Accelerating Hadoop Applications with the MapR Distribution
Using Flash Storage and High-Speed Ethernet

Contents

Background................................................................................................................................. 2
The MapR Distribution.............................................................................................................. 2
Mellanox Ethernet Solution...................................................................................................... 3
Test Setup and Results ............................................................................................................ 4
Conclusion............................................................................................................................... 6
Trademarks and special notices............................................................................................. 6
The adoption of mobile technology and e-commerce with their pervasive use in everyday life has resulted in a massive influx of data. This provides the opportunity for companies to transform the flood of data (often referred to as “Big Data”) to actionable insights for competitive advantage. Organizations that successfully leverage this have increased their innovation, operational efficiency and consequently, revenue growth. However, the practical usage of the data depends on the ability to store, manage and analyze it efficiently and quickly.

Apache Hadoop was born out of the need to process Big Data. Hadoop, which was initially designed for web indexing, has moved far beyond its original purpose and is increasingly becoming the go-to framework for large-scale, data-intensive deployments. Hadoop includes the MapReduce framework, that provides distributed analytics processing for mining the data; and the Hadoop Distributed File System (HDFS), which provides a scalable storage mechanism. Apache HDFS is written in Java and runs on different operating systems. However, Hadoop was designed from the beginning to accommodate multiple filesystem implementations.

MapR Technologies, an enterprise software company provides a complete distribution of Apache Hadoop and continues to be the fastest Hadoop distribution in the market. It combines over twenty different open source packages from the Hadoop ecosystem along with enterprise-grade features that provide unique capabilities for data management, protection and business continuity. The MapR Distribution includes several distinct advantages including:

- Data ingestion via NFS as a real-time stream
- Analysis performed immediately on the data as it is loaded
- Automated response execution

These advanced capabilities are largely made possible due to the innovation in the underlying file system: MapR-FS.
Due to the advantages of MapR-FS and the increasing usage of flash storage in production grade Hadoop environments, network bandwidth often becomes the bottleneck during data ingestion and analytics. Investing in fast servers and flash storage for MapR does not make sense if performance is restricted by the network. In this whitepaper, we evaluate the need for 40Gb/s networks in such environments.

**Mellanox Ethernet Solution**

Mellanox offers a complete product line of end-to-end 10/25/40/56/100GbE Ethernet solutions tailored for Big Data applications like Hadoop and NoSQL. Mellanox deliver industry-leading performance, scalability, reliability and power savings for advanced data center applications. Mellanox Ethernet switches feature consistently low latency and can support a variety of non-blocking, lossless fabric designs. Further with the Mellanox NEOTM network orchestration platform, network administrators can leverage existing data center fabric management solutions to deploy, orchestrate and monitor a large-scale cluster easily.

Big Data applications utilizing TCP or UDP over IP transport can achieve the highest throughput and application density using hardware-based stateless offloads and flow steering engines in ConnectX-3 Pro network adapters. These advanced stateless offloads reduce CPU overhead in IP packet processing thereby allowing completion of heavier analytic workloads in less time in the Big Data cluster. Sockets acceleration software further increases performance for latency sensitive applications, and faster network speeds such as 40 and 56Gb/s support greater throughput.

To connect the clusters, Mellanox copper, active optical cables, and transceivers offer reliable connections at speed from 10 to 100Gb/s with the highest quality, featuring error rates up to 100x lower than industry standards.
Test Setup and Results

The test setup consists of 5 servers connected with a Mellanox SX1036 40Gb Ethernet switch. The servers run RedHat Enterprise Linux 6.5 and the MapR Distribution 5.0 with hyper-threading turned off. The data was stored on the SSDs (on flash). Table 1 shows the configuration of the setup.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Master Node</th>
<th>Data Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>2x 8 Core Intel(R) Xeon(R) CPU E5-2680 @ 2.70GHz</td>
<td>2x 8 Core Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.70GHz</td>
<td>2x HGST FlashMax II Capacity SSDs (4800GB)</td>
</tr>
<tr>
<td>96GB DDR3 RAM</td>
<td>64GB DDR3 RAM</td>
<td>2x HGST FlashMax II Capacity SSDs (4800GB)</td>
</tr>
<tr>
<td>1x HGST FlashMax II Capacity SSD (4800GB)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Network Fabrics

- Mellanox ConnectX-3 40Gb Ethernet (MCX314A-BCBT)
- Mellanox ConnectX-3 10Gb Ethernet (MCX312A-XCBT)
- Intel 10Gb Ethernet
- Mellanox SX1036 36-port Ethernet switch

Software Stack

- Operating System: RedHat Enterprise Linux 6.5
- MapR Community Edition - 5.0 (Hadoop YARN 2.7.0)
- MLNX_OFED_LINUX-3.0-2.0.1

Figure 7. Map-R Hadoop 5-node 40GbE deployment
For the convenience of testing, we used standard QSA module with QSFP cable to connect 10 Gb/s adapters to SX1036 switch. There are many options in terms of adapters, cables and switches. Refer to Mellanox website for more details.

**TeraSort Benchmark Results**

TeraSort is probably the best known Hadoop benchmark. The goal of TeraSort is to sort 1TB of data as quickly as possible. It stresses both the MapReduce layer as well as the underlying distributed file system. Figure 8 below shows the total execution time to sort 1TB of data in our MapR cluster. The total execution time improved by 10% by moving from Intel 10GbE to Mellanox ConnectX-3 Pro 10GbE adapters and by a further 60% by moving to higher speed 40GbE network. In addition to the 3.2X improvement in execution time, the CPU usage also decreases when migrating from Intel to Mellanox network as shown in Figure 9. This allows more Hadoop jobs to run on the server while simultaneously generating faster analytics results. Our investigation led to the finding that there were significant pagefaults while running on Intel 10Gb/s network, which explains the reason for poor execution time and CPU usage.

![Figure 8. TeraSort Execution Time](image)

![Figure 9. TeraSort CPU Utilization](image)

1The test was done with out-of-box driver from RedHat Enterprise Linux 6.5. Further, no network centric tuning was done with Mellanox or Intel drivers.
**TestDFSIO Benchmarking Results**

TestDFSIO is a standard example of a benchmark that measures the capacity of HDFS for reading and writing bulk data. The test measures the time taken to create a number of large files, and then uses those same files as inputs as a test to measure the read performance an HDFS instance can sustain. Figure 10 shows the aggregated throughput generated by TestDFSIO benchmark. It can be seen that the 10Gb/s network becomes the bottleneck and in order to fully utilize the cluster one needs to switch to a higher speed 40Gb/s network, in this case using the Mellanox ConnectX-3 Pro 40Gb Ethernet adapter.

![Figure 10. TestDFSIO Per-Node Bandwidth](image)

**Conclusion**

Both TeraSort and TestDFSIO demonstrated that the network often becomes the bottleneck and limits the capabilities provided by MapR-FS. They further show the need to migrate to higher speed network to take the full advantage of flash storage, which is increasingly becoming common on a production grade Hadoop cluster due to their overall favorable ROI. The above benchmarks use standard TCP and UDP transport and additional performance gains are likely by enabling RDMA/RoCE (RDMA over Converged Ethernet) in the future. Our analysis also opens up the opportunity for future testing to understand the benefits of 25, 50 and 100G Ethernet speeds for Hadoop applications like Spark in such environment.