Network Switch Impact on "Big Data" Hadoop-Cluster Data Processing

Comparing the Hadoop-Cluster Performance with Switches of Differing Characteristics

DR160301C
March 2016

Miercom
www.miercom.com

Contents

1 - Introduction

The trend is clear: Monolithic, big-iron computer performance has limits. Recent years have seen the re-deployment of massive data-processing jobs – for scientific and even business applications – onto scalable clusters of tightly networked servers.

Those responsible for their organization's next-gen data-center infrastructure are asking: What are the most important components of clustered computing that affect performance? Is it the servers themselves – processing power, disk read/write speed?

But how about the data-center network that interconnects the servers and server clusters? Are the links connecting the computing nodes a bottleneck? How about the data-center switches? Do their architectures, and aspects such as buffer depth, matter?

Miercom was engaged by Cisco Systems to conduct independent testing of two vendors' top-of-the-line, data-center switch-routers. The objective: Given the same 'big-data' application, topology and benchmark tests, how do they compare in terms of clustered-computing performance?

This report describes the findings in the first set of test that focused on performance of the cluster with uni-dimensional benchmarks (one single benchmark at a time). However, Big Data clusters are more complex with multiple workloads of varying requirements, from compute intensive Analytics to latency-sensitive in-memory queries, running simultaneously. This has an impact on the requirements from the network. Future reports on other test cases will cover this aspect of Big Data clusters.

The topology tested here ran Hadoop – the most popular software framework for the distributed processing of large data sets across clusters of networked servers.

- The Hadoop distribution was deployed across the many servers and server clusters in the test bed.
- The package includes several popular benchmark tests for assessing distributed performance, data flow and bottlenecks. We applied three in this testing:
    - TestDFSIO, a test of distributed file system input/output time,
    - TeraSort; which measures the amount of time to sort 1 Terabyte of randomly distributed data, and
    - TeraGen, which measures the time to generate the TeraSort input data set.
- In addition, for each of these multi-cluster tests, buffer utilization by each switch was closely monitored.

What was tested, and how, and the results relating to the performance of this Hadoop environment are detailed in this paper.

Robert Smithers

CEO

Miercom

2 - Hadoop, and Measuring Hadoop Performance

This study focused primarily on the performance of several of the standard benchmark tests that are included with the Hadoop distribution to complete. We ran Hadoop release 2.7.0, the MapR distribution, from San Jose, CA-based MapR Technologies, Inc. Concurrently, the testers carefully monitored the percent buffer utilization of each of the Leaf switches; see the following test-bed section, to which all the computing nodes directly connected.

Originally considered a scientific experiment, the Hadoop software framework has already been deployed in numerous production environments – including, increasingly, for mission-critical applications. Because it is open-source, Hadoop has become the preferred software base for storing data and running applications on clusters of servers. The software is especially adept at working around server, even whole site, failures. As SAS puts it: "Hadoop provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs." (For an excellent backgrounder on Hadoop, see: http://www.sas.com/en_us/insights/bigdata/hadoop.html?keyword=hadoop&matchtype=p&publisher=google&gclid=CMrNmL--lcsCFYIfhgoda1kEkQ

Standard benchmarks. The focus of this study was to see what effect, if any, the selection of data-center network equipment has on Hadoop performance. Building a realistic Hadoop environment itself is a daunting task, given the number of servers, proper software deployment and installation, and the sheer complexity of orchestrating everything in executing the standard benchmark tests.

The Hadoop distribution includes numerous standard benchmark tests. Our testing applied and focused on three of them:

- **TestDFSIO** – for testing distributed-file-system I/O, namely the interfaces and disks.
- **TeraSort** – a popular benchmark that measures the amount of time to sort one terabyte (TB) of randomly distributed data.
- **TeraGen**, which measures the time it takes to generate the terabyte data set that is processed by the TeraSort test.

As explained more in the following Test Bed section, three different switches were tested. That means the test bed was built, the first time with a pair of Cisco Nexus 9272Q switches, and all the tests were run.

Then the two switches were replaced with a pair of Cisco Nexus 92160YC switches, and all the tests were re-run, in the same order and in exactly the same way. Finally, these two Cisco switches were replaced with a pair of Arista 7280 switches, everything was reconnected and the tests re-run again, in exactly the same manner.

Metrics. The main test objective was to record the Test Completion Time (TCT) for the TestDFSIO, TeraGen and TeraSort tests. The TCT embodies all the I/O and/or processing tasks, and constraints, of the systems executing the test, as well as the network infrastructure that interconnects the servers and clusters.

Other, secondary measurements recorded during and after each test included:

- Interface throughput on the switches
- Buffer utilization on the switch interfaces
- Disk I/O throughput on the servers
- Packet drops
- TCP segment resets and segment retransmissions
- CPU/RAM utilization, this to make sure that these were not bottlenecks.

Some additional details are useful in understanding the tests that were run and the control-parameter settings applied.

TestDFSIO

TestDFSIO is a read and write test for Hadoop clusters, which simulates the functionality of moving large data in and out of the cluster. It is used for initial stress testing, to discover performance bottlenecks, to shake out the cluster configuration, and to get a first impression of how fast the cluster I/O is. The test imposes load on the I/O elements, namely the interfaces and disks, and not as much on CPU and RAM, as processing is minimal.

For the DFSIO testing, various large file sizes totaling 6 TB (terabytes) were tried: 1,000 files of 6 GB each; 2,000 files of 3 GB each; and 3,000 files of 2 GB per file. No appreciable changes were observed in test completion times between the different file-size combinations, so subsequent tests – and all the tests for the record – were carried out with the combination of 3,000 files of 2 GB per file.

Hadoop creates batches of files, so we optimized the number and size of files for the best batching across our 60 servers. We found that 600 MB per batch, yielding five batches, was measured to deliver the best performance. So this setting was used for all the DFSIO tests.

TeraSort

TeraSort is a popular benchmark test that measures the amount of time it takes a computing node to sort one terabyte of randomly distributed data. Its purpose is to sort a large amount of data as fast as possible. TeraSort simulates a typical big-data workload that reads from disk, performs intensive processing, and then writes back to disk – imposing stress on I/O as well as processing elements.

TeraSort is commonly used to measure the MapReduce performance of a Hadoop cluster. MapReduce is a program model and associated software for processing and generating large data sets on a server cluster with a parallel, distributed algorithm.

Running the TeraSort benchmark test is a two-step process:

1. The input data is generated via TeraGen (see below).
2. Running the actual TeraSort on the input data.
3. Once the TeraGen is run, the TeraSort is run on the generated data independent of the TeraGen run.

TeraGen

TeraGen measures the time it takes to generate the terabyte data set that is processed by the TeraSort test. The user specifies the number of rows and the output directory and the process runs a MapReduce program to generate the data.

TeraGen's workload is I/O intensive and entails minimal processing.

Control parameters

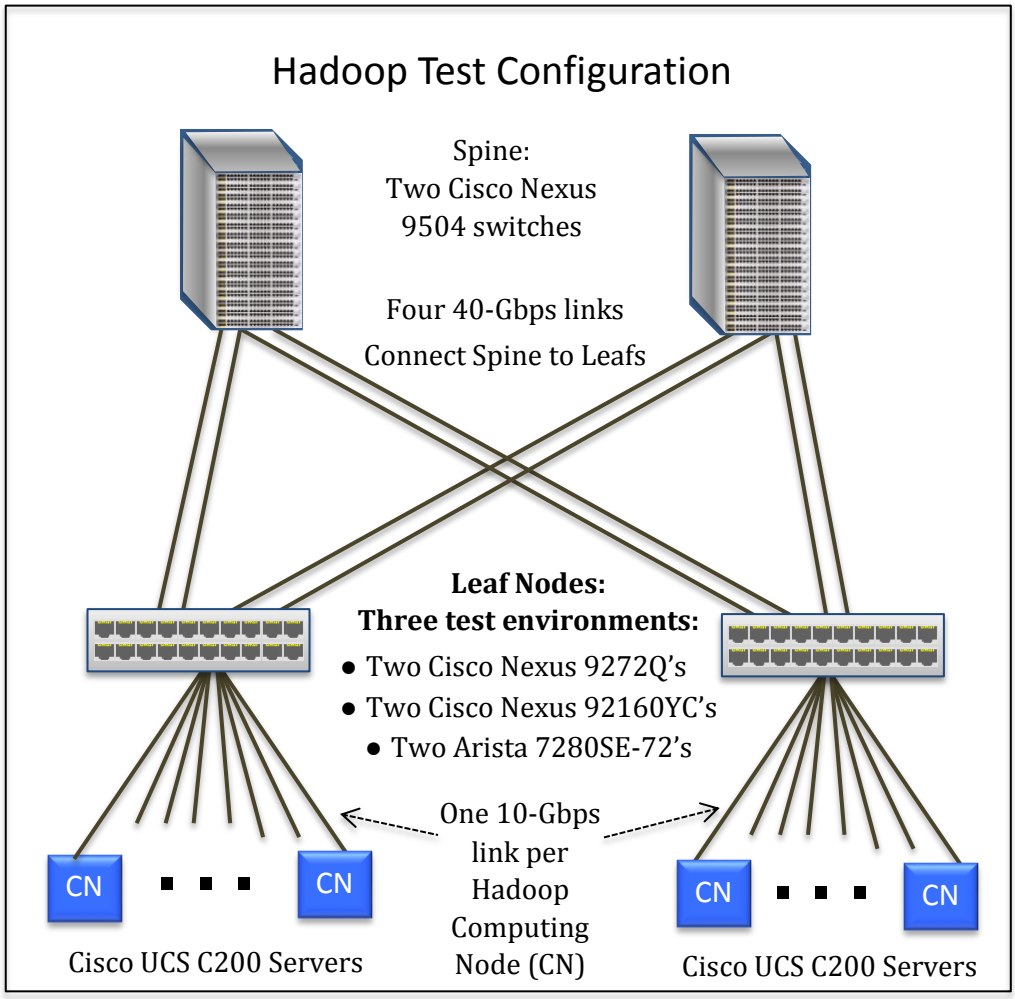
One of the relevant control parameters that had to be set for the tests was TCP Window Size. We found that changing TCP buffer size between 4 MB, 16 MB and 32 MB all yielded the same results in test completion times, so all tests were run using a 32MB TCP buffer.

Hadoop spawns one map task for each file in a DFSIO task. The total number of map tasks per node = no. of disks x overprovisioning factor per disk. This translates to approximately six map tasks per node and five batches of execution, for each of the 50 to 60 nodes.

For TeraSort and TeraGen, 1 TB of data was generated using TeraGen and subsequently used for sorting in the TeraSort test. TeraGen generated 512 files of 2 MB each, translating to 512 map tasks. TeraSort results were optimized for 2,048 Maps and 700 to 800 reducers.

3 - Test Bed: The Hadoop Configuration Tested

To see how different switch-routers affect performance in terms of the Hadoop benchmark tests applied, a realistic test bed representing a data-center network infrastructure was assembled, see below diagram.



The test-bed data-center network featured two spine switches – Cisco Nexus 9504 modular switches with Cisco Nexus 9636Q Linecards. Each supported four 40-Gbps links, two to each "leaf" node, as shown above.

The leaf nodes were the actual devices tested. Three pairs of switches were tested, one pair at a time. The switches tested were:

- Cisco Nexus 92160YC
- Cisco Nexus 9272Q
- Arista 7280SE-72.

The leaf switches connected to a total of 63 blade servers in Cisco UCS C200 High-Density Rack-Mount Server systems. One switch connected 31 servers, each via a 10-Gbit/s link. The other switch similarly connected 32 servers.

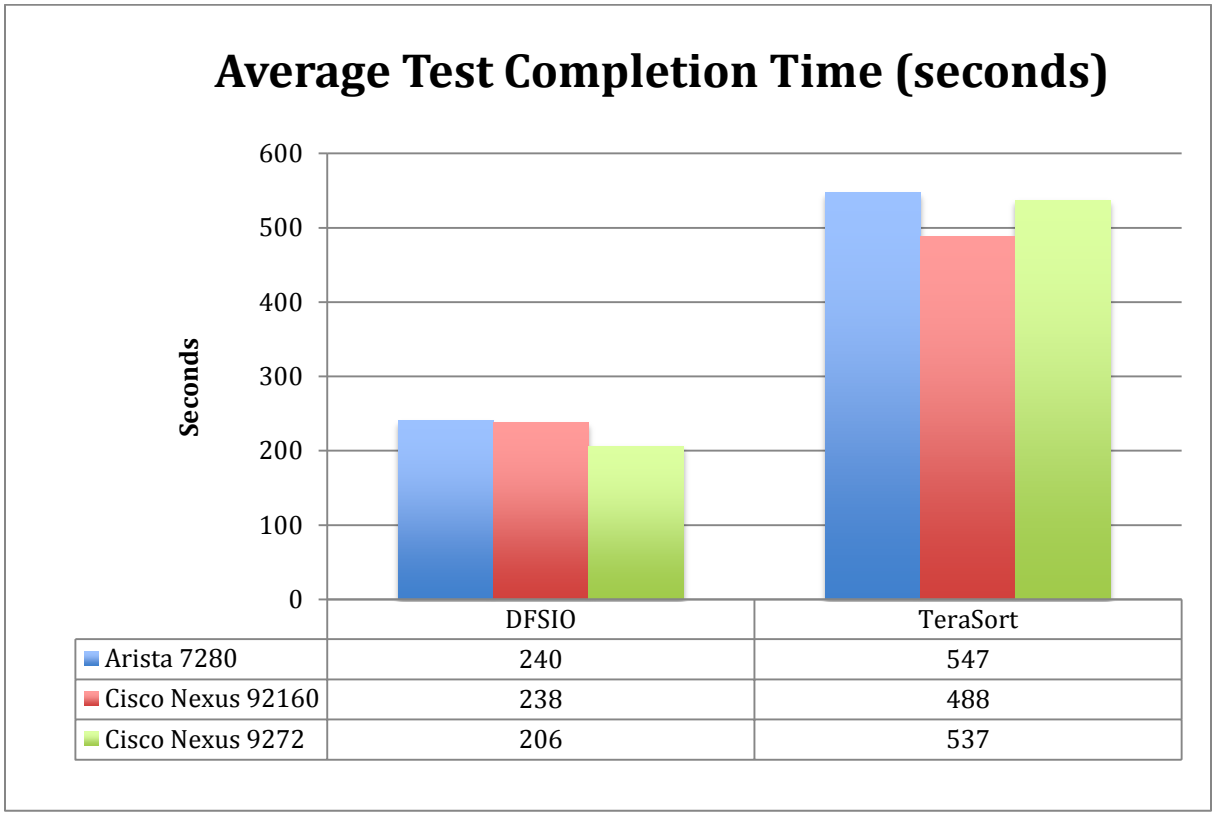
Though 63 servers were available, only 58 took an active role in the Hadoop configuration at a time.

For the Hadoop configuration, each server – or Computing Node (CN) – had three 2-TB disks assigned to Hadoop. Six of the servers were designated Hadoop Control Nodes, while 52 were set-up as Hadoop Data Nodes.

4 - Results

Test Completion Time

The chart below shows the average completion times for each switch, for each of the three Hadoop benchmark tests. Each test for each switch was conducted multiple times.



	DFSIO	TeraSort
Arista 7280	240	547
Cisco Nexus 92160	238	488
Cisco Nexus 9272	206	537

The test completion times across the platforms are comparable, even though some aspects of the switches – such as per-port buffer depth – are quite varied.

The differences observed from one test run to the next are due to the dynamic behavior of the Hadoop configuration across the servers during the tests. Hadoop spawns jobs across various servers and the results can vary based on the server chosen for critical maps/reducer and the state of the server during that time. Results could still vary during different test runs even with the exact same Hadoop setup.

Other Key Observations

As noted, several other switch metrics were observed and recorded in addition to the Hadoop benchmark tests. Among the key observations:

- Buffer utilization was low across the board, for all the switches tested, during the Hadoop benchmark tests.
- The throughput achieved during the tests were typically around 1Gbps, but less than line rate.
- Disk I/O reads and writes, consumed the most time in the tests.
- Very few TCP Resets and Segment Retransmits occurred, indicating low congestion levels.

Buffer Utilization. Terasort was the most intensive of tests in terms of buffer utilization. Below table summarizes the buffer consumption across the different switches.

The data shows that while plenty of buffer space was available for Arista switches, the buffer went largely unused for all of these standard Hadoop benchmark tests. Size of buffers, then, was never an issue.

	Cisco Nexus 9272Q	**Cisco Nexus 92160**	**Arista 7280**
Documented Buffer Size	Max 5 MB/Port	Max 10 MB/Port	50+ MB per port
Max Buffer in Use in Testing	1.2 MB /Port (terasort)	1.9 MB/Port (terasort)	< 5 MB/Port
Intelligent Buffer Capacity	Yes	Yes	No

Line rates. Link utilization (load) and disk I/O rates were also monitored during test runs. The maximum and average link loads – of the 10-Gbit/s capacity available to each server, are shown below for the three switches for the TeraGen tests.

Max and average link loads during TeraGen tests

	TeraGen Line Rate (Mbit/s)					
	Cisco 92160YC		Cisco 9272Q		Arista 7282	
	Rx	Tx	Rx	Tx	Rx	Tx
Max	1542	1657	907	874	1327	1684
Average	1208	1214	587	585	889	958

During all the tests, maximum and average interface traffic of less than 2 Gbit/s was observed across the interfaces. The Cisco Nexus 92160YC, on average, used just 0.6 Gbit/s of the link's 10-Gbit/s capacity.

Disk rates. The TeraGen test makes substantial use of disk writes, there are no disk reads in the TeraGen test. The table below shows the amount of disk I/O – in MBps of data, for the TeraGen test.

Disk I/O – MBps (megabytes per second) of disk writes, during TeraGen tests

TeraGen Disc Rate - megabytes (MB) per second			
	Cisco 92160YC	Cisco 9272Q	Arista 7282
	Writes	Writes	Writes
Max	223	191	147
Average	82	83	129

The TeraGen test moved an average of about 100 MBps of disk I/O data, disk writes, through the Cisco Nexus 92160YC, with peaks of about 220 megabytes per second. The TeraGen tests exhibit few periods of inactivity, which is why the average values approach the maximums.

Packet drops. No switch packet drops were observed with either Cisco or Arista switches, indicating there were no buffer overloads. Packet drops were observed during some tests with Arista, but are attributable more to Hadoop processes and excess re-transmits than to network switch or link issues.

5 - Conclusions

- Big Data does not need big buffers
  - With the benchmarks that were run as part of this study, there was no discernable difference in the performance of the cluster with different switching platforms. Performance was measured primarily as job completion time for the DFSIO, TeraGen and TeraSort tests.
  - A large buffer is not required for dealing with the Hadoop tests. All buffer-occupancy figures revealed that average and maximum buffer utilization in all Hadoop test cases were very low. Across the entire set of benchmarks, the maximum instantaneous buffer occupancy that was noticed was under 15% of available buffers.
  - Based on the study conducted, it is clear that buffer sizes of the networking infrastructure has minimal impact on the performance of the Hadoop clusters, compared to other elements such as server RAM, CPU and disk I/O, which more directly impact the number of maps/reduce containers that can be spawned on a server.

Additional Comments

- There are several more important network design considerations for a Hadoop cluster:
  - Availability and resiliency: Network design should ensure a high network availability for the cluster
  - Node and network connectivity: Hadoop clusters need to standardize on 10G or better server connectivity and move away from 1G networks that were common in the past. Low over-subscription in the network with utilization of 40G/100G technologies will provide better network performance
  - Burst handling and queue management: Network devices need adequate buffers that are combined with intelligent buffer management algorithms to recognize congestion buildup and handle it proactively.
- Intelligent buffers are more important than big buffers
  - Numerous academic studies have pointed to the need for intelligent congestion management techniques as being more important than big buffers (E.g., http://simula.stanford.edu/~alizade/papers/conga-sigcomm14.pdf)
  - With multiple simultaneous workloads running in a cluster, it is important to be able to detect and manage various congestion scenarios, like elephant-mice contention, network-failures, etc. Technologies like Dynamic Load Balancing and Dynamic Packet Prioritization implemented on the Cisco Nexus switches can provide a distinct advantage in these scenarios. A recent Miercom study concluded that switches with intelligent buffer management algorithms outperform switches with big buffers: Miercom Report: Cisco Systems Speeding Applications in Data Center Networks

6 – References

Why flow-completion time is the right metric for congestion control
http://yuba.stanford.edu/techreports/TR05-HPNG-112102.pdf

Experimental Study of Router Buffer size
http://yuba.stanford.edu/~nickm/papers/IMC08_buffersizing.pdf

On the Data Path Performance of Leaf-Spine Datacenter Fabrics
https://people.csail.mit.edu/alizadeh/papers/hoti13.pdf

DCTCP: Efficient Packet Transport for the Commoditized Data Center
http://research.microsoft.com/pubs/121386/dctcp-public.pdf

VL2: A Scalable and Flexible Data Center Network
http://research.microsoft.com/pubs/80693/vl2-sigcomm09-final.pdf

PIE: A lightweight control scheme to address the bufferbloat problem
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=6602305&abstractAccess=no&userType=inst

Hadoop Benchmarking

http://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench/

Nexus 9000 configuration Guide

http://www.cisco.com/c/en/us/td/docs/switches/datacenter/nexus9000/sw/7-x/interfaces/configuration/guide/b_Cisco_Nexus_9000_Series_NX-OS_Interfaces_Configuration_Guide_7x.pdf

7 - About Miercom Testing

This report was sponsored by Cisco Systems, Inc. The data was obtained completely and independently by Miercom engineers and lab-test staff as part of our performance verification testing. Testing such as this is based on a methodology that is jointly co-developed with the sponsoring vendor. The test cases are designed to focus on specific claims of the sponsoring vendor, and either validate or repudiate those claims. The results are presented in a report such as this one, independently published by Miercom.

8 - About Miercom

Miercom has published hundreds of network-product-comparison analyses in leading trade periodicals and other publications. Miercom's reputation as the leading, independent product test center is undisputed.

Private test services available from Miercom include competitive product analyses, as well as individual product evaluations. Miercom offers comprehensive certification and test programs including: Certified Interoperable, Certified Reliable, Certified Secure and Certified Green. Products may also be evaluated under the Performance Verified program, the industry's most thorough and trusted assessment for product performance.

9 - Use of This Report

Every effort was made to ensure the accuracy of the data contained in this report but errors and/or oversights can occur. The information documented in this report may also rely on various test tools, the accuracy of which is beyond our control. Furthermore, the document relies on certain representations by the vendors that were reasonably verified by Miercom but beyond our control to verify to 100 percent certainty.

This document is provided "as is," by Miercom and gives no warranty, representation or undertaking, whether express or implied, and accepts no legal responsibility, whether direct or indirect, for the accuracy, completeness, usefulness or suitability of any information contained in this report.