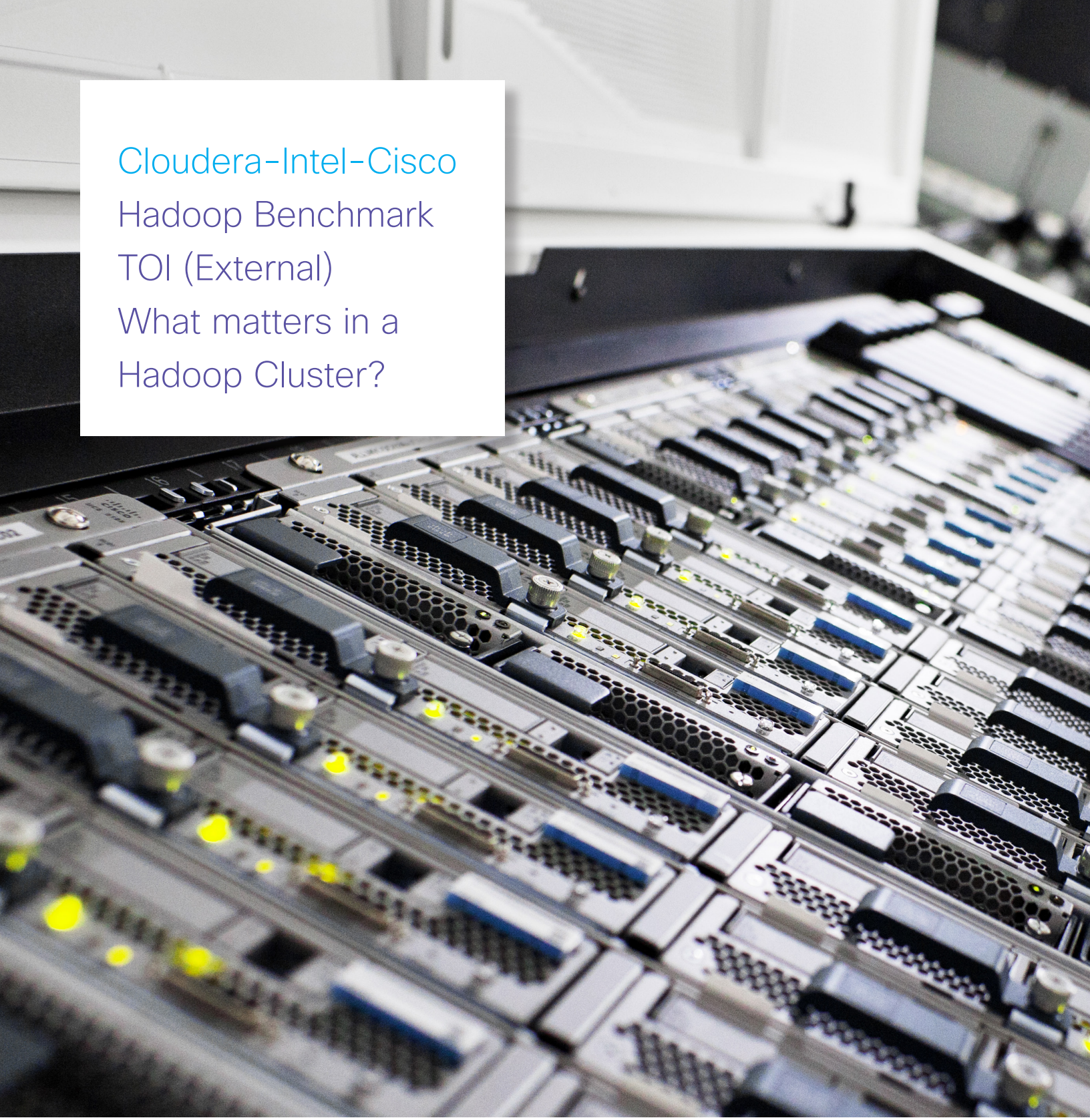


Cloudera-Intel-Cisco Hadoop Benchmark TOI (External) What matters in a Hadoop Cluster?

Floris Grandvarlet (Cisco) floris.grandvarlet@cisco.com
Patrick Schotts (Intel) patrick.schots@intel.com
Woody Christy (Cloudera) wchristy@cloudera.com

cisco

cloudera

intel

Acknowledgments

The authors acknowledge the contributions of:

Intel:

Stephen G. Anderson, stephen.g.anderson@intel.com
Rob Kypriotakis, rob.kypriotakis@intel.com
Jacob A. Ohara, jacob.a.ohara@intel.com
Gert Pauwels, Gert.Pauwels@intel.com
Richard B. Pilling, richard.b.pilling@intel.com

Cisco:

Arnaud Bassaler, abassale@cisco.com
Peter Ruttens, pruttens@cisco.com
Michel Sumbul, msumbul@cisco.com
Karthik Kulkarni, kkulkar@cisco.com

Cloudera:

Sandeep Brahmarouthu, sandeep@cloudera.com
Jonathan Cooper, jcooper@cloudera.com
Rob Johnson, rj@cloudera.com
Kunal Kusoorkar, kkusoorkar@cloudera.com
Dwai Lahiri, dlahiri@cloudera.com
Jonathan Seidman, jseidman@cloudera.com

Contents

Cloudera-Intel-Cisco Hadoop Benchmark TOI (External) What matters in a Hadoop Cluster?

1. Introduction

This benchmark intends to look at the physics of Hadoop. Today, it is pretty easy to set up a working Hadoop environment where the base default configuration works and the jobs are completed.

But, are you really happy if the job completes in three hours at 10% utilisation?

Also, success can easily be a curse, with management realizing the benefits and looking for richer insight, meaning larger more complex and more numerous jobs, but Hadoop does not self-optimize.

You now have 3 options :

1. Buy more hardware
2. Buy expertise
3. Optimize the platform as-is to get the best performance

Because of this, an idea came about in April 2014 as we (Intel/Cisco) discussed the setup of a demonstration Hadoop cluster. We decided to extend the initial setup work to include a benchmark of the physics of Hadoop before launching the demo-platform. When we talk about the physics of Hadoop, we are looking at various aspects of the platform:

- Platform processor choice
- Base Hardware platform
- Network stack
- Disk partitioning/filesystem
- Base Operating System
- Cloudera software
- Data workload

This paper provides all the necessary information to reproduce the same test-bed, as well as a summary of the results and the lesson learned. This information is valuable in designing and building Hadoop clusters.

The intent of this benchmarking effort is not compete with other published record benchmarks; rather it done mainly to demonstrate the relative impacts of the choices of CPU, Network BW, HDD, etc. for your Hadoop platform. The final results are expressed as a relative percentage of the baseline. (Raw data available in the Appendix section).

Executive Summary

This benchmark intends to look at the physics of Hadoop and how to optimize your Hadoop platform.

When we talk about the physics of Hadoop, we are looking at various aspects of the platform, including:

- Platform processor choice
- Base Hardware platform
- Network stack
- Disk partitioning/filesystem
- Base Operating System
- Cloudera software
- Data workload

The intent of this benchmarking effort is not to compete with other published record benchmarks; rather it was done mainly to demonstrate the relative impacts of the choices of CPU, Network BW, HDD, etc., for your Hadoop platform. The final results are expressed as a relative percentage of the baseline. (Raw data available in the Appendix section).

As such, we benchmarked CPU, HyperThreading, Networking BW (10/5/1 Gb), HDD (12/24) as well as scalability (4/5/6 data nodes), using the usual suspects–TeraGen, TeraSort, TeraValidate (1TB and 2TB), as well as wordcount (~12GB). We adopted a multi-run (x3, average value) benchmarking strategy with a max deviation of 20% between runs, which if it occurred triggered a full re-test.

The test-bed :

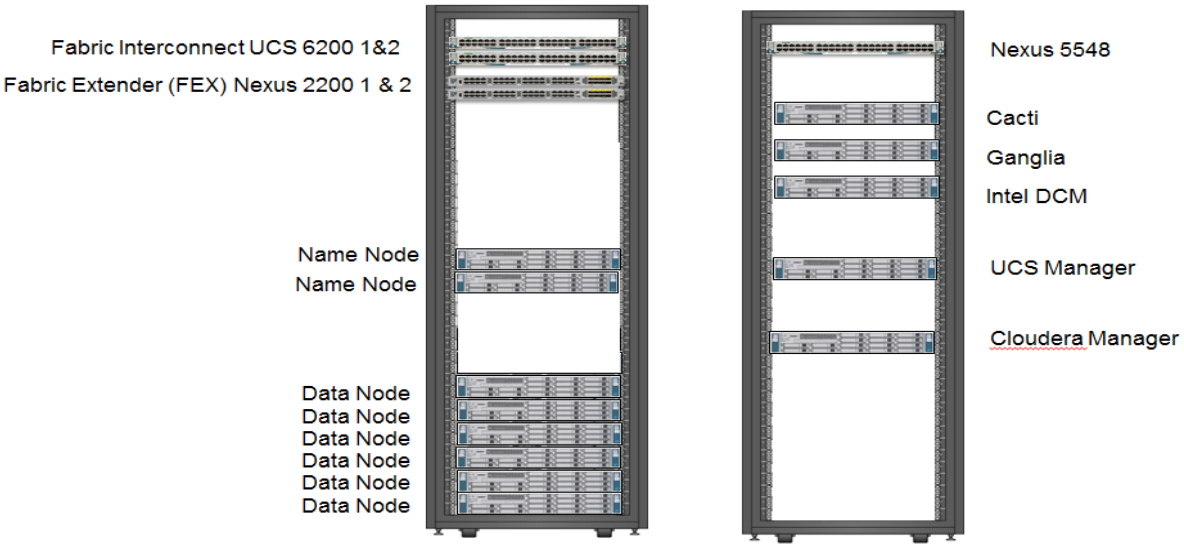


Figure 1 Test-Bed

All details have been provided to enable you to reproduce/modify these tests for your own purposes, with the results expressed as a relative percentage of the baseline. (Raw data available in the Appendix section).

In conclusion, it appears that:

The best CPU is the E5-2690 v2, which presents the best balance of cores and frequency.

Network BW is definitively on the 10 GE side.

Hyper-Threading needs to be set to "ON".

HDD of 24 provides a benefit, as expected.

Scalability is proven, also as expected with Hadoop.

Security is transversal and provided as a gentle reminder, as we did not consider security when setting up the benchmark, but it is mandatory for the operational environment.

Overall, the main lesson learned is that knowledge and review of the entire Hadoop platform is a must for decent performance (feel free to read the epilogue to learn about our initial results).

2. Benchmark Test bed

2.1. Hardware

The Cloudera-Intel-Cisco cluster for this benchmark is composed of 6 Data Nodes and 2 Name Nodes (active/standby).

- 8 x Cisco C240 M3 rack server.
- 256GB total memory for each node (16GB DDR3-1600-MHz RDIMM/PC3-12800/dual rank/1.35v)
- Data Nodes: 24 individual RAID0 volume with 1 HDD per volume.
- Name Nodes: 8 HDDs configured as RAID6 with 4 hot-spare HDDs (12 HDDs free)
- Each HDD is a 600GB 6 GbpsSAS 10K RPM SFF HDD

We also used a couple of UCS rack servers to host Monitoring/Management platform in the backend (Cacti, Ganglia, Intel DCM, UCS Manager, Cloudera Manager).

The network is architectured around:

- **Cisco UCS 6200 Series Fabric Interconnects** provide high-bandwidth, low-latency connectivity for servers, with integrated, unified management provided for all connected devices by Cisco UCS Manager.
- **Cisco Nexus 2200 Series Fabric Extenders (FEX)** provide an optional extended single point of management (extending the network into each rack, acting as remote line cards for fabric interconnects and providing highly scalable and extremely cost-effective connectivity for a large number of nodes). For the high-performance BW test, we did not use the FEX.
- **Cisco UCS Virtual Interface Cards (VICs)** are unique to Cisco. Cisco UCS Virtual Interface Cards incorporate next-generation converged network adapter (CNA) technology from Cisco, and offer dual 10-Gbps ports designed for use with Cisco UCS C-Series Rack-Mount Servers.

Using a Nexus 5548 switch to connect all management workstations (Cloudera Manager, UCS Manager, monitoring tools/suites…)

2.2. Software

Cloudera Distribution of Hadoop (known as CDH) is a popular enterprise-grade, hardened distribution of Apache Hadoop and related projects. CDH is 100 percent Apache-licensed open source and offers unified batch processing, interactive SQL, and interactive search, and role-based access controls.

Similar to Linux distribution, which gives you more than Linux, CDH delivers the core elements of Hadoop; scalable storage and distributed computing, along with additional components such as a user interface, plus necessary enterprise capabilities, such as security and integration with a broad range of hardware and software solutions.

The integration and the entire solution is thoroughly tested and fully documented. By taking the guesswork out of building a Hadoop deployment, CDH provides a streamlined path to success in solving real business problems.

For more information about what projects are included in CDH, see CDH Version and Packaging information:
http://www.cloudera.com/content/cloudera-content/cloudera-docs/CDH5/latest/CDH-Version-and-Packaging-Information/CDH-Version-and-Packaging-Information.html

- CentOS: 6.5 (Final)
- UCSM version: 2.2(1d)
- BIOS C240 M3: 3.1.5.4f.0.1113201
- Cloudera Hadoop 5.0.1

2.3. Software post-installation configuration

All other values are at their specified defaults except for the following

Configuration item	Final tuning
dfs.namenode.handler.count	160
fs.trash.interval	1440
io.file.buffer.size	64kb
yarn.nodemanager.resource.memory-mb (all resource groups)	152GB
yarn.nodemanager.resource.cpu-vcores	TBC during some tests
yarn_scheduler_minimum_allocation_mb	1GB
yarn.scheduler.maximum-allocation-mb	96GB
mapreduce.map.memory.mb	1GB
mapreduce.reduce.memory.mb	1GB
mapreduce.map.java.opts	CD manager Defaults
mapreduce.reduce.java.opts	CD manager Defaults
mapred_compress_map_output	TRUE
mapred_map_output_compression_codec	org.apache.hadoop.io.compress.SnappyCodec
mapred_reduce_parallel_copies	32
mapreduce.task.io.sort.mb	512
mapreduce.map.sort.spill.percent	0.8
mapreduce.task.io.sort.factor	64
CDH version	5.0.1

Additional Linux and cluster changes:

- CPU scaling set to performance to force the CPU to always run at the highest frequency
  - Please note that installing unnecessary services will cause resources to be assigned to these services rather than be free for use by services needed to process the workload. We chose to disable these unnecessary services for all phases of testing;unused services such as impala and Hbase were disabled;Impala/Hbase systems need to be set to performance
- vm.swappiness set to vm.swapiness=0
- Transparent Huge Pages Defrag set to OFF.
- Network tuning
  - Set ring buffer to 4096
  - Tuned memory settings in sysctl.conf
  - Enabled jumbo frames on the LAN and the servers
  - lowered IRQ rate and increased throughput

Changes in sysctl.conf:

```
#10 Gbps settings
TCP/IP memory tuning
net.ipv4.tcp_timestamps = 0
net.ipv4.tcp_sack = 0
net.ipv4.tcp_rmem = 4096 134217728 134217728
net.ipv4.tcp_wmem = 4096 134217728 134217728
Core memory tuning
net.core.rmem_max = 134217728
net.core.wmem_max = 134217728
net.core.rmem_default = 134217728
net.core.rmem_default = 134217728
net.core.optme_max = 134217728
net.core.netdev_max_backlog = 250000
# recommended default congestion control is htcp
net.ipv4.tcp_congestion_control=htcp
# recommended for hosts with jumbo frames enabled
net.ipv4.tcp_mtu_probing=1
vm.swappiness = 0
```

- onBoot.sh script added to /etc/rc.local to apply on every boot.

onBoot.sh script:

```
#Set all CPU to max freq
for i in {0..47};do echo performance> /sys/devices/system/cpu/cpu$i/cpufreq/scaling_
governor;done
#Disable Transparent Huge Page defrag
echo never> /sys/kernel/mm/transparent_hugepage/defrag
#Transmit to NIC card 4096 Bytes at a time
#setpci -v -d 1137:0043 e6.b=2e
#Jumbo frames
ifconfig eth0 mtu 9000
ifconfig eth0 txqueuelen 1000
```

2.4. Architecture

This benchmark architecture is based on CPAv2 for Big Data with Cloudera-based architecture, scaled down to a total of 8 nodes.

The Cisco UCS solution for Cloudera is based on Cisco UCS Common Platform Architecture Version 2 (CPAv2) for Big Data, a highly scalable architecture designed to meet a variety of scale-out application demands with seamless data integration and management integration capabilities built using the following components:

- **Cisco UCS 6200 Series Fabric Interconnects** provide high-bandwidth, low-latency connectivity for servers, with integrated, unified management provided for all connected devices by Cisco UCS Manager. Deployed in redundant pairs, Cisco fabric interconnects offer the full active-active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications. Cisco UCS Manager enables rapid and consistent server configuration using service profiles, automating ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

- **Cisco UCS 2200 Series Fabric Extenders** extend the network into each rack, acting as remote line cards for fabric interconnects and providing highly scalable and extremely cost-effective connectivity for a large number of nodes. For the high-performance BW test, we did not us the FEX(see 3.5).
- **Cisco UCS C240 M3 Rack-Mount Servers** are 2-socket servers based on Intel Xeon E5-2600 v2 series processors and supporting up to 768 GB of main memory. The 24 Small Form Factor (SFF) disk drives are supported in a performance optimized option and 12 Large Form Factor (LFF) disk drives are supported in the capacity option, along with 4 Gigabit Ethernet LAN-on-motherboard (LOM) ports. While the UCS Manager provides global manageability of both Rack and Blade servers (same profiles, templating, etc.), we use a rack-mounted solutionaligned to the essence of Hadoop, clustering tightly coupled CPU-Storage capacity for data locality and processing efficiency.
- **Cisco UCS Virtual Interface Cards (VICs)** are unique to Cisco. Cisco UCS Virtual Interface Cards incorporate next-generation converged network adapter (CNA) technology from Cisco, and offer dual 10-Gbps ports designed for use with Cisco UCS C-Series Rack-Mount Servers. Optimized for virtualized networking, these cards deliver high performance and bandwidth utilization and support up to 256 virtual devices.
- **Cisco UCS Manager** resides within the Cisco UCS 6200 Series Fabric Interconnects. It makes the system self-aware and self-integrating, managing all of the system components as a single logical entity. Cisco UCS Manager can be accessed through an intuitive graphical user interface (GUI), a command-line interface (CLI), or an XML application-programming interface (API). Cisco UCS Manager uses service profiles to define the personality, configuration and connectivity of all of the resources within Cisco UCS, radically simplifying provisioning of resources so that the process takes minutes instead of days. This simplification allows IT departments to shift their focus from constant maintenance to strategic business initiatives.

2.5. Server Configuration and Cabling

The Cisco UCS C240 M3 Rack Server is by default equipped with Intel Xeon E5-2660 v2 processors, 256 GB of memory, Cisco UCS Virtual Interface Card 1225 Cisco, Cisco LSI MegaRAID SAS 9271 CV-8i storage controller and 24 x 1TB 7.2K SAS disk drives.

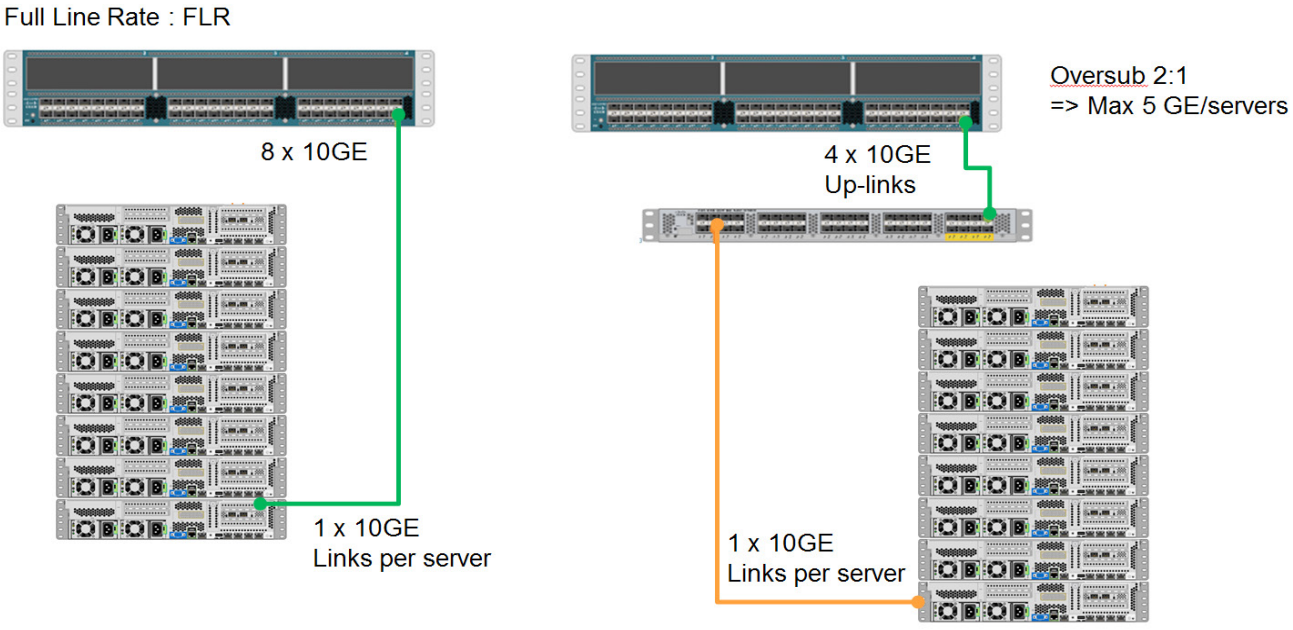


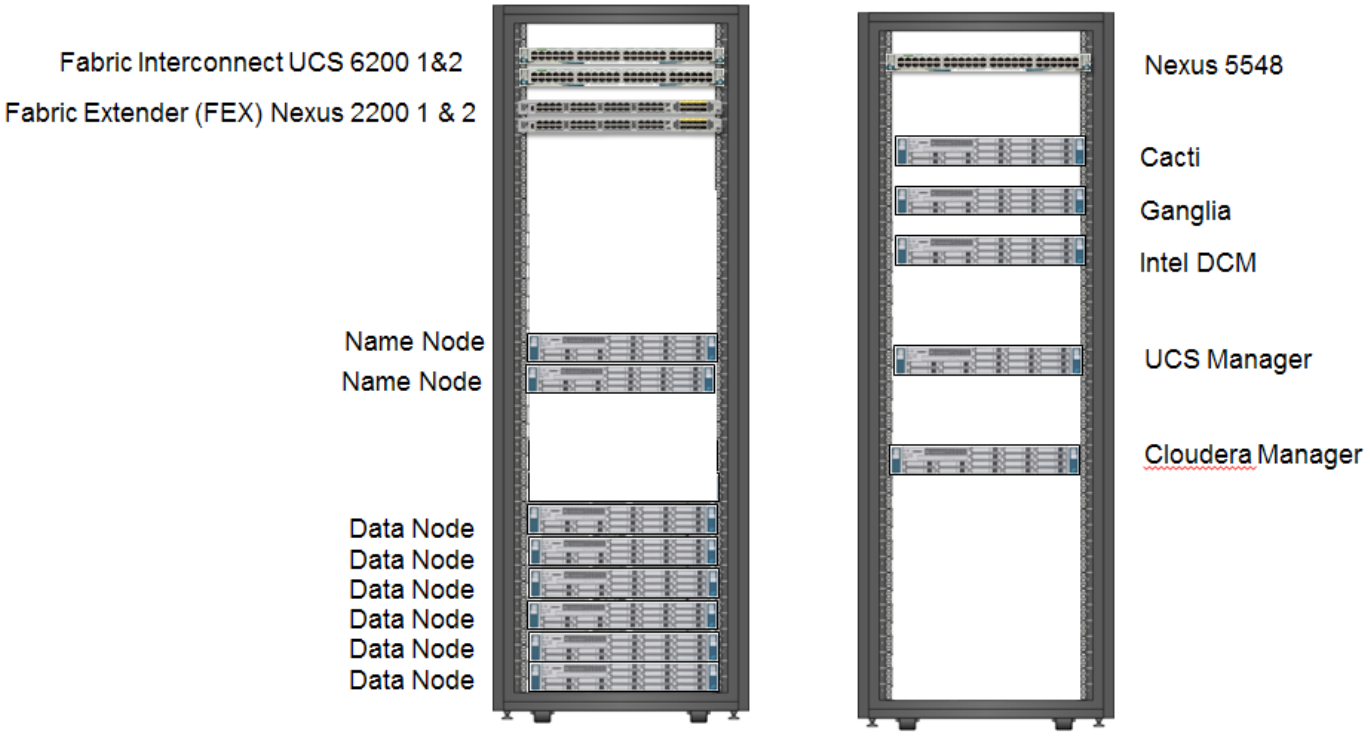
Figure 2 Network Cabling both 10GE and 5/1GE

For more information on physical connectivity and single-wire management see:
http://www.cisco.com/en/US/docs/unified_computing/ucs/c-series_integration/ucsm2.1/b_UCSM2-1_C-Integration_chapter_010.html

For more information on physical connectivity illustrations and cluster setup, see:
http://www.cisco.com/en/US/docs/unified_computing/ucs/c-series_integration/ucsm2.1/b_UCSM2-1_C-Integration_chapter_010.html#reference_FE5B914256CB4C47B30287D2F9CE3597

Figure 3 depicts a 64-node cluster, and each link represents 8 x 10 Gigabit links.

2.6. Rack



3. CPU Benchmark

3.1. Overview

The CPU selection process considered the attribute of: core count, base and effective turbo frequency as well as power consumption in a limited set of processors to be tested, and GPL pricing[1] added on the side for information. You can see below the 5 CPU selected, trying to be representative of Highest-frequency, Highest number of core and mean pricing.

Processor SKU	TDP (W)	Base Freq. (GHz)	# of Cores	Maximum Frequency (+ x00 MHz over base frequency)												
				Number of Active Cores												
				1	2	3	4	5	6	7	8	9	10	11	12	
E5-2697 v2	130W	2.7	12	+8	+7	+6	+5	+4	+3	+3	+3	+3	+3	+3	+3	
E5-2695 v2	115W	2.4	12	+8	+7	+6	+5	+4	+4	+4	+4	+4	+4	+4	+4	$2336
E5-2690 v2	130W	3.0	10	+6	+5	+4	+3	+3	+3	+3	+3	+3	+3	N/A	N/A	$2057
E5-2687W v2	150W	3.4	8	+6	+5	+4	+3	+2	+2	+2	+2	N/A	N/A	N/A	N/A	
E5-2680 v2	115W	2.8	10	+8	+7	+6	+5	+4	+3	+3	+3	+3	+3	N/A	N/A	
E5-2670 v2	115W	2.5	10	+8	+7	+6	+5	+4	+4	+4	+4	+4	+4	N/A	N/A	
E5-2667 v2	130W	3.3	8	+7	+6	+5	+4	+3	+3	+3	+3	N/A	N/A	N/A	N/A	$2057
E5-2660 v2	95W	2.2	10	+8	+7	+6	+5	+4	+4	+4	+4	+4	+4	N/A	N/A	$1389
E5-2650 v2	95W	2.6	8	+8	+7	+6	+5	+4	+4	+4	+4	N/A	N/A	N/A	N/A	
E5-2650L v2	70W	1.7	10	+4	+3	+2	+2	+2	+2	+2	+2	+2	+2	N/A	N/A	
E5-2643 v2	130W	3.5	6	+3	+2	+1	+1	+1	+1	N/A	N/A	N/A	N/A	N/A	N/A	$1552
E5-2640 v2	95W	2.0	8	+5	+4	+3	+3	+3	+3	+3	+3	N/A	N/A	N/A	N/A	
E5-2637 v2	130W	3.5	4	+3	+2	+1	+1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	
E5-2630 v2	80W	2.6	6	+5	+4	+3	+3	+3	+3	N/A	N/A	N/A	N/A	N/A	N/A	
E5-2630L v2	60W	2.4	6	+4	+3	+2	+2	+2	+2	N/A	N/A	N/A	N/A	N/A	N/A	
E5-2620 v2	80W	2.1	6	+5	+4	+3	+3	+3	+3	N/A	N/A	N/A	N/A	N/A	N/A	
E5-2609 v2	80W	2.5	4	Intel® Turbo Boost Technology not supported												
E5-2603 v2	80W	1.8	4	Intel® Turbo Boost Technology not supported												

See below some information about the TurboBoost technology and effect between # of core active and frequency.

Processor SKU (TDP (W))	Base Frequency (GHz)	Cores	1 Core Active	2 Cores Active	3 Cores Active	4 Cores Active	5 Cores Active	6 Cores Active	7 Cores Active	8 Cores Active
E5-2690 (135W)	2.9	8	3.8 (+9)	3.6 (+7)	3.6 (+7)	3.4 (+5)	3.4 (+5)	3.3 (+4)	3.3 (+4)	3.3 (+4)

The E5-2660 v2 was chosen as the baseline, all other results have been scored relative to the baseline.

For example, ifthe time to completion for the E5-2660 v2 is 00:40:00 (so 40 minutes),then the following results:

E5-2695 v2 00:44:00

E5-2690 v2 00:20:00

The comparative results are given as follows:

E5-2695 v2 : +10%

E5-2690 v2 : - 50%

3.2. CPU Test Architecture

All CPUs used in testing have the same architecture; only yarn.nodemanager.resource.cpu-vcores Cloudera configuration setting was changed to match the Logical cores available in the processor installed for that specific test sequence

[1] This is pricing we found at the time of the benchmark, we can reasonably assume that pricing will go down in time or can be different per region/theatre; however, we can also assume that differences/evolutions would be proportional for each CPU/solution.

The entire platform Config was locked down, with the only changes being those being measured – tests 1-5 on cpu (with only Vcores changes), test 6+ platform attributes (BW, # of disks, HyperThreading, #of NameNode)

Testname	CPU	DataNode/ node manager	1Gb /10Gb	HT ON / OFF	Cluster Tuning Vcores (default group)	Nbr. of HDD	Line rate	DCM power logging
Test 1	E5-2695v2	6/5	10Gb	ON	38	24	4 links - 5Gb/s	GroupBD
test2	E5-2660v2	6/5	10Gb	ON	32	24	4 links - 5Gb/s	GroupBD
test3	E5-2643v2	6/5	10Gb	ON	19	24	4 links - 5Gb/s	GroupBD
test4	E5-2667v2	6/5	10Gb	ON	26	24	4 links - 5Gb/s	GroupBD
test5	E5-2690v2	6/5	10Gb	ON	32	24	4 links - 5Gb/s	GroupBD

3.3. CPU Benchmarks Caveats

The open-public benchmark was selected so that results could be reproduced by anyone. In addition, one could tune or modify them to better reflect their specific needs and be able to make the best choice of Hadoop cluster architecture.

The following well-known and popular benchmark processes/protocols were used: TeraGen, TeraSort, TeraValidate[2], and Pig Wordcount[3].

Note: The Transaction Processing Performance Council recently published TPCx-HS for standard Hadoop benchmarking. TPC Express Benchmark(tm)HS (TPCx-HS) was developed to provide an objective measure of the hardware, operating system and commercial Apache Hadoop File System API-compatible software distributions, and to provide the industry with verifiable performance, price-performance and availability metrics. The benchmark models continuous system availability 24 hours a day, 7 days a week. More information on TPCx-HS can be found here: http://www.tpc.org/tpcx-hs/default.asp

There are two required Jar files[4] to execute for this benchmark

- Jar 1 - hadoop-mapreduce-client-jobclient.jar
- Jar 2 - hadoop-mapreduce-examples.jar

Copy these files from /opt/cloudera/parcels/<distribution>/lib/hadoop-mapreduce to the working directory where the tests will be run (where benchmark-test.sh is located)

Unpack and copy the inputtext.zip file to hdfs as follows:

- Unzip inputtext.xzip

Note the location where the files are extracted.

Edit the benchmark-v2.sh script line as follows:

- For hadoop fs put <pattern> /user/root/wordcount/in/
- Once script has run once, put a # in front of the command to avoid the copy action happening for every script run

[2] https://hadoop.apache.org/docs/current/api/index.html?org/apache/hadoop/examples/terasort/package-summary.html

[3] https://pig.apache.org/docs/r0.7.0/tutorial.html

[4] All Jar scripts can be found in the Appendix section.

The benchmark script started 3 Hadoop reference benchmarks:

- TeraGen, TeraSort, TeraValidate (10 billion rows, ~1TB)
- TeraGen, TeraSort, TeraValidate (20 billion rows, ~2TB)
- Pig Wordcount (~12GB).

The above sequence was repeated three times, the result of each was kept, then a consolidated, specific test result was determined as the mean value of the 3 runs (RUN1, RUN2, RUN3).We also secured the validity of our results measuring the Standard deviation (StdDev) between the 3 runs, while a StdDev higher than 20% between test runs would trigger a full 3 run retest. This approach helped with identifying potential anomalies from misconfigurations or hardware issues with the testbed. (Especially when one keeps manipulating the HW, swapping CPU multiple times, best practise here is to use the special Intel insertion tool:  , to prevent pin deformation on insert).

TeraGen benchmark (1 ..2)

Parameters:

- hadoop jar hadoop-mapreduce-examples.jar teragen -Dmapred.map.tasks=119[5] <teragensize> /user/root/terasort<1..2>-out

TeraSort benchmark

Parameters:

- hadoop jar hadoop-mapreduce-examples.jar terasort /user/root/terasort<1..2>-out/ /user/root/terasort<1..2>-sorted/

TeraValidate benchmark

Parameters:

- hadoop jar hadoop-mapreduce-examples.jar teravalidate /user/root/terasort<1..2>-sorted/ /user/root/terasort<1..2>-validated/

Pig Wordcount

Parameters:

- Datasets of 11.5GB of text files
  - hadoop fs -rm -r -skipTrash wordcount/out-ba
  - hadoop fs -expunge
  - pig wordcount.pig

The following metrics were measured :

Time to Complete: provided by Cloudera Manager (see 4.3.1), time for the job to be completed as well as Ganglia

CPU Time: provided by Cloudera Manager,

Power consumption (see 4.3.2): Provided by DCM-ED (Intel Datacenter Manager : Energy Director[6])

[5] Based on #core/mem and # of HDD, this changes for 12 HDD test12.

[6] http://www.intel.com/content/www/us/en/software/intel-energy-director-product-detail.html/DEVICE1/GB

Each Test has its full details: example of Test 1

Description: Testing of Ivy Bridge CPU E5-2695 v2 with Cloudera optimization to12C

Settings	
CPU model	E5-2695 (Ivy Bridge)
Cluster tuning	yarn.nodemanager.resource.cpu-vcores = 38
Hyper-threading On / OFF	Set to ON (default setting)
Network (1Gbps or 10Gbps)	10Gbps(default setting)
Number of DataNodes / node managers	6 DataNodes - 5 node managers
Number of NameNodes switched ON	2 (default setting)
Number of HDDs used by a DataNode	24 (default setting)
Line rate	4 cables to FEX - 5Gb/S

Test1: Start script benchmark-v2.sh		
	Surname	<>.
	Testname	Test1

Collect Ganglia information	Link: Collect_Ganglia_information
Collect Cloudera report data	Link: Collect_Cloudera_information
Collect Power measurement	Link: Collect_Power_information (GroupBD)
Collect Network measurement	Link: Collect_Network_information

Copy the three directories <surname>-test1-1, <surname>-test1-2, <surname>-test1-3 to the SAN.

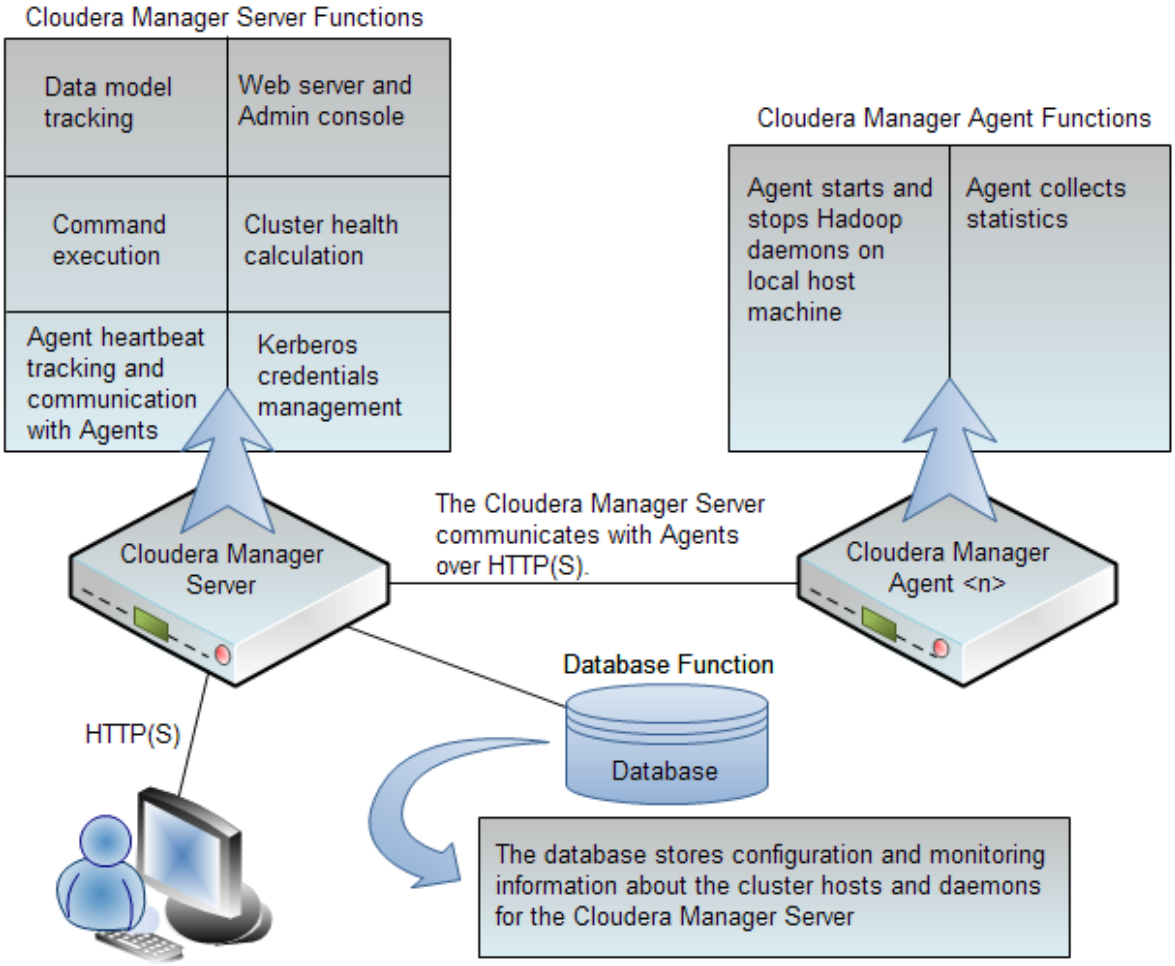
3.3.1. Cloudera Manager Architecture

The Cloudera Manager Server performs the following functions:

- Tracks the Cloudera Manager data model, which is stored in the Cloudera Manager Server database. The data model is a catalog of the available host machines in the cluster, and the services, roles, and configurations assigned to each host.
- Communicates with Agents to send configuration instructions and track Agents' heartbeats
- Performs command execution to do tasks
- Provides an Admin console for the operator to perform management and configuration tasks
- Creates, reads, validates, updates, and deletes configuration settings
- Calculates and displays the health of the cluster and its components
- Tracks host metrics such as disk usage, CPU, and RAM
- Provides a comprehensive set of APIs for the various features supported in Cloudera Manager
- Manages Kerberos credentials
- Monitors the health of Hadoop daemons, and dozens of service performance metrics, and alerts you when you approach critical thresholds.

- Keeps a history of activity, monitoring data and configuration changes
- Keeps a history all jobs executed over time + all host monitoring attributes

Each Agent starts and stops Hadoop daemons on the local host machine and collects statistics (overall and per-process memory usage and CPU usage, log tailing) for health calculations and status in the Admin console.



3.3.2. Power measurements

The power consumption has been tracked and recorded using intel DCM, providing extended reporting of real-time power consumption per Core.

Intel® Datacenter Manager: Energy Director (Intel® DCM: Energy Director) provides high value power management features that address power and thermal issues challenging IT organizations.

Intel Datacenter Manager: Energy Director features

Monitoring

- Real-time monitoring of actual power and inlet temp data aggregated to rack, row, and room
- User-defined physical or logical groups
- Receives alerts based on custom power and thermal events
- Power estimation engine for legacy servers lacking power monitoring
- Displays server asset tag and serial number for HP, IBM, and Dell
- Cisco rack and UCS support

Trending

- Logs power and thermal data, query trend data using filters
- Saves one year of history data for capacity planning

Control

- Intelligent and patented group policy engine
- Supports multiple concurrent active power policy types at multiple hierarchy levels
- Accepts workload priority as policy directive
- Allows scheduling of policies including power capping, by time of day and/or day of week
- Maintains group power capping while dynamically adapting to changing server loads
- Intel® Node Manager 2.0 support for memory power limiting and dynamic core allocation

Agent-less

- Does not require installation of any software agents on managed nodes

Easy integration and coexistence

- Device inventory prescan using IP ranges
- Exposes high-level Web Services Description Language (WSDL) APIs
- Can reside on an independent management server or coexist with ISV product on same server
- Power thermal aware scheduling: airflow and outlet temperature modeling (OEM-dependent)
- Outlet temperature sensor (OEM-dependent)

Scalability

- Manages tens of thousands of servers

Security

- Secured APIs
- Secured communication with managed nodes
- Encryption of all sensitive data

Parameter	E5-2695v2	E5-2660v2	E5-2643v2	E5-2667v2	E5-2690v2
Average (W)	3212	2938	3066	3318	3121
Duration test (h)	5.70	6.35	8.70	5.75	5.55
Energy (kWh)	18.3	18.7	26.7	19.1	17.3
CPU TDP (W)	115	95	130	130	130
Delta Energy (%)	-1.87%	0.00%	42.99%	2.26%	-7.15%

Average Power of (1)

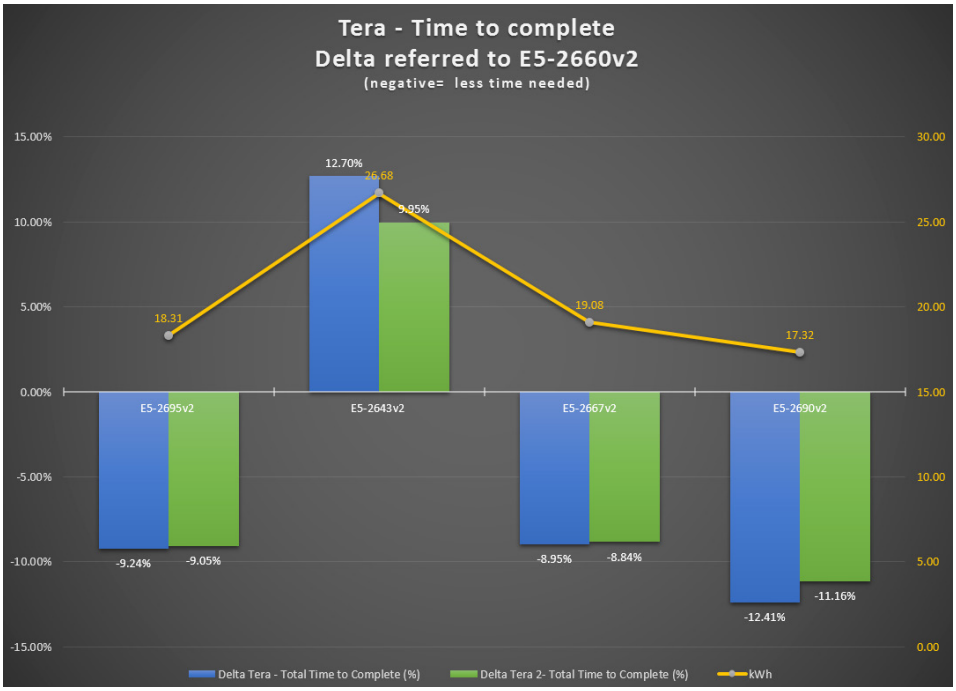
Duration of test

Energy = Average Power x Duration of test / 1000

3.4. Results

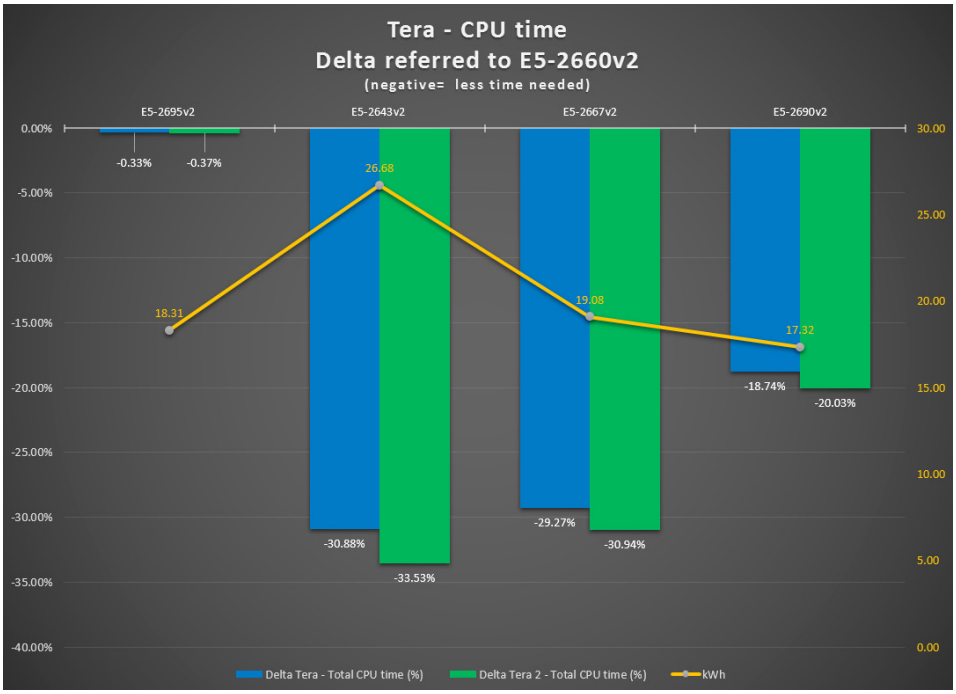
Tera results are the consolidation of Tera-Gen, Tera-Sort and Tera-Validate as the details level did not provide any additional/ relevant information for CPU best choice. The second section of the benchmark addresses the Networks, HDD and other aspects of the solution.

3.4.1. Tera Results for CPU



The baseline is E5-2660 v2, the E5-2690 v2 is clearly the best choice from a time to complete perspective.

Completing jobs in the shortest time returns the system to idle, reducing the runtime power using C states or run one or more additional tasks.

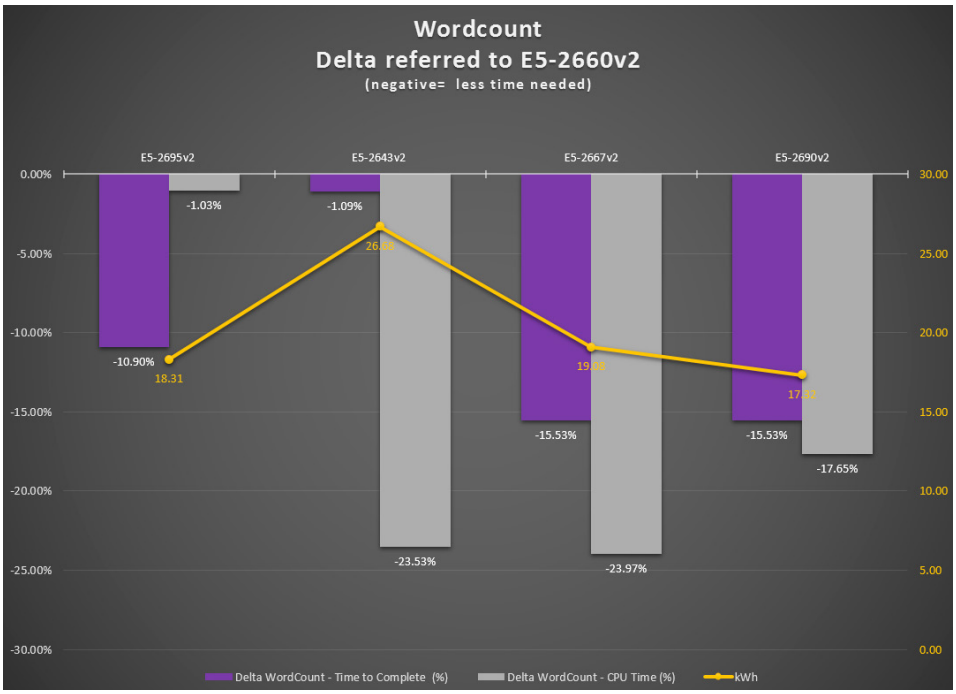


This CPU time perspective is interesting because it allows us to see the effect of high-frequency CPU and power consumption.

From an energy/CPU time ratio perspective, E5-2667 v2 is the best, showing about the same power consumption for up to 33% less CPU time than the E5-2660 v2, while the E5-2690 v2 shows minimum power consumption but with up to 20% less CPU time than reference E5-2660v2.

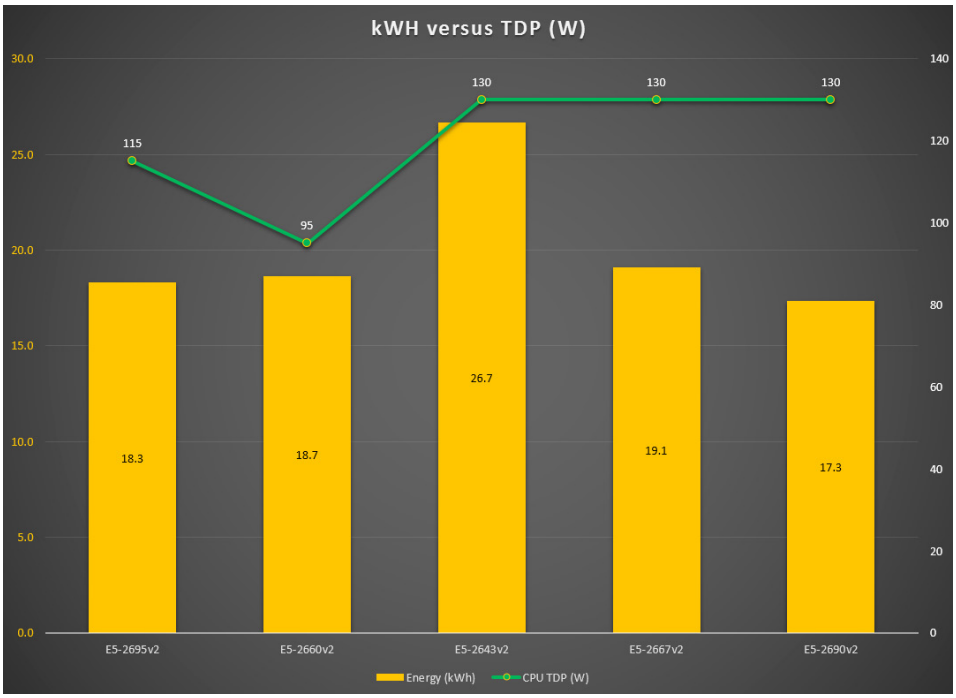
FYI, E5-2660v2 power consumption is: 18.66Kwh

3.4.2. Word Count for CPU



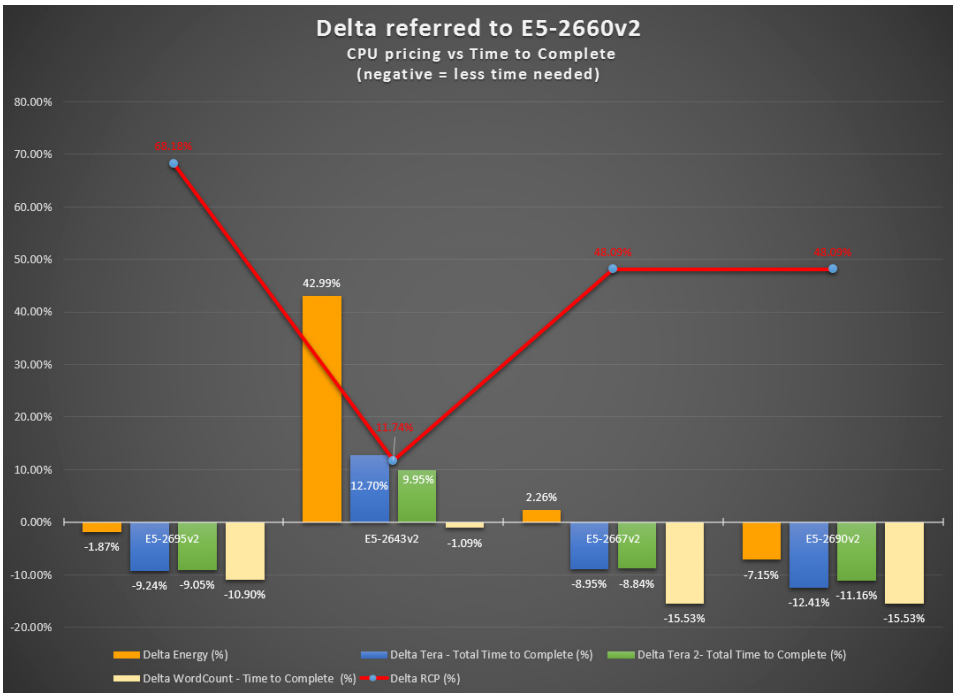
For best time to complete Word-count, E5-2667 v2 / E5-2690 v2 are the best 2 SKUs taking also into account the Energy factor (yellow line, showing 8-9 Kwh less than the E5-2643 v2 by example).

3.4.3. Power Results for CPU

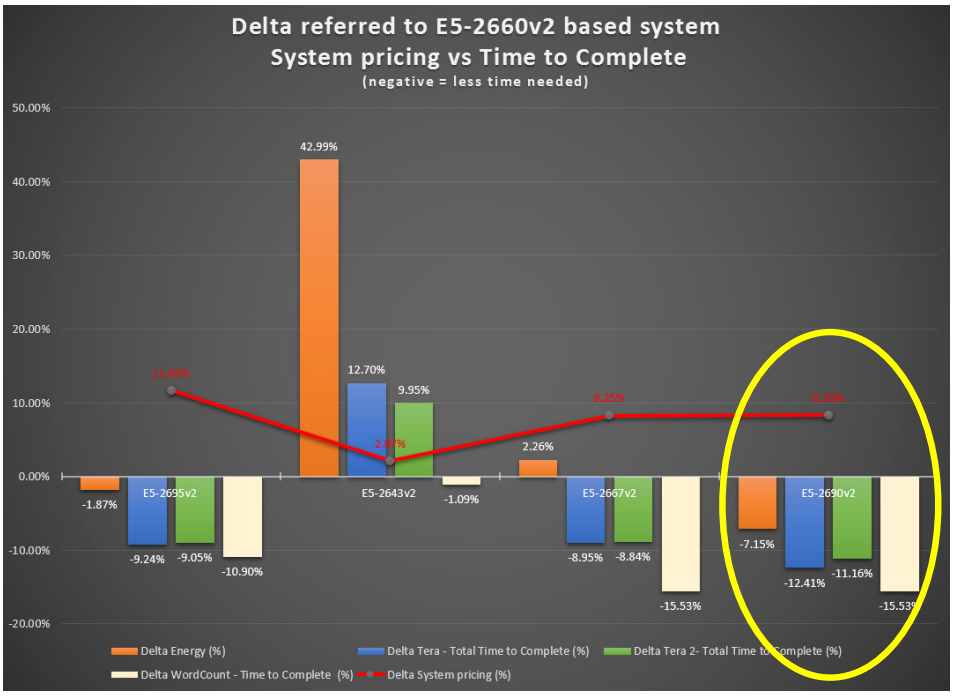


The E5-2690 v2 is the best choice for power consumption, as CPU can be potentially more Watts hungry (130W vs 115W) but consuming for much less time…

3.4.4. Consolidated Results with Pricing



E5-2690 v2 provide good scoring but has its price impact, see in 4.2 in perspective of the overall system pricing/configuration.



The E5-2690v2 is clearly the best CPU selection overall.

Extra system investment give double-digit performance increase and power saving. One of the main arguments was to discuss the number of cores versus CPUfrequency, it appears that the E52690v2 gives the right balance between cores and frequency – cores for capacity, frequency for execution speed, resulting in having a job done faster with best power performance (see lowest W).

The key metric here is wall clock time,time to complete, which E5-2690v2 is the best choice. In other words, it completes the workloads the fastest.

When looking at the sub-level of the various Tera-jobs, the processing benefit is even more apparent:

Max difference of time (for 1Tb) is 8.47% in TeraGen, 28.87% in Terasort,

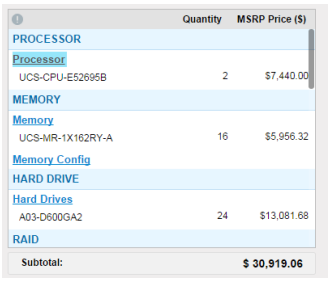
Max difference of time (for 2Tb) is 4.7% in TeraGen, 27.41% in Terasort.

3.5. CPU Benchmark Results Conclusion

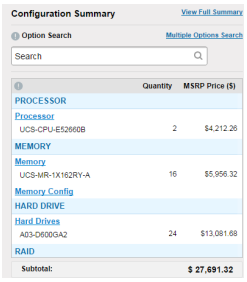
Looking at the MSRP[7] of system pricing, to give some perspective of CPU cost alone, system prices range from $30,919 to $27,691.

Major disclaimer here, this is pricing we found at the time of the benchmark, we can reasonably assume that pricing will go down in time or can be different per region/theatre; however, we can also assume that differences/evolutions would be proportional for each CPU/solution.

[7] MSRP value must be considered for the sake of this exercise as a factor to be considered during evaluation of the best architecture. These values would obviously be obsolete by the time of publication, but it is still a valid principle.



	Quantity	MSRP Price ($)
PROCESSOR		
Processor		
UCS-CPU-E52695B	2	$7,440.00
MEMORY		
Memory		
UCS-MR-1X162RY-A	16	$5,956.32
Memory Config		
HARD DRIVE		
Hard Drives		
A03-D600GA2	24	$13,081.68
RAID		
Subtotal:		$ 30,919.06

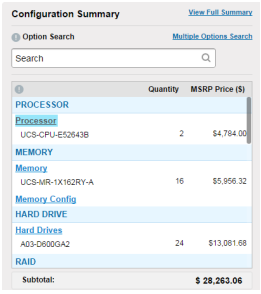


Configuration Summary — View Full Summary

Option Search — Multiple Options Search

Search

	Quantity	MSRP Price ($)
PROCESSOR		
Processor		
UCS-CPU-E52660B	2	$4,212.26
MEMORY		
Memory		
UCS-MR-1X162RY-A	16	$5,956.32
Memory Config		
HARD DRIVE		
Hard Drives		
A03-D600GA2	24	$13,081.68
RAID		
Subtotal:		$ 27,691.32



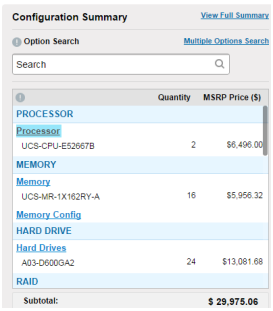
Configuration Summary — View Full Summary

Option Search — Multiple Options Search

Search

	Quantity	MSRP Price ($)
PROCESSOR		
Processor		
UCS-CPU-E52643B	2	$4,784.00
MEMORY		
Memory		
UCS-MR-1X162RY-A	16	$5,956.32
Memory Config		
HARD DRIVE		
Hard Drives		
A03-D600GA2	24	$13,081.68
RAID		
Subtotal:		$ 28,263.06

UCSC-C240-M3S

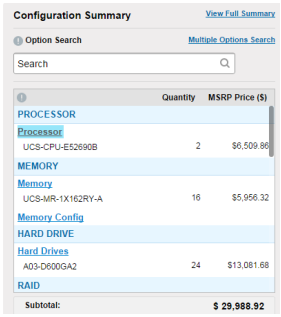


Configuration Summary — View Full Summary

Option Search — Multiple Options Search

Search

	Quantity	MSRP Price ($)
PROCESSOR		
Processor		
UCS-CPU-E52667B	2	$6,496.00
MEMORY		
Memory		
UCS-MR-1X162RY-A	16	$5,956.32
Memory Config		
HARD DRIVE		
Hard Drives		
A03-D600GA2	24	$13,081.68
RAID		
Subtotal:		$ 29,975.06



Configuration Summary — View Full Summary

Option Search — Multiple Options Search

Search

	Quantity	MSRP Price ($)
PROCESSOR		
Processor		
UCS-CPU-E52690B	2	$6,509.86
MEMORY		
Memory		
UCS-MR-1X162RY-A	16	$5,956.32
Memory Config		
HARD DRIVE		
Hard Drives		
A03-D600GA2	24	$13,081.68
RAID		
Subtotal:		$ 29,988.92

Same system as in the cluster – with different CPUs

In conclusion, between all five systems, the prices have a maximum difference of ~10.4% while the performance benefit, i.e., compared with Terasort1&2 total time to compute, can be as much as 30% processing time benefit. The best ratio of price/performance/power/time is definitively the E5-2690 v2 despite its high-price, though different workload requirements may influence the ultimate CPU selection for the cluster.

4.Cluster Benchmark

4.1. Overview

In the second block of the test, the CPU was fixed from the E5-2690 v2 – the CPU with the best price/performance ratio. The other aspects of the architecture were tested: Networking design/bandwidth, HDD, Hyper-Threading effect (test 5-8) as well as the scale-out behaviour of the solution to complete the overall test (test 9-12).

Testname	CPU	DataNode/ node manager	1Gb /10Gb	HT ON / OFF	Cluster Tuning Vcores (default group)	Nbr. of HDD	Line rate	DCM power logging
test5 E5-2690V2 5Gb/s	E5-2690v2	6/5	10Gb	ON	32	24	4 links - 5Gb/s	GroupBD

Testname	CPU	DataNode/ node manager	1Gb / 10Gb	HT ON / OFF	Cluster Tuning Vcores (default group)	Nbr. of HDD	Line rate	DCM power logging
test6 E5-2690V2 FDR	E5-2690v2	6/5	10Gb	ON	32	24	FE removed	GroupBD
test7 E5-2690V2 HT-off	E5-2690v2	6/5	10Gb	OFF	16	24	FE removed	GroupBD
test8 E5-2690V2 1Gb	E5-2690v2	6/5	1Gb	ON	32	24	FE removed	GroupBD
test9 E5-2690V2 5ND	E5-2690v2	5/5	10Gb	ON	32	24	FE removed	GroupBD-8
test10 E5-2690V2 4ND	E5-2690v2	4/4	10Gb	ON	32	24	FE removed	GroupBD-8&7
test11 E5-2690V2 6NM	E5-2690v2	6/6	10Gb	ON	32	24	FE removed	GroupBD
test12 E5-2690V2 12HDD	E5-2690v2	6/6	10Gb	ON	32	12	FE removed	GroupBD

10Gb/s

Ref1. for test 5..8

Ref2. for test 9..12

4.2. Benchmark Caveat

Note (or not to do…): In the raw data node/name node columns, a 5/6, then 6/6, it indicated one data node did not start. This has no impact on the first set of benchmark as it was CPU focused. However, the full six nodes were needed for networking and scale-out tests. Consequently, a new baseline had to be re-run for tests 9 to 12, where the number of data nodes is key to the results. Tests 6 to 8 had the same setup, even though the cluster was rebuilt due to a misconfiguration in the RAID setup. (see5.2.1)

4.2.1. Benchmark Caveat : Raid Configuration

The UCS servers are configured with an LSI 9266 controller.

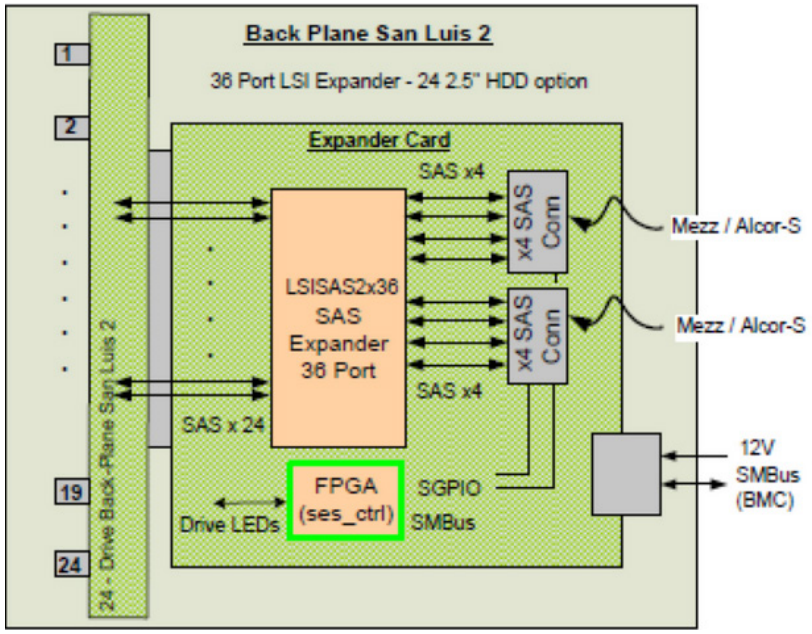
LSI 9266 controller specs can be found here :
http://www.lsi.com/downloads/Public/MegaRAID%20SAS/MegaRAID%20SAS%209266-8i/LSI_MR-SAS9266-8i_PB.pdf

Basically :

- PCIe 2.0 x8 RAID ctrl
- 8 SAS ports 6Gb/s
- 1 GB write cache (NAND)
- up to 200,000 iops

In our configuration :

- C240 M3
- 1 LSI RAID controller in x8 PCIe slot
- 24 drives configured as 24 individual RAID0
- 1 GB write cache
- used SAS expander from the 2 LSI SAS connectors to control the 24 drives



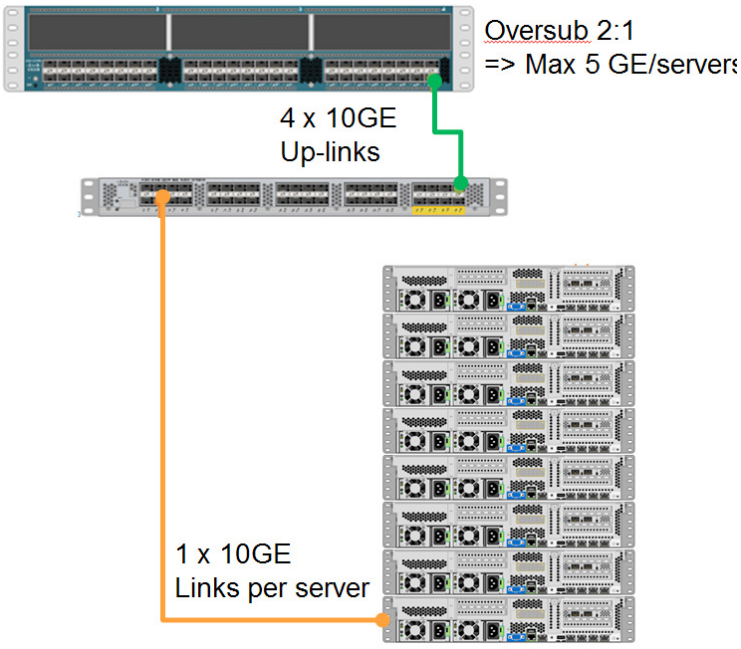
Here is our implementedsolution:

- individual RAID 0
- 1MB stripe size
- Read-ahead cache enabled
- writecache enabled while battery is present

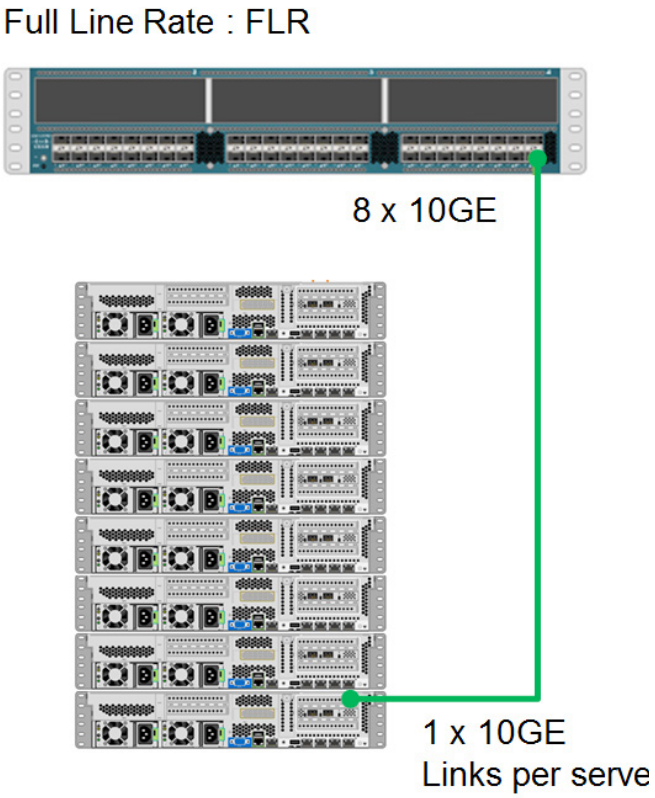
4.2.2. Benchmark Caveat : Network Bandwidth

Between the columns, stating 1/10 Gbpsand the line rate stating 5 Gb, let's review the network design/setup that was used to check impact of the network on the Hadoop cluster capacity.

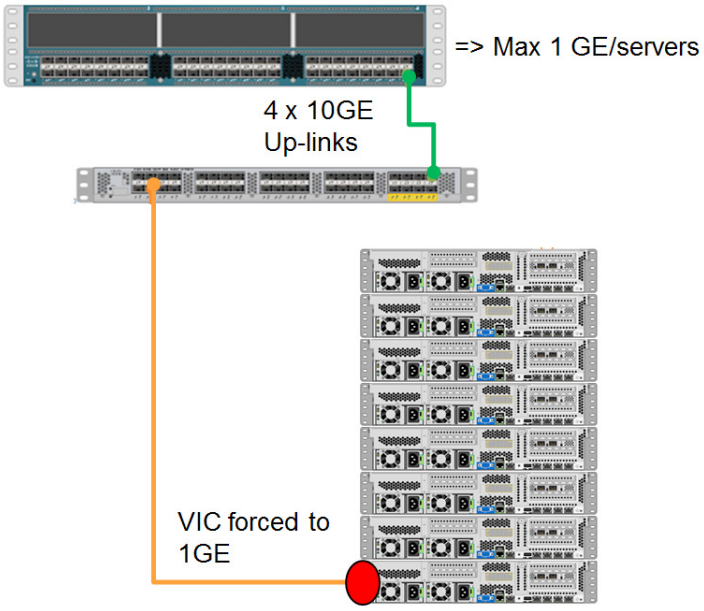
Tests 1-5 were using a typical CPAv2 architecture. It was designed to scale up to 160 servers per domain and, consequently, using a FEX betweenserver and FI switch.



Removing the FEX and cabling the server directly to the FI (an option now in CPA v3, highperformance) provide full-line rate at 10 Gbps for the whole server.

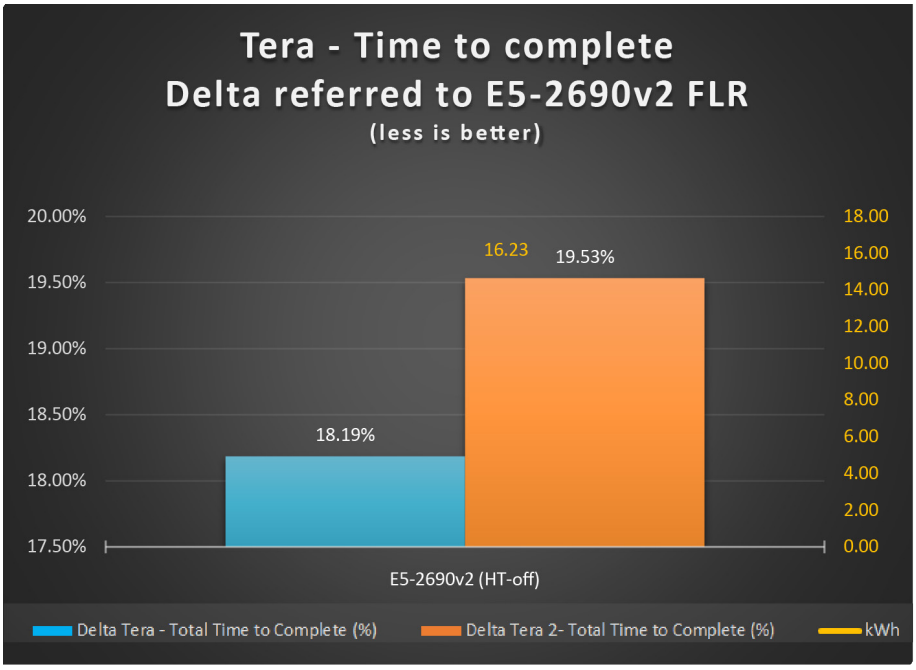


1 Gbps test has been managed by forcing the NIC Ethernet port to work at 1Gbps instead of 10Gbps.



These configurations made testing of the 1Gbps, 5Gbpsand 10Gbpsnetworks possible.

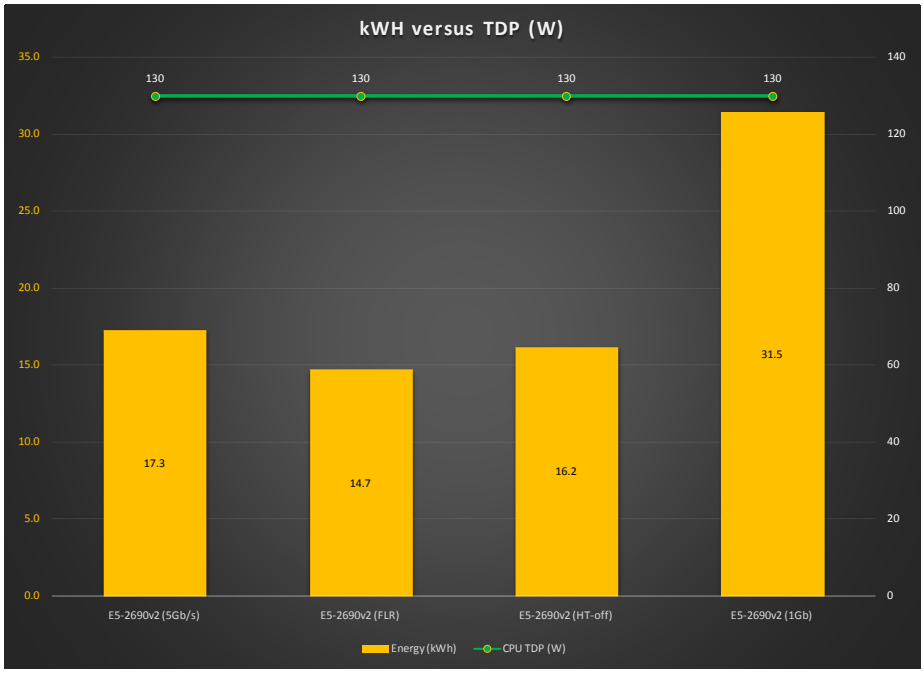
4.3. Benchmark Hyper-Threading



We can see a clear impact of Hyper-Threading OFF, about 20% more time to complete.

Baseline result used: E5-2690v2, Hyper-threading ON, FLR (full-line rate 10Gb) with 24 HDD.

Enabling Hyper-Threading allows for two execution threads per core. This enables additional processing resources and makes them available for use. The result being the job executed in a shorter time, fewer execution cores to process, so fewer containers active to do the data processing.

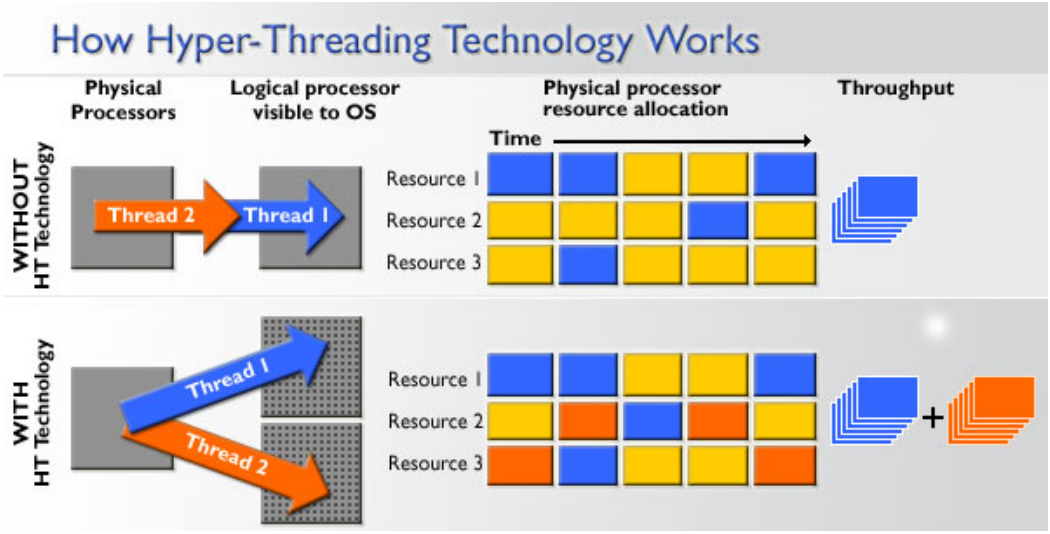


From a power consumption perspective, HT makes no differences, A processor with a good core/frequency balance completes tasks quicker overall, therefore reducing overall platform power consumption

Eliminating IO and network bottlenecks allows the processor to execute optimally, resulting in an overall reduction in job completion time and therefore overall platform power reduction.As a s result, we recommend enabling HT on any system that supports it to maximise the compute resources available.

4.3.1. Hyper-Threading details

For each processor core that is physically present, the operating system addresses two virtual or logical cores, and shares the workload between them when possible. The main function of hyper-threading is to increase the number of independent instructions in the pipeline; it takes advantage of superscalar architecture, in which multiple instructions operate on separate data in parallel. With HTT, one physical core appears as two processors to the operating system, which can use each core to schedule two processes at once. In addition two or more processes can use the same resources. If resources for one process are not available, then another process can continue if its resources are available.

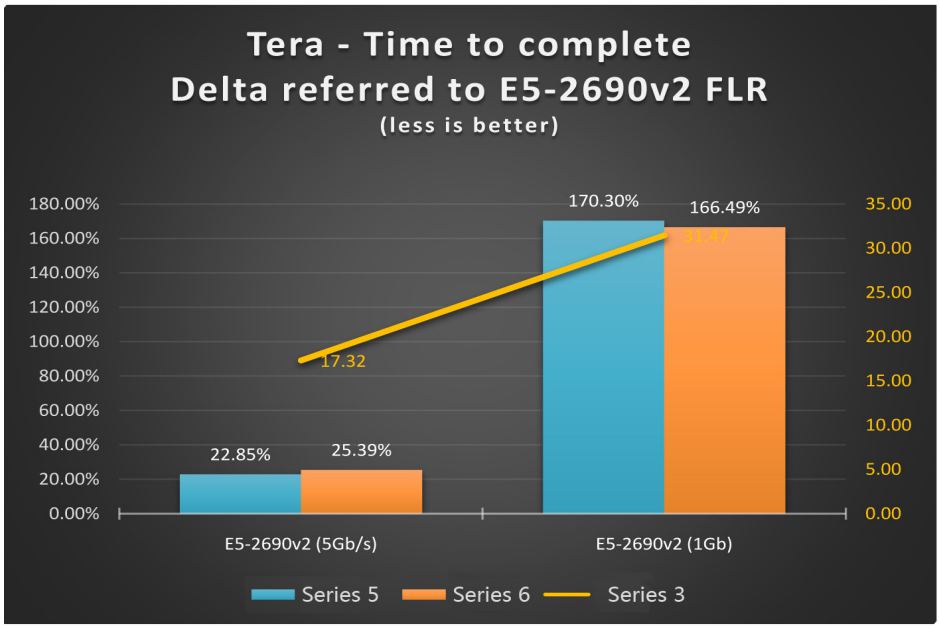


4.4. Benchmark Network Bandwidth

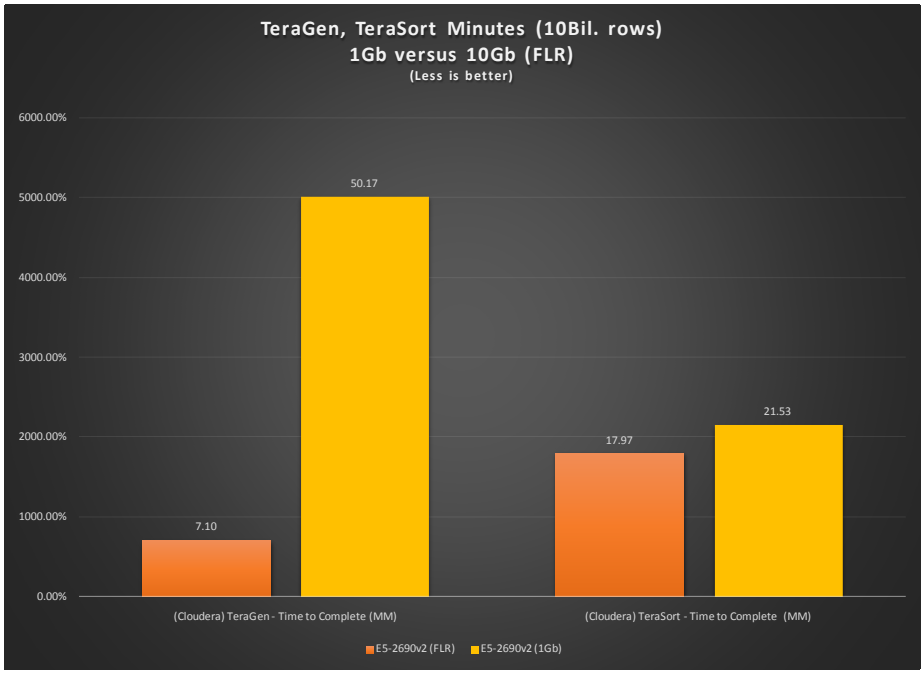
The network has a direct impact on the cluster's performance. 1Gbpscreates limitation in transferring the data between the data nodes during the Teragen generation phase. Time to complete this take is higher due to 1Gbpsbut CPU utilization is lower.

Hadoop is especially sensitive to network performance, especially with larger datasets.

We have a clear demonstration of "Networking bound" architecture for the 5Gbpsand the 1Gbpstime did not permit, but we suspect that we could get some slight improvements of results with 2 x 10 Gbpsper servers FLR, based on disk IO capacity with 24 HDD (to be completed by iops calculation with 24 HDD raid, i.e., 200 iops x 24).



4.4.1. TeraGen and TeraSort details



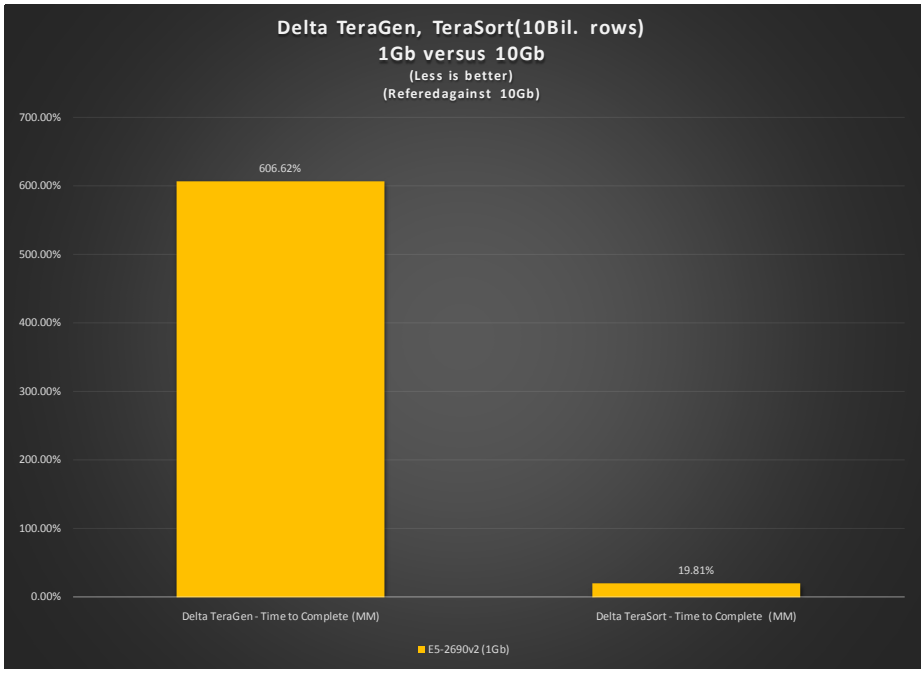
Please note, full numbers in this chart, usual percentage difference on the following figure.

Teragen 1GB impact is huge due the nature of the workload generating 1TB of data and distributing across all datanodes. Network performance is critical to this action

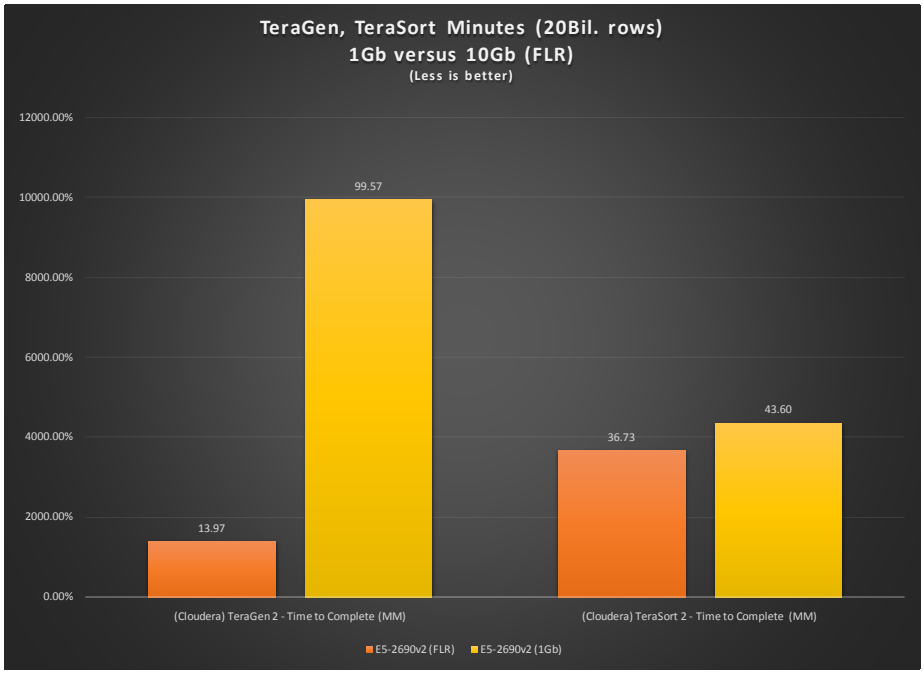
Teragen – The write pipeline with default replication of 3 means that 1 TB of data would be 3 TB of data on disk and 2 TB of network traffic(1 replica written local and 2 replicas remote to the first node which creates 2 TB of network traffic).

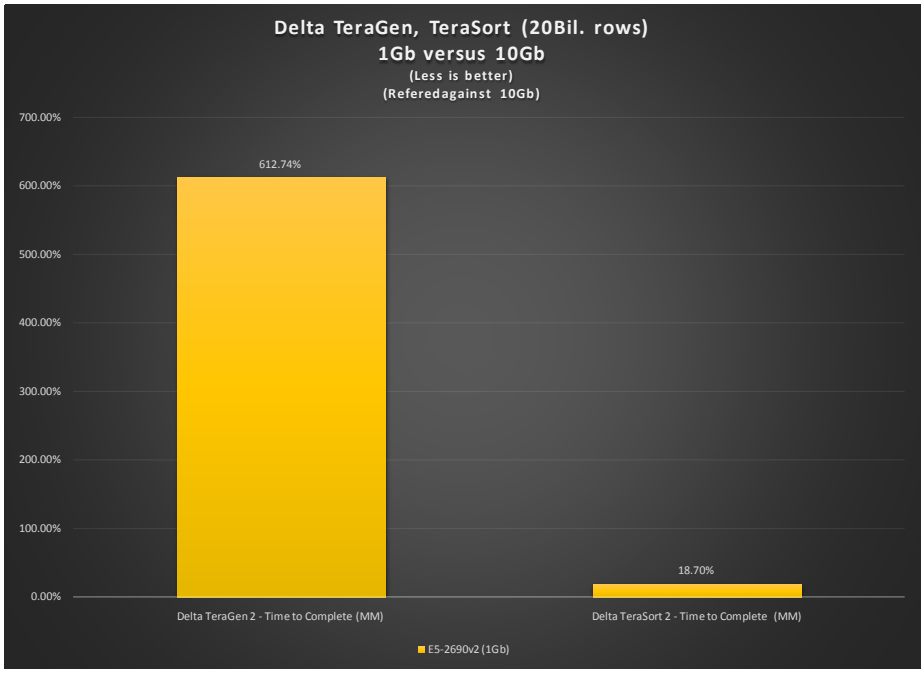
Terasort is interesting in that it only has replication factor of 1 so there is no network traffic in the HDFS write pipeline as the first replica in the pipeline is always written locally.

1. Map phases write to local disk for shuffle phase –1 TB read from disk -> 1 TB written to disk
2. Reduce phase copies data from output of map phase to local disk of the reducer for input to that step – 1 TB read from Disk Map Phase read ->1 TB written to disk Reduce Phase intermediate results -> 1 TB read reduce phase input, 1 TB of network traffic Map phase copy to Reduce Phase
3. The final output of the reduce phase is written to HDFS with replication factor of 1 (no network traffic as all data is written locally) – 1 TB of written to HDFS

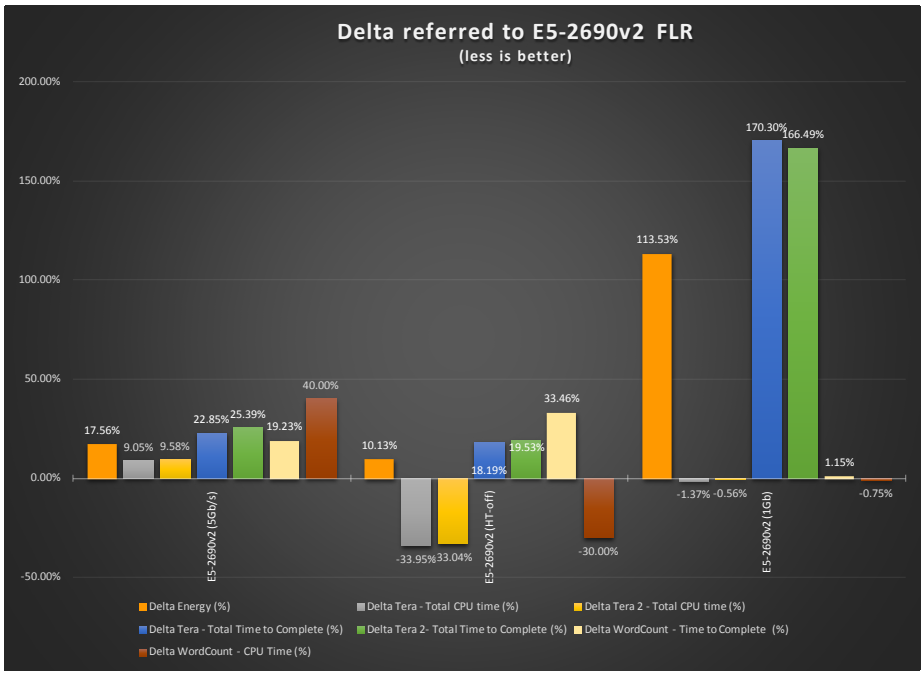


Similar behaviour for the 20 Billion rows (2Tb)





4.5. Benchmark Hyper-Threading/Networking results conclusion



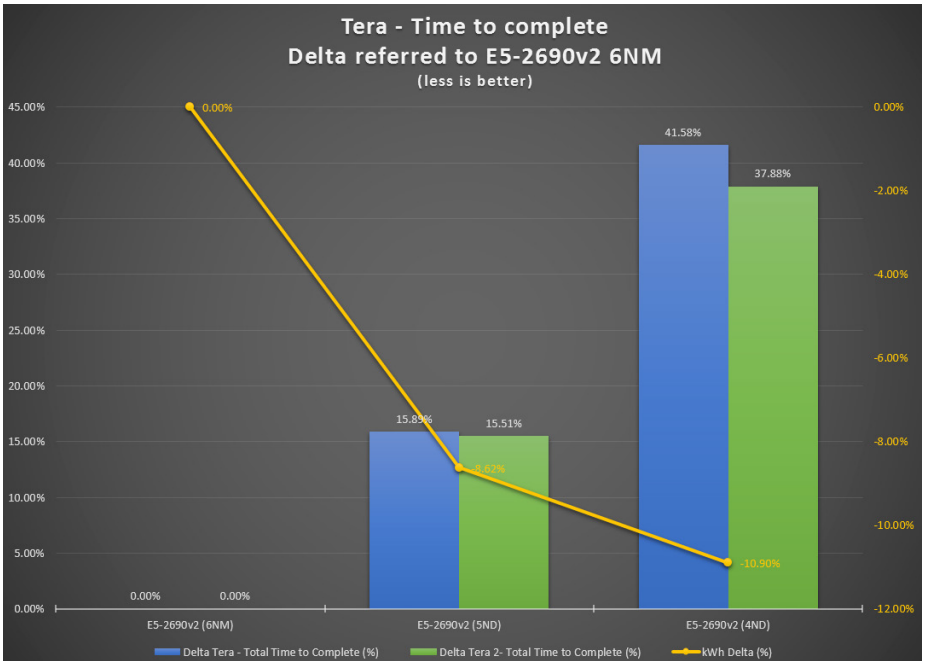
A cluster at FLR with HyperThreading-ON is the recommended configuration

4.6. Benchmark Data Nodes Scale-out

This little series of tests validate that the Hadoop task can be processed in a time inversely proportional to the number of DataNodes. So if it takes 4 hours with 10 Data Nodes, it should take 2 hours with 20 Data Nodes.

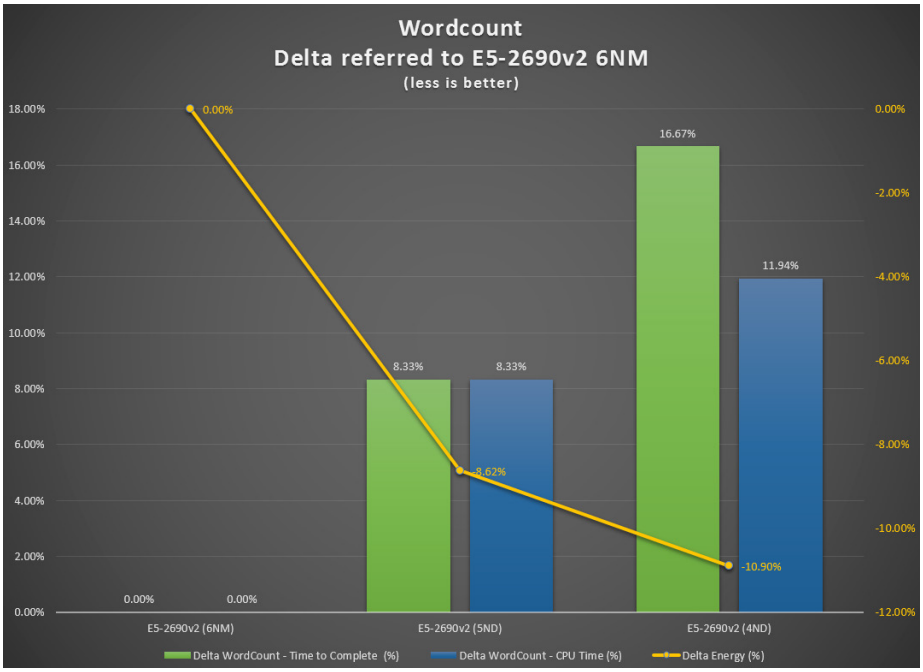
Which is validated by our results, from 6 to 5 nodes, this is 1/16 less ⇔ 16%, from 5 to 4 nodes, this is 1/5 less ⇔ 20%.

Power numbers are slightly different as they are impacted by a fixed powercost for the redundant Name-Node. The power numbers are to be moderated, as with a small number of data node versus name node, the named node power consumption sharing if pretty high between 3.4 or 5 data nodes …in the operational environment, 10-100's of data node, the power consumption of the Name-Node become negligible.



Tera is a very distributed application that relies on the aggregate of compute, network and the IO in the cluster. Removing a node has immediate impact on the performance of the cluster.

Wordcount exhibits linear behavior between 6 vs 5 nodes.This is probably due to the load not high enough to stress the cluster.



The Wordcount dataset at 12GB is too small to fully test the scalability of the cluster. The overhead is mostly due to scheduling.

4.7. Benchmark HDD scaling

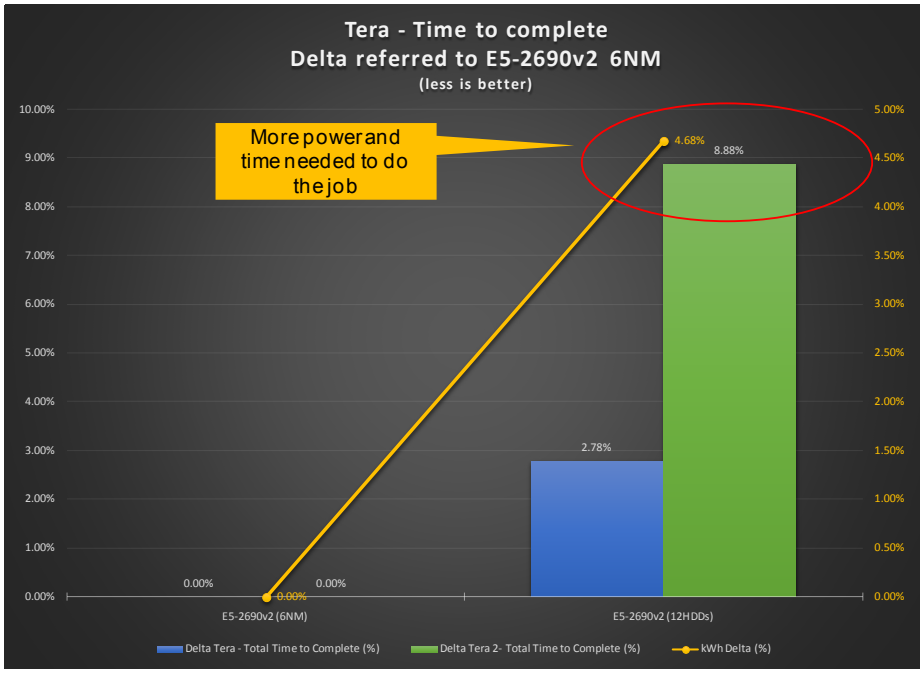
Workload with high I/O usage will benefit from the increased HDD count.

More HDD results in:

Overall better data throughput and processing

Reduction of overall system power consumption.

Having the job completed in a shorter time frame



1 TB of data in HDFS written out with 128 MB blocks = ~8000 HDFS blocks. The more individual hard drives exist to serve this data in parallel, the more throughput can be observed in the read phase. The same is true on the write phase. Less contention will be seen as the load is spread across more disks. Read and write throughput is impacted as more drives are available.

The write pipeline with default replication of 3 means that 1 TB of data would be 3 TB of data on disk and 2 TB of network traffic(1 replica written local and 2 replicas remote to the first node which creates 2 TB of network traffic).

Terasort is interesting in that it only has a replication factor of 1 so there is no network traffic in the HDFS write pipeline as the first replica in the pipeline is always written locally.

1. Map phases writes to local disk for shuffle phase –1 TB read from disk -> 1 TB written to disk
2. Reduce phase copies data from output of map phase to local disk of the reducer for input to that step – 1 TB read from Disk Map Phase read ->1 TB written to disk Reduce Phase intermediate results -> 1 TB read reduce phase input, 1 TB of network traffic Map phase copy to Reduce Phase
3. The final output of the reduce phase is written to HDFS with replication factor of 1 (no network traffic as all data is written locally) – 1 TB of written to HDFS

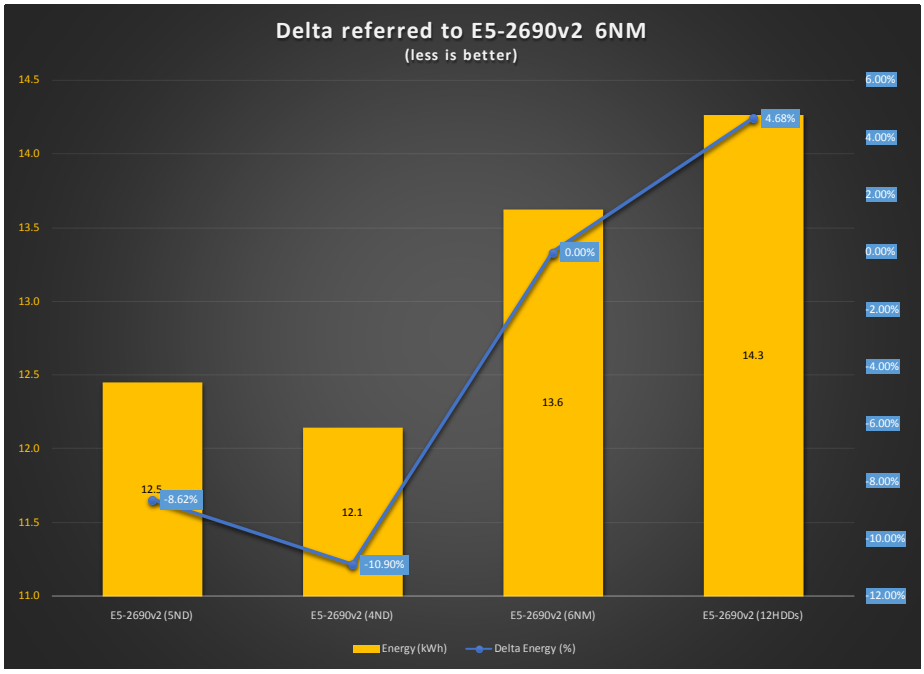
We did not include here the word count, as the size of it is not relevant to trigger 12/24 HDD demand.

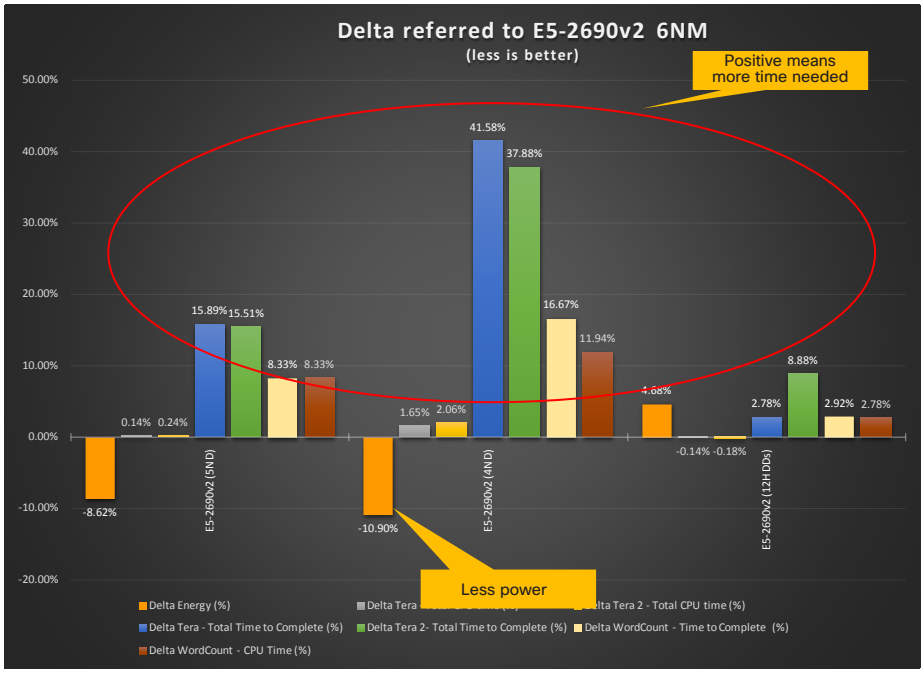
The extra hard drives will help in mixed workloads of read/write as the entire drive's bandwidth won't be just used for write as what was seen in Teragen. Having more drives as expected scales very well. However, there are upper bounds ~24 where the payoffs diminish. Driven by the form factor of two-rack unit (2RU) server can fit at most 24 drives.If you change to a larger server the costs become prohibitive.

Also when planning the node design ensure adequate bandwidth is available for installed drive controllers to serve drive data – whether they be add-in or embedded controllers, ensure there are adequate data lanes present in the bus and the controllers chosen need to be able to support concurrent random IO to all connected drives, which is the case for our UCS Cisco LSI MegaRAID SAS 9271 CV-8i

4.8. Benchmark HDD/Scaling results conclusion

For the power consumption results, the overhead of the masters is less of a percentage of the entire cluster as your cluster grows. Performance follows a scaling curve, more data nodes means better results, same for doubling the number of HDD from 12 to 24.

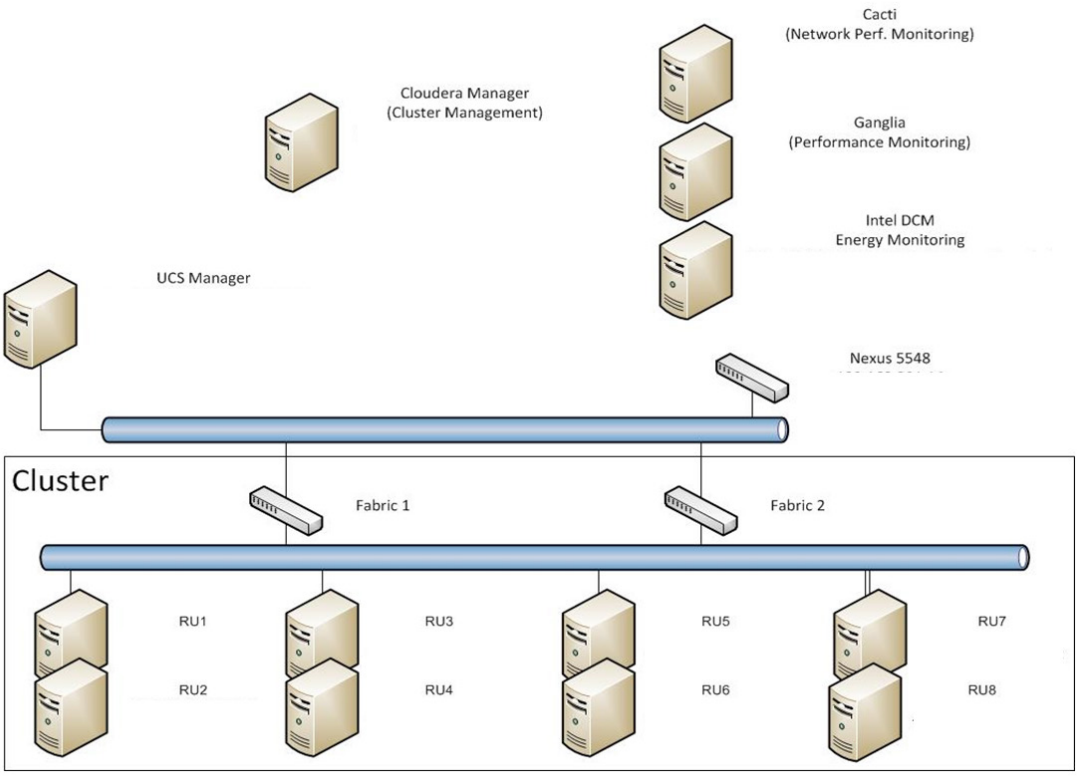




5. Security assessment

5.1. Overview

Security is mandatory for any operational environment; Intel was invited to review the benchmark environment and found many security non-complaint issues. It is recommended to follow the guidelines for network/server/OS security when building a secured Hadoop cluster.



5.2. Servers

This is a summary of the full Centos security test (full report available in Appendix section: server1.testlab.local-20141209T134758Z.zip).

Summary

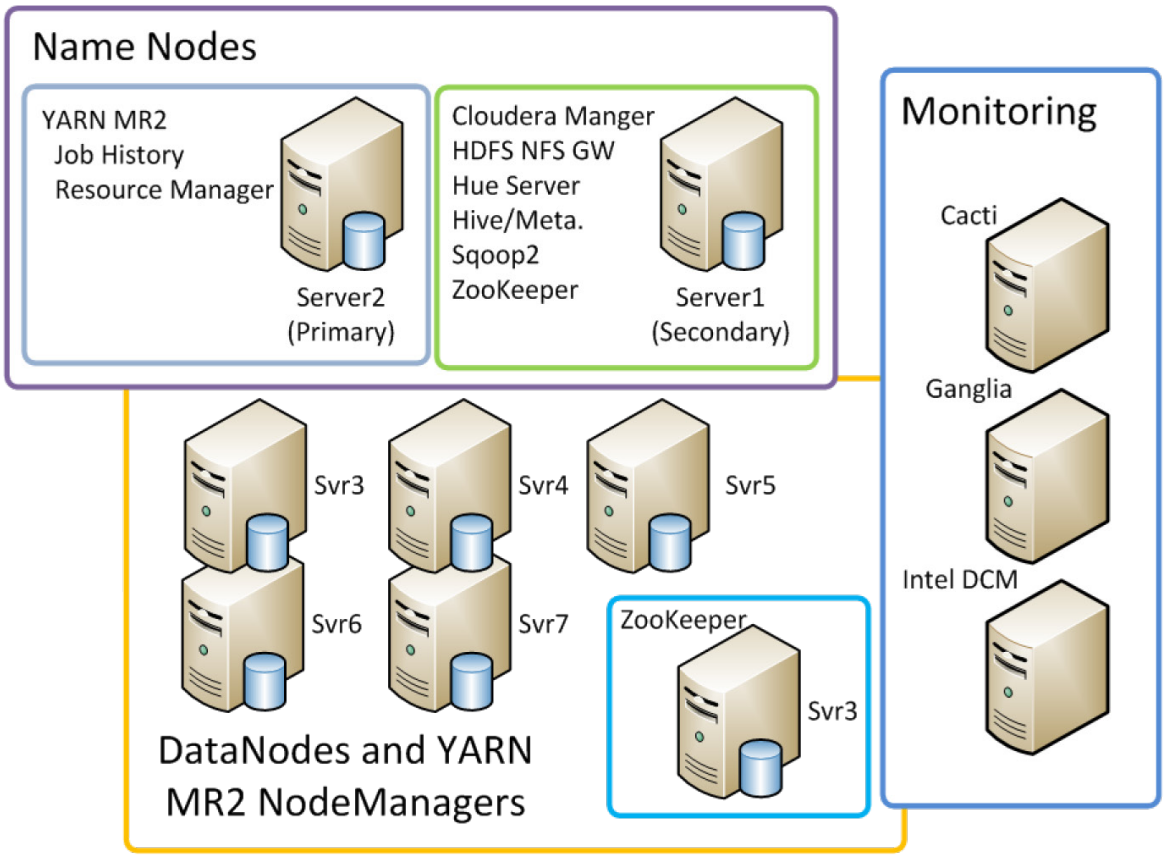
Description	Tests				Scoring		
	Pass	Fail	Error	Unkn.	Score	Max	Percent
1 Install Updates, Patches and Additional Security Software	**7**	**17**	**0**	**0**	**7.0**	**24.0**	**29%**
1.1 Filesystem Configuration	1	13	0	0	1.0	14.0	7%
1.2 Configure Software Updates	2	0	0	0	2.0	2.0	100%
1.3 Advanced Intrusion Detection Environment (AIDE)	0	0	0	0	0.0	0.0	0%
1.4 Configure SELinux	0	0	0	0	0.0	0.0	0%
1.5 Secure Boot Settings	2	3	0	0	2.0	5.0	40%
1.6 Additional Process Hardening	2	1	0	0	2.0	3.0	67%
2 OS Services	**17**	**0**	**0**	**0**	**17.0**	**17.0**	**100%**
2.1 Remove Legacy Services	17	0	0	0	17.0	17.0	100%
3 Special Purpose Services	**13**	**2**	**0**	**0**	**5.0**	**6.0**	**83%**
4 Logging and Auditing	**3**	**2**	**0**	**0**	**2.0**	**4.0**	**50%**
4.1 Configure rsyslog	2	2	0	0	2.0	4.0	50%
4.2 Configure System Accounting (auditd)	0	0	0	0	0.0	0.0	0%
4.2.1 Configure Data Retention	0	0	0	0	0.0	0.0	0%
5 Network Configuration and Firewalls	**8**	**6**	**0**	**0**	**7.0**	**11.0**	**64%**
5.1 Modify Network Parameters (Host Only)	1	1	0	0	1.0	2.0	50%
5.2 Modify Network Parameters (Host and Router)	4	2	0	0	4.0	6.0	67%
5.3 Wireless Networking	1	0	0	0	0.0	0.0	0%
5.4 Disable IPv6	0	2	0	0	0.0	0.0	0%
5.4.1 Configure IPv6	0	2	0	0	0.0	0.0	0%
5.5 Install TCP Wrappers	2	0	0	0	2.0	2.0	100%
5.6 Uncommon Network Protocols	0	0	0	0	0.0	0.0	0%
6 System Access, Authentication and Authorization	**6**	**24**	**0**	**0**	**6.0**	**29.0**	**21%**
6.1 Configure cron and anacron	3	8	0	0	3.0	11.0	27%
6.2 Configure SSH	2	12	0	0	2.0	14.0	14%
6.3 Configure PAM	1	3	0	0	1.0	3.0	33%
7 User Accounts and Environment	**2**	**5**	**0**	**0**	**2.0**	**7.0**	**29%**
7.1 Set Shadow Password Suite Parameters (/etc/login.defs)	1	2	0	0	1.0	3.0	33%
8 Warning Banners	**2**	**1**	**0**	**0**	**1.0**	**2.0**	**50%**
9 System Maintenance	**29**	**2**	**0**	**0**	**29.0**	**29.0**	**100%**
9.1 Verify System File Permissions	10	2	0	0	10.0	10.0	100%
9.2 Review User and Group Settings	19	0	0	0	19.0	19.0	100%
Total	**87**	**59**	**0**	**0**	**76.0**	**129.0**	**59%**

5.3. Hadoop

Security with the Hadoop environment is commonly controlled at the perimeter, regulating who can communicate and submit jobs to the nodes of the cluster. While this is a good first level defense, the relatively open design of the Hadoop ecosystem makes compromising the entire system somewhat trivial once this single layer is defeated. To help better guide security within Hadoop environments, typical deployments must be examined and recommendations made that will lessen the impact in the case of a perimeter breach.

5.3.1. Environment

Typical to most environments, the servers within ours serve multiple roles:



5.3.2. Attack Surface

When considering security, the most urgent step in a deployment is to limit the available ports and services exposed by a system. Each open port offers a potential avenue to system compromise for an attacker. This can be a complex task in a Hadoop environment given the variety of roles any particular node may have. To illustrate this point, the following lists what is available on the first three servers in the benchmark lab environment:

Host	Protocol	Port	Purpose
Server 1	TCP	22	SSH
	TCP	111	RPC
	TCP	2049	NFS
	TCP	2181	ZooKeeper Client Port
	TCP	4181	ZooKeeper Election Port
	TCP	4242	HDFS NFS Gateway
	TCP	5678	Cloudera Reports Manger Server Port
	TCP	7180	Cloudera Manager Admin Console
	TCP	7182	Cloudera Agent Connect Port
	TCP	7184	Cloudera Event Publish Port
	TCP	7185	Cloudera Event Query Port
	TCP	7432	Postgres
	TCP	8083	Cloudera Reports Manager Debug WebUI
	TCP	8084	Cloudera Event Server Debug WebUI
	TCP	8086	Cloudera Service Monitor Debug WebUI
	TCP	8087	Cloudera Activity Monitor Debug WebUI
	TCP	8091	Cloudera Host Monitor Debug WebUI
	TCP	8649	Ganglia
	TCP	8888	Hue Server
	TCP	9000	Cloudera Agent
	TCP	9010	ZooKeeper JMX Remote Port
	TCP	9083	Hive Metastore Server Port
	TCP	9994	Cloudera Host Monitor Nozzle Port
	TCP	9995	Cloudera Host Monitor Listen Port
	TCP	9996	Cloudera Service Monitor Nozzle Port
	TCP	9997	Cloudera Service Monitor Listen Port
	TCP	9998	Cloudera Activity Monitor Nozzle Port
	TCP	9999	Cloudera Activity Monitor Listen Port
	TCP	10000	Hive Server2 Port
	TCP	10101	Cloudera Alert Publisher Listen Port
	TCP	12000	Sqoop2 HTTP Port
	TCP	40164	ZooKeeper
	TCP	50090	Secondary NameNode Web UI Port

Server 2	TCP	22	SSH
	TCP	111	RPC
	TCP	8040	Node Manager Localizer Port
	TCP	8041	NodeManager IPC Port
	TCP	8042	NodeManager Web UI Port
	TCP	8649	Ganglia
	TCP	9000	Cloudera Agent
	TCP	13562	MR2 Shuffle Port
	TCP	50010	DataNode Transceiver Port
	TCP	50070	NameNode WebUI Port
	TCP	50075	DataNode HTTP Web UI Port
Server 3	TCP	22	SSH
	TCP	111	RPC
	TCP	8040	Node Manager Localizer Port
	TCP	8041	NodeManager IPC Port
	TCP	8042	NodeManager Web UI Port
	TCP	8649	Ganglia
	TCP	9000	Cloudera Agent
	TCP	13562	MR2 Shuffle Port
	TCP	50010	DataNode Transceiver Port
	TCP	50020	DataNode Protocol Port
	TCP	50075	DataNode HTTP Web UI Port

The list indicates a considerable amount of potential entry points for an attacker, a vulnerability in any one of these services could lead to local system compromise and ultimately cluster data.

The full report in the appendix section (Intel Hadoop UCS Lab – Security Review Whitepaper v0.1.doc), outlines the major points an attacker would target given a route to the target environment. These sections are mostly the more appealing to an attacker and shouldn't be considered an exhaustive list. The key points to secure Hadoop, following General Hadoop Security Configuration guidelines.

5.3.3. Additional Notes

This section contains various observations about the environment that could not be expanded on within the timeframe of the initial security assessment.

- **Ganglia** – Ganglia's WebUI is unauthenticated and does not require TLS. Additionally, the server queries the gmond daemon running on each node using a clear text protocol and the server itself is written in C, which makes it a candidate for fuzzing to identify memory corruption vulnerabilities
- **Hue** – Hue was left unconfigured. When connecting to the WebUI, Hue required that a username and password be set. Tester set it to root:password
- **User Accounts** – All job submissions via CLI and interaction with the nodes of the cluster was using the root user, this was not highlighted in the above sections but is an obvious security concern.

6. Appendix

All RAW data files, scripts, excel… can be found in this GitHub repository:
https://github.com/fgrandva/Cloudera-Intel-Cisco-Hadoop-Benchmark-TOI-

7. References

Cisco Design Guides :

- Cisco UCS Common Platform Architecture (CPA) for Big Data with Cloudera
- Cisco UCS Common Platform Architecture Version 2 (CPAv2) for Big Data with Cloudera (PDF - 7 MB) Updated
- Cisco UCS Integrated Infrastructure for Big Data : Cisco Solution Brief

For more information about Cisco Big Data, please visit:
http://blogs.cisco.com/datacenter/

For more information about Cisco UCS, please visit :
www.cisco.com/go/ucs

Intel references:

- More detailed information about Intel® Processors: http://ark.intel.com/
  - ** Please note the Intel® Xeon® V3 processor family has been released
- Intel® processor insertion/removal tool
  - http://www.intel.com/support/motherboards/server/sb/CS-032587.htm
- More detailed information about Intel® Datacenter Manager: Energy Director:
  https://www-ssl.intel.com/content/www/us/en/software/intel-dcm-energy-director-technical-information.html
- For more information about Intel Big Data, please visit :
  https://www-ssl.intel.com/content/www/us/en/big-data/big-data-analytics-turning-big-data-into-intelligence.html
- Blogs, communities:
  - https://www-ssl.intel.com/content/www/us/en/blogs-communities-social.html
  - https://communities.intel.com/search.jspa?q=big+data

Cloudera references:

- More detailed information about Cloudera offering: http://www.cloudera.com/content/cloudera/en/home.html
- Detailed documentation: http://www.cloudera.com/content/cloudera/en/documentation.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices**.

LE-38701-00 03/13