Cisco live!

What You Make Possible

TOMORROW
starts here.

CISCO

Expanding your Enterprise Application Ecosystem with Big Data

BRKDCT-1661

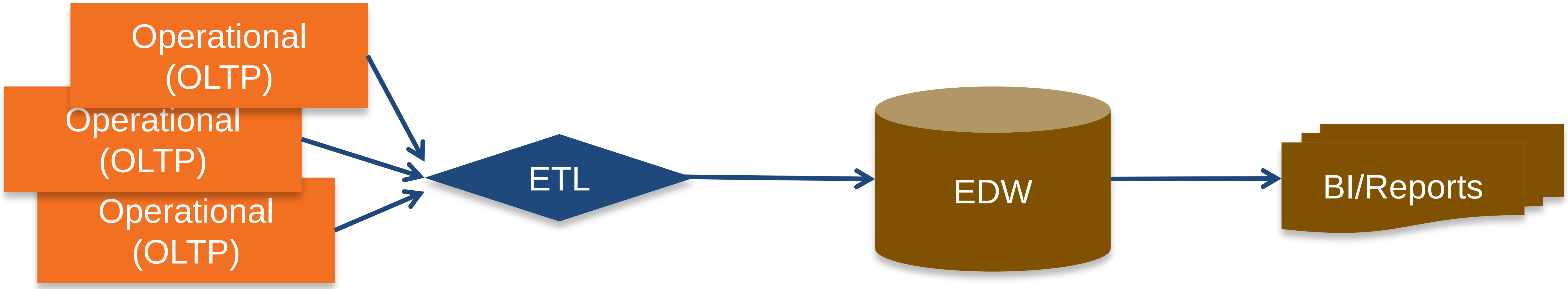
TOMORROW starts here.

"In my career I don't think I've seen such a fantastic opportunity or capability that's emerging now [with Hadoop]… we're doing things that were previously impossible to do on a scale you just couldn't imagine before [Hadoop]."

- Phil Shelley, Sears CTO

Enterprise Data Management and Big Data

Traditional Enterprise Data Management



Online Transactional Processing

Extract, Transform, and Load (batch processing)

Enterprise Data Warehouse

Business Intelligence

Traditional Business Intelligence Questions

Transactional Data (e.g. OLTP)

Real-time, but limited reporting/analytics

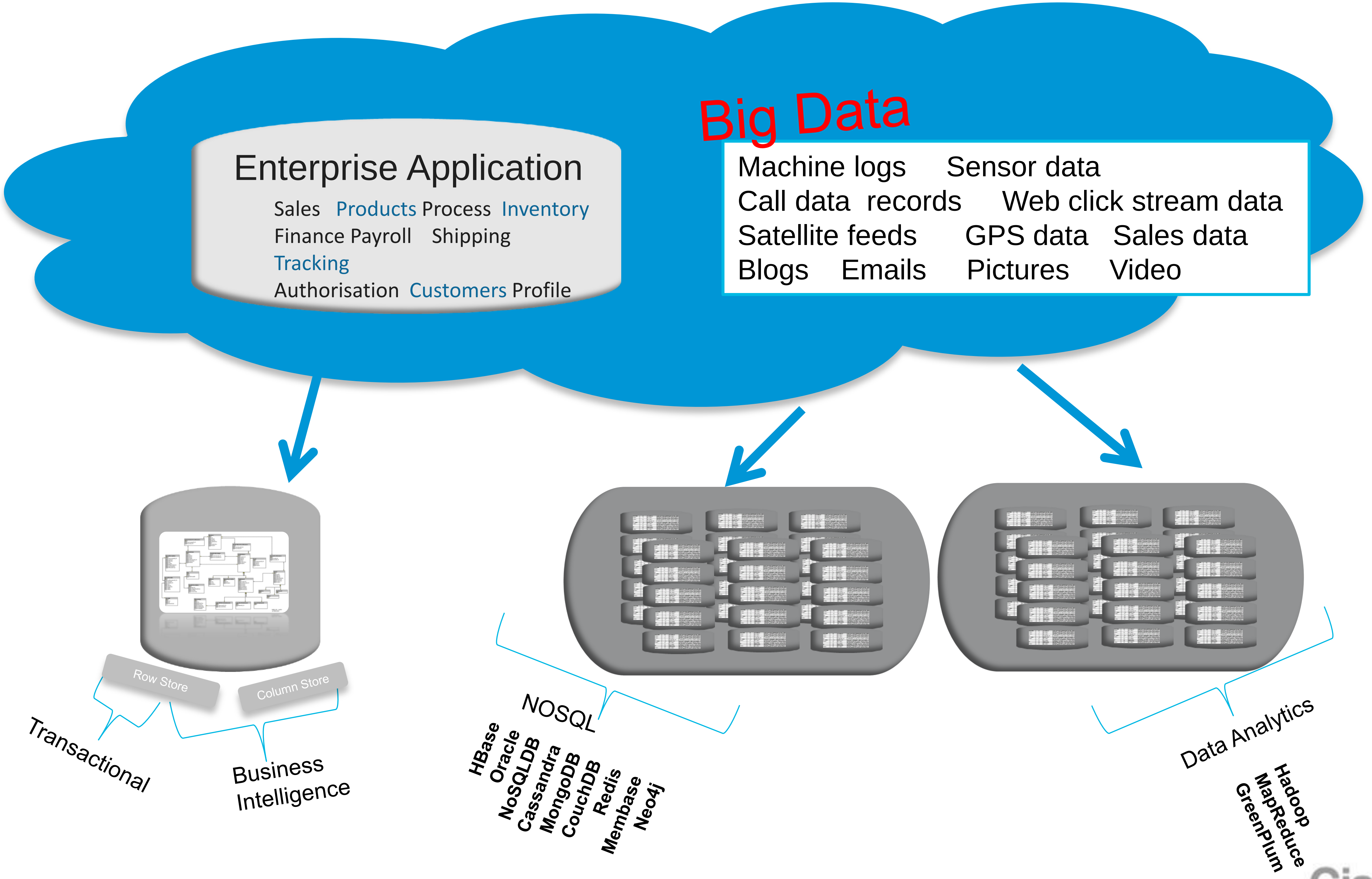
- What are the top 5 most active stocks traded in the last hour?
- How many new purchase orders have we received since noon?

Enterprise Data Warehouse

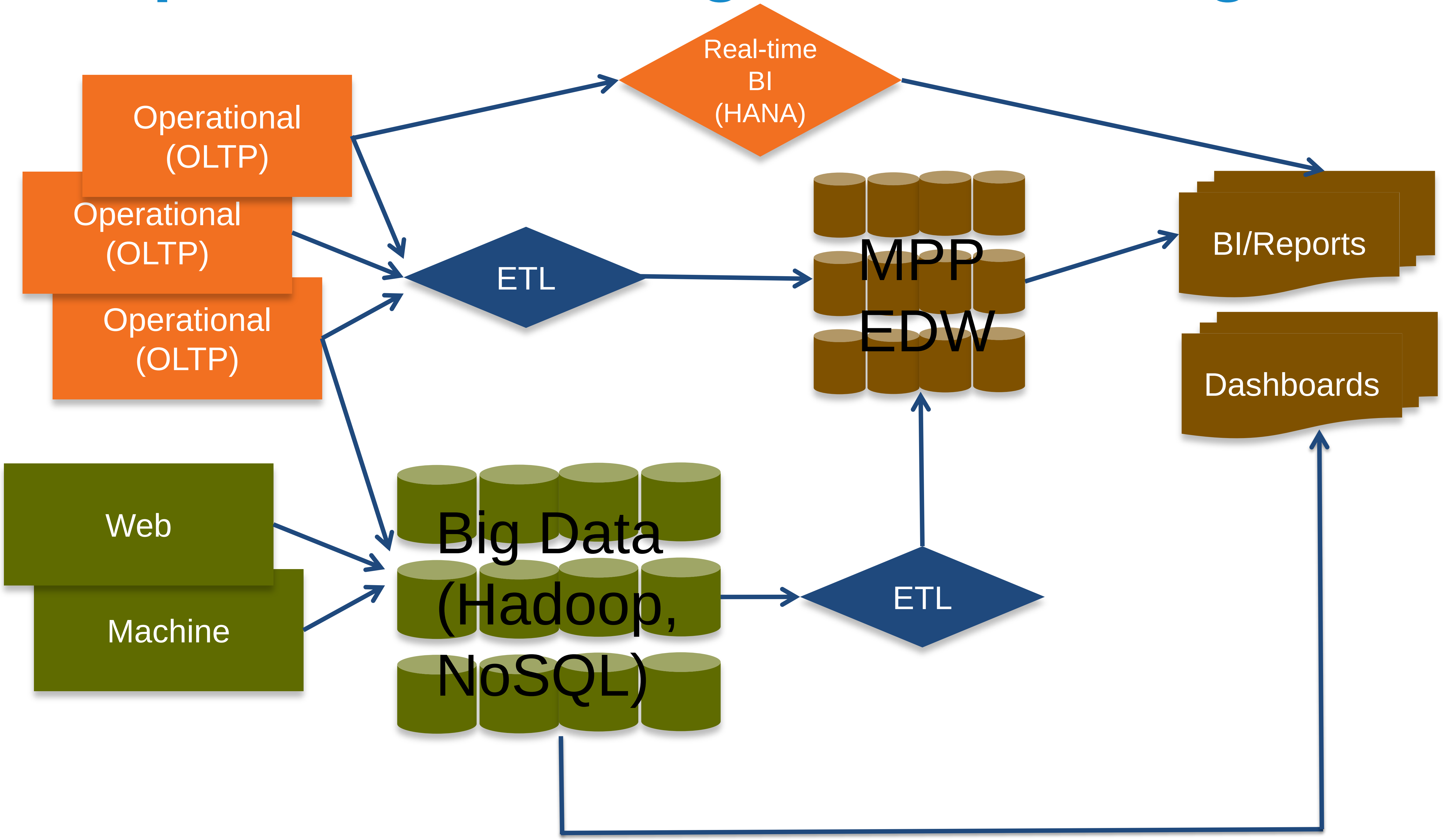
High value, structured, indexed, cleansed

- How many more hurricane windows are sold in Gulf-area stores during hurricane season vs. the rest of the year?
- What were the top 10 most frequently back-ordered products over the past year?

Data Sources



Enterprise Data Management with Big Data



Traditional Business Intelligence Questions

Transactional Data (e.g. OLTP, now SAP HANA)

Fast data, real-time

- What are the top 5 most active stocks traded in the last hour?
- How many new purchase orders have we received since noon?
- With HANA: How to optimally route delivery trucks based on morning sales and stock levels?

Enterprise Data Warehouse

High value, structured, indexed, cleansed

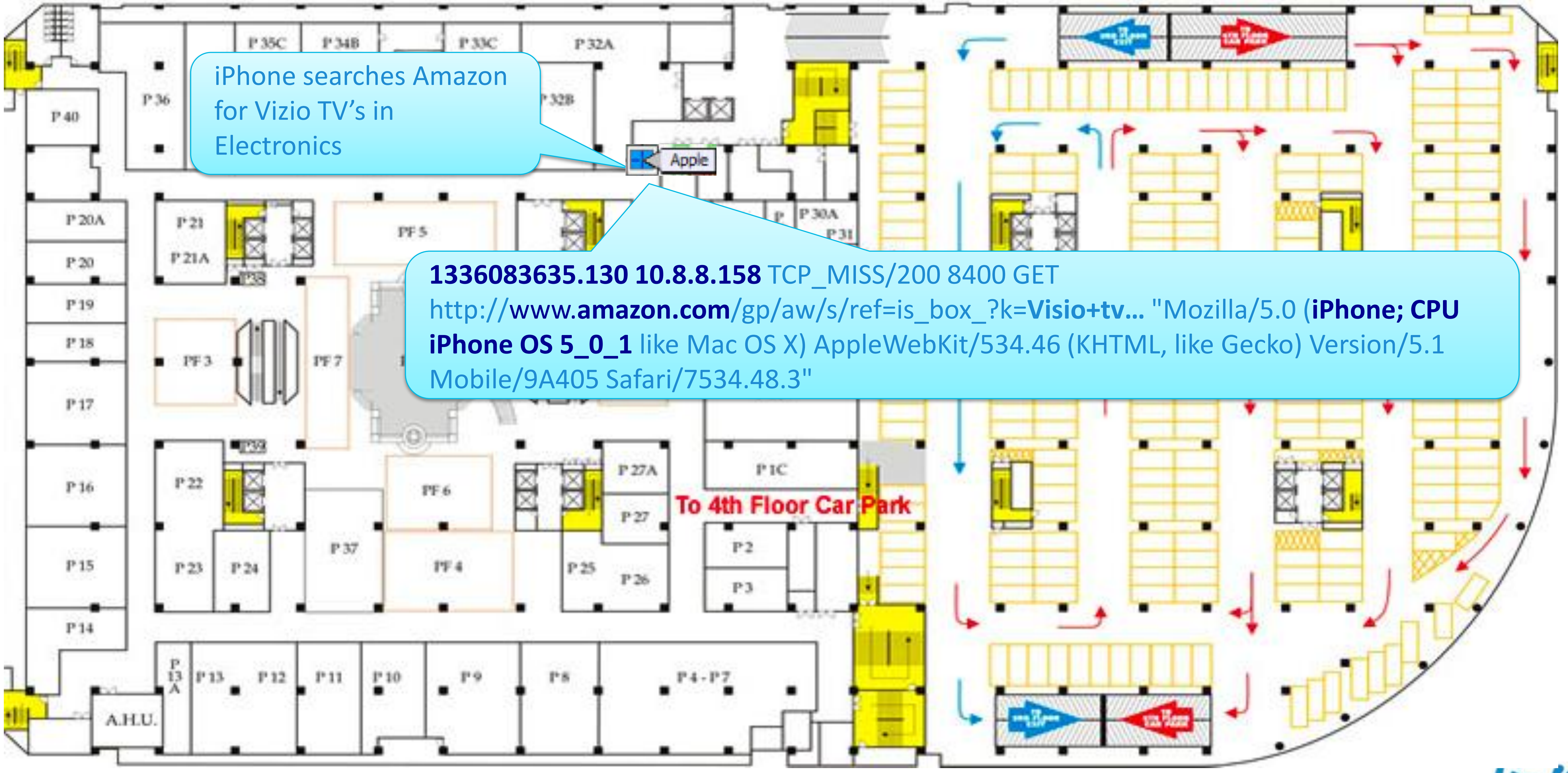
- How many more hurricane windows are sold in Gulf-area stores during hurricane season vs. the rest of the year?
- What were the top 10 most frequently back-ordered products over the past year?

Big Data

Lower value, semi-structured, multi-source, raw/"dirty"

- Which products do customers click on the most and/or spend the most time browsing *without* buying?
- How do we optimally set pricing for each product in each store for *individual customers* everyday?
- Did the recent marketing launch generate the expected online buzz, and did that translate to sales?

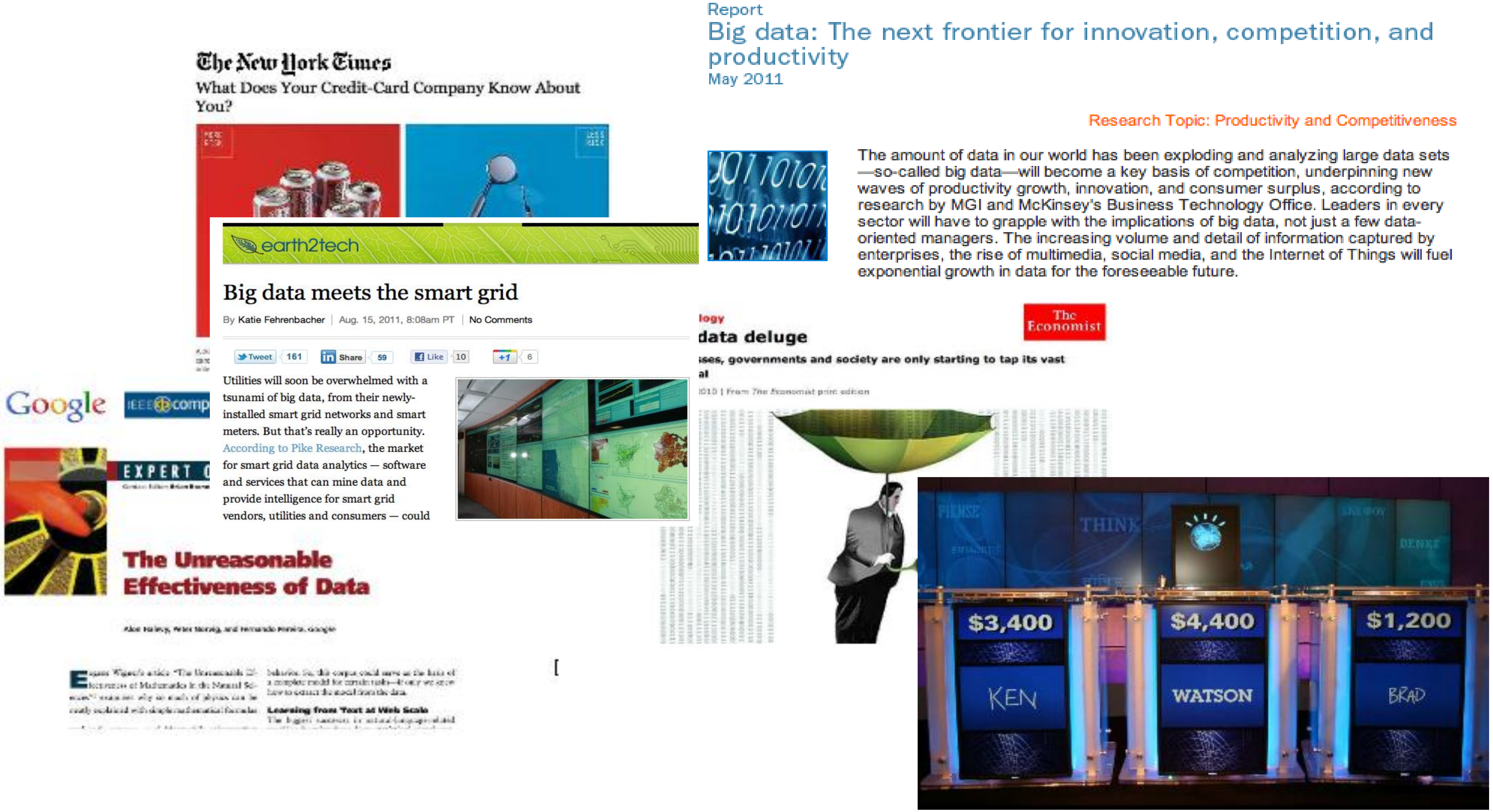
Example: Web and Location Analytics



iPhone searches Amazon for Vizio TV's in Electronics

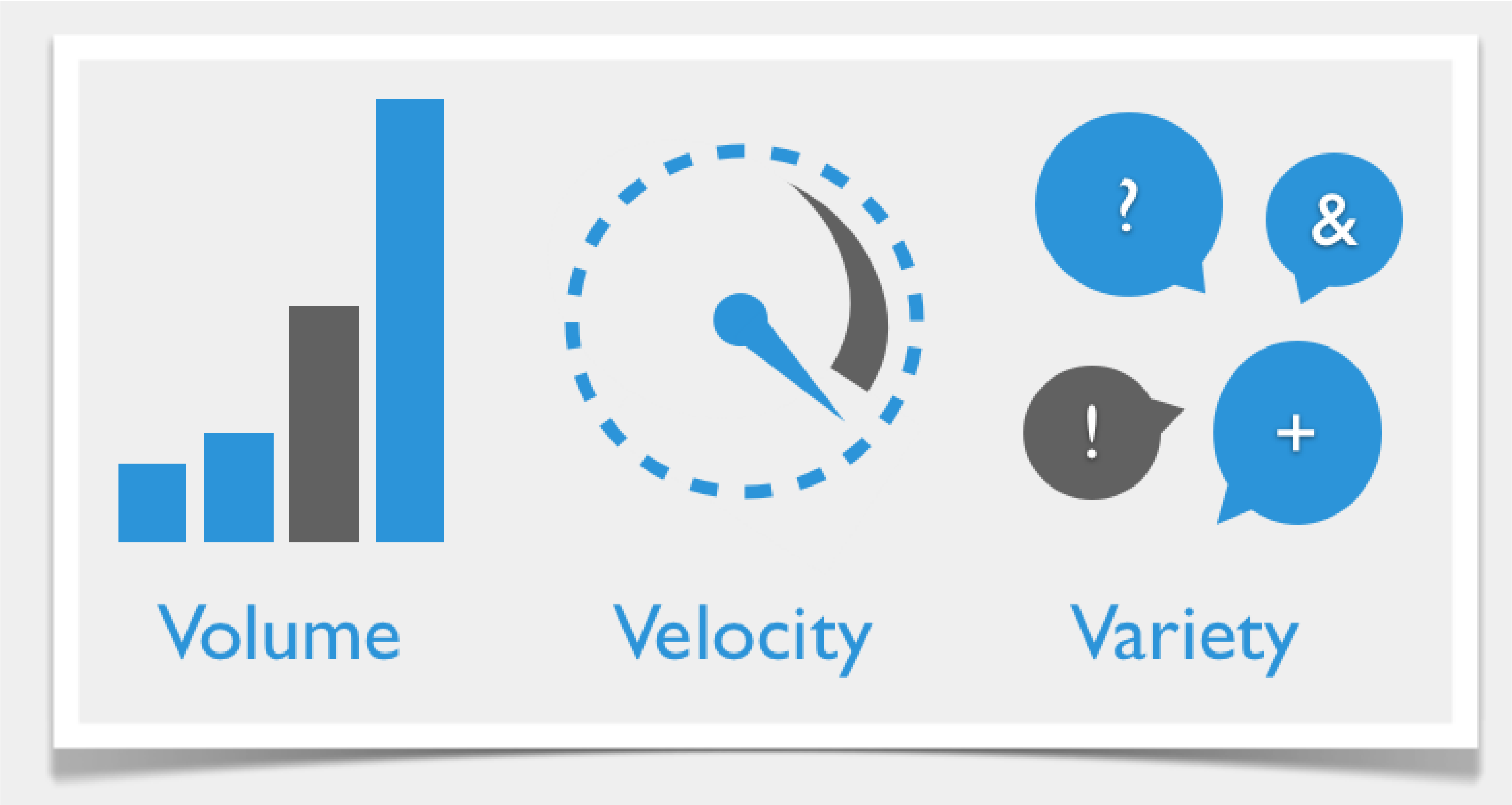
1336083635.130 10.8.8.158 TCP_MISS/200 8400 GET http://**www.amazon.com**/gp/aw/s/ref=is_box_?k=**Visio+tv…** "Mozilla/5.0 (**iPhone; CPU iPhone OS 5_0_1** like Mac OS X) AppleWebKit/534.46 (KHTML, like Gecko) Version/5.1 Mobile/9A405 Safari/7534.48.3"

Big data is top of mind



How Do We Define Big Data?

For our purposes, big data means wide scale-out rack-mount, bare-metal, DAS-centric, spindle-dense distributed computing architectures aimed at the “3 V’s” of data: Volume, Velocity, Variety

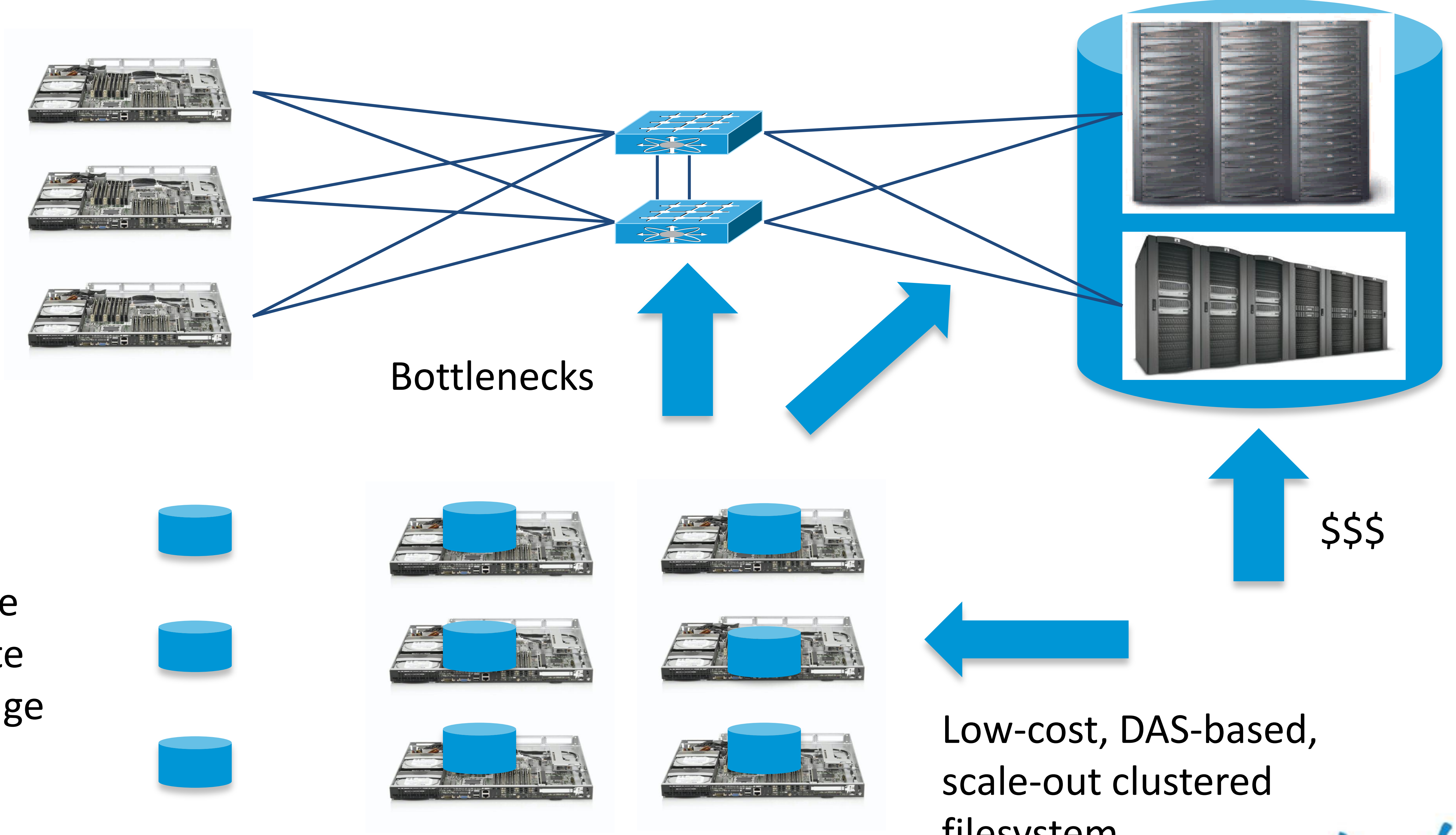


Is It Big Data?

Yes	No	Borderline, but still no
• Hadoop/MapR • Oracle NoSQL • Greenplum DB • ParAccel • Cassandra • HBase	• Classic RDBMS (even if very large) data warehouses • Oracle Exalytics • Actian • Vblock, Flexpod	• Teradata • Oracle Exadata • HPC • SAP HANA (fast data, but not big data)

Classic NAS/SAN vs. new scale-out DAS

Traditional – separate compute from storage



Bottlenecks

$$$

New – move the compute to the storage

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

The Economics of DAS

SAN Storage

$2 - $10/Gigabyte

$1M gets:
0.5Petabytes
200,000 IOPS
1Gbyte/sec

NAS Filers

$1 - $5/Gigabyte

$1M gets:
1 Petabyte
400,000 IOPS
2Gbyte/sec

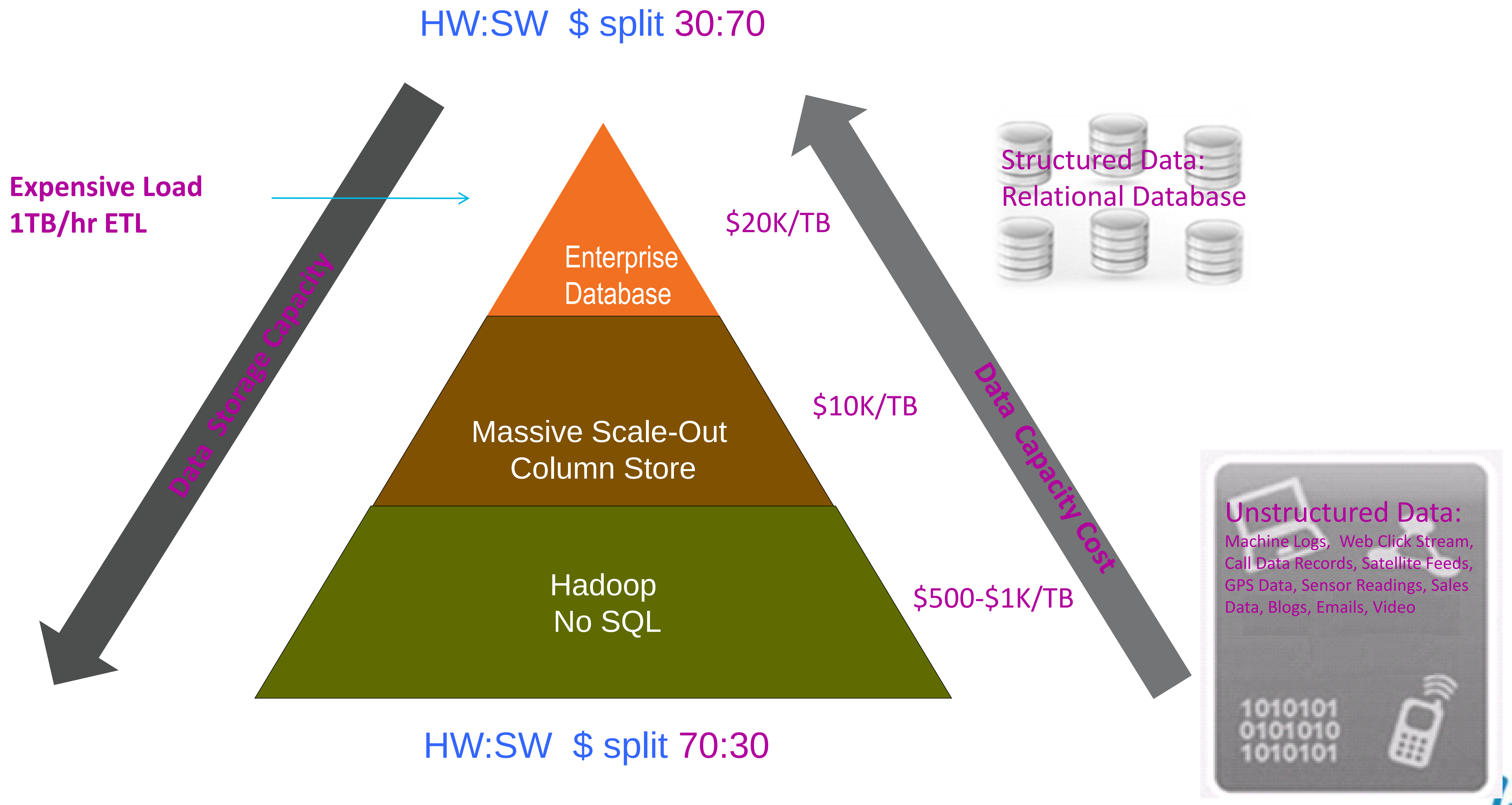
Local Storage

$0.10/Gigabyte

$1M gets:
10 Petabytes
800,000 IOPS
800 Gbytes/sec

Source: VMWare Strata Conference

Cost, Performance, and Capacity



Three Basic Categories of Big Data Architectures

MPP Relational Columnar Database

"Scale-out BI/DW"

- Structured Data
- Optimised for DW/OLAP, some OLTP (ACID-compliant)
- Data stored via frequently-access columns rather than rows for faster retrieval
- Rigid schema applied to data on insert/update
- Read and write (insert, update) many times
- Somewhat limited linear scaling
- Queries often involve a smaller subset of data set vs. Hadoop
- TB to low PB size

Three Basic Categories of Big Data Architectures

MPP Relational Columnar Database

“Scale-out BI/DW”

- Structured Data
- Optimised for DW/OLAP, some OLTP (ACID-compliant)
- Data stored via frequently-access columns rather than rows for faster retrieval
- Rigid schema applied to data on insert/update
- Read and write (insert, update) many times
- Somewhat limited linear scaling
- Queries often involve a smaller subset of data set vs. Hadoop
- TB to low PB size

Batch-oriented Hadoop

Heavy lifting, processing

- Unstructured Data – emails, syslogs, health and science raw data
- Optimised for large streaming reads of large blocks (128-256 MB) comprising large files
- Dynamic schema effectively applied on read
- Optimised to compute data locally at cost of normalisation
- Write once, read many
- Linear scaling to thousands of nodes and tens of PB
- Entire data set at play for a given query

Three Basic Categories of Big Data Architectures

MPP Relational Columnar Database

"Scale-out BI/DW"

- Structured Data
- Optimised for DW/OLAP, some OLTP (ACID-compliant)
- Data stored via frequently-access columns rather than rows for faster retrieval
- Rigid schema applied to data on insert/update
- Read and write (insert, update) many times
- Somewhat limited linear scaling
- Queries often involve a smaller subset of data set vs. Hadoop
- TB to low PB size

Batch-oriented Hadoop

Heavy lifting, processing

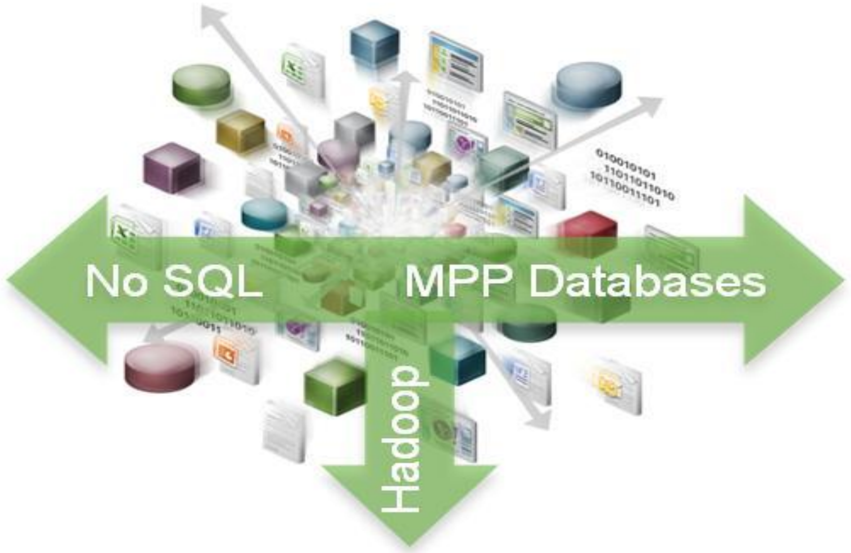
- Unstructured Data – emails, syslogs, health and science raw data
- Optimised for large streaming reads of large blocks (128-256 MB) comprising large files
- Dynamic schema effectively applied on read
- Optimised to compute data locally at cost of normalisation
- Write once, read many
- Linear scaling to thousands of nodes and tens of PB
- Entire data set at play for a given query

Real-time NoSQL

Fast store/retrieve

- Unstructured Data – tweets, sensor data, clickstream
- Data typically stored and retrieved as key-value pairs in flexible column families
- High transaction rates, many reads and writes, small block/chunk sizes (1K-1MB)
- Less well-suited for ad-hoc analysis than Hadoop
- TB to PB scale

Three Basic Big Data Software Architectures



Real-time NoSQL

Fast key-value store/retrieve

- HBase (part of Apache Hadoop)
- Cassandra
- Oracle NoSQL*
- Amazon Dynamo

Batch-oriented Hadoop

Heavy lifting, processing

- Cloudera*
- MapR*
- HortonWorks*
- IBM BigInsights

MPP Relational Database

Scale-out BI/DW

- Greenplum DB (EMC)*
- ParAccel*
- Vertica (HP)
- Netezza (IBM)
- Teradata
- Arguably Exadata (Oracle)

*Cisco Partners

Hadoop Deep Dive

What Is Hadoop?

- 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.

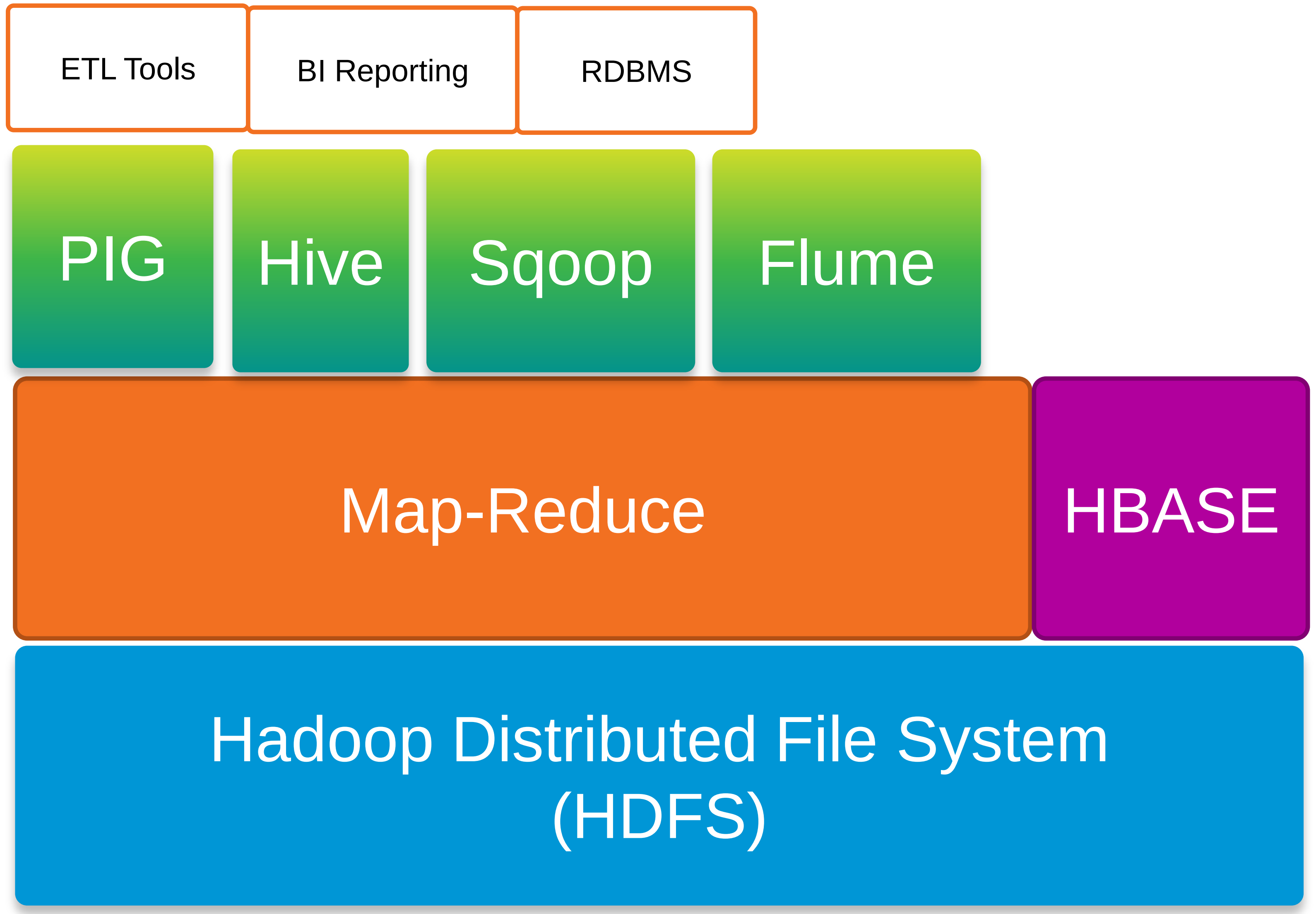
Main Hadoop Building Blocks

Top level Interfaces

Mid-level abstractions

Distributed Data Processing

At the base is a self-healing clustered storage system



Database with Real-time access

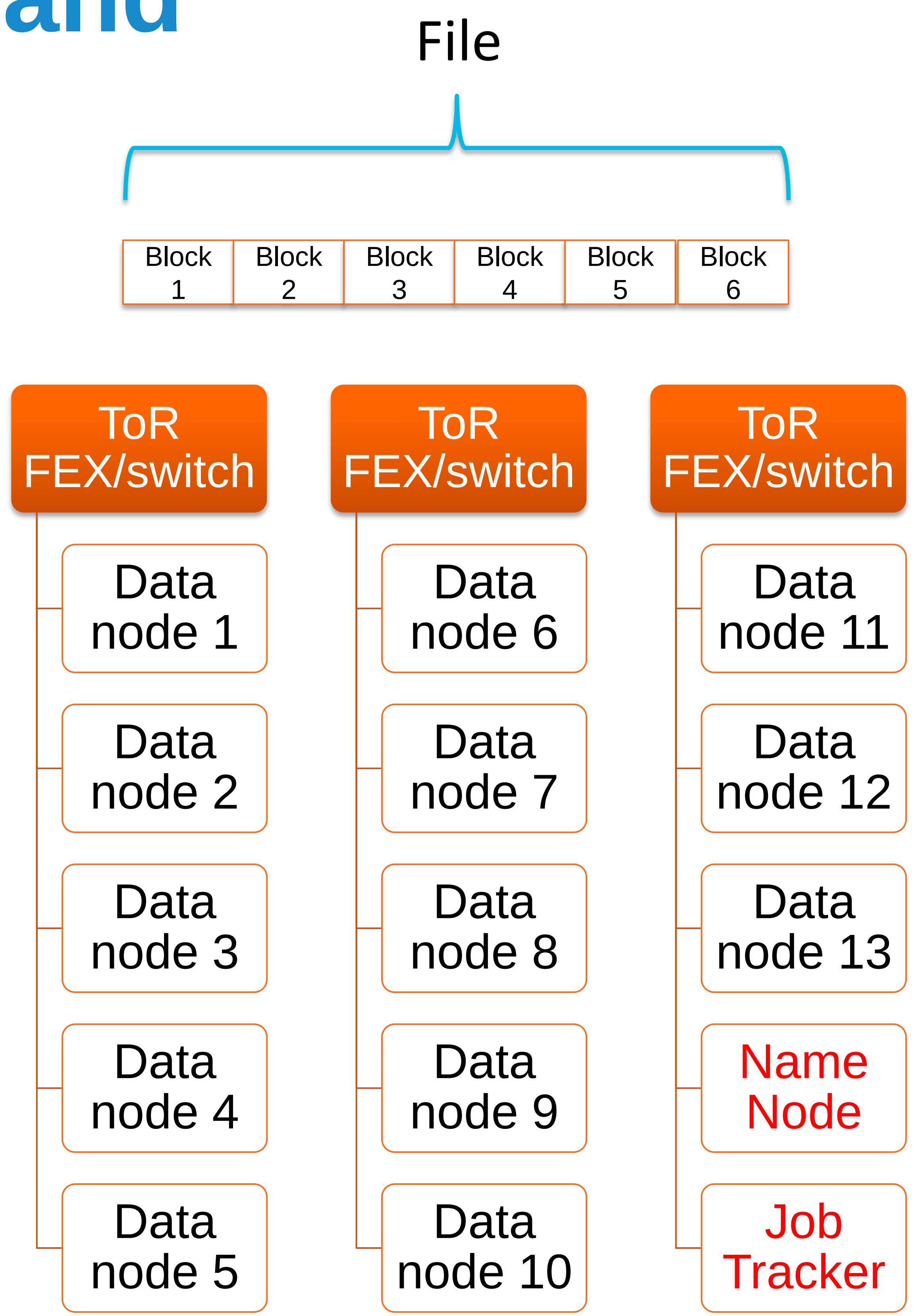
“Failure is the defining difference between distributed and local programming”

- Ken Arnold, CORBA designer

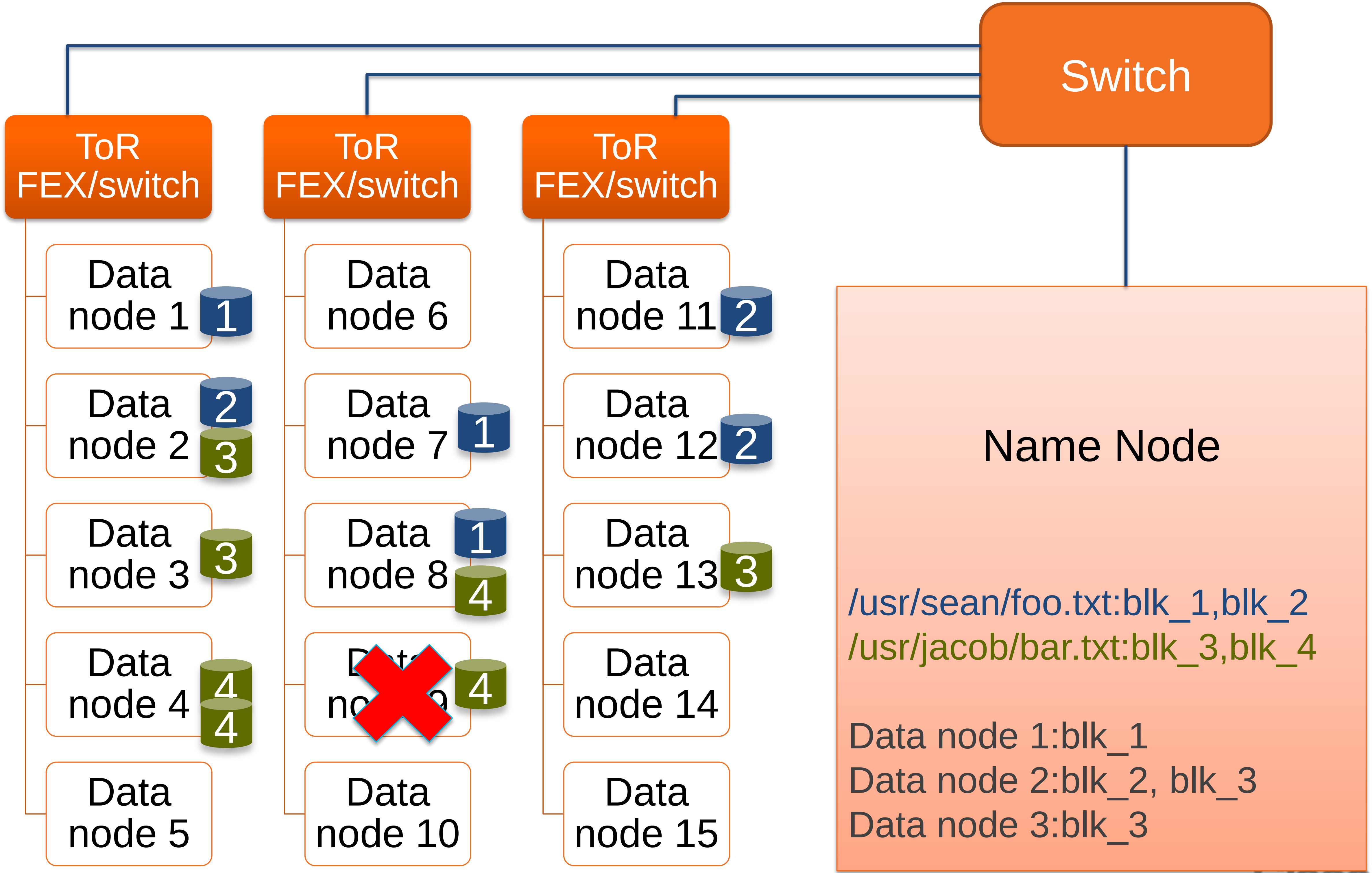
Hadoop Components and Operations

Hadoop Distributed File System

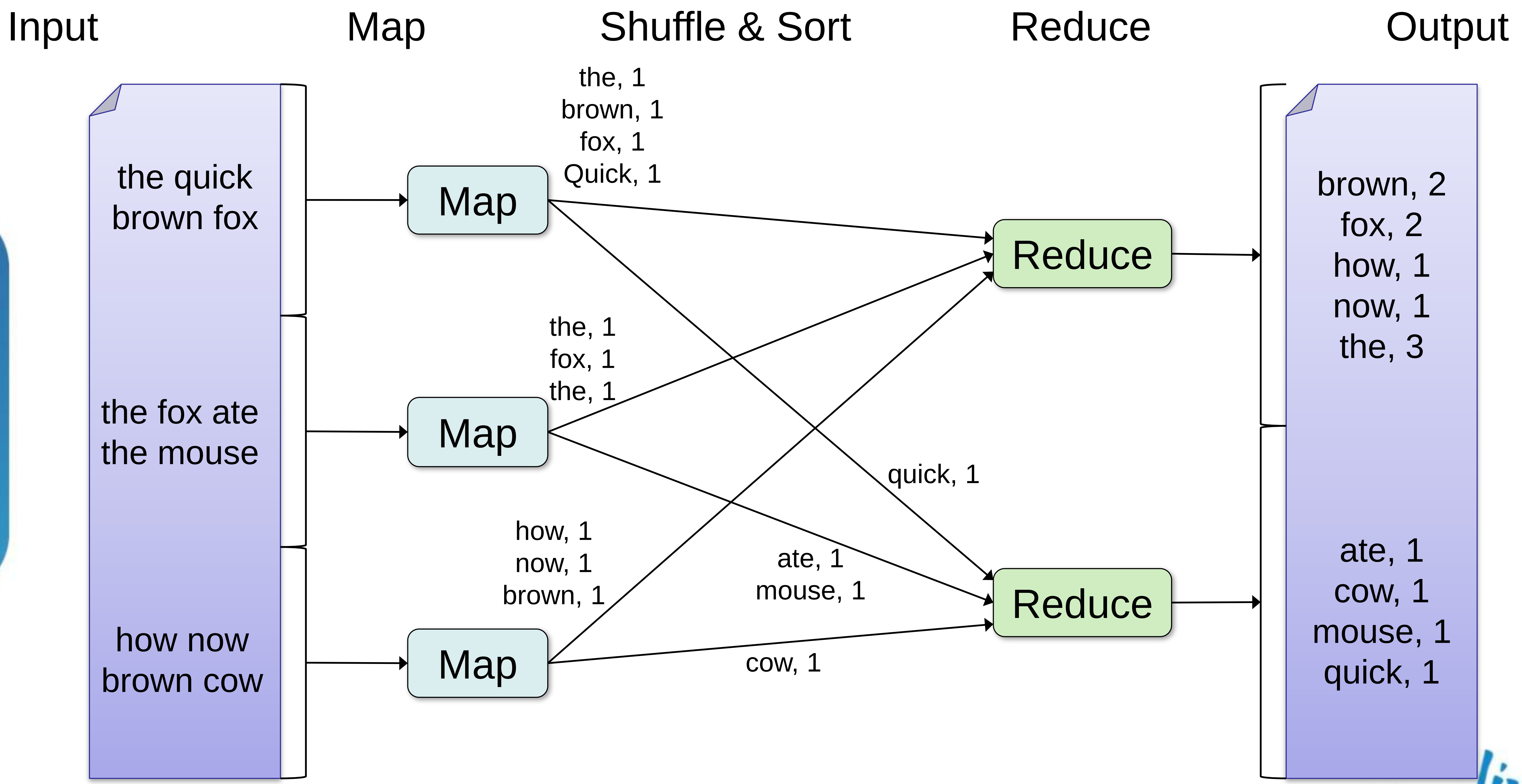
- Scalable & Fault Tolerant
- Filesystem is distributed, stored across all data nodes in the cluster
- Files are divided into multiple **large blocks** – 64MB default, typically 128MB – 512MB
- Data **is stored reliably.** Each block is replicated 3 times by default
- Types of Node Functions
  - Name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



HDFS Architecture



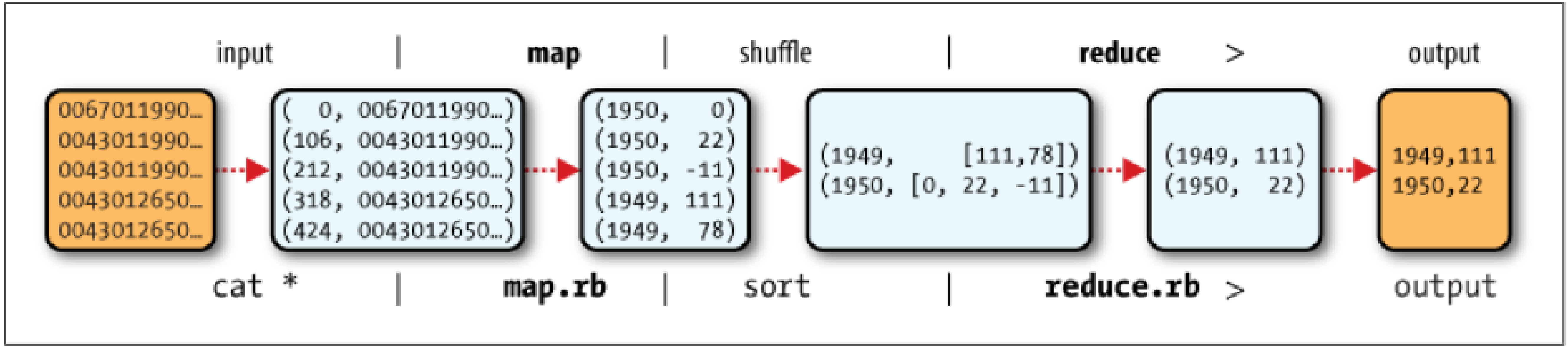
MapReduce Example: Word Count



MapReduce Example: Max Temperature

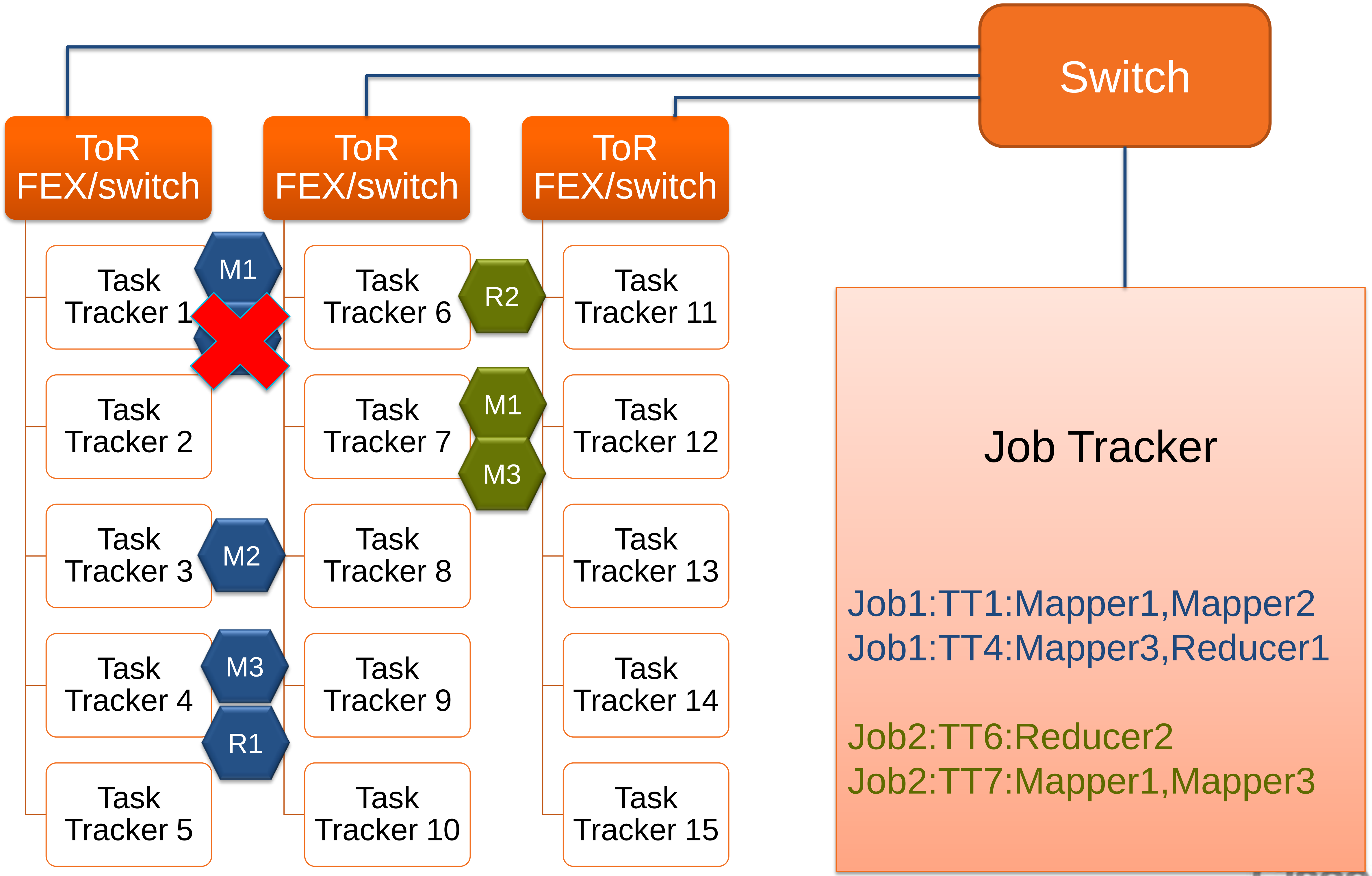
Example:
Historic Weather Data (max temperatures/Year)

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



Source: O'Reilly Hadoop: A Definitive Guide

MapReduce Architecture



Hadoop on the Network

Hadoop Network Traffic Types

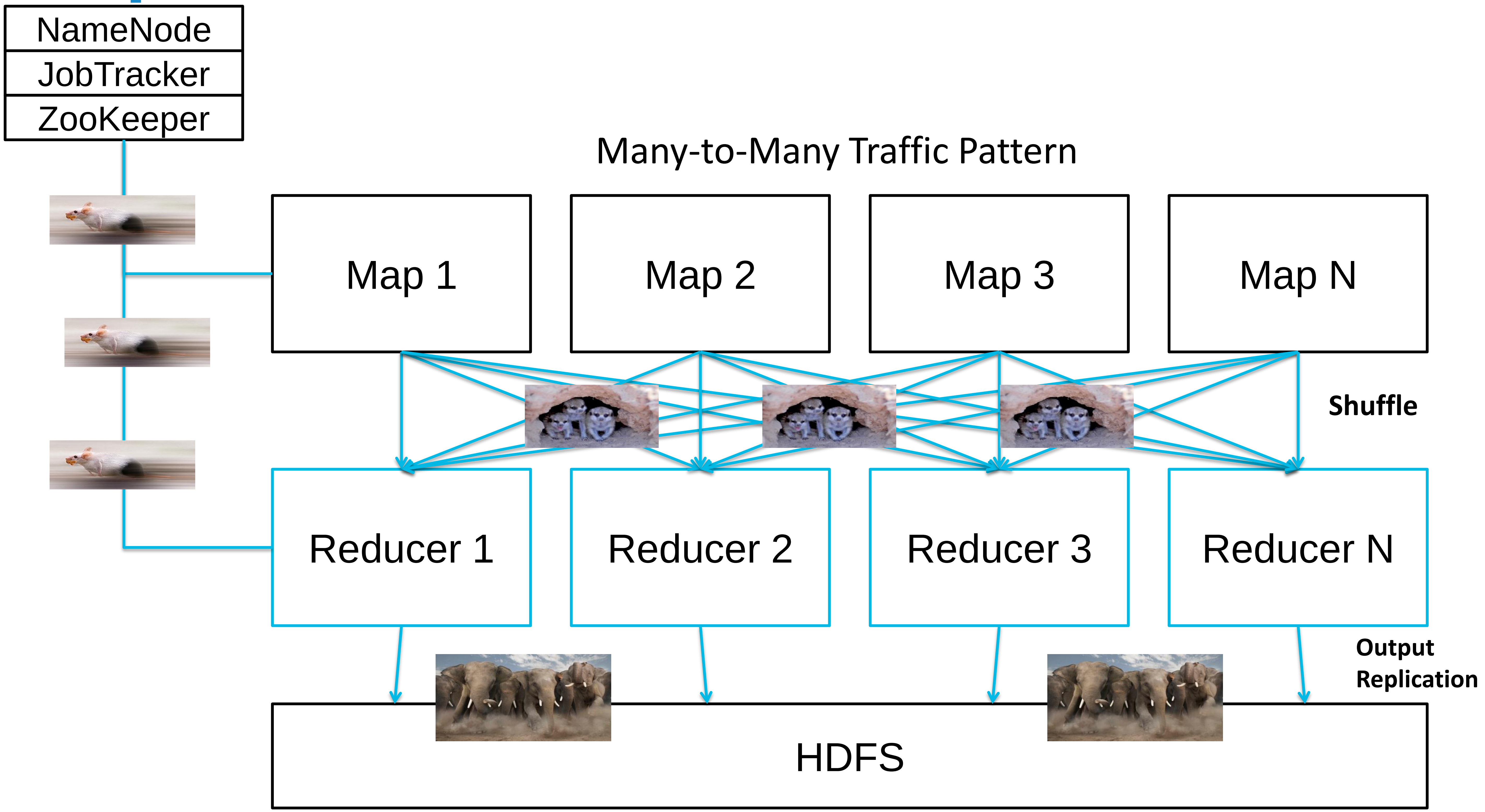
Small Flows/Messaging
(Admin Related, Heart-beats, Keep-alive, delay sensitive application messaging)

Small – Medium Incast
(Hadoop Shuffle)

Large Flows
(HDFS egress)

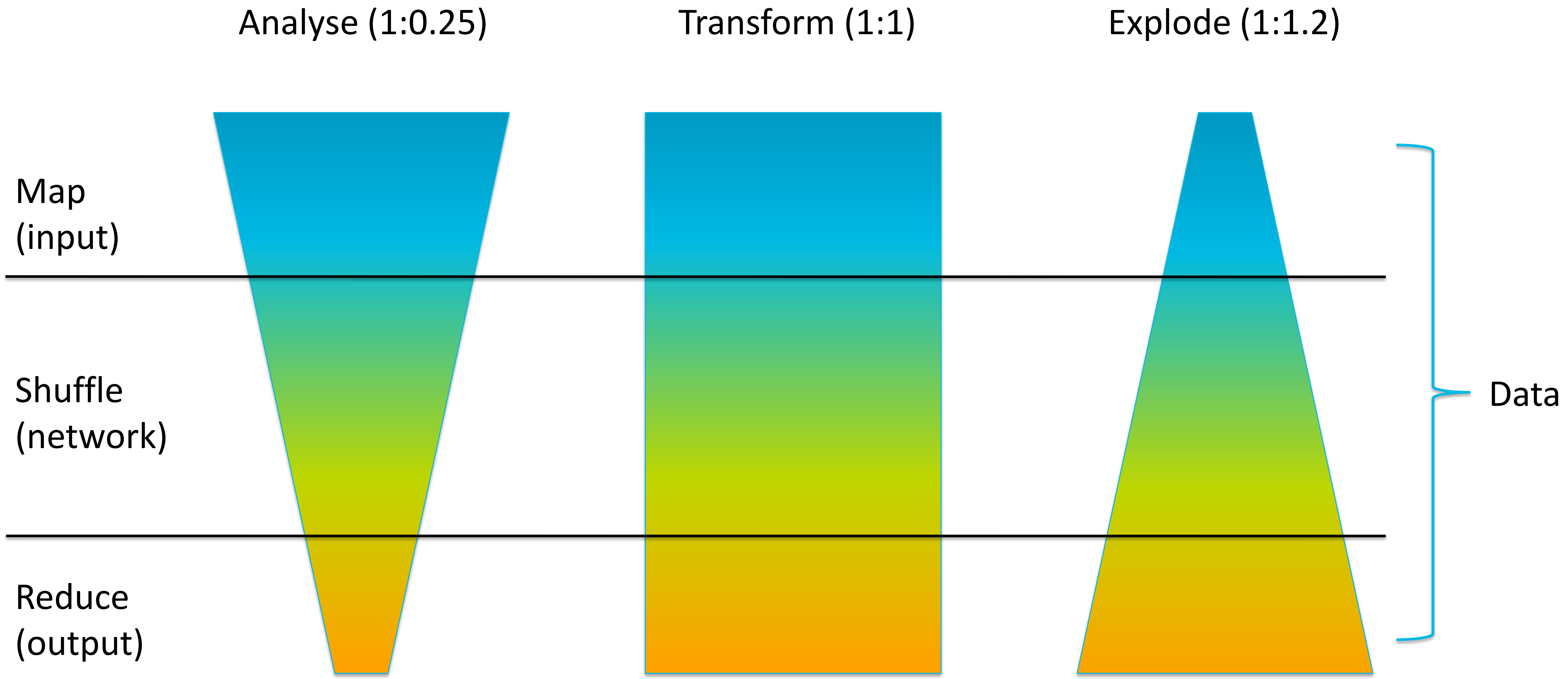
Large Incast
(Hadoop Replication)

Map and Reduce Traffic



Typical Hadoop Job Patterns

Different workloads can have widely varying network impact

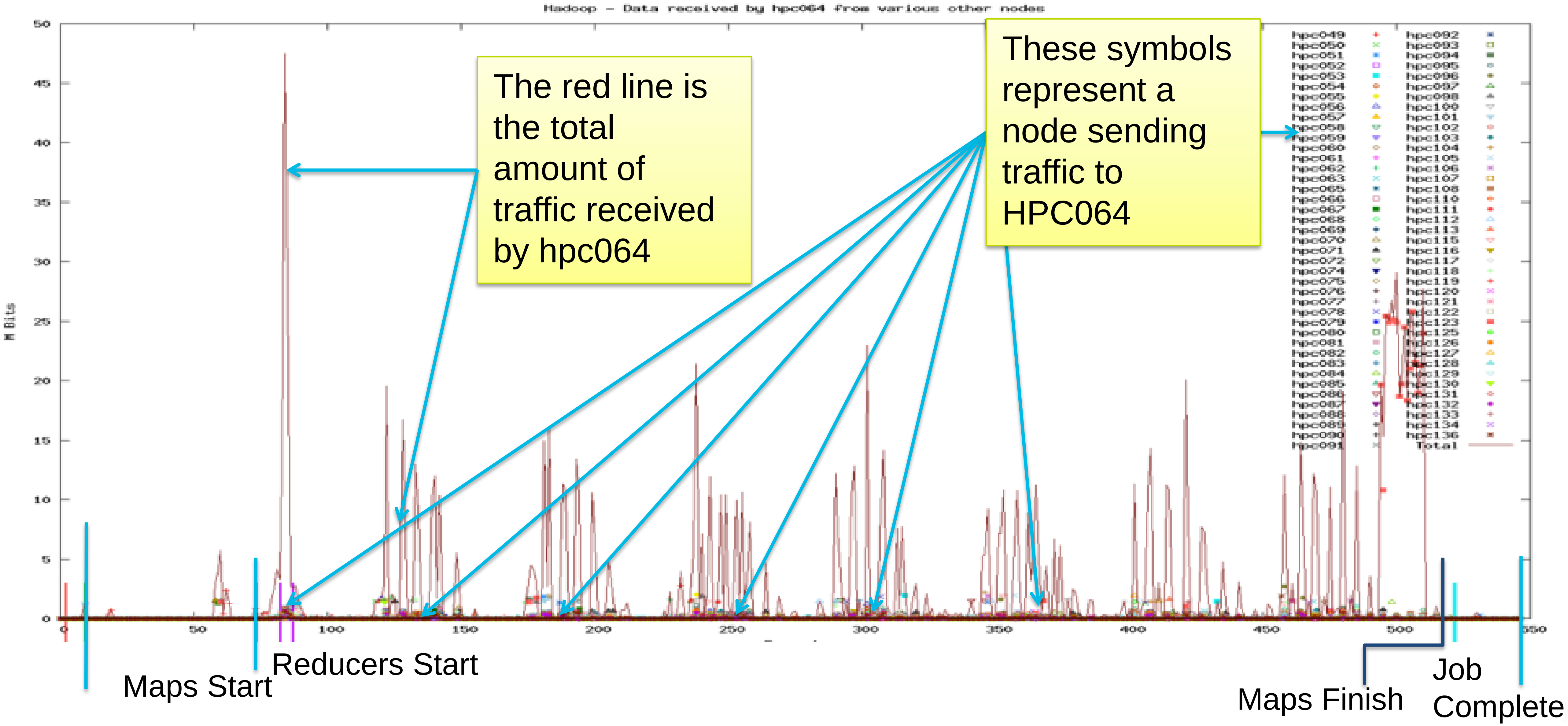


Analyse Workload

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

Wordcount on 200K Copies of complete works of Shakespeare

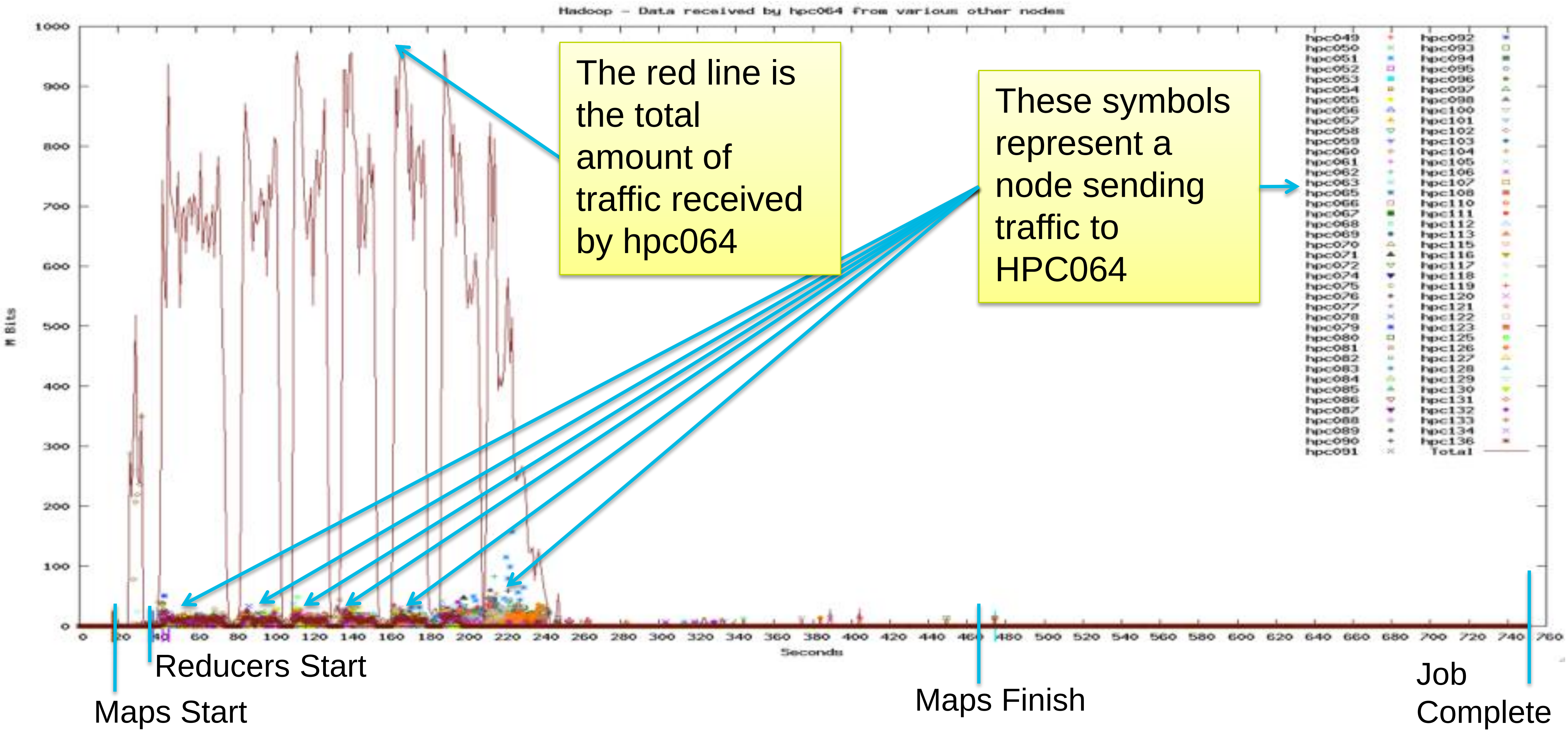
Note:
Due the combination of the length of the Map phase and the reduced data set being shuffled, the network is being utilised throughout the job, but by a limited amount.



Transform Workload (1TB Terasort)

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

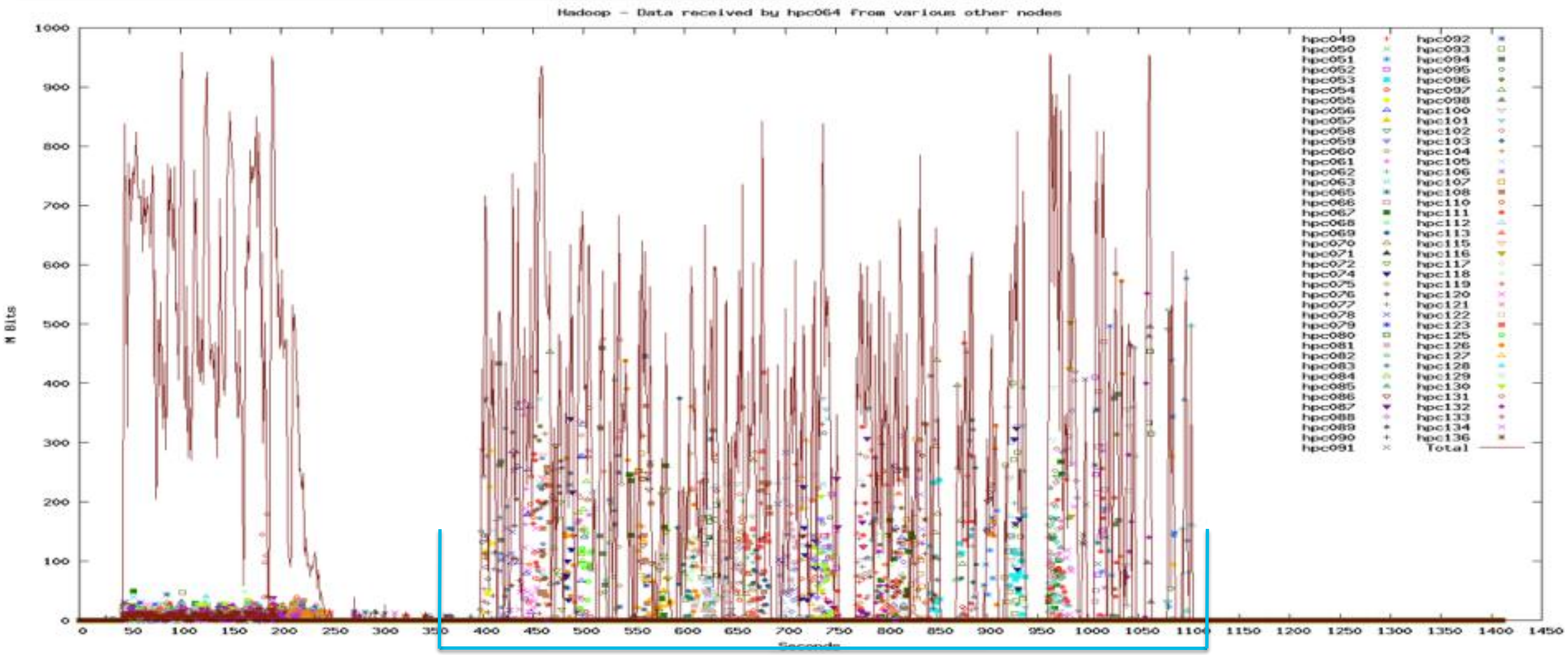
Note:
Shortly after the Reducers start Map tasks are finishing and data is being shuffled to reducers
As Maps completely finish the network is no loner used as Reducers have all the data they need to finish the job



Transform Workload
(1TB Terasort with output replication)

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

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

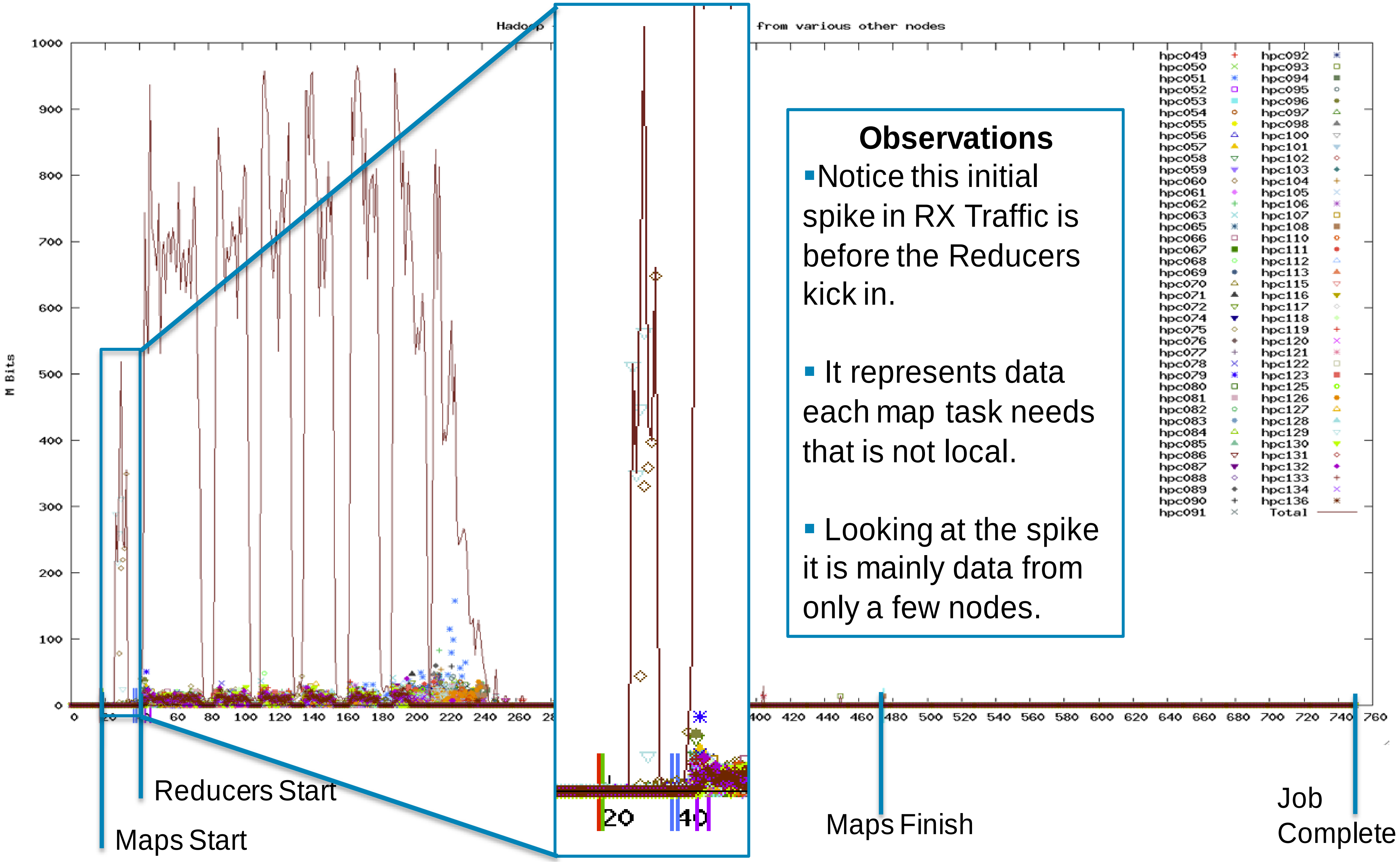


Output Data Replication Enabled

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

Data Locality in Hadoop

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



Observations

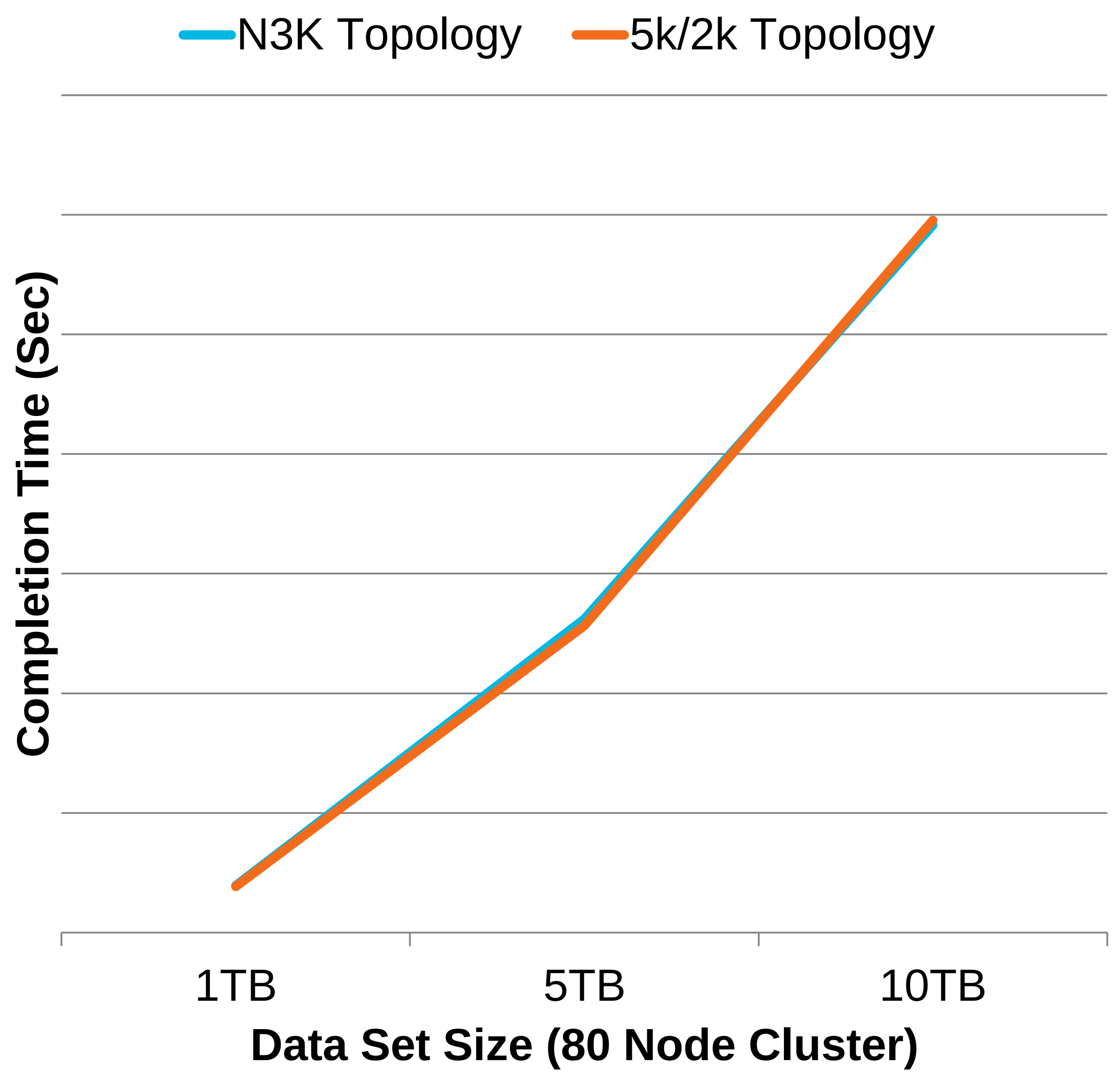
- Notice this initial spike in RX Traffic is before the Reducers kick in.
- It represents data each map task needs that is not local.
- Looking at the spike it is mainly data from only a few nodes.

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.

Network Latency

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

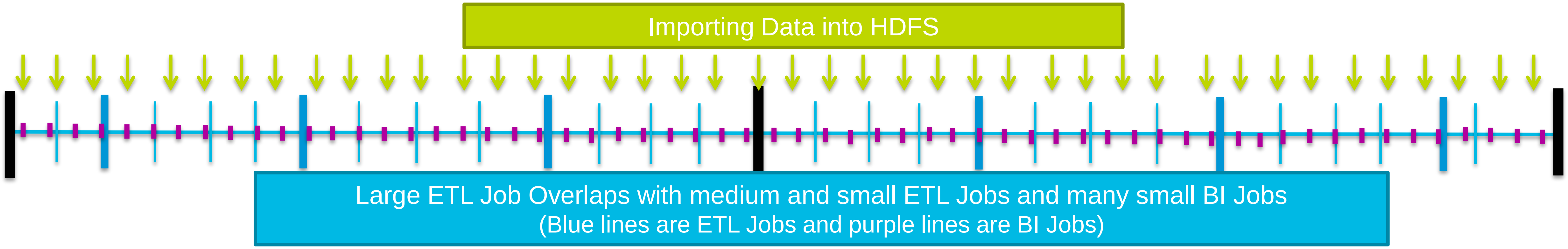
Note:
There is a difference in network latency vs. application latency. Optimisation in the application stack can decrease application latency that can potentially have a significant benefit.



Multi-use Cluster Characteristics

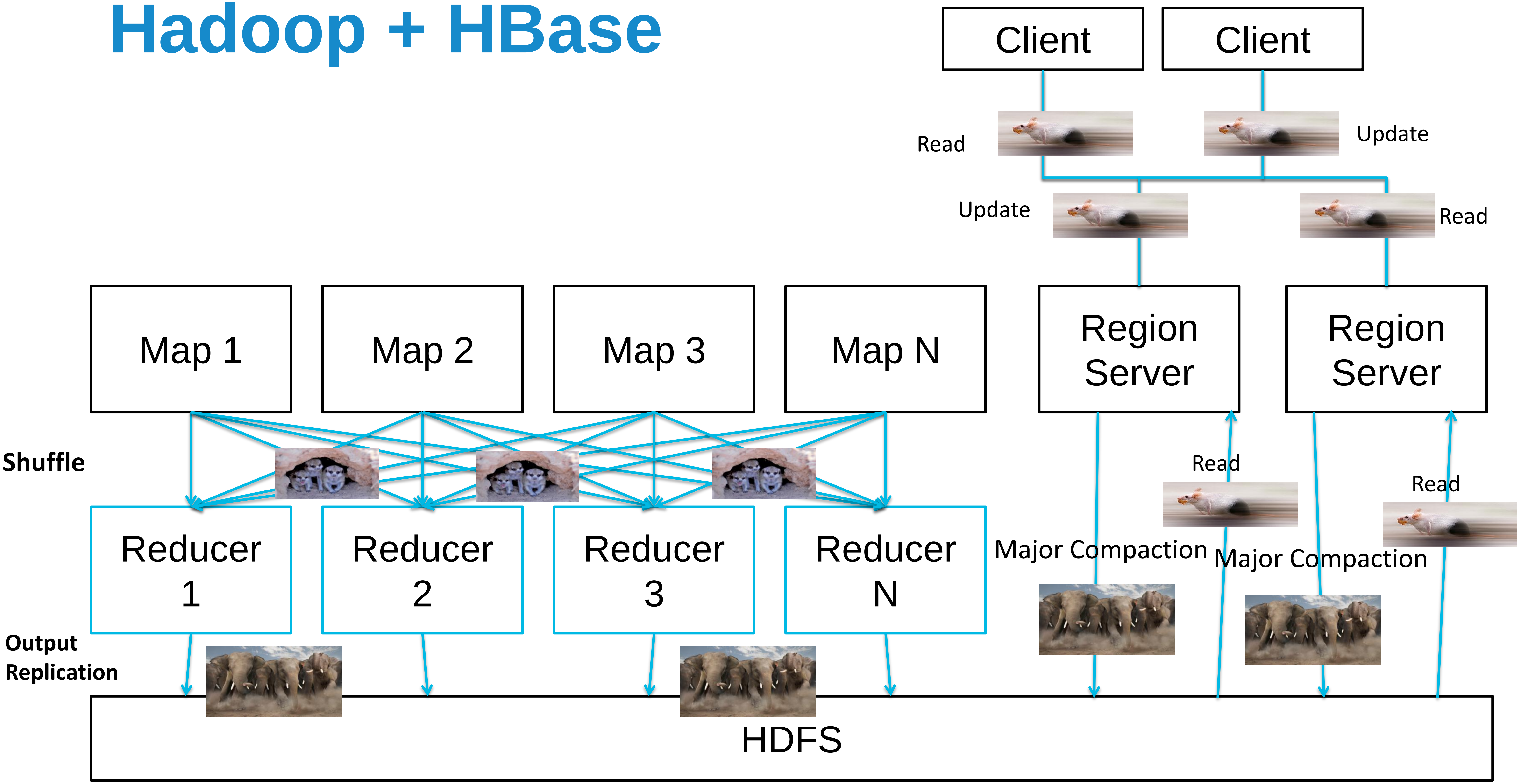
Hadoop clusters are generally multi-use. The effect of background use can affect any single job's completion time

Note:
A given cluster, running many different types of jobs, importing into HDFS, etc.

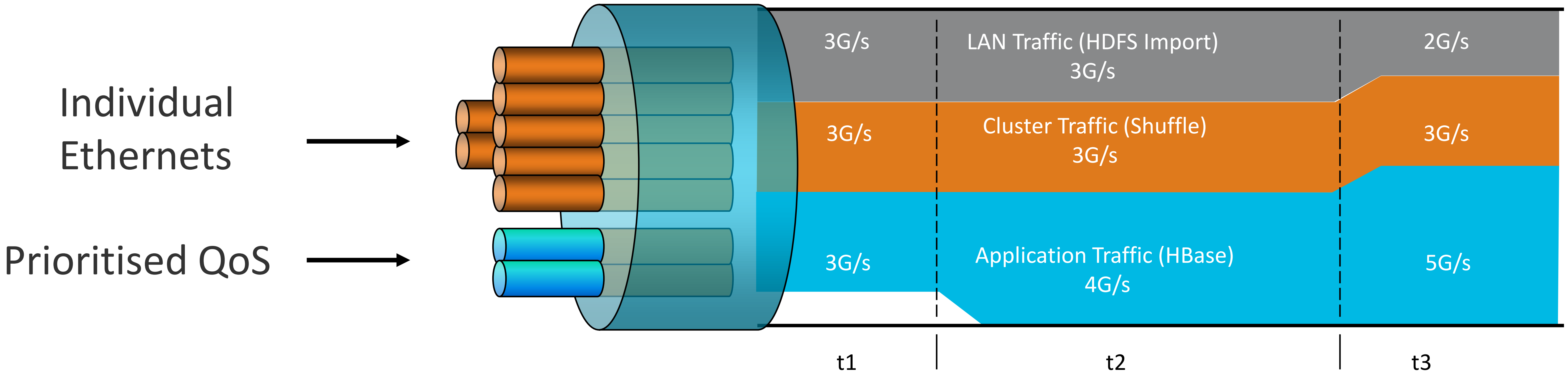


Example View of 24 Hour Cluster Use

Hadoop + HBase

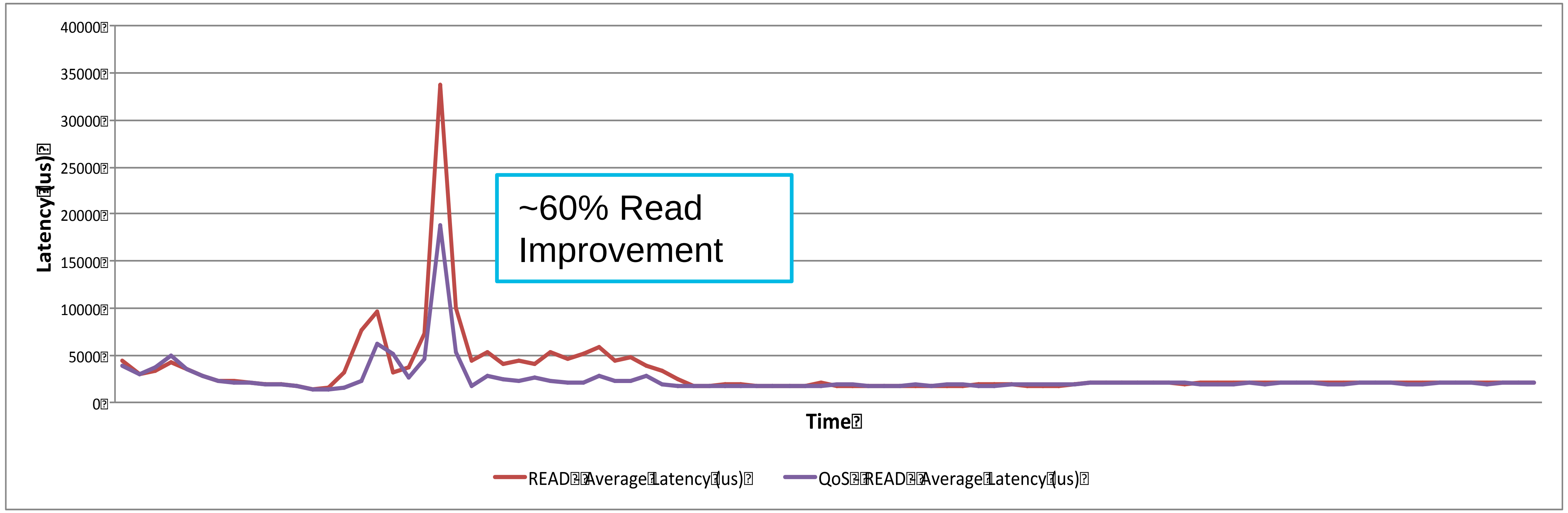


Cisco Unified IO Grant Bandwidth

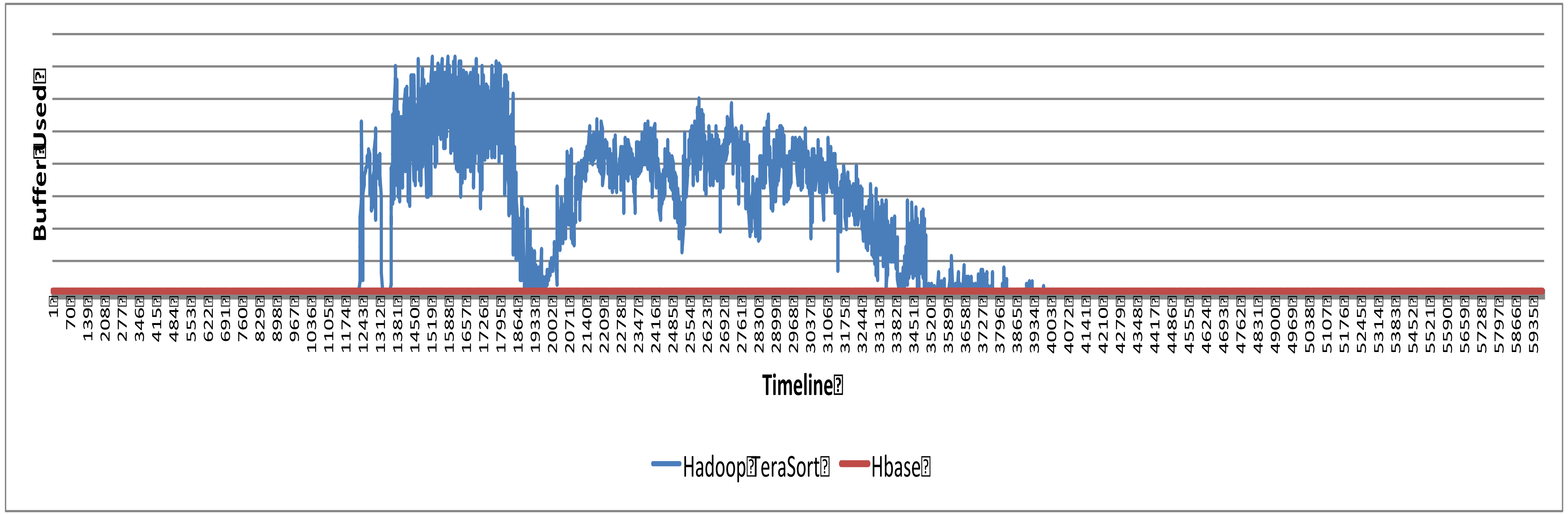


- Near Wire Speed without CPU load
- Dynamic bandwidth management according to SLA's

HBase + Hadoop Map Reduce

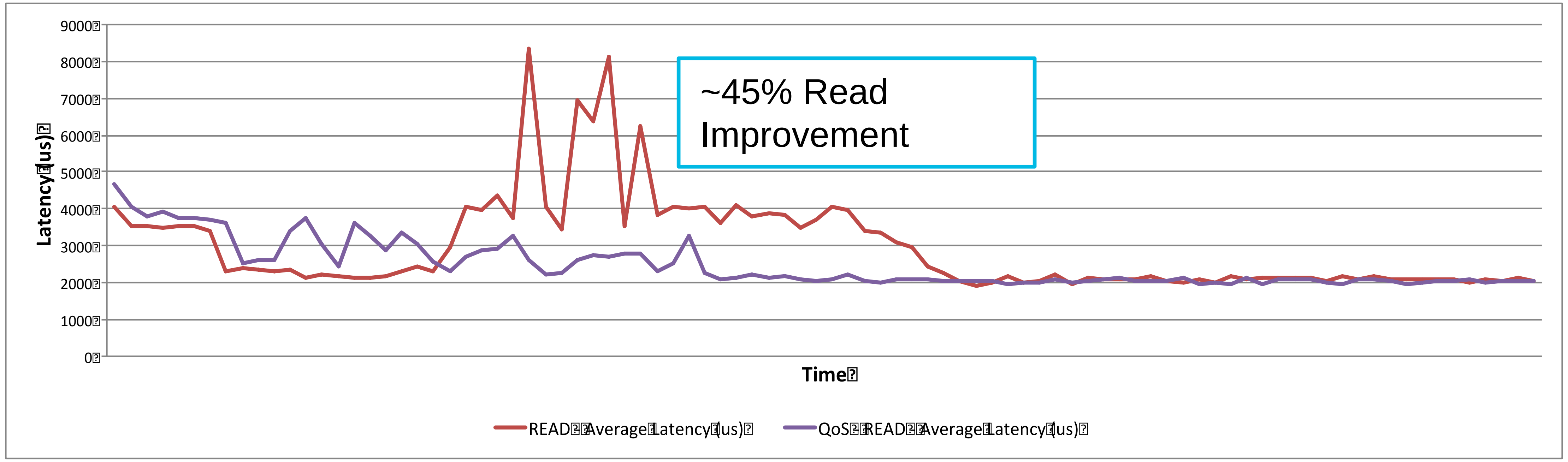


Read Latency
Comparison of Non-QoS vs. QoS Policy

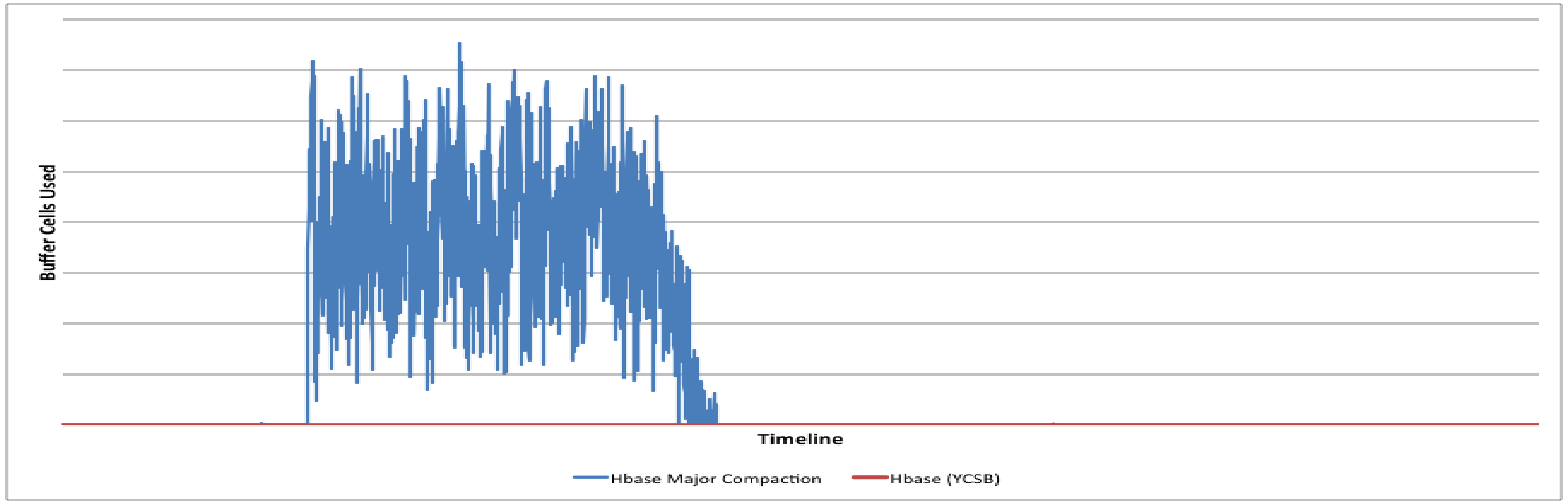


Switch Buffer Usage
With Network QoS Policy to prioritise Hbase Update/Read Operations

HBase During Major Compaction



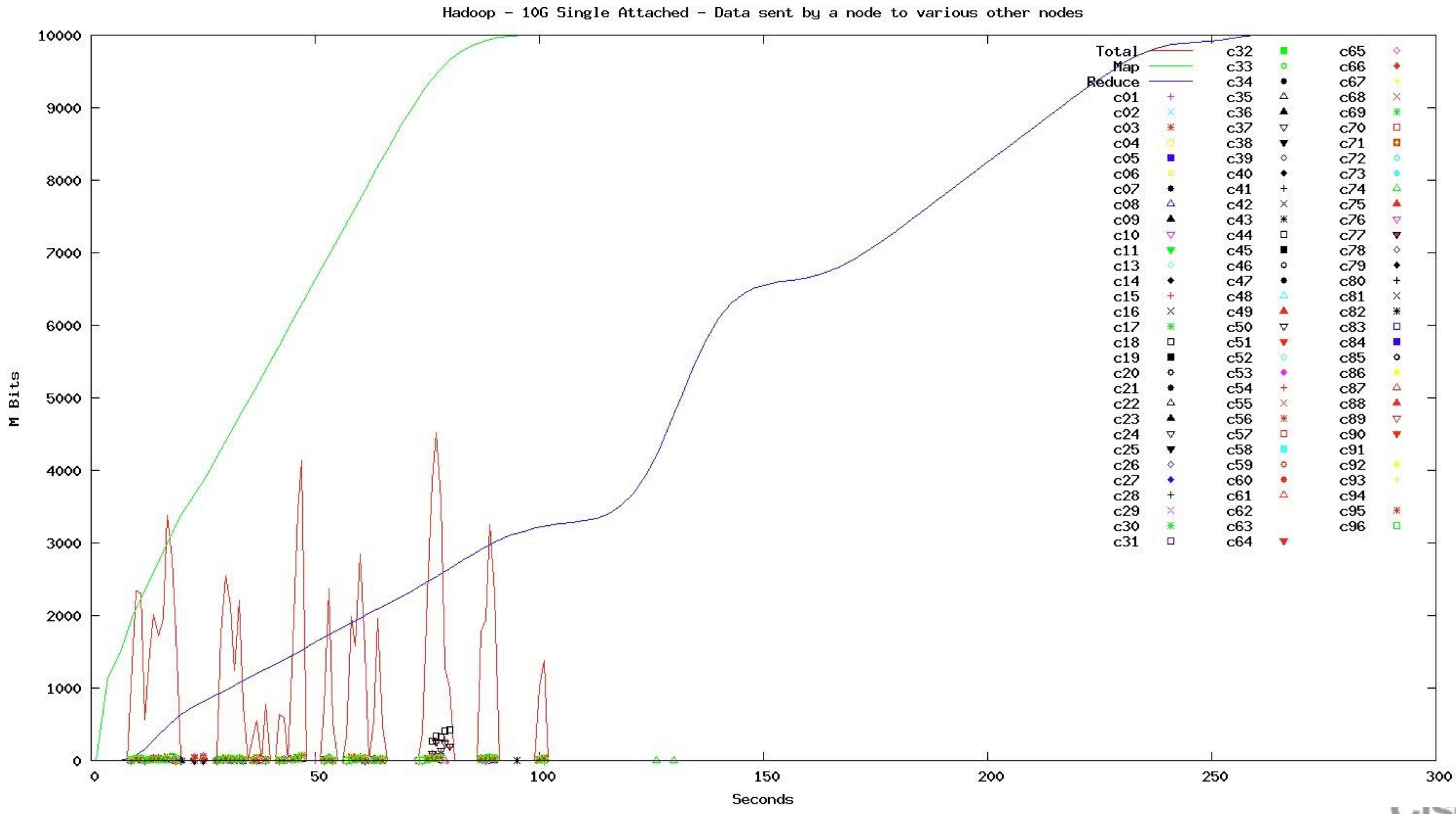
Read Latency
Comparison of Non-QoS vs. QoS Policy



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

10GE Data Node Speed

TCPDUMP of Reducers TX



Hadoop Parameters with Network Impact

- The following parameters can directly affect network utilisation and performance of the cluster
- The defaults and even the commonly accepted values for these settings are often based on the assumption of a 1GE network
- When a faster network is available, their values may need to be reconsidered

Hadoop Parameters with Network Impact

mapred.reduce.slowstart.completed.maps

- Governs when reducers start; goal is to have reducers start early enough to be almost done copying data when last mapper finishes, but not so early that they sit idle.
- More bandwidth allows a later start, increasing cluster efficiency

dfs.balance.bandwidthPerSec

- Specifies how much b/w each data node can consume for HDFS rebalancing; desirable to complete rebalance as fast as possible with minimal impact to running jobs.
- More b/w allows for higher values for faster rebalance

mapred.compress.map.output

- Compresses map task output when spilled to disk, lowering disk and network (shuffle) I/O at the price of CPU cycles
- If cluster is more CPU bound than network- and I/O-bound, can consider leaving this off

Hadoop Parameters with Network Impact

mapred.reduce.parallel.copies

- Controls how many parallel copy processes reducers use to retrieve intermediate map output
- Higher values can improve shuffle times with a fast network

mapred.reduce.tasks and mapred.tasktracker.reduce.tasks.maximum

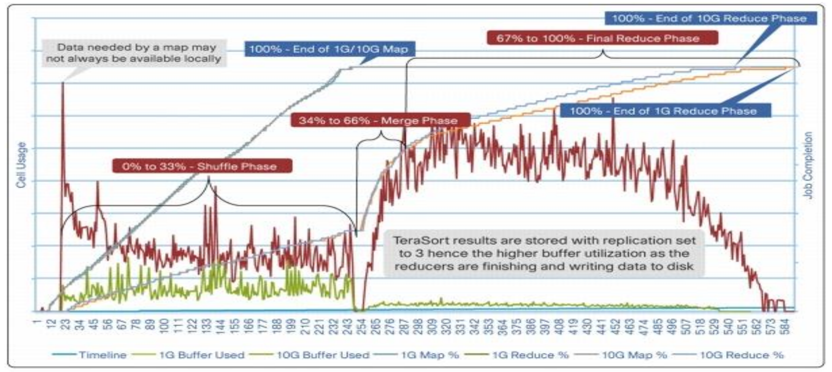
- Controls default/maximum number of slots assigned/available on a worker node for reduce tasks
- Additional reducers can improve job completion time at the cost of increased network utilisation

Nexus Solutions for Big Data

128 Node/1PB test cluster

Certifications with Nexus 5500+22xx

- Cloudera Hadoop Certified Technology
  - Cloudera Hadoop Solution Brief
- Hortonworks Reference architecture
  - Presented at 2012 Hadoop Summit



Multi-month network and compute analysis testing
(In conjunction with Cloudera)

- Network/Compute Considerations Whitepaper
- Presented Analysis at Hadoop World
- Couchbase solution brief

UCS for Big Data

Installing Servers Today

- RAID settings
- Disk scrub actions

- Number of vHBAs
- HBA WWN assignments
- FC Boot Parameters
- HBA firmware

- FC Fabric assignments for HBAs

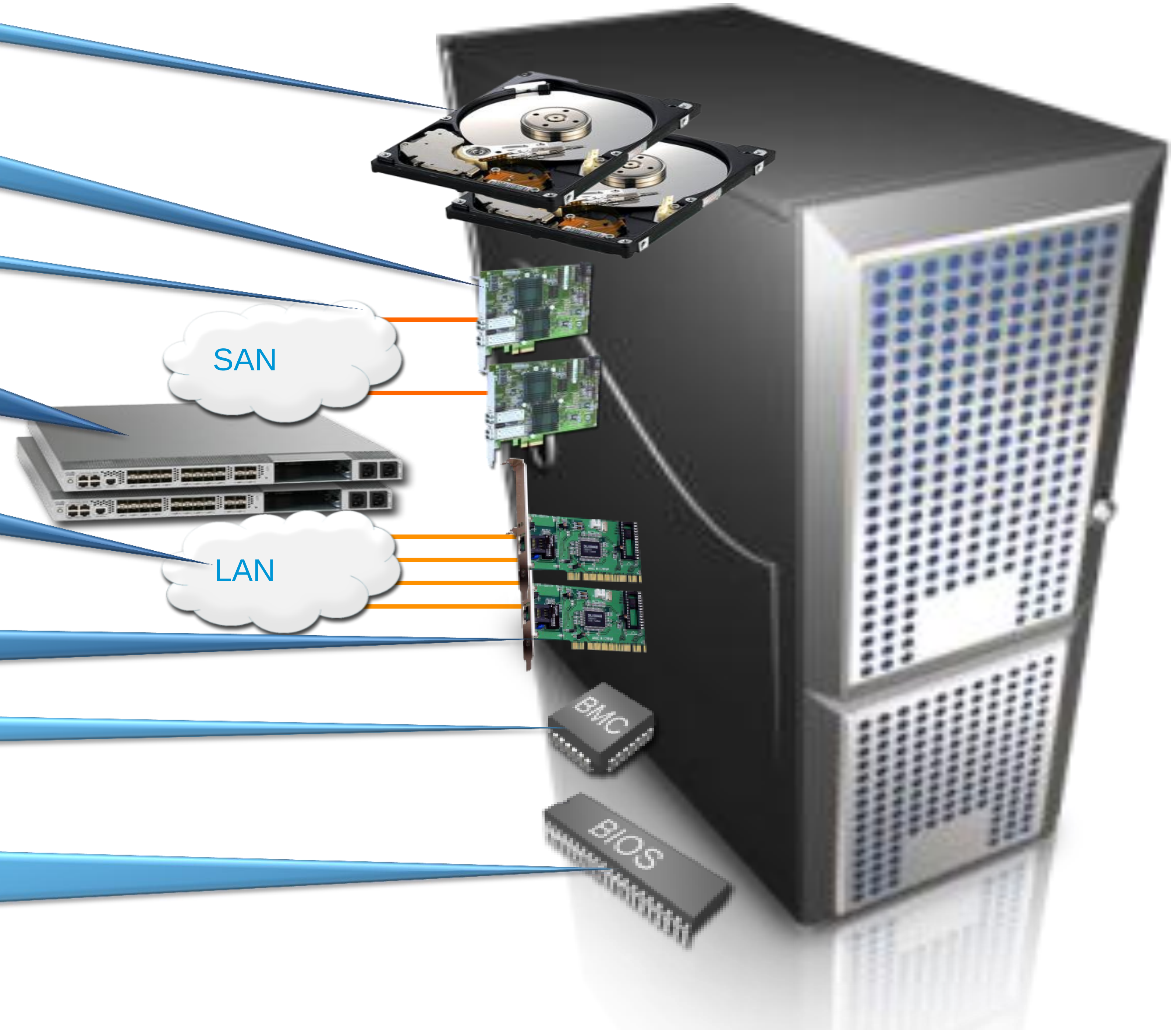
- QoS settings
- Border port assignment per vNIC
- NIC Transmit/Receive Rate Limiting

- VLAN assignments for NICs
- VLAN tagging config for NICs

- Number of vNICs
- PXE settings
- NIC firmware
- Advanced feature settings

- Remote KVM IP settings
- Call Home behaviour
- Remote KVM firmware

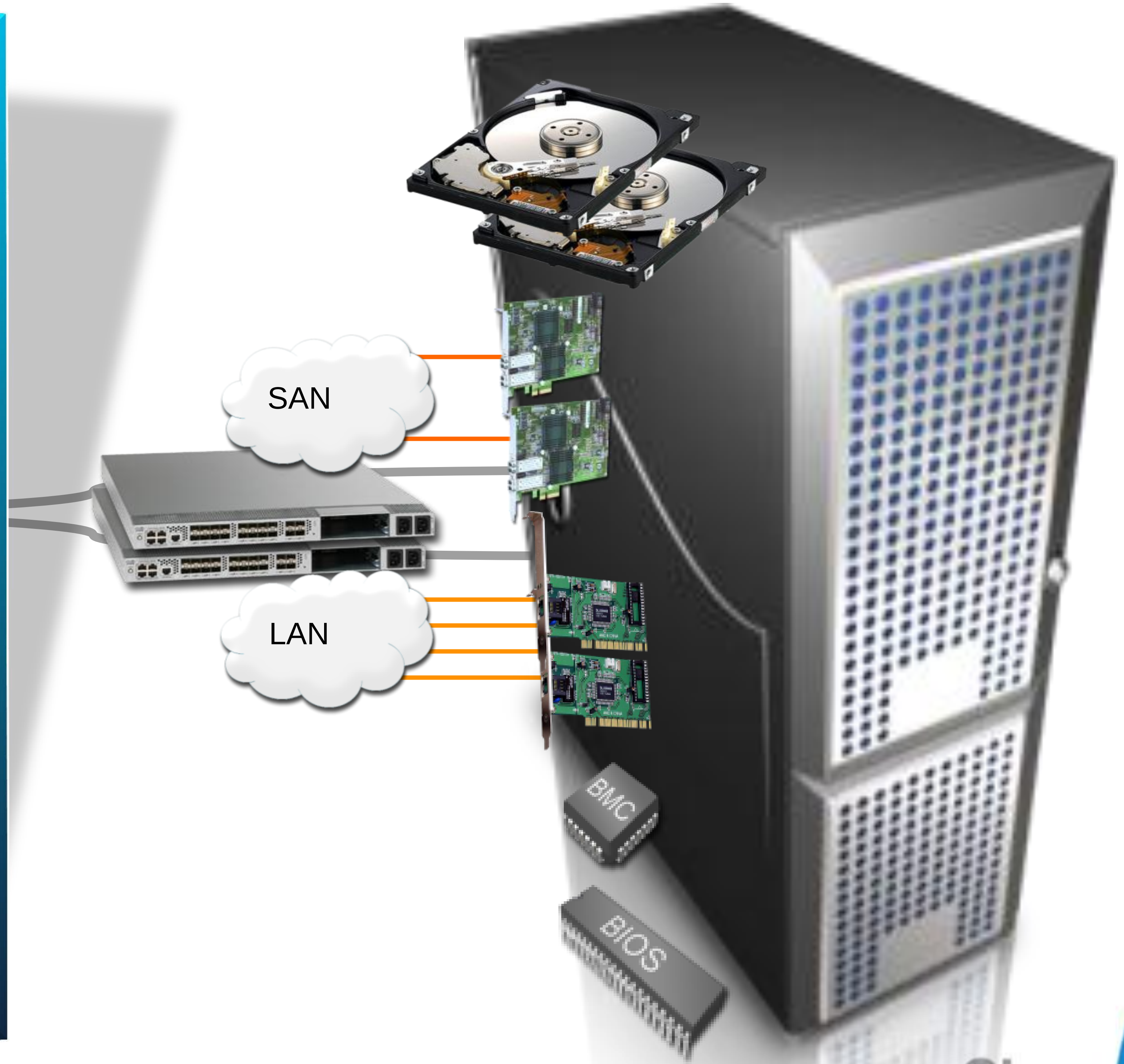
- Server UUID
- Serial over LAN settings
- Boot order
- IPMI settings
- BIOS scrub actions
- BIOS firmware
- BIOS Settings



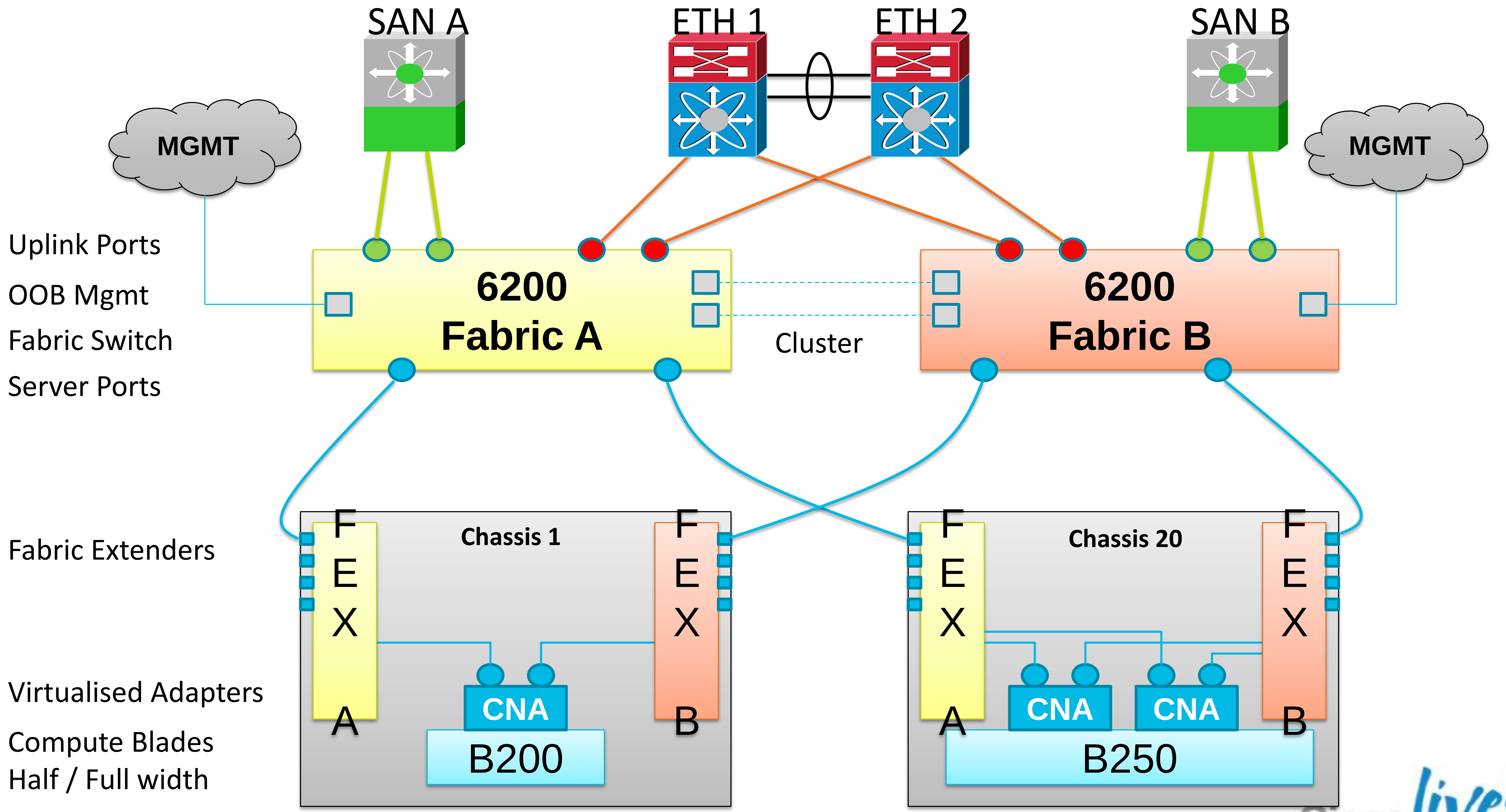
UCS Service Profiles

Service Profile

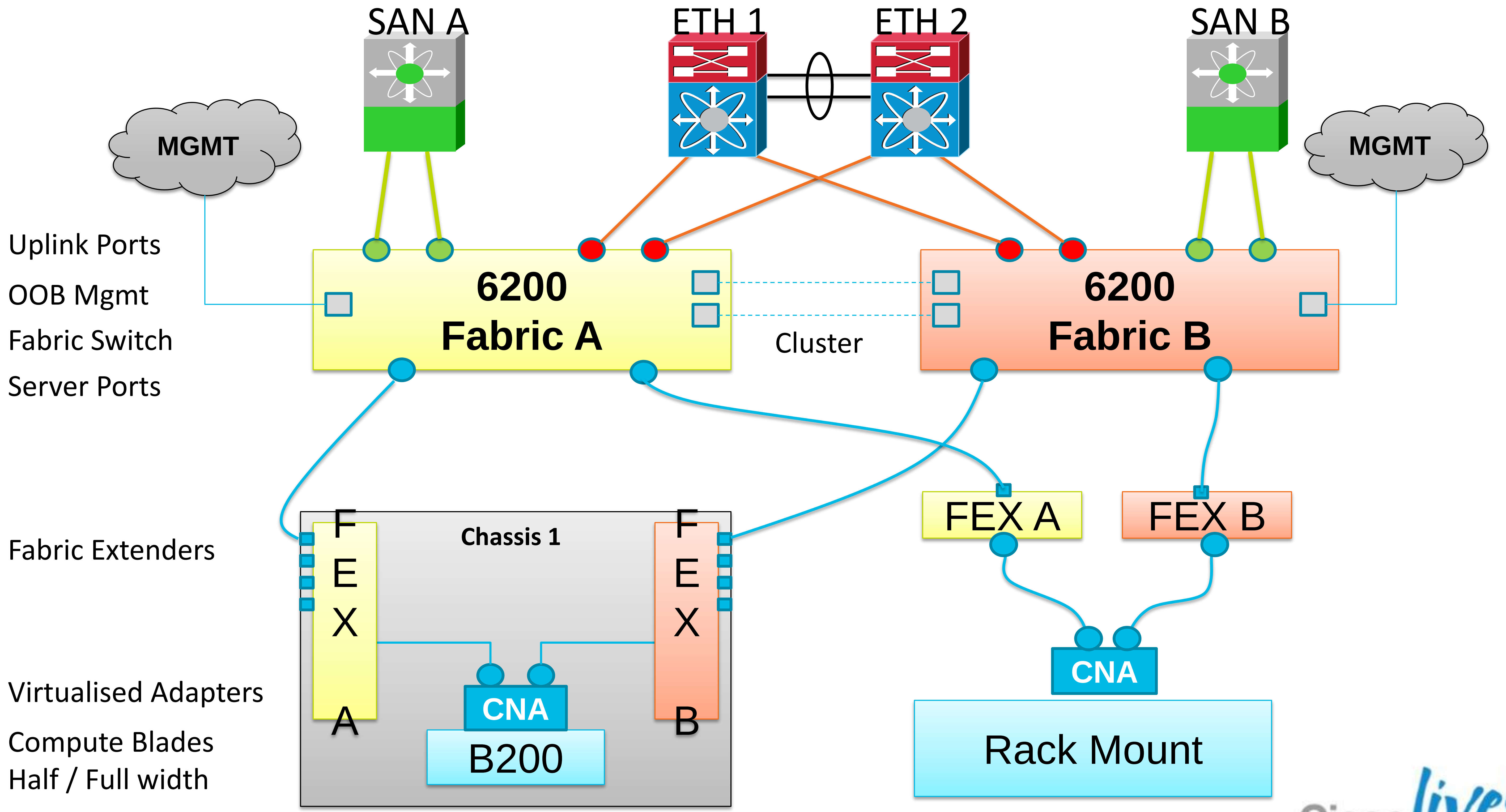
- NIC MACs
- HBA WWNs
- Server UUID
- VLAN Assignments
- VLAN Tagging
- FC Fabrics Assignments
- FC Boot Parameters
- Number of vNICs
- Boot order
- PXE settings
- IPMI Settings
- Number of vHBAs
- QoS
- Call Home
- Template Association
- Org & Sub Org Assoc.
- Server Pool Association
- Statistic Thresholds
- BIOS scrub actions
- Disk scrub actions
- BIOS firmware
- Adapter firmware
- BMC firmware
- RAID settings
- Advanced NIC settings
- Serial over LAN settings
- BIOS Settings



Cisco UCS Networking: Physical Architecture

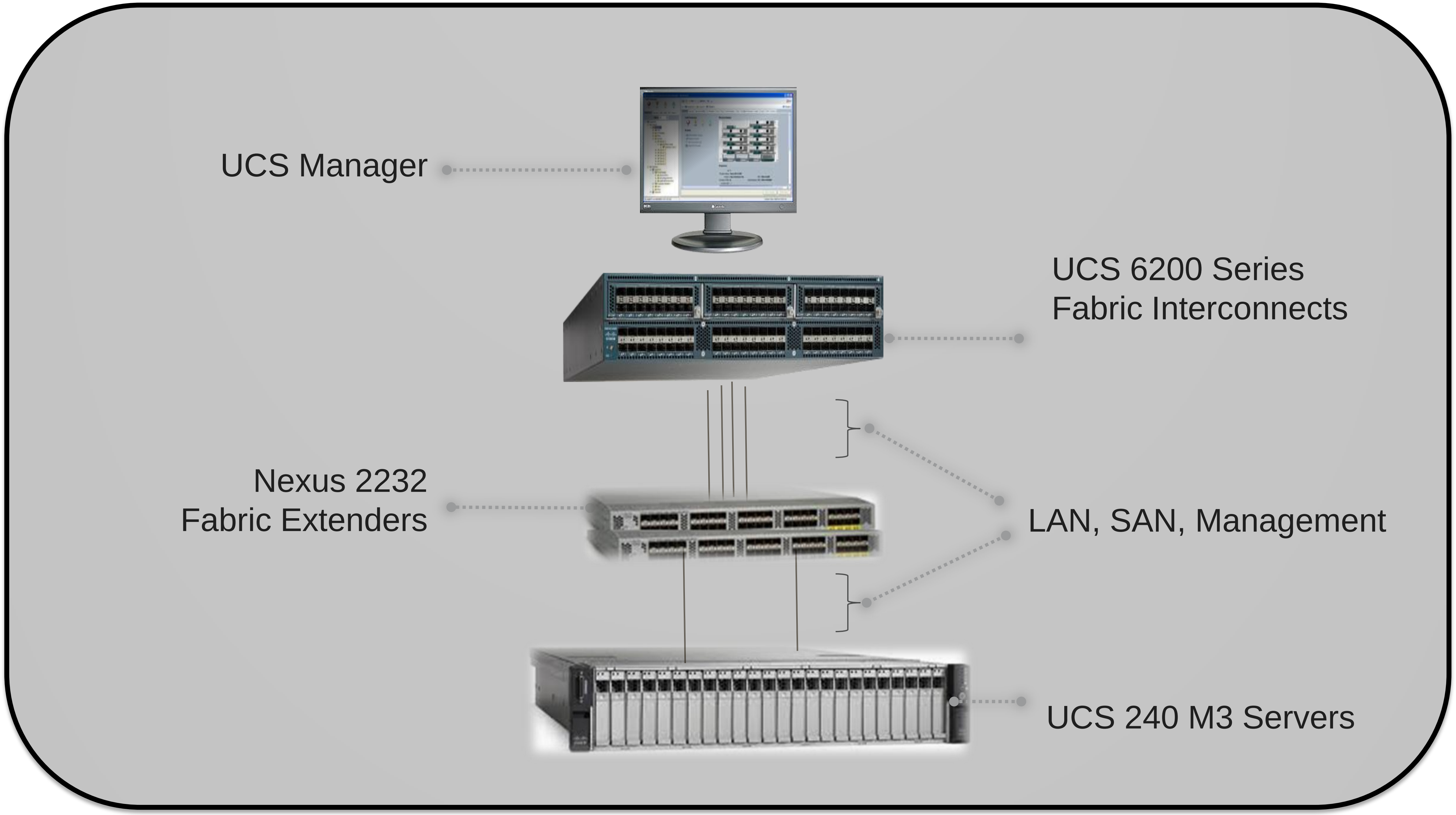


Cisco UCS Networking: Physical Architecture



Cisco UCS Big Data Common Platform Architecture (CPA)

Building Blocks for Big Data



Hadoop Hardware Evolving in the Enterprise

Typical 2009 Hadoop node

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

Economics favor "fat" nodes

- 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

Typical 2012 Hadoop node

- 2RU server
- 12 x 3TB 3.5" or 24 x 1TB 2.5" spindles
- 2 x 8-core CPU
- 1-2 x 10GE
- 128 GB RAM
- Dual PSU
- Running Cloudera or MapR or HortonWorks
- $$$

Balanced Node Configuration Example

Network impact on single node failure and recovery

41 data nodes in the cluster, 12 x 3TB SATA = 36 TB/node

- Assume 75%/25% ratio of file to temp space
- Assume 75% utilisation of file space
- 36 TB * 0.75 * 0.75 = ~ 20TB actually used for HDFS blocks

40 data nodes remain after 1 fails

- Assume even distribution of replicas – each node has 0.5 TB it needs to send elsewhere
- At 1 Gbps, that's ~ 1.1 hrs. theoretical best case assuming no other network traffic, no oversub, no HDFS overhead, etc.
- At 10 Gbps, the bottleneck is likely the disks (12 x 80MB/s ~ 8 Gbps)

Cisco UCS Bundles for Big Data

Half-Rack UCS Solutions Bundle for MPP Configuration	Single Rack UCS Solutions Bundle for Hadoop Capacity	Single Rack UCS Solutions Bundle for Hadoop Performance
2 x UCS 6248 2 x Nexus 2232 PP 8 x C240 M3 (SFF) 2x E5-2690 256GB 24x 600GB 10K SAS	2 x UCS 6296 2 x Nexus 2232 PP 16 x C240 M3 (LFF) E5-2640 (12 cores) 128GB 12x 3TB 7.2K SATA	2 x UCS 6296 2 x Nexus 2232 PP 16 x C240 M3 (SFF) 2x E5-2665 (16 cores) 256GB 24 x 1TB 7.2K SAS

Cluster Management Pain Points

UCSM provides:

- Speed
- Ease of experimentation
- Consistency
- Simplicity
- Visibility

Big Data Infrastructure

UCS Management (160 Nodes per UCS Managed Cluster/Domain)

Inventory & Asset Mgmt

- Cluster Layout and Inventory
- Per-Server Inventory
- ID Pools (MAC, IP, UUID) Management

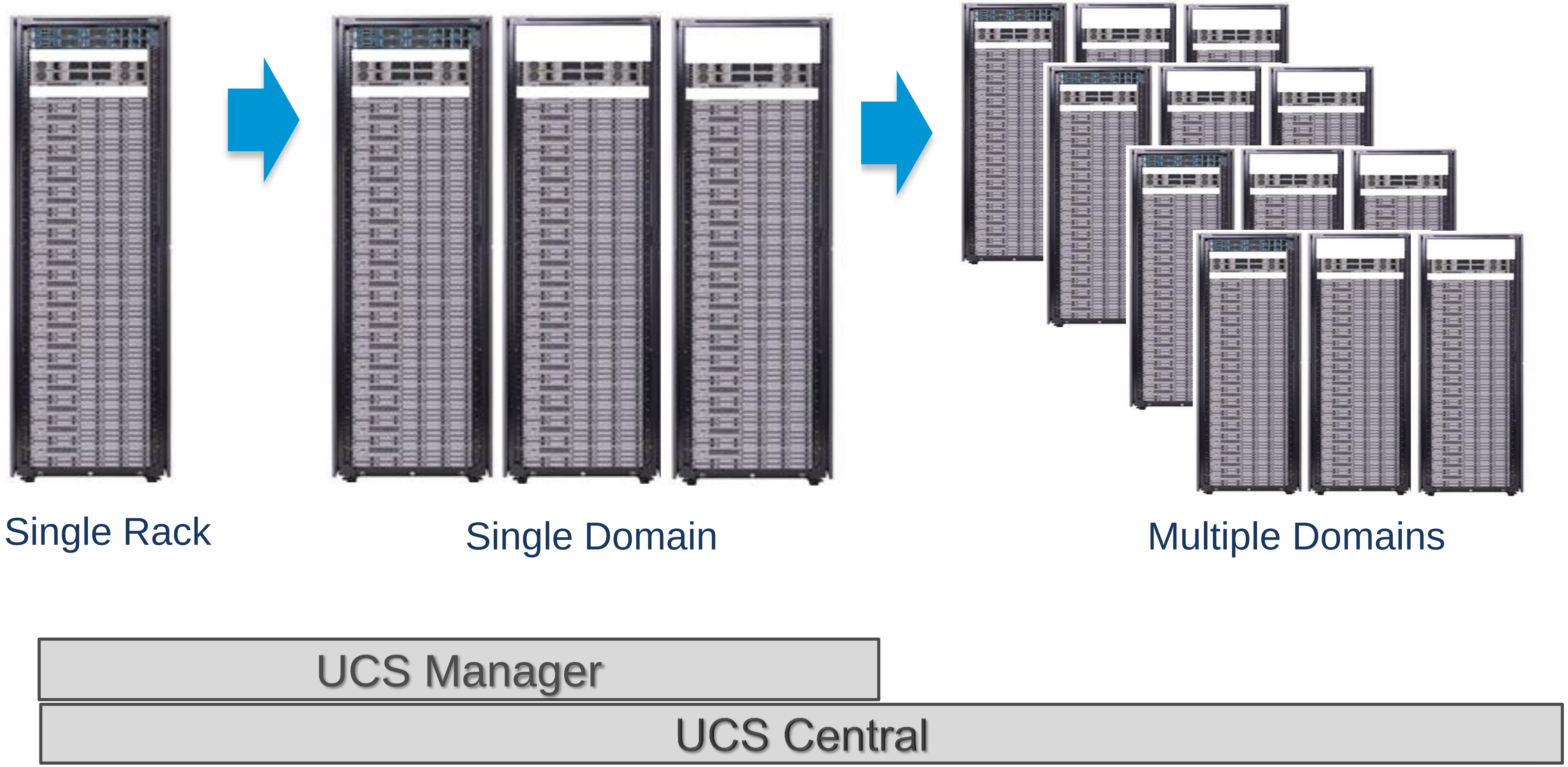
Fault Detection & SW Updates

- Fault detection & Logs
- Event Aggregation
- System software updates

QoS Policies & Power Capping

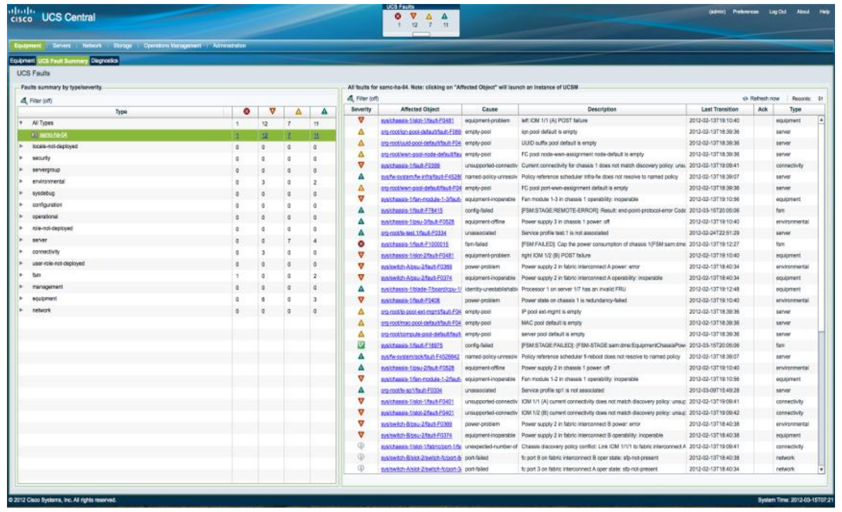
- QoS Policy definition
- Policy driven framework
- Policy Based Power Capping

Growth of Infrastructure



Big Data Infrastructure

UCS Multi-Domain (UCS Central Manages up to 10,000 nodes)



- Inventory, Fault, Log, Event Aggregation
- Global ID Pools, Firmware Updates, Backups and Global Admin Policies

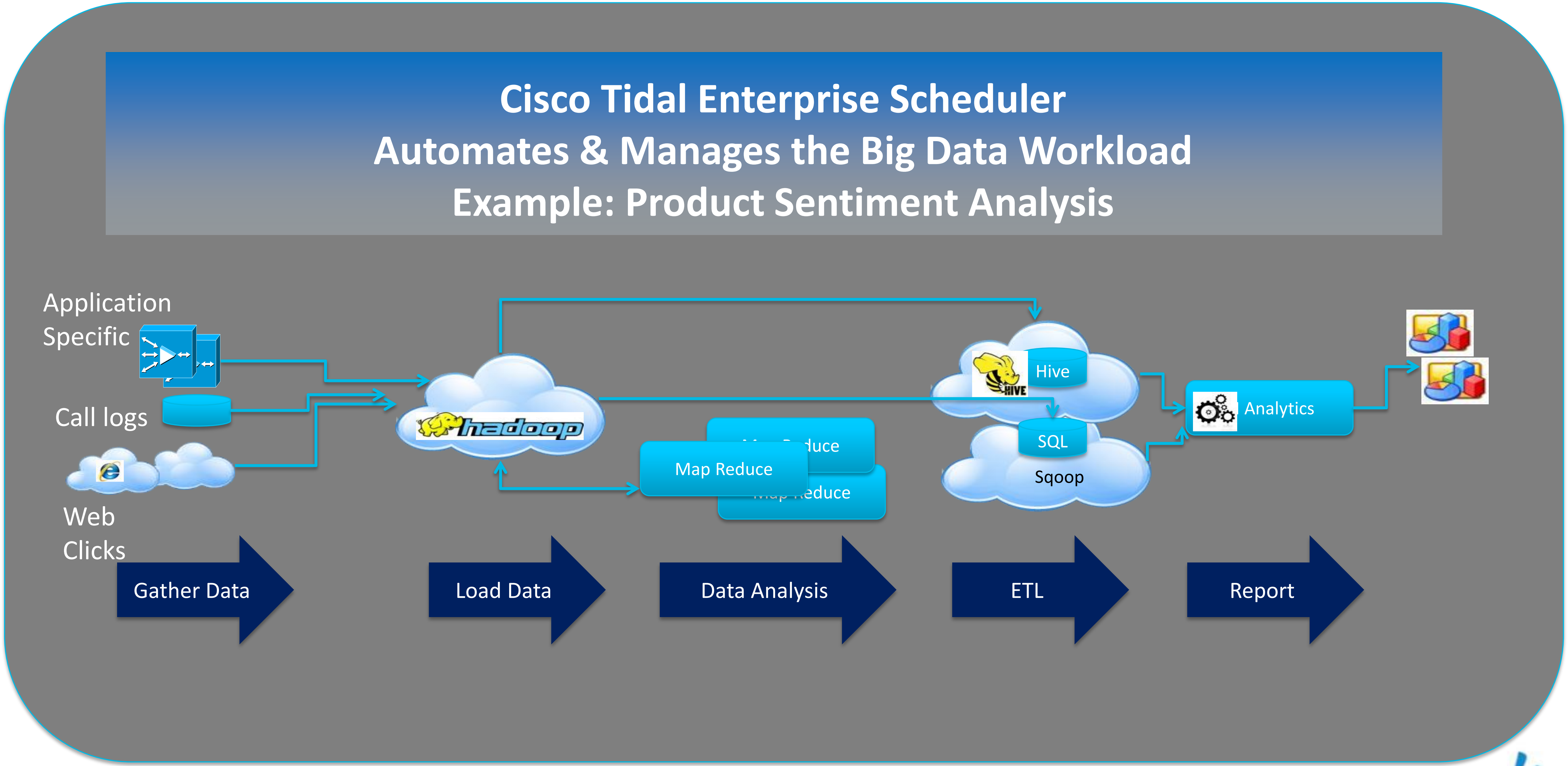
Available

- Global Service Profiles, Templates & Policies
- Statistics Aggregation
- HA for UCS Central Virtual Machine with shared storage

Coming soon

Tidal Enterprise Scheduler

Holistic Approach to Big Data Automation



Benefits of Cisco UCS for Big Data

ARCHITECTURE	Modular architecture common across different domains
MANAGEMENT	Simplified and centralised management across domain
PERFORMANCE	Industry leading performance & scalability with UCS rack mount servers & 10G flexible networking
TIME TO VALUE	Rapid, consistent deployment with reduced risk
SUPPORT	Enterprise-class service and support

Thank You For Listening

Cisco.com Big Data

www.cisco.com/go/bigdata

Cisco Unified Data Centre

UNIFIED FABRIC

Highly Scalable, Secure Network Fabric

www.cisco.com/go/nexus

UNIFIED COMPUTING

Modular Stateless Computing Elements

www.cisco.com/go/ucs

UNIFIED MANAGEMENT

Automated Management

Manages Enterprise Workloads

http://www.cisco.com/go/workloadautomation

Q & A

Complete Your Online Session Evaluation

Give us your feedback and receive a Cisco Live 2013 Polo 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 www.ciscoliveaustralia.com/mobile
- Visit any Cisco Live Internet Station located throughout the venue

Polo Shirts can be collected in the World of Solutions on Friday 8 March 12:00pm-2:00pm

Don't forget to activate your Cisco Live 365 account for access to all session material, communities, and on-demand and live activities throughout the year. Log into your Cisco Live portal and click the "Enter Cisco Live 365" button.
www.ciscoliveaustralia.com/portal/login.ww