dome.na.infn.it**/dpm/**roma1.infn.it**/home/atlas/user.angianni/user.angianni.14404934._000001.CxAOD.root

At the first file retrival, the "cold" cache contacts Rucio to get a file replica from the grid and returns it to the requester.
Any other file access (warm cache) is local and about 10 times faster.



For very small files, the overhead of Rucio setup and retrieval prevails over the file transfer time. However it can't be reduced below 10 sec. With warm cache, the transfer times of 2M files are lower then a second.

Multi-site setup tests

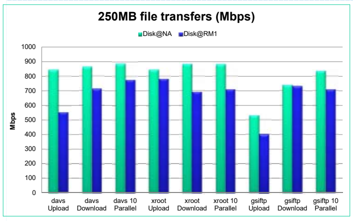
The distributed permanent pool was tested with 3 protocols: davs, xroot and gsiftp.

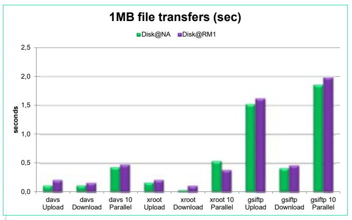
File upload, download and upload of 10 files in parallel threads were performed, with files of different sizes.

The different locations of head and disk nodes and the different underlying filesystems (CEPH @Roma, posix @ NA and LNF) might add an overhead in transfers.

In order to point it out, the tests were repeated with different setups:
- all the file systems enabled for RW,
- only Napoli (local) disks writable,
- only Roma (remote) disks writable.

The results of the two last cases are shown.





Conclusions

The feasibility of the proposed architecture and its functionality have been proven. The results of the transfer tests show that different storage hw (CEPH, Posix file systems) and geographical distributed storage areas don't affect significantly the data access and transfer. Any ATLAS file can be automatically retrieved in the cache, this could considerably improve local analysis on frequently used datasets.

Next steps of this study:
- Extend to larger infrastructures and to more heterogeneous storage media
- Perform stress tests in a production-like environment
- Test caches with real user analysis.
- Improve the integration with Rucio Data Management.

CHEP2018 – 23rd International Conference on Computing in High Energy and Nuclear Physics 2018 **9-13 July 2018, Sofia, Bulgaria**