Deeper Insight into Operational BigData Cluster

Strata/Hadoop World 2013

Samuel Kommu

Technical Marketing Engineer

sakommu@cisco.com

Session Objectives

- Back to Basics
- Hadoop and Sorting
- Tuning – Validated parameters that help
- Visibility and Monitoring
- Integration with Splunk
- Recommendations
- Q & A

Big Data @ Cisco - www.cisco.com/go/bigdata

Multi-year network and compute analysis testing
(In conjunction with partners)

Hadoop World 2011 on Hadoop Network and Compute Considerations:
http://bit.ly/18s6h8y

Hadoop Summit 2012 on Network Reference Architectures (Best practices).
Slides: http://slidesha.re/1aNt3sJ Youtube: http://bit.ly/16ENk2y

Hadoop World 2012 Designing Hadoop for the Enterprise Data Center
http://bit.ly/1gTkCow

Hadoop Summit 2013
http://bit.ly/1aiqu7j

Certifications and Solutions with UCS C-Series and Nexus Series Switches

Cloudera Hadoop Certified Technology

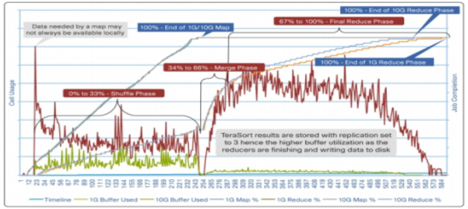
Oracle NoSQL Validated Solution

O'REILLY® Strata CONFERENCE + HADOOP WORLD

Tools and Techniques That Make Data Work

2012 HADOOP SUMMIT

Visibility & Monitoring



cloudera CERTIFIED TECHNOLOGY

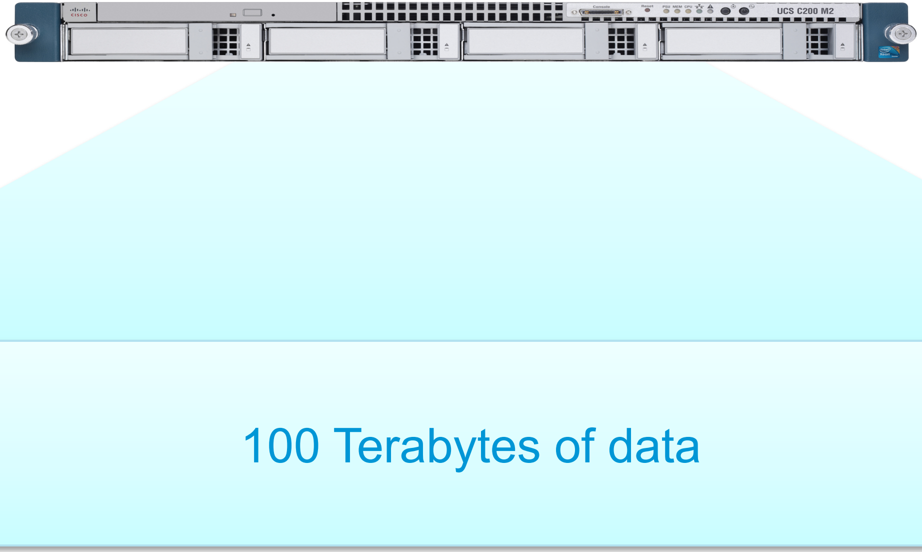
Back to Basics

HDFS & MapReduce

Hadoop HDFS

Problem with the monolithic system

Could take several days just to read
Per a test it took 11 days

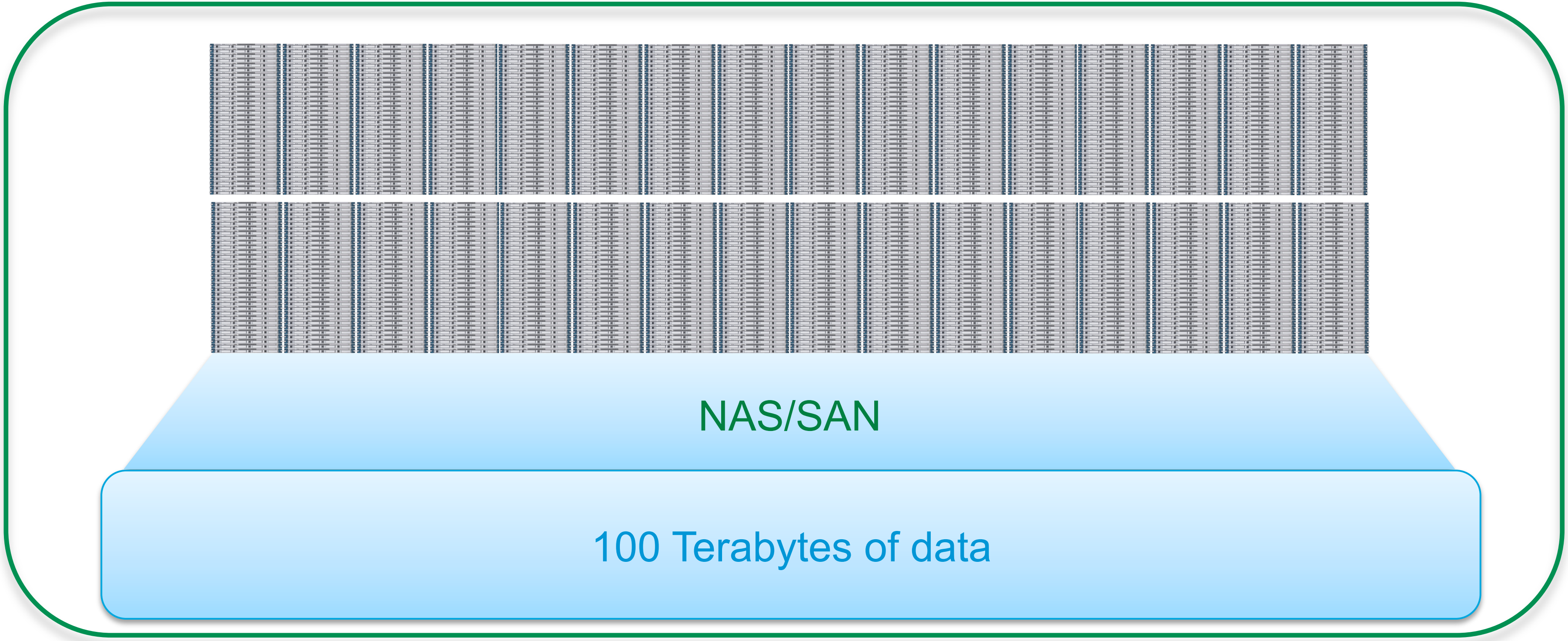


100 Terabytes of data

Hadoop HDFS

Scaling compute is not enough

Since you are still limited by the Throughput of the storage systems



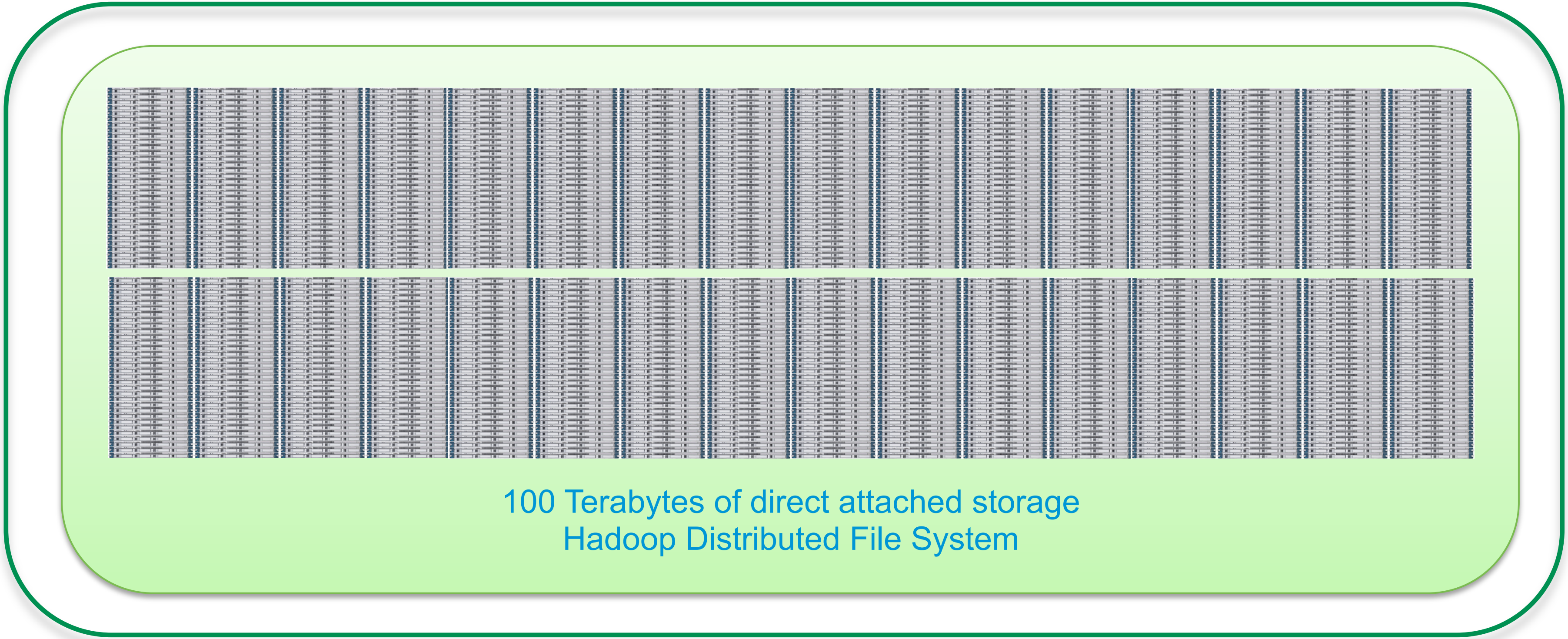
NAS/SAN

100 Terabytes of data

Hadoop HDFS

Solution – Go Parallel and use DAS
Local Data Access!

Same job took 15 minutes
once they went parallel
spreading the load across 1000 nodes



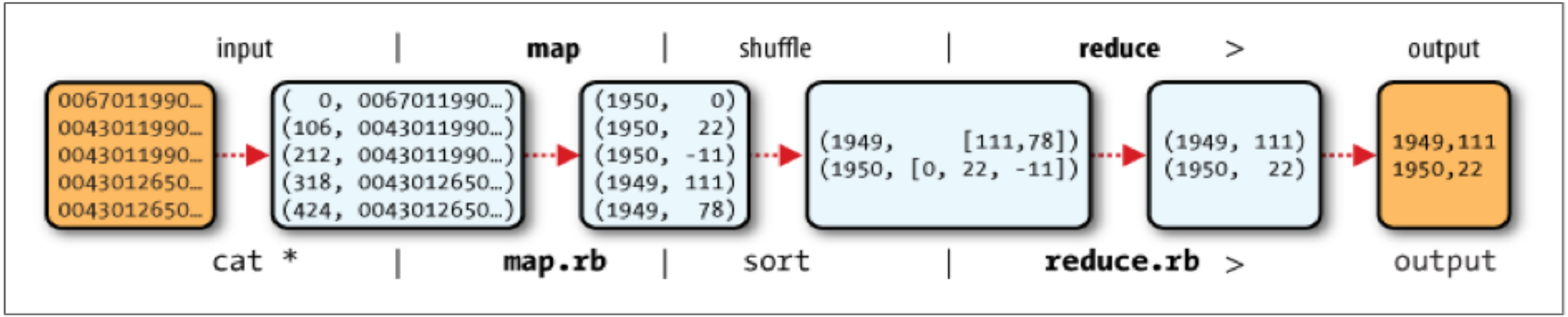
100 Terabytes of direct attached storage
Hadoop Distributed File System

Hadoop Map/Reduce

How does it work

Example:
Historic Weather Data (max temperatures/Year)

- Maps: Separates temperatures and year out of huge historical database
- Reducers: Finds the max per year



Source: O'Reilly Hadoop A definitive Guide

Hadoop Components and Operations

Hadoop Distributed File System

- The Data Ingest & Replication

  External Connectivity

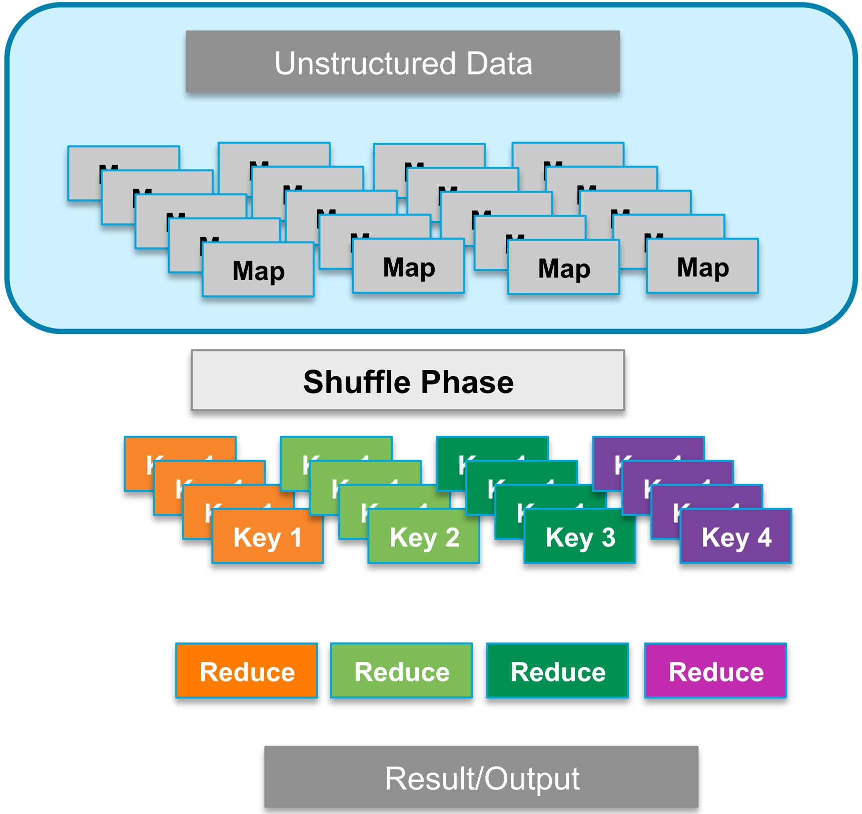
  East West Traffic (Replication of data blocks)

- **Map Phase –** Raw data Analyzed and converted to name/value pair.

  Workload translate to multiple batches of Map task

  Reducer can start the reduce phase ONLY after the entire Map set is complete

- Mostly a IO/compute function



Hadoop Components and Operations

Hadoop Distributed File System

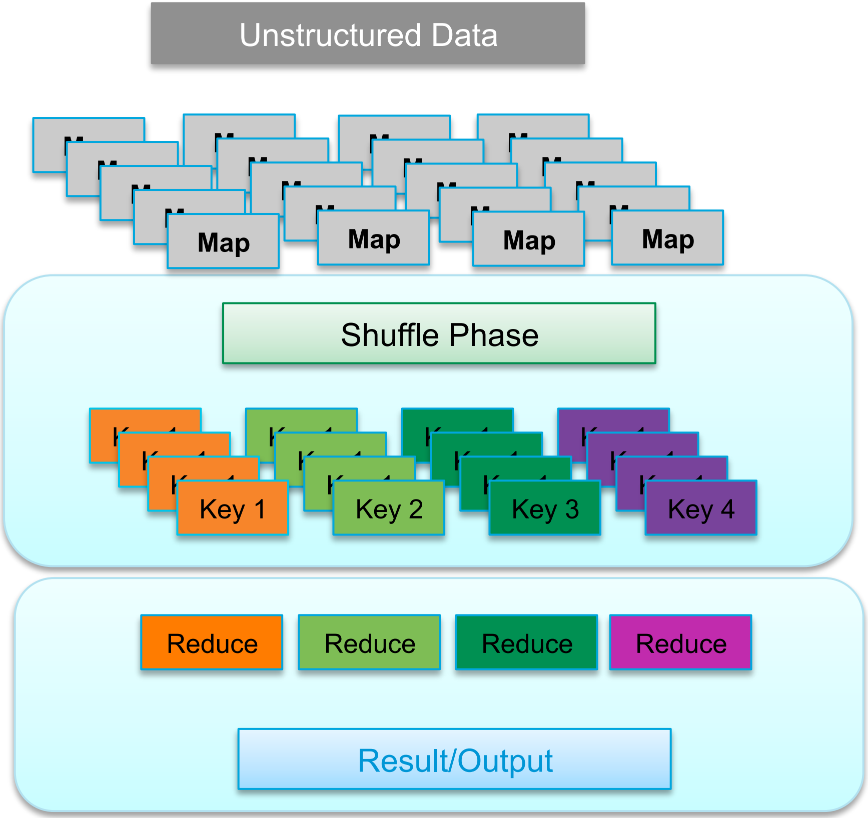
- **Shuffle Phase -** All name/value pair are sorted and grouped by their keys.
- Reducer is PULLING the data from the Mapper Nodes
- **High Network Activity**
- **Reduce Phase** – All values associates with a key are process for results, three phases

  Copy - get intermediate result from each data node local disk

  Merge - to reduce the number of files

  Reduce method
- **Output Replication Phase** - Reducer replicating result to multiple nodes

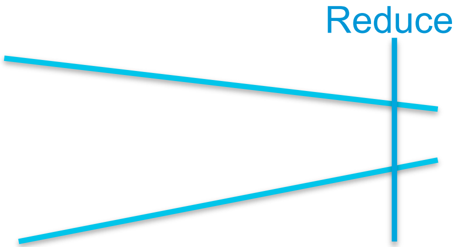
  Highest Network Activity
- Network Activities Dependent on Workload Behavior



Job Patterns

Analyze

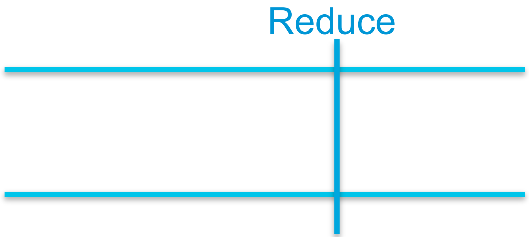
Ingress vs. Egress Data Set
1:0.3



The Time the reducers start is dependent on:
mapred.reduce.slowstart.completed.maps
It doesn't change the amount of data sent to Reducers, but may change the timing to send that data

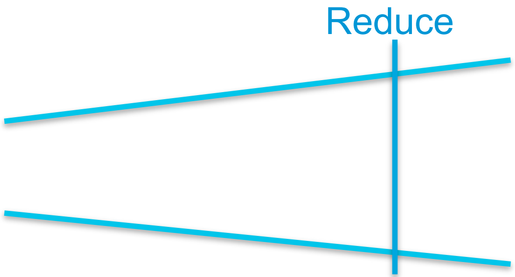
Extract Transform Load (ETL)

Ingress vs. Egress Data Set
1:1



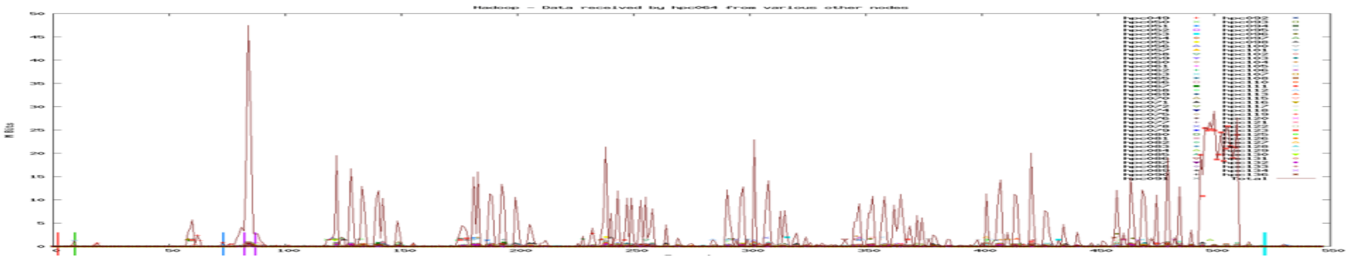
Explode

Ingress vs. Egress Data Set
1:2



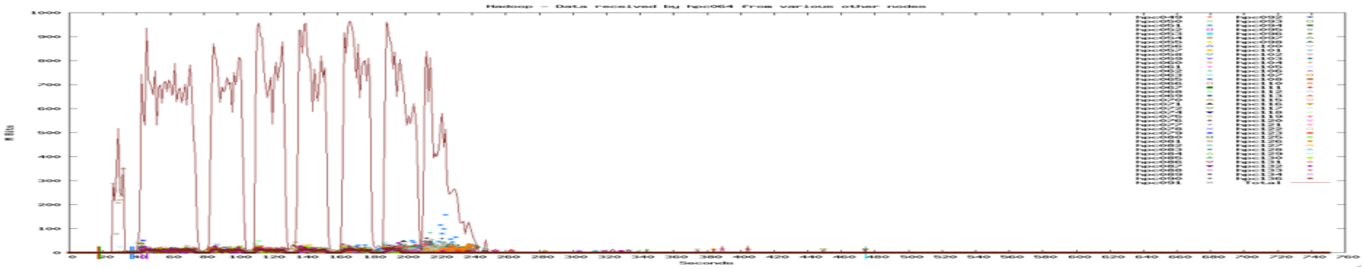
Job Patterns

Job Patterns have varying impact on network utilization



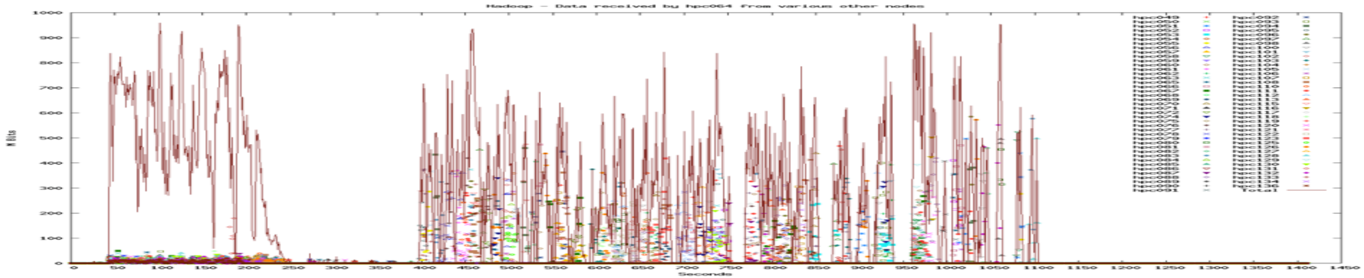
Analyze

Simulated with Shakespeare Wordcount



Extract Transform Load (ETL)

Simulated with TeraSort



Extract Transform Load (ETL)

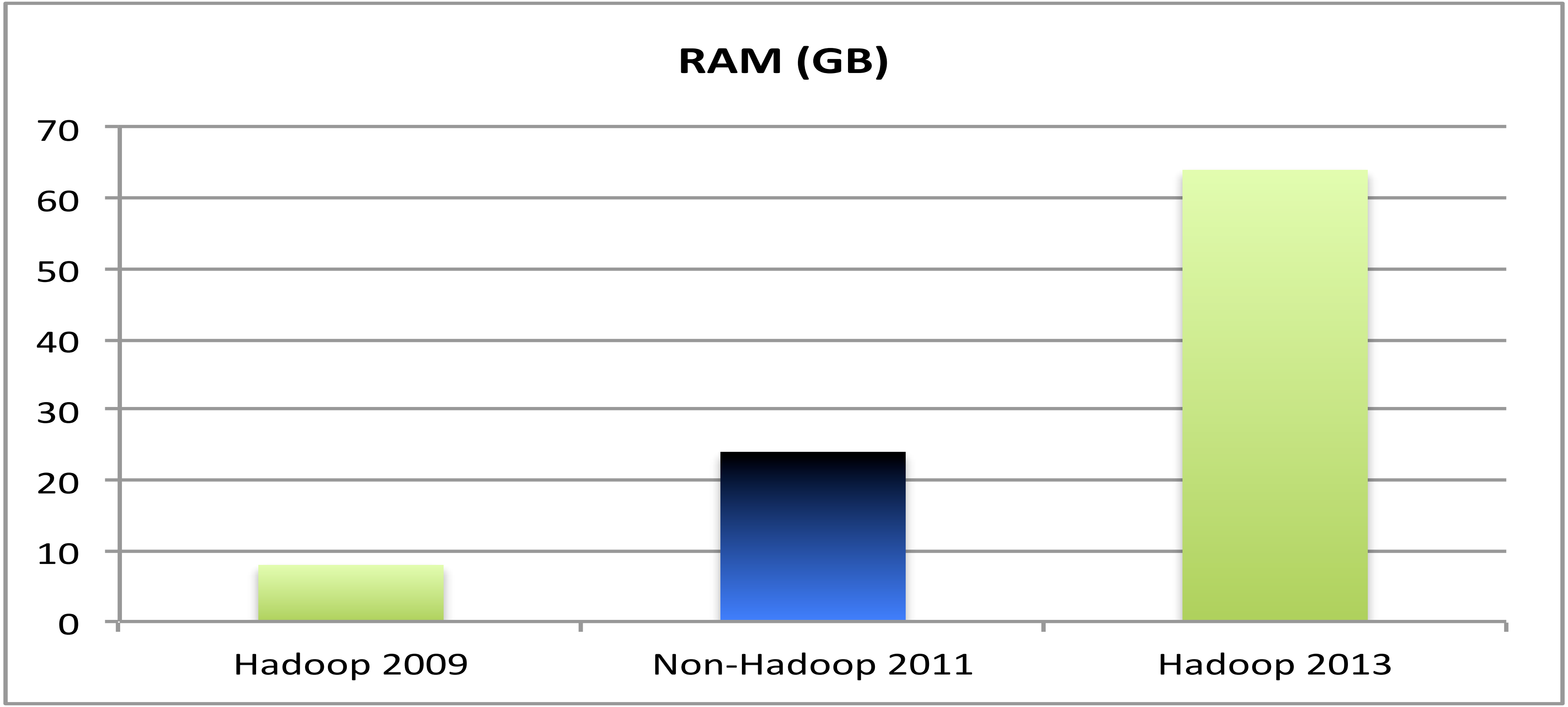
Simulated with TeraSort + output replication

Hadoop and Sorting

www.sortbenchmark.org

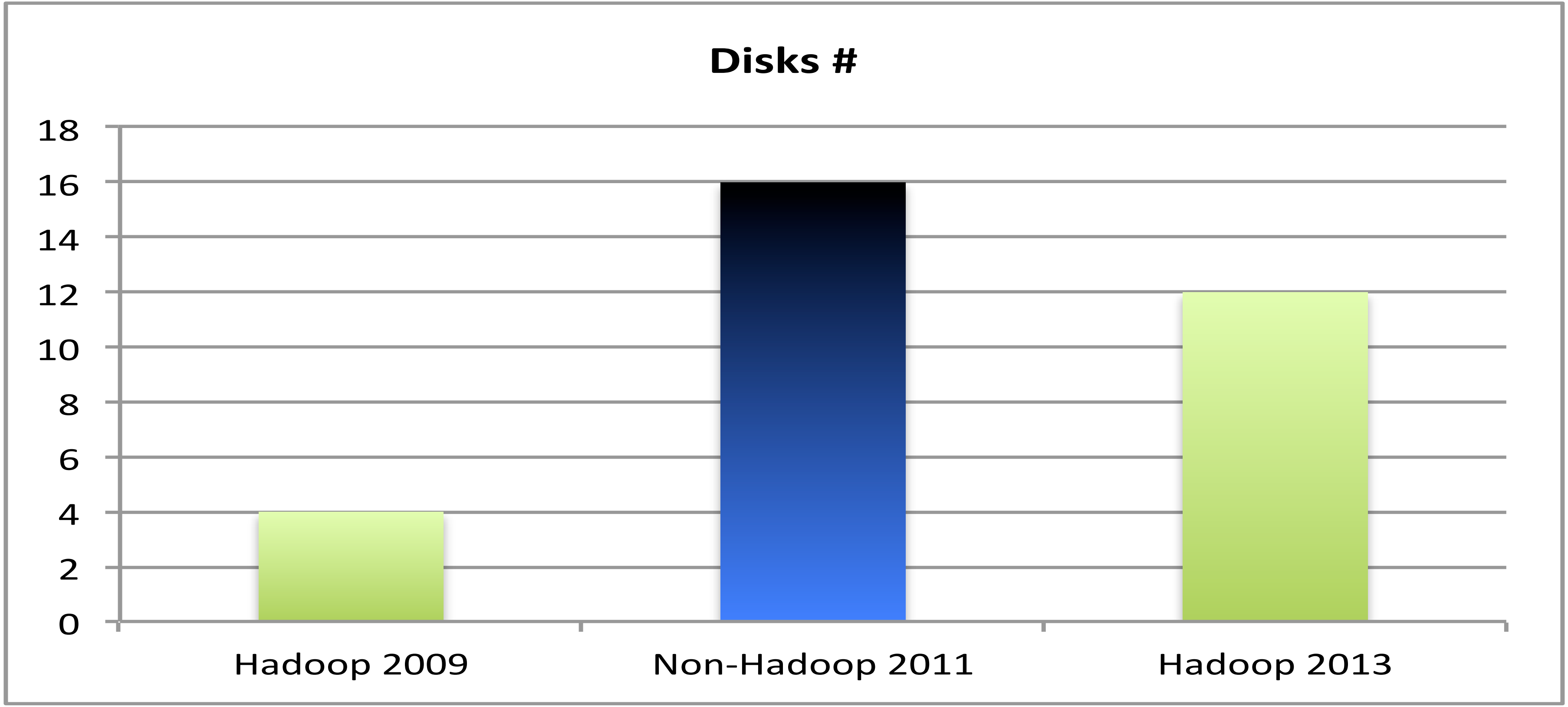
Hadoop for ETL?

Terasort Story



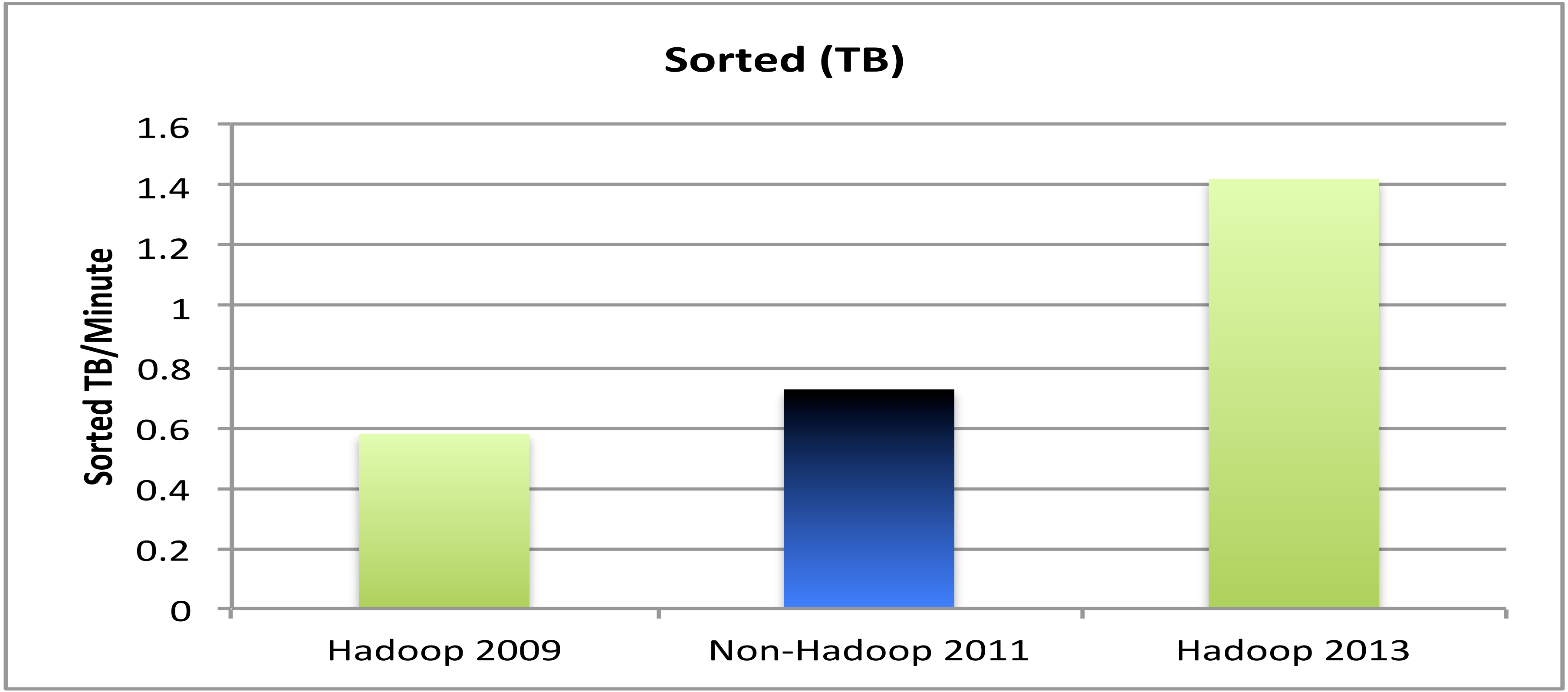
Hadoop for ETL?

Terasort Story



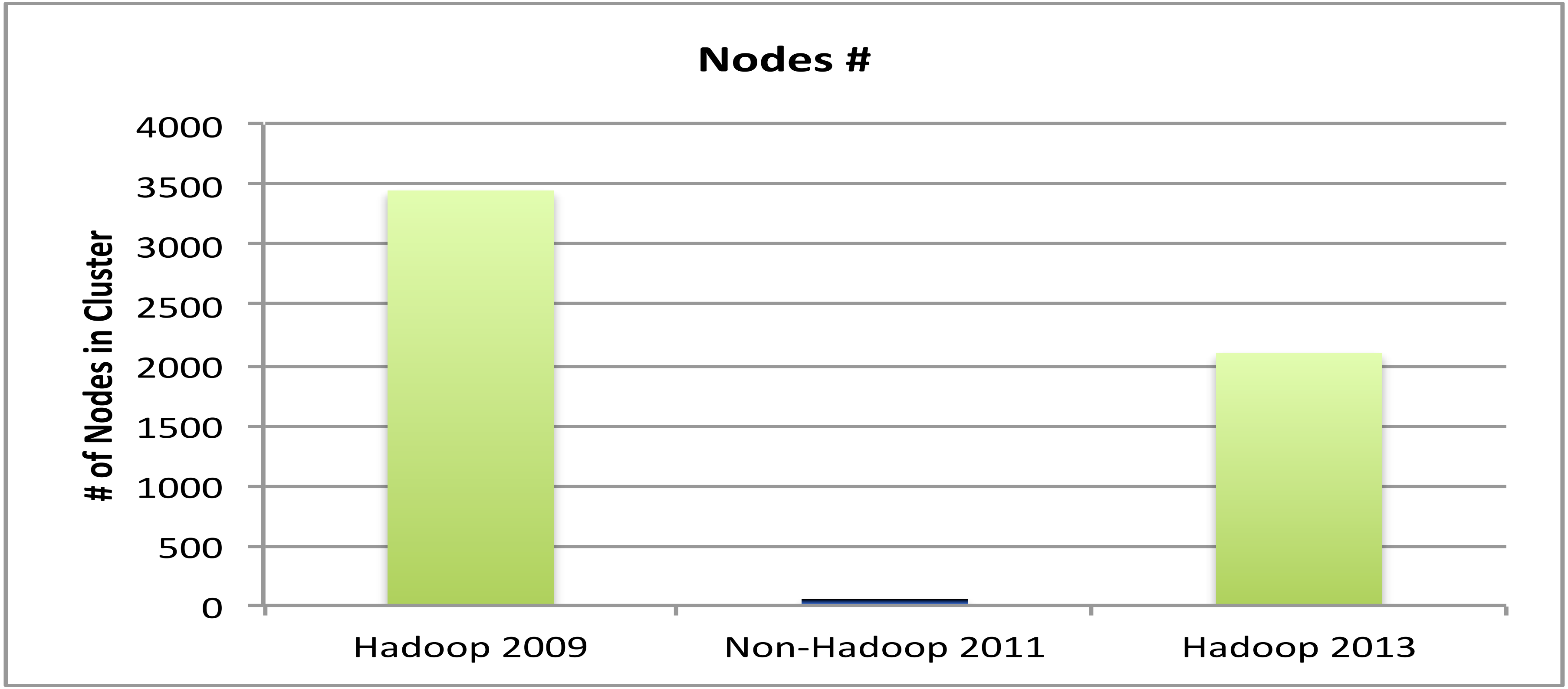
Hadoop for Extract Transform Load jobs

Sort Benchmark Results – TB Sorted/Minute



Hadoop for Extract Transform Load jobs

No. of Nodes used

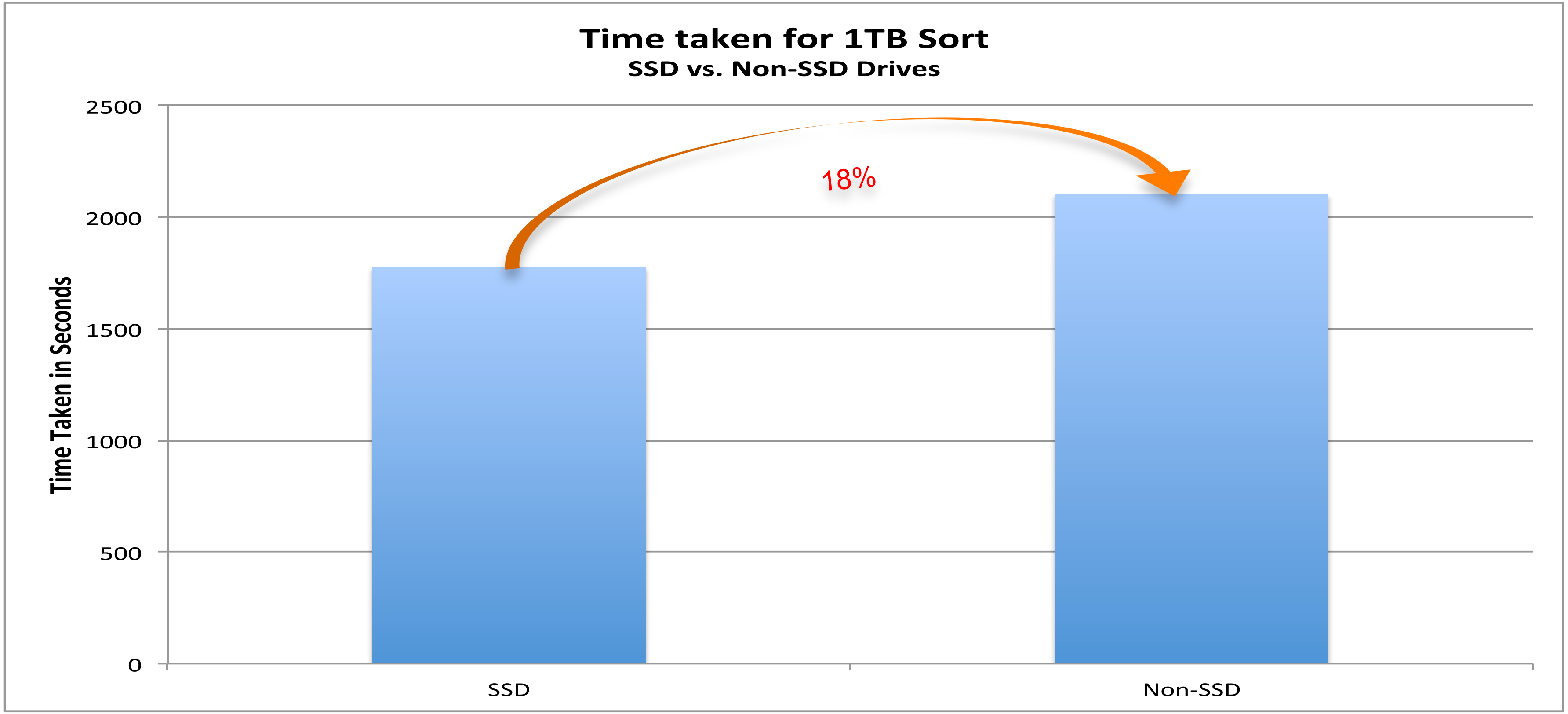


Hadoop Optimization/Tuning

SSDs & Transparent Hugepages

SSD Drives

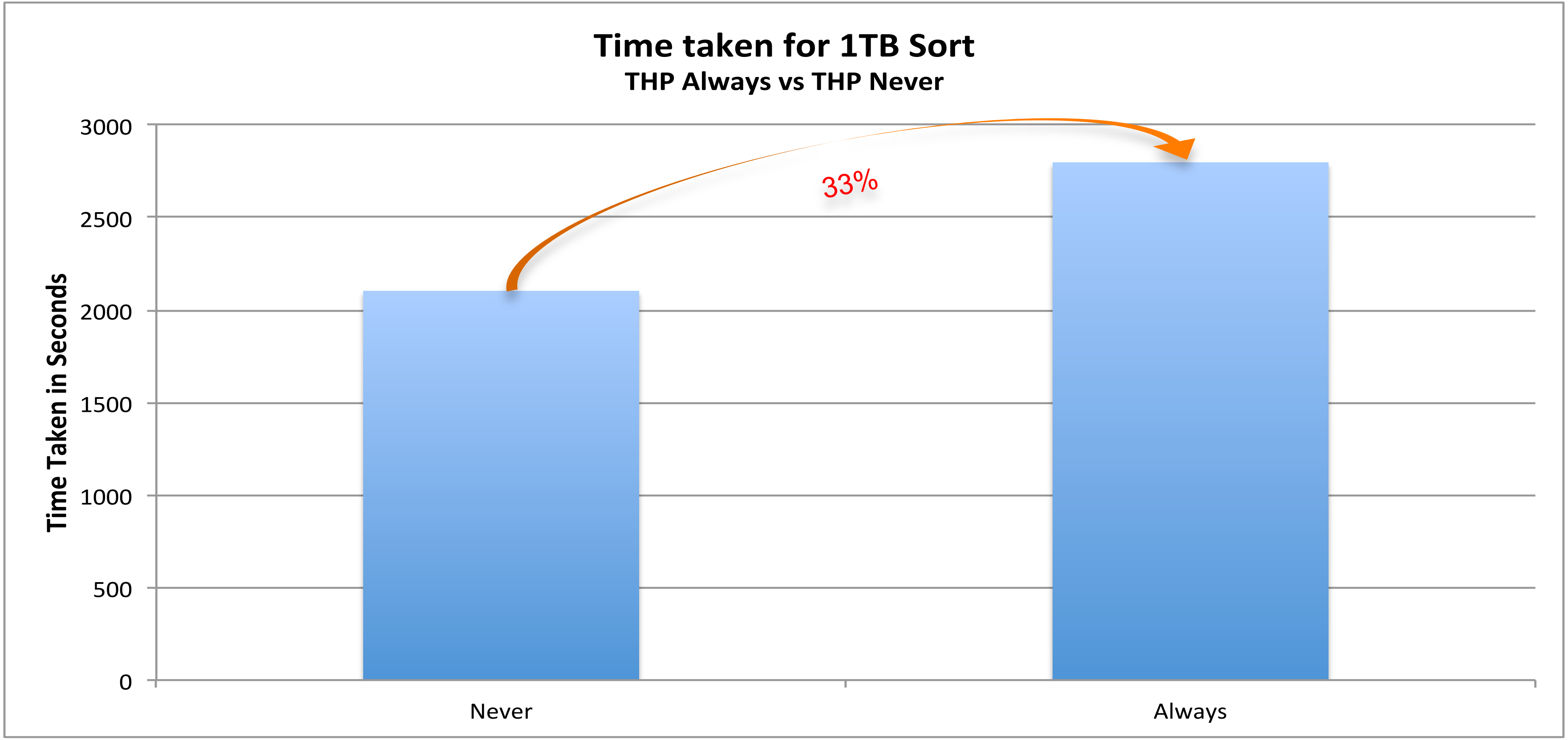
Running 1TB Terasort on 8 nodes – Lower is better



Transparent Hugepages

RedHat 6.2+ Parameter – Lower is better

```
To disable:
echo never > /sys/kernel/mm/redhat_transparent_hugepage/enabled
```



Hadoop + Network

Integration

Which port is connected?

```
n3548-001# show interface brief

--------------------------------------------------------------------------------
Ethernet      VLAN   Type Mode   Status  Reason                   Speed     Port
Interface                                                                   Ch #
--------------------------------------------------------------------------------
Eth1/1        1      eth  access up      none                       10G(D) --
Eth1/2        1      eth  access up      none                       10G(D) --
Eth1/3        1      eth  access up      none                       10G(D) --
Eth1/4        1      eth  access up      none                       10G(D) --
Eth1/5        1      eth  access up      none                       10G(D) --
.
.
Eth1/33       1      eth  access up      none                       10G(D) --
Eth1/34       1      eth  access up      none                       10G(D) --
Eth1/35       1      eth  access down    SFP not inserted           10G(D) --
Eth1/36       1      eth  access down    SFP not inserted           10G(D) --
Eth1/37       1      eth  access down    Administratively down      10G(D) -
.
```

What is connected there?

Classic Network View

```
n3548-001# show mac address-table dynamic
Legend:
        * - primary entry, G - Gateway MAC, (R) - Routed MAC, O - Overlay MAC
        age - seconds since first seen,+ - primary entry using vPC Peer-Link
   VLAN     MAC Address      Type      age       Secure NTFY    Ports
---------+-----------------+--------+---------+------+----+------------------
* 1          e8b7.484d.a208   dynamic   60570        F    F   Eth1/31
* 1          e8b7.484d.a20a   dynamic   60560        F    F   Eth1/31
* 1          e8b7.484d.a73e   dynamic   60560        F    F   Eth1/34
* 1          e8b7.484d.a740   dynamic   60560        F    F   Eth1/34
* 1          e8b7.484d.ad15   dynamic   60560        F    F   Eth1/28
* 1          e8b7.484d.ad17   dynamic   60560        F    F   Eth1/28
* 1          e8b7.484d.b3e9   dynamic   60570        F    F   Eth1/25
* 1          e8b7.484d.b3eb   dynamic   60560        F    F   Eth1/25
.
.
```

MAC Addresses of the connected devices … and the port they are on…

What is actually connected there?

Which server is connected to which port on the switch …

Note: Eth1/10 is missing because there is nothing connected to it

```
n3548-001# portServerMap
=======================================
Port    Server FQDN
---------------------------------------
Eth1/1  c200-m2-10g2-001.cluster10g.com
Eth1/2  c200-m2-10g2-002.cluster10g.com
Eth1/3  c200-m2-10g2-003.cluster10g.com
Eth1/4  c200-m2-10g2-004.cluster10g.com
Eth1/5  c200-m2-10g2-005.cluster10g.com
Eth1/6  c200-m2-10g2-006.cluster10g.com
Eth1/7  c200-m2-10g2-031.cluster10g.com
Eth1/8  c200-m2-10g2-008.cluster10g.com
Eth1/9  c200-m2-10g2-009.cluster10g.com
Eth1/11 c200-m2-10g2-011.cluster10g.com
.
.
.
```

What is running on those servers?

Hadoop - TaskTracker List

Note:
Eth1/1 is not on the list because it’s the namenode and is not running a tasktracker

Eth1/10 is not on the list because there is nothing connected to it

```
n3548-001# trackerList
============================================
Port    Server                          Server Port
--------------------------------------------
Eth1/2  c200-m2-10g2-002                50544
Eth1/3  c200-m2-10g2-003                41909
Eth1/4  c200-m2-10g2-004                36480
Eth1/5  c200-m2-10g2-005                38179
Eth1/6  c200-m2-10g2-006                51375
Eth1/7  c200-m2-10g2-031                41915
Eth1/8  c200-m2-10g2-008                50983
Eth1/9  c200-m2-10g2-009                37056
Eth1/11 c200-m2-10g2-011                35882
Eth1/12 c200-m2-10g2-012                44551
.
.
.
```

Which node is using the buffer?

```
n3548-001# bufferServerMap
==================================================================
Port     Server                   1sec     5sec     60sec    5min     1hr
------------------------------------------------------------------
Eth1/1   c200-m2-10g2-001         0KB      0KB      0KB      0KB      0KB
Eth1/2   c200-m2-10g2-002         384KB    384KB    1536KB   2304KB   2304KB
Eth1/3   c200-m2-10g2-003         384KB    384KB    1152KB   1536KB   1536KB
Eth1/4   c200-m2-10g2-004         384KB    384KB    2304KB   2304KB   2304KB
Eth1/5   c200-m2-10g2-005         384KB    384KB    768KB    1536KB   1536KB
Eth1/6   c200-m2-10g2-006         384KB    2304KB   2304KB   2304KB   2304KB
Eth1/7   c200-m2-10g2-031         384KB    384KB    3456KB   3840KB   3840KB
Eth1/8   c200-m2-10g2-008         768KB    768KB    2688KB   2688KB   2688KB
Eth1/9   c200-m2-10g2-009         384KB    384KB    2304KB   2304KB   2304KB
Eth1/11  c2[illegible]                     384KB    1920KB   1920KB   1920KB
.
.
.
```

Eth1/1(c200-m2-10g2-001) has 0 buffer usage because it's the name node

What’s running on this cluster + Buffer usage per server …

```
n3548-001# jobsBuffer
Hadoop Job Info ...
===================================================================
1 jobs currently running
JobId           RunTime(secs)   User    Priority
job_201306131423_0009   120     hadoop  NORMAL
===================================================================
Buffer Info - Per Port
Port    Server                  1sec    5sec    60sec   5min    1hr
-------------------------------------------------------------------
Eth1/1  c200-m2-10g2-001        0KB     0KB     0KB     0KB     0KB
Eth1/2  c200-m2-10g2-002        384KB   384KB   768KB   768KB   768KB
Eth1/3  c200-m2-10g2-003        384KB   384KB   1152KB  1152KB  1152KB
Eth1/4  c200-m2-10g2-004        384KB   1536KB  1536KB  1536KB  1536KB
Eth1/5  c200-m2-10g2-005        384KB   768KB   1152KB  1152KB  1152KB
.
.
```

What jobs were running during peak buffer usage … and for how long were they running

What’s running on this cluster + Buffer usage per server …

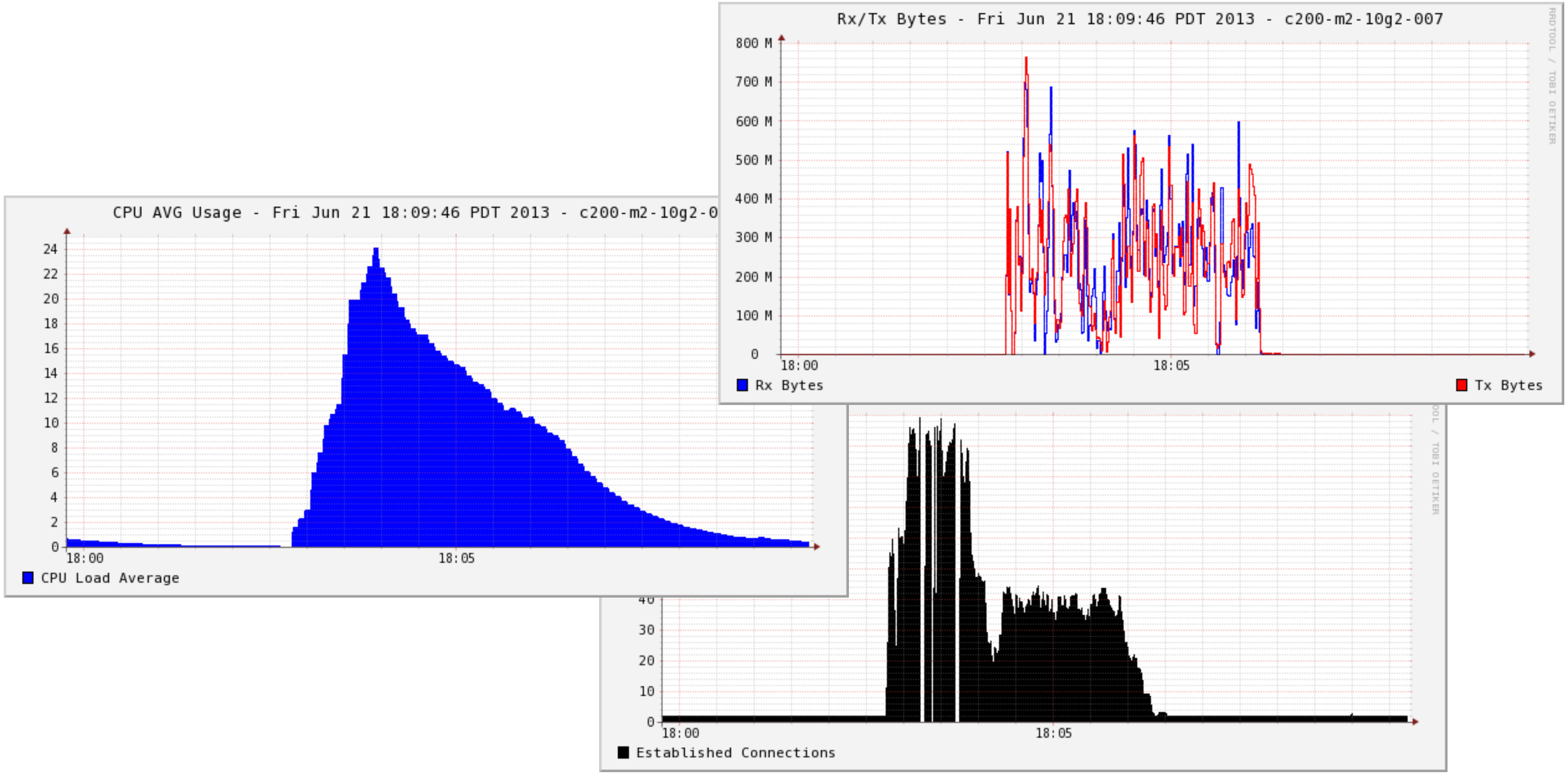
Historic look at the buffer usage …

```
n3548-001(config)# jobsBuffer
Hadoop Job Info ...
====================================================================
0 jobs currently running
JobId                RunTime(secs)    User     Priority
====================================================================
Buffer Info - Per Port
Port     Server                       1sec     5sec     60sec    5min     1hr
--------------------------------------------------------------------
Eth1/1   c200-m2-10g2-001             0KB      0KB      0KB      0KB      0KB
Eth1/2   c200-m2-10g2-002             0KB      0KB      0KB      1920KB   1920KB
Eth1/3   c200-m2-10g2-003             0KB      0KB      0KB      2304KB   2304KB
Eth1/4   c200-m2-10g2-004             0KB      0KB      0KB      2688KB   2688KB
Eth1/5   c200-m2-10g2-005             0KB      0KB      0KB      2304KB   2304KB
Eth1/6   c200-m2-10g2-006             0KB      0KB      0KB      2304KB   2304KB
Eth1/7   c200-m2-10g2-031             0KB      0KB      0KB      1920KB   2688KB
.
```

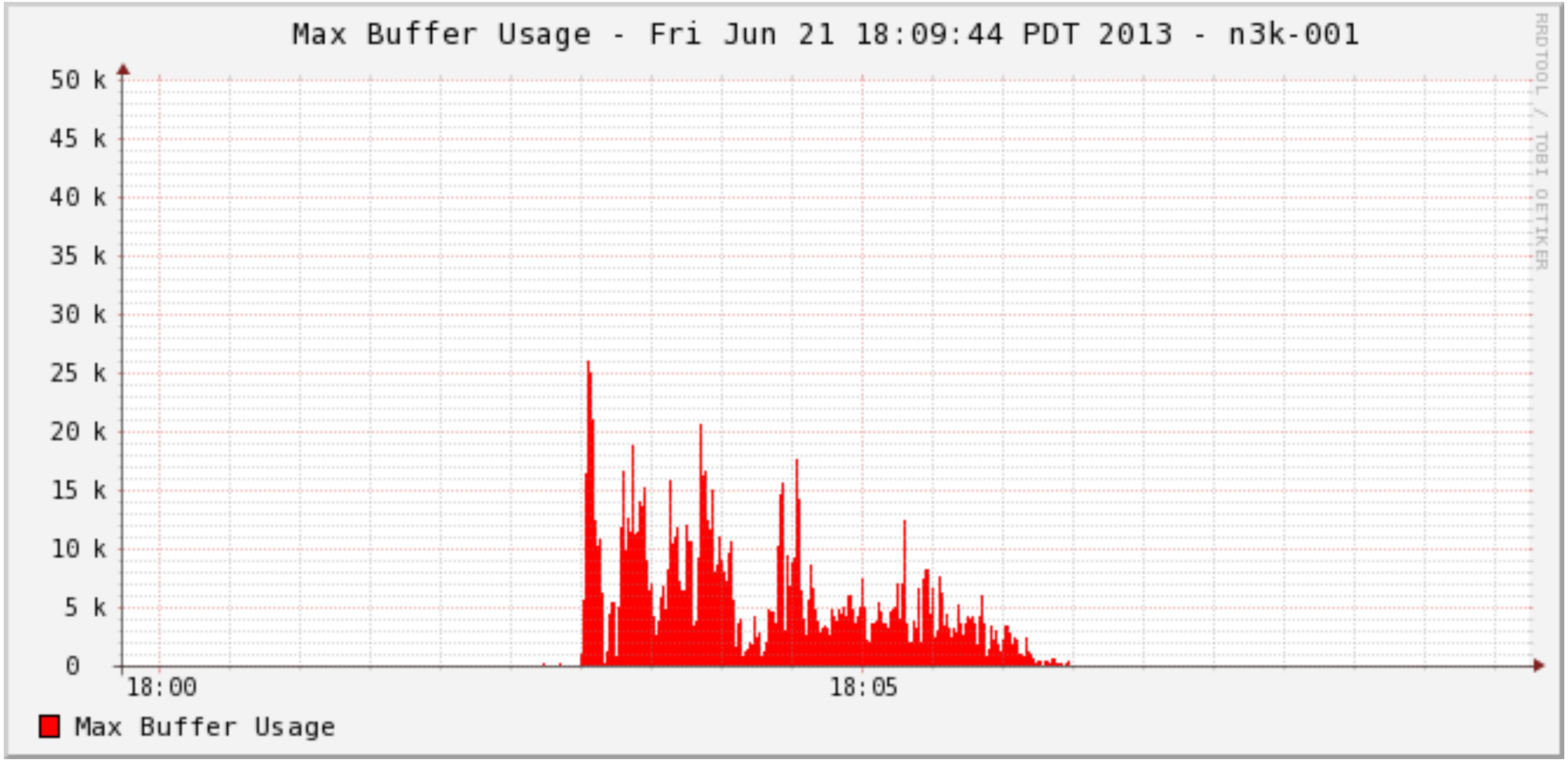
Hadoop + Network

Monitoring

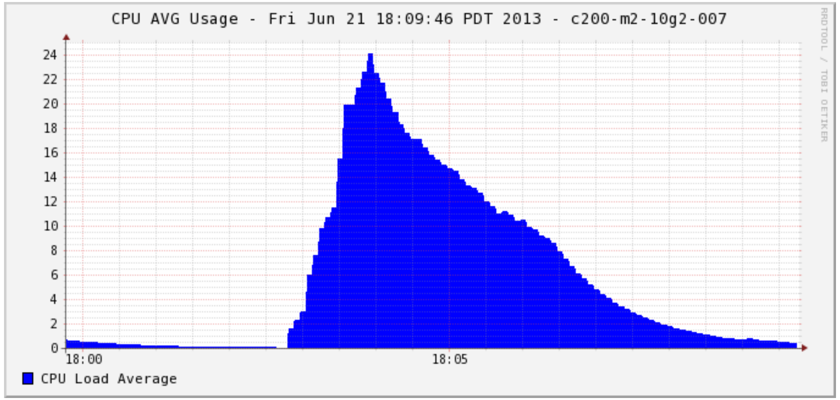
Server Resource Monitoring – CPU, Connections, etc.,

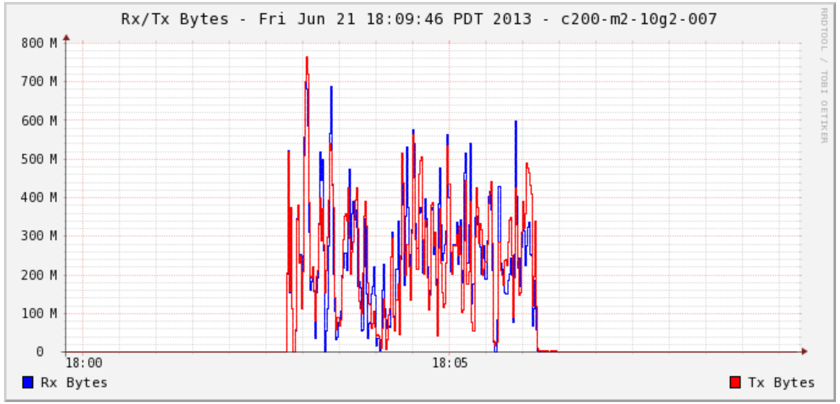


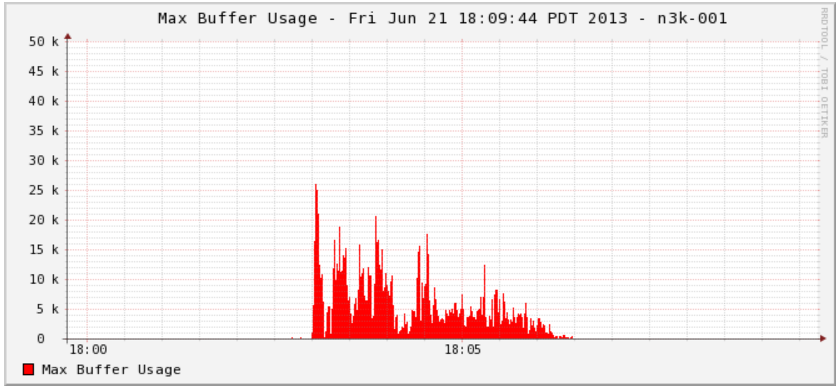
Network Resource Monitoring – Buffer Counters etc.,

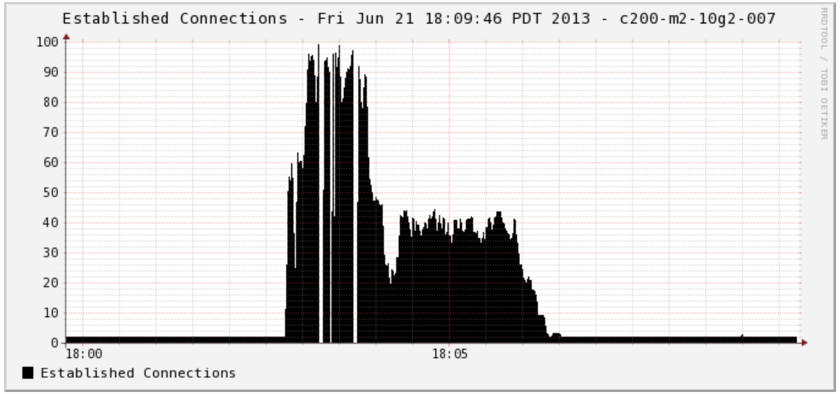


Server + Network



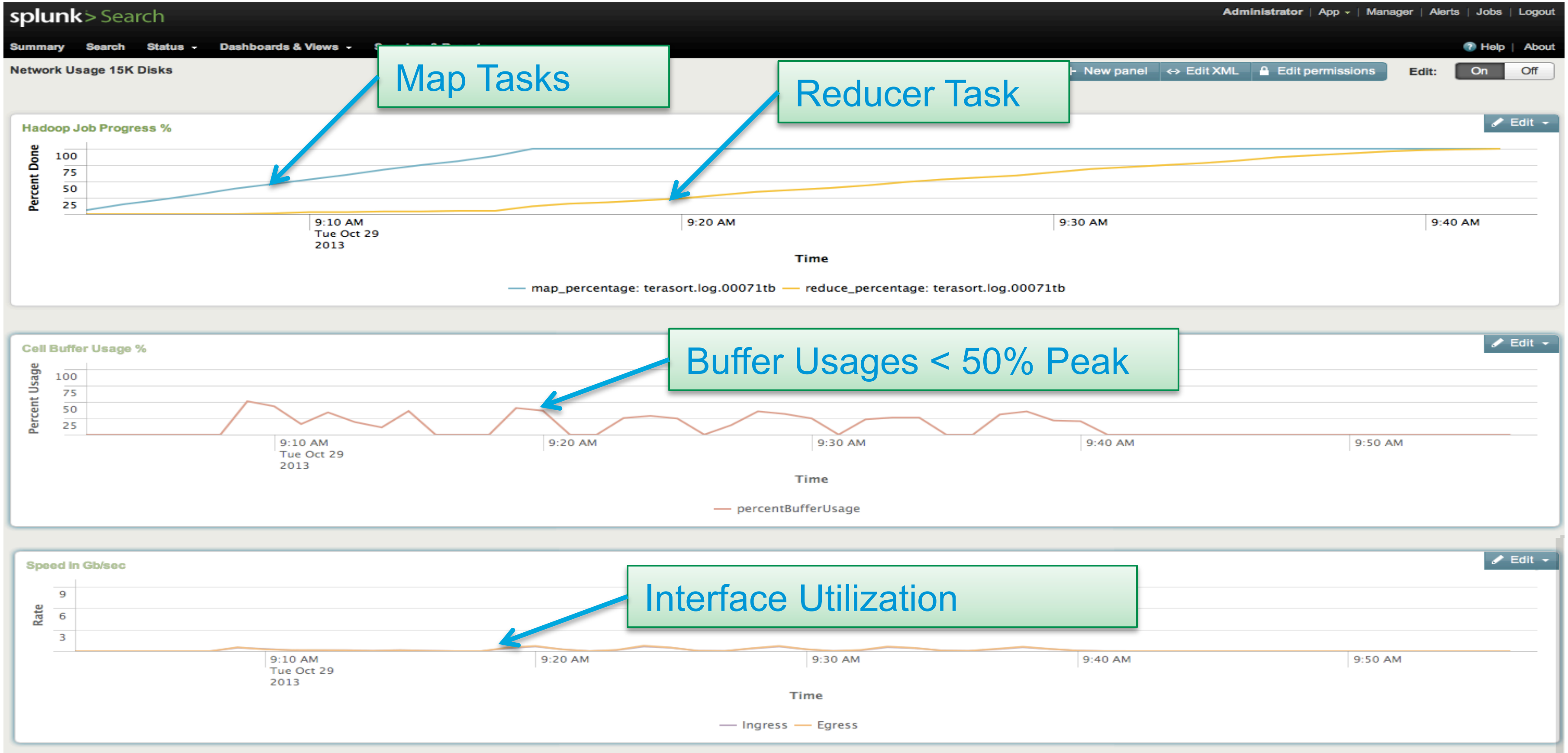






Integration with Splunk

Hadoop Process/Network Correlation



Map Tasks

Reducer Task

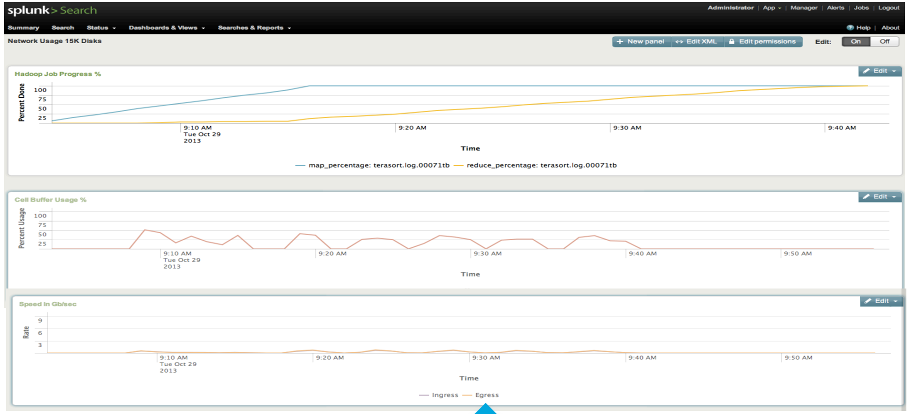
Buffer Usages < 50% Peak

Interface Utilization

Splunk Building Blocks

How do we go from raw data to graphs

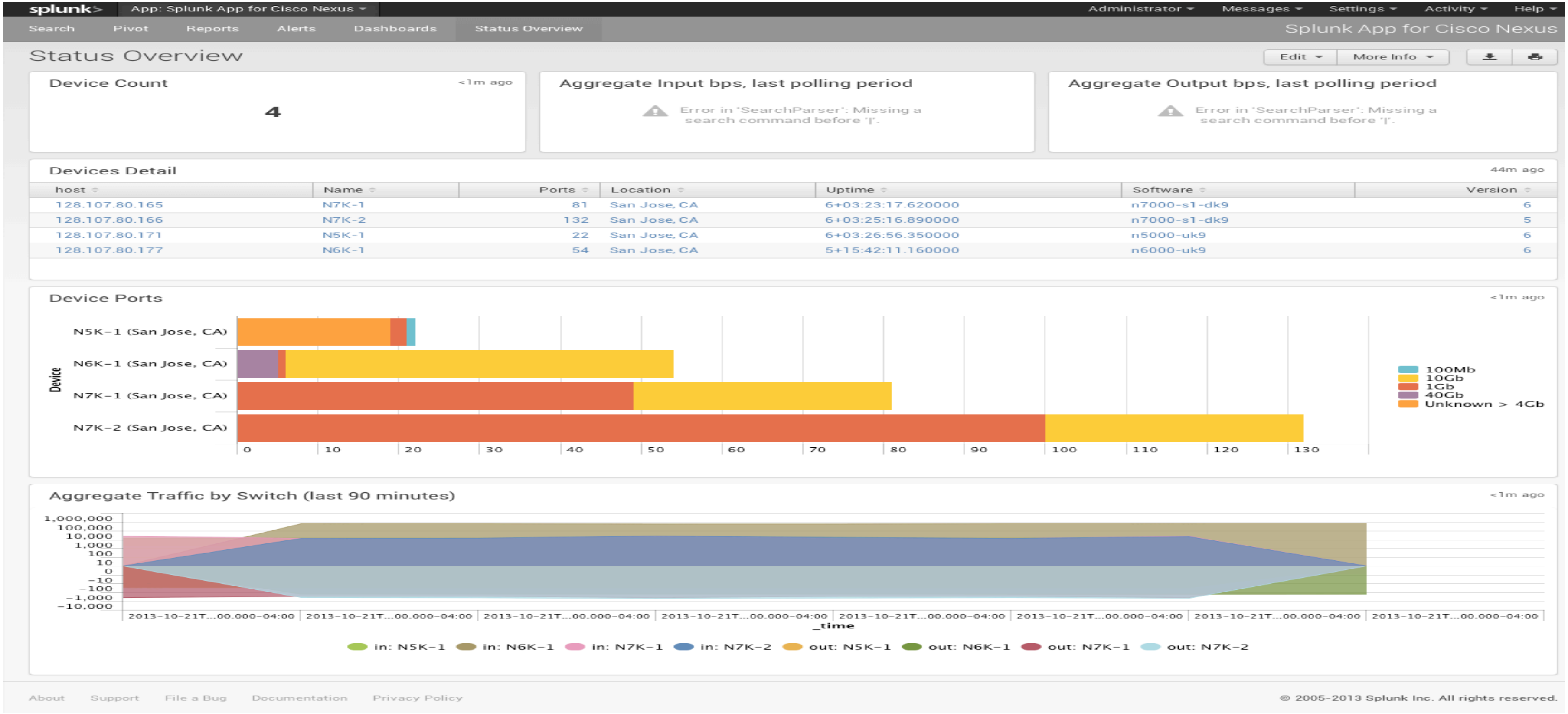
Python/TCP Dump/SNMP/Syslogs/Hadoop logs ….



splunk>

```
sourcetype="hadoop_jobs" host=c240-m3-017 map reduce |
rex field=source "(?<job>[^/]+)$" |
timechart first(map_perc) as map_percentage
first(reduce_perc) as reduce_percentage by job
```

Nexus Native Apps with Splunk



Status Overview

Device Count

4

Aggregate Input bps, last polling period

Error in 'SearchParser': Missing a search command before '|'.

Aggregate Output bps, last polling period

Error in 'SearchParser': Missing a search command before '|'.

Devices Detail

host	Name	Ports	Location	Uptime	Software	Version
128.107.80.165	N7K-1	81	San Jose, CA	6+03:23:17.620000	n7000-s1-dk9	6
128.107.80.166	N7K-2	132	San Jose, CA	6+03:25:16.890000	n7000-s1-dk9	5
128.107.80.171	N5K-1	22	San Jose, CA	6+03:26:56.350000	n5000-uk9	6
128.107.80.177	N6K-1	54	San Jose, CA	5+15:42:11.160000	n6000-uk9	6

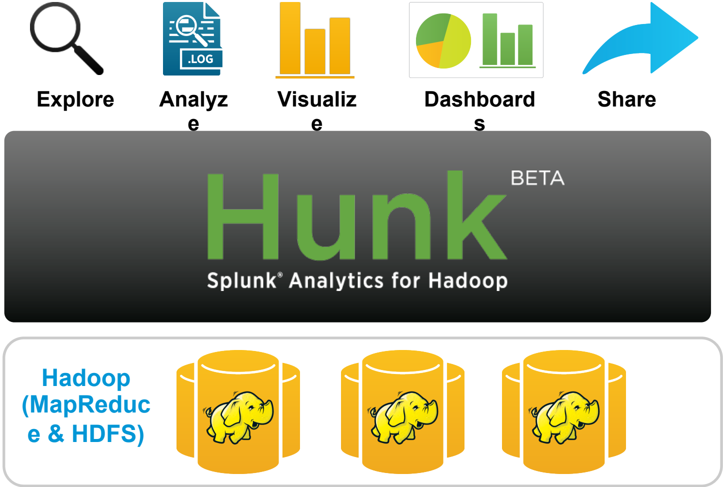
Device Ports

Aggregate Traffic by Switch (last 90 minutes)

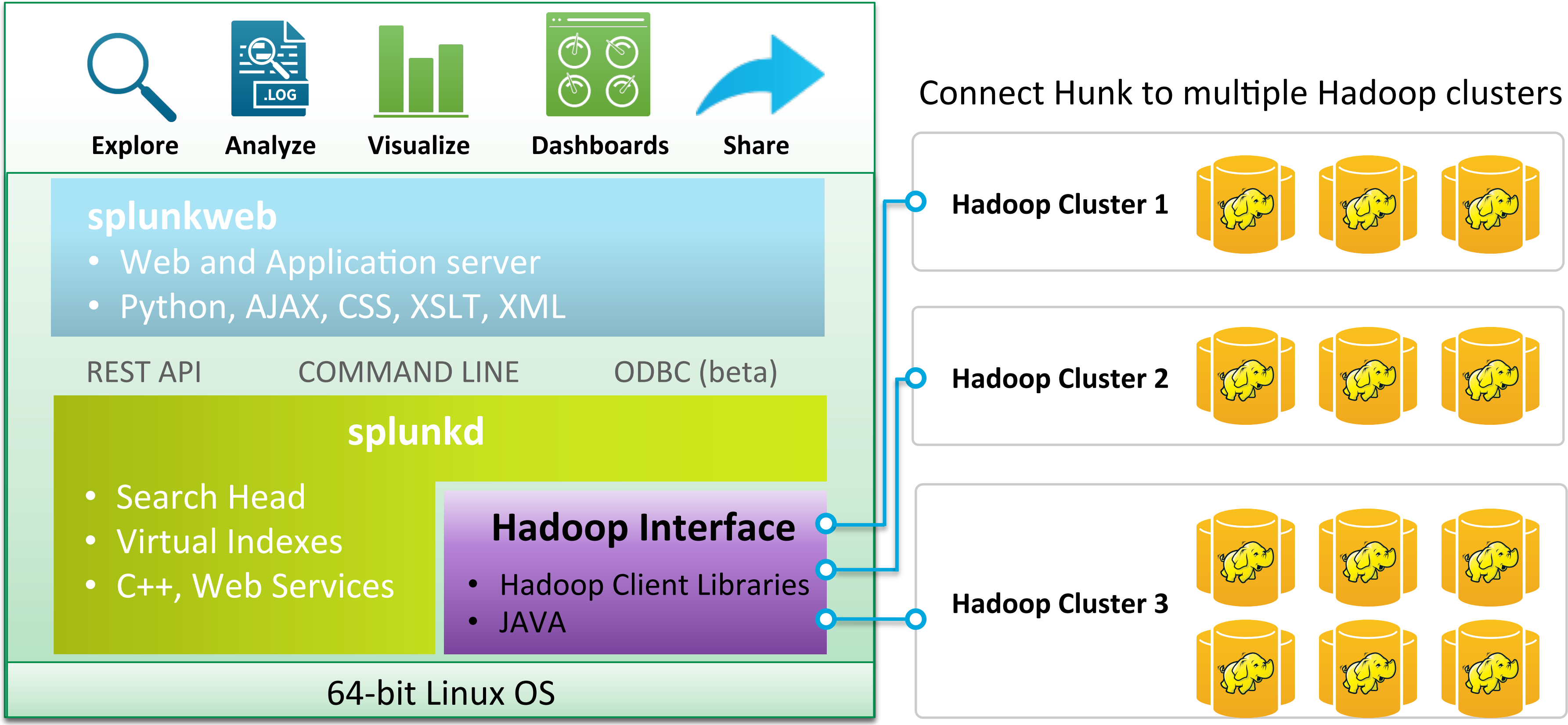
Hadoop for Hadoop

Using Hadoop to profile/optimize/analyze Hadoop applications

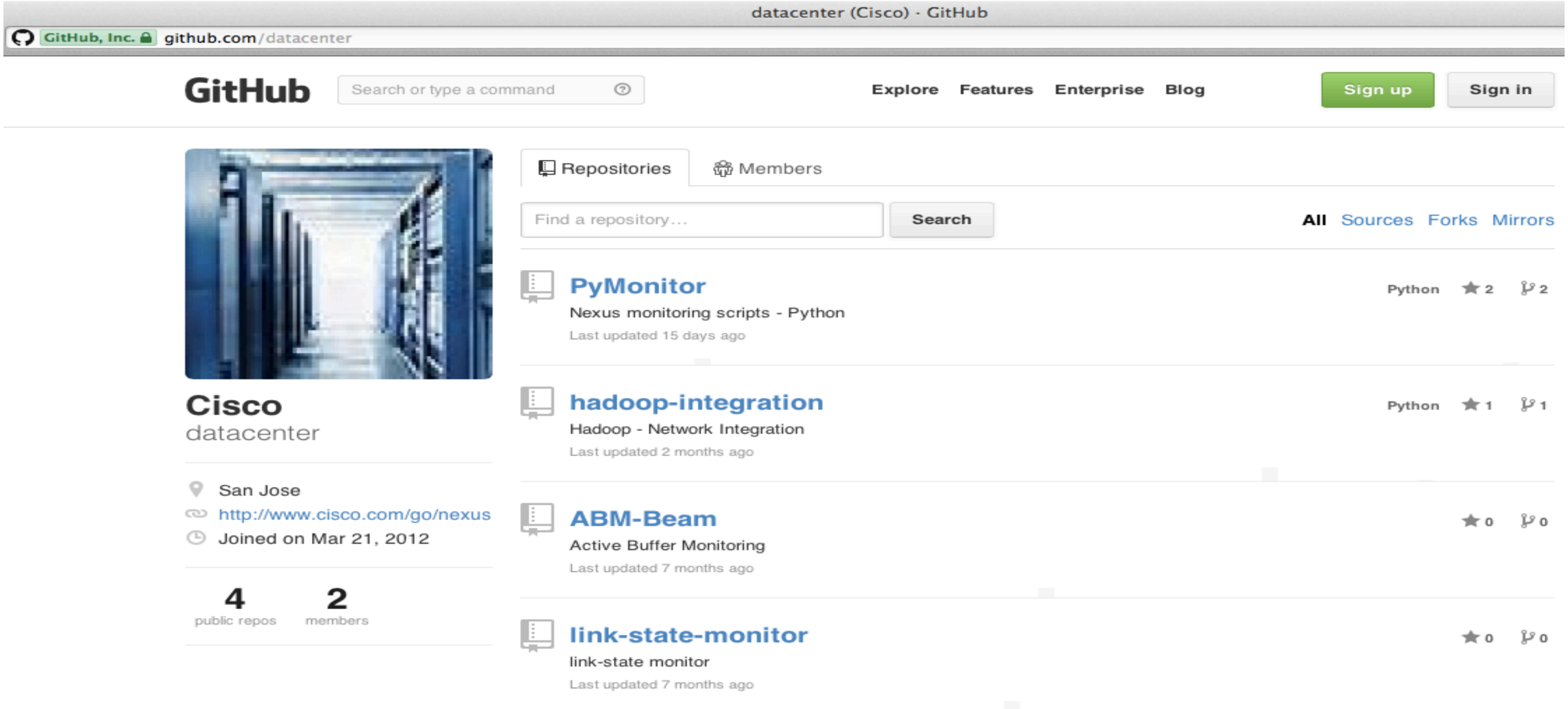
- Enabling the same visual analytics for hadoop data just like any other business analytics function
- Hunk allows multi-sourced data search and uses the Map/Reduce natively
- Any popular ingest method that is available can be used – Flume, Scribes, Chukwa etc
- Name node, data node and network traffic activities can be imported using standard methods – TCPDUMP, SNMP, Python Poller
- Insight into multi-cluster operation & multi-workload tuning



Hunk Scales with your Hadoop Deployments



Cisco Nexus on
www.github.com/datacenter



datacenter (Cisco) · GitHub

GitHub, Inc. github.com/datacenter

Cisco
datacenter

San Jose
http://www.cisco.com/go/nexus
Joined on Mar 21, 2012

4 public repos
2 members

PyMonitor — Python ★ 2, forks 2
Nexus monitoring scripts - Python
Last updated 15 days ago

hadoop-integration — Python ★ 1, forks 1
Hadoop - Network Integration
Last updated 2 months ago

ABM-Beam — ★ 0, forks 0
Active Buffer Monitoring
Last updated 7 months ago

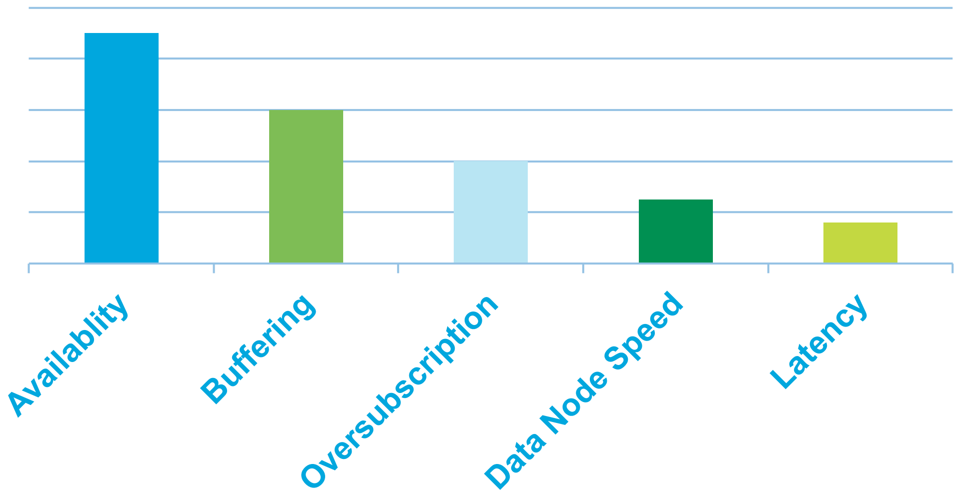
link-state-monitor — ★ 0, forks 0
link-state monitor
Last updated 7 months ago

Recommendations

www.cisco.com/go/bigdata

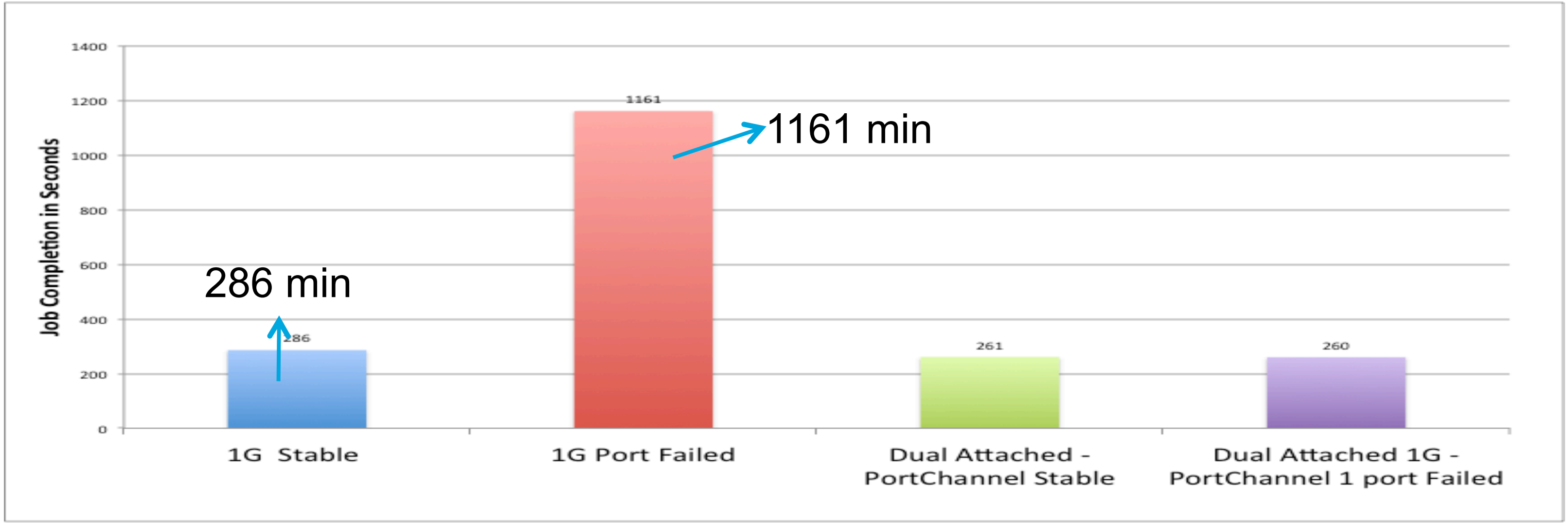
Integration Considerations

- Network Attributes
- Architecture
- Availability
- Capacity, Scale & Oversubscription
- Flexibility
- Management & Visibility



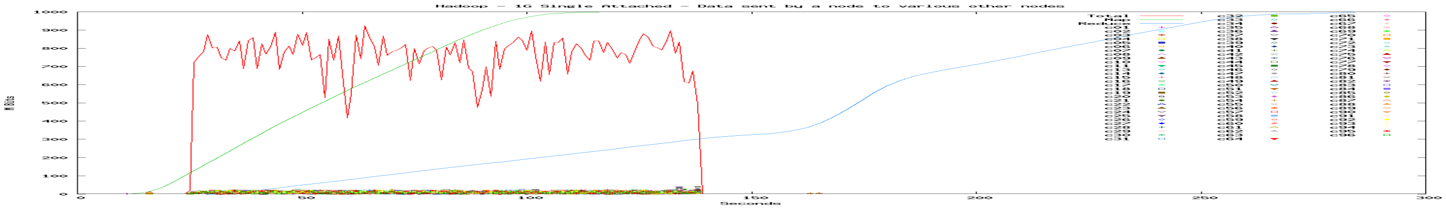
Availability

- Single NIC failure doubles the job completion time.
- Dual NIC has no impact on job completion time
- Effective load-sharing of traffic flow on two NICs. NIC bonding configured at Linux – with LACP mode of bonding
- Recommended to change the hashing to src-dst-ip-port (both network and NIC bonding in Linux) for optimal load-sharing

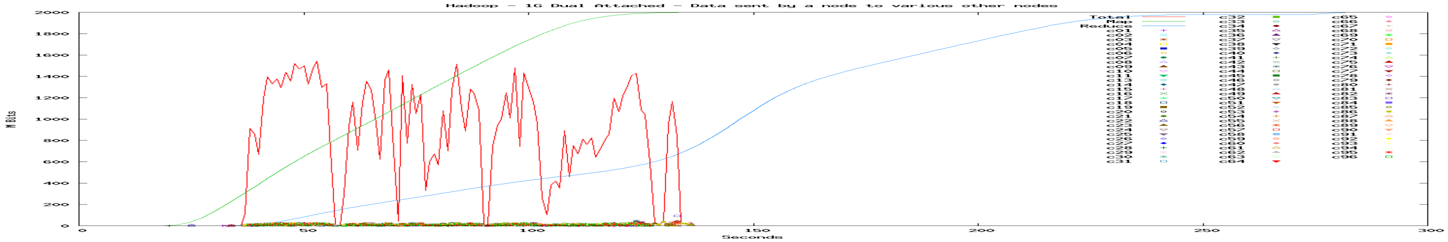


Data Node Speed Differences

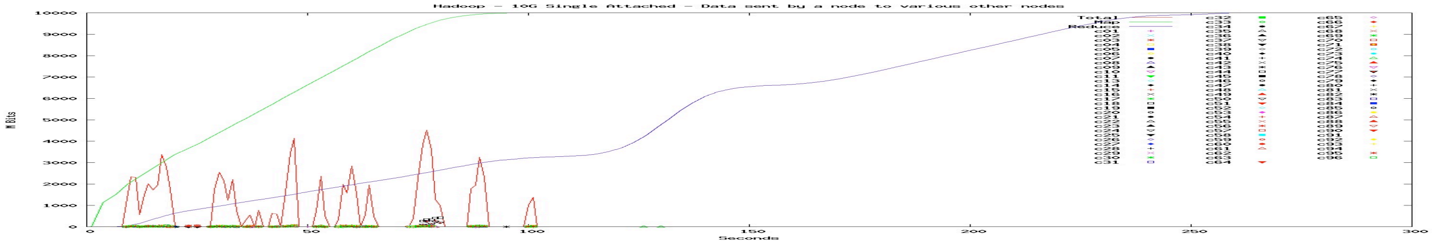
Generally 1G is being used largely due to the cost/performance trade-offs. Though 10GE can provide benefits depending on workload



Single 1GE
100% Utilized

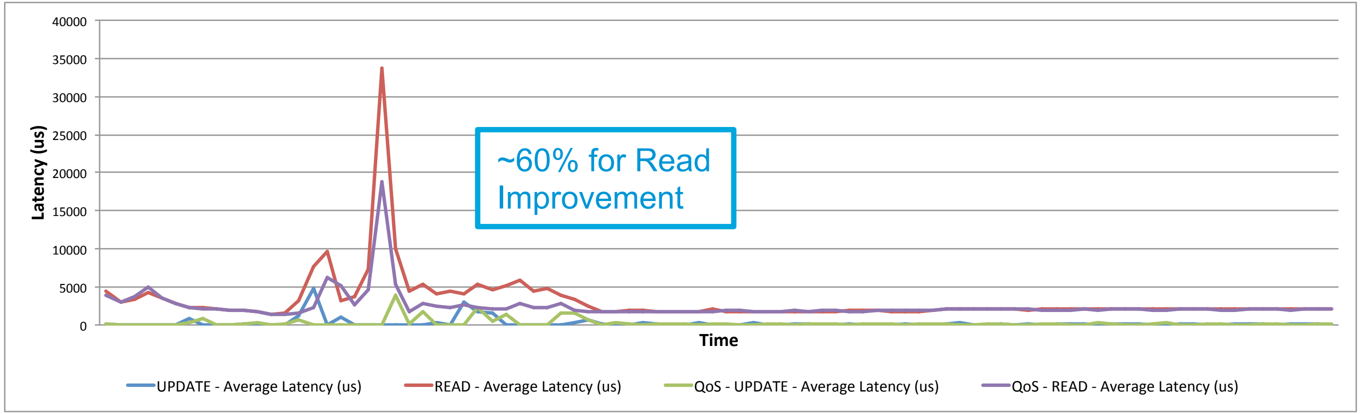


Dual 1GE
75% Utilized



10GE
40% Utilized

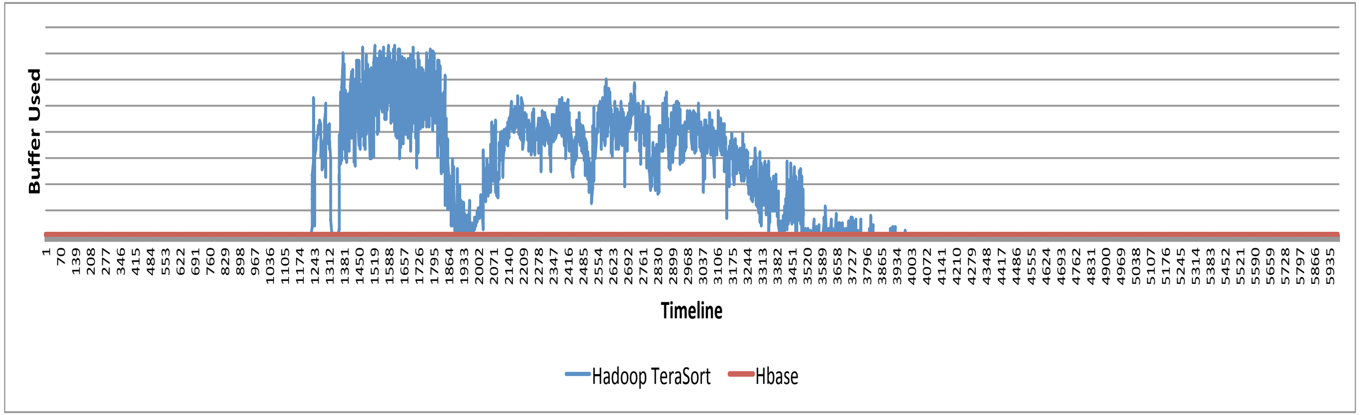
Hbase + Hadoop Map Reduce



~60% for Read Improvement

Read/Update Latency

Comparison of Non-QoS vs. QoS Policy

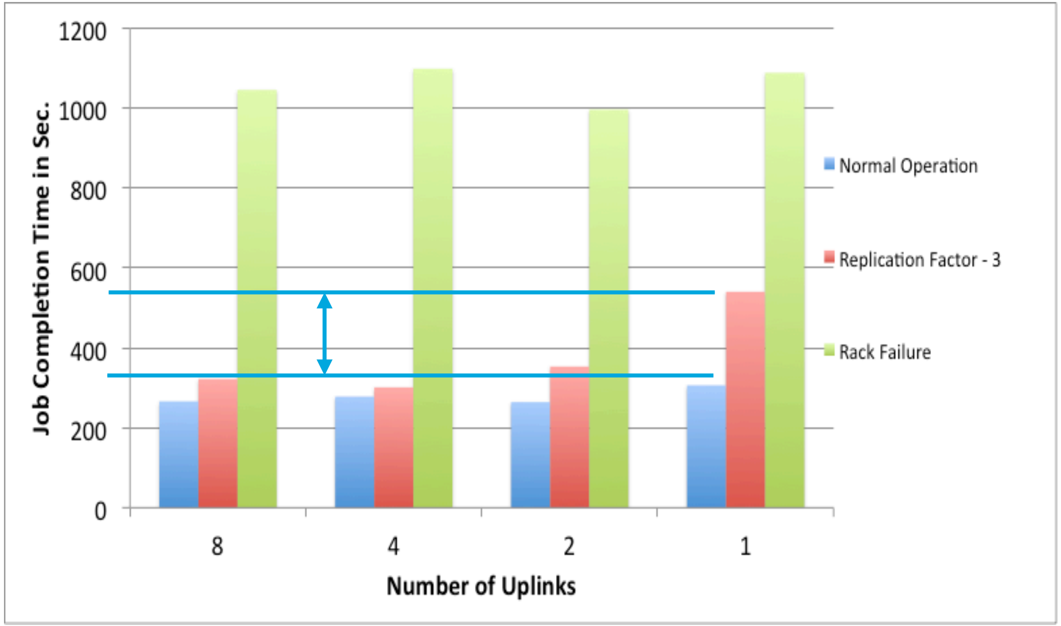


Switch Buffer Usage

With Network QoS Policy to prioritize Hbase Update/Read Operations

Non-Blocking Network vs Non-Blocking Design

- There is no such thing as non-blocking design
  - Even thought network is designed with no oversubscription – the in-casting, limitation on IO and compute negates the cost
  - Higher the spindle count, higher network traffic – but not always linear
  - Eventually the application itself reach to a limitation of concurrency, threads etc.
- Failure impact in the context of job completion time
- Normal Job Run – not much impact
- Result Replication with 1,2,4, & 8 10G uplink(s) - larger relative impact
- Rack failure is immune to oversubscriptions – IOW the rack failure impact hides the oversubscription loss



Big Data @ Cisco - www.cisco.com/go/bigdata

Multi-year network and compute analysis testing
(In conjunction with partners)

Hadoop World 2011 on Hadoop Network and Compute Considerations:
http://bit.ly/18s6h8y

Hadoop Summit 2012 on Network Reference Architectures (Best practices).
Slides: http://slidesha.re/1aNt3sJ Youtube: http://bit.ly/16ENk2y

Hadoop World 2012 Designing Hadoop for the Enterprise Data Center
http://bit.ly/1gTkCow

Hadoop Summit 2013
http://bit.ly/1aiqu7j

Certifications and Solutions with UCS C-Series and Nexus Series Switches

Cloudera Hadoop Certified Technology

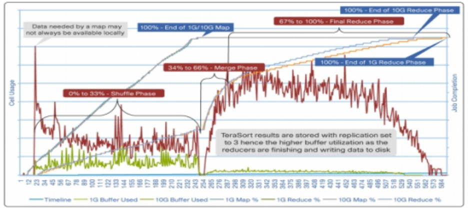
Oracle NoSQL Validated Solution

O'REILLY® Strata CONFERENCE + HADOOP WORLD

Tools and Techniques That Make Data Work

2012 HADOOP SUMMIT

Visibility & Monitoring



cloudera CERTIFIED TECHNOLOGY