Contribution ID: **110** Type: **Talk**

Time to Revisit Erasure Codes in Data-intensive Clusters

Friday, 2 December 2022 12:30 (30 minutes)

Replication has been successfully employed and practiced to ensure high data availability in large-scale distributed storage systems. However, with the relentless growth of generated and collected data, replication has become expensive not only in terms of storage cost but also in terms of network cost and hardware cost. Traditionally, erasure coding (EC) is employed as a cost-efficient alternative to replication when high access latency to the data can be tolerated. However, with the continuous reduction in its CPU overhead, EC is performed on the critical path of data access. For instance, EC has been integrated into the last major release of Hadoop Distributed File System (HDFS) which is the primary storage backend for data analytic frameworks such as Hadoop and Spark. This talk explores some of the potential benefits of erasure coding in data-intensive clusters and discusses aspects that can help to realize EC effectively for data-intensive applications.

Primary author: IBRAHIM, Shadi (Inria)

Presenter: IBRAHIM, Shadi (Inria)

Session Classification: HPC

Track Classification: Storage and IO