Multi-site DPM – The Bern case

Gianfranco Sciacca

AEC - Laboratory for High Energy Physics, University of Bern, Switzerland

DPM Workshop 2019 - University of Bern - 14 June 2019

UNIVERSITÄT BERN
AEC
ALBERT EINSTEIN CENTER
FOR FUNDAMENTAL PHYSICS

LABORATORIUM FÜR HOCHENERGIEPHYSIK
LHEP
UNIVERSITÄT BERN

OUTLINE

- Introduction
- Motivation
- Layout
- Configuration

INTRODUCTION

WLCG STORAGE IN SWITZERLAND

Tier-2 storage

- CSCS-LCG2: **dCache** (ATLAS, CMS, LHCb)
- UNIBE-LHEP: **DPM** (ATLAS)

Tier-3 storage

- T3_CH_PSI: **dCache** (CMS)
- UNIGE-DPNC: **DPM** (ATLAS)

MOTIVATIONS

- **Loss of funding / manpower at Geneva**
  - ATLAS_LOCALGROUPDISK heavily used by local T3 users at DPNC
  - ATLAS_DATADISK used for the ATLAS Production jobs (PanDA queue)
  - The DPM ran unattended for several months

- **Pressure from ATLAS to reduce number of storage endpoints**
  - Well, ***one*** or ***two*** doesn't really make much difference … *still* … :-)

  - **Geneva T3 investigating a lightweight storage solution**
    - Users need to keep working meanwhile

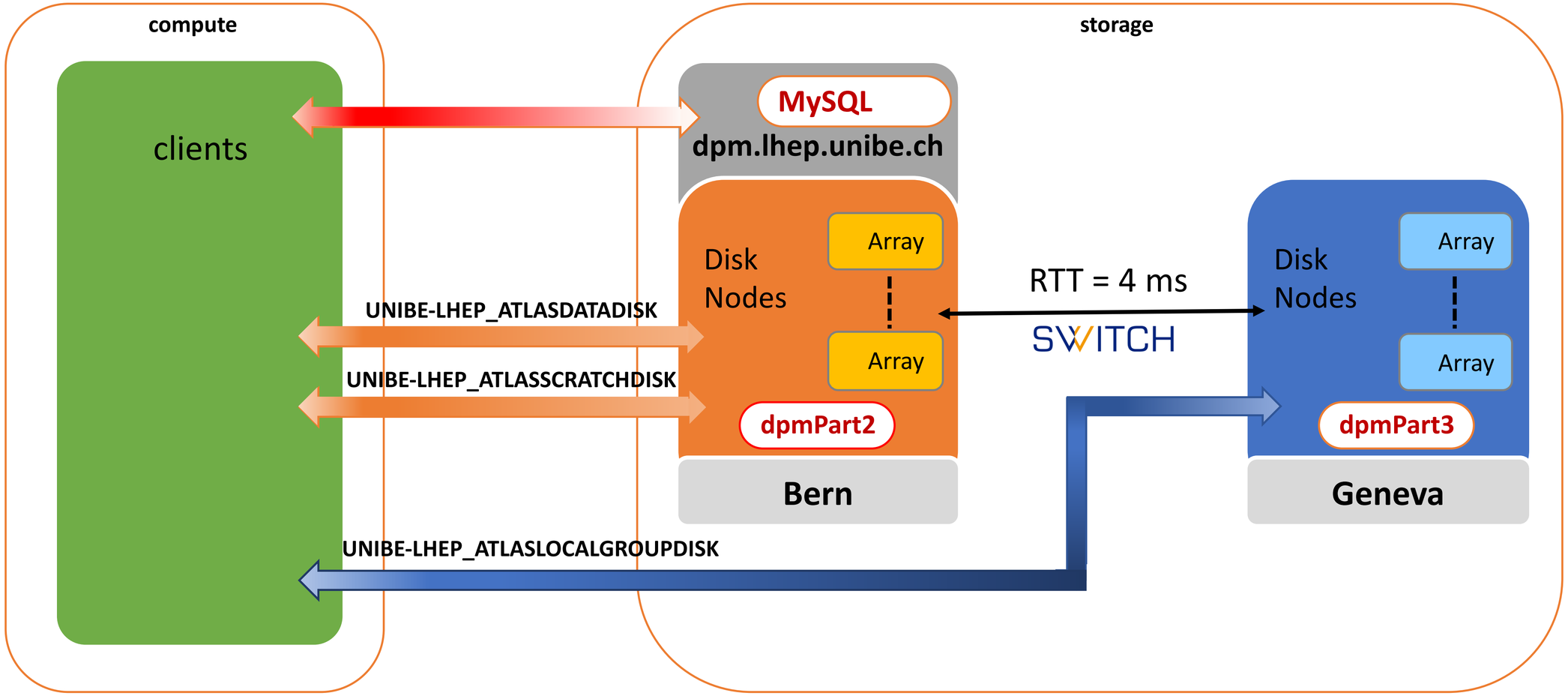
INTER-SITE AGREEMENT

- **We agreed to try running Geneva's storage pools under the Bern head node**
  - Geneva run the hardware
    - Server provisioning and maintenance (failed disks, etc.)
    - Network, firewall …
    - OS updates, security …
    - Host certificates
  - Bern in charge of the DPM stack
    - *root access by ssh key granted from one Bern IP address*

INTER-SITE AGREEMENT

- **We agreed to try running Geneva's storage pools under the Bern head node**
  - Geneva run the hardware
    - Server provisioning and maintenance (bad disks, etc.)
    - Network, firewall …
    - OS updates, security …
    - Host certificates
  - Bern in charge of the DPM stack
    - *root access by ssh key granted from one Bern IP address*
- **Make a common** ATLAS_LOCALGROUPDISK **area available to T3 users**
- **Dismiss the** Geneva ATLAS_DATADISK **and bind the Geneva compute resources to the** Bern ATLAS_DATADISK (in AGIS)

DPM LAYOUT


compute
storage
clients
MySQL
dpm.lhep.unibe.ch
Disk
Nodes
Array
Array
dpmPart2
Bern
RTT = 4 ms
SWITCH
Disk
Nodes
Array
Array
dpmPart3
Geneva
UNIBE-LHEP_ATLASDATADISK
UNIBE-LHEP_ATLASSCRATCHDISK
UNIBE-LHEP_ATLASLOCALGROUPDISK

CONFIGURATION

BEFORE

```
[root@dpm ~]# dpm-qryconf
POOL dpmPart1 DEFSIZE 100.00M GC_START_THRESH 0 GC_STOP_THRESH 0 DEF_LIFETIME 7.0d DEFPINTIME 2.0h MAX_LIFETIME 1.0m MAXPINTIME 12.0h FSS_POLICY
maxfreespace GC_POLICY lru RS_POLICY fifo GIDS 102,103,104,105,106,107,112,113,141,233,234,235,239 S_TYPE - MIG_POLICY none RET_POLICY R
                           CAPACITY 0 FREE 33.44T (  0.0%)
  dpm.lhep.unibe.ch /mnt/storage1 CAPACITY 9.99G FREE 9.96G ( 99.7%) RDONLY
  dpm.lhep.unibe.ch /mnt/storage2 CAPACITY 99.95G FREE 93.65G ( 93.7%) RDONLY
POOL dpmPart2 DEFSIZE 100 GC_START_THRESH 0 GC_STOP_THRESH 0 DEF_LIFETIME 7.0d DEFPINTIME 2.0h MAX_LIFETIME 1.0m MAXPINTIME 12.0h FSS_POLICY
maxfreespace GC_POLICY lru RS_POLICY fifo GIDS 101,102,103,104,105,106,107,109,112,113,114,115,117 S_TYPE - MIG_POLICY none RET_POLICY R
                           CAPACITY 571.12T FREE 32.74T (  5.7%)
  dpmdisk01.lhep.unibe.ch /mnt/storage1 CAPACITY 18.19T FREE 395.31G (  2.1%)
  dpmdisk01.lhep.unibe.ch /mnt/storage2 CAPACITY 18.19T FREE 372.61G (  2.0%)
  dpmdisk01.lhep.unibe.ch /mnt/storage3 CAPACITY 18.19T FREE 376.76G (  2.0%)
  dpmdisk02.lhep.unibe.ch /mnt/storage1 CAPACITY 18.19T FREE 372.52G (  2.0%)
  dpmdisk02.lhep.unibe.ch /mnt/storage2 CAPACITY 18.19T FREE 372.52G (  2.0%)
  dpmdisk02.lhep.unibe.ch /mnt/storage3 CAPACITY 18.19T FREE 389.66G (  2.1%)
  dpmdisk03.lhep.unibe.ch /mnt/storage1 CAPACITY 18.19T FREE 372.69G (  2.0%)
  dpmdisk03.lhep.unibe.ch /mnt/storage2 CAPACITY 18.19T FREE 372.49G (  2.0%)
  dpmdisk03.lhep.unibe.ch /mnt/storage3 CAPACITY 18.19T FREE 374.15G (  2.0%)
  dpmdisk04.lhep.unibe.ch /mnt/storage1 CAPACITY 43.65T FREE 15.97T ( 36.6%)
  dpmdisk05.lhep.unibe.ch /mnt/storage1 CAPACITY 27.28T FREE 805.87G (  2.9%)
  dpmdisk05.lhep.unibe.ch /mnt/storage2 CAPACITY 27.28T FREE 796.44G (  2.9%)
  dpmdisk05.lhep.unibe.ch /mnt/storage3 CAPACITY 27.28T FREE 837.00G (  3.0%)
  dpmdisk06.lhep.unibe.ch /mnt/storage1 CAPACITY 27.28T FREE 806.96G (  2.9%)
  dpmdisk06.lhep.unibe.ch /mnt/storage2 CAPACITY 27.28T FREE 893.80G (  3.2%)
  dpmdisk06.lhep.unibe.ch /mnt/storage3 CAPACITY 27.28T FREE 912.66G (  3.3%)
  dpmdisk07.lhep.unibe.ch /mnt/storage1 CAPACITY 36.38T FREE 1.25T (  3.4%)
  dpmdisk07.lhep.unibe.ch /mnt/storage2 CAPACITY 36.38T FREE 1.11T (  3.0%)
  dpmdisk07.lhep.unibe.ch /mnt/storage3 CAPACITY 36.38T FREE 745.05G (  2.0%)
  dpmdisk08.lhep.unibe.ch /mnt/storage1 CAPACITY 90.95T FREE 38.39T ( 42.2%)
```

CONFIGURATION

CONFIGURING

Add a new pool and bind the file systems to it

```
[root@dpm ~]# dpm-addpool --poolname dpmPart3 --def_filesize 100

[root@dpm ~]# for x in 11 12 13 14 16 18 19 23 24 25 26; do dpm-addfs --poolname dpmPart3 --server atlasfs$x.unige.ch --fs /DATA; done
[root@dpm ~]# for x in 11 12 13 14 16 18 19 23 24 25 26; do dpm-modifyfs --server  atlasfs$x.unige.ch --fs /DATA --st RDONLY; done

[root@dpm ~]# dpm-modifypool --poolname dpmPart3 --gid 101,109,114,115
```

Add the new disk nodes to the head node manifest and apply it

```
[root@dpm modules]# grep disk_nodes dpm.pp | grep unige
$disk_nodes = "dpmdisk01.lhep.unibe.ch dpmdisk02.lhep.unibe.ch dpmdisk03.lhep.unibe.ch dpmdisk04.lhep.unibe.ch
dpmdisk05.lhep.unibe.ch dpmdisk06.lhep.unibe.ch dpmdisk07.lhep.unibe.ch dpmdisk08.lhep.unibe.ch atlasfs11.unige.ch
atlasfs12.unige.ch atlasfs13.unige.ch atlasfs14.unige.ch atlasfs16.unige.ch atlasfs18.unige.ch atlasfs19.unige.ch
atlasfs23.unige.ch atlasfs24.unige.ch atlasfs25.unige.ch atlasfs26.unige.ch"

[root@dpm modules]# puppet apply dpm.pp
```

Add the nodes to `/etc/shift.conf`

CONFIGURATION

AFTER

```
[root@dpm ~]# dpm-qryconf
POOL dpmPart1 DEFSIZE 100.00M GC_START_THRESH 0 GC_STOP_THRESH 0 DEF_LIFETIME 7.0d DEFPINTIME 2.0h MAX_LIFETIME 1.0m MAXPINTIME 12.0h FSS_POLICY
maxfreespace GC_POLICY lru RS_POLICY fifo GIDS 102,103,104,105,106,107,112,113,141,233,234,235,239 S_TYPE - MIG_POLICY none RET_POLICY R
                          CAPACITY 0 FREE 33.44T (  0.0%)
  dpm.lhep.unibe.ch /mnt/storage1 CAPACITY 9.99G FREE 9.96G ( 99.7%) RDONLY
  dpm.lhep.unibe.ch /mnt/storage2 CAPACITY 99.95G FREE 93.65G ( 93.7%) RDONLY
POOL dpmPart2 DEFSIZE 100 GC_START_THRESH 0 GC_STOP_THRESH 0 DEF_LIFETIME 7.0d DEFPINTIME 2.0h MAX_LIFETIME 1.0m MAXPINTIME 12.0h FSS_POLICY
maxfreespace GC_POLICY lru RS_POLICY fifo GIDS 101,102,103,104,105,106,107,109,112,113,114,115,117 S_TYPE - MIG_POLICY none RET_POLICY R
                          CAPACITY 571.12T FREE 32.74T (  5.7%)
  dpmdisk01.lhep.unibe.ch /mnt/storage1 CAPACITY 18.19T FREE 395.31G (  2.1%)
  <snip>
  dpmdisk08.lhep.unibe.ch /mnt/storage1 CAPACITY 90.95T FREE 38.39T ( 42.2%)
POOL dpmPart3 DEFSIZE 100 GC_START_THRESH 0 GC_STOP_THRESH 0 DEF_LIFETIME 7.0d DEFPINTIME 2.0h MAX_LIFETIME 1.0m MAXPINTIME 12.0h FSS_POLICY
maxfreespace GC_POLICY lru RS_POLICY fifo GIDS 101,109,114,115 S_TYPE - MIG_POLICY none RET_POLICY R
                          CAPACITY 402.72T FREE 702.71G (  0.2%)
  atlasfs11.unige.ch /DATA CAPACITY 32.64T FREE 23.85T ( 73.1%)
  atlasfs12.unige.ch /DATA CAPACITY 32.64T FREE 23.48T ( 71.9%)
  atlasfs13.unige.ch /DATA CAPACITY 32.64T FREE 23.77T ( 72.8%)
  atlasfs14.unige.ch /DATA CAPACITY 32.64T FREE 23.89T ( 73.2%)
  atlasfs16.unige.ch /DATA CAPACITY 32.64T FREE 23.83T ( 73.0%)
  atlasfs18.unige.ch /DATA CAPACITY 32.64T FREE 23.55T ( 72.2%)
  atlasfs19.unige.ch /DATA CAPACITY 32.64T FREE 23.75T ( 72.8%)
  atlasfs23.unige.ch /DATA CAPACITY 43.56T FREE 34.41T ( 79.0%)
  atlasfs24.unige.ch /DATA CAPACITY 43.56T FREE 34.29T ( 78.7%)
  atlasfs25.unige.ch /DATA CAPACITY 43.56T FREE 34.34T ( 78.8%)
  atlasfs26.unige.ch /DATA CAPACITY 43.56T FREE 34.58T ( 79.4%)
```

CONFIGURATION

SPACE TOKEN

```
[root@dpm ~]# dpm-getspacemd
7b3b2207-2bee-482c-baf7-bd27adbdb957 ATLASLOCALGROUPDISK dpmPart2
    atlas/ch,atlas/Role=production
    71.29T 71.44T Inf REPLICA ONLINE
cbd8449a-83aa-4544-9127-254c3c8a7323 ATLASDATADISK dpmPart2
    atlas/Role=production
    441.00T 44.79T Inf REPLICA ONLINE
0f505be3-0e48-4e44-8dab-d56b2380441a ATLASSCRATCHDISK dpmPart2
    atlas
    14.65T 460.48k Inf REPLICA ONLINE

[root@dpm ~]# dpm-releasespace --space_token 7b3b2207-2bee-482c-baf7-bd27adbdb957
```

Before

Release

```
[root@dpm ~]# dpm-reservespace --gspace 402T --lifetime Inf --poolname dpmPart3 --token_desc ATLASLOCALGROUPDISK
4241cf50-cee6-495d-abd4-1b95cfe9b9fd
[root@dpm ~]# dpm-updatespace --token_desc ATLASLOCALGROUPDISK  --group atlas/ch,atlas/Role=production

[root@dpm ~]# dpm-getspacemd
cbd8449a-83aa-4544-9127-254c3c8a7323 ATLASDATADISK dpmPart2
    atlas/Role=production
    512.00T 54.93T Inf REPLICA ONLINE
0f505be3-0e48-4e44-8dab-d56b2380441a ATLASSCRATCHDISK dpmPart2
    atlas
    14.65T 3.65T Inf REPLICA ONLINE
4241cf50-cee6-495d-abd4-1b95cfe9b9fd ATLASLOCALGROUPDISK dpmPart3
    atlas/ch,atlas/Role=production
    402.00T 324.50T Inf REPLICA ONLINE
```

Reserve

After

PERFORMANCE CONSIDERATIONS

With the BE-GE system architecture: no concerns anticipated

- Data (tokens) are close to where they are mainly accessed from
- Sites not too distant anyway
- So we did not consider running a stress test

Even with a different architecture: would not expect concerns

- Bern T2 runs ARC in caching mode, data staging is asynchronous
- Data staged from (much more) remote sites too (distributed NDGF model)
- For the T3 users, (*possibly increased*) latency is not a big issue anyhow

CONCLUSIONS

- **Setup of a multi-site DPM is technically straightforward**

- **Perhaps more challenging is the cross-site administrative side**
  - Role definition, appropriate privileges, etc.
  - Excellent communication between the site admins / technical teams
  - Minimise response time to incidents

- **Performance**
  - To be investigated for the pilot mode use case
  - Not a concern for the ARC caching mode