CONFIGURATION BRIEF

Big Data

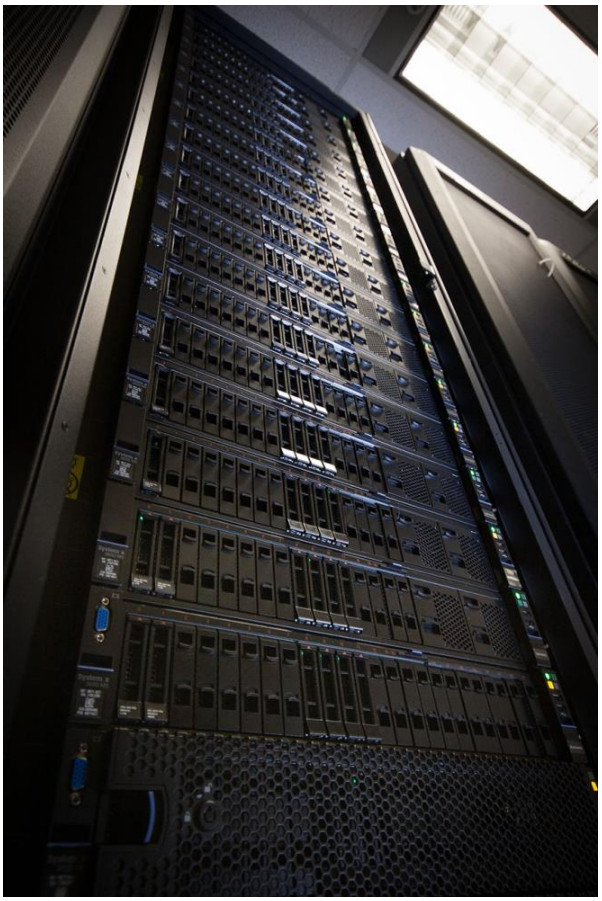
Lenovo Big Data Configuration for Apache Spark

Industry First 100TB Spark SQL Benchmark

Summary

Apache Spark is gaining adoption both for its features and its performance. Apache Spark enables the organizations to realize the real-time analytics and gain faster insights in today's competitive business environment. It provides significant performance advantages to analyze massive datasets with its in-memory processing engine and support for running standard SQL natively on Apache Spark platform. This configuration brief describes the performance and scalability benefits of deploying Apache Spark cluster on the Lenovo server platform using Intel NVMe drives and Mellanox network interface cards. Some of the key advantages of this solution stack are listed below:

- Serves as a reference for optimized Apache Spark Cluster setup
- Illustrates performance benefits using Hadoop-DS[1] benchmark
- Perfectly suited for high data ingest rates
- Delivers extreme performance in a small data center footprint



HIGHLIGHTS

- Unleash extreme Apache Spark cluster performance
- Faster time to analytics with sub-millisecond I/O latency
- Leverage scalability from 1 to 100 TB dataset and beyond
- Optimized solution driven by reliable infrastructure stack
- Standard SQL support with Apache Spark 2.1
- Targeted workloads include SQL & Analytics

CONFIGURATION BRIEF

Lenovo Big Data Configuration for Apache Spark

Intel® Solid State Drive DC P3600 Series AIC (Add-in Card) and 2.5" form factors both include:

- Consistently high IOPS and throughput
- Sustained low latency
- Variable Sector Size and End-to-End data path protection
- Power loss protection capacitor self-test
- Out of band management
- Thermal throttling and monitoring

Mellanox Ethernet Solution delivers:

- World's fastest interconnect with lowest latency and highest bandwidth
- Advanced hardware offloads for IP packet processing
- Uncompromised data integrity with zero-packet loss switches and cables with lowest BER rate of $1e^{-15}$

"The rapid evolution of Spark and its readiness for the next generation of Big Data Spark clusters is demonstrated by these 100TB performance results"

Berni Schiefer, IBM

Lenovo Big Data Configuration for Apache Spark

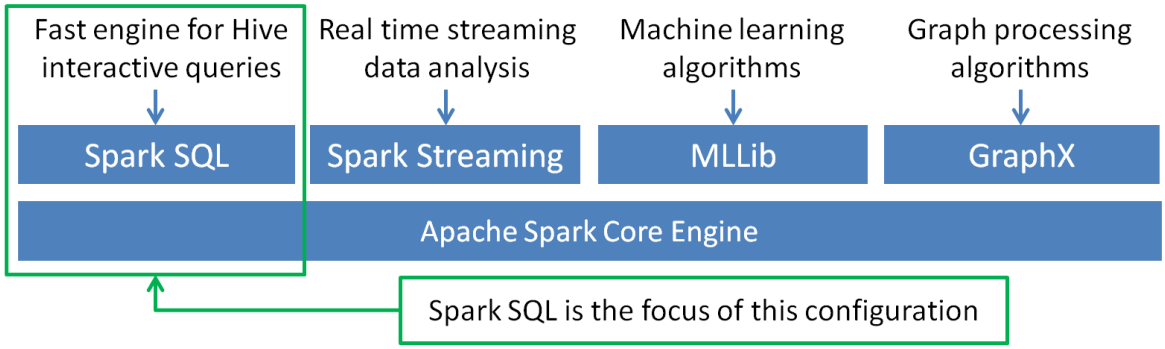
The system is architected with high performance servers featuring high performance storage and networking components. This architecture offers a balanced configuration to manage the compute, storage and networking demands of enterprise scale Apache Spark clusters. . The aggregate capability is provided by 1008 cores, 42TB DRAM, 448TB Flash storage and a 100Gbps Ethernet network. The cluster is running Linux, IBM Open Platform with Apache Hadoop and Apache Spark 2.1 (pre-release).

As a challenging SQL workload we selected the Hadoop-DS workload (a TPC-DS derivative) at a scale of 100TB on a cluster of 28 Lenovo x3650 data nodes along with a pair of Lenovo x3550 management nodes to provide HDFS services. All servers were connected with a 100GbE Mellanox network. Each data node had a pair of Intel E5-2697 v4 (18-core) processors, 1.5TB of DRAM and 8 x 2.0TB Intel NVMe SSD drives.

The goal is to determine how much the very latest CPUs, NVMe solid state storage and high bandwidth networking would improve Data Center Space and energy efficiency while handling a demanding SQL workload with multiple concurrent users. To ensure the latest Spark SQL was used, we took a snapshot of the Apache Spark 2.1 trunk and added select Spark SQL enhancements from the IBM Spark Technology Center.

Performance Overview

Early results are extremely encouraging. With the major uplift in Apache Spark 2.0, Spark SQL was able to parse and obtain plans for all 99 SQL queries using TPC-DS compliant syntax. This is an achievement no other company has demonstrated with open source SQL on Hadoop. 91 of 99 queries were successfully run against 100TB of data in a single user test. During the execution of this test, the cluster achieved peak CPU utilization of 100% as well as I/O throughput of 358 GB/s.



Apache Spark

There is an important new cluster computing framework called Apache Spark in the Big Data world. Over the past year it has become a white-hot Apache project based on a number of measures[2, 3]. Many view Spark as a natural successor to Hadoop MapReduce that preserves the elasticity and resiliency of MapReduce while being easier to program. It also has a richer and better set of programmer APIs and is much better at exploiting abundant dynamic memory.

In addition to the Apache Spark core there are a set of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. Users can combine these libraries seamlessly in the same application. Of these one of the most significant is Spark SQL since SQL remains one of the world's most popular database access technologies[4].

So far, most performance studies of Spark SQL have used more traditional Hadoop cluster topologies, with relatively small servers, modest amounts of memory and large capacity SAS/SATA drives and 10GbE networks. IBM, Intel, Lenovo and Mellanox set out to investigate a topology that leverages advances made in CPU, memory, storage and networking to assess the readiness of Spark SQL to harness these new capabilities.

Key Technologies

Advantages of using Intel® NVMe™ storage – The Intel® Solid State Drive (SSD) Data Center Family for PCIe® brings extreme data throughput directly to Intel® Xeon® processors with up to six times faster data transfer speed than 6 Gbps SAS/SATA SSDs. The performance of a single drive from the Intel SSD Data Center Family for PCIe®, specifically the Intel® SSD DC P3600 Series (450K IOPS), can replace the performance of 7 SATA SSDs aggregated through an HBA (~500K IOPS). The P3600 Series is a PCIe® Gen3 SSD architected with the new high performance controller interface – NVMe™ (Non-Volatile Memory Express™) delivering leading performance, low latency and Quality of Service.

Advantages of using Mellanox 100GbE network – With an increasing need to use faster storage (SSDs, NVMe) for IOPS intensive big data workloads like Spark SQL over HDFS, a faster higher bandwidth network becomes a paramount. While it takes ~10 HDDs per data node to exceed what a 10GbE network can do, it takes just 1 NVMe drive to cross that limit and just 4 to demand a 100GbE network. Network latency also becomes crucial for such transactional-centric workload to deliver consistently low query latency. With 100GbE based network including ConnectX®-4 adapters, Spectrum™ switches and LinkX® cables, Mellanox provides the highest bandwidth and lowest latency to deliver the fastest network needed for faster storage.

Performance Test Methodology

The Transaction Processing Performance Council (TPC) has served as the pre-eminent industry-standard benchmarking organization since 1988.

The TPC-DS benchmark models the decision support functions of a retail product supplier. The supporting schema contains vital business information, such as customer, order, and product data. TPC-DS models industries that must manage, sell and distribute products. It utilizes the business model of a large retail company having multiple stores located nation-wide. The benchmark currently supports database sizes at 1TB, 3TB, 10TB, 30TB and 100TB. A workload consisting of 99 standard ANSI/ISO SQL queries are used to exercise the system.

In 2014, IBM defined Hadoop-DS, a SQL on Hadoop-oriented benchmark inspired by and derived from TPC-DS. Hadoop-DS executes a workload that is largely consistent with the requirements of TPC-DS. It uses the same schema, the same queries and the same execution rules but it omits the data maintenance functions, the second throughput run and does not include any pricing. It also does not go through a formal TPC audit process and as a result it is not a formal TPC result. To date, no company has officially audited and published a TPC-DS result.

Test Results

When running the 4-stream multi-user (throughput) test, a total of 360 queries completed requiring a max of 87% CPU utilization across the cluster and with an average CPU utilization of 49% The 100GbE switch moved a remarkable 13.5 GB/s of traffic, while the flash storage handled 12.8 GB/s of I/O bandwidth per data node (358 GB/s cluster wide). Overall we were able to largely saturate the entire cluster.

These results demonstrate that Spark SQL is well on its way to being able to master challenging SQL workloads such as Hadoop-DS. We can also see that using new advanced servers allows our cluster to handle a large, complex workload rapidly while saving energy and space compared with traditional Hadoop Clusters. As memory and flash storage improve and drop in price, we expect the Spark F1 Cluster design to become pervasively deployed.

In summary, advances in hardware and software are paving the way for a new class of Spark cluster that offers the resiliency and elasticity found in larger Hadoop clusters, the powerful ability to integrate advanced analytic processing, including machine learning into elegant Spark applications, and at the same time able to process demanding SQL-based business intelligence queries.

Solution Configuration

Data Nodes – Lenovo x3650 M5 Servers
The Lenovo System x3650 M5 server is an enterprise class 2U two-socket versatile server that incorporates outstanding reliability, availability, and serviceability (RAS), security, and high-efficiency for business-critical applications. It offers a flexible, scalable design and simple upgrade path to various storage configurations and up to 1.5 TB of TruDDR4 Memory. This reference configuration uses 8 NVMe PCIe SSDs and a dual-port 100GbE network adapter.

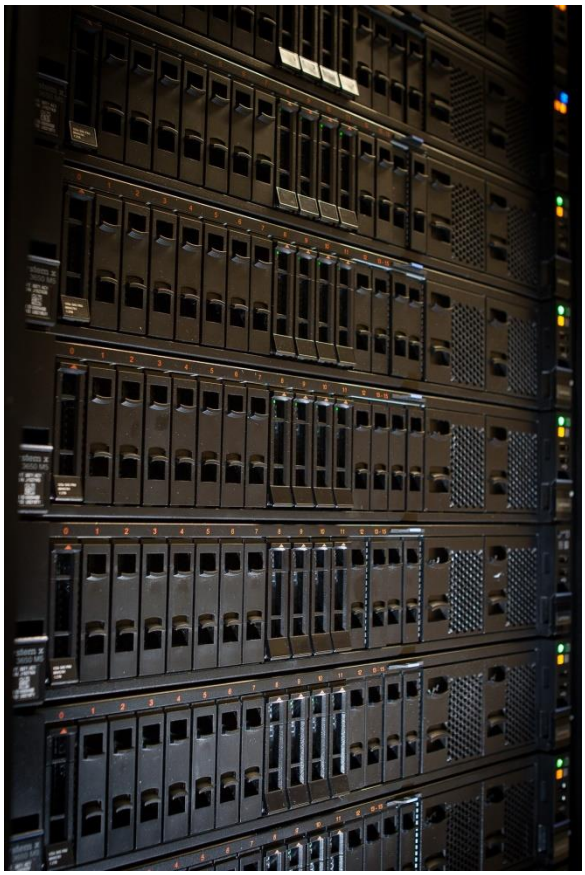
With standards-based Intel® infrastructure that allows for innovation across a wide choice of leading analytics platforms, now you can confidently begin to move your business forward with actionable, real-time insights from advanced analytics. The powerful new Intel® Xeon® processor E5-2600 v4 product family offers versatility across diverse data center workloads including a wide range of scale-out data analytics frameworks.

Management Nodes – Lenovo x3550 M5 Servers
The Lenovo System x3550 M5 server is a cost- and density-balanced 1U two-socket rack server. The x3550M5 features a new, innovative, energy-smart design with up to two Intel Xeon processors of the high-performance E5-2600 v4 product family processors a large capacity of faster, energy-efficient TruDDR4 Memory up to twelve 12Gb/s SAS drives, and up to three PCI Express (PCIe) 3.0 I/O expansion.

Storage - Intel® P3600 NVMe SSD
The Intel® SSD DC P3600 Series is a PCIe* Gen3 SSD architected with the new high performance controller interface, Non-Volatile Memory Express* (NVMe*), delivering leading performance, low latency, and quality of service. Matching the performance with world-class reliability and endurance, Intel SSD DC P3600 Series offers a range of capacity—400 GB, 800 GB, 1.2 TB, 1.6 TB and 2 TB in both add-In card and 2.5-inch form factor.

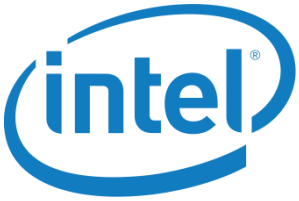
Networking – Mellanox Spectrum SN2700 switch
Mellanox Spectrum-based SN2700 (32x 100GbE ports) switch provides the highest performance fabric solution in a 1U form factor delivering non-blocking throughput for big data workloads, with predictable low-latency and zero-packet loss (ZPL). Due to the bursty network traffic in big data workloads, non-blocking switches play a crucial role in delivering predictable SQL query completion time. In addition, LinkX DAC cables offers reliable connections at speed from 10 to 100Gb/s with highest quality, featuring error rates up to 100x lower than industry standards.

Network Adapters - Mellanox ConnectX-4 100GbE adapters
Mellanox ConnectX-4 Ethernet adapter provides the highest performing and most flexible interconnect solution for Big Data, Cloud and HPC applications at various speed of 10/25 and 40/50/100Gbps. Big Data applications utilizing TCP or UDP over IP transport can achieve the highest efficiency and application density with the hardware-based stateless offload and flow steering engines. These advanced offloads reduce CPU overhead in packet processing and improving application query latency.



Spark F1 Cluster: 28 Data Nodes each provisioned with Intel E5-2697 v4 processors, 16TB of NVMe storage and 1.5 TB DDR4 memory and interconnected with a 100 GbE network yields one incredibly fast Apache Spark platform

The Lenovo x3650 M5 server Data Node and Lenovo x3550 M5 server Management Node bring high performance and high reliability in a small footprint



Why Lenovo

Lenovo is a leading provider of x86 servers for the data center. Featuring rack, tower, blade, dense and converged systems, and the Lenovo server portfolio provides excellent performance, reliability and security. Lenovo also offers a full range of networking, storage, software, solutions, and comprehensive services supporting business needs throughout the IT lifecycle. With options for planning, deployment, and support, Lenovo offers expertise and services needed to deliver better service-level agreements and generate greater end-user satisfaction.

For More Information

To learn more about the Lenovo Big Data Configuration for Apache Spark Platform, contact your Lenovo Business Partner or visit: lenovo.com/systems/solutions.

[1] Hadoop-DS is based on TPC-DS. While results have been reviewed for correctness, they have not been audited or published

[2] https://upside.tdwi.org/articles/2016/08/22/spark-is-hot-hot-hot.aspx

[3] https://adtmag.com/blogs/dev-watch/2016/05/asf-big-data-projects.aspx

[4] http://stackoverflow.com/research/developer-survey-2016