EOS, DPM and FTS developments and plans

Andrea Manzi - on behalf of the IT Storage Group, AD section

HEPIX Fall 2016 Workshop – LBNL
Outline

- CERN IT Storage group, AD section
- EOS
  - Namespace on Redis
- DPM
  - DOME
- FTS
  - New optimizer
  - Object Stores Integration
CERN IT-ST group, AD Section

• 16 members
• Main activities
  • Development
    • EOS, DPM, FTS, Data management clients (Davix, gfal2, Xrootd client)
  • Operations
    • FTS
  • Analytics WG
  • Effort in WLCG ops
Outline

- CERN IT Storage group, AD section
- EOS
  - Namespace on Redis
- DPM
  - DOME
- FTS
  - New optimizer
  - Object Stores Integration
**EOS architecture**

- Disk only file storage
- Designed for **Massive storage** with high performances ( > 50PB, 500 M files and KHz metadata operations)
- In memory hierarchical namespace
- File layouts (default 2 replicas)
- Low latency access
EOS World-Wide Deployment

CERN

ASGC

AARNET

EOS, DPM and FTS developments and plans

19/10/16
EOS releases and branches

- **Production version**
  - Branch: *beryl_aquamarine*
  - Release number: $\geq 0.3.210$

- **Development version** (master)
  - Branch: *citrine*
  - Release number: $\geq 4.1.4$
  - Requires *XRootD 4.4.0*

- **Feature branches** get merged into master e.g. kinetic, geo-scheduling, namespace devel. etc.
What is the EOS namespace?

- C++ library used by the EOS MGM node single-threaded
- Provides API for dealing with hierarchical collections of files

**Filesystem elements**
- Containers & files

**Views**
- Aggregate info about filesystem elem.
- E.g QuotaView, FileSystemView etc.

**Persistence objects**
- Objects responsible for reading and storing filesystem elements
- Implemented as binary change-logs
Namespace architectures: pros/cons

**Pros:**
- Using hashes all in memory $\rightarrow$ extremely fast
- Every change is logged $\rightarrow$ low risk of data loss
- Views rebuilt at each boot $\rightarrow$ high consistency

**Cons:**
- For big instances it requires a lot of RAM
- Booting the namespace from the change-log takes long
EOS Namespace Interface

- Prepare the setting for different namespace implementations
- Abstract a **Namespace Interface** to avoid modifying other parts of the code

- **EOS citrine 4.***
  - **Plugin manager** – able not only to dynamically load but also stack plugins if necessary
  - **libEosNsInMemory.so** – the original in-memory namespace implementation
  - **libEosNsOnFilesystem.so** – not existing based on a Linux filesystem
Why Redis?

- **Redis** – in-memory **data structure store**
- Separate data from the application logic and user interface
- Supports various data structures: strings, hashes, lists, sets, sorted sets etc.

- Namespace implementation: **libEosOnRedis.so**
XRootD and Redis

- Replace Redis backend with XRootD
- Implemented as an XRootD protocol plugin – to be contributed upstream
- XRootD can use RocksDB as persistent key-value store
Namespace HA

• Ensure high-availability using the Raft consensus algorithm
Outline

• CERN IT Storage group, AD section
• EOS
  • Namespace on Redis
• DPM
  • DOME
• FTS
  • New optimizer
  • Object stores Integration
Disk Pool Manager

- DPM is a system for managing disk storage at small and medium sites
- Over 70PB stored in the system at around 150 sites
- DPM has just finished a little revolution.
  - We now have our definitive platform for the future
    - SRM-less operations
    - Caching
The 3rd Generation DPM

• **1st generation** – derived from Castor

• **2nd generation** – introduced dmlite
  • Internal framework abstracting many functions, enabling multiple new frontends

• **3rd generation** – introduces Dome
  • DPM 1.9 “Legacy flavour”
    • Legacy services still running
  • DPM 1.9 “Dome flavour”
    • Legacy services retired
    • No more dpmd, dpns, rfio, srm, csec
DPM Headnode architecture

Legacy daemons

- dpns
- rfio
- srm
- CSec
- dpm (daemon)

Communication with DPMd can now be dropped

DMLite section

- gsiftp
- WebDAV
- Xrootd
- dmlite core
- mem cache
- profiler
- mysql
- DOME
- DOME Adapter

Legacy clients

DOME Adapter
DPM 1.9 “Dome flavour”

• Activate Dome, what do you get?
  • SRM free operation
    • Quotas
      • “space token” concept generalised and mapped into namespace
  • Space reporting – used/free via HTTP/DAV
    • Reporting on “space tokens” and subdirectories
  • GridFTP redirection enables scalability
• Caching hooks
• Simplified system
  • All internal communication over HTTP
    • Control and data tunneling
• Improved dmlite-shell
Disk server (simplified)

DOME
- Timed Logic
- Workers
- AuthN
- Request logic
- Task Executor
- Dome_checksum Script (bundled with DOME)

Apache HTTPd
- mod_proxy_fcgi
- Dmlite:
  - IO
  - DiskCatalog
- lcgdm_dav

xrootd

gsiftp

Disks

Requests to HEAD

Ext Pull() requests (configurable script)

/Domedisk/

/<diskpaths>/

Dome can do IO tunnelling if clients don’t support redirection
Used/free space: WebDAV

Going SRM-less with your DPM

1.9.0 in EPEL

Site updates the system
ADM Configures quotatokens AND primes the space calcs
DmLite::DomeAdapter substitutes Adapter (config change)
Experiments gradually stop using SRM
Load on SRM goes to zero

Sysadmin
Experiments
DPM team
Q4/2016

...continuous support...

DPM/dpns/rfio still used (through DMLite::Adapter and SRM)
DOME used only for dmlite-shell
New powerful dmlite-shell is available
Can report used space per path
Free space reporting is like before (Only SRM can report on free spc)

All the DOME features are enabled
WebDAV reports used AND free spc per path
Quotas are properly enforced
Can create the ATLAS space report
SRM/DPM still report on spacetokens
SRM/DPM/dpns/rfio are now optional

DPM/dpns/rfio/SRM can be uninstalled

Good moment for a minor release
Caching laboratory

- DPM 1.9 with Dome will allow investigation of operating **WLCG storage as a cache**
- Scenarios
  - Data origin a local federation of associate sites
  - Data origin the global federation
  - Hybrid cache/conventional setup
- **A volatile pool** can be defined which calls out to a stager on a miss
  - Caching logic implemented in a pluggable way
- Questions to investigate
  - Cache management logic
  - Different client strategies on miss
    - blocking read, async read, redirection to origin
  - Authentication solutions
  - Workflow adaptation for locality

19/10/16 EOS, DPM and FTS developments and plans
Outline

• CERN IT Storage group, AD section
• EOS
  • Namespace
• DPM
  • DOME
• FTS
  • New optimizer
  • Object stores Integration
**FTS**

- **FTS** is the service responsible for distributing the majority of LHC data across the WLCG infrastructure.
- It is a low level data movement service, responsible for moving sets of files from one site to another while allowing participating sites to control the network resource usage.
- **WLCG stats:**
  - Installed at: CERN, RAL, BNL, FNAL
  - ~20PB monthly volume / ~2.2M files per day
FTS Architecture

MySQL

Server
Bring online
REST

Server
Bring online
REST

Optimizer
Scheduler

FTS-Url-Copy
Gfal2 plugins

...
FTS 3.5: Optimizer changes

- **Min/max active ranges** can be configured per link
- Improved **throughput calculation**
- Softener *Exponential Moving Average* to reduce noise effects
- Throughput taken into account even with high success rates
- **1 stream per transfer by default**
  - Reduces resource consumption on the storages
FTS 3.5: Optimizer evolution
Plans for FTS 3.6

- Targeted towards Jan 2017
- **Remove SOAP**
  - Only CMS is using it and it’s already migrating to REST
- **Remove Bulk Deletions**
- New algorithm for **multiple replicas jobs**
- Database profiling and optimizations
Object stores Integration

- Advantages
  - **Scalability** and performance achieved through relaxing or abandoning many aspects of posix
  - Applications must be aware or adapted
- How can such resources be plugged into existing WLCG workflows?
  - Can apply to public or private cloud
    - NB ceph at sites
- Data transfer -> FTS integration via davix/gfal2
davix

- `davix-put /etc/services`
  `https://objbkt1.s3.amazonaws.com/file01 --s3secretkey <secret> --s3accesskey <access>`

- `davix-cp -P grid`
  `davs://dpm.cern.ch/dpm/cern.ch/home/dteam/file01`
  `s3s://objbkt1.s3.amazonaws.com/file01 --s3secretkey <secret> --s3accesskey <access>`

3rd party copy!!
gfal2/davix

- gfal-copy file:///etc/services
  s3://objbkt1.s3.amazonaws.com/file01

- gfal-copy
davs://dpm.cern.ch/dpm/cern.ch/home/dteam/file01 s3://objbkt1.s3.amazonaws.com/file01

3rd party copy!!
FTS: Pre-signed URL

fts-transfer-submit --strict-copy -s
https://fts3.cern.ch:8446
https://dpm.cern.ch/dpm/cern.ch/home/dteam/file01
'https://objbkt1.s3.amazonaws.com/tf_04?Signature=eFAyXMWISY%2BWEVcqfvGvuxZF6ZQ%3D&Expires=2105774242&AWSAccessKeyId=A
KIAJZZQ2TYSEBKNVWKA'
FTS: key management

You can also allow FTS to hold the keys to your cloud storage

```
$ curl[...] https://fts3devel01.cern.ch:8446/config/cloud_storage -H "Content-Type: application/json" -X POST -d '{"storage_name":"S3:s3.domain.com"}'

$ curl[...]"https://fts3devel01.cern.ch:8446/config/cloud_storage/S3:s3.domain.com" -H "Content-Type: application/json" -X POST -d "$\{config\}"

```
FTS: transport

- Solutions for import to and export from clouds
  - Several S3 variants supported
- Various architectures possible
  - FTS gateway
    - SRM<->S3
  - 3rd party transfer
  - Multi-hop with tactical storage
References

• EOS
  • http://eos.web.cern.ch/

• DPM
  • http://lcgdm.web.cern.ch/
  • DPM Workshop 2016, 23-24 Nov, Paris
    • https://indico.cern.ch/event/559673/

• FTS
  • http://fts3-service.web.cern.ch/