Lessons Learned: Performance Tuning Hadoop Systems

A study based on TPCx-HS

TPCTC 2016

Raghunath Nambiar, Cisco

Manankumar Trivedi, Cisco

42nd International Conference on
VERY LARGE DATA BASES
New Delhi, India • September 5 – 9, 2016

Very Large Data Bases INDIA 2016

Agenda

- Benchmarking Big Data Systems
- Experiments and Analysis
- Q & A

Benchmarking Big Data Systems

Big Data Benchmark ! Motivation

- Big Data (especially Hadoop) has become an integral part of enterprise IT ecosystem across major verticals
- Industry demanded standards. Top challenge for enterprise customers - What platform to choose in terms of performance, price-performance, and energy efficiency ?
- All major vendors (HW and SW) have invested in Big Data practice. Traditional standards are inadequate to benchmark Big Data Systems
- There are claims (not discrediting them) but not easily variable - Situation was not any different from the 1980’s what motivated industry experts to establish TPC and SPEC

TPC-Big Data Standard Initiatives

- Big Data was identified as one of the top areas for industry standard benchmark developments at the VLDB 2014, TPCTC 2014 , WBDB 2014 and other conferences
- Continuing TPC's commitment to developing relevant benchmark standards
- TPC-BD Working Group formed in October 2013 to evaluate big data workload(s) and make recommendations to the TPC general council
- TPC-BD Subcommittee formed in February 2014 to develop an Express benchmark based on already popular TeraSort workload
- In July 2014 TPCx-HS became industry's first standard for benchmarking Big Data Systems
- TPC to continued to work on other benchmark(s). TPC announced TPC-DS v2 in 2015 and TPCx-BB in 2016

TPC Big Data Benchmark Standards

- TPC Express benchmark HS (TPCx-HS), 2014
  - Sort benchmark for Hadoop Systems
- TPC Enterprise benchmark DS (TPC-DS v2), 2015
  - Hadoop friendly version of TPC-DS
- TPC Express benchmark BB (TPCx-BB), 2016
  - Express benchmark based on Big Bench

TPC Express Benchmark HS

- Industry's first standard for benchmarking big data systems to provide the industry with verifiable performance, price-performance and availability metrics of hardware and software systems dealing with big data
- First benchmark developed through the Express benchmark category
- http://www.tpc.org/tpcx-hs/default.asp

TPC “Express” Benchmark Standards

- To keep pace with rapidly changing industry demands
- Easy to implement, run and publish, and less expensive
- Test sponsor is required to use the TPC provided kit
- The vendor may choose an independent audit or peer audit
- 60 day review/challenge window apply (as per TPC policy)
- Approved by super majority of the TPC General Council, No Mail Ballot
- All publications are required to follow the TPC Fair Use Policy

TPCx-HS Benchmark

- x: Express, H: Hadoop, S:Sort
- Provides verifiable performance, price/performance, general availability, and optional energy consumption metrics of big data systems
- Enable measurement of both hardware and software including Hadoop Runtime, Hadoop Filesystem API compatible systems and MapReduce layers
- Primary audience is enterprise customers (not public clouds)

TPCx-HS Workload

- Based on TeraSort workload
- TeraSort is part of Apache Hadoop distribution. org.apache.hadoop.examples.terasort
- A valid run consists of five separate phases run sequentially
- The benchmark test consists of two runs and run with lower metric is reported
- No configuration or tuning changes or reboot are allowed between the two runs

TPCx-HS Scale Factors

- The TPCx-HS follows a stepped Scale factor model (like in TPC-H and TPC-DS)
- The test dataset must be chosen from the set of fixed Scale Factors defined as follows:
- 1TB, 3TB, 10TB, 30TB, 100TB, 300TB, 1000TB, 3000TB, 10000TB.
- The corresponding number of records are
- 10B, 30B, 100B, 300B, 1000B, 3000B, 10000B, 30000B, 100000B, where each record is 100 bytes generated by HSGen
- The TPC will continuously evaluate adding larger Scale Factors and retiring smaller Scale Factors based on industry trends

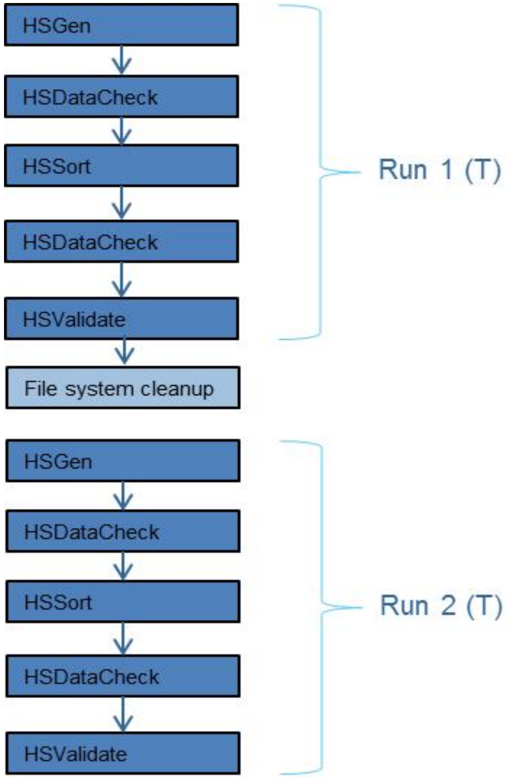
TPCx-HS Kit

- The TPCx-HS kit contains the following:
- TPCx-HS Specification document
- TPCx-HS Users Guide documentation
- Scripts to run the benchmark
- Java code to execute the benchmark load
- More Information http://www.tpc.org/tpcx-hs/default.asp

TPCx-HS Contributors

- Developing an industry standard benchmark for a new environment like Big Data has taken the dedicated efforts of experts across many companies. Thanks to:
- Andrew Bond (Red Hat), Andrew Masland (NEC), Avik Dey (Intel), Brian Caufield (IBM), Chaitanya Baru (SDSC), Da Qi Ren (Huawei), Dileep Kumar (Cloudera), Jamie Reding (Microsoft), John Fowler (Oracle), John Poelman (IBM), Karthik Kulkarni (Cisco), Meikel Poess (Oracle), Mike Brey (Oracle), Mike Crocker (SAP), Paul Cao (HP), Reza Taheri (VMware), Simon Harris (IBM), Tariq Magdon-Ismail (VMware), Wayne Smith (Intel)and Yanpei Chen (Cloudera)

TPCx-HS Execution

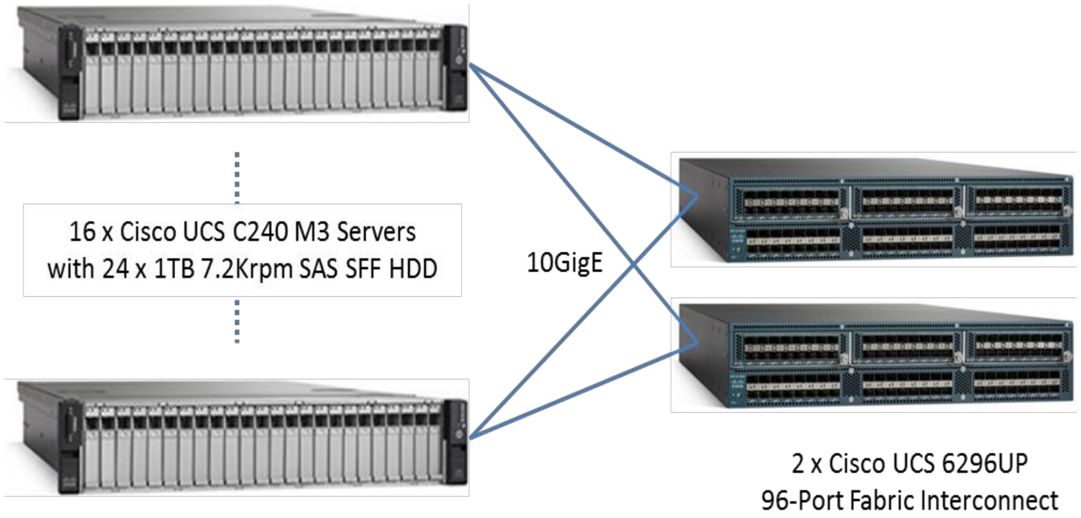


- HSGen is a program to generate the data at a particular Scale Factor
- HSDataCheck is a program to check the compliance of the dataset and replication
- HSSort is a program to sort the data into a total order
- HSValidate is a program that validates the output is sorted

The **performance run** is defined as the run with the lower Performance Metric. The **repeatability run** is defined as the run with the higher Performance Metric

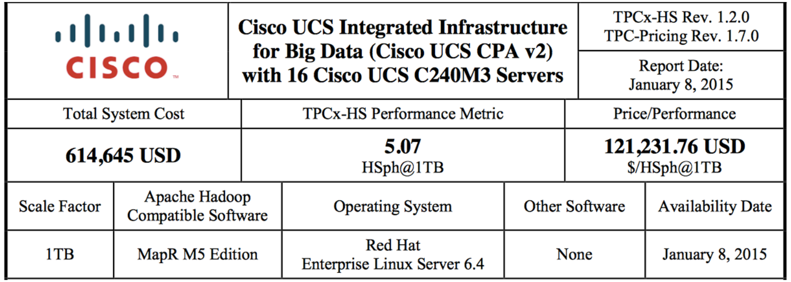
Experiments and Analysis

First TPCx-HS Publication



16 x Cisco UCS C240 M3 Servers
with 24 x 1TB 7.2Krpm SAS SFF HDD

2 x Cisco UCS 6296UP
96-Port Fabric Interconnect



cisco	**Cisco UCS Integrated Infrastructure for Big Data (Cisco UCS CPA v2) with 16 Cisco UCS C240M3 Servers**	TPCx-HS Rev. 1.2.0 TPC-Pricing Rev. 1.7.0 Report Date: January 8, 2015
Total System Cost	TPCx-HS Performance Metric	Price/Performance
614,645 USD	**5.07** HSph@1TB	**121,231.76 USD** $/HSph@1TB

Scale Factor	Apache Hadoop Compatible Software	Operating System	Other Software	Availability Date
1TB	MapR M5 Edition	Red Hat Enterprise Linux Server 6.4	None	January 8, 2015

Observations

- Significant performance improvement with tuning CPU, Memory, IO and Network
- x-HS did not perform out of the box on some commercial Hadoop distributions, unveiled interesting problems
- 50% Performance improvement in 6 months, 2x Performance improvement in 12 months
- Workload is simple but does exercise major subsystems ‘fairly’ equally
- SUT – realistic configurations
- Performance improvements haven't been following SPECintRate

CPU and Memory Tuning: Example

Parameters	Settings
Turbo Boost:	Enabled
Enhanced Intel Speedstep	Enabled
Hyper threading	Enabled
Core Multiprocessing	All
Executive Disabled Bit	Platform Default
Virtualization Technology	Disabled
Hardware Pre-fetcher	Enabled
Adjacent Cache Line Pre-fetcher	Enabled
DCU Streamer Pre-fetcher	Enabled
DCU IP Pre-fetcher	Enabled
Direct Cache access	Enabled
Processor C state	Disabled
Processor CIE	Disabled
Processor C3 Report	Disabled
Processor C6 Report	Disabled
Processor C7 Report	Disabled
CPU Performance	Enterprise
Max Variable MTRR Setting	Platform Default
Local X2 APIC	Platform Default
Power Technology	Performance
Energy Performance	Performance
Frequency Floor Override	Enabled
P-State Coordination	Hw-all
DRAM Clock Throttling	Performance
Channel Interleaving	Platform Default
Rank Interleaving	Platform Default
Demand Scrub	Disabled
Patrol Scrub	Disabled

Parameters	Settings
Memory RAS Configuration	Maximum-Performance
NUMA	Enabled
LV DDR Mode	Performance-mode
DRAM Refresh Rate	1x
DDR 3 Voltage Selection	Platform Default

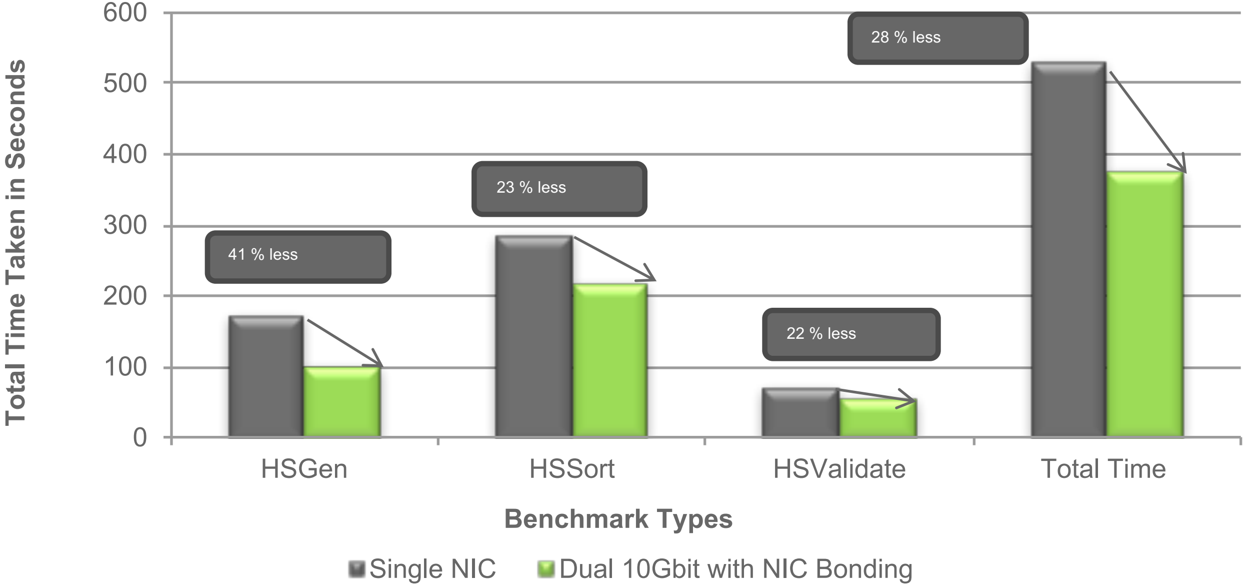
Network Tuning: Example

Parameters	Tuned Value
net.core.somaxconn	1024
net.ipv4.tcp_retries2	5
net.ipv4.ip_forward	0
net.ipv4.conf.default.rp_filter	1
net.ipv4.conf.all.rp_filter	1
net.ipv4.conf.default.accept_source_route	0
net.ipv4.tcp_syncookies	1
net.ipv4.conf.all.arp_filter	1
net.ipv4.tcp_mtu_probing	1
net.ipv4.icmp_echo_ignore_broadcasts	1
net.ipv4.conf.default.promote_secondaries	1
net.ipv4.conf.all.promote_secondaries	1
net.core.rmem_max	16777216
net.core.wmem_max	16777216
net.ipv4.tcp_rmem	4096 87380 16777216
net.ipv4.tcp_wmem	4096 65536 16777216
net.core.netdev_max_backlog	10000
net.core.netdev_max_backlog	10000

IO Tuning: Example

Parameters	Settings
RAID	RAID 0 of individual disk drives
Controller Cache	Always Write Back ,NoCacheBadBBU, Read Ahead
Stripe Size	1024K
Disk Drive Cache	Enabled (Read) Disable (Write)

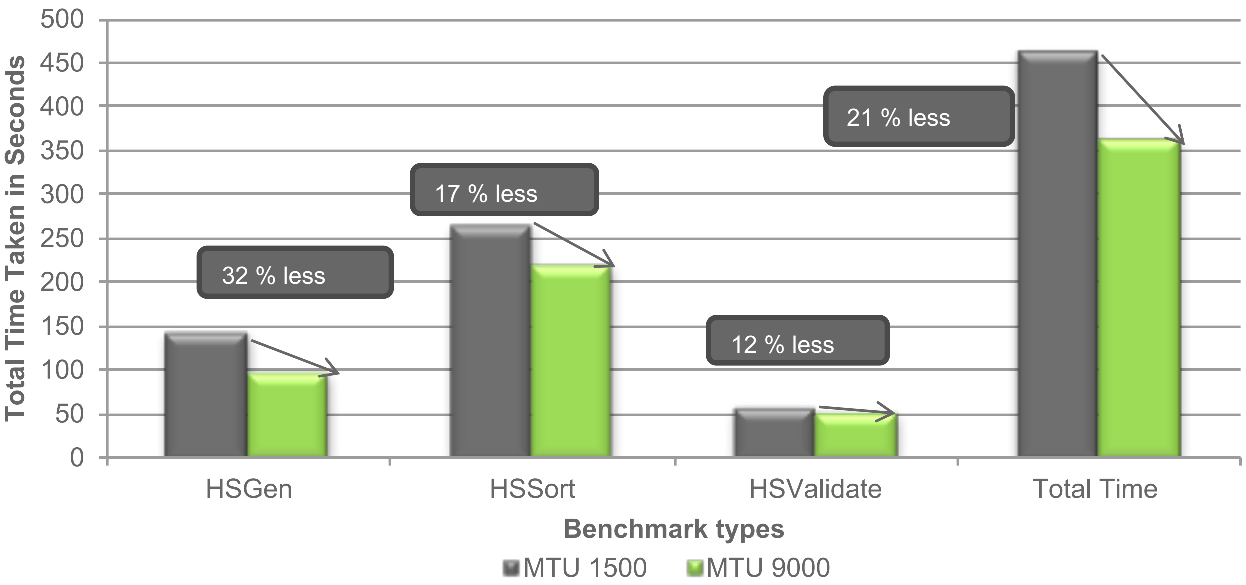
Single NIC vs Dual 10Gbit with NIC Bonding



	10Gbits	2x 10Gbits
HSGen	173	102
HSSort	286	218
HSValidate	69	55
Total Time	528	375
HSph@SF	**5.2**	**7.4**

28% better performance

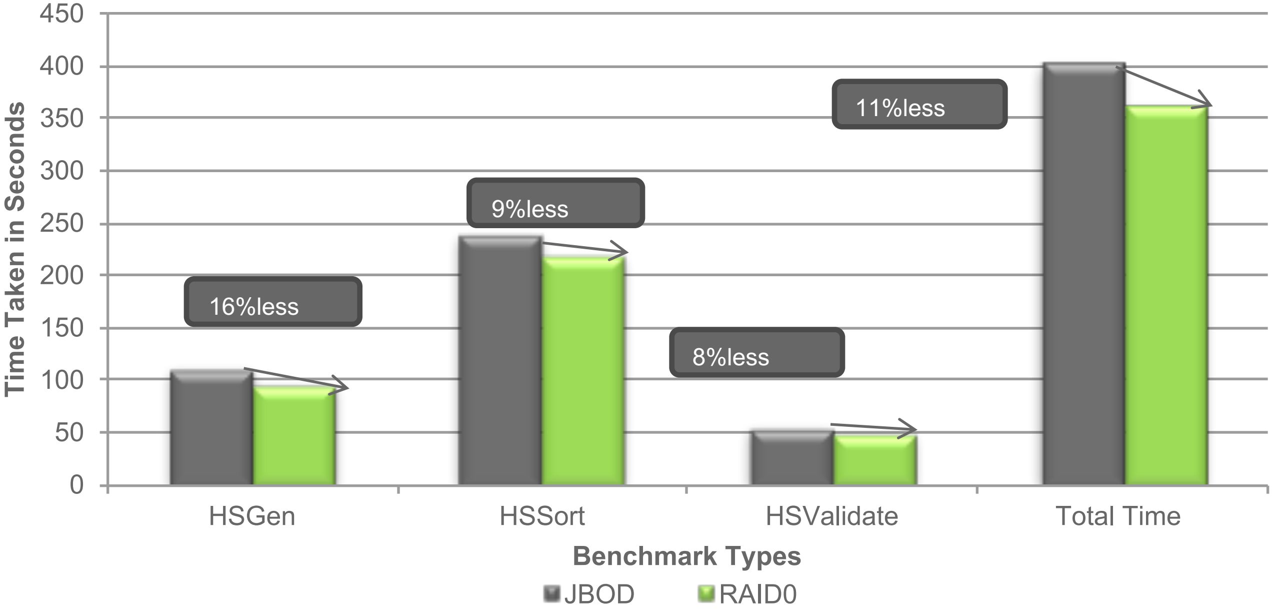
MTU 1500 vs MTU 9000



	MTU 1500	MTU 9000
HSGen	140	95
HSSort	264	217
HSValidate	56	49
Total Time	460	361
HSph@SF	**6.0**	**7.7**

21% better performance

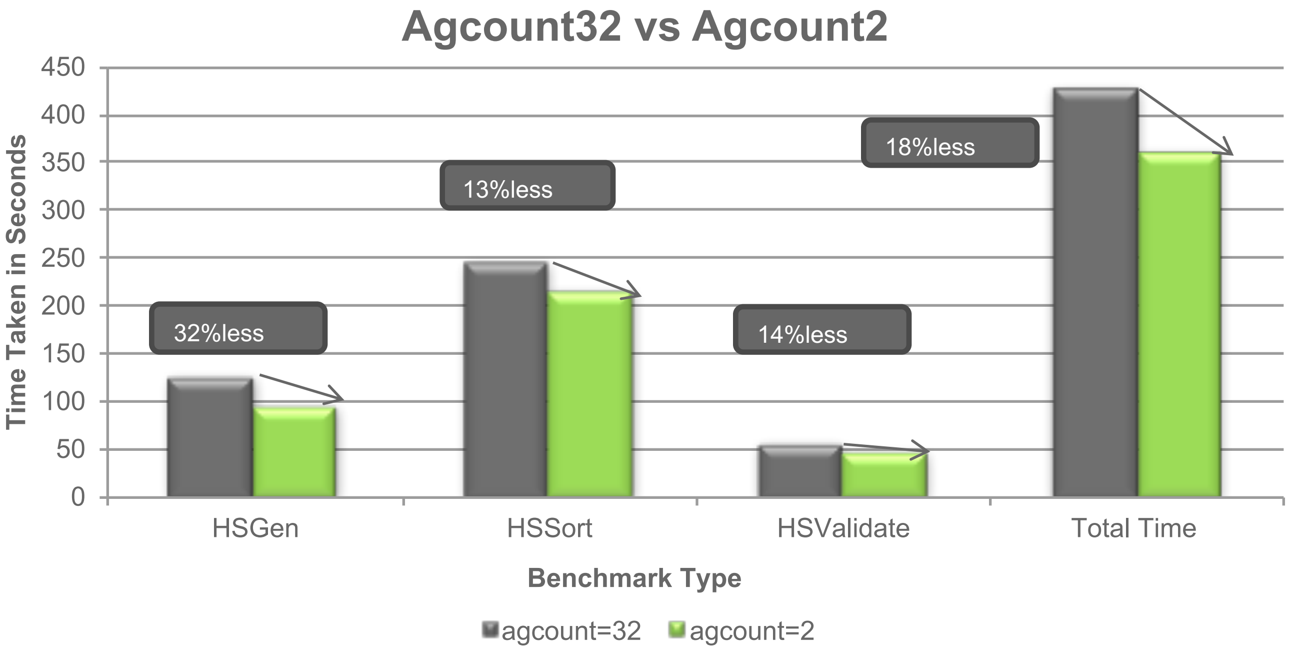
JBOD vs RAID 0



	JBOD	RAID0
HSGen	111	95
HSSort	237	217
HSValidate	53	49
Total Time	401	361
HSph@SF	**6.9**	**7.7**

11% better performance

XFS agcount32 vs agcount2


Agcount32 vs Agcount2

32%less

13%less

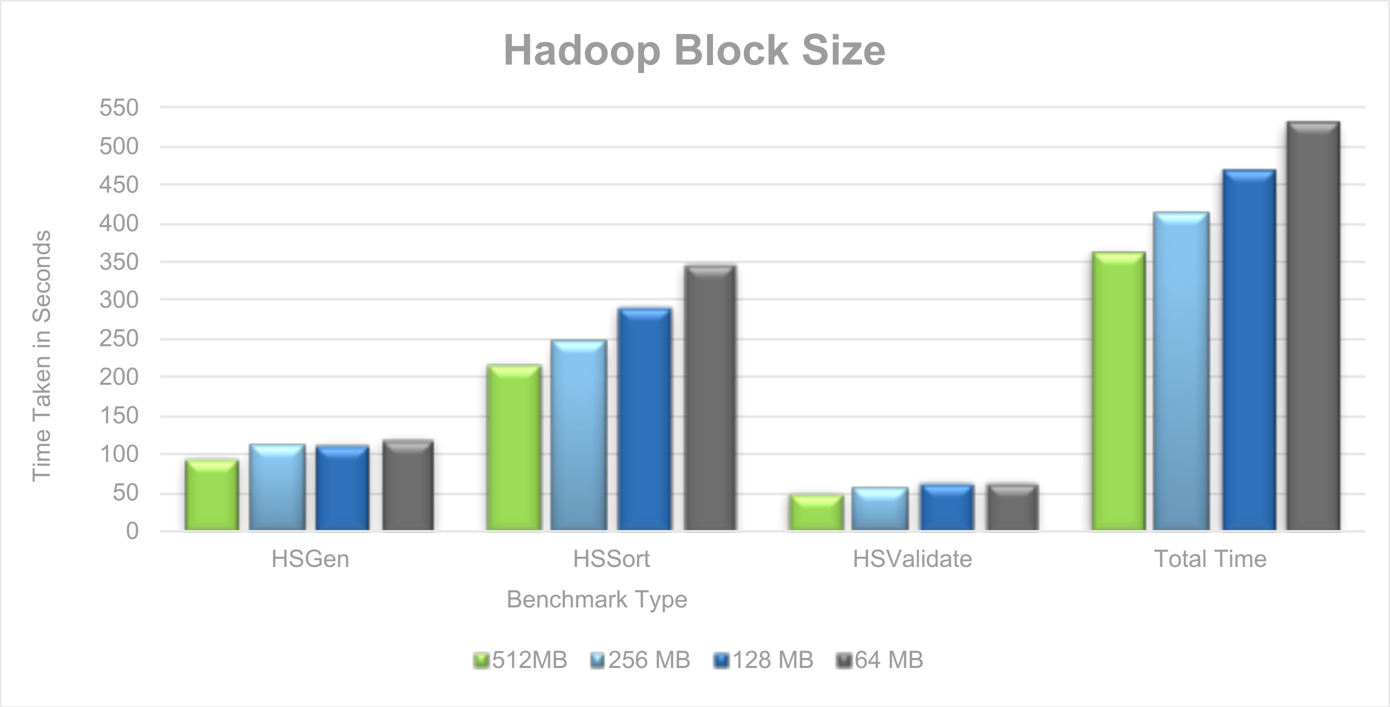
14%less

18%less

	agcount32	Agcount2
HSGen	126	95
HSSort	246	217
HSValidate	56	49
Total Time	428	361
HSph@SF	**6.5**	**7.7**

18% better performance

Hadoop Block Sizes: 512, 256, 128, 64MB



	512MB	256MB
HSGen	95	110
HSSort	217	246
HSValidate	49	55
Total Time	361	411
HSph@SF	7.7	6.8

	128MB	64MB
HSGen	111	119
HSSort	291	344
HSValidate	65	65
Total Time	467	528
HSph@SF	5.9	5.3

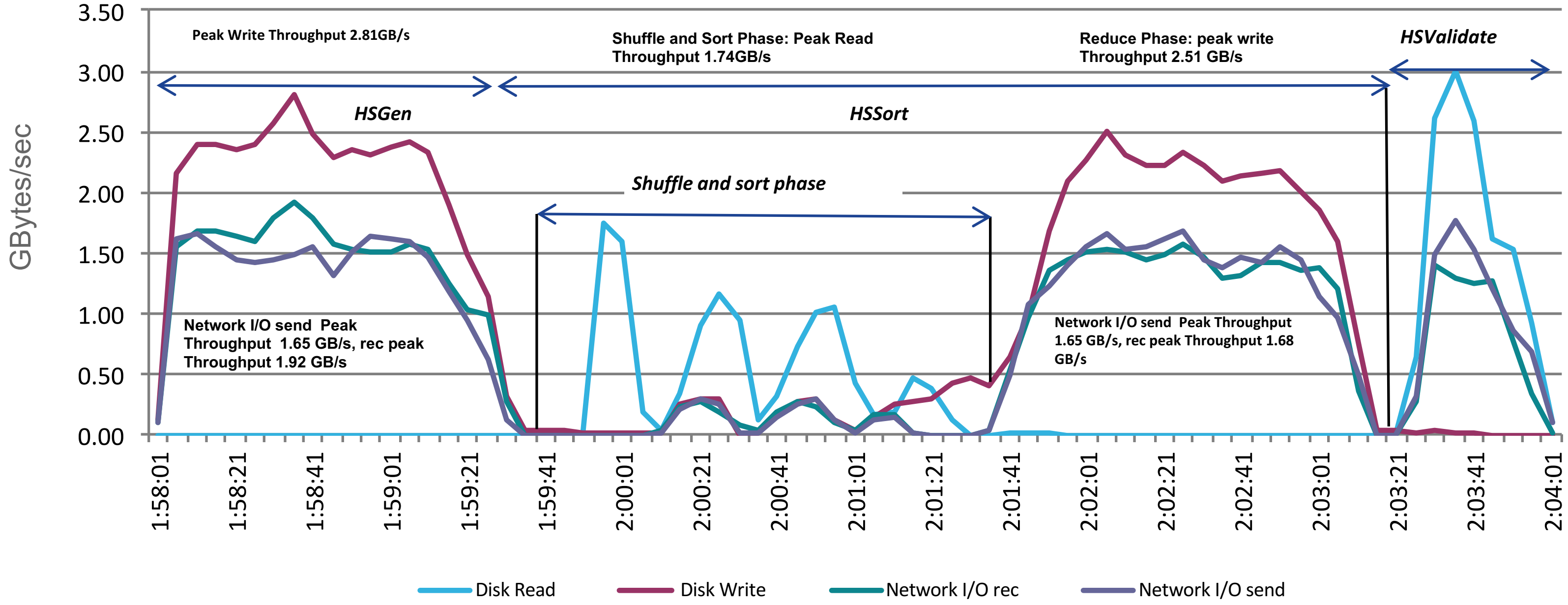
Hadoop Tuning: Example

Tuning Parameter	Values
Mapred.map.tasks	540
Mapred.reduce.tasks	450
mapred.tasktracker.map.tasks.maximum	36
mapred.tasktracker.reduce.tasks.maximum	30
mapred.map.child.java.opts	-Xmx800m -Xms800m -Xmn256m
mapred.reduce.child.java.opts	-Xmx1200m -Xmn256m
mapred.child.ulimit	4096MB
io.sort.mb	1024MB
io.sort.factor	64
io.sort.record.percent	0.15
Io.sort.spill.percent	0.98
mapred.job.reuse.jvm.num.tasks	-1
mapred.reduce.parallel.copies	20
mapred.reduce.slowstart.completed.maps	0
tasktracker.http.threads	120
mapred.job.reduce.input.buffer.percent	0.7
mapreduce.reduce.shuffle.maxfetchfailures	10
mapred.job.shuffle.input.buffer.percent	0.75
mapred.job.shuffle.merge.percent	0.95
mapred.inmem.merge.threshold	0
mapreduce.ifile.readahead.bytes	16777216
mapred.map.tasks.speculative.execution	False

Tuning Parameter	Values
dfs.blocksize	512MB
dfs.datanode.drop.cache.behind.writes	True
dfs.datanode.sync.behind.writes	True
dfs.datanode.drop.cache.behind.reads	True

Experiments were conducted on a 16 node cluster, one server configured as name node and 15 servers configured as data nodes, with two CPUs with a total of 24 cores/48 threads

IO and Network Utilization



Peak Write Throughput 2.81GB/s

Shuffle and Sort Phase: Peak Read Throughput 1.74GB/s

Reduce Phase: peak write Throughput 2.51 GB/s

HSValidate

HSGen

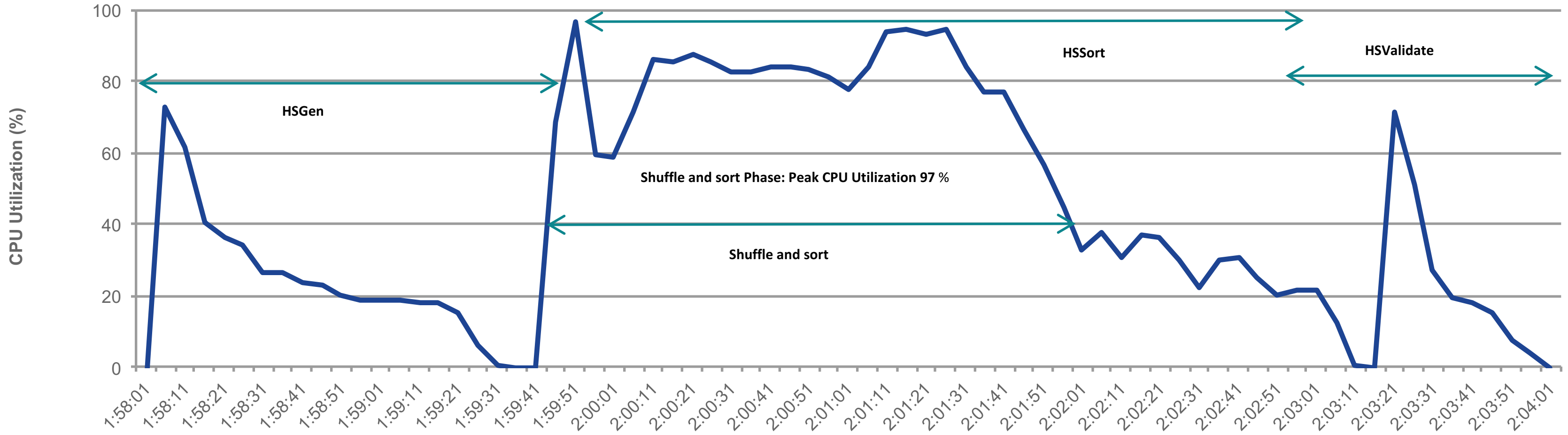
HSSort

Shuffle and sort phase

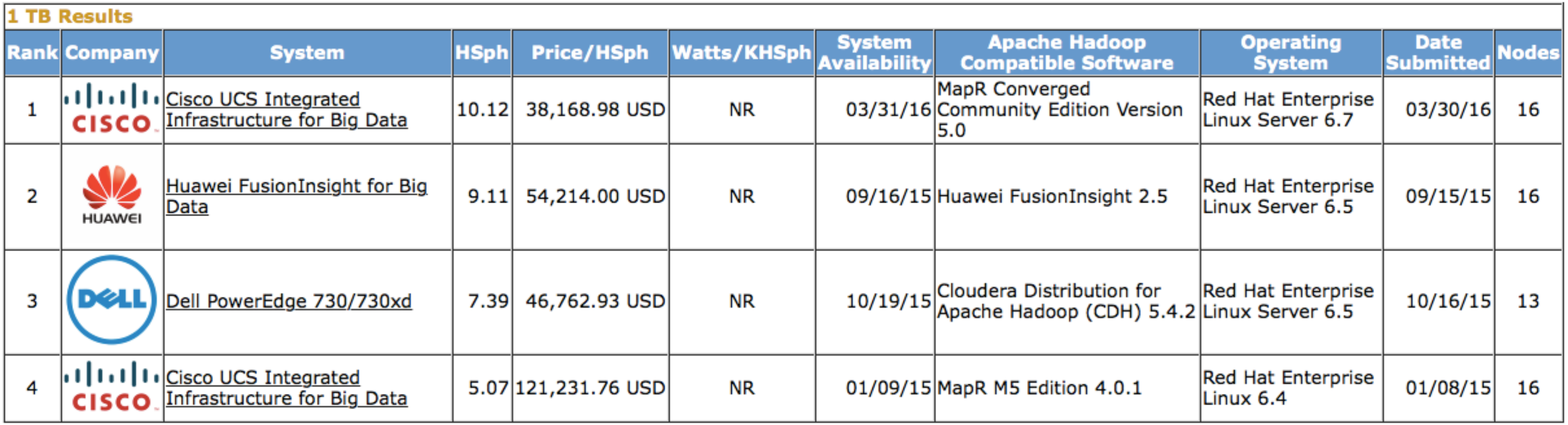
Network I/O send Peak Throughput 1.65 GB/s, rec peak Throughput 1.92 GB/s

Network I/O send Peak Throughput 1.65 GB/s, rec peak Throughput 1.68 GB/s

CPU Utilization



Analysis of Results and Trends

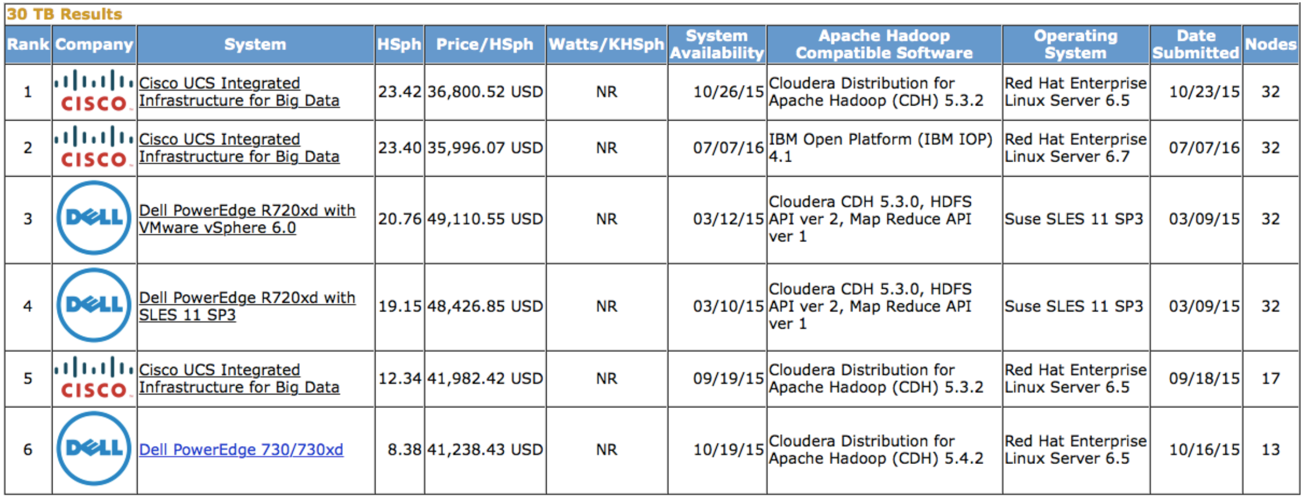


1 TB Results

Rank	Company	System	HSph	Price/HSph	Watts/KHSph	System Availability	Apache Hadoop Compatible Software	Operating System	Date Submitted	Nodes
1	CISCO	Cisco UCS Integrated Infrastructure for Big Data	10.12	38,168.98 USD	NR	03/31/16	MapR Converged Community Edition Version 5.0	Red Hat Enterprise Linux Server 6.7	03/30/16	16
2	HUAWEI	Huawei FusionInsight for Big Data	9.11	54,214.00 USD	NR	09/16/15	Huawei FusionInsight 2.5	Red Hat Enterprise Linux Server 6.5	09/15/15	16
3	DELL	Dell PowerEdge 730/730xd	7.39	46,762.93 USD	NR	10/19/15	Cloudera Distribution for Apache Hadoop (CDH) 5.4.2	Red Hat Enterprise Linux Server 6.5	10/16/15	13
4	CISCO	Cisco UCS Integrated Infrastructure for Big Data	5.07	121,231.76 USD	NR	01/09/15	MapR M5 Edition 4.0.1	Red Hat Enterprise Linux 6.4	01/08/15	16

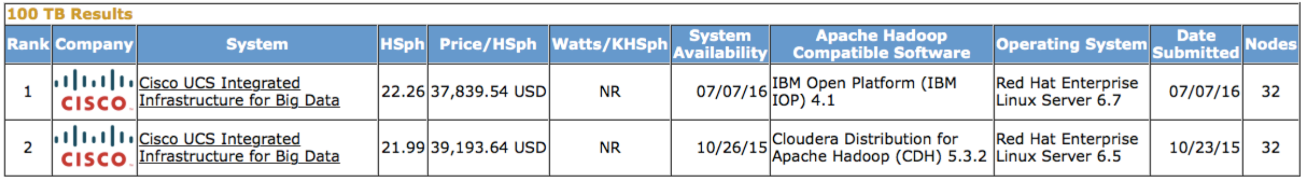
- 3 HW vendors, 4 Hadoop ISVs, Bare-metal and virtualized
- Significant performance improvement by tuning, applicable to real-life

Analysis of Results and Trends



30 TB Results

Rank	Company	System	HSph	Price/HSph	Watts/KHSph	System Availability	Apache Hadoop Compatible Software	Operating System	Date Submitted	Nodes
1	CISCO	Cisco UCS Integrated Infrastructure for Big Data	23.42	36,800.52 USD	NR	10/26/15	Cloudera Distribution for Apache Hadoop (CDH) 5.3.2	Red Hat Enterprise Linux Server 6.5	10/23/15	32
2	CISCO	Cisco UCS Integrated Infrastructure for Big Data	23.40	35,996.07 USD	NR	07/07/16	IBM Open Platform (IBM IOP) 4.1	Red Hat Enterprise Linux Server 6.7	07/07/16	32
3	DELL	Dell PowerEdge R720xd with VMware vSphere 6.0	20.76	49,110.55 USD	NR	03/12/15	Cloudera CDH 5.3.0, HDFS API ver 2, Map Reduce API ver 1	Suse SLES 11 SP3	03/09/15	32
4	DELL	Dell PowerEdge R720xd with SLES 11 SP3	19.15	48,426.85 USD	NR	03/10/15	Cloudera CDH 5.3.0, HDFS API ver 2, Map Reduce API ver 1	Suse SLES 11 SP3	03/09/15	32
5	CISCO	Cisco UCS Integrated Infrastructure for Big Data	12.34	41,982.42 USD	NR	09/19/15	Cloudera Distribution for Apache Hadoop (CDH) 5.3.2	Red Hat Enterprise Linux Server 6.5	09/18/15	17
6	DELL	Dell PowerEdge 730/730xd	8.38	41,238.43 USD	NR	10/19/15	Cloudera Distribution for Apache Hadoop (CDH) 5.4.2	Red Hat Enterprise Linux Server 6.5	10/16/15	13



100 TB Results

Rank	Company	System	HSph	Price/HSph	Watts/KHSph	System Availability	Apache Hadoop Compatible Software	Operating System	Date Submitted	Nodes
1	CISCO	Cisco UCS Integrated Infrastructure for Big Data	22.26	37,839.54 USD	NR	07/07/16	IBM Open Platform (IBM IOP) 4.1	Red Hat Enterprise Linux Server 6.7	07/07/16	32
2	CISCO	Cisco UCS Integrated Infrastructure for Big Data	21.99	39,193.64 USD	NR	10/26/15	Cloudera Distribution for Apache Hadoop (CDH) 5.3.2	Red Hat Enterprise Linux Server 6.5	10/23/15	32

- Publications at larger scale factors
- MR1 vs MR2

Summary

- Significant performance improvement with tuning CPU, Memory, IO, and Network
- 50% Performance improvement in 6 months, 2x Performance improvement in 12 months – based on published results
- Workload is simple but does exercise major subsystems ‘fairly’ equally
- Hadoop systems do not perform out of the box even on commercial Hadoop distributions, and unveiled interesting problems
- Applicability across broad range of system topologies and implementation methodologies
- The paper provides several BIOS, operating system (OS), Hadoop, and Java tunings that can maximize the performance of Hadoop cluster

Questions ?

Announcement from VLDB

7:00 to 9:00 PM Today

Mr. Ravi Shankar Prasad, Government of India's Union Minister of Information Technology, How data is instrumental for making India Digital

Prof. Deepak Phatak, IIT Bombay, The journey from VLDB 1996 till VLDB 2016 and vision beyond

Спасибо
Gracias
شكرا
Obrigado
Спасибо
Dank U
Grazie
Ευχαριστώ
Ευχαριστώ
Danke
Merci
Diolch
Obrigado
Thank You
Ngiyabonga
Dank U
Thank You
Ngiyabonga
Tack
Merci
Ευχαριστώ
Dzięki
Dzięki
Danke
Grazie
תודה
Dank U
Terima Kasih
Diolch
Grazie
Merci
Tack
Tack
Diolch
Gracias
Merci
תודה

धन्यवाद

നന്ദി