Report from 2018 DPM workshop

Philippe Seraphin, Andrea Sartirana

8th DPM Workshop.

- **31st May - 1st June '18**, CESNET, **Praha** - Czech Republic
  - (very well)organized by **CESNET** and Institute of Physics of the Czech Academy of Sciences;
- ~15 participants (smaller wrt LPNHE '16 WS)
  - communities: FR (P.S. & A.S), IT, UK, CZ, CH. No much sites/communities reports but **focused on issues and technical items** (on dome in particular);
  - experiments: Atlas, CMS, Belle II;
  - themes: **1.10** release, **dome**, **space reporting**, **caching**;
- this is just a partial **summary** more info at
  - https://indico.cern.ch/event/699602/overview .

community/project status.

- **130 instances, 90 PB** (used to be 70PB)
  - lost a tail of small sites;
  - several larger than 2PB, 20 larger than 1PB, larger site is 6,5PB;
- very **active** and pro-active **community**
  - try to keep sysadmins cost low;
  - new wiki: https://twiki.cern.ch/twiki/bin/view/DPM/WebHome;
  - also moving from SVN to GIT;
- one **1.9.x** maint. release and one **1.10.x** feature release
  - focus on consolidation and scalability;
  - high level support of standard protocols: gsiftp, http, xrootd.

Disk Pool Manager

dpm 1.10.x

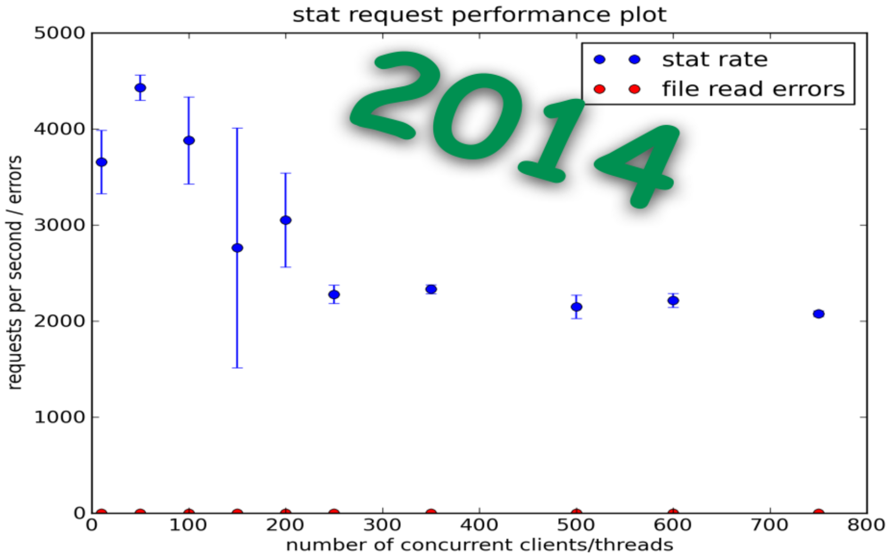
- In epel **since 06/2018**
  - already deployed at several sites (~27)
    - FR sites: GRIF (IRFU, LLR);
  - **no config changes** required for lcgdm (**legacy**) components
    - if you stay **in legacy mode** this is a **trivial upgrade**;
- quite some **changes in DOME** (even in legacy mode)
  - some **major changes/evolutions** (see next slides);
  - some **config changes**
    - host_dn => 'your headnode host cert DN';
    - token_password parameter MUST be now a string with more than 32 chars;
- more: https://twiki.cern.ch/twiki/bin/view/DPM/DpmSetup1100
  - https://indico.cern.ch/event/699602/contributions/2941791/attachments/1660057/2659148/EnablingDOME.pdf .

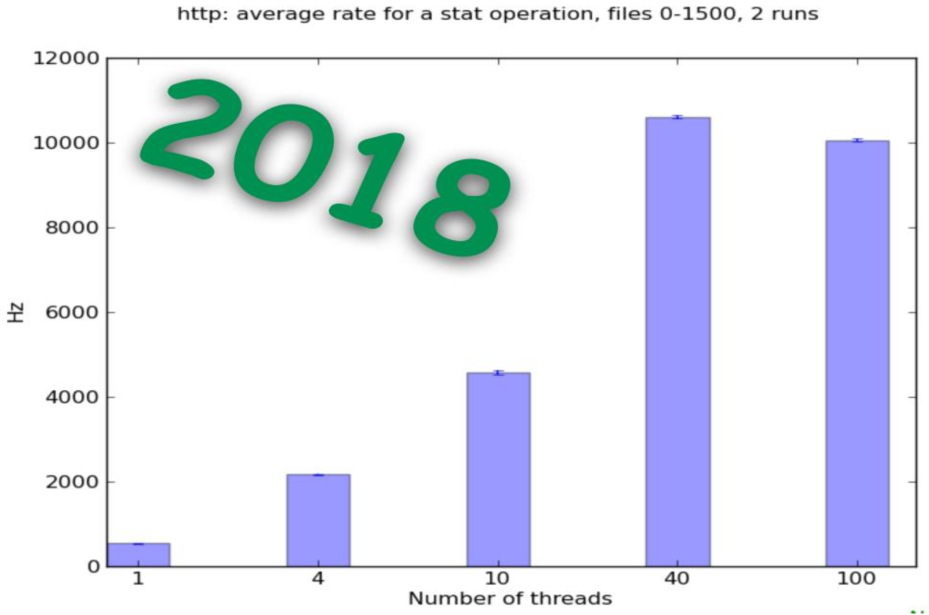
dome in 1.10.x...

- DOME and DOMEAdapter now have **all the primitives**
  - many tiny limitations have been removed, the behavior is very linear and has much less code;
- DMLite now loads only DOMEAdapter
  - **dmlite-memcache**, **dmlite-mysql** are **no** longer **necessary**;
  - X10 less res. consumption, complexity and cost for us all;
- **write-through metadata cache** (taken from Dynafed)
  - **eliminates** the strange race conditions that made the **memcache** plugin become complex and less effective (basically discarding its content way too often);
  - write-through: warm entries are never purged when modified;
  - internal thing no **need to touch/tune** its **parameters** so far.

…dome in 1.10.x

- ➢ From fastCGI to **XrdHttp for REST interface**
  - ❖ **in 1.9 fastCGI and apache.** But mod_fcgid disappeared and mod_proxy_fcgi has connection reuse **broken** (and will stay so)
    - ❑ perf blocked at 100Hz (almost worse than SRM);
  - ❖ **in 1.10 adopted XrdHttp.** HTTP/WebDAV impl. of Xrootd
    - ❑ much **better performances** more than 10KHz;
    - ❑ ready for SciTokens and HTTP third party copy;





migration to dome

- **stop LCGDM support from 01/Jun/2019 !!**
  - they will not fix it if it breaks
    - ❑ in EPEL as long as it compiles untouched;
  - this means passing to SRM-less operations!!
    - ❑ everything should be ready… more or less;
- start thinking **how to migrate**
  - first to **DOME mode with SRM** still there
    - ❑ can still use srm (no need for gftp redir)the tricky part is def. QT;
  - …**then** move to **SRM-less**
    - ❑ needs srm-less and a number of tools ready
    - ❑ …and to validate VO's workflows;
  - bunch of **pilot sites**
    - ❑ see **next talk**;

puppet setup

- modules and deps **via EPEL:** *dmlite-dpm-puppet pkg*
  - version **validated by developers**
    - installed in */usr/share/dmlite/puppet/modules* ;
    - also **still in puppet forge**;
  - for **quattor** (quappet) straightforward to use these modules
    - made PR to QWG. Already in use at LLR;

- **improvements** and **new features**
  - support both **puppet 4 and 5** (use pp 5);
  - support **CentOS7**;
  - fully **integrated with** the new **DOME**
    - implementing **lcgdm-free** configuration;
    - better **use last versions** to migrate to DOME;
  - https://twiki.cern.ch/twiki/bin/view/DPM/DpmSetupPuppetInstallation
  - new AAA conf, bug fixes, …

dmlite-shell in 1.10…

- **Tool for** dpm **administration**
  - for admins and devs **not for users**;
  - https://twiki.cern.ch/twiki/bin/view/DPM/DpmAdminDmliteShells ;
- **improvements**
  - add recursive option to acl command;
  - qryconf command reports the status of the FS (ONLINE, DOWN);
  - dmlite-shell does not just exit on simple ^C ;
  - add extension of the current folder to quotatoken* commands;
  - add print for quotatoken* commands (to restart DPM daemon);
  - add a check replica status before running draining;
  - add force parameter drain with more than 10 threads;
  - functional tests;

DPM
Disk Pool Manager

…dmlite-shell in 1.10

- Some **bug fixes**
  - fixing userid/groupids in chown and chgrp;
  - fix drain "last replica" problem;
  - fix replicadel behavior when multiple replicas;
  - qryinfo/poolinfo errors on an empty DPM without Legacy stack;
  - fixed return codes of some commands;
- warning: qryconf, fs* and quotatoken* **only work in DOME**
  - use the lcgdm equivalents in legacy mode;
- **To do:**
  - dpm-disk-to-dpns, dpm-list-disk, dpm-dpns-to-disk, dpm-dbck…;
  - recursion on more commands.

checksums

- Cksums are **fetched from the DB or calculated/verified**
  - **queueing** logic (same as the volatile pools)
    - □ no more than N retrievals per server;
    - □ no more than M retrievals overall;
  - **clients** peacefully/transparently **wait** their turn;
  - supports a number of checksum types;
  - makes checksum **storms safer**;
- **different protocols** supported
  - **HTTP/WebDAV** works **fine**;
  - **xrootd will**, at the next dpm-xrootd minor version;
  - **gridftp can't use DOME** for checksums
    - □ globus misses checksum callbacks in the DSI plugins;
    - □ gridftp remains vulnerable to checksum storms.

more stuff

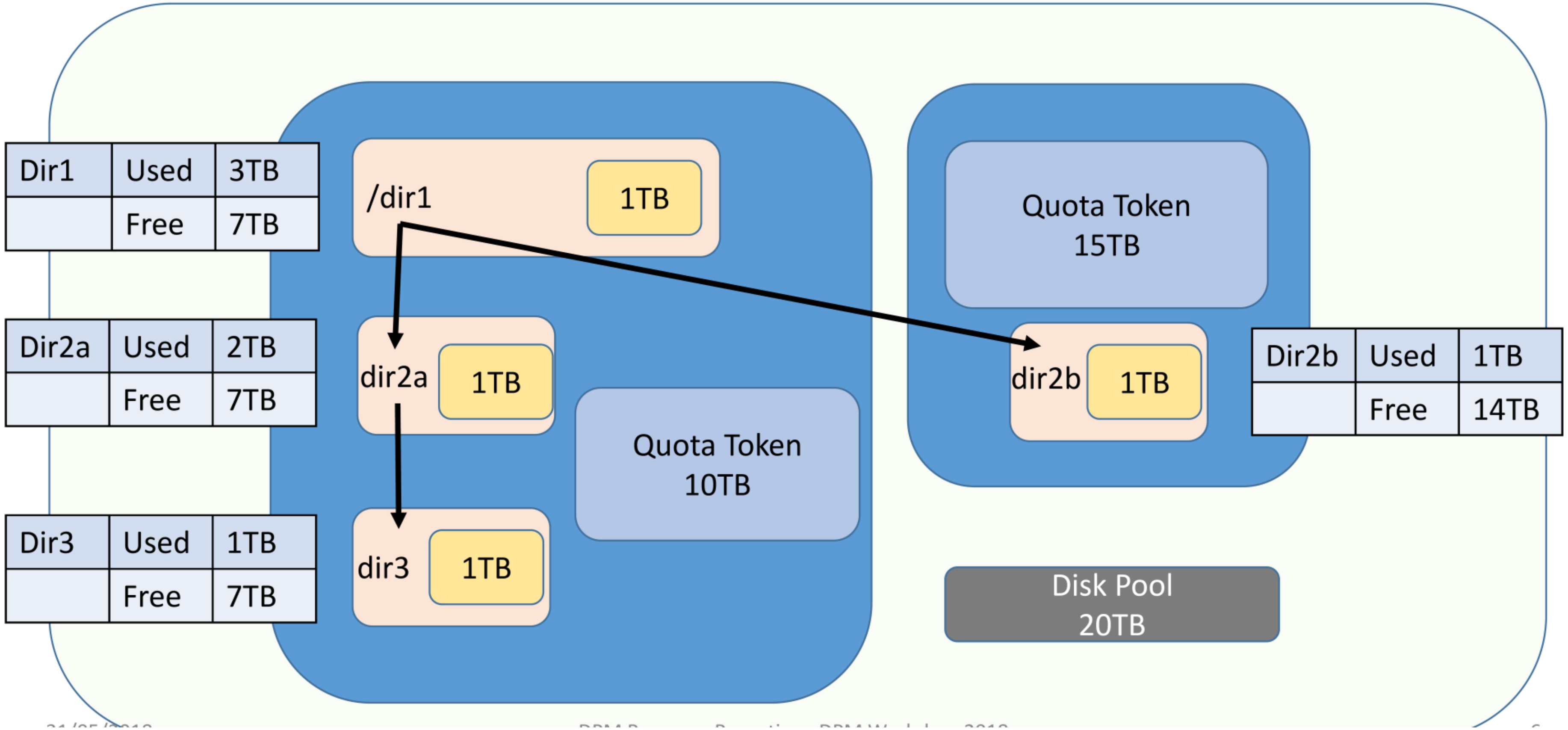
- **dpm-xrootd** release
  - next release will be **harmonized with the version of DMLite**;
  - this cuts the cost of managing the Fedora/EPEL releases;
  - consequence: the next dpm-xrootd version will be 1.11.x (epoch 2) for all the xrootd plugins that we provide
    - now it's 3.6.x! the rpm name will change;
- dmlite **C++ interface**
  - big **source of cost and not used** by other packages or components
    - sets many constraints that apply only to its authors;
    - difficult to evolve and simplify, due to ABI things… academical because the user are only the developers;
  - solution: **republish it as private headers**, to cut the unnecessary cost of a public C++ interface not designed to be public;
  - only C++… the C interface is just fine instead.

space reporting

- **WLCG Resource Reporting**
  - **recommendations** on storage reporting for accounting and exp. ops
    - https://twiki.cern.ch/twiki/bin/view/LCG/AccountingTaskForce#Storage_Space_Accounting;
    - https://docs.google.com/document/d/1yzCvKpxsbcQC5K9MyvXc-vBF1HGPBk4vhjw3MEXoXf8/edit?usp=sharing;
    - https://twiki.cern.ch/twiki/pub/LCG/AccountingTaskForce/storage_service_v4.txt;
  - allows resource reporting **in the absence of SRM**
    - used/free space for independent areas (Quota Tokens in DPM);
- **DPM** has **fully implemented** these
  - you need 1.10 DOME mode;
  - based on quota tokens accounting;

space reporting

- Get used/free space for QTs
  - via DAV (RFC 4331);
  - via a summary file called storagesummary.json in the namespace;



Dir1	Used	3TB
	Free	7TB

Dir2a	Used	2TB
	Free	7TB

Dir3	Used	1TB
	Free	7TB

Dir2b	Used	1TB
	Free	14TB

space reporting

- Get used/free space for QTs
  - **via DAV** (RFC 4331);
  - via a summary file called storagesummary.json in the namespace;

```
$ davix-http -P grid -X PROPFIND --header 'Depth: 0' --header 'Content-Type: text/xml;
charset=UTF-8' "https://domehead-trunk.cern.ch/dpm/cern.ch/home/dteam" --data '<?xml
version="1.0" ?><D:propfind xmlns:D="DAV:"><D:prop><D:quota-used-bytes/><D:quota-
available-bytes/></D:prop></D:propfind>’
```

```
<?xml version="1.0" encoding="utf-8"?>
<D:multistatus xmlns:D="DAV:" xmlns:ns0="DAV:">
<D:response xmlns:lp1="DAV:" xmlns:lp2="http://apache.org/dav/props/"
xmlns:lp3="LCGDM:">
<D:href>/dpm/cern.ch/home/dteam/</D:href>
<D:propstat>
<D:prop>
<lp1:quota-used-bytes>24677181319</lp1:quota-used-bytes>
<lp1:quota-available-bytes>75322818681</lp1:quota-available-bytes>
</D:prop>
<D:status>HTTP/1.1 200 OK</D:status>
</D:propstat>
</D:response>
</D:multistatus>
```

space reporting

- ➢ Get used/free space for QTs
  - ❖ via DAV (RFC 4331);
  - ❖ **via a summary file called storagesummary.json in the namespace;**

```
# cat /etc/cron.hourly/dpm-storage-summary
#!/bin/bash
/usr/bin/dpm-storage-summary.py --path /dpm/domain.org/home/dteam
/usr/bin/dpm-storage-summary.py --path /dpm/domain.org/home/atlas
```

- ❖ to be consumed by WLCG storage accounting and/or experiment portals (e.g. AGIS fro Atlas);
- ❖ contact dimitrios.christidis@cern.ch;
- ❖ https://monit-grafana.cern.ch/d/mHqFLAbik/wlcg-storage-space-accounting?orgId=20 .

namespace dump

- WLCG **also** made **recommendations on storage dumps**
- /usr/bin/dpm-dump has been updated to support these requirements
  - it should be **useable for all experiment** namespace dumps;
  - available in dpm-contrib-admintools package.

info provider

- **New** info provider for **SRM-less** DOME installations
  - **available with 1.10** but not yet ready
    - not attempting to mimic dpm-listspaces;
    - **glue-2.0** only;

- in /var/lib/bdii/gib/provider/dome-info-exec
  - available in the dmlite-shell package;
  - configure /etc/sysconfig/dpminfo
    - ensure DPM_INFO_PROVIDER=“dome”;
    - dpm-listspace can be invoked setting DOM_INFO_PROVIDER=“dpm-listspaces”;
  - remove the earlier info providers
    - /var/lib/bdii/gip/provider/se-dpm;
    - /var/lib/bdii/gip/provider/service-srm2.2;
    - there will be puppet support.

caching…

- **In DOME** define a **Volatile pool** to trigger cache behavior
  - works seamlessly with **http**, **xroot**, **gsiftp** (SRM not supported);
  - **files** not present are recovered **from an external source**
    - ❑ the client will wait;
  - **older** (ctime) files are **deleted if** space is **needed**;
- **setup** is simple
  - make the volatile pool and make sure the default files size > max size you want to cache;
  - **create a QT** and attribute it to your **caching path**;
  - put a **stat script on the HN** that stats the external file;
  - put a **pull script on the DS** that pulls the file;
  - more at https://twiki.cern.ch/twiki/bin/view/DPM/DpomSetupDpmCache .

…*caching*…

Scenarios

- **cache + primary storage**
  - satellite site can **accelerate access to a nearby custodial storage**;
  - this could allow a group of nearby sites to consolidate their storage;
- **cached access to a federation**
  - the **upstream server** can in fact be **a federator** such as Dynafed;
  - this would transparently **accelerate access to a federation**;

evolutions

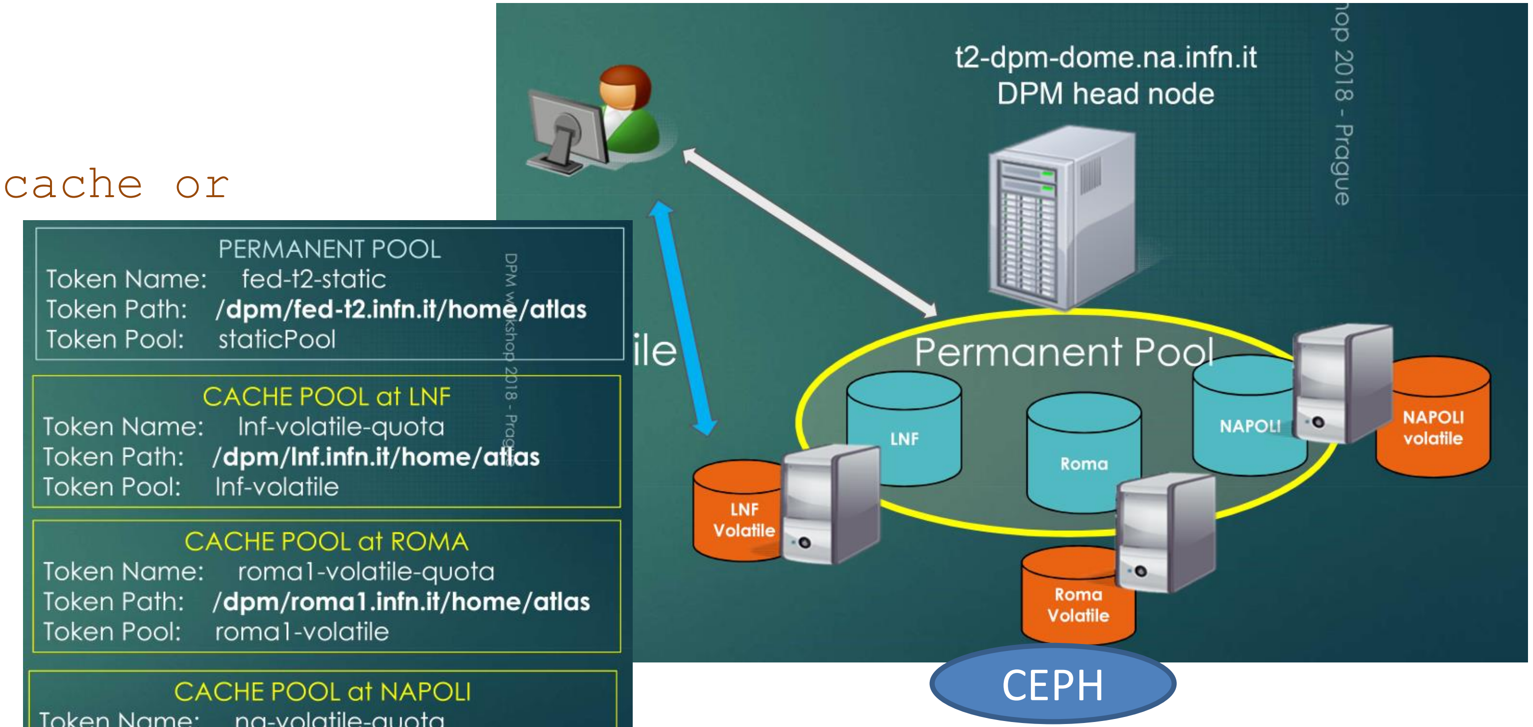
- **redirect clients** rather than blocking (lower latency/more WAN traffic);
- **federating the cache**
  - a federator always sees the cache as full;
  - if it redirects a client there, the pull is triggered.

https://indico.cern.ch/event/699602/contributions/2941779/attachments/1659518/2658569/DPMCache.pdf

INFN testbed for **distributed DPM and caching**

- ❖ permanent **distributed pool as data source**
  - ❑ the pull script make a **davix-get** of the file from the permanent pool;
  - ❑ **plan to interact with rucio** to get any ATLAS file in the cache;
- ❖ **different domains in path**
  - ❑ required bit of hacking;
  - ❑ different paths to local cache or permanent storage;
- ❖ open questions
  - ❑ **pinning/lifetime**;
  - ❑ test **NW requirements**;
  - ❑ test **CEPH vs local** FS.

https://indico.cern.ch/event/699602/contributions/2941786/attachments/1660250/2659490/Italy_DPM2_2018.pdf



PERMANENT POOL
Token Name: fed-t2-static
Token Path: **/dpm/fed-t2.infn.it/home/atlas**
Token Pool: staticPool

CACHE POOL at LNF
Token Name: lnf-volatile-quota
Token Path: **/dpm/lnf.infn.it/home/atlas**
Token Pool: lnf-volatile

CACHE POOL at ROMA
Token Name: roma1-volatile-quota
Token Path: **/dpm/roma1.infn.it/home/atlas**
Token Pool: roma1-volatile

CACHE POOL at NAPOLI
Token Name: na-volatile-quota
Token Path: **/dpm/na.infn.it/home/atlas**
Token Pool: na-volatile

…caching.

SCOReS project: **HTTP caching** for HEP. **Belle II** pilot VO

- **DynaFed** + volatile pool
  - file metalink always at least the real URL + the (virtual) cache copy;
  - **GeoPlugin** prioritize the cache copy if close to the Client;
- stat and pull **scripts**
  - **stat:** sees it is file/dir. Gets the size of the real file;
  - **pull:** if not in cache, downloads from grid. **Localizing it via Dynafed**;
- client gets 202 if file is not ready and waits n secs;
- made **performance** tests downloading and reading files
  - the solution seems to be **stable and well performing**;
- working on a **FilterPlugin** implementing a **cost function**
  - e.g. prioritizing cheaper S3 cloud storages.

Belle II

Consortium GARR

https://indico.cern.ch/event/699602/contributions/2953001/attachments/1660160/2659474/HTTP-Caching-01-06-2018.pdf

summary

- **1.10 released** just after the workshop in June 2018
  - major **DOME** refactoring: **XrdHTTP**, only domeadapter, internal cache, checksum, std ST writing for CMS;
  - other evolutions: **puppet** modules **in epel**, checksum, dmlite-shell…;
- speedup of the **migration to DOME** and to SRM-less
  - LCGDM EOL on June 2019;
  - SRM-less SRR and info-provides;
- evolutions/r&d's
  - **distributed pools**;
  - **storage caches**;
  - TPC beyond globus.