Evaluation of Erasure Coding, Triple NameNode High Availability and HDFS Router-based Federation features of Hadoop 3 (v3.2)

Nazerke Seidan[1,2], Emil Kleszcz[1], Zbigniew Baranowski[1]

[1)]CERN openlab, Geneva, Switzerland; [2)] Eötvös Loránd University, Budapest, Hungary

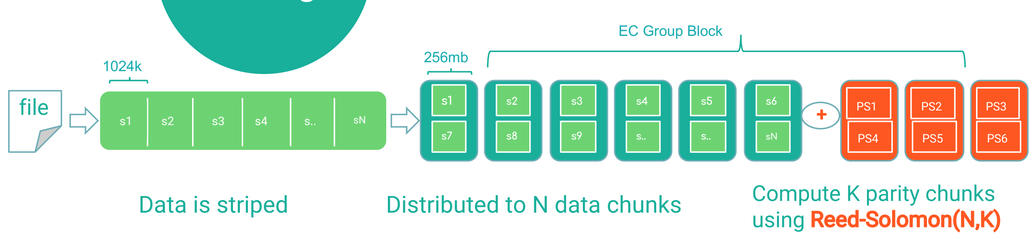
IT Information Technology Department

CERN openlab

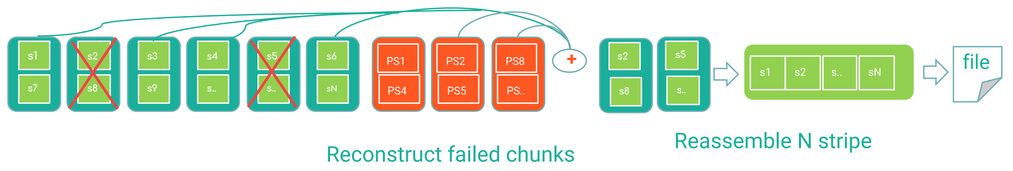
Project Objectives

- **The goal** of this work is to evaluate new features of Hadoop 3 and make an assessment of its readiness for production systems
- **Features to be evaluated:**
  - Erasure Coding
  - Triple NameNode High Availability
  - HDFS Router-based Federation
- **Evaluation methods:**
  - Raw storage performance (write/read) on small datasets
  - Analytics performance using TPC-DS benchmark tool on Spark SQL queries in Parquet, JSON file formats

Erasure Coding



- **Erasure Coding (EC)** uses RAID 5/6 concept to protect data which gives the same level of fault tolerance as 3x replication but with much less storage space
- Uses a codec to generate K parity data chunks eg., Reed-Solomon, RS(N,K) where N = data chunks, K = parity chunks,
- EC block group = data chunks + parity chunks
- It can tolerate up to K DataNode failures
- Current supported EC policy types: RS(3,2), RS(6,3) and RS(10,4)

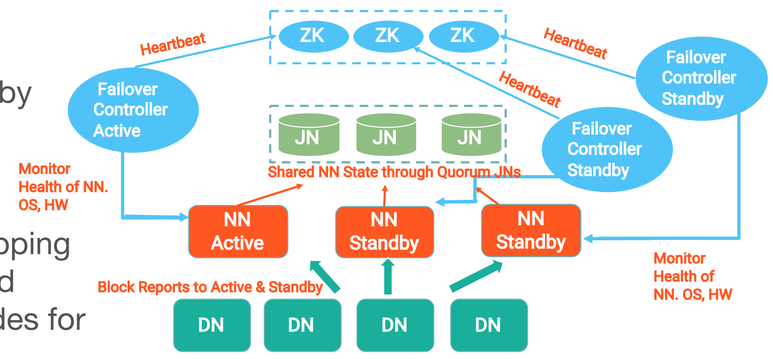


Triple NameNode

- **Triple NameNode architecture** provides a solution for two-point failure I.e., both NameNodes are down
- It is guaranteed that only one NameNode is active at a time while others being standby to avoid Split-Brain Scenario

Evaluation

- Tested the feature by adding the third NameNode and measured the performance by flipping between Active and Standby NameNodes for failover
- It acts very quickly for failures

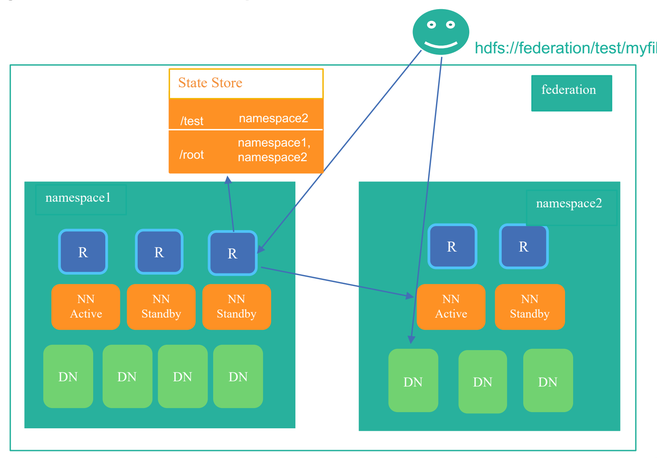


Router-based Federation

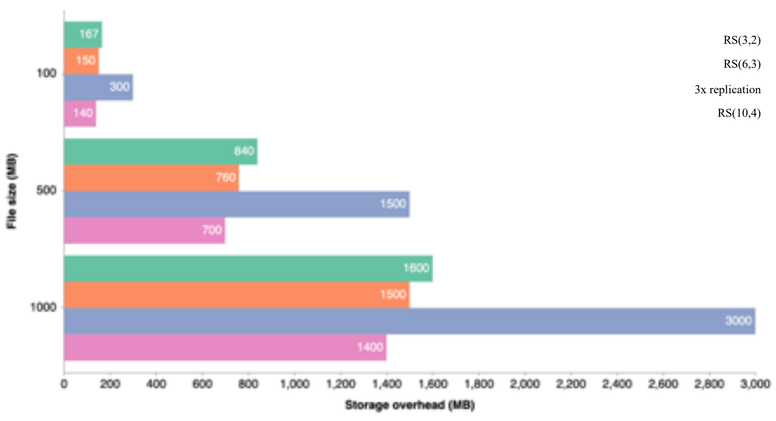
- **HDFS Router- based federation** overcomes NameNode scalability limits by introducing extra layer - Router components

Evaluation

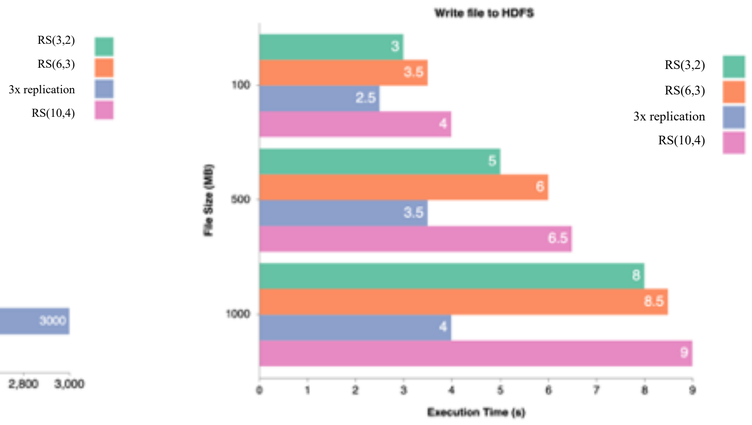
- Tested the feature by running federation of two clusters
- Secured clusters are supported from Hadoop v3.3
- The federation appears to its users as a single coherent system



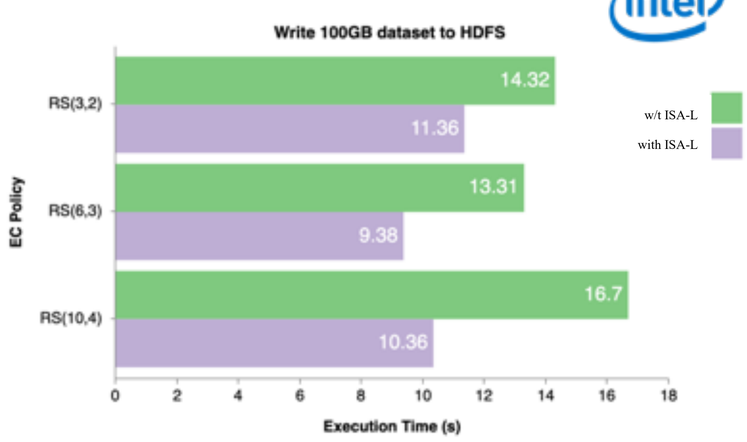
Measured performance of HDFS Erasure Coding



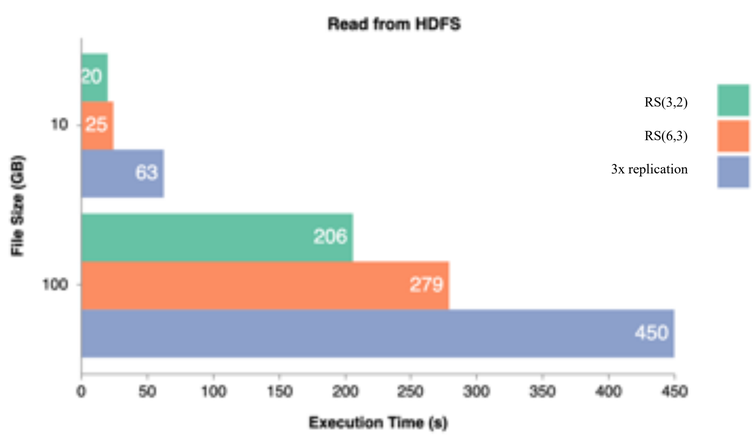
Storage overhead with EC is **~60% cheaper**



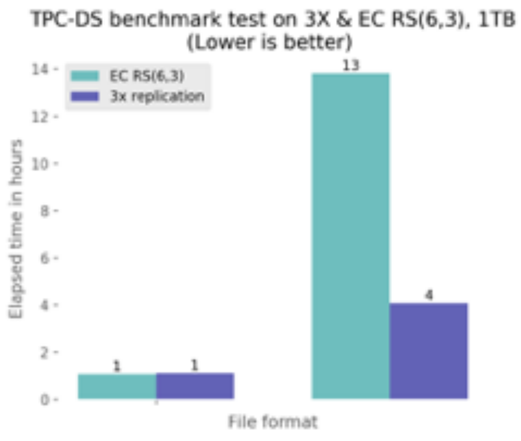
Write on EC dir. in HDFS is **50% slower** because of parity computation time



With **Intel's ISA-L** library write operation on EC dir. improved to **~30%**



Reading from erasure-coded directory is **twice faster** than 3x replication as EC leverages **parallelism**



Unoptimised file formats eg., JSON should be avoided with EC, while the performance with **optimised file formats** such as Parquet is the same for both 3x replication and EC

Summary

Erasure Coding

- gives an advantage in **storage savings**
- does not compromise analytics performance on **smart formats** (Parquet, Orc, etc)
- offers **flexible configuration**: can be selectively deployed on datasets
- can be **gradually enabled** on production systems

https://openlab.cern/