ALICE and DOMA

QUICK RECAP ON ALICE DATA MANAGEMENT ELEMENTS AND PRINCIPLES

CENTRAL FILE CATALOGUE

- All files on the Grid are annotated in the catalogue
    - LFN namespace with ACLs

-rwxr-xr-x alidaq alidaq 264403565 Sep 09 22:10 /alice/data/2016/LHC16n/000261088/raw/16000261088034.205.root

    - Pointer to file location and GUIDs for SE filenames

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

- 5billion LFNs, 3billion physical files (archives of multiple logically combined LFNs)
- Running on MySQL (master-slave) DB with cache
- For 150k running jobs – 15KHz read / 1KHz change/delete

DATA ACCESS

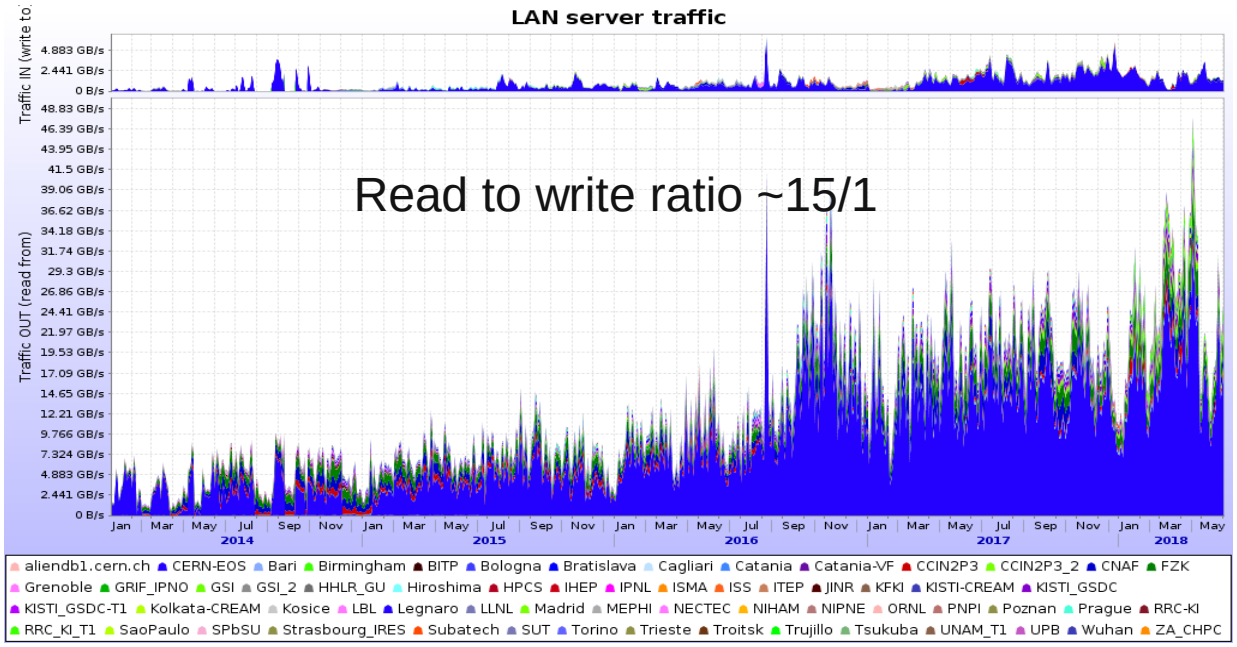
- Job is sent to data, but remote reading is OK
  - 97% of data volume is read locally
  - 3% remote reading due to local issue - server down/overloaded or corrupted/missing file
- File access
  - Authz (access envelope) – shared cypher between catalogue and storage, unique envelope per file and operation
  - Developed by A.Peters and D.Feichtinger 13 years ago
  - Inspiration for SciTokens
- Protocol – xrootd

DATA TRANSFERS

- Central transfer queue
  - All transfers are managed by admin
- Protocol – xrootd or xrd3cp (server to server)
  - 50% of data transfer is RAW data replication
  - 25% is storage management - occupancy equilibration or volume replacement
  - 25% is replica management

DATA ACCESS

- Most of the read access is from organized analysis (20% of Grid CPU capacity)
- Storage capacity and rates
  - Currently managing ~73PB of disk and ~60PB of tape
  - Read rates (from disk) 27GB/sec => 2.3PB/day



ALICE

UPGRADE AND CHALLENGES

GENERAL CHALLENGES

- File catalogue
  - An outdated technology by today standards
  - Scalability with higher load is doubtful
  - Should last another 10+ years
- File management tools
  - Low number of replicas (single copy is a norm nowadays)
  - Data placement policy to minimize effects of inaccessible storage
  - Migration of data from old to new storage
  - Self-reporting in case of data loss

CHALLENGES FOR RUN3 (ALICE UPGRADE)

- Order of magnitude more storage and files to manage
- Query rates of ~200 kHz
- Simplify the catalogue schema and improve caching (given the read/write ratio)
- Rock-solid tools for data migration
    - Including standardized site reporting
- Wish for fewer storage types (one protocol works great)
    - Less storage endpoints too, but not at the expense of capacity or performance

ONGOING WORK – HORIZON 2021

NEW AND IMPROVED CATALOGUE

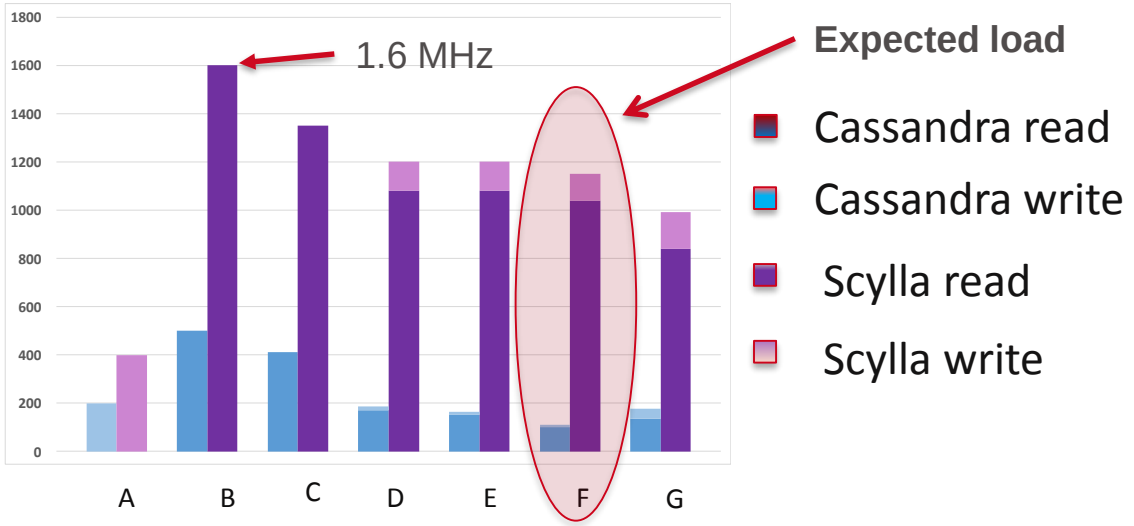
- Apple-ALICE collaborative work on Cassandra
- Answers all general and specific requirements:
  - Horizontal scaling, no single point of failure
  - High query rate, high availability
  - Consistency, easy setup
  - Drawback – SQL to NoSQL operations

SETUP AND BENCHMARKING

- 6-node test ring, replication factor 3, ~cheap hardware
- Any node (+1) can go away without performance degradation

=> Success!

KHZ
(ops/second)



	Benchmark (default cassandra-stress)	Cassandra	Scylla	Diff
A	Insert only	18 %util, 2% iowait	100 %util	x2
B	Read-only Gauss(5B,2.5B,10K)	Disk idle, 50% cpu	100 %cpu	x3
C	Read-only Gauss(5B,2.5B,1M)	11 %util, 40% cpu	11 %util, 100 %cpu	x3.28
D	Mixed (10r,1w) Gauss(5B,2.5B,100K)	2 %util, 45% cpu	10 %util, 100 %cpu	x5.8
E	Mixed (10r,1w) Gauss(5B,2.5B,1M)	5 %util, 50% cpu	16 %util, 100 %cpu	x6.4
F	**Mixed (10r,1w) Gauss(5B,2.5B,10M)**	**9% util, 45% cpu**	**40 %util, 100 %cpu**	**x6.1**
G	Mixed 2K thrd. read, 200 write, G(5B,2.5B,100K)	8 %util, 40% cpu, no iowait	26 %util, 100 %cpu	x5.62

ONE STEP FURTHER

- Even more promising results with Cassandra spin-off **ScyllaDB**
- Better DB management tools, fully asynchronous, no thread locking
- In summary – we have accumulated a lot of experience with a NoSQL solution for distributed FC, down to technical details for efficient operation

- ***Anyone interested in a common project?***

STORAGE MANAGEMENT TOOLS

- Reduce the time and human involvement for
  - Removal of selected datasets or dark data
  - Data loss – temporary or permanent
  - Data migration – partial (server) or full (entire SE)
- Volume of storage will increase, number of SEs will not decrease (in the observable future)

  => More need for efficient management and standardized reporting format

- ***Anyone interested in a common project?***

SUMMARY

- ALICE is gearing toward the 2021 Run 3 upgrade
  - Much larger storage capacity, higher load on individual storages
- Development of new catalogue ongoing
  - Addresses frequency of access and horizontal scalability requirements
- Looking into ways to standardize bulk storage management operations
  - Reporting formats, tools for site admins
- Potential for common activities, perhaps within DOMA
  - Note that ALICE timescale is basically ‘tomorrow’…