
TOMORROW
starts here.
CISCO
Cisco live!

Big Data Architecture and Deployment

BRKAPP-2033

Sean McKeown - Technical Solution Architect

#clmel

Cisco live!

Agenda

- Big Data Concepts and Overview
  - Enterprise data management and big data
  - Infrastructure attributes
  - Hadoop, NOSQL and MPP Architecture concepts
- Hadoop and the Network
  - Network behaviour, FAQs
- Cisco UCS for Big Data
  - Building a big data cluster with the UCS Common Platform Architecture (CPA)
  - UCS Networking, Management, and Scaling for big data
- Q & A

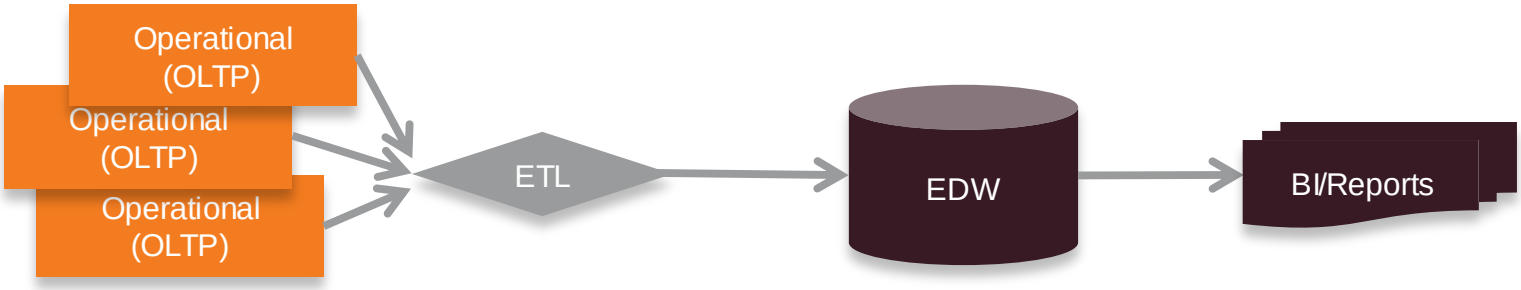
Big Data Concepts and Overview

Cisco live!

“More data usually beats better algorithms.”

-Anand Rajaraman, SVP @WalmartLabs

Traditional Enterprise Data Management



Online Transactional Processing

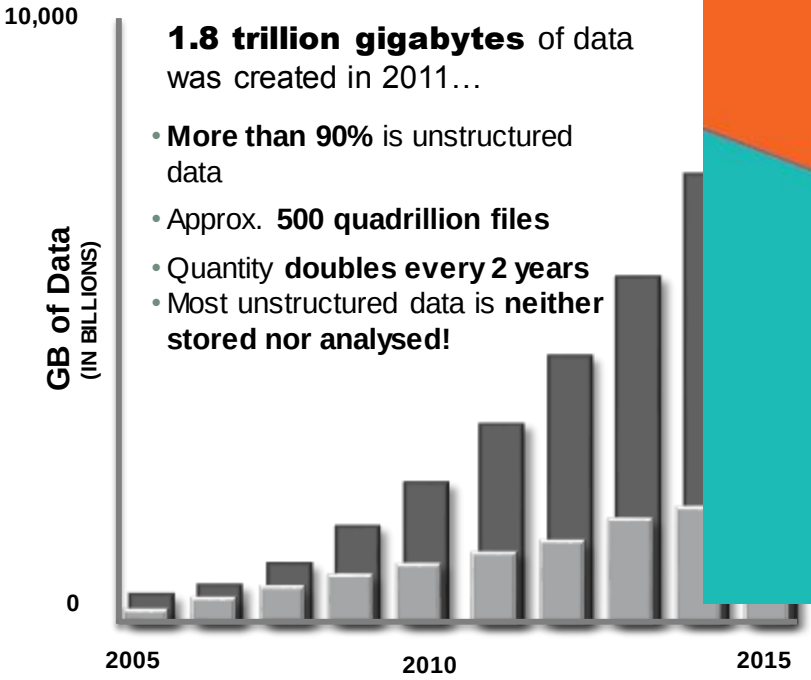
Extract, Transform, and Load (batch processing)

Enterprise Data Warehouse

Business Intelligence

So What Has Changed?

The Explosion of Unstructured Data



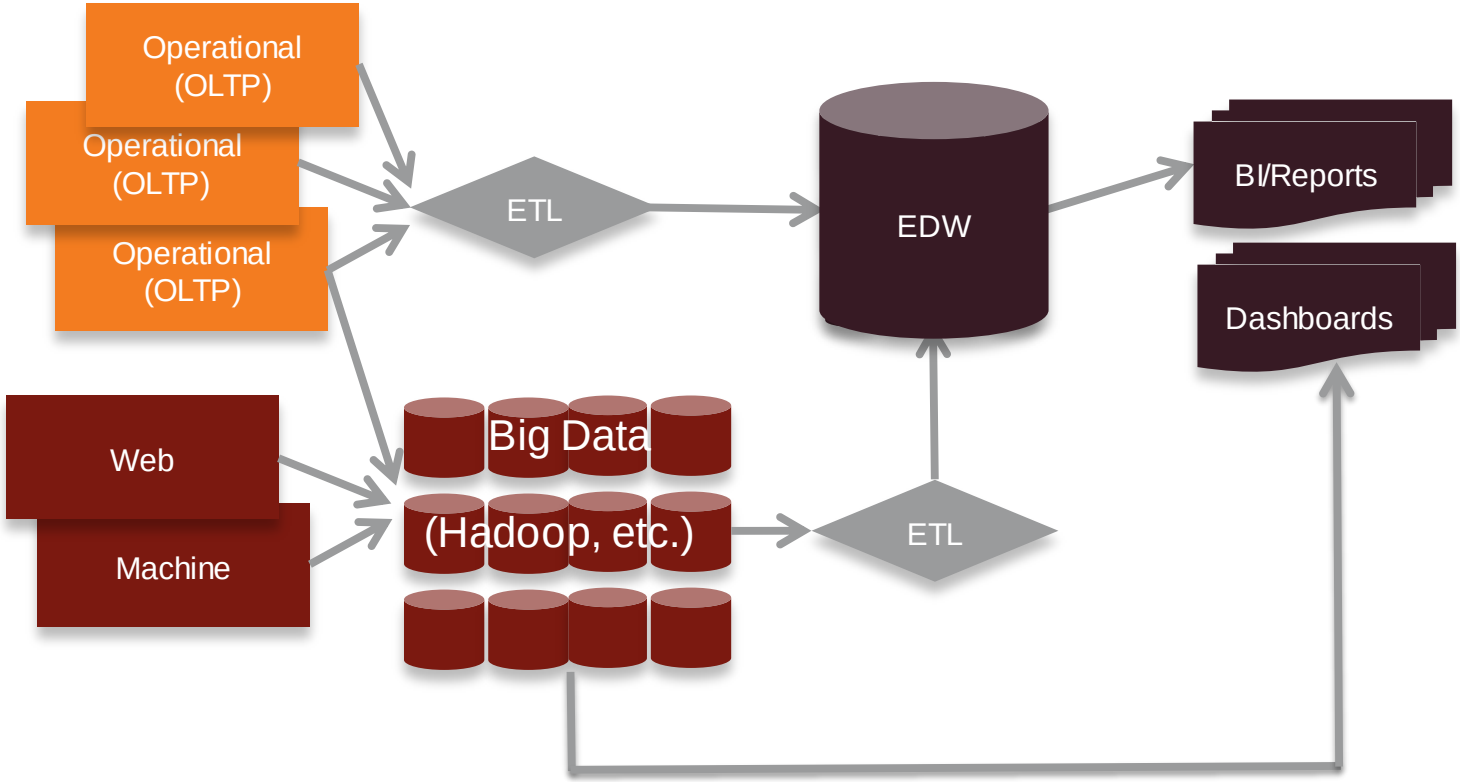
1.8 trillion gigabytes of data was created in 2011…

- **More than 90%** is unstructured data
- Approx. **500 quadrillion files**
- Quantity **doubles every 2 years**
- Most unstructured data is **neither stored nor analysed!**

Source: Cloudera

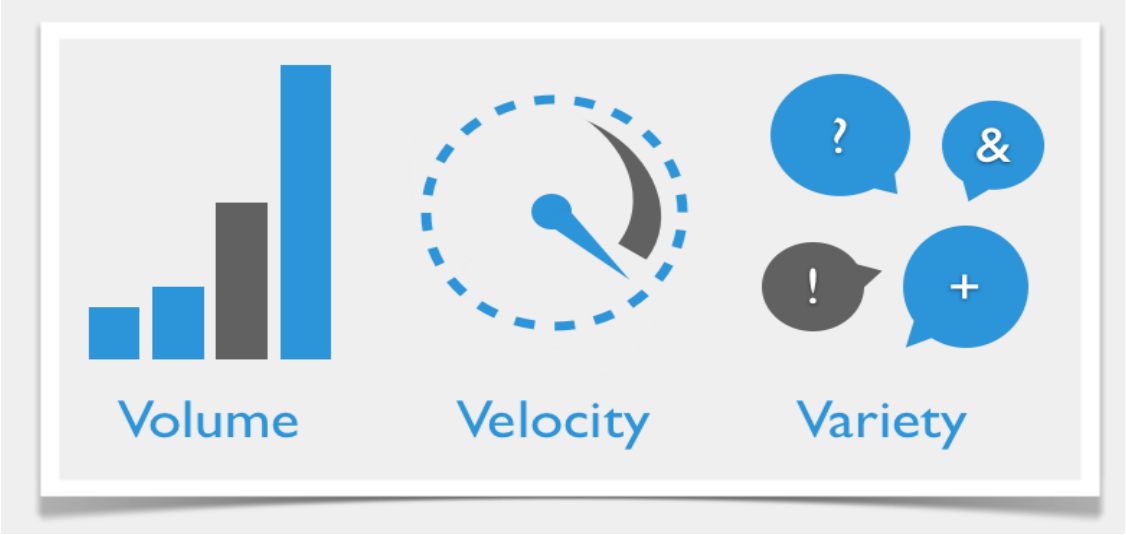
Cisco live!

Enterprise Data Management with Big Data



What is Big Data?

When the size of the data itself is part of the problem.

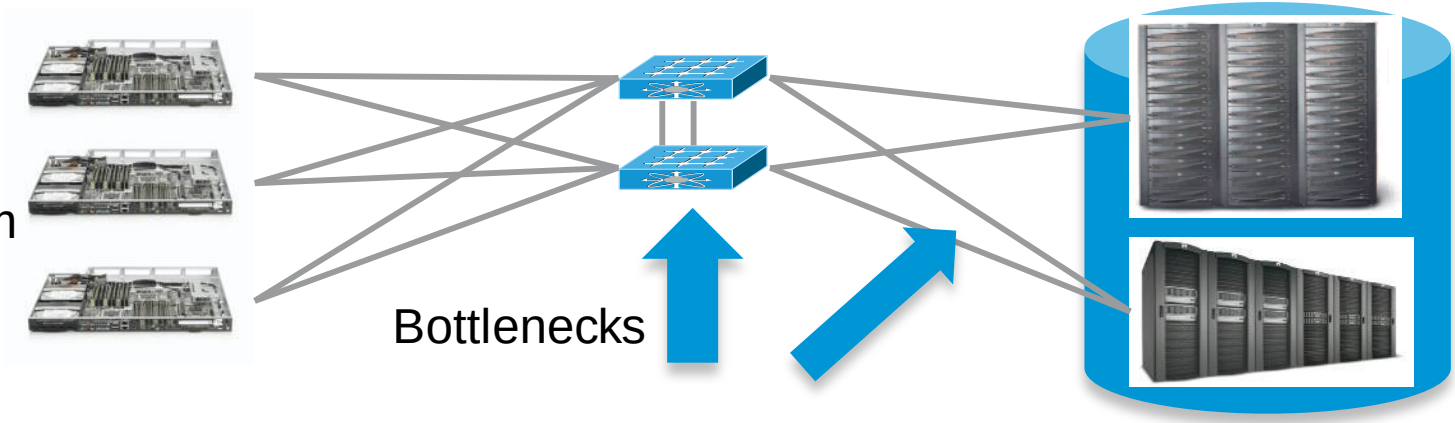


What isn't Big Data?

- Usually not blade servers
  (not enough local storage)
- Usually not virtualised
  (hypervisor only adds overhead)
- Usually not highly oversubscribed
  (significant east-west traffic)
- Usually not SAN/NAS

Classic NAS/SAN vs. New Scale-out DAS

Traditional – separate compute from storage



New – move the compute to the storage

$$$

Low-cost, DAS-based, scale-out clustered filesystem

Big Data Software Architectures and Design Considerations

Cisco live!

Three Common Big Data Architectures

DATASTAX

splunk>

mongoDB

NoSQL

Fast key-value store/retrieve in real time

actian

Pivotal Greenplum

MPP Relational Database

Scale-out BI/DW

HBASE

cloudera

PIVOTAL HD

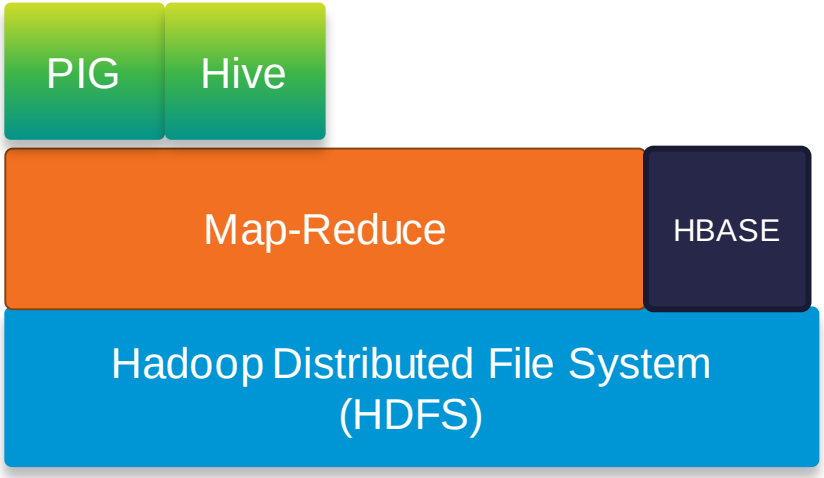
MAPR

Hadoop

Distributed batch, query, and processing platform

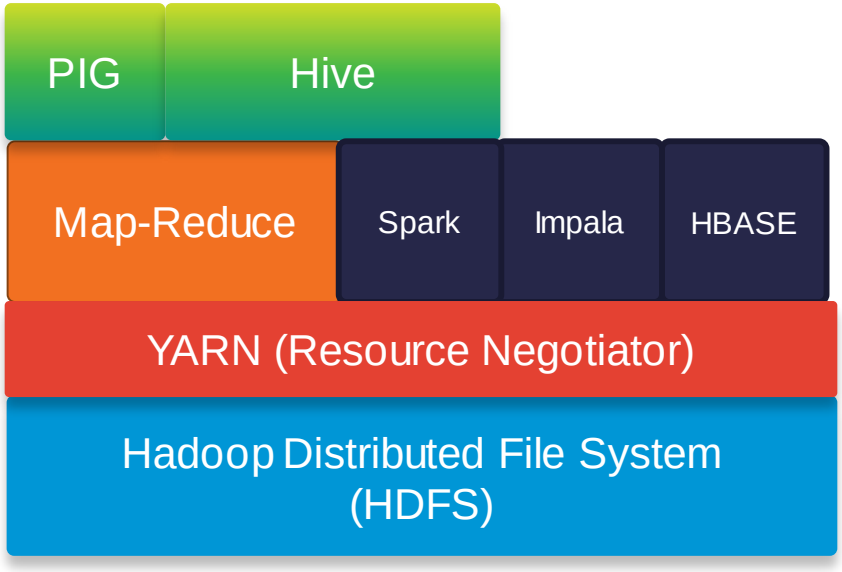
hortonworks

Hadoop: A Closer Look



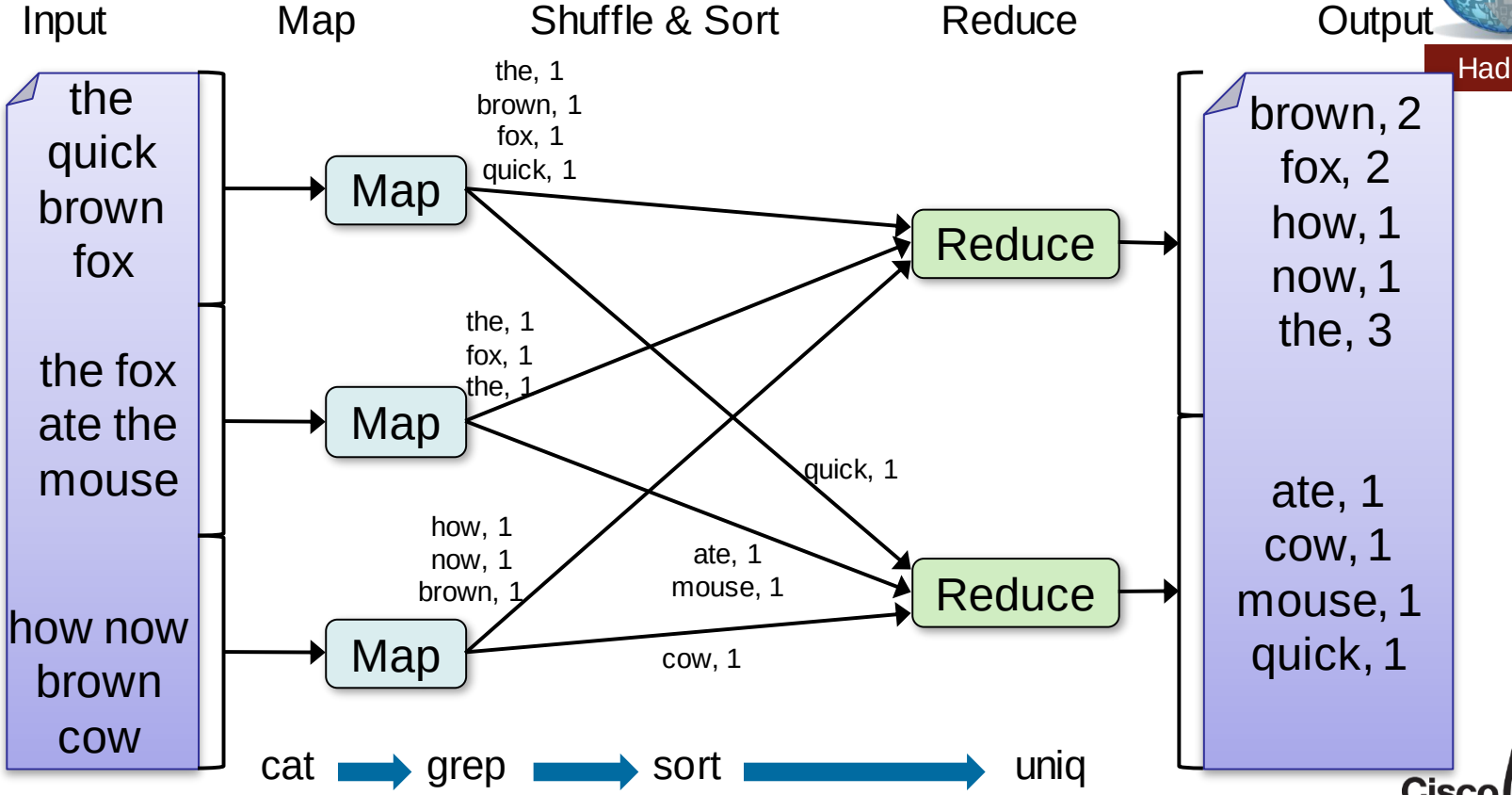
- Hadoop is a distributed, fault-tolerant framework for storing and analysing data
- Its two primary components are the Hadoop Filesystem (HDFS) and the MapReduce application engine

Hadoop: A Closer Look



- Hadoop 2.0 (with YARN) adds the ability to run additional distributed application engines concurrently on the same underlying filesystem

MapReduce Example: Word Count Frequency



Cisco live!

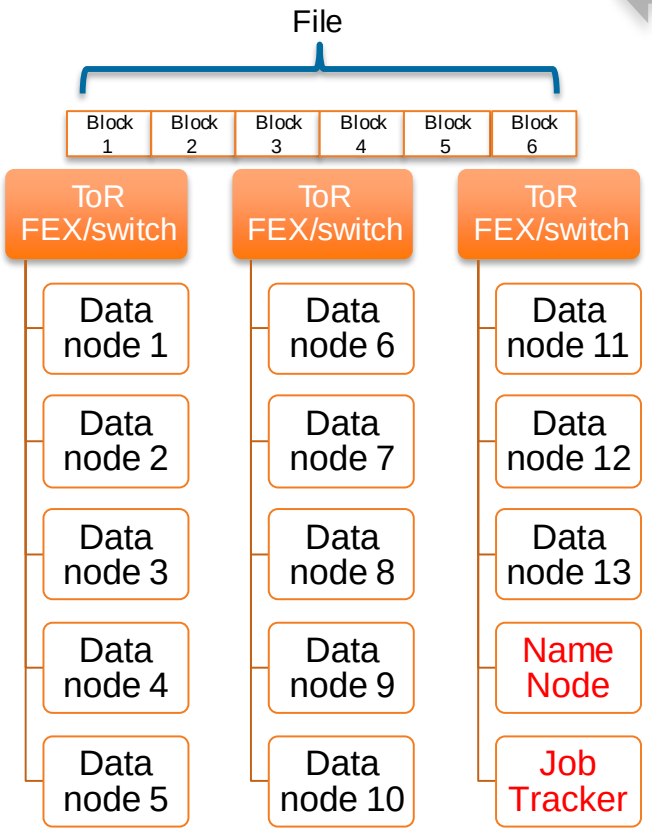
“Failure is the defining difference between distributed and local programming”

- Ken Arnold, CORBA designer

Hadoop Components and Operations

Hadoop Distributed File System

- Scalable & Fault Tolerant
- Filesystem is distributed, stored across all data nodes in the cluster
- Files are divided into multiple **large blocks** – 64MB default, typically 128MB – 512MB
- Data **is stored reliably.** Each block is replicated 3 times by default
- Types of Nodes
  - Name Node - Manages HDFS
  - Job Tracker – Manages MapReduce Jobs
  - Data Node/Task Tracker – stores blocks/does work



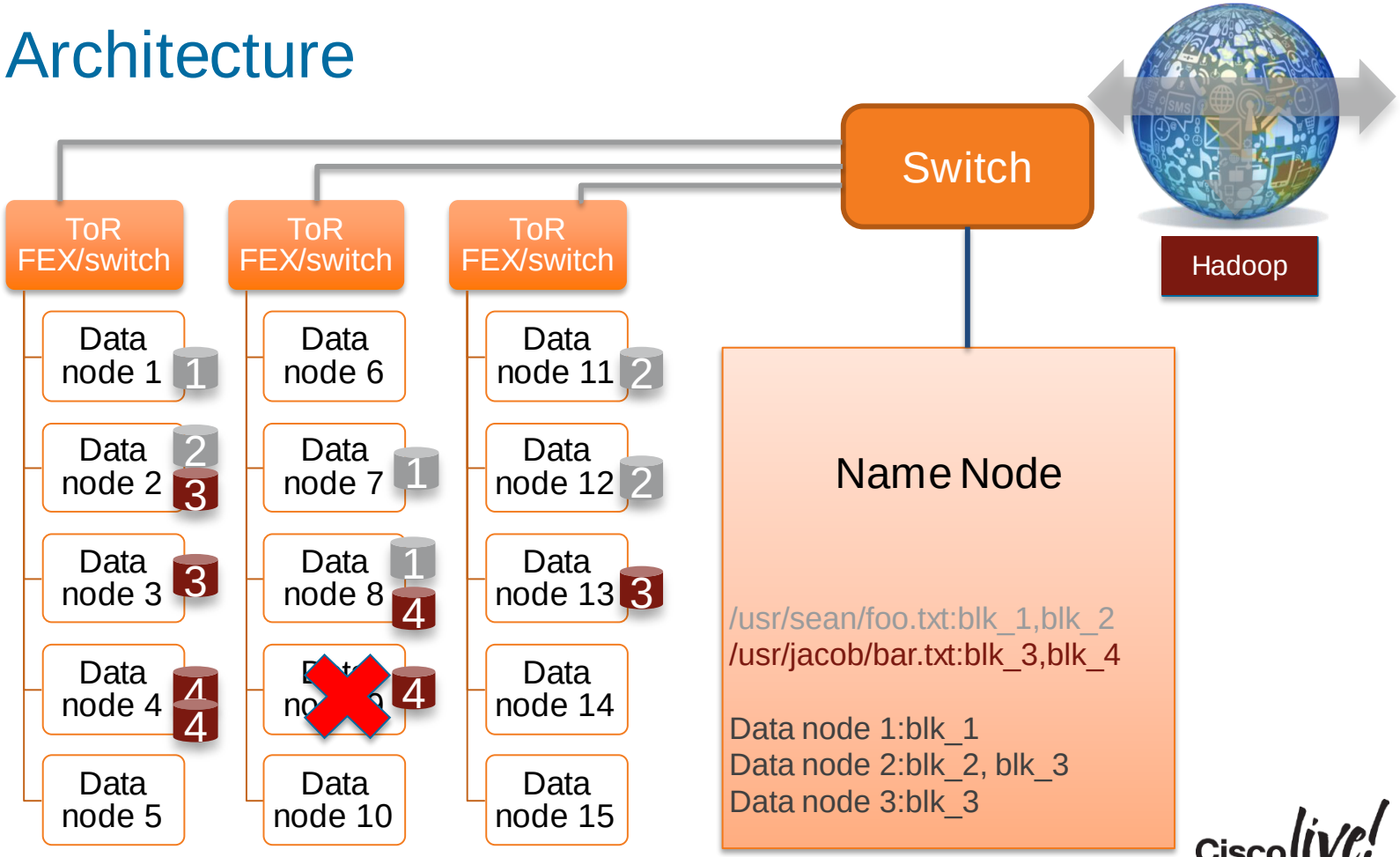
Why Replicate?

Two key reasons

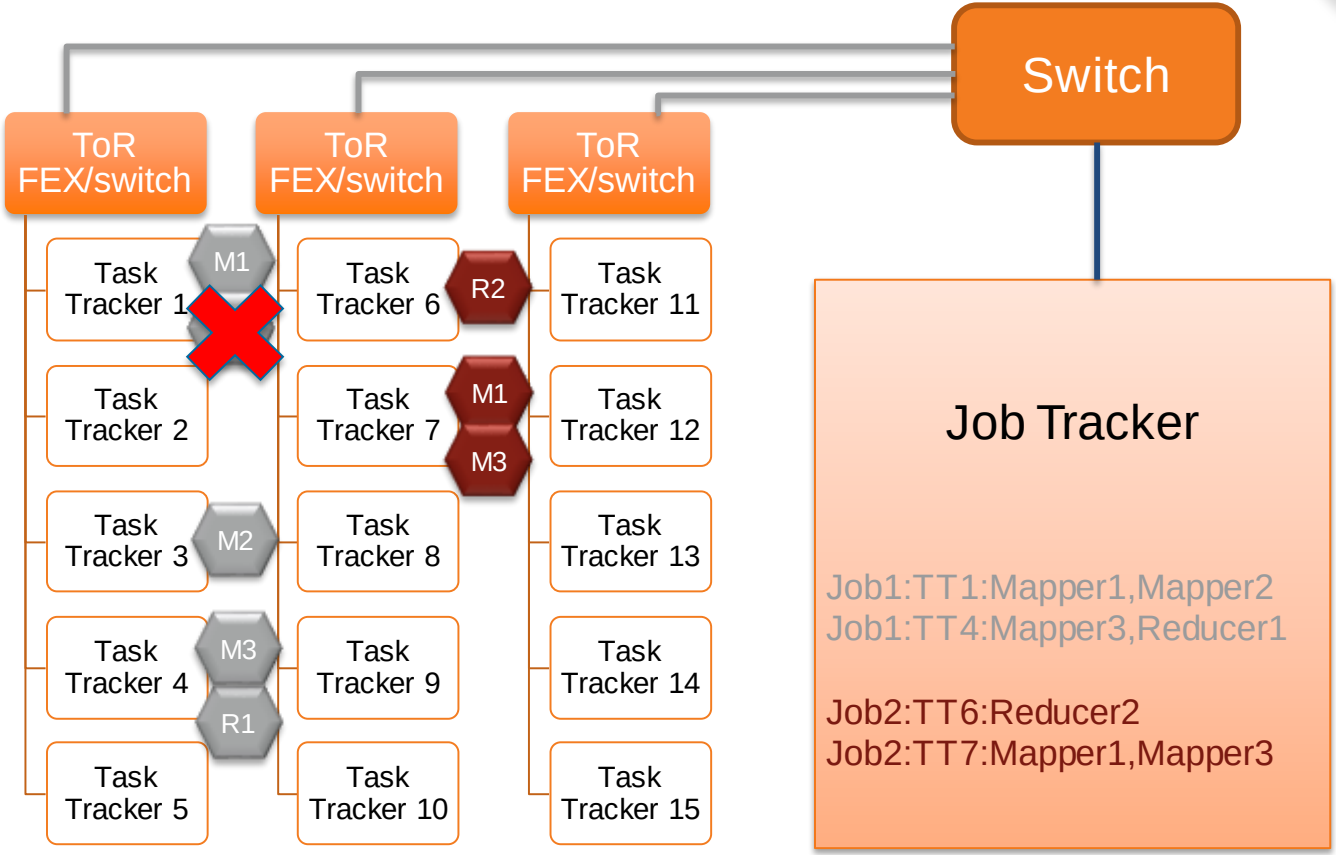
1. Fault tolerance

2. *Increase data locality*

HDFS Architecture



MapReduce (Hadoop 1.0) Architecture



Design Considerations for NoSQL Databases

Design Considerations

- Scale-out with Shared-Nothing
- Data Redundancy Options
  - Key-Value: JBOD + 3-Way Replication
  - Document-Store: RAID or Replication

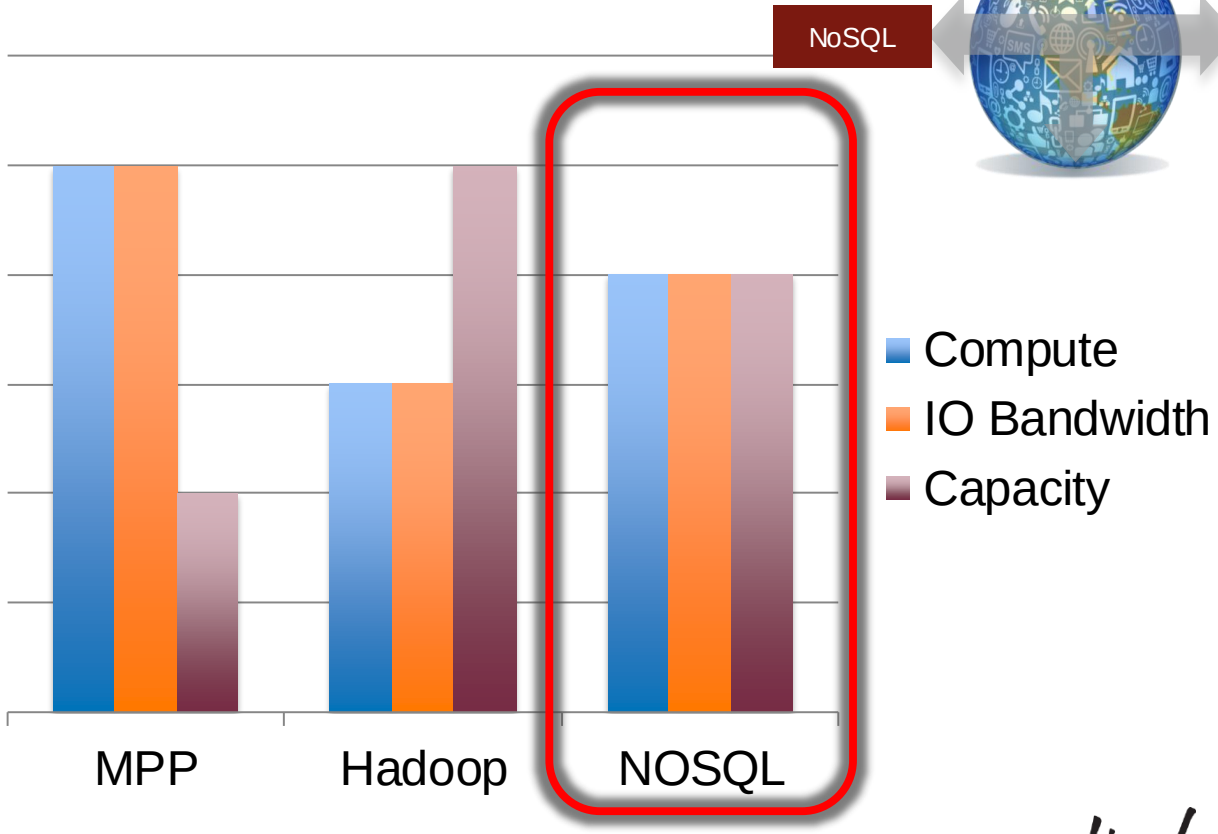
Configuration Considerations

(1) Moderate Compute

(2) Balanced IOPS (Performance vs. Cost)

10K RPM HDD

(3) Moderate to High Capacity



Design Considerations for MPP Database

Design Considerations

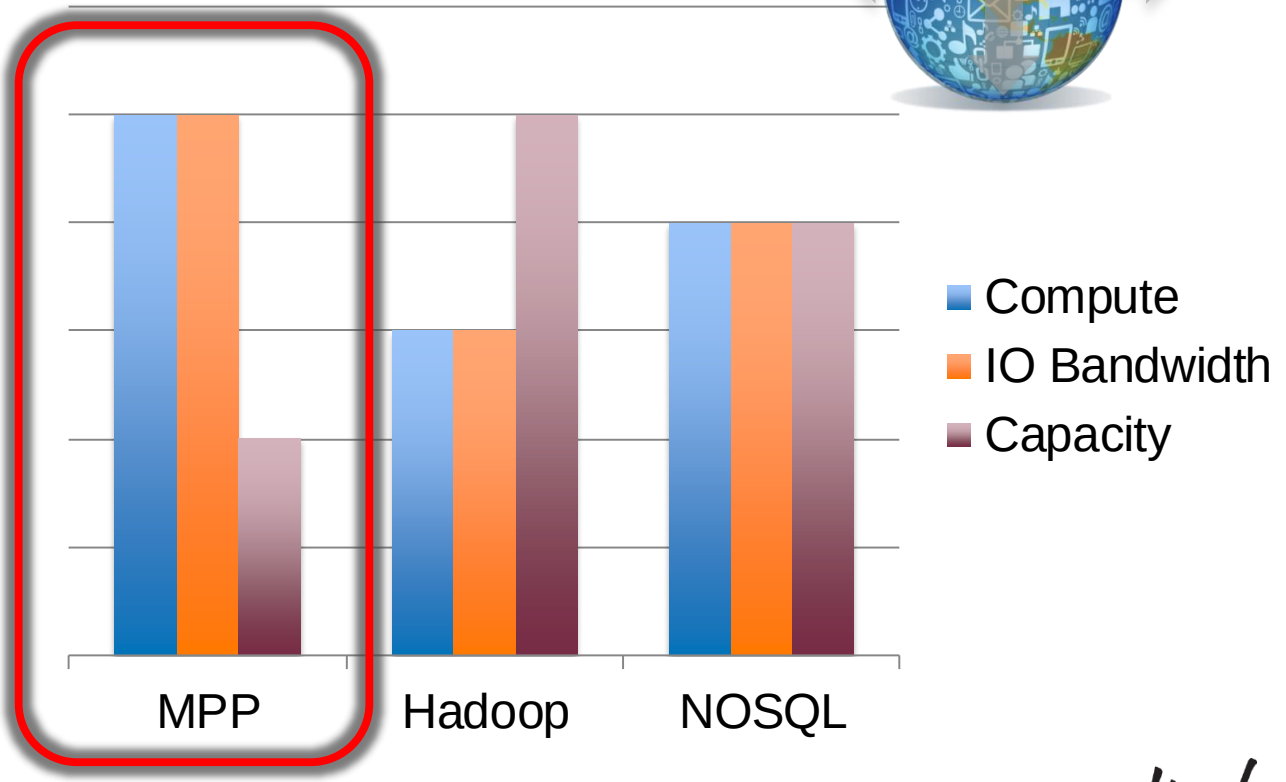
- Scale-out with Shared nothing
- Data Redundancy with Local RAID 5

Configuration Considerations

(1) High Compute (Fastest CPU)

(2) High IO Bandwidth
Flash/SSD and In-memory

(3) Moderate Capacity



Greenplum

Actian

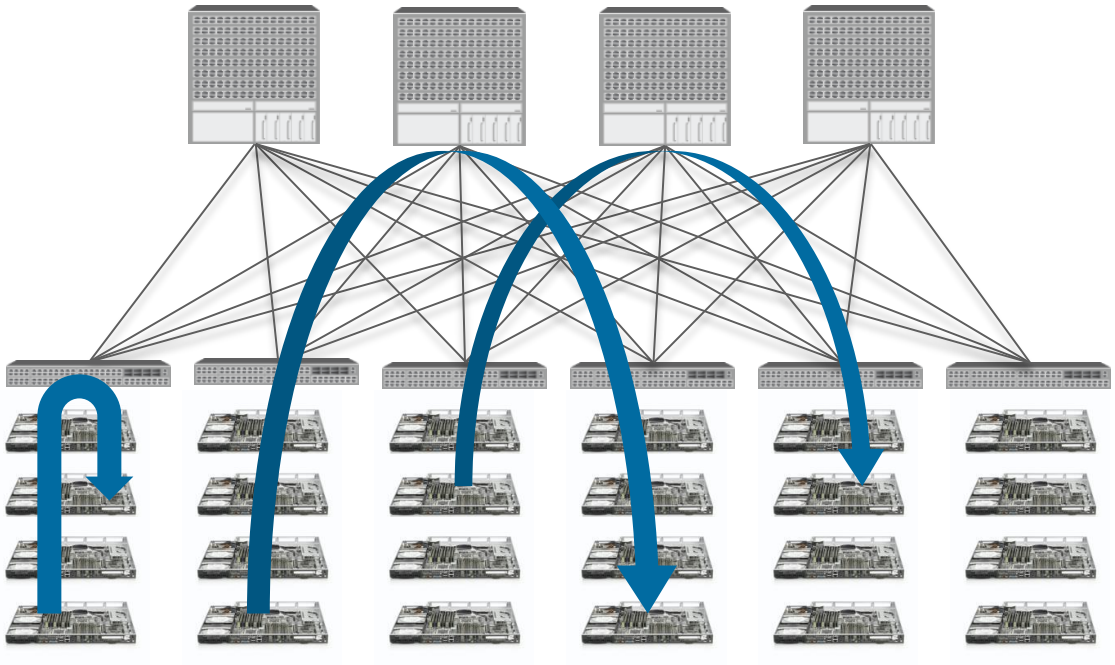
Cisco live!

Hadoop and The Network

Cisco live!

Hadoop Network Design

- The network is the fabric – the ‘bus’ - of the ‘supercomputer’
- Big data clusters often create **high east-west, any-to-any** traffic flows compared to traditional DC networks
- Hadoop networks are typically isolated/dedicated; simple leaf-spine designs are ideal
- 10GE typical from server to ToR, low oversubscription from ToR to spine
- With Hadoop 2.0, clusters will likely have heterogeneous, multi-workload behaviour



Hadoop Network Traffic Types

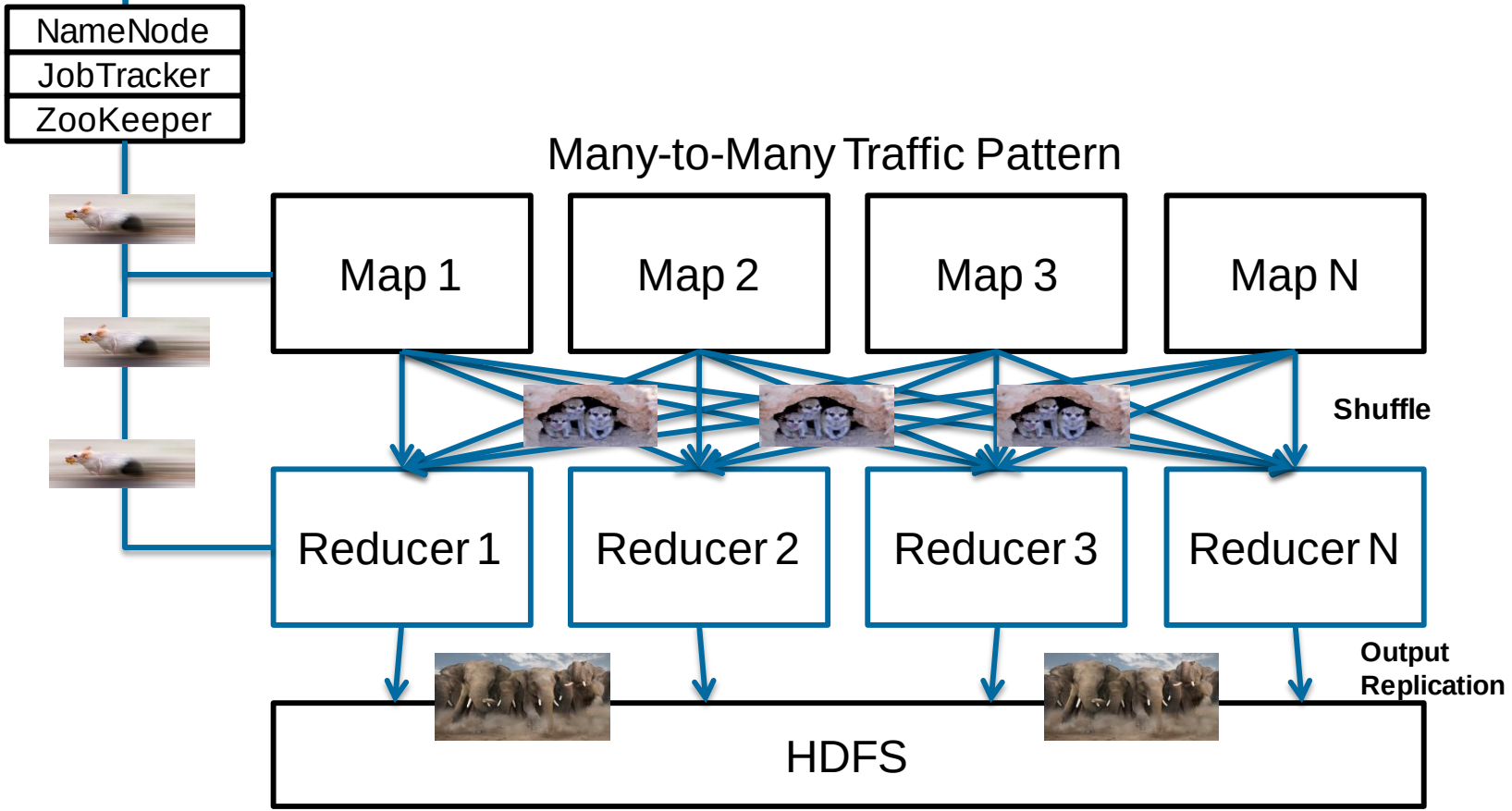
Small Flows/Messaging
(Admin Related, Heart-beats, Keep-alive, delay sensitive application messaging)

Small – Medium Incast
(Hadoop Shuffle)

Large Flows
(HDFS egress)

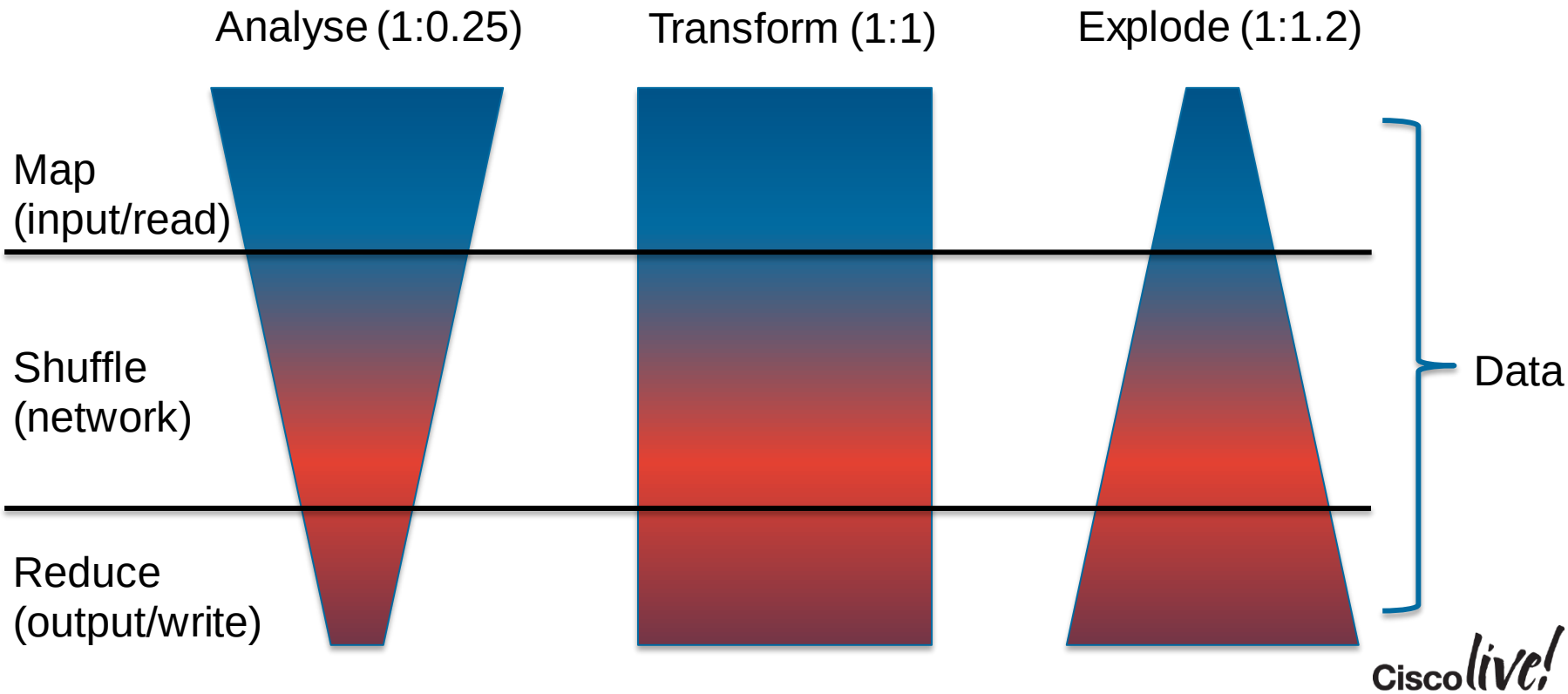
Large Pipeline
(Hadoop Replication)

Map and Reduce Traffic



Typical Hadoop Job Patterns

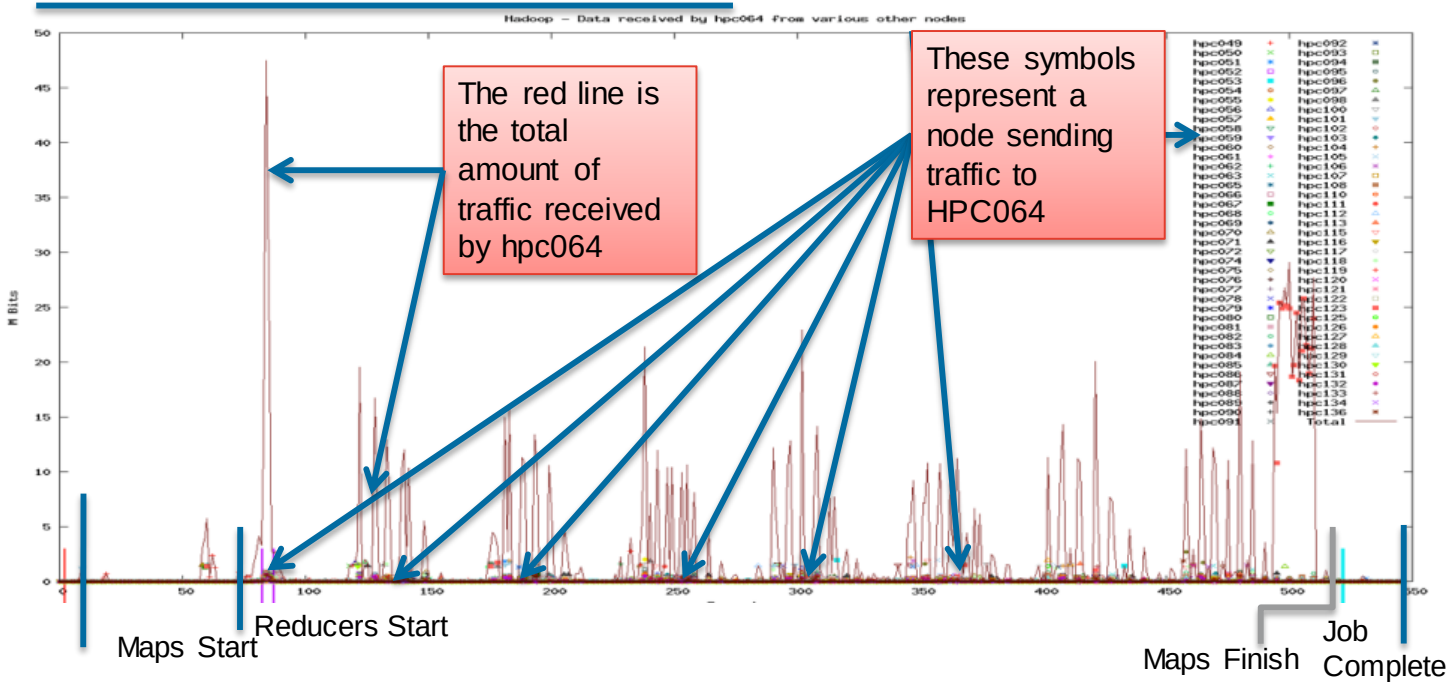
Different workloads can have widely varying network impact



Analyse Workload

Wordcount on 200K Copies of complete works of Shakespeare

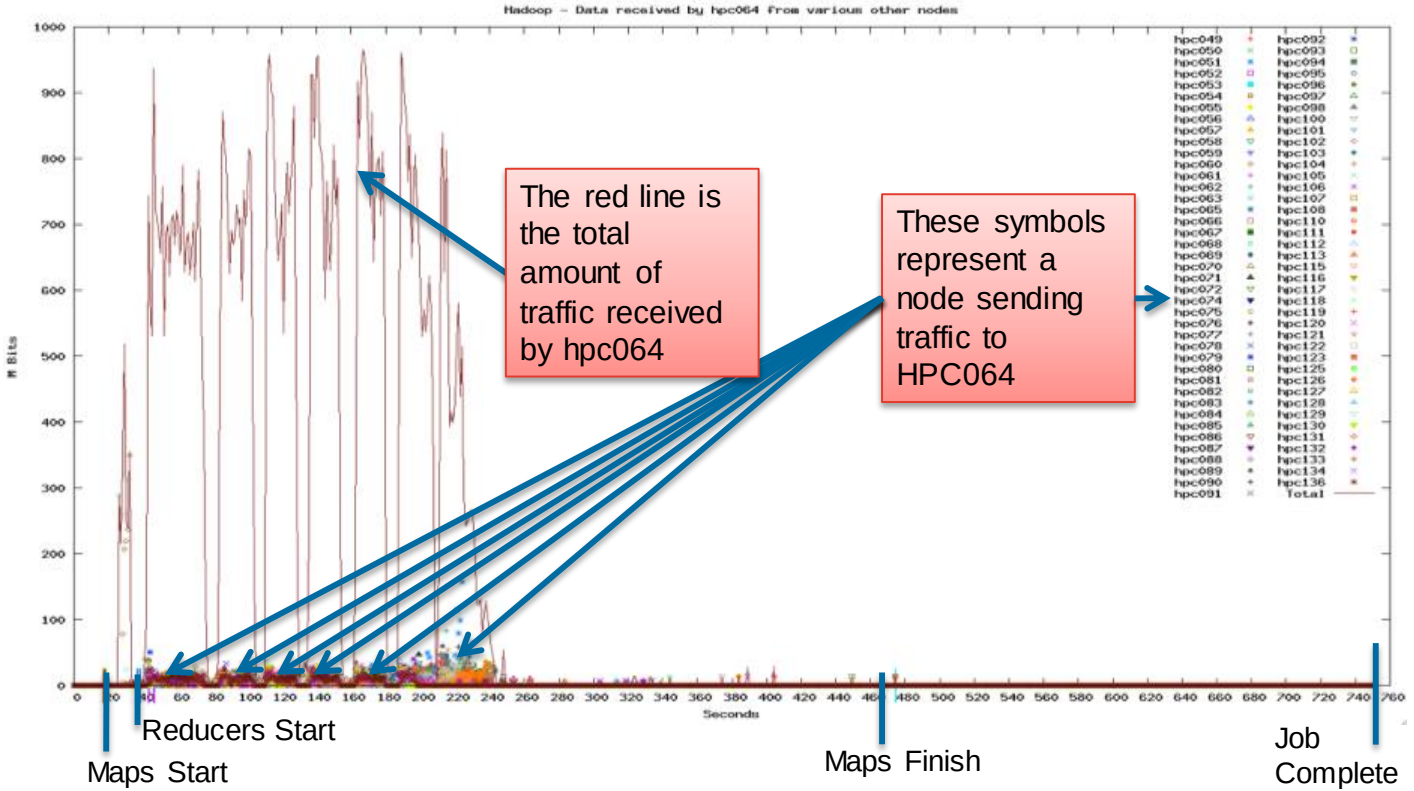
Note:
Due the combination of the length of the Map phase and the reduced data set being shuffled, the network is being utilised throughout the job, but by a limited amount.



Network graph of all traffic received on a single node (80 node run)

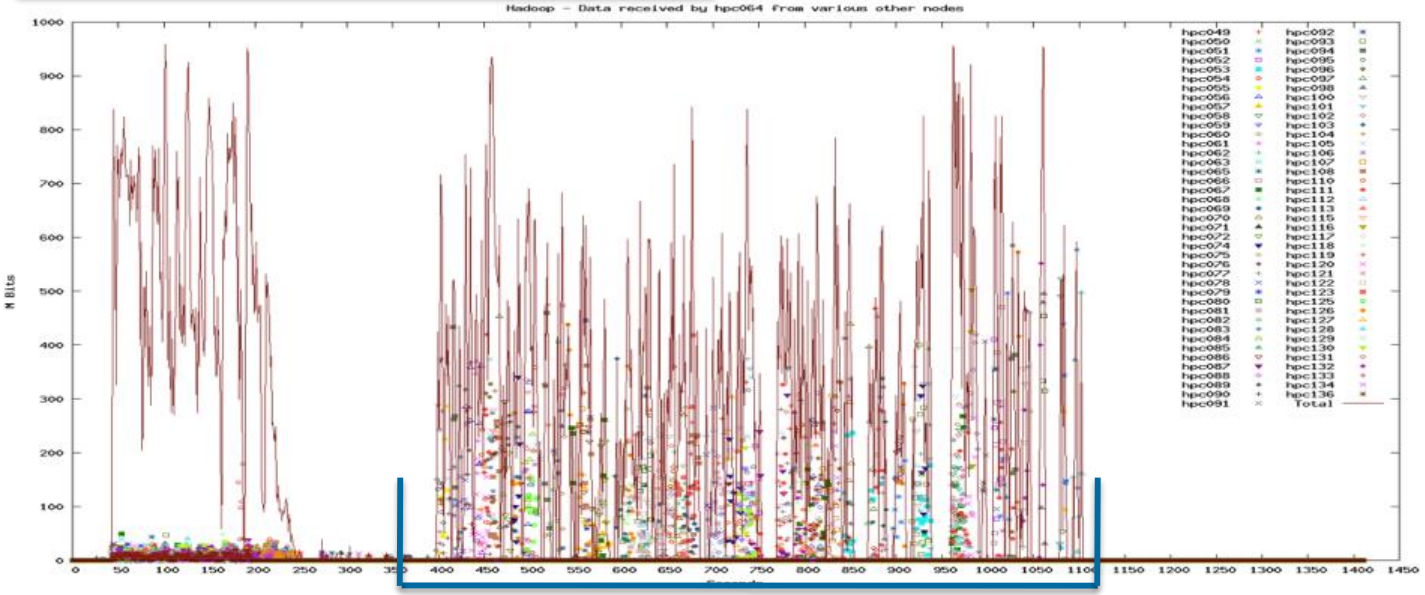
Transform Workload (1TB Terasort)

Network graph of all traffic received on a single node (80 node run)



Transform Workload (1TB Terasort With Output Replication)

Note:
If output replication is enabled, then at the end of the job HDFS must store additional copies. For a 1TB sort, additional 2TB will need to be replicated across the network.

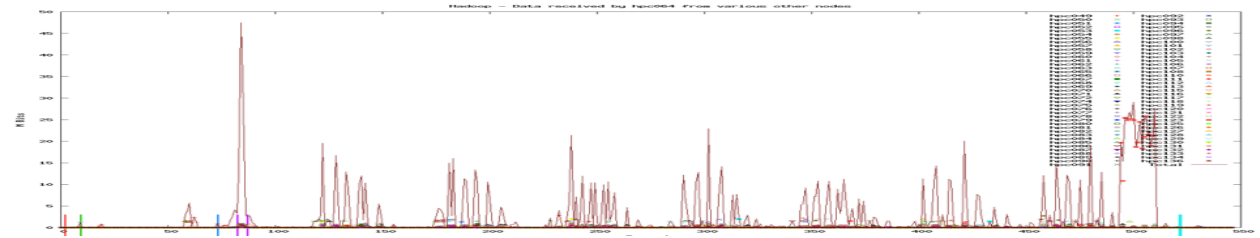


Output Data Replication Enabled

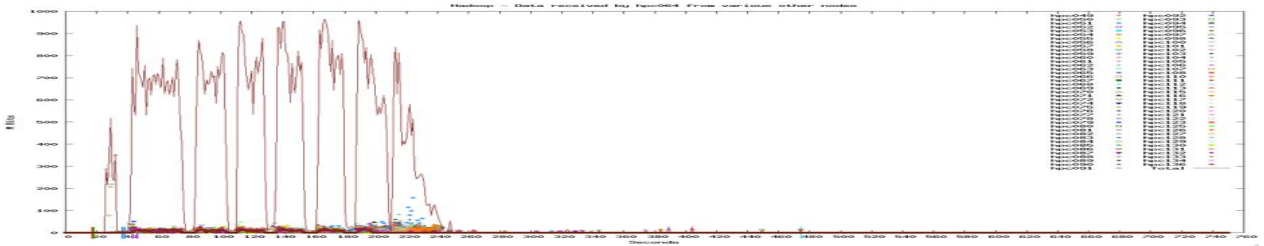
- Replication of 3 enabled (1 copy stored locally, 2 stored remotely)
- Each reduce output is replicated now, instead of just stored locally

Job Patterns - Summary

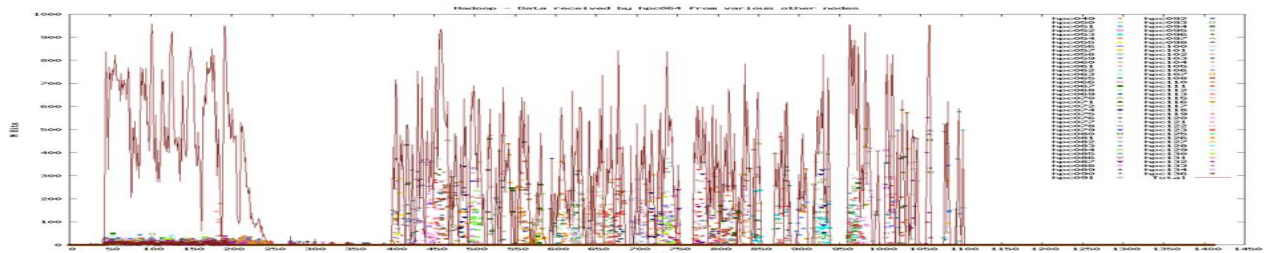
Job Patterns have varying impact on network utilisation



Analyse
Simulated with Shakespeare Wordcount



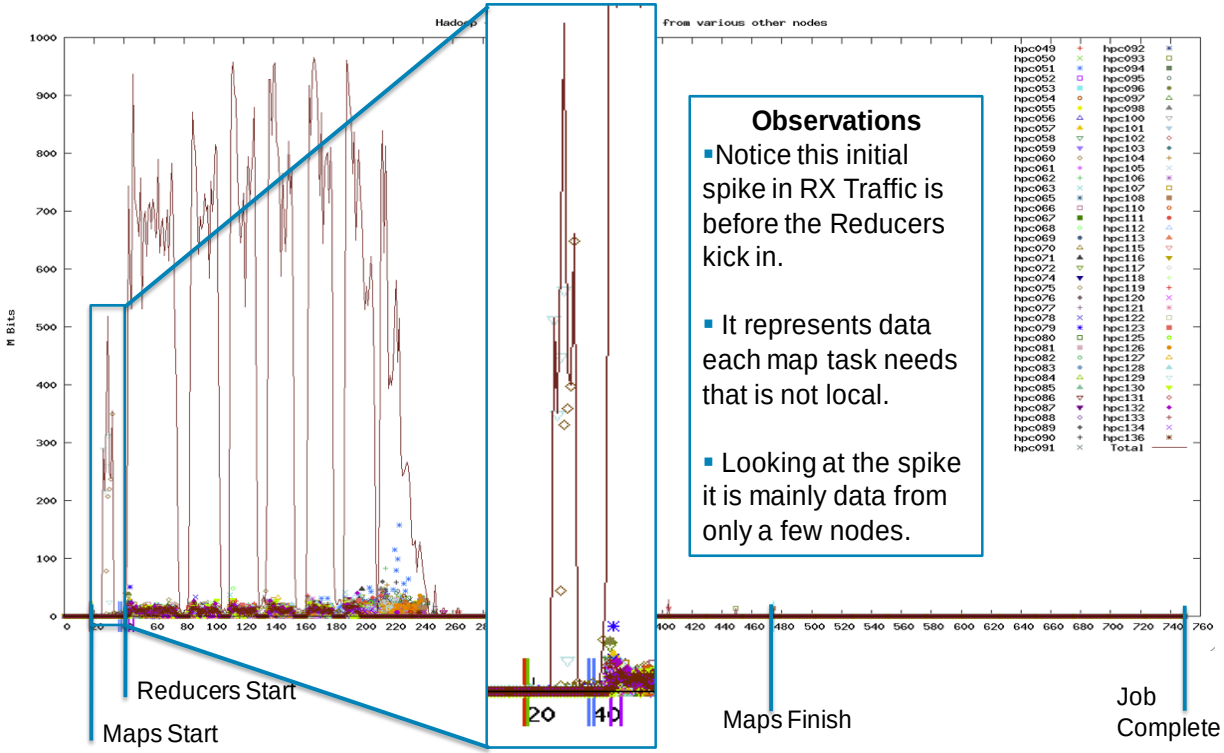
Extract Transform Load (ETL)
Simulated with Yahoo TeraSort



Extract Transform Load (ETL)
Simulated with Yahoo TeraSort with output replication

Data Locality in Hadoop

Data Locality – the ability to process data where it is locally stored



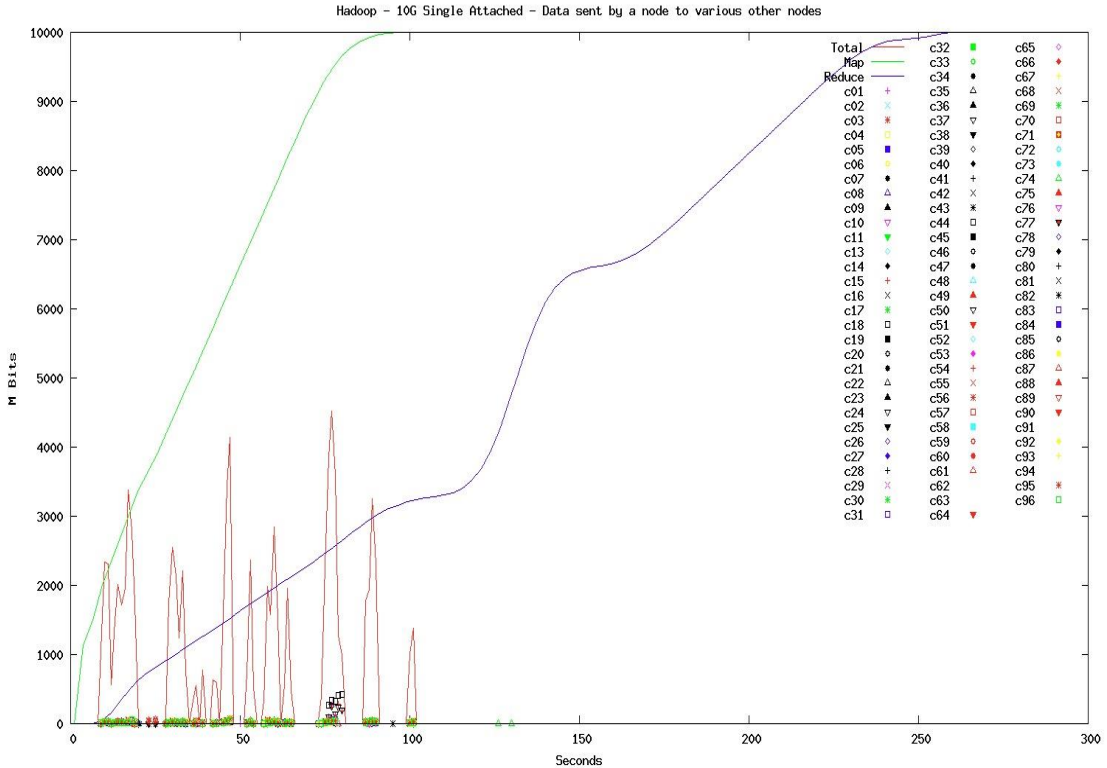
Map Tasks: Initial spike for non-local data. Sometimes a task may be scheduled on a node that does not have the data available locally.

Frequently Asked Questions

Cisco live!

Does Hadoop Really Need 10GE?

Definitely, so tune for it!

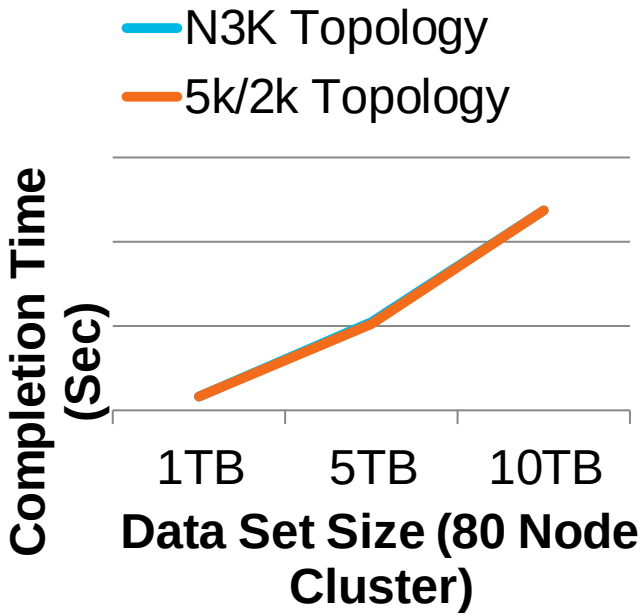


- Analytic workloads tend to be lighter on the network
- Transform workloads tend to be heavier on the network
- Hadoop has numerous parameters which affect network
- Take advantage of 10GE:
  - mapred.reduce.slowstart.completed.maps
  - dfs.balance.bandwidthPerSec
  - mapred.reduce.parallel.copies
  - mapred.reduce.tasks
  - mapred.tasktracker.reduce.tasks.maximum
  - mapred.compress.map.output

How Important is Network Latency for Hadoop?

Consistent, low network latency is desirable, but ultra low latency does not represent a significant factor for typical Hadoop workloads.

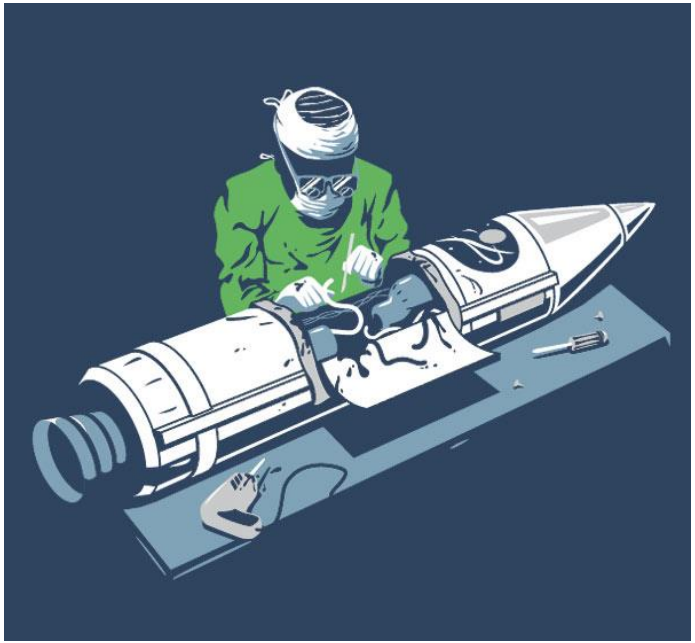
Note:
There is a difference in network latency vs. application latency. Optimisation in the application stack can decrease application latency that can potentially have a significant benefit.



“How Do I Size My Network?”

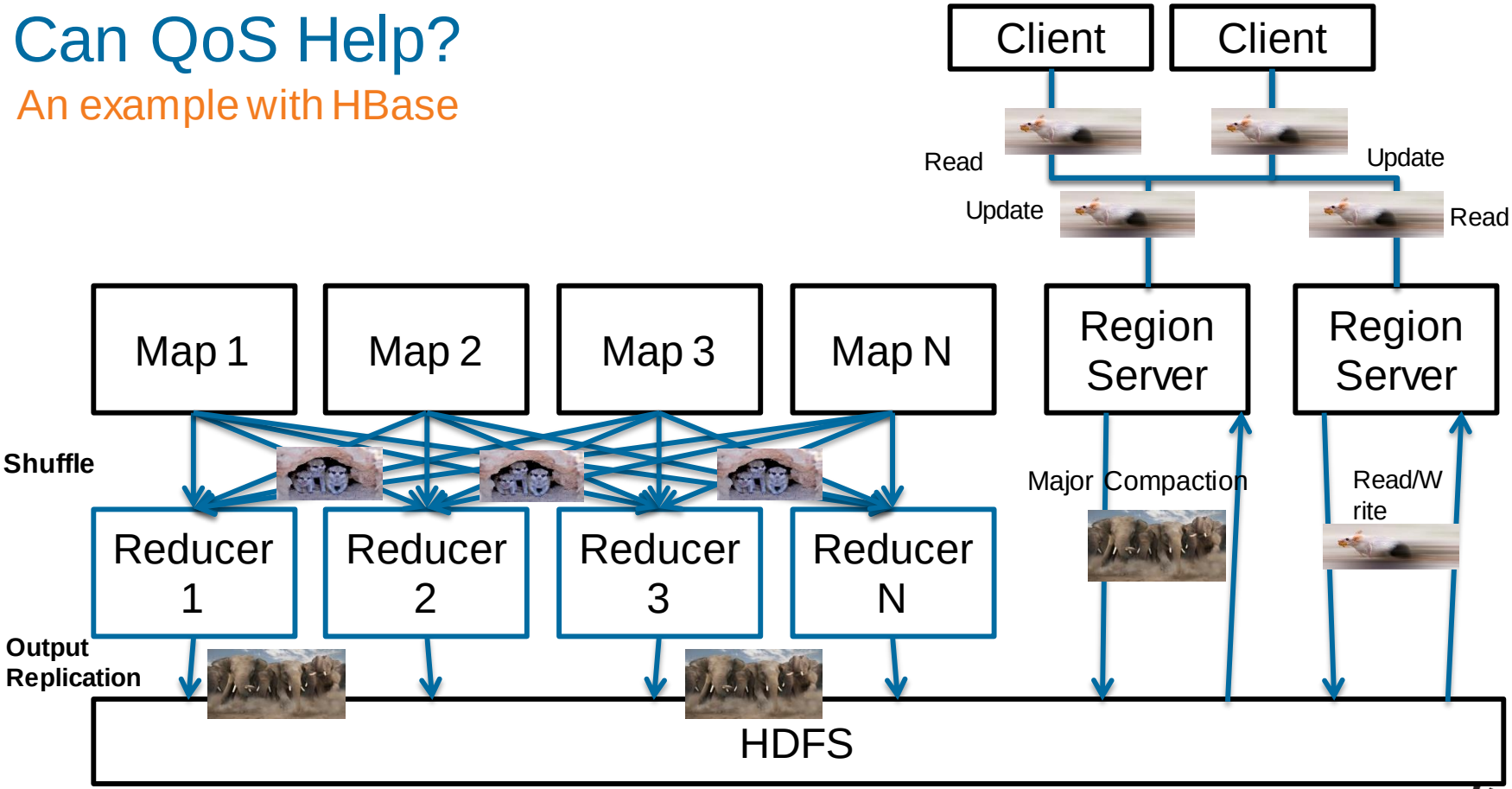
Don’t panic! It’s not rocket surgery…

- Basic port math
- Very homogenous design exercise – same for all nodes
- Start with total node count
  - Be sure to understand desired server-side bonding (active-active, active-passive, vPC, etc.)
  - Factor in projected growth
- Keep an eye on oversubscription
- Ask about server config
  - Fat or thin node?
  - Which software distribution?

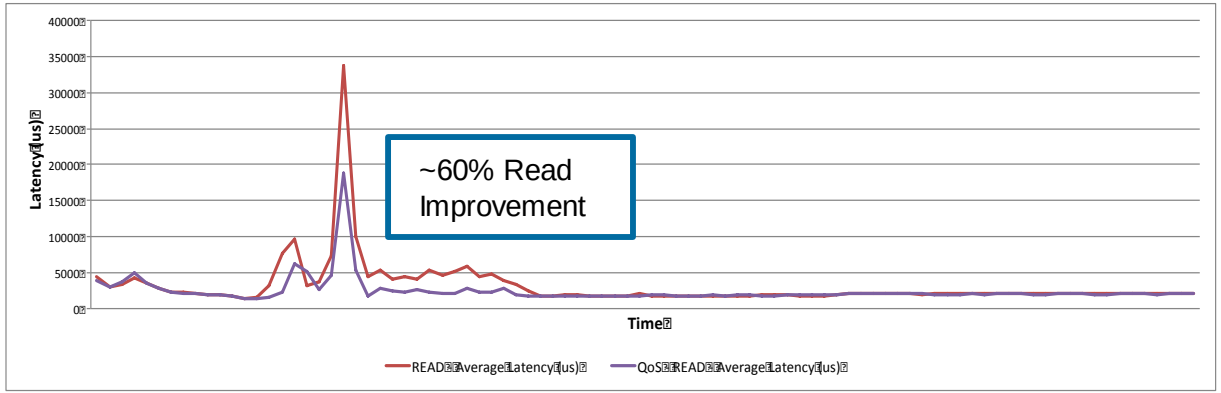


Can QoS Help?

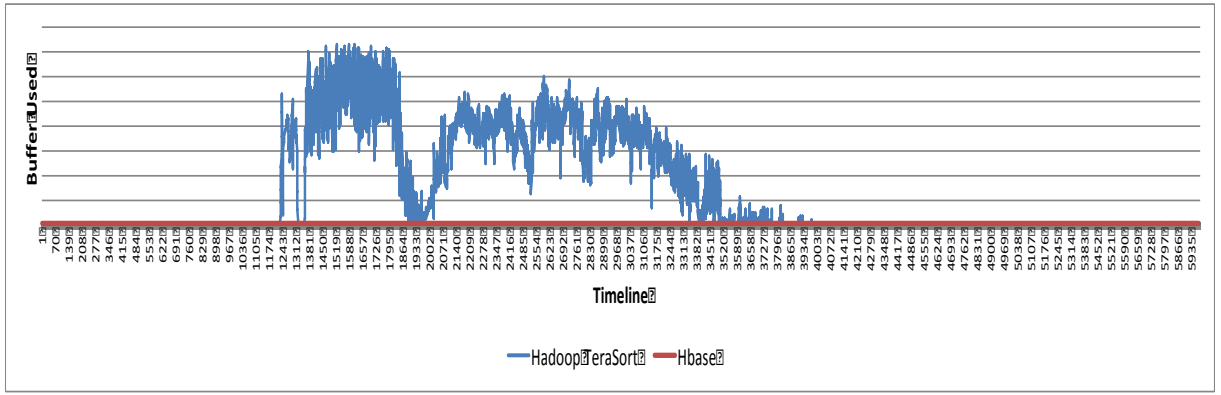
An example with HBase



HBase + MapReduce with QoS



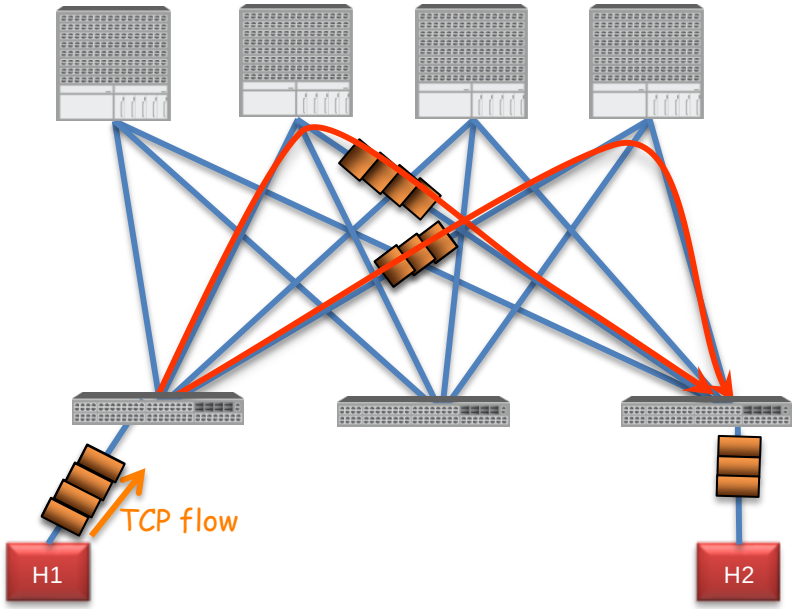
Read Latency
Comparison of Non-QoS vs. QoS Policy



Switch Buffer Usage
With Network QoS Policy to prioritise HBase Update/Read Operations

ACI Fabric Load Balancing

Flowlet Switching

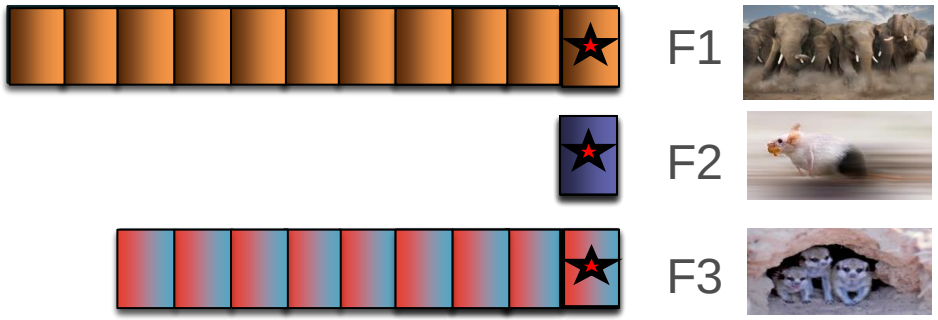


- ***Flowlet switching*** routes bursts of packets from the same flow independently, based on measured congestion of both external wires and internal ASICs
- Allows packets from the same flow to take different paths, while maintaining packet ordering
- **Provides better (more evenly distributed) utilisation of available paths**
- **Does all this transparently – *nothing to modify at the host/app level***

ACI Fabric Load Balancing

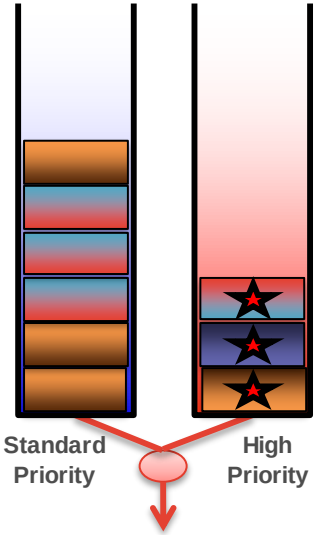
Dynamic Packet Prioritisation

Real traffic is a mix of large (elephant) and small (mice) flows.



Key Idea:
Fabric detects initial few flowlets of each flow and assigns them to a high priority class.

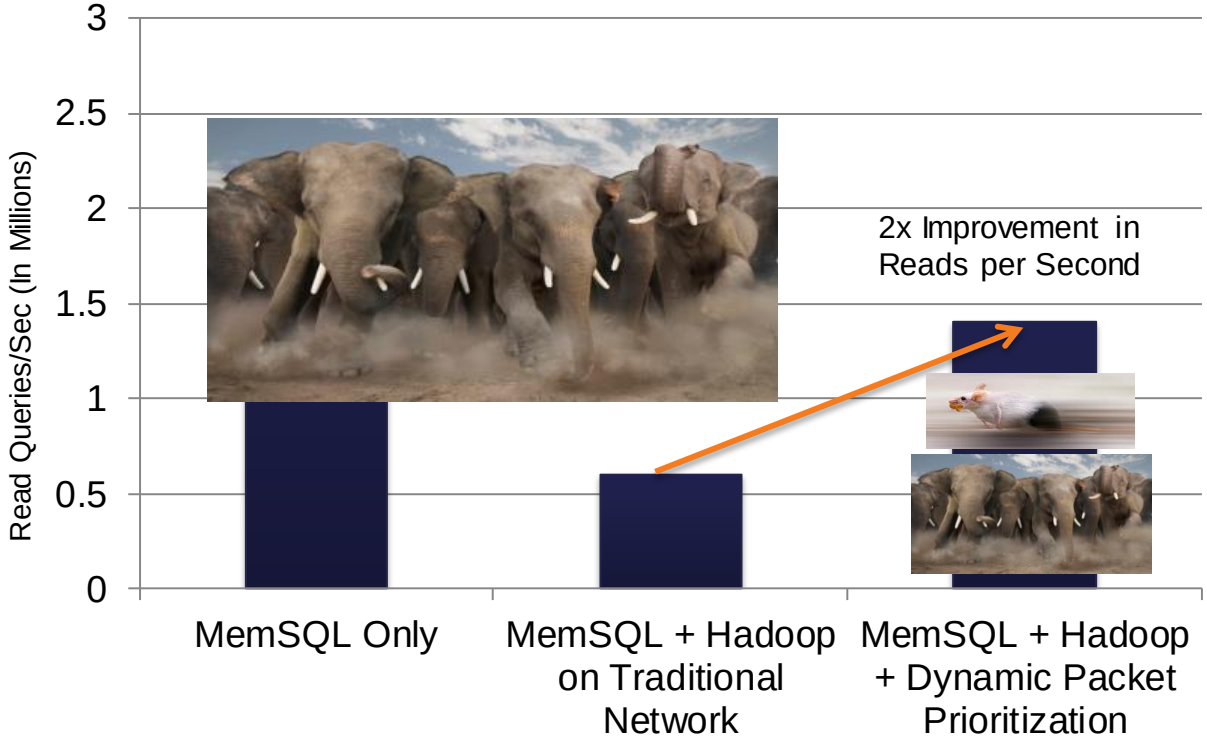
Standard (single priority):
Large flows severely impact performance (latency & loss). for small flows



Dynamic Flow Prioritisation:
Fabric automatically gives a higher priority to small flows.

Dynamic Packet Prioritisation

Helping heterogeneous workloads



- 80-node test cluster
- MemSQL used to generate heavy #'s of small flows - mice
- Large file copy workload unleashes elephant flows that trample MemSQL performance
- DPP enabled, helping to "protect" the mice from the elephants

Network Summary

- The network is the “system bus” of the Hadoop “supercomputer”
- Analytic- and ETL-style workloads can behave very differently on the network
- Ultra-low latency probably not critical for typical Hadoop workloads
- Minimise oversubscription, leverage QoS and DPP, and tune Hadoop to take advantage of 10GE – *distribute fairly*

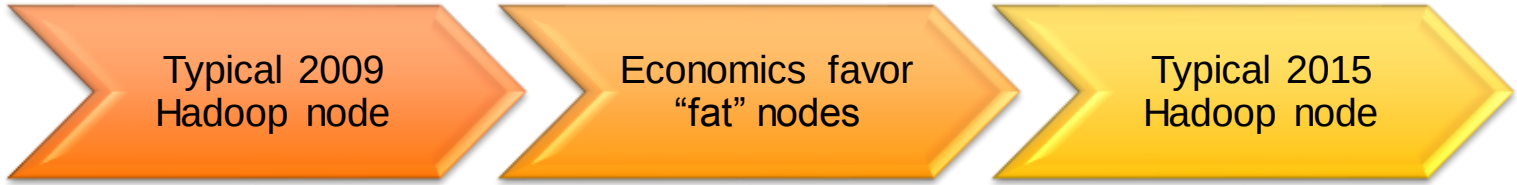
Cisco UCS and Big Data

Cisco live!

“Life is unfair, and the unfairness is distributed unfairly.”

-Russian proverb

Hadoop Server Hardware Evolving in the Enterprise



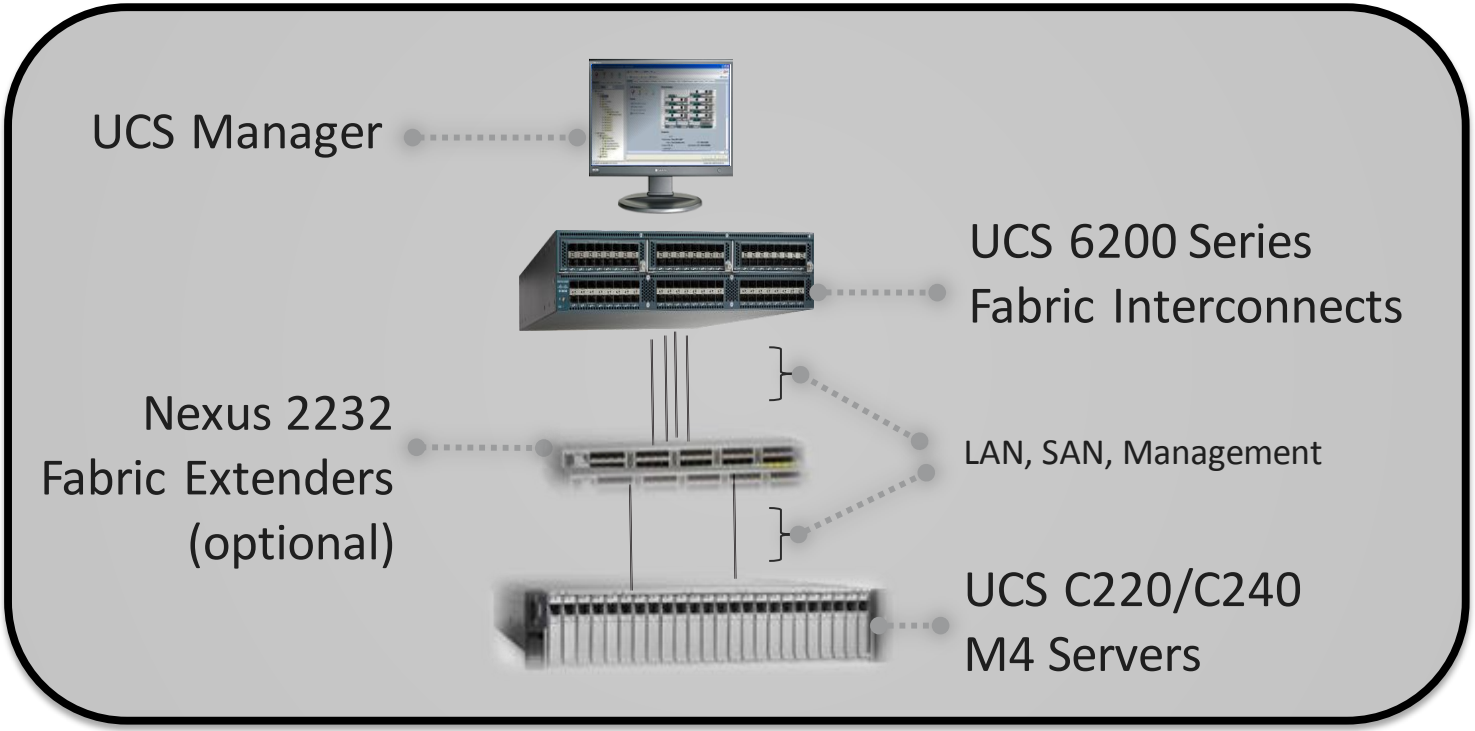
- 1RU server
- 4 x 1TB 3.5" spindles
- 2 x 4-core CPU
- 1 x GE
- 24 GB RAM
- Single PSU
- Running Apache
- $

- 6x-9x more data/node
- 3x-6x more IOPS/node
- Saturated gigabit, 10GE on the rise
- Fewer total nodes lowers licensing/support costs
- Increased significance of node and switch failure

- 2RU server
- 12 x 4TB 3.5" or 24 x 1TB 2.5" spindles
- 2 x 6-12 core CPU
- 2 x 10GE
- 128-256 GB RAM
- Dual PSU
- Running commercial/licensed distribution
- $$$

Cisco UCS Common Platform Architecture (CPA)

Building Blocks for Big Data



Cisco live!

New UCS Reference Configurations for Big Data

Quarter-Rack UCS Solution for MPP, NoSQL – High Performance

2 x UCS 6248
8 x C220 M4 (SFF)
2 x E5-2680v3
256GB
6 x 400-GB SAS SSD

Full Rack UCS Solution for Hadoop, NoSQL – Balanced

2 x UCS 6296
16 x C240 M4 (SFF)
2 x E5-2680v3
256GB
24 x 1.2TB 10K SAS

Full Rack UCS Solution for Hadoop Capacity-Optimised

2 x UCS 6296
16 x C240 M4 (LFF)
2 x E5-2620v3
128GB
12 x 4TB 7.2K SATA

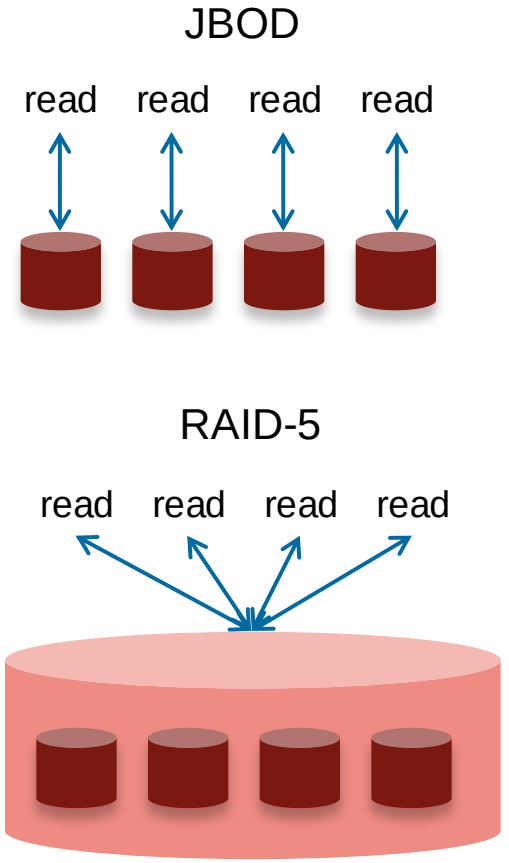
Frequently Asked Questions

Cisco live!

Hadoop and JBOD

Why not use RAID-5?

- It hurts performance:
  - RAID-5 turns parallel sequential reads into slower *random* reads
  - RAID-5 means speed limited to the *slowest* device in the group
- It's wasteful: Hadoop already replicates data, no need for more replication
  - Hadoop block copies serve **two** purposes: 1) redundancy and 2) performance (more copies available increases data locality % for map tasks)



Can I Virtualise?

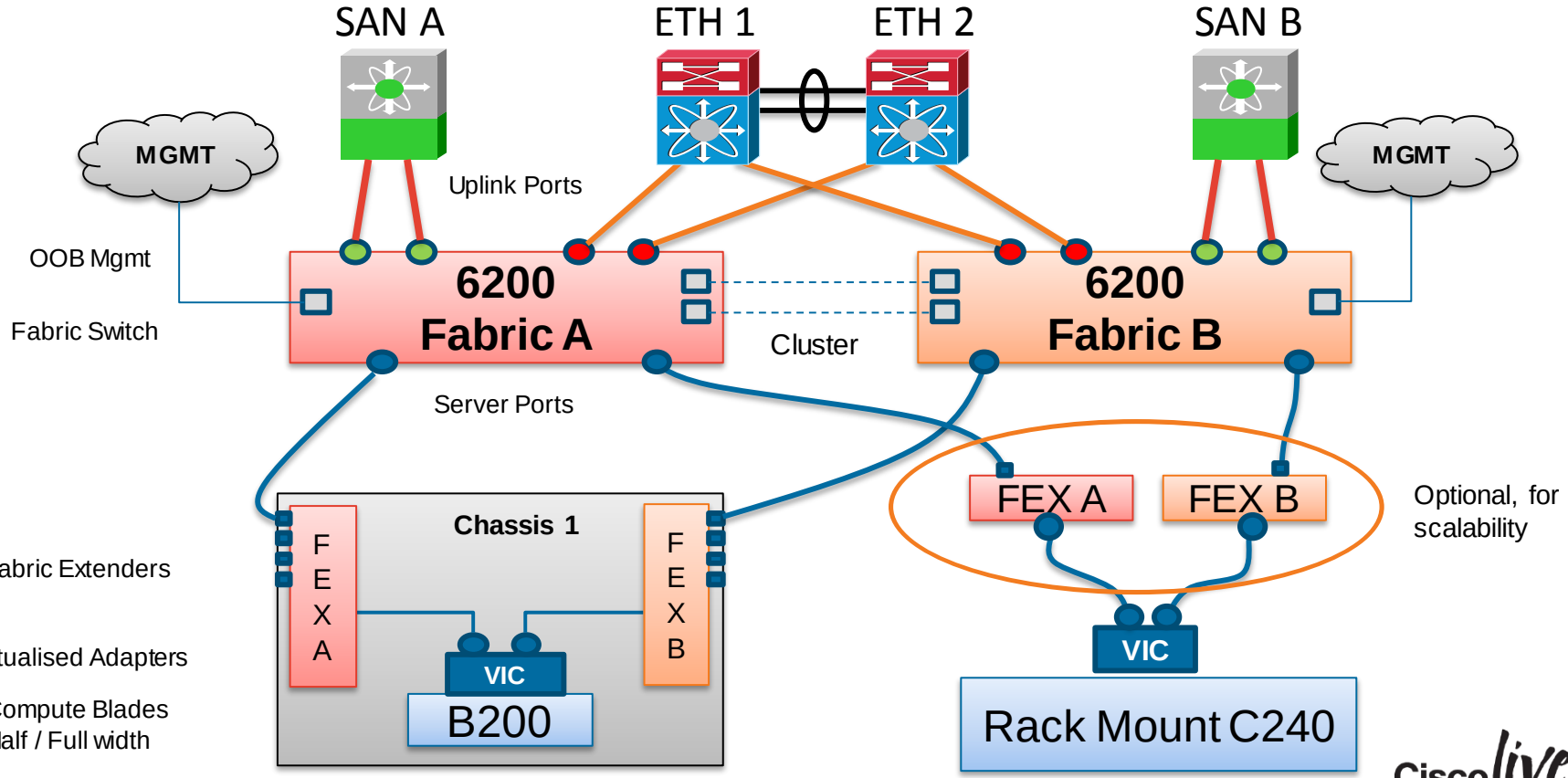
Yes you can (easy with UCS), but should you?

- Hadoop and most big data architectures can run virtualised
- However this is typically not recommended for performance reasons
  - Virtualised data nodes will contend for storage and network I/O
  - Hypervisor adds overhead, typically without benefit
- Some customers are running master/admin nodes (e.g. Name Node, Job Tracker, Zookeeper, gateways, etc.) in VM's, but consider single point of failure
- UCS is ideal for virtualisation if you go this route

CPA Network Design for Big Data

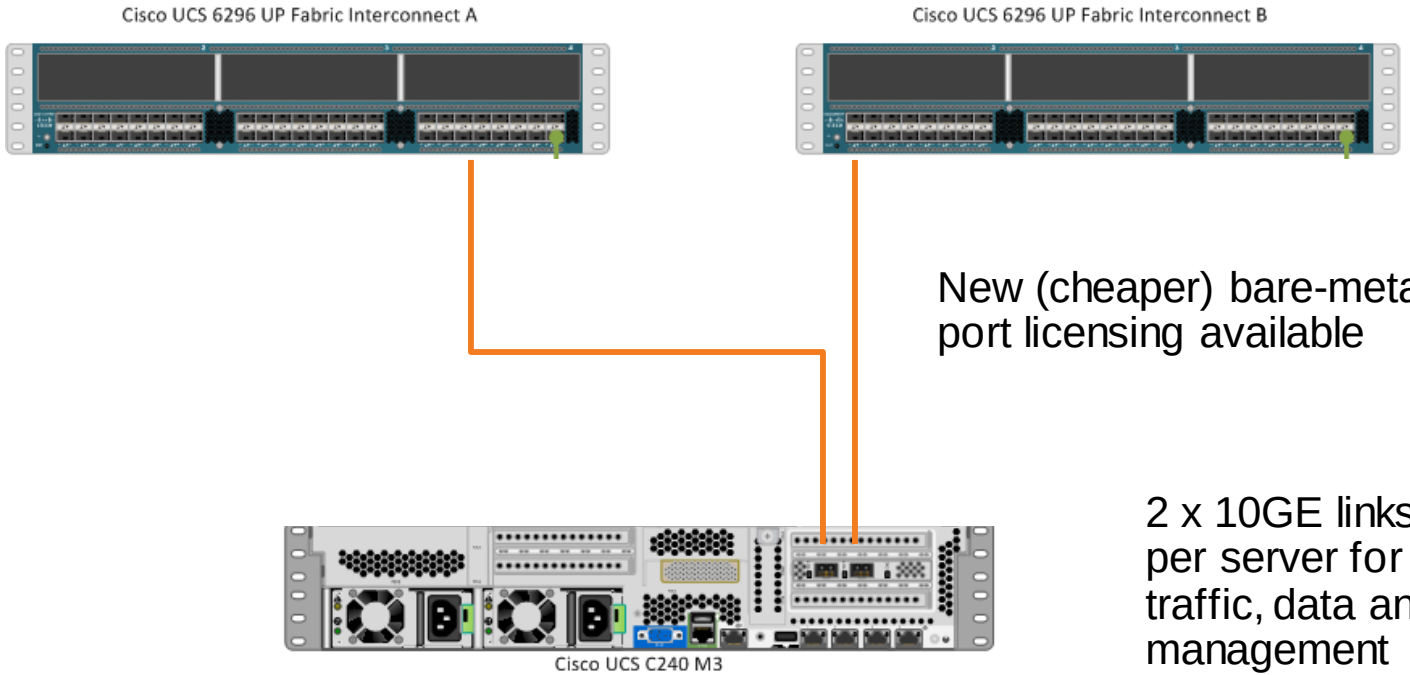
Cisco live!

Cisco UCS: Physical Architecture



CPA: Single-connect Topology (No Oversubscription)

Single wire for data and management

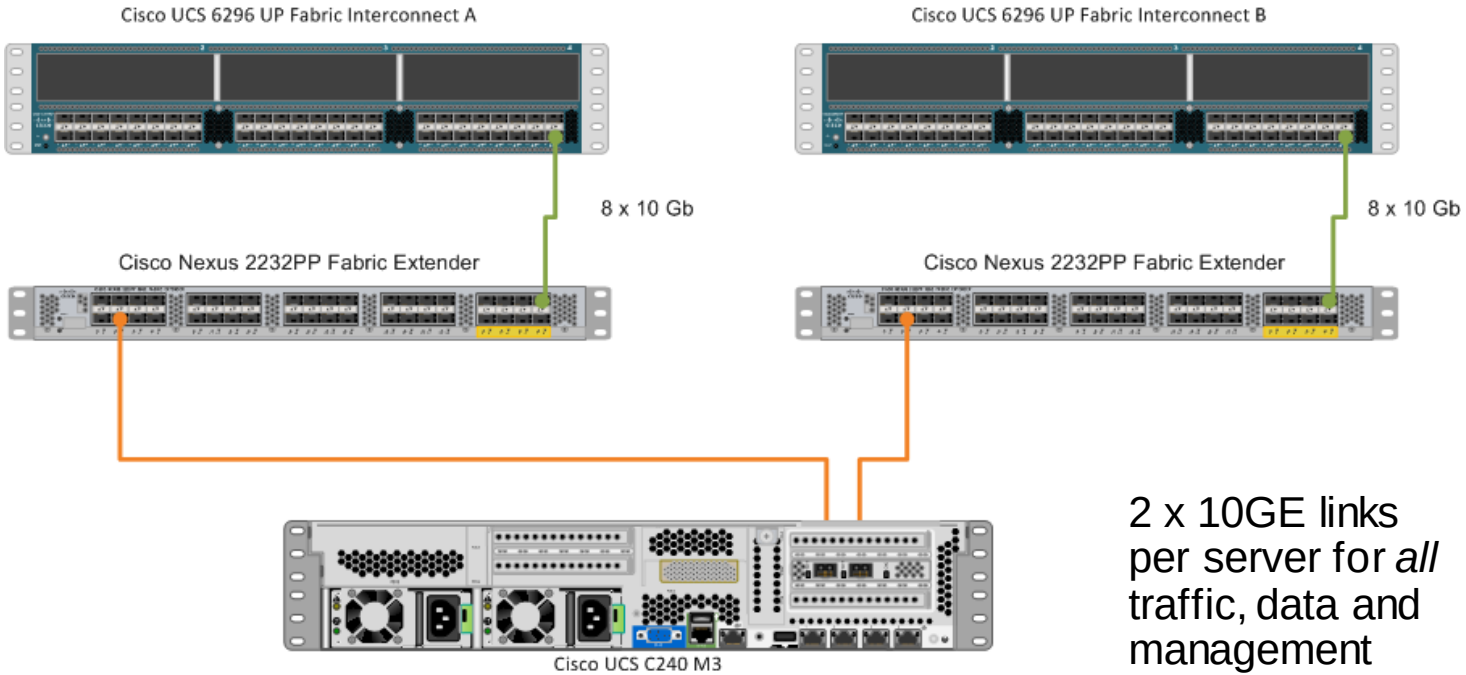


New (cheaper) bare-metal port licensing available

2 x 10GE links per server for *all* traffic, data and management

CPA: FEX Topology (Optional, For Scalability)

Single wire for data and management

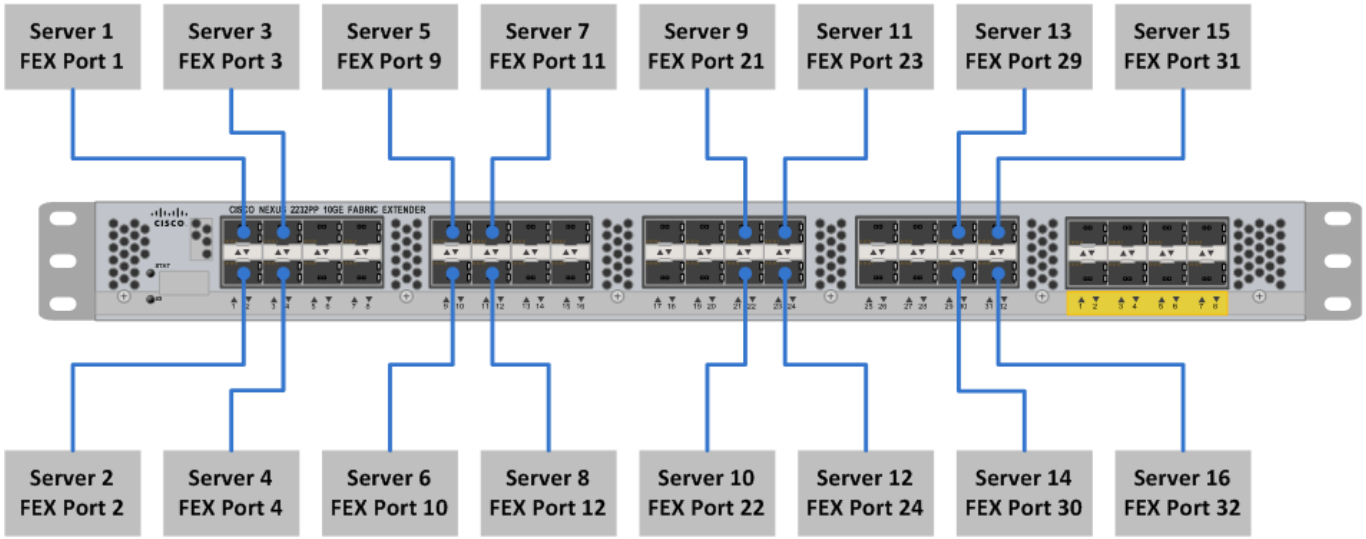


8 x 10GE uplinks per FEX= 2:1 oversub (16 servers/rack), no portchannel (static pinning)

2 x 10GE links per server for *all* traffic, data and management

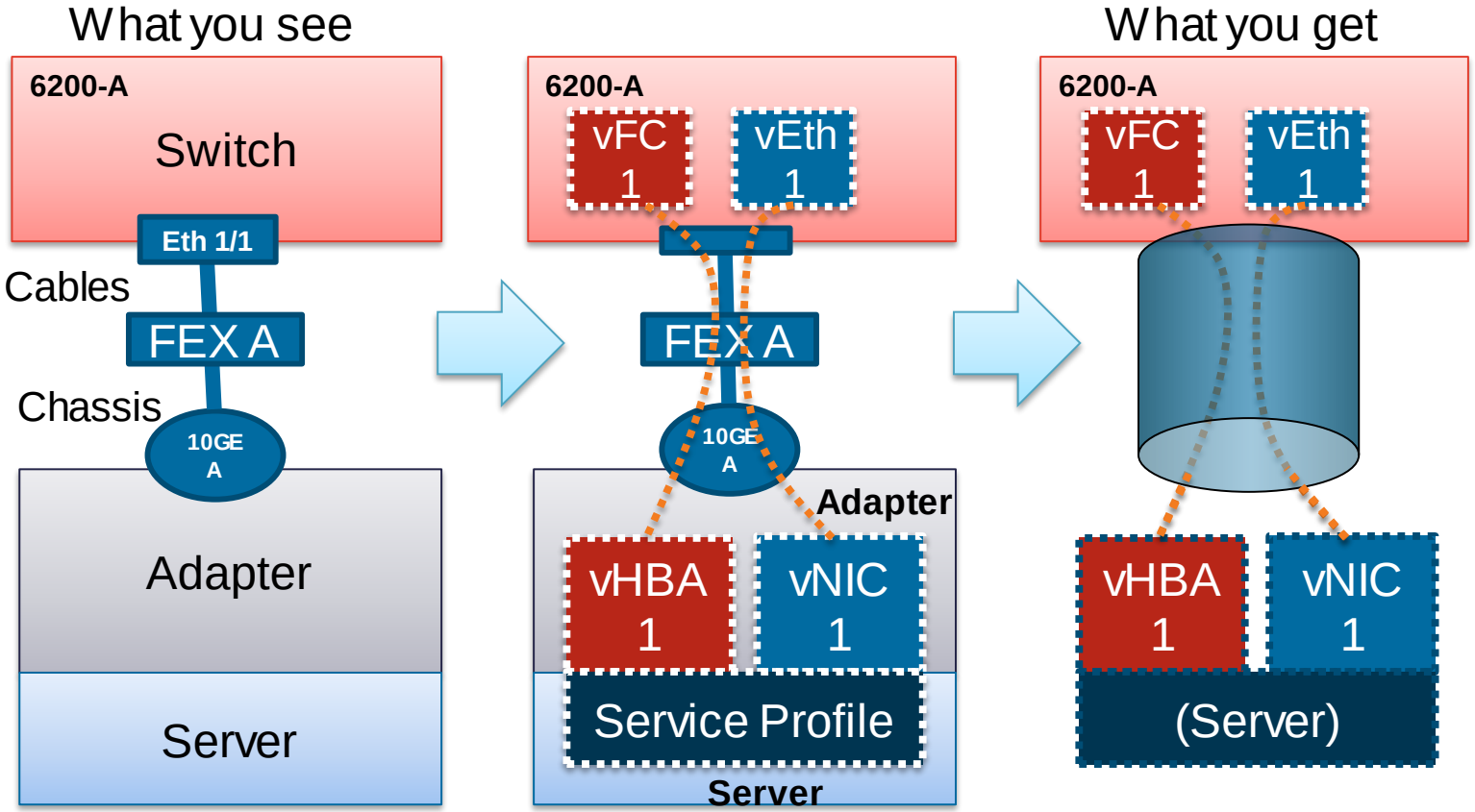
Cisco live!

CPA Recommended FEX Connectivity



- 2232 FEX has 4 buffer groups: ports 1-8, 9-16, 17-24, 25-32
- Distribute servers across port groups to maximise buffer performance and predictably distribute static pinning on uplinks

Virtualise the Physical Network Pipe



- ✓ Dynamic, Rapid Provisioning
- ✓ State abstraction
- ✓ Location Independence
- ✓ Blade or Rack

Physical Cable

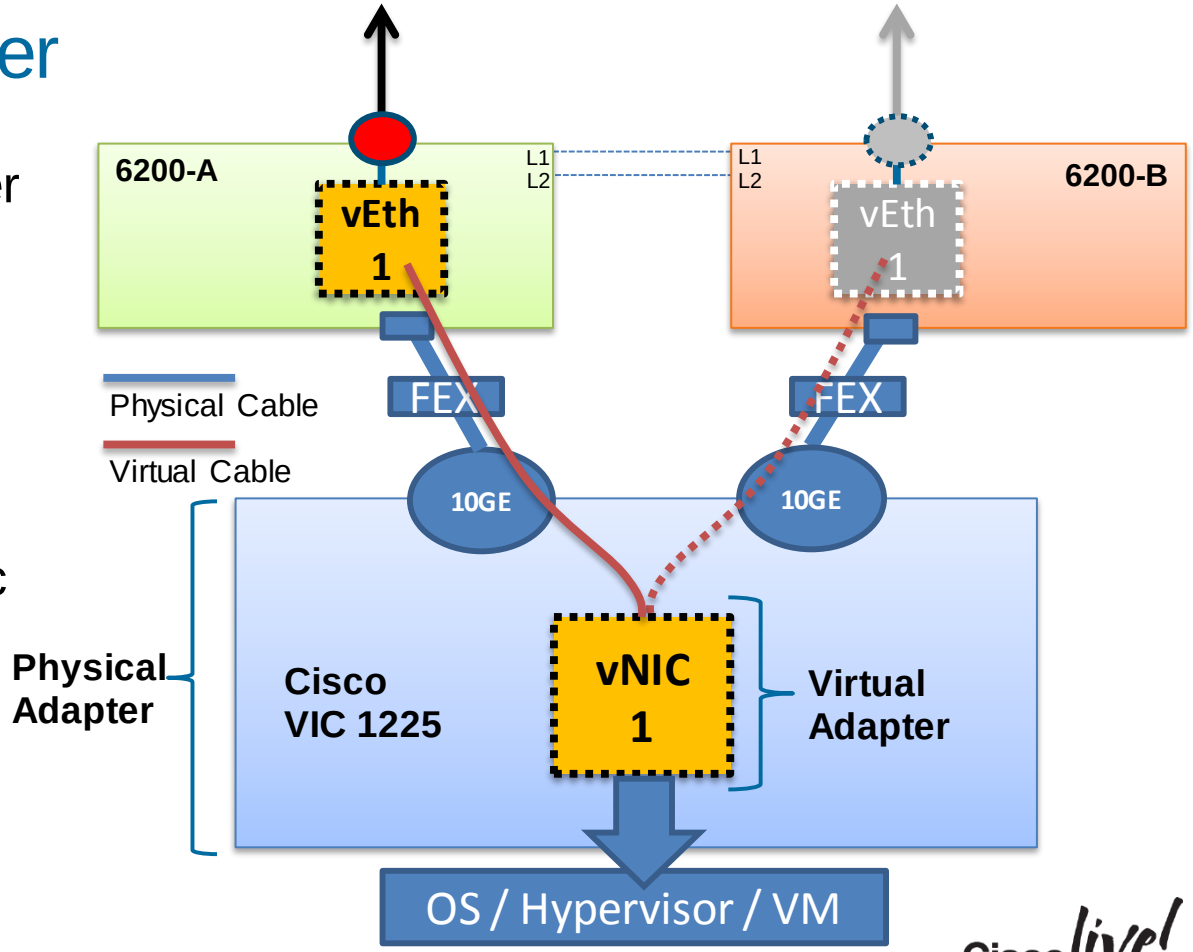
Virtual Cable (VN-Tag)

“NIC bonding is one of Cloudera’s highest case drivers for misconfigurations.”

http://blog.cloudera.com/blog/2015/01/how-to-deploy-apache-hadoop-clusters-like-a-boss/

UCS Fabric Failover

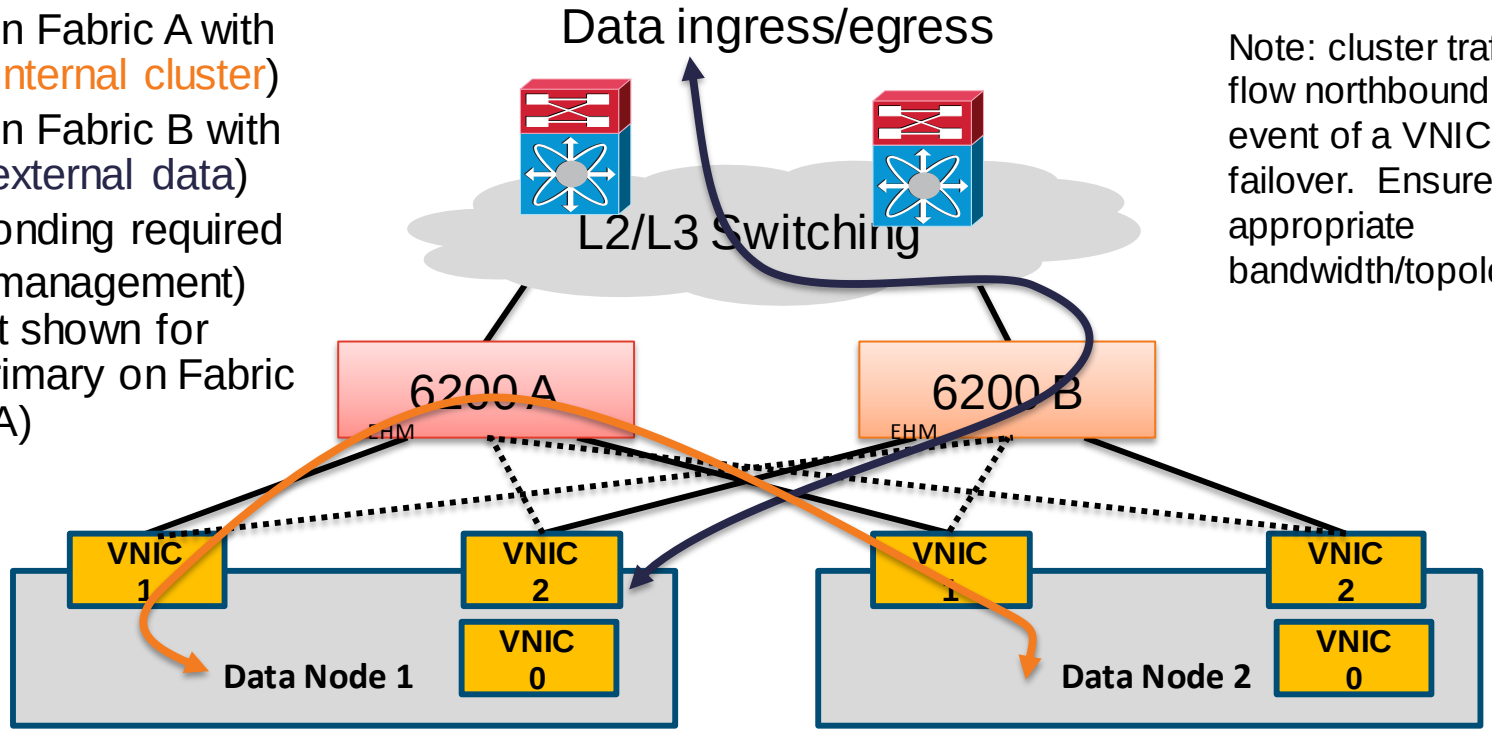
- Fabric provides NIC failover capabilities chosen when defining a service profile
- **Avoids traditional NIC bonding in the OS**
- Provides failover for both unicast and multicast traffic
- Works for any OS on bare metal
- (Also works for any hypervisor-based servers)



Recommended UCS Networking with Apache Hadoop

Use 2 VNICs with Fabric Failover on opposite fabrics for internal and external traffic

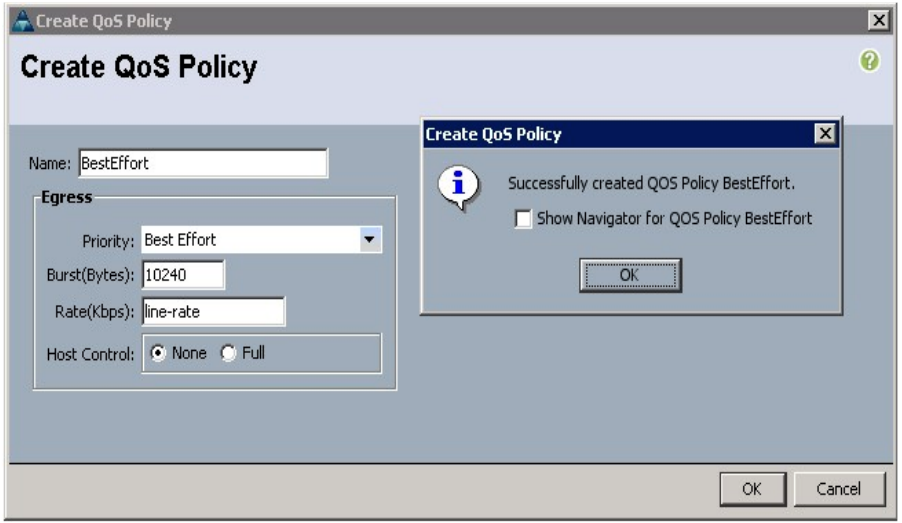
- VNIC 1 on Fabric A with FF to B (internal cluster)
- VNIC 2 on Fabric B with FF to A (external data)
- No OS bonding required
- VNIC 0 (management) wiring not shown for clarity (primary on Fabric B, FF to A)



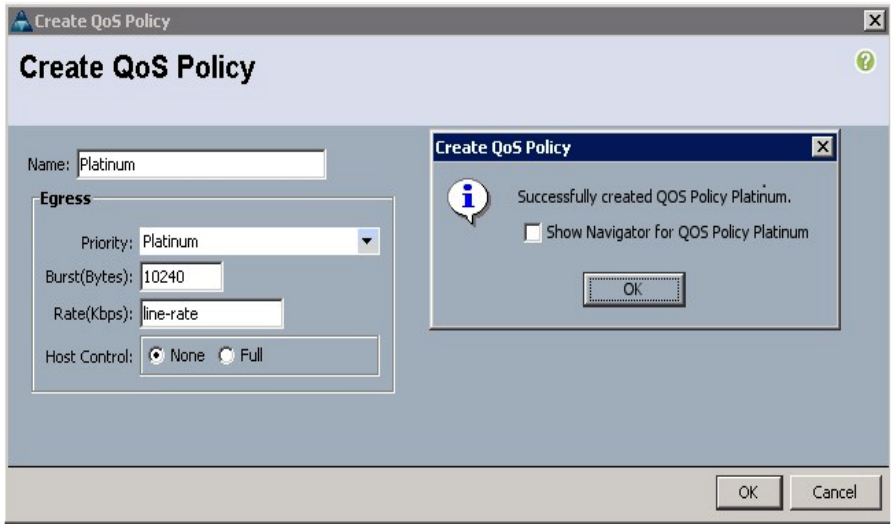
Note: cluster traffic will flow northbound in the event of a VNIC1 failover. Ensure appropriate bandwidth/topology.

Create QoS Policies

Leverage simplicity of UCS Service Profiles

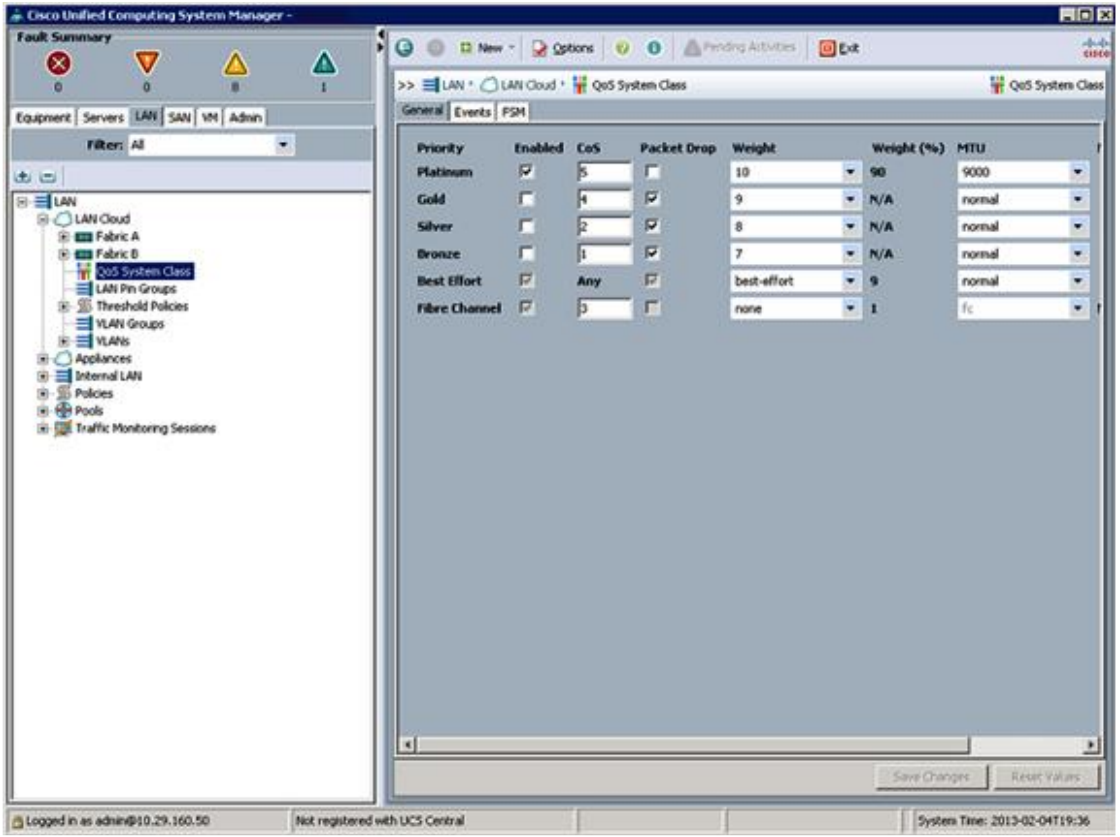


Best Effort policy for management VLAN



Platinum policy for cluster VLAN

Enable JumboFrames for Cluster VLAN



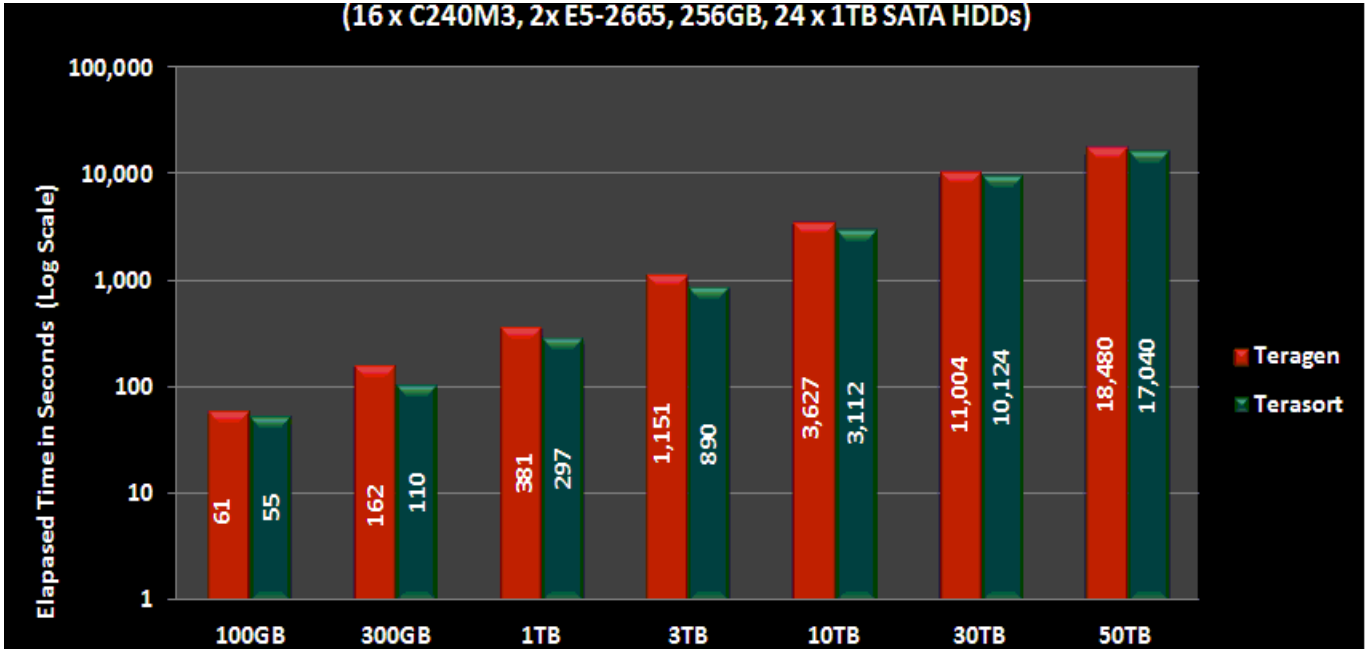
1. Select the LAN tab in the left pane in the UCSM GUI.
2. Select LAN Cloud > QoS System Class.
3. In the right pane, select the General tab
4. In the Platinum row, enter 9000 for MTU.
5. Check the Enabled Check box next to Platinum.
6. Click Save Changes.
7. Click OK.

CPA Sizing and Scaling for Big Data

Cisco live!

Cluster Scalability

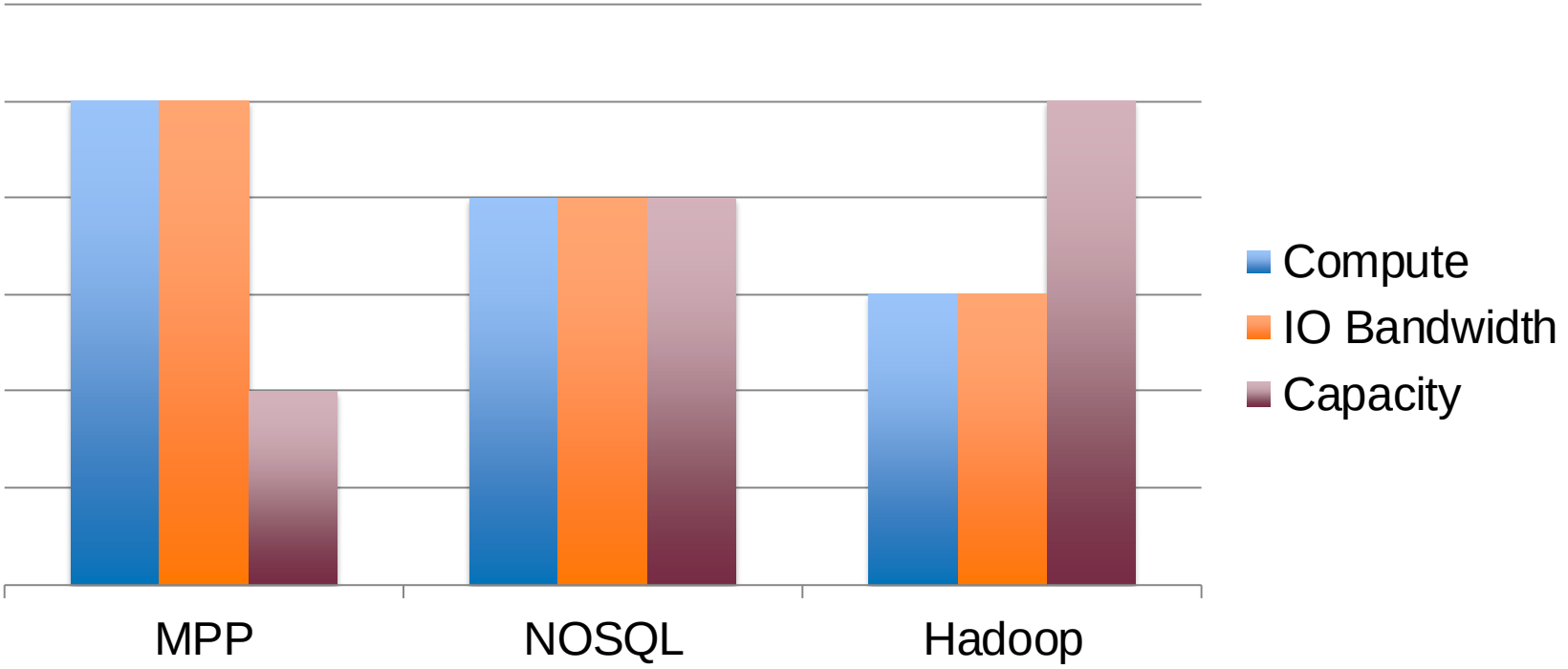
A general characteristic of an optimally configured cluster is a linear relationship between data set sizes and job completion times



Sizing
Part science, part art

- Start with current storage requirement
  - Factor in replication (typically 3x) and compression (varies by data set)
  - Factor in 20-30% free space for temp (Hadoop) or up to 50% for some NoSQL systems
  - Factor in average daily/weekly data ingest rate
  - Factor in expected growth rate (i.e. increase in ingest rate over time)
- If I/O requirement known, use next table for guidance
- Most big data architectures are very linear, so more nodes = more capacity and better performance
- Strike a balance between price/performance of individual nodes vs. total # of nodes

Remember: Different Apps With Different Needs



CPA Sizing and Application Guidelines

Server	CPU	2 x E5-2680v3	2 x E5-2680v3	2 x E5-2620v3
	Memory (GB)	256	256	128
	Disk Drives	6 x 400GB SSD	24 x 1.2TB 10K SFF	12 x 4TB 7.2K LFF
	IO Bandwidth (GB/Sec)	2.6	2.6	1.1
Rack-Level (32 x C220 or 16 x C240)	Cores	768	384	192
	Memory (TB)	8	4	2
	Capacity (TB)	64	460	768
	IO Bandwidth (GB/Sec)	192	42	16
Applications		MPP DB NoSQL	Hadoop NoSQL	Hadoop

Best Performance  Best Price/TB

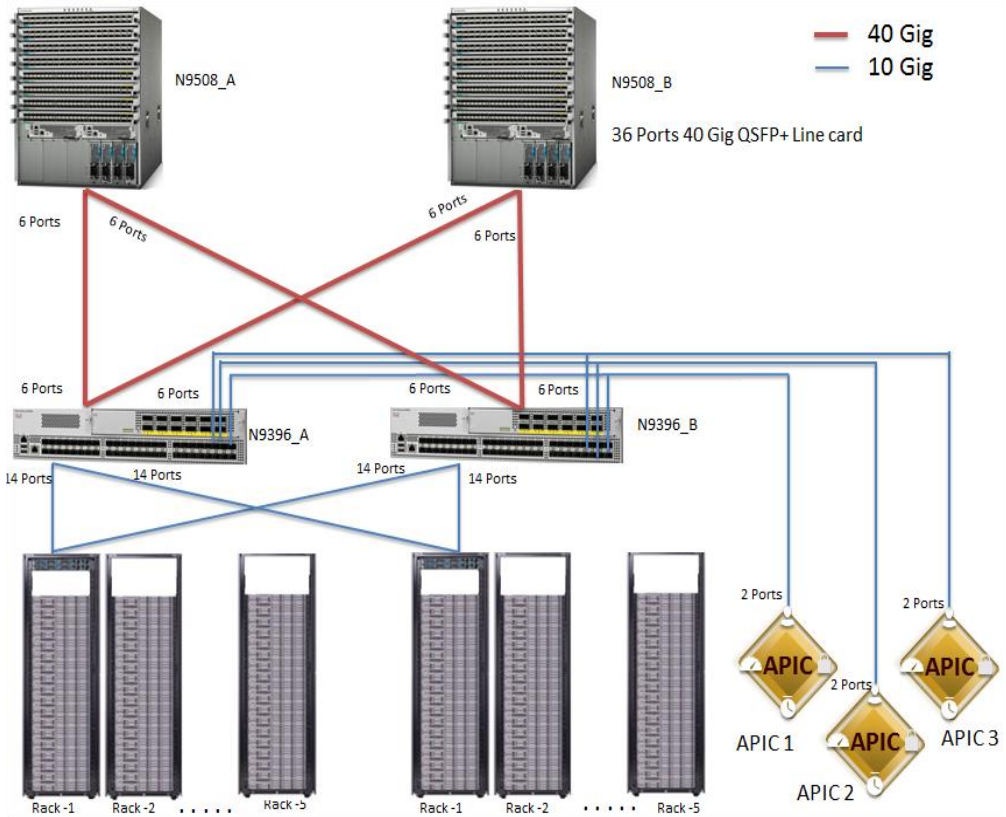
Scaling the CPA

Single Rack
16 servers

Single Domain
Up to 10 racks, 160 servers

Multiple Domains

Scaling Example: Nexus 9000 Validated Design



Use Nexus 9000 with ACI to scale out multiple UCS CPA domains (1000's of nodes) and/or to connect them to other application systems

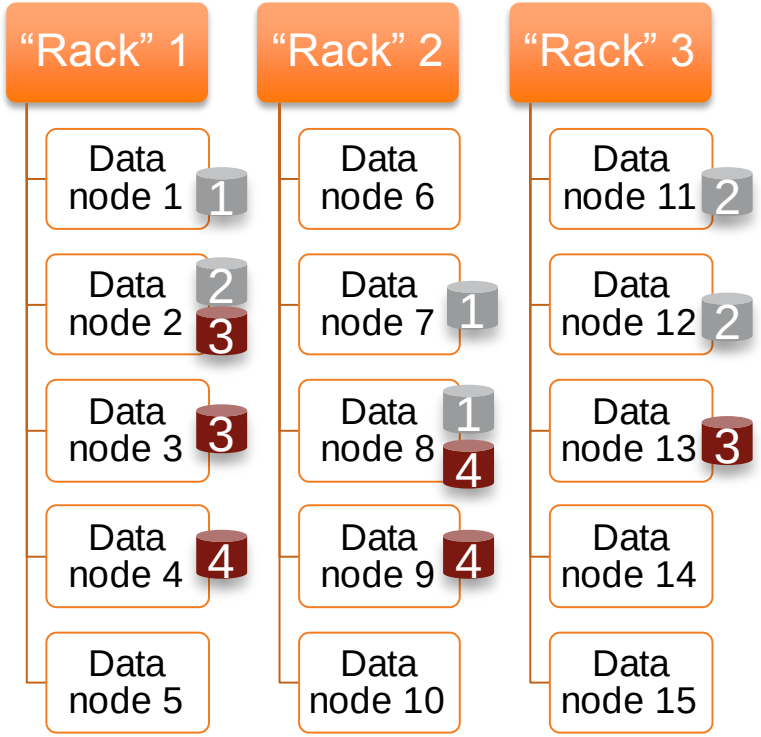
Enable ACI's Dynamic Packet Prioritisation and Dynamic Load Balancing to optimise multi-workload traffic flows

Scaling the Common Platform Architecture

Consider intra- and inter-domain bandwidth:

Servers Per Domain (Pair of Fabric Interconnects)	Available North-Bound 10GE ports (per fabric)	Southbound oversubscription (per fabric)	Northbound oversubscription (per fabric)	Intra-domain server-to-server bandwidth (per fabric, Gbits/sec)	Inter-domain server-to-server bandwidth (per fabric, Gbits/sec)
160	16	2:1 (FEX)	5:1	5	1
128	32	2:1 (FEX)	2:1	5	2.5
80	16	1:1 (no FEX)	5:1	10	2
64	32	1:1 (no FEX)	2:1	10	5

Rack Awareness



- Rack Awareness provides Hadoop the optional ability to group nodes together in logical "racks"
- Logical "racks" may or may not correspond to physical data centre racks
- Distributes blocks across different "racks" to avoid failure domain of a single "rack"
- It can also lessen block movement between "racks"
- Can be useful to control block placement and movement in UCSM integrated environments

Recommendations: UCS Domains and Racks

Single Domain Recommendation
Turn off or enable at physical rack level

- For simplicity and ease of use, leave Rack Awareness off
- Consider turning it on to limit physical rack level fault domain (e.g. localised failures due to physical data centre issues – water, power, cooling, etc.)

Multi Domain Recommendation
Create one Hadoop rack per UCS Domain

- With multiple domains, enable Rack Awareness such that each UCS Domain is its own Hadoop rack
- Provides HDFS data protection across domains
- Helps minimise cross-domain traffic

“The future is here, it’s just not evenly distributed.”

- William Gibson, author

Summary

Leverage UCS and Nexus to integrate big data into your data centre operations

- Think of big data clusters as a single “supercomputer”
- Think of the network as the “system bus” of the supercomputer
- Strive for consistency in your deployments
- The goal is an even distribution of load – *distribute fairly*
- Cisco Nexus and UCS Common Platform Architecture for Big Data can help!

Cisco live!

Call to Action

- Visit the World of Solutions for
  - Cisco Campus – (speaker to add relevant demos/areas to visit)
  - Walk in Labs – (speaker to add relevant walk in labs)
  - Technical Solution Clinics
- Meet the Engineer (Speaker to specify when they will be available for meetings)
- Lunch time Table Topics
- DevNet zone related labs and sessions
- Recommended Reading: for reading material and further resources for this session, please visit www.pearson-books.com/CLMilan2015

Cisco live!


Q & A

Complete Your Online Session Evaluation

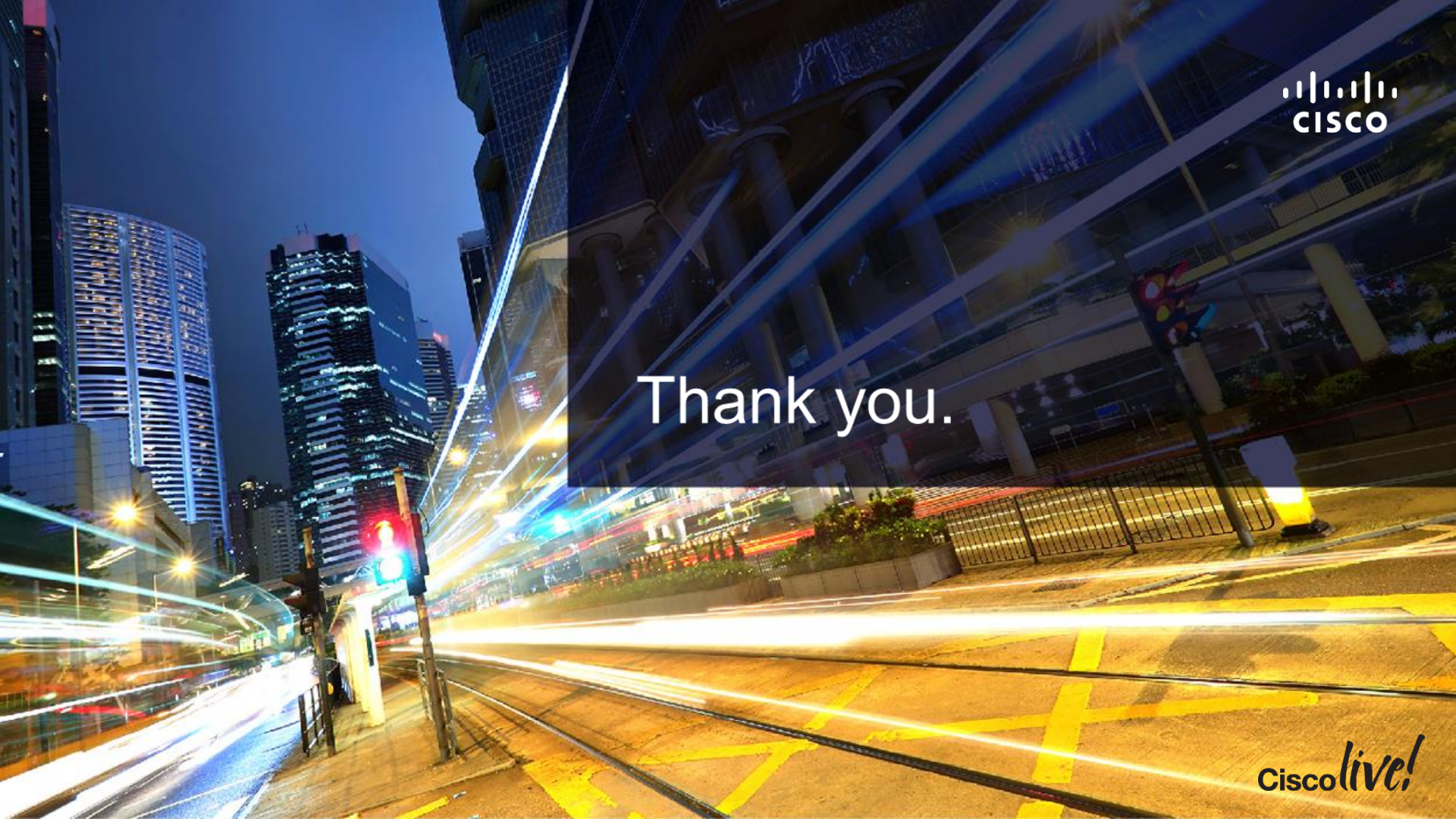
Give us your feedback and receive a Cisco Live 2015 T-Shirt!

Complete your Overall Event Survey and 5 Session Evaluations.

- Directly from your mobile device on the Cisco Live Mobile App
- By visiting the Cisco Live Mobile Site http://showcase.genie-connect.com/clmelbourne2015
- Visit any Cisco Live Internet Station located throughout the venue

T-Shirts can be collected in the World of Solutions on Friday 20 March 12:00pm - 2:00pm

Learn online with Cisco Live!
Visit us online after the conference for full access to session videos and presentations. www.CiscoLiveAPAC.com


Thank you.


CISCO