Multi-criteria partitioning on distributed file systems for efficient accelerator data analysis and performance optimization

S. Boychenko, M.A. Galilee, J.C. Garnier, M. Zerlauth, CERN, Switzerland
M.Z. Rela CISUC, University of Coimbra, Portugal

Motivation

In order to collect the requirements for the next generation storage solution, a detailed analysis of the CERN Accelerator Logging and Post Mortem systems was conducted. Despite the fact that modern distributed storage solutions can solve most of the identified issues, there are a few which require an individual approach in order to allow the infrastructure to reach its maximum potential:

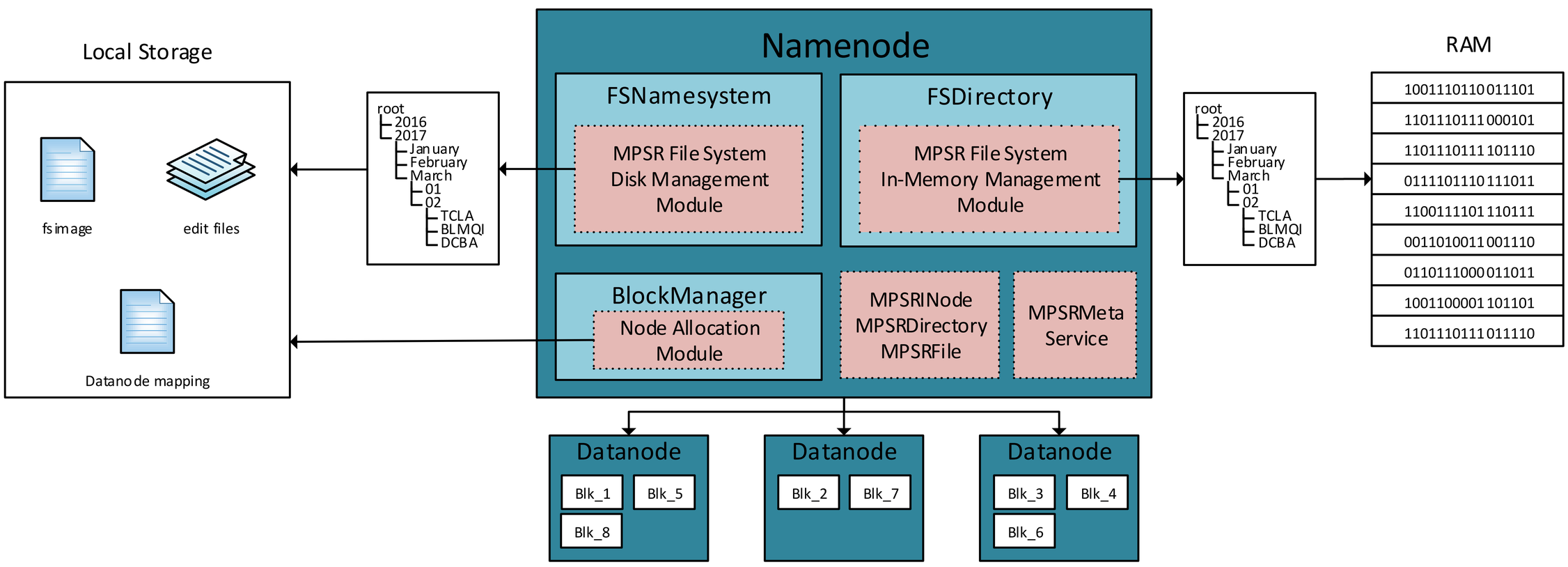
- The storage and processing solution must be optimized for heterogeneous workloads, as different users are interested in different analysis.
- The solution should be able to cope with seasonality in the executed query profiles, as the operations vary according to the accelerator state.
- The solution must be resilient to workload deviations, as constant upgrades of the LHC hardware systems can render an once efficient approach obsolete.

Mixed Partitioning Scheme Replication

To overcome the identified challenges, a novel approach for distributed storage and processing solutions was developed - Mixed Partitioning Scheme Replication (MPSR). The core functionalities of the proposed approach are presented below:

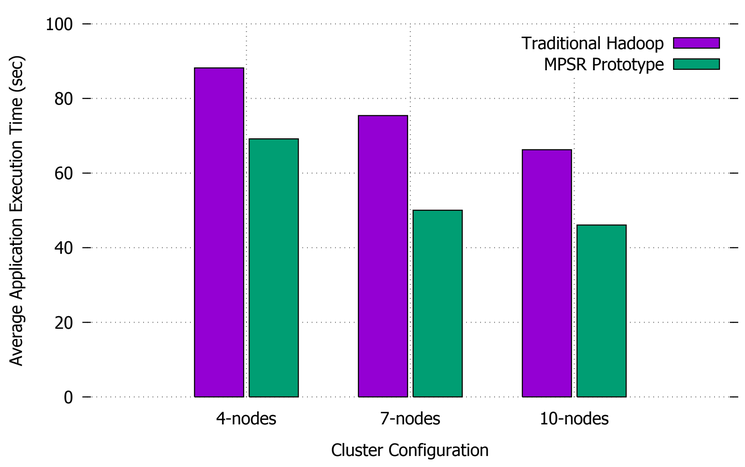
- The user requests executed on the system are classified based on the stored object attributes and grouped into several categories (which should be consistent with the storage replication factor).
- The data partitioning criteria are optimized for the previously determined workload categories, allowing multiple representations to be distributed throughout the cluster.
- Unlike in traditional solutions, the replication is performed using different data representations, rather than distributing the identical scheme on each of the replica groups.
- The elastic replica management temporary boosts the performance of the system for determined workload categories.

MPSR Hadoop Prototype



Average Query Execution Time

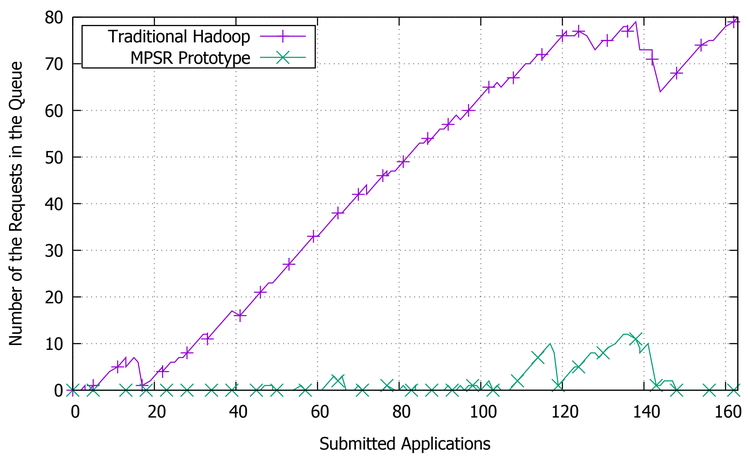
We started by conducting benchmarks to study and compare the average query execution time of the traditional Hadoop deployments and the MPSR prototype. This metric is absolutely critical for assessing the usefulness of the proposed approach.



It can be also observed that the performance gains of the MPSR prototype were higher in larger clusters, leading to a reduction in the average execution time by 21-42% on a 10-node infrastructure, respectively 19-35% on 7-node and 15-34% on 4-node infrastructures.

Average Queue Size

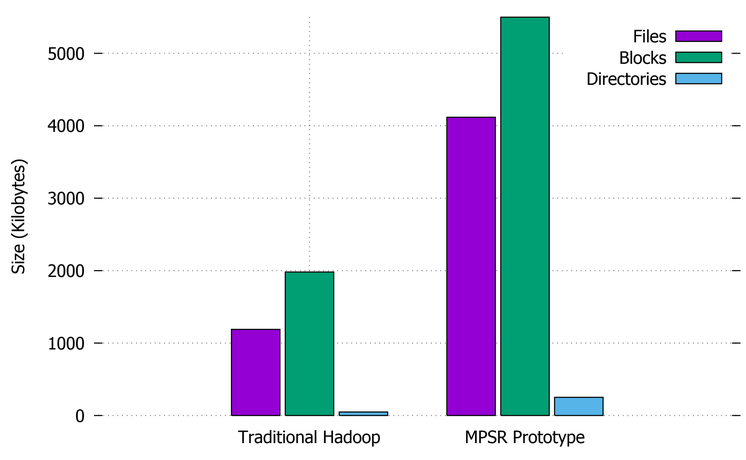
Despite the fact that the average execution time is a good measure to determine the performance of the system, the queue size cannot be neglected, as the request pile-up can render the infrastructure unusable at some point and therefore severely impact the application waiting time.



The MPSR approach was able to maintain the queue size close to zero throughout the entire runtime of the experiment, with the exception of a successive submission of a few applications with large input size.

Namenode Memory

The analysis of the heap memory revealed that the final data scheme of the standard Hadoop installation was represented by a total of 5342 files, requiring 5512 blocks to store its actual contents, while the MPSR prototype namespace is represented by 18577 files, which require 25128 blocks to store the contents



The size calculations were based on the estimations discussed by the Hadoop architects, where each of the Java objects (taking into account the Hadoop configuration) was assigned on approximate size.