+

Scalable Global Grid Catalogue for Run3 and beyond

Miguel Martinez Pedreira

ALICE

A Large Ion Collider Experiment

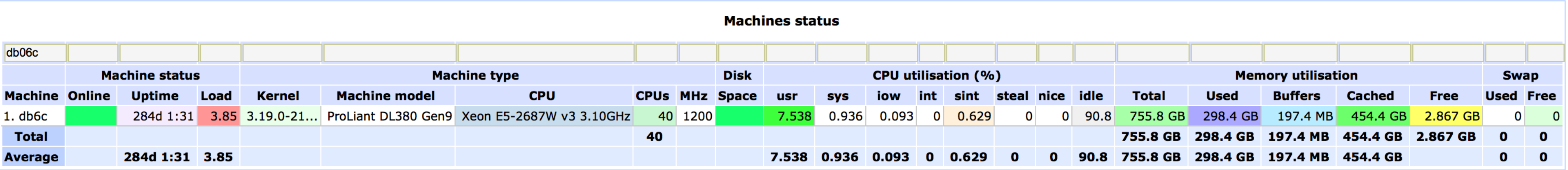
CERN

European Organisation for Nuclear Research

AliEn 2 @GRID

+ Current implementation

- MySQL-based AliEn File Catalogue
  - 2.5B logical entries
- One (powerful) DB master
  - 1.5TB RAM, 2.4TB on disk size
- DB slaves for hot standby / backups
  - 4h to dump, ~2 days to restore

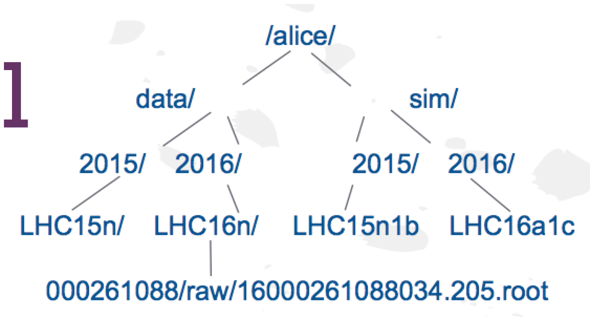


Machines status

db06c

	Machine status			Machine type					Disk	CPU utilisation (%)								Memory utilisation					Swap	
Machine	Online	Uptime	Load	Kernel	Machine model	CPU	CPUs	MHz	Space	usr	sys	iow	int	sint	steal	nice	idle	Total	Used	Buffers	Cached	Free	Used	Free
1. db6c	🟩	284d 1:31	3.85	3.19.0-21...	ProLiant DL380 Gen9	Xeon E5-2687W v3 3.10GHz	40	1200	🟩	7.538	0.936	0.093	0	0.629	0	0	90.8	755.8 GB	298.4 GB	197.4 MB	454.4 GB	2.867 GB	0	0
Total							**40**											**755.8 GB**	**298.4 GB**	**197.4 MB**	**454.4 GB**	**2.867 GB**	**0**	**0**
Average		**284d 1:31**	**3.85**							**7.538**	**0.936**	**0.093**	**0**	**0.629**	**0**	**0**	**90.8**	**755.8 GB**	**298.4 GB**	**197.4 MB**	**454.4 GB**		**0**	**0**

Catalogue in a nutshell



/alice/
data/
sim/
2015/
2016/
2015/
2016/
LHC15n/
LHC16n/
LHC15n1b
LHC16a1c
000261088/raw/16000261088034.205.root

- LFN namespace
  - **/alice/data/2016/LHC16n/000261088/raw/16000261088034.205.root**
  - 1180 tables (max 55M), 2.5B entries, namespace split into tables
  - Metadata

-rwxr-xr-x alidaq alidaq 264403565 Sep 09 22:10 0f24bce32446ea22840d188e035b11a9

- GUID namespace
  - 76CEBD12-76A0-11E6-9717-0D38A10ABEEF
  - 170 tables (max 210M), 2.4B entries, split by time intervals (append)
  - Version 1 UUIDs (MAC+timestamp)

- Physical File Pointers

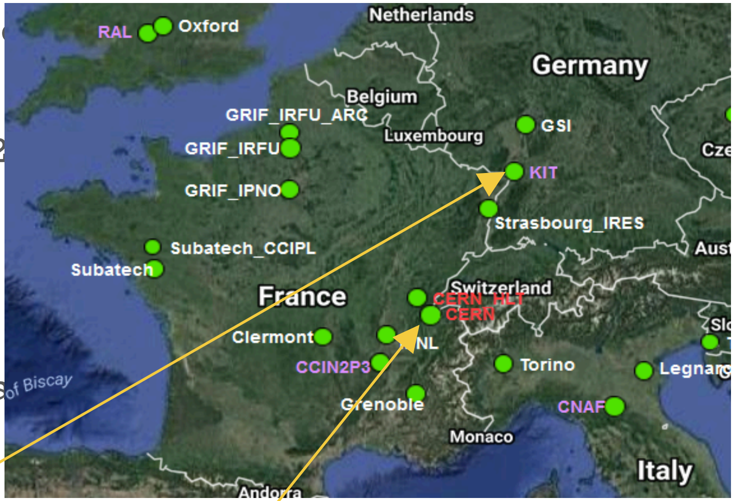
root://alice-tape-se.gridka.de:1094//10/33903/76cebd12-76a0-11e6-9717-0d38a10abeef
root://voalice10.cern.ch//castor/cern.ch/…/16000261088034.205.root

  - 3B entries, 920M physical files, pointers to ZIP members, 70PB over 70 Storage Elements

Catalogue in a nutshell

- LFN namespace
  - **/alice/data/2016/LHC16n/000261088/raw/16000261088034.205.root**
  - 1180 tables (max 55M), 2.5B entries, namespace split int
  - Metadata

-rwxr-xr-x alidaq alidaq 264403565 Sep 09 22:10 0f24bce32



- GUID namespace
  - 76CEBD12-76A0-11E6-9717-0D38A10ABEEF
  - 170 tables (max 210M), 2.4B entries, split by time interva
  - Version 1 UUIDs (MAC+timestamp)

- Physical File Pointers

root://alice-tape-se.gridka.de:1094//10/33903/76cebd12-76a0-11e6-9717-0d38a10abeef
root://voalice10.cern.ch//castor/cern.ch/…/16000261088034.205.root

  - 3B entries, 920M physical files, pointers to ZIP members, 70PB over 70 Storage Elements

DB query rates

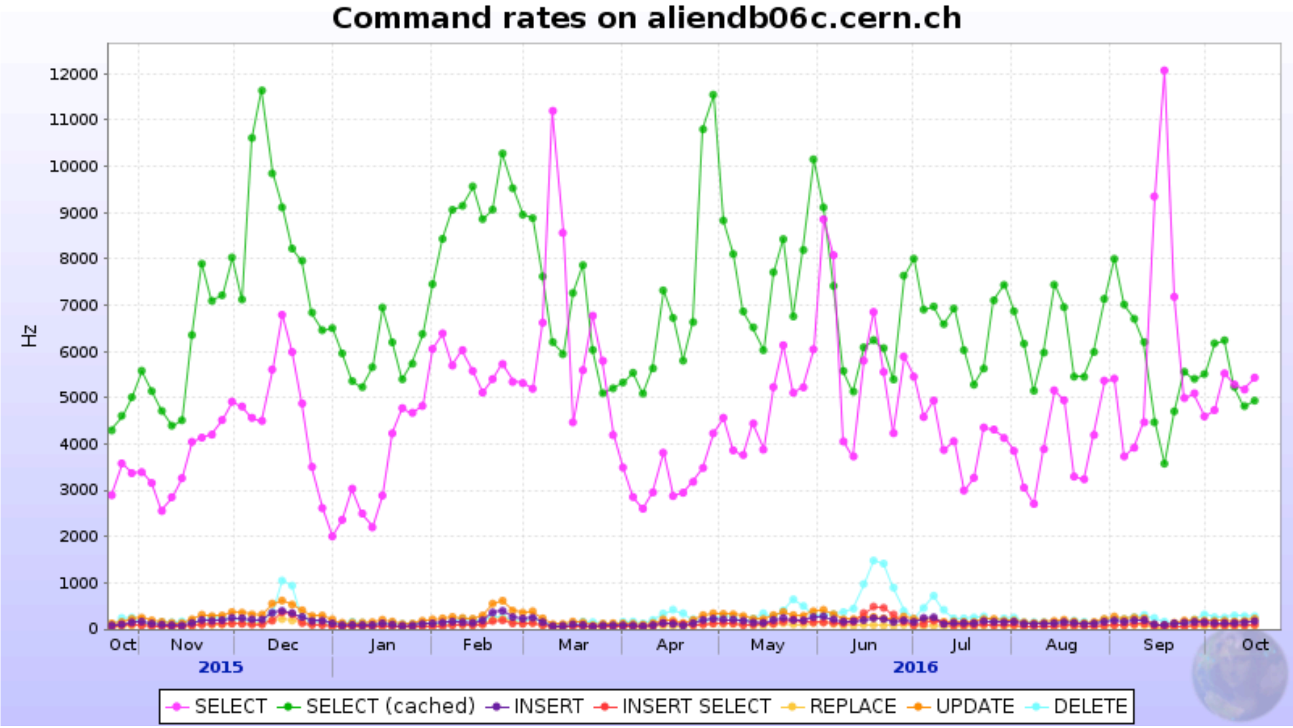
- Averages (1y)

11500 Hz Reads
570 Hz Changes
260 Hz Deletes

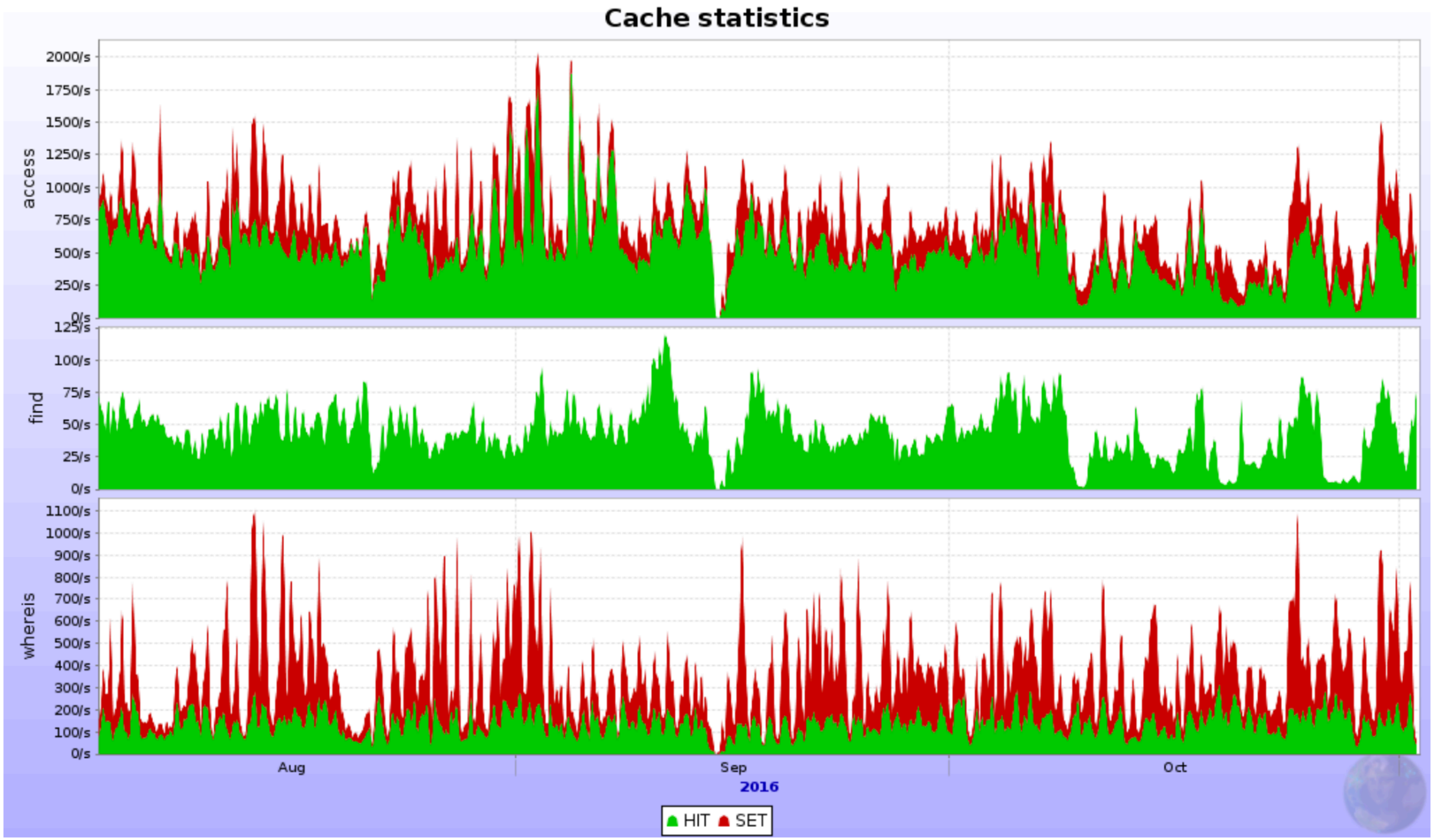
71500 running jobs

20:1 select/change ratio

10:1 read/write data volume

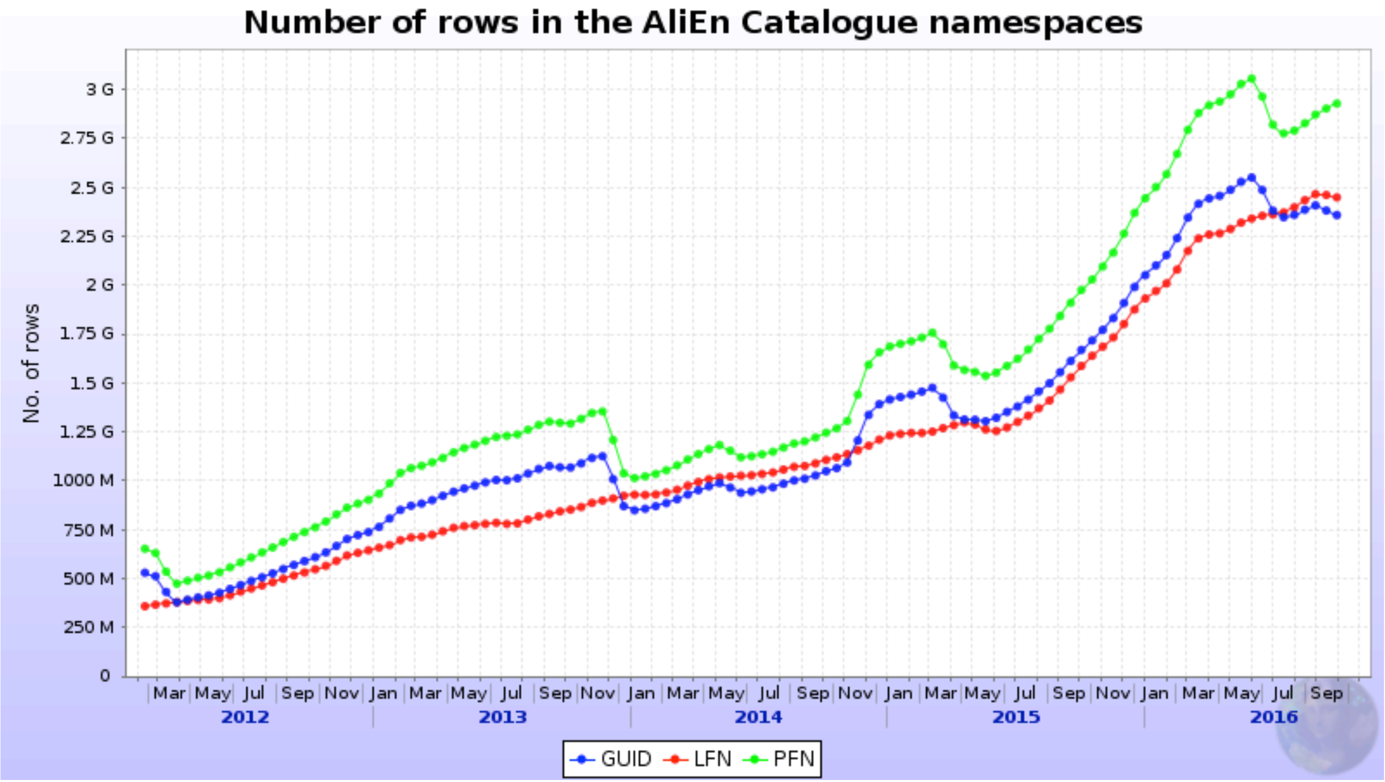


+ Cache


Cache statistics
access
2000/s
1750/s
1500/s
1250/s
1000/s
750/s
500/s
250/s
0/s
find
125/s
100/s
75/s
50/s
25/s
0/s
whereis
1100/s
1000/s
900/s
800/s
700/s
600/s
500/s
400/s
300/s
200/s
100/s
0/s
Aug
Sep
Oct
2016
HIT
SET

The AliEn catalogue in time

- 2x more files in 1 year



Future needs

- In Run3 we will have 5x more computing resources (300K CPUs + 5000GPUs)
- 10x more disk and tape storage =>
  ~10x more files to manage
- The goal is to sustain ~200kHz queries (stable)
  ~1Mhz queries (peaks)

- Looking for a solution providing:
  - Horizontal scaling
  - No single point of failure
  - High query rate
  - HA

Project proposal

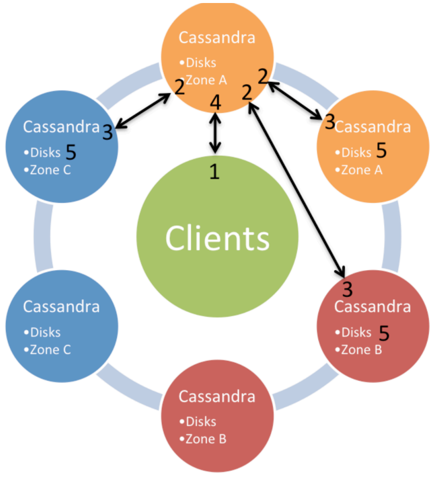
- Full proposal written as CHEP2016 project

- Apache Cassandra as main DB backend
- Exploitation of new memory technologies
- Optimization of framework workflows
- CVMFS enhanced user catalogue

Apache Cassandra

- Provides all the requirements mentioned before:
  - Horizontal scale
    - Add nodes to keep up ops/s
  - HA – No point of failure
  - Performance (see later initial benchmarks)
- Consistency
  - Tunable levels, key factor for us
- We move from N to 1 tables for the namespace
  - Simplification
- Mapping certain SQL operations not trivial
  - Groupings, quota calculations, 'where' possibilities…
  - NoSQL re-implementation, CQL helps too

Cassandra Write Data Flows



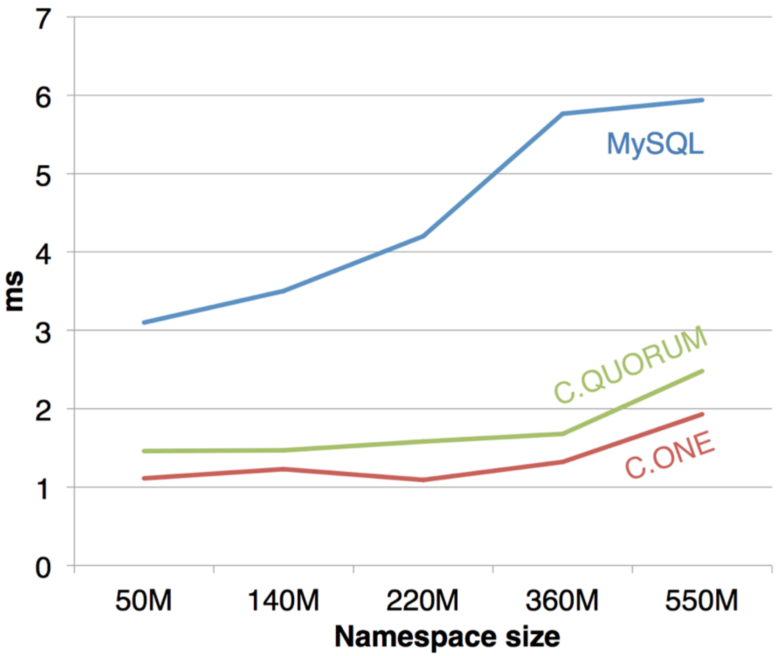
[1] Netflix techblog

Cassandra benchmarks

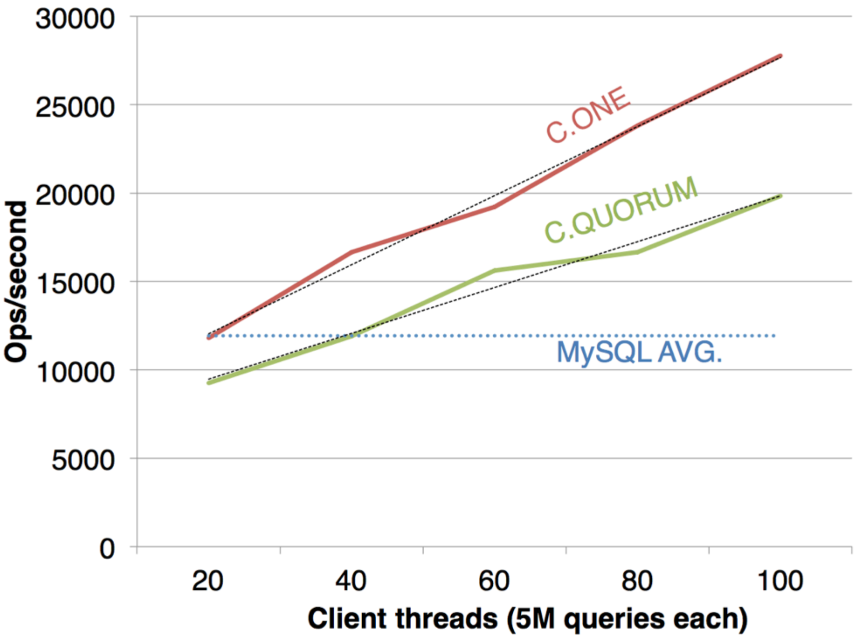
- Setup a 5-node ring
  - Server power: 16-48 cores, 100-350GB RAM
  - Java 8 Oracle, no swap, nofile/memlock limits (no degraded mode)
  - Discovered we need to be careful with a set of things thanks to developers!
  - Mapped namespace into a column family that is able to do `whereis` and `ls`: entry contains lfn+pfns metadata
    - Starting a new round of benchmarking on a implementation that allows `find` as well
  - Replication factor 3
- Data dump
  - MySQL to Cassandra -> slow
  - Artificial lfns and dirs -> very quick!
- Execution
  - Java sized thread pool, configure hierarchies, number of LFNS, etc…

Cassandra benchmarks

- Initial benchmarking shows promising results

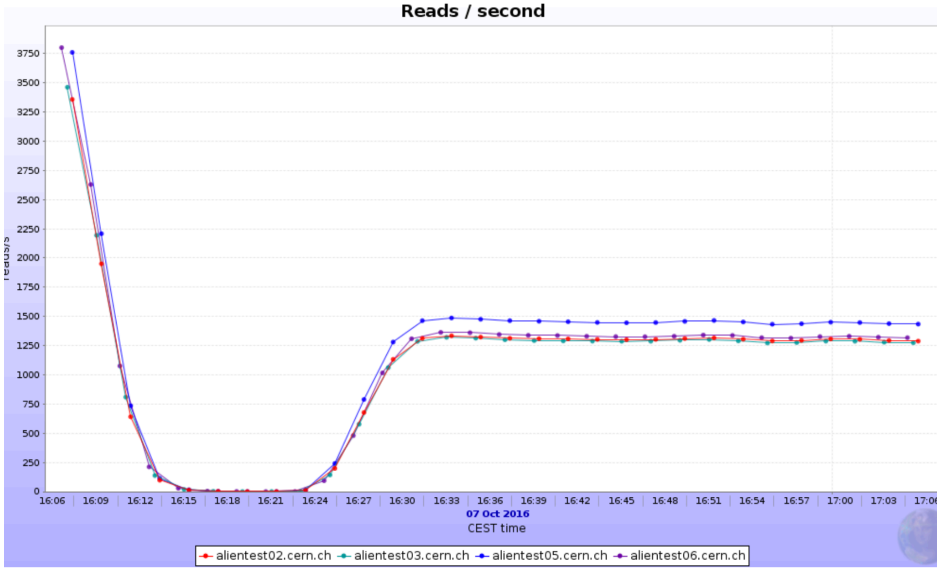


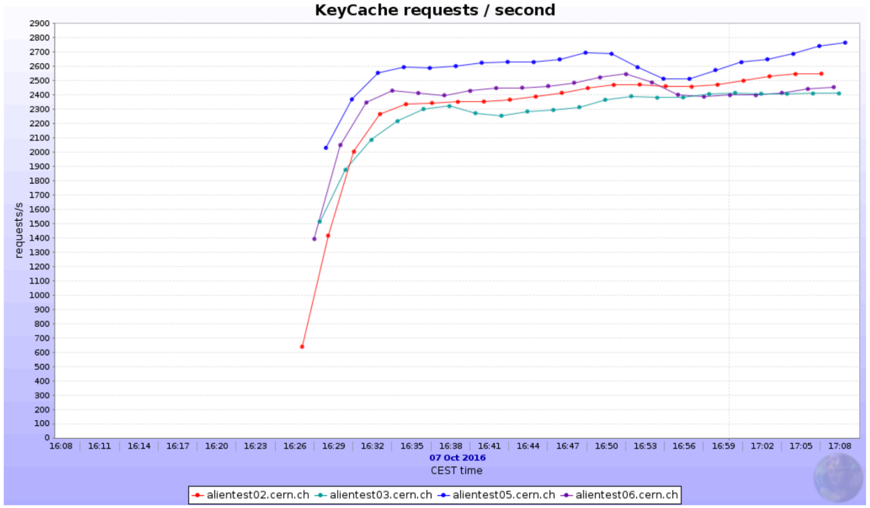
Time to retrieve logical and physical information of a file

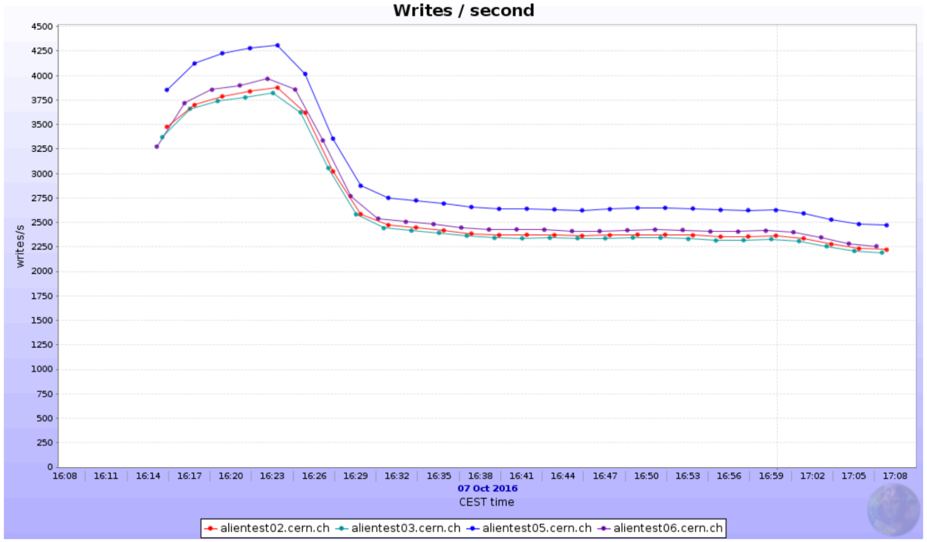


Operations per second based on number of clients

Cassandra in ML

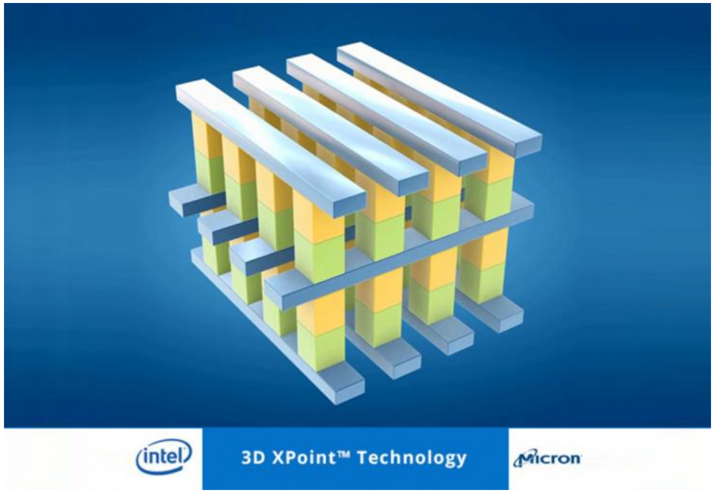






3D Crosspoint

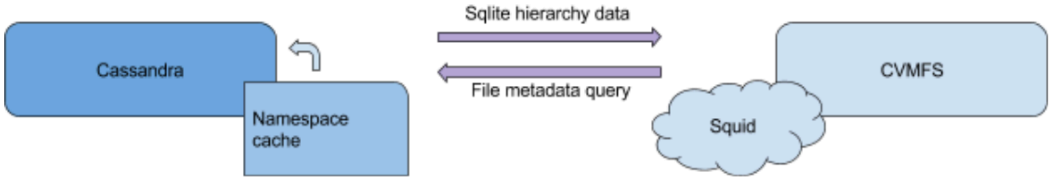
- Biggest memory breakthrough in 25 years
- Not very clear yet how it will work, but provides:
  - Higher data volume than RAM
  - Low latency
  - ½ price RAM?
  - Persistent?
- If non-persistent:
  - Bigger RAM -> bigger caching
- Persistent RAM
  - Remove slow I/O layers -> In-memory DB?
  - Booking area…



CVMFS

- CVMFS hierarchies pulled from Cassandra
  - "idea" stage, talking with developers
  - goal to provide a POSIX style filesystem of the AliEn Catalogue
    - To browse with a tool users are familiar with (CVMFS)
    - Re-using established infrastructure

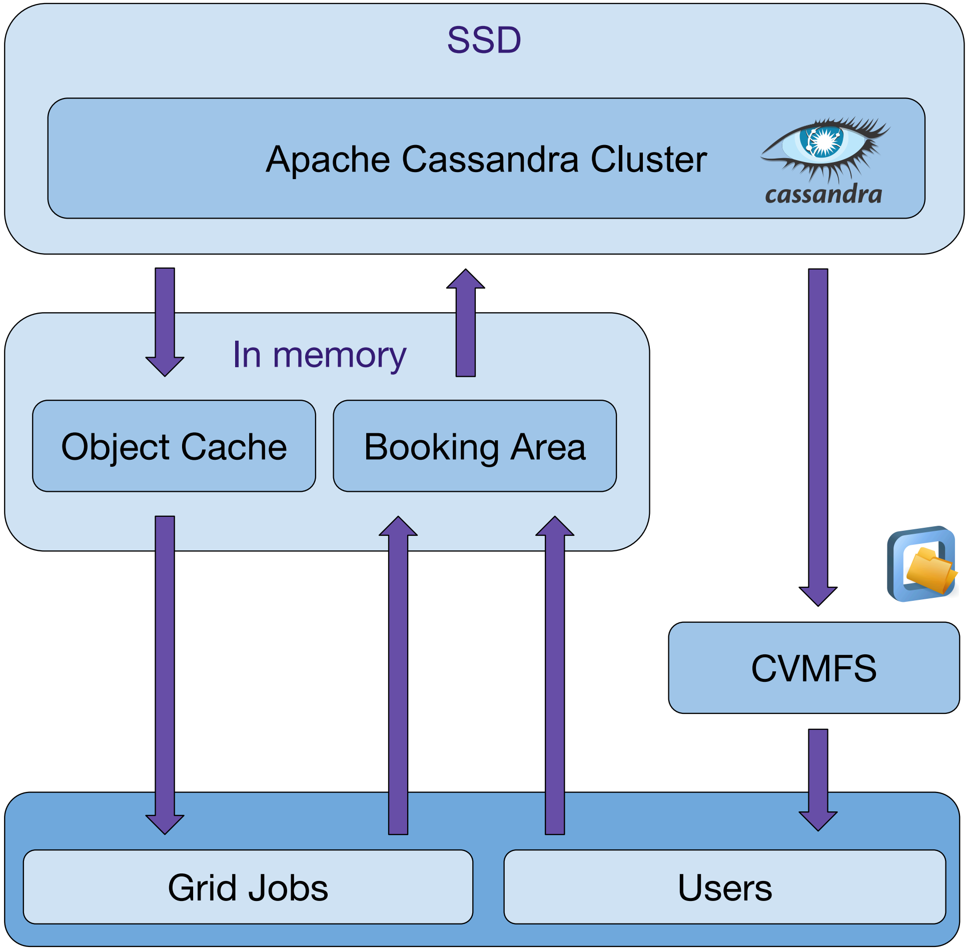
- Developments to be discussed and done:
  - Authentication+authorization
  - Sqlite creation plugin
  - Fine-grained cache



Framework

- Prepare envelopes for jobs in advance
- GUIDless would make even more sense in unique namespace
- Booking area
- Two layered-architecture:
  - REST services handling translation of AliEn to Cassandra calls
  - Thin client library to integrate in applications
    - EOS - namespace plugin libEosNsCassandra - direct mapping of AliEn FC to EOS namespace object types and views
    - AliEn user interface (shell, Web browser)
    - Experiment software (direct data access from ROOT)
- More…

Global schema



Thanks

- Questions?