Author Miguel Martínez Pedreira | miguel.martinez.pedreira@cern.ch
Co-author Costin Grigoraș | costin.grigoras@cern.ch
for the ALICE Collaboration

SCALABLE GLOBAL GRID CATALOGUE FOR RUN3 AND BEYOND

ABSTRACT

The AliEn[1] file catalogue provides mapping between a UNIX-like logical name structure and the physical files stored in 80+ storage elements worldwide. In production since 2005, the size today is more than 2 billion logical file names. The ALICE upgrade in Run3[5] will bring 20 fold increase of namespace and access frequency. To anticipate this growth, we are investigating new DB technologies as replacement of the current set of classical relational databases.

CURRENT CATALOGUE

The catalogue grows smoothly with year-on-year namespace size increase by a factor 1.6 Today it holds above 2 billion LFNs[3] and 3 billion PFNs[4]. The performance is tuned for the most frequent 'read' operation. Full list of commands and their relative frequency are shown in table 1.

MySQL query rates

Command	Average rate	Max rate	Group rate
INSERT	155 Hz	4747 Hz	577 Hz
INSERT SELECT	115 Hz	1711 Hz	
REPLACE	79 Hz	957 Hz	
UPDATE	228 Hz	2339 Hz	
SELECT	4802 Hz	23333 Hz	11930 Hz
SELECT (cached)	7128 Hz	35470 Hz	
DELETE	291 Hz	6799 Hz	291 Hz
TOTAL	**12798 Hz**		

Table 1. Catalogue queries Jan-Jul 2016.

JOB WORKFLOW

Grid jobs rely on sets of configuration files and payload data. Jobs explore file hierarchy through `find` queries and direct file access. The catalogue load is proportional to the number of running jobs and number of files in the queries. Figure 2 shows the job and the catalogue DB interaction schema, passing through the intermediate layers of the AliEn middleware.

CASSANDRA BENCHMARKING RESULTS

The benchmarks use a MySQL replica of the AliEn catalogue and a 5-node Cassandra ring on similar hosts. The keyspace replication factor was 3 for strong consistency when using QUORUM for read and write. Cassandra is loaded with a set of the AliEn catalogue data in a single column family, similar to key-value pairs for LFN to PFN translation. The benchmark `*whereis*` command accesses the information related to the logical filename and its physical locations. 20 simultaneous clients query the databases with a frequency of several hundred Hz. Figures 4 and 5 show the MySQL and Cassandra performance. Figure 6 shows the linear scalability of Cassandra with increasing load.

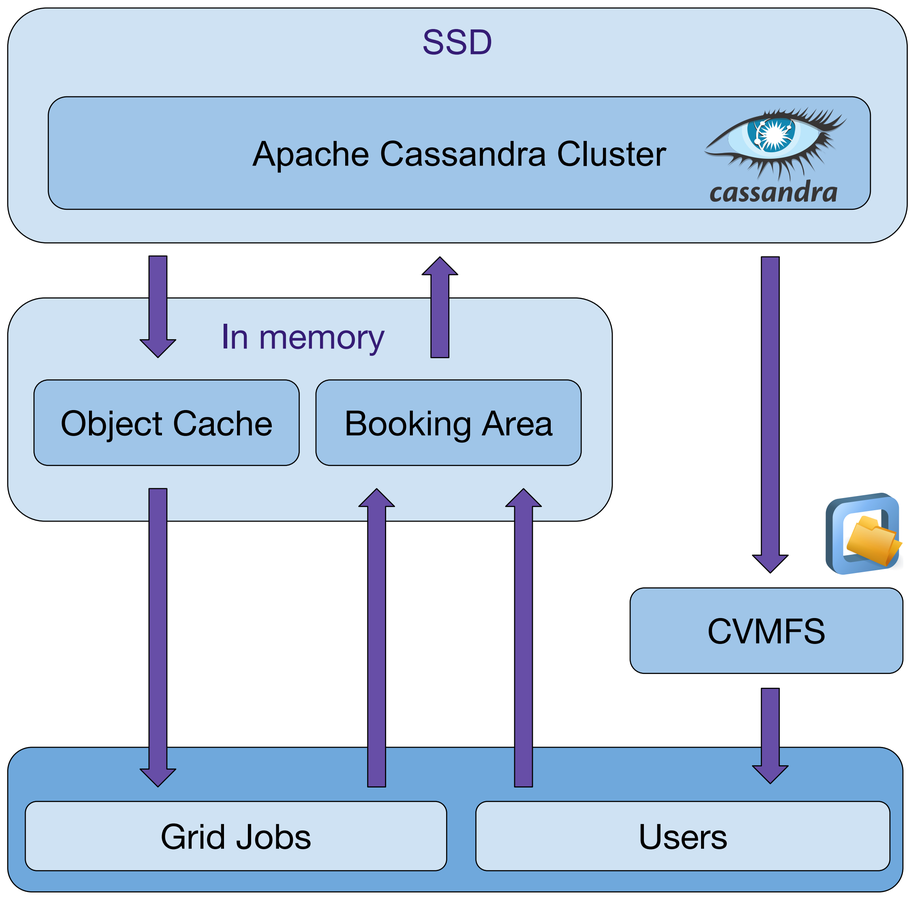

Fig.1 Interactions between the basic elements of the new catalogue – Cassandra and CVMFS.

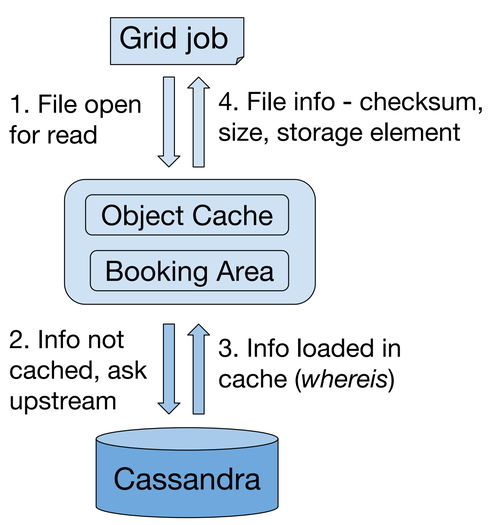

Fig.2 Job to catalogue Interaction.

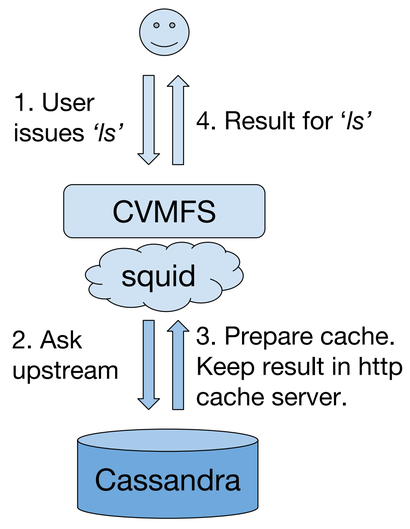

Fig.3 User to catalogue interaction.

GLOSSARY

[1] AliEn (ALICE Environment): Middleware for distributed computing operations
[2] CVMFS: Scalable, reliable and low-maintenance software distribution service
[3] LFN (Logical filename): Human-readable filename in the file catalogue
[4] PFN (Physical filename): Filename in the storage element filesystem
[5] Run3: Data taking period at the CERN Large Hadron Collider starting in 2021

TECHNOLOGY CHOICES

Database selection: Apache Cassandra is a free and open-source distributed database for managing large amounts of structured data across many servers. It provides the crucial key points needed for our use-case:

- Fault-tolerant and linear scalability
- High availability and operational simplicity
- Consistency

The project's main challenge is migrating from RDBMS schema to the Cassandra model which relies strongly on the application queries. To address this point we have analyzed the queries made by Grid jobs and users in all workflows.

User interface: CVMFS[2] is used across the Grid as a software distribution service. Our project will re-purpose CVMFS to present the file catalogue to the user with a familiar view of a standard filesystem.

ENHANCED CATALOGUE INTERFACE

CVMFS will provide snapshots of directory hierarchies on demand, querying the Cassandra DB. The process will be transparent to the end user, file registration will work in the same way as for Grid jobs.

CVMFS offers the possibility to have a common catalogue platform for all collaborations already using this tool. The existing stratum and squid caches infrastructure will be re-used.
Figure 3 shows the basic elements of the user – catalogue interaction with the new CVMFS + Cassandra based schema. More information in poster: "*New Directions in the CernVM File System"* from Track 4.

FUTURE WORK

We will focus on applying the results from the query workflow analysis to develop optimal Cassandra schema design. The AliEn code will be refactored to work with Cassandra. The full catalogue information will be inserted in the new database followed by extensive performance and stability tests. CVMFS will be enhanced with a client-side authentication and authorization; a catalogue snapshot creation plugin and a fine-grained system for cached data.

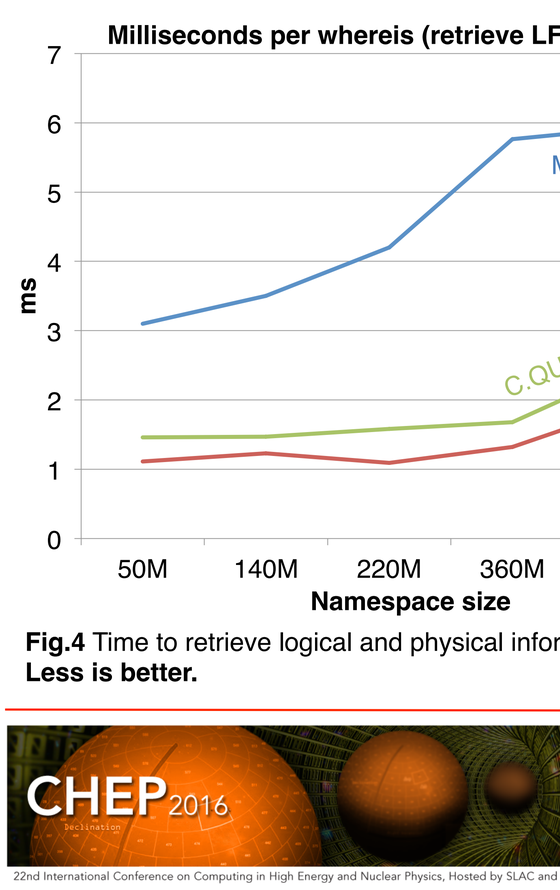

Fig.4 Time to retrieve logical and physical information of a file. **Less is better.**

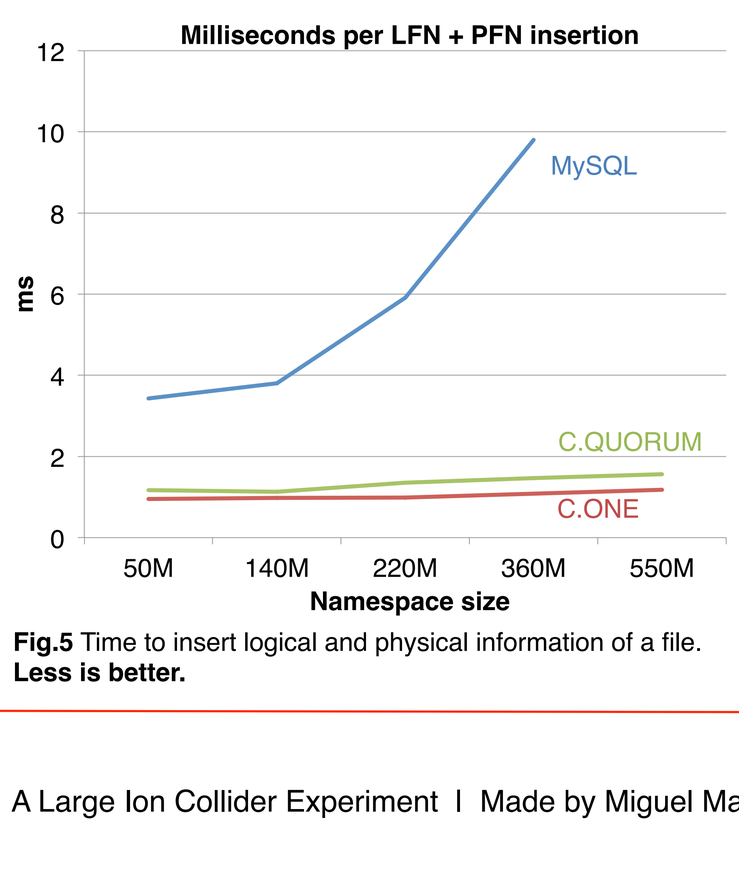

Fig.5 Time to insert logical and physical information of a file. **Less is better.**

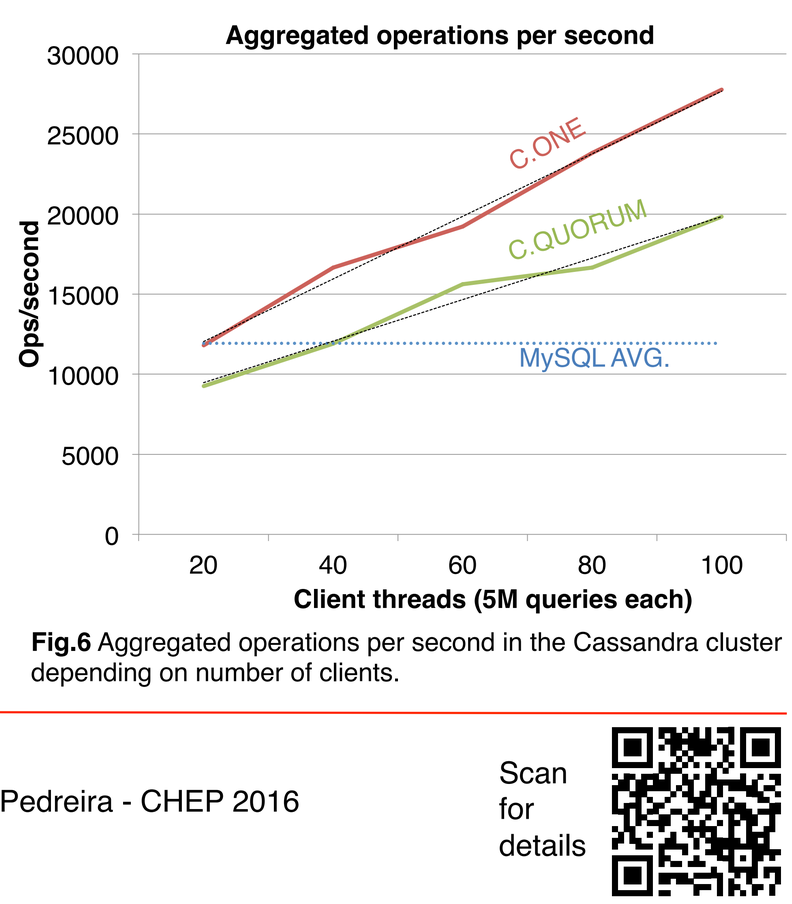

Fig.6 Aggregated operations per second in the Cassandra cluster depending on number of clients.