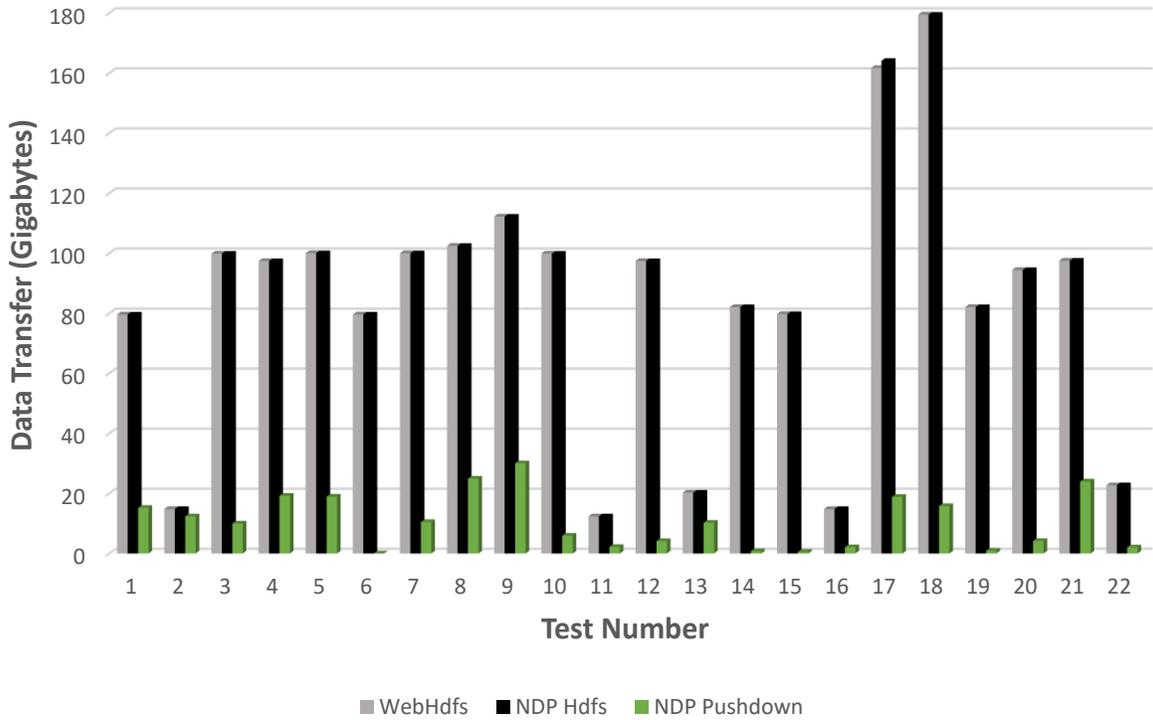


Preliminary Performance Evaluation of Near-Data Processing in Caerus

This document summarizes the preliminary performance results of the near-data processing (NDP) capabilities in Project Caerus. The experiments measured the amount of network I/O and application latency when running the TPCB benchmark on a Spark cluster, with data residing on a separate HDFS cluster. To enable NDP, we introduced our own DataSource module in Spark to plan for and perform computation pushdown. We also made slight modifications to the HDFS protocol to allow NDP-related information to be passed to the HDFS data nodes. The study compared three protocols: standard WebHDFS, modified HDFS without actual pushdown, and modified HDFS with HDFS pushdown.

The input dataset is 100GB in size, generated by `./dbgen -s 100`. All tests use 4 Spark workers.

Data Transfer Per Test (bytes)



Time per Test (seconds)

