Offloading Calculations to Computational Storage Devices: Spark and HDFS

(LA-UR-21-27951)

Abstract

As the amount of data in the world grows, researchers have sought solutions to process data more quickly. One potential solution that has been explored in a High Performance Computing (HPC) context is using Computational Storage Devices (CSDs) to process data closer to where it is stored. In order to evaluate the functionality of these devices in a HPC environment, our team used Apache Spark and Hadoop Filesystem in order to offload computations and benchmark the drives. We wrote tests that perform arithmetic and matrix operations on Trinity sensor data. We ran our Spark experiments while scaling up the number of CSDs and cores used for each benchmark, and compared the amount of time elapsed for each computation. Our results show that the CSDs can be effectively used to offload tasks from the host machine because their compute power scales well and they can effectively work on workloads of various file sizes. However, we also determined that these devices have hardware reliability and workflow issues that make them unready for a production HPC environment.