dCache @IN2P3-CC

Adrien Georget

Welcome to Lyon

IN2P3 Computing Center

Resources
- 80 people (**65 IT engineers**)
- Budget : 7.3M€ (HR excluded)
  - 2.5M€ buildings running costs (incl. **1.2M€ electricity**)
  - 4M€ IT investments (incl. 2M€ for WLCG)

Facilities
- **1700 m2** over two computer rooms
- 1,4 MW total (PUE 1.4)

Computing
- ~850 servers, 55k HTC, **931 kHS23**

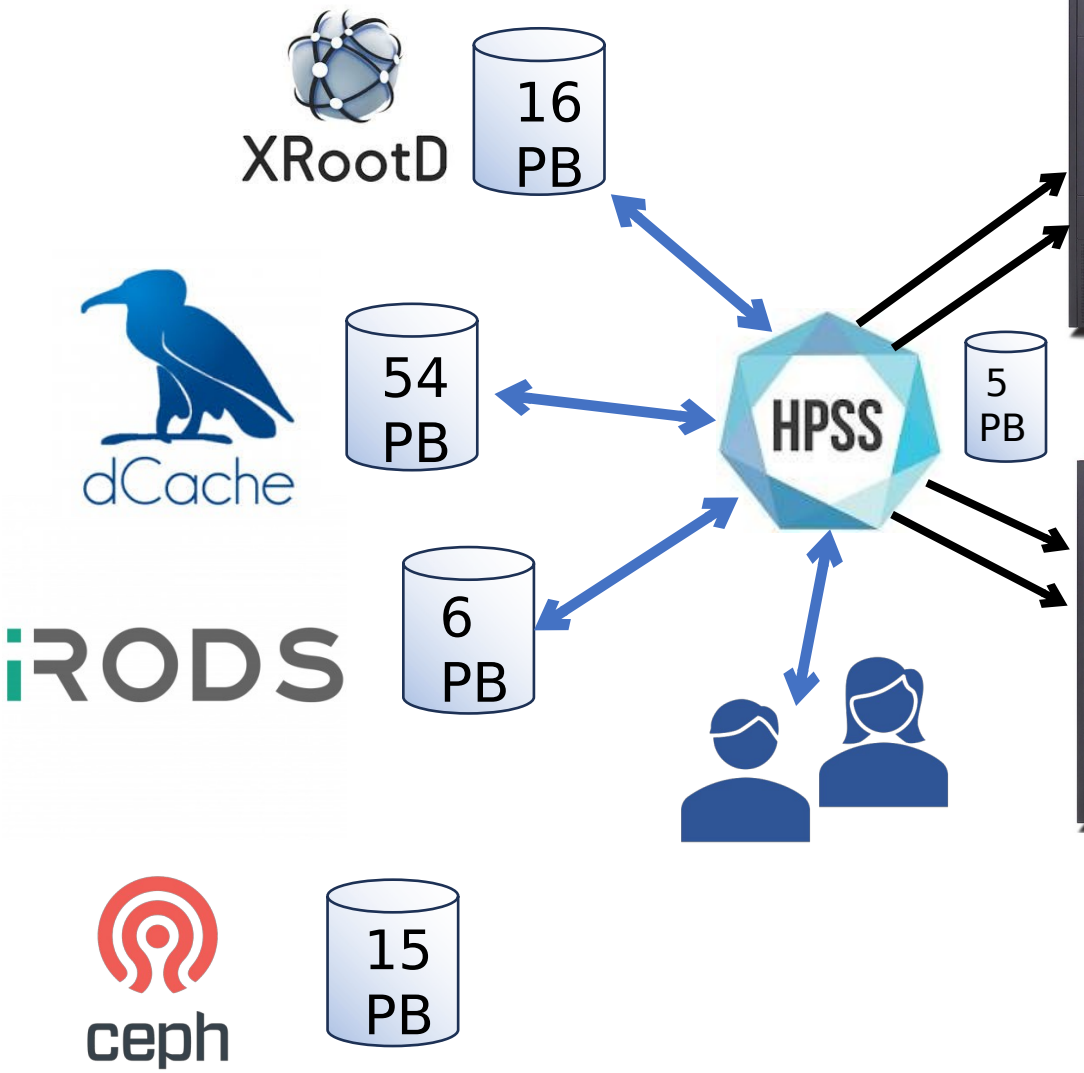
Storage
- Total allocated storage : **~240 PB** (62% tapes)

Networking
- 2x 200Gbps for WLCG (LHCOP & LHCONE)
- 1x 100 Gbps dedicated to IDRIS
- 1x 100 Gbps as backup and general purpose

100+ scientific collaborations

Who is using IN2P3-CC

IN2P3-CC Storage



200 PB on tape (single copy)

In production since 2005, 20th birthday 🥳 from 35TB to 72PB

IN2P3dCachesetup

for the Tier II dCache workshop, June 2006
by Lionel Schwarz, IN2P3

1. Head node setup

Right now all head node services are located on a single machine which is a (V20Z bi-opteron 2GHz, 2GB RAM). There are plans to separate the pnfs server and its DB to another host, same hardware. The backup is done once a day with pg_dump and saved to our TSM backup system.

2. Pool Nodes

We have 13 disk servers in dCache serving about 35TB. We use various disk configuration (direct attached disk/disk array) on various hardware (Transtec bi-Xeon 4GB RAM/V40Z quad-pro 8GB RAM...). All nodes are installed under SL3. We plan to install nodes under SL4 and Solaris10 in the future. All nodes have 2 1Gb interface, 1 on the outside and 1 on the inside (workers and HPSS connection), so that GridFTP traffic does not mix with migration/stage.

3. Installation

All installations/upgrades are done manually. We plan to use some automatic tools like yaim in the future.

IN2P3 experience

Conclusions

- We are confident in using dCache in production but aware of the amount of work in order to reach a good level of administration
- We feel that we would benefit from having the sources...
- We are ready to participate to any effort (development, tools, documentation...)
- Suggestions
  - Have a dCache admins mailing list
  - Have a regular (once a year?) meeting with dCache developers and admins (and users?)

dCache workshop – 30 Aug – 1 Sept 2005 11

dCache @IN2P3-CC

3 dCache instances (v9.2)

- **LCG (Atlas / CMS / LCHb)**

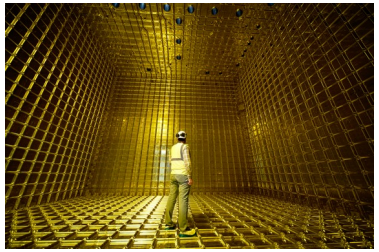
  - 51PB / 129M files
  - 165 servers (Dell R740XD2, HPE Apollo 4200)
  - weakly : 3PB imported, 5PB exported, 4PB read analysis
  - up to 300TB staged from tapes per day

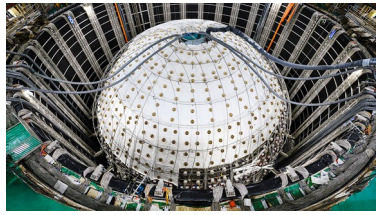
- **Rubin Observatory (LSST)**

  - 18PB allocated, 25% used but 277M files
  - 65 servers
  - 2500 images per night (20TB), +5PB per year
  - *See Fabio's presentation tomorrow*

- **EGEE (Dune, Belle2, Juno, Xenon, ...)**

  - 2.5PB / 36M files
  - 13 servers

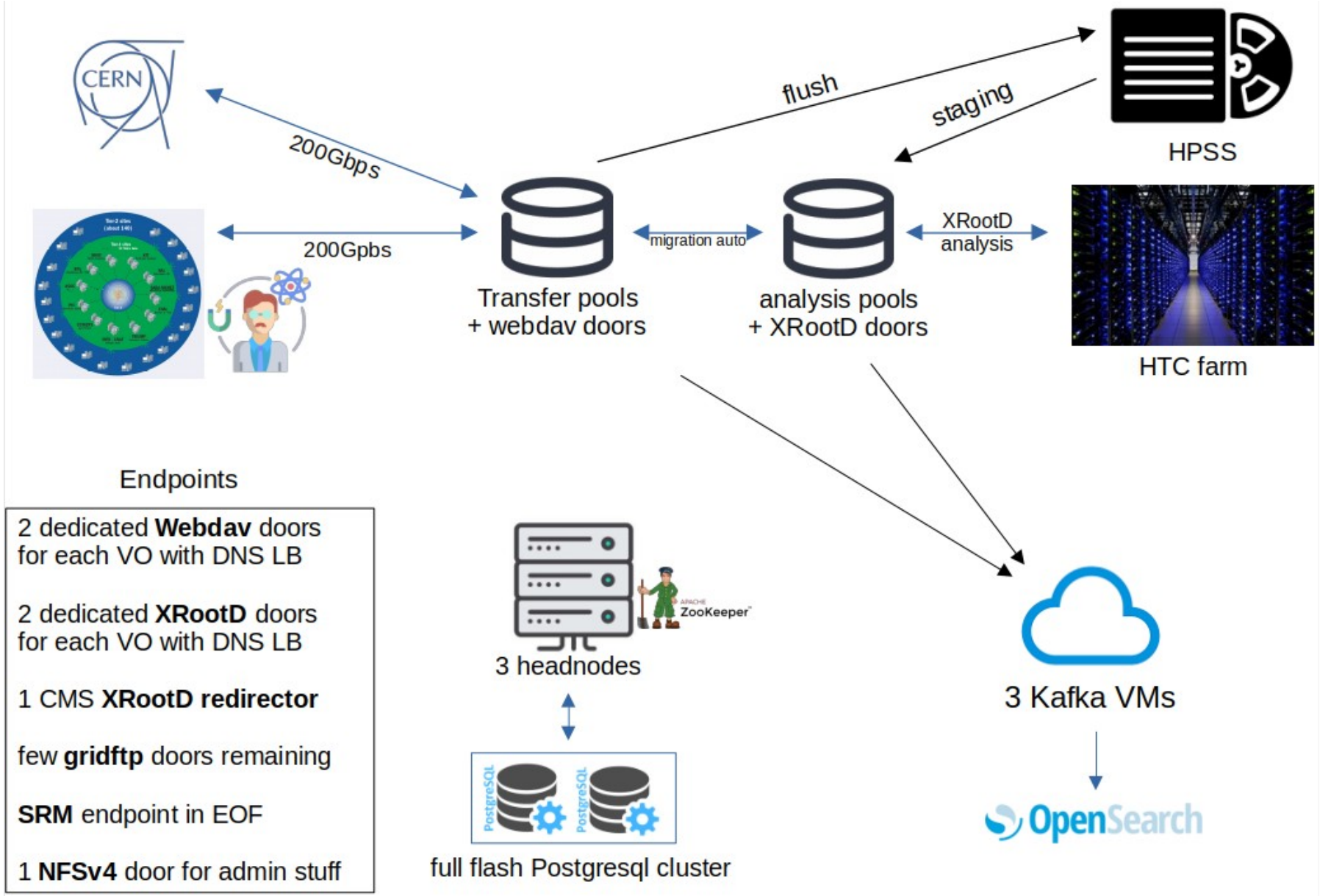




dCache architechure

Disk servers conf :
- Dell R740XD2, HPE Apollo 4200
- 24x 16TB disks, RAID6 XFS
- 128G RAM
- 2x Intel Xeon-S 4310 12c 2.10GHz
- 25Gbps network

Core servers conf :
- HPE DL360 Gen11
- 128G RAM
- Intel Xeon-G 5415+ 8c 2.9GHz



Highlights

- **SRM decommissionned for WLCG, still in use for others (EGEE)**

- **bulk API is running well, used by Atlas, CMS, LHCB (1 frontend / VO)**

```
--- statistics (Tracks request and target states (counts), sweeper state, etc.)
Running since: Thu Mar 13 14:10:50 CET 2025
Uptime 66 days, 19 hours, 21 minutes, 57 seconds

Last job sweep at Mon May 19 10:32:47 CEST 2025
Last job sweep took 0 seconds

------------------ TARGETS BY STATE ------------------
         (cumulative from last service start)
CANCELLED            :         13800
COMPLETED            :      48183941
FAILED               :          5932
SKIPPED              :             0

------------ REQUEST TOTALS (since start) ------------
Requests received    :       1581371
Requests completed   :       1581175
Requests cancelled   :             0
```

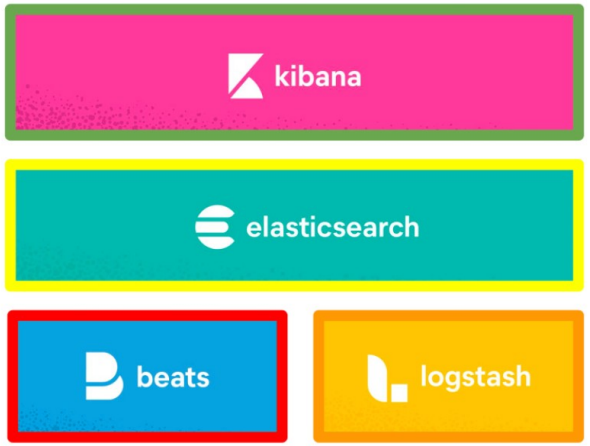
- **Still affected sometimes by pools stuck flushing to tape #7511 or #6426**

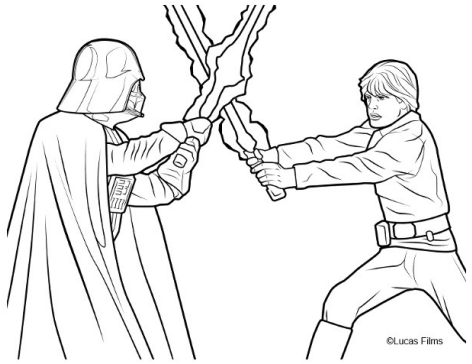
Highlights

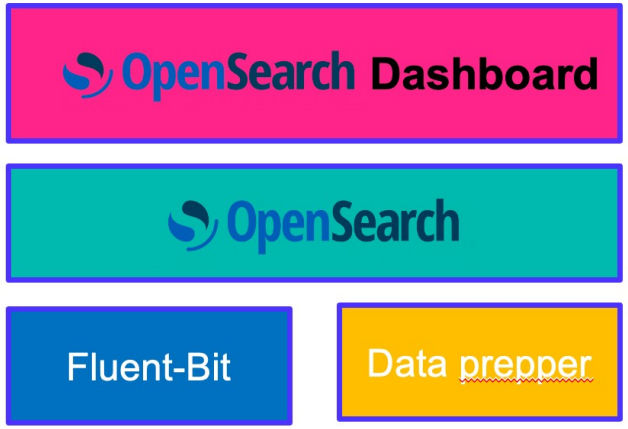
- **71 CentOS7 disk servers migrated to RHEL9 in 1 day** (without loosing files)
- **Some issues with RHEL9 + Java17** (SHA1 clients, SRM access fails with « unknown SOAP error ») -> Back to Java11 for problematic endpoints
- **JVM error: java.lang.OutOfMemoryError** : Cannot reserve x bytes of direct buffer memory

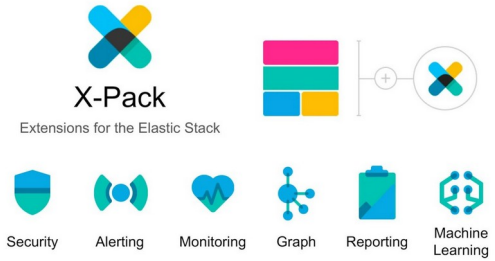
Highlights

- **Kafka events monitoring migrated from ELK to Opensearch**









X-pack tools under commercial license

- **Events flush with Opensearch connector for Kafka**
  https://github.com/Aiven-Open/opensearch-connector-for-apache-kafka
- **New feature in Opensearch 3.0 : Pull-bases ingestion**
  Pull-based ingestion enables OpenSearch to ingest data from streaming sources such as Apache Kafka

Highlights

LCG write

268,111	293.16TB	3.45GB
Events	Write	rate /s

LCG read

128,732	498.47TB	5.87GB
Events	Read	rate /s

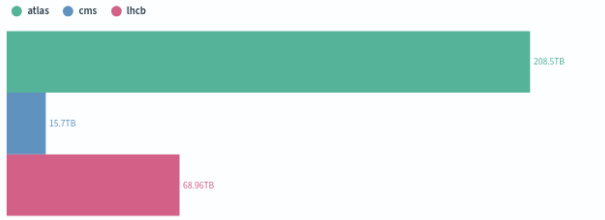
LCG XRootD analysis

974,367	386.4TB	4.5GB
Events	XRootD read size	rate /s

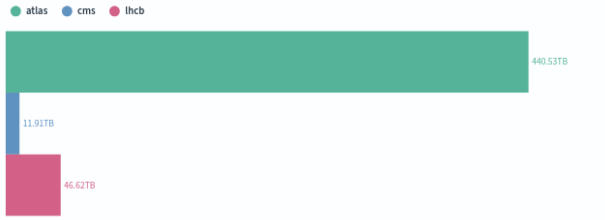
LCG Pool2Pool

220,033	512.72TB
Events	P2P volume

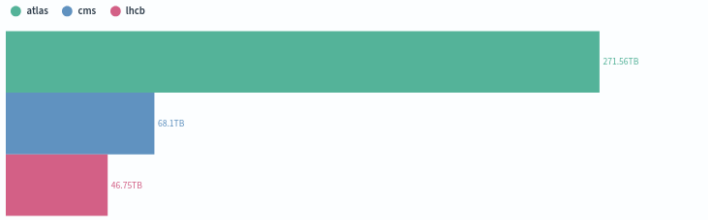
LCG write volume / VO



LCG read volume / VO



LCG XRootD analysis



LCG deletion

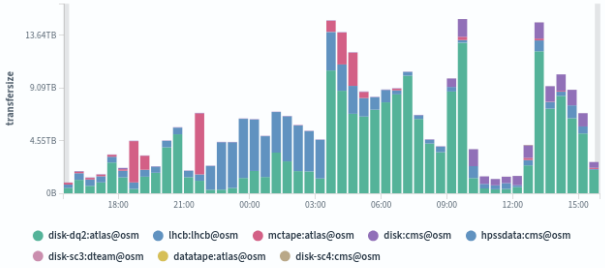
330,086	318.7TB
Events	Deletion Size

LCG hsm migration details

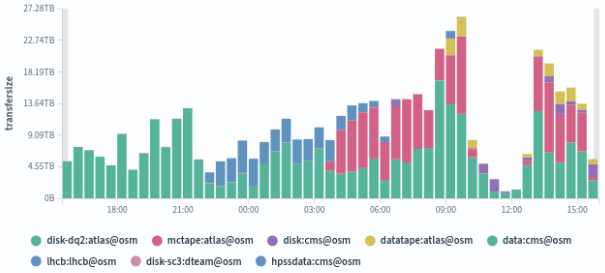
storageInfo: Desc...	Count	migration volume
mctape:atlas@osm	23,759	18.9TB
lhcb:lhcb@osm	2,244	5.6TB
datatape:atlas@osm	29	31.5GB
	26,032	**24.6TB**

1

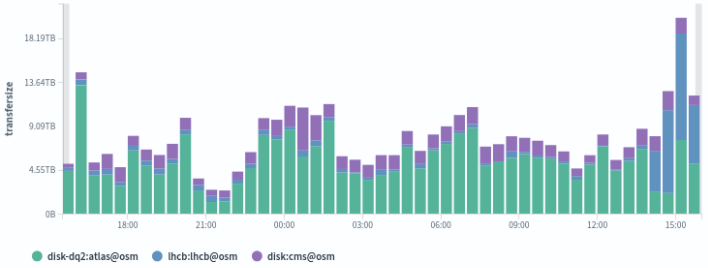
lcg write size / VO



lcg read size / VO



lcg xrd size / VO

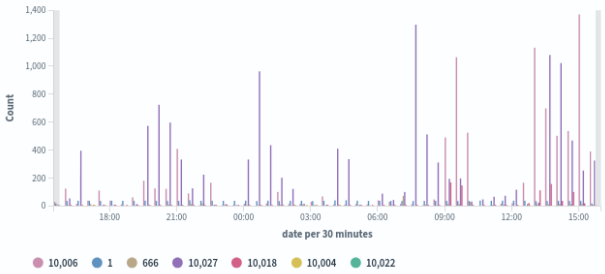


LCG hsm staging details

storageInfo: Desc...	Count	staging volume
mctape:atlas@osm	12,905	93.4TB
lhcb:lhcb@osm	10,225	46.5TB
datatape:atlas@osm	677	3.7TB
[illegible]	41	104.8GB
	23,848	**143.8TB**

1

errorcode



error table

error code	error msg	Count
10,006	No connection from client after 300 se	6,476
10,027	rejected PUT: 507 UNKNOWN_RESPO	3,345
10,006	No connection from client after 5 seco	2,012
10,027	rejected PUT: 501 NOT_IMPLEMENTE	1,715
10,018	File not online. Staging not allowed.	696
10,006	Request to [>PoolManager@dCacheD	411
10,027	rejected GET: 500 Server Error	333

Top 10 DN

DN	Count
atlagrid	235,086
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo2/CN=663	182,536
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atlpilo1/CN=614	175,402
cmsgrid	173,970
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=chiw/CN=87859	124,940
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=atulupov/CN=84	119,162
/DC=ch/DC=cern/OU=Organic Units/OU=Users/CN=amaltaro/CN=71	106,699

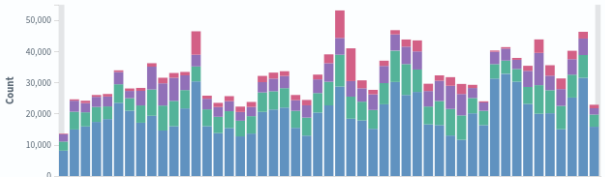
LCG hsm errors count

1,775	1
store - errors	restore - errors

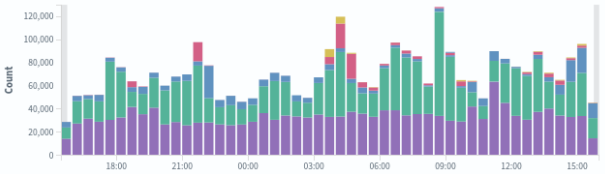
Top 10 doors

Door	Count
xrootd-ccdcacli432Domain	252,296
xrootd-ccdcacli433Domain	220,865
xrootd-ccdcacli423Domain	219,495
webdav-ccdcatli367Domain	206,349
xrootd-ccdcacli521Domain	118,937
webdav-ccdcatli406Domain	114,136
webdav-ccdcacli537Domain	113,447

lcg event by protocol



LCG event / VO



LCG Top 10 clients

Host	Count
2620:6a:0:8420:0:0:0:132	11,193
2001:1458:d00:24:0:0:100:1b	11,031
2001:638:50a:125:1:7cd8:0:1	10,222
2001:660:5009:9:193:48:99:6	10,175

Highlights

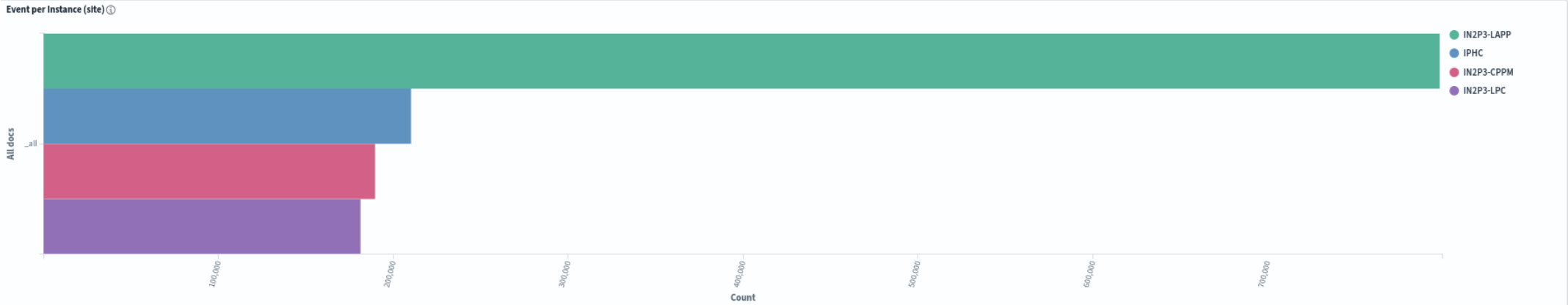
- **French dCache T1+T2 sites global monitoring (T1 + 4 T2)**
  - **1 dCache and Kafka instance with connector on each site**
  - **1 global Opensearch/Kibana instance running @IN2P3-CC**

Table JSON

@timestamp	May 7, 2025 @ 11:42:58.835
Instance	IN2P3-CC
VERSION	1.0
_id	billingLCG+1+1460617980
_index	.ds-lcg-france-dcache-ccin2p3-billinglcg-000408
_score	-
_type	-
billingPath	/pnfs/in2p3.fr/data/lhcb/LHCb-Disk/lhcb/buffer/lhcb/MC/2018/SIM/00234590/0145/


```
transforms=InsertField
transforms.InsertField.type=org.apache.kafka.connect.transforms.InsertField$Value
transforms.InsertField.static.field=Instance
transforms.InsertField.static.value=IN2P3-CC
```

Highlights



LCG deletion

230,827	48.8TB
Events	Deletion Size

Top 10 doors

Door	Count
doorsDomain	400,049
lapp-dccentral01_doorsDo	196,089
lapp-dp3s03_doorsDomai	7
lapp-dp10_doorsDomain	6
lapp-dp16_doorsDomain	6

LCG write

96,531	129.58TB	1.53GB
Events	Write	rate /s

LCG read

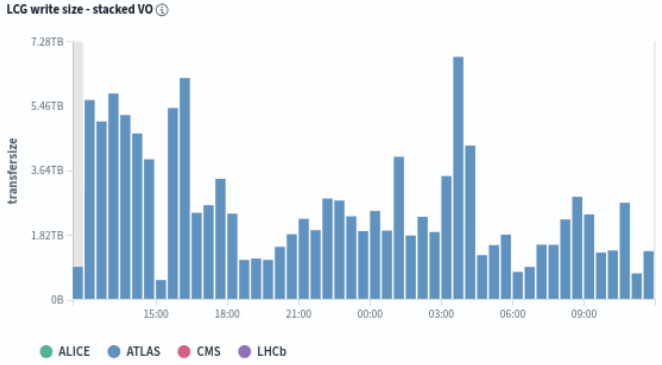
72,509	131.85TB	1.55GB
Events	Read	rate /s

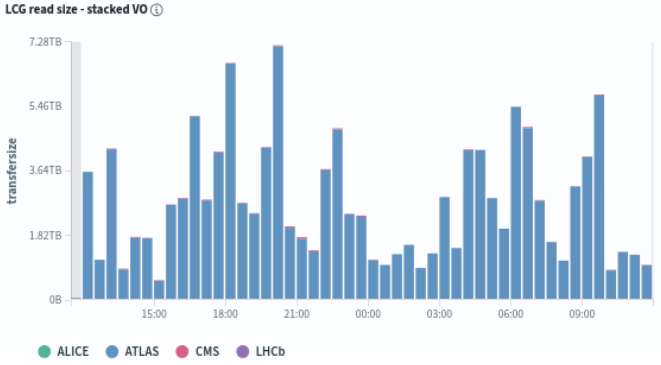
LCG XRootD analysis

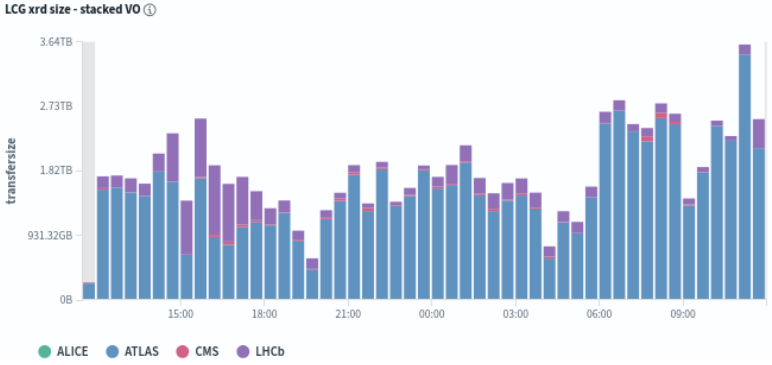
337,558	125.2TB	1.5GB
Events	XRootD read size	rate /s

LCG Top 10 pools

Pool	Count
DP_06-ATLAS-1	305,908
DP_06-ATLAS-2	64,837
DP_37-ATLAS-2	6,530
Sbgpool1_201	6,509
DP_302-OTHER-2	6,083
DP_09-ATLAS-1	6,048
DP_36-ATLAS-2	6,020
DP_37-ATLAS-1	5,879
DP_12-ATLAS-2	5,835
DP_12-ATLAS-1	5,759







Deployment

- dCache install and conf deployed with Ansible playbook
- ecosystem with Puppet (sys conf, grid stuff, Nagios probes)

```yaml
# Host specific settings for ccdcatli537
# Pools infos
dcache_poolinfo:
  pool-lhcb-dst:
    poolgroup: "pgroup-lhcb-dst"
    poolname: "pool-lhcb-dst-li537a"
    poolsize: "162000G"
  pool-atlas-dq2:
    poolgroup: "pgroup-atlas-import-disk"
    poolname: "pool-atlas-dq2-li537a"
    poolsize: "6000G"
  pool-cms-hpssdata:
    poolgroup: "pgroup-cms-hpssdata"
    poolname: "pool-cms-hpssdata-li537a"
    poolsize: "8000G"

# Doors infos
dcache_doorinfo:
  webdav:
    doorname: "webdav-ccdcacli537"
    root: "/pnfs/in2p3.fr/data/cms/"
    tag: "glue storage-descriptor"
```

Nagios monitoring

- **Nagios probes to monitor :**

  - check dCache cells
  - certificates validity
  - Zookeeper/Kafka health
  - PostgreSQL cluster
  - CPU load
  - Filesystem partitions
  - mountpoints
  - sysadmin stuff (server status, puppet last run age, sssd, ...)

Service	Status	Last Check
Check dcache lsst cells	OK	15:18:26
Service status	OK	15:20:14
Service status with hamac webhook	OK	15:10:47
Check /pbs/home mountpoint	OK	15:10:14
Check cpu load	OK	15:23:05
Check host system integration test suite	OK	2025-05-18 18:11:10
Check sssd backends status	OK	15:10:42
Check system_partitions filesystem	OK	15:13:38
Check tls certificate expiration for check_dcache_certificate	OK	15:13:44
Check zookeeper health for ccdcamcli23.in2p3.fr	OK	15:22:34
Puppet catalog status	OK	14:55:33
Puppet last run age	OK	15:00:03
Sampler post data	OK	15:15:02

What’s next ?

- **Prepare for HL-LHC workload**
- **Migrate to dCache 10.2**
- **Rewrite our HSM script interface between dCache and HPSS or reuse KIT’s script *dc2Hpss***
- **Set up QOS for data movement on demand between disk and tape (LSST)**


THANK YOU
DCACHE
DCACHE
DCACHE
DCACHE
DCACHE