



# Not Just for the Web Giants – The Intelligent Enterprise

---

Google



eBay



Walmart 



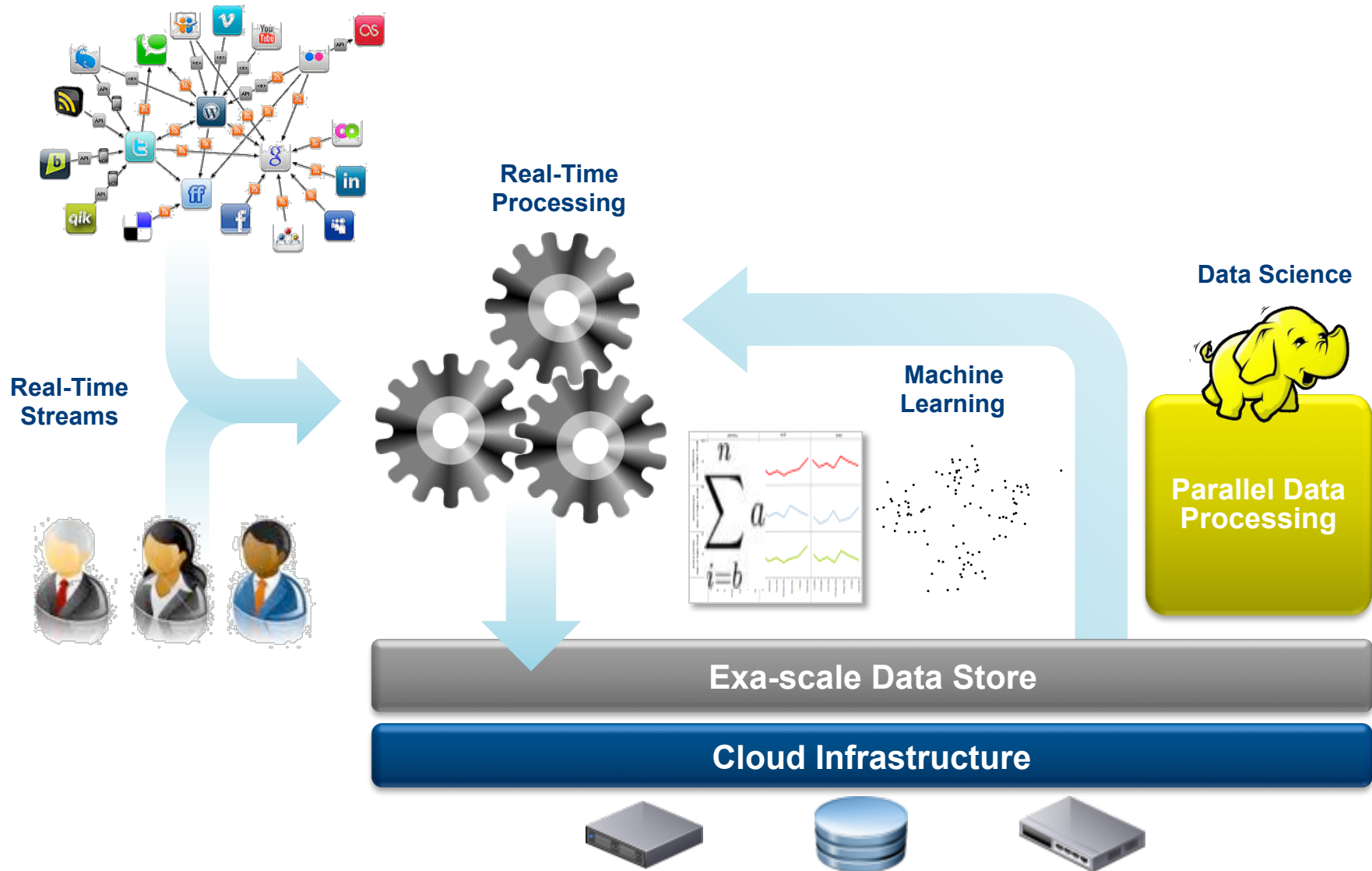
**Real-time analysis allows instant understanding of market dynamics.**

**Retailers can have intimate understanding of their customers needs and use direct targeted marketing.**



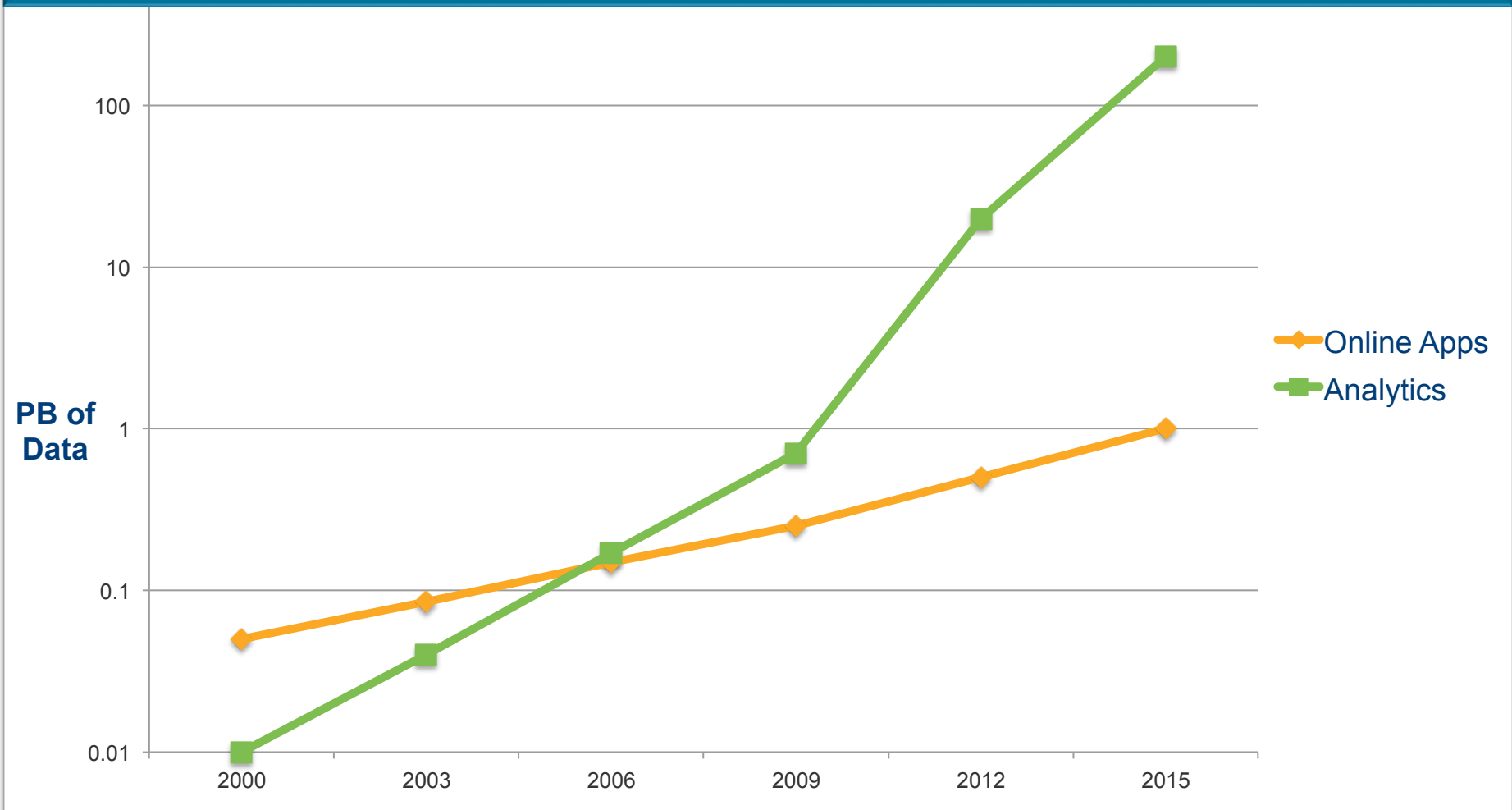
**Market Segment Analysis → Personalized Customer Targeting`**

# The Emerging Pattern of Big Data Systems: Retail Example



# Storage: Plan for Peta-scale Data Storage and Processing

## Analytics Rapidly Outgrows Traditional Data Size by 100x



# Unprecedented Scale

**\$3398**  
**10MB**

## THE HARD DISK YOU'VE BEEN WAITING FOR

**XCOMP** introduces a complete micro-size disk subsystem with more...

- MORE STORAGE
- MORE SPEED
- MORE VALUE
- MORE SUPPORT

**OUTPERFORMS OTHER HARD DISKS**  
Floppy disk and larger, more expensive hard disks are no match for this powerful little system. More data is available on every seek. 64K on 10MB and 32K on 5MB. Faster seek time too — an average of 70MS. It provides solid performance anywhere with only 20 watts of power. Data is protected in the sealed enclosure, and the landing zone for heads provides another margin of safety. The optional power board plugs directly into the S100 bus and provides power for the drive.

**MORE SOFTWARE**  
Included with the system is software for testing, formatting, I/O drivers for CP/M<sup>®</sup>, plus an automatic CP/M driver attach program. Support software and drivers for MP/M<sup>®</sup> and Oasis<sup>®</sup> are also available. The sophisticated formatting program assigns alternate sectors for any weak sectors detected during formatting, assuring the lowest possible error rate — at least ten times better than floppies.

**WARRANTY**  
The system has a full one-year warranty on parts and workmanship.

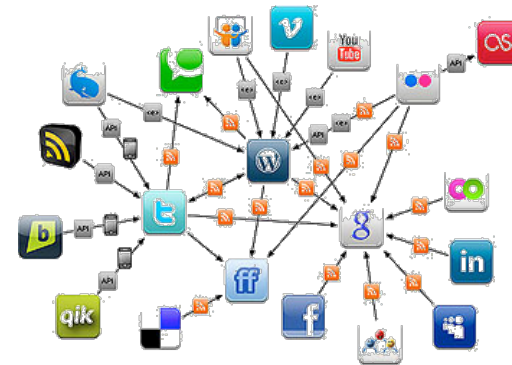
**ALSO AVAILABLE FROM XCOMP**

- General Purpose controllers (8 bit interface), with easy interface to microprocessor-based systems
- GP controller adapter that plugs directly into most Z80 computers
- S.T.R. GP controller for the 5MB and 10MB drive above, with ST506 type interface
- S.G.R. GP controller for SA1000 interface
- S.M.R. GP controller for storage module drives
- S.T.S. S.G.S. and S.M.S. same as above, for the S100 bus

Quantity discounts available. Distributor, Dealer, and OEM inquires invited.  
See your local Dealer, or call XCOMP, Inc. 14500 14th Street

We are creating an Exabyte of data every minute in 2013

**Yottabyte by 2030**



“Data transparency, amplified by Social Networks generates data at a scale never seen before”  
- *The Human Face of Big Data*

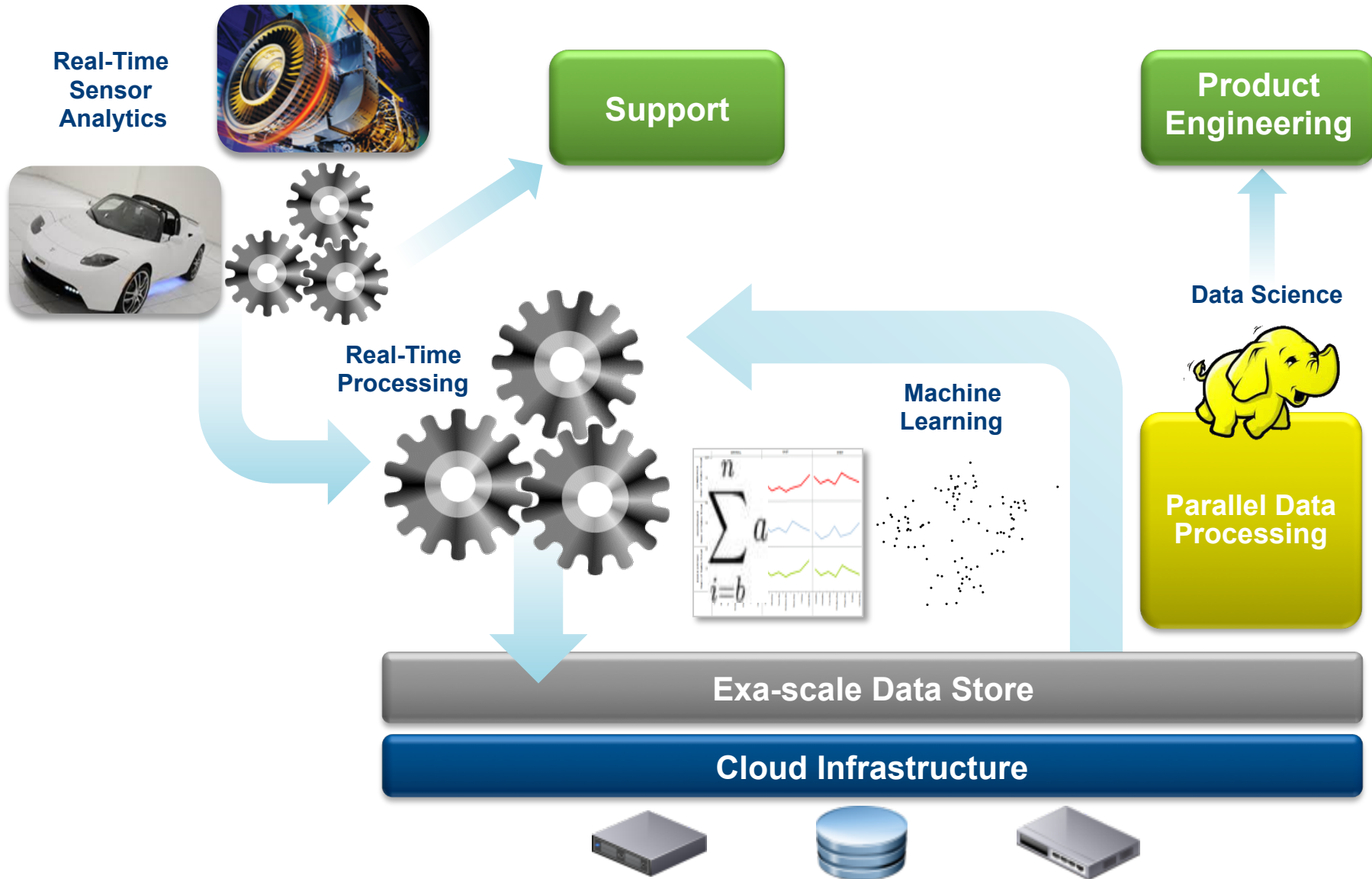


**A single GE Jet Engine produces  
10 Terabytes of data in one hour  
– 90 Petabytes per year.**

**Enabling early detection of  
faults, common mode failures,  
product engineering feedback.**

**Post Mortem → Proactively Maintained Connected Product**

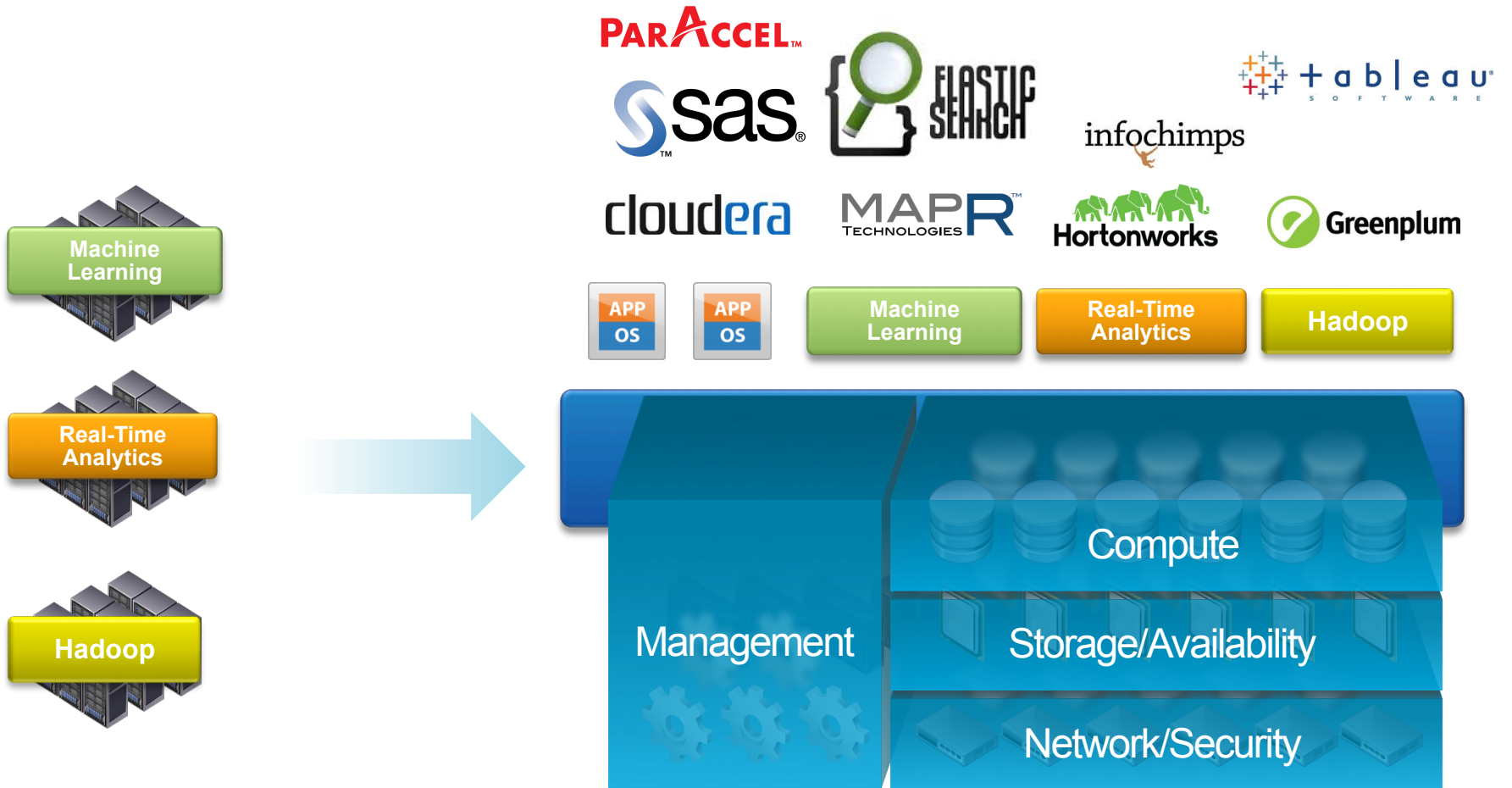
# The Emerging Pattern of Big Data Systems: Manufacturing



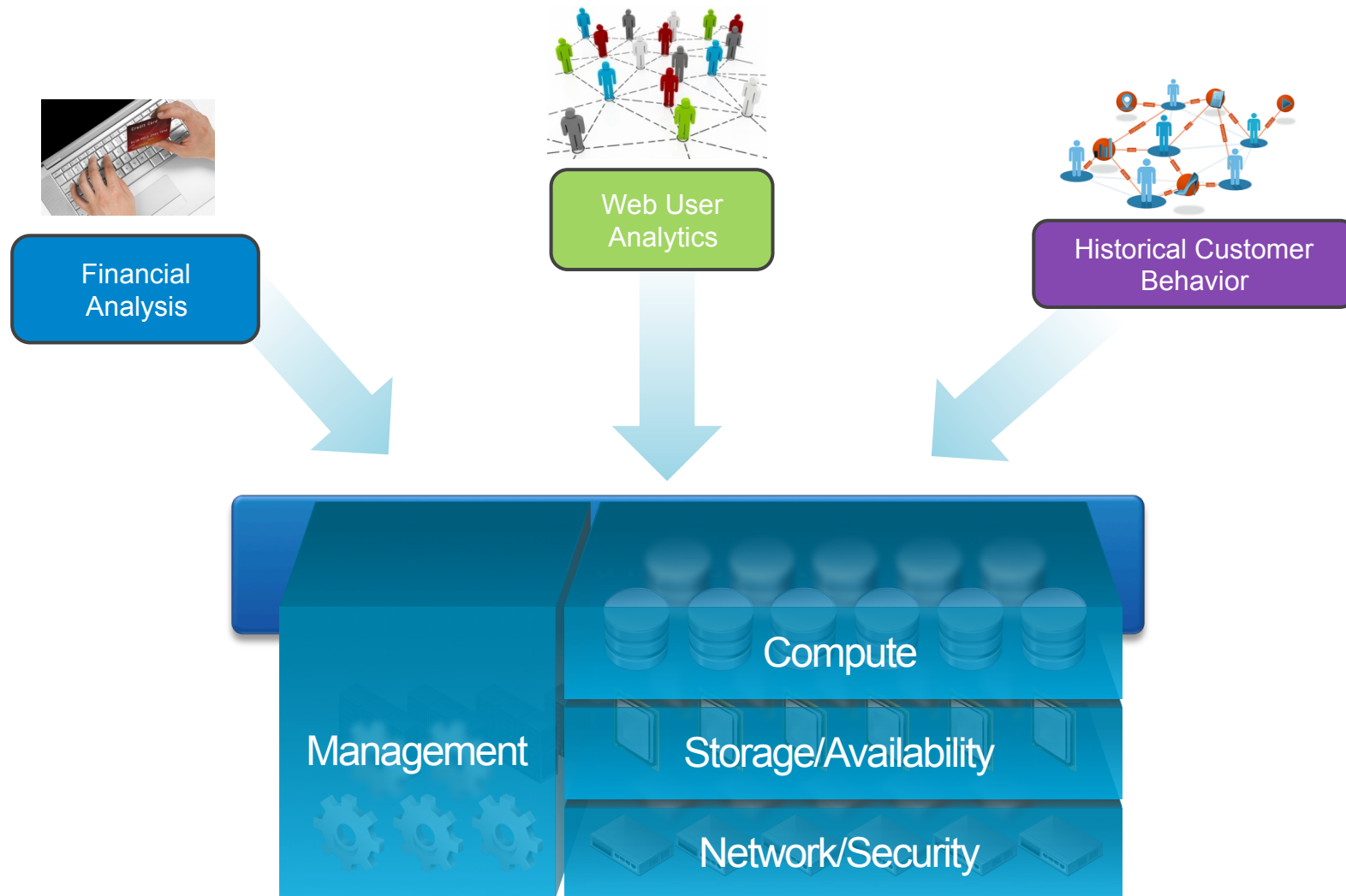




# Cloud Platform: Supporting Mixed Big Data Workloads



# Cloud Platform: Supporting Multiple Tenants



# What if you can...

Recommendation engine

Production



Test



Experimentation



Ad targeting

Production



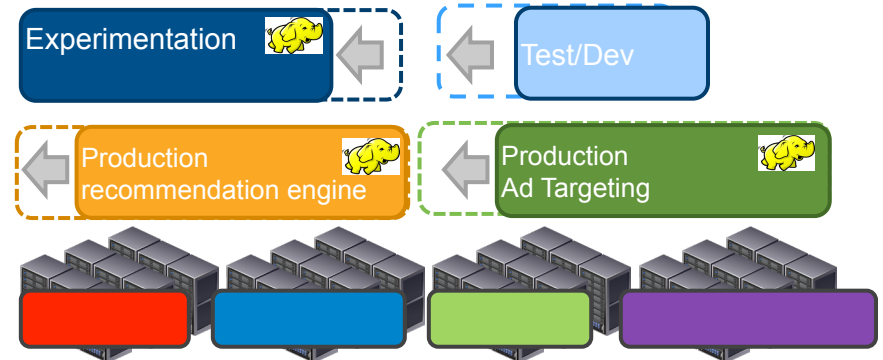
Test



Experimentation



One physical platform to support multiple virtual big data clusters



# Values of a Cloud Platform for Big Data

---



- 1 Agility / Rapid deployment
- 2 Isolation for resource control and security
- 3 Lower Capex
- 4 Operational efficiency

# Hadoop as a Service

---

## Operational Simplicity

- Rapid deployment
- One stop command center
- Easy to configure/reconfigure

## High Availability

- High availability for entire Hadoop stack
- One click to setup
- Battle-tested

## Elasticity & Multi-tenancy

- Shrink and expand cluster on demand
- Independent scaling of Compute and data
- Strong multi-tenancy

# Self Service Access to Big Data Environments

---



## Developer

- 3 Hadoop nodes
- Cloudera, Pivotal MapR
- Small VM
- Local storage
- No HA
- ...



## Data Scientist

- 5 Hadoop nodes
- Cloudera, Pivotal
- Hive, Pig
- Medium VM
- HA
- ...



## High priority

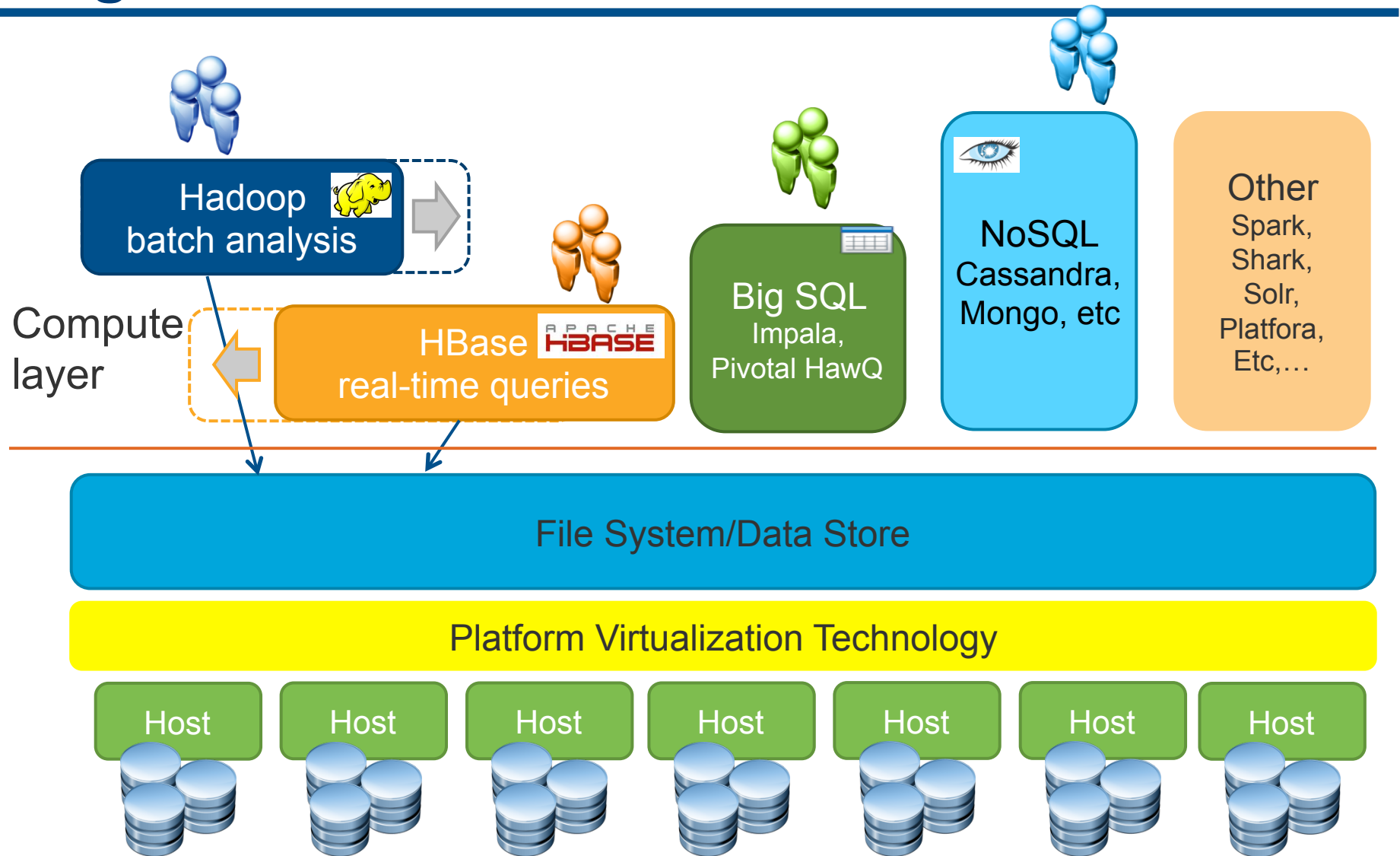
- 50 Hadoop nodes
- Cloudera
- Hive, Pig
- Large VM
- HA
- ...

...

- ...
- ...

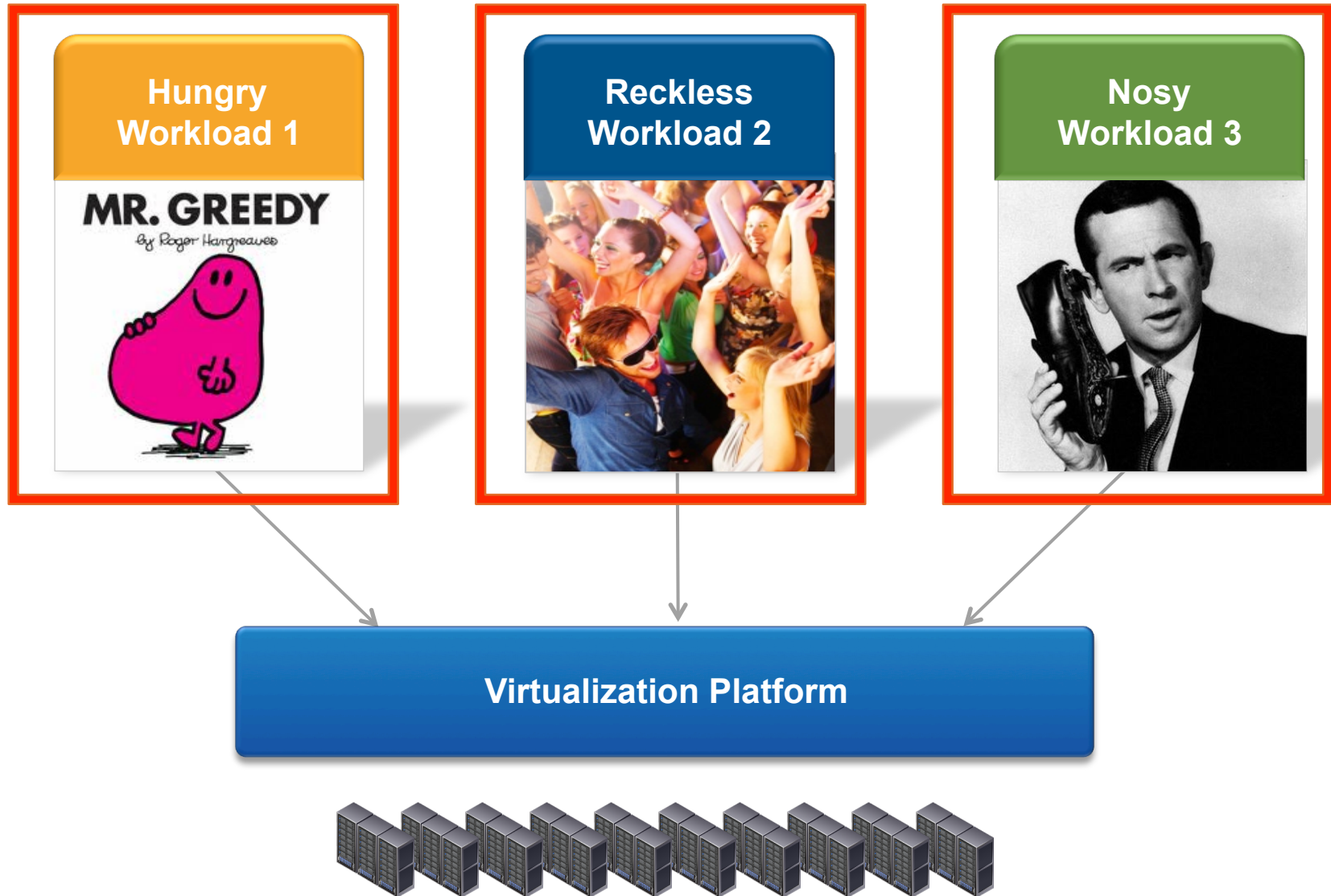
Templates for Different Cloud Users

# Big Data needs a Mix of Workloads





# Strong Isolation between Workloads is Key



# Community activity in Isolation and Resource Management



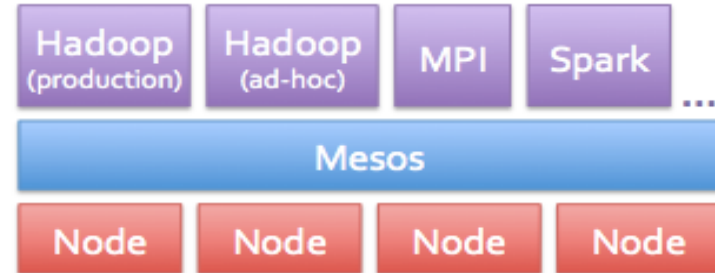
## ■ YARN

- Goal: Support workloads other than M-R on Hadoop
- Initial need is for MPI/M-R from Yahoo
- Non-posix File system self selects workload types



## ■ Mesos

- Distributed Resource Broker
- Mixed Workloads with some RM
- Active project, in use at Twitter
- Leverages OS Virtualization – e.g. cgroups

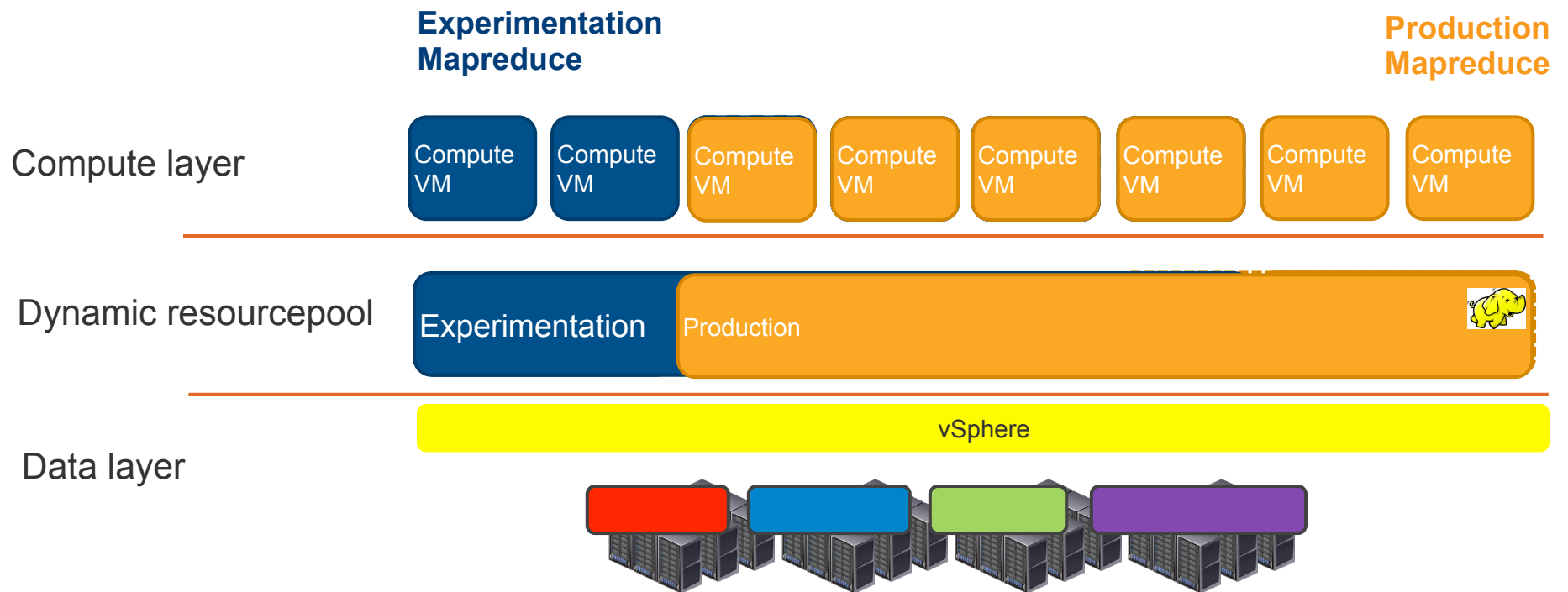


## ■ Virtualization

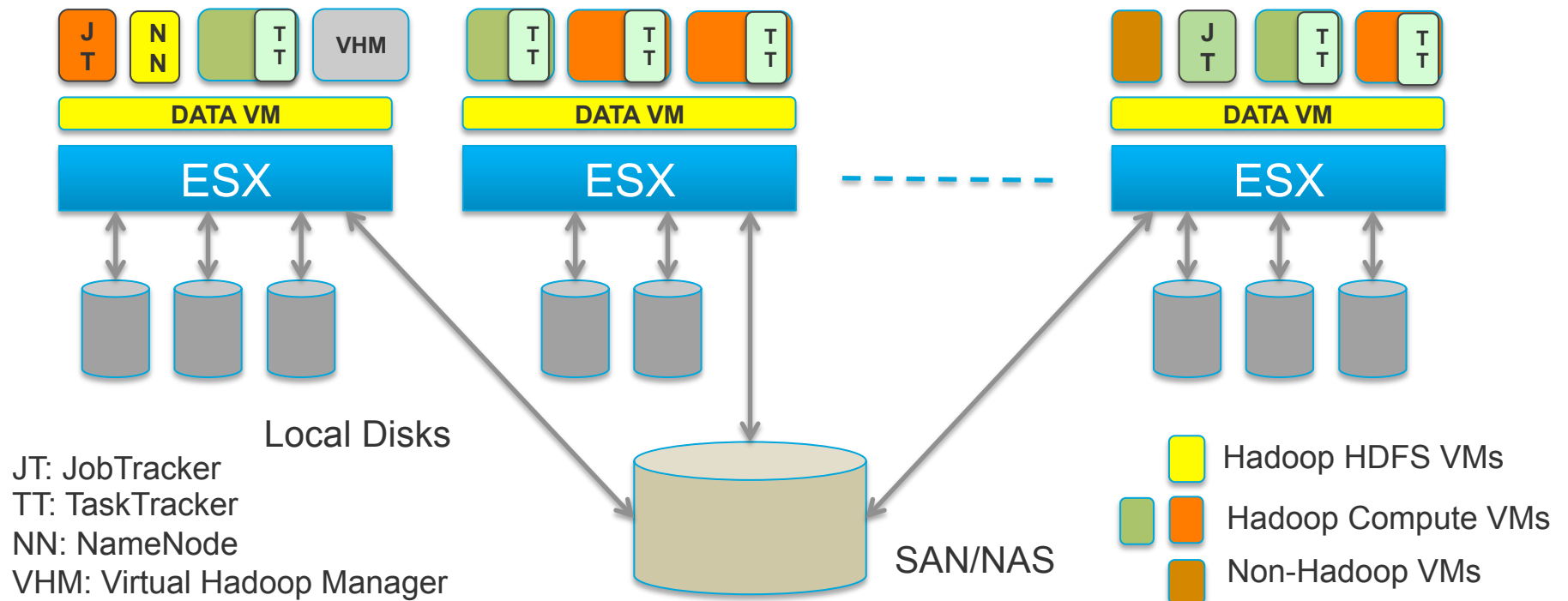
- Virtual machine as the primary isolation, resource management and versioned deployment container
- Basis for Project Serengeti

# Use case: Elastic Hadoop with Tiered SLA

- Production workloads has high priority
- Experimentation workloads has lower priority

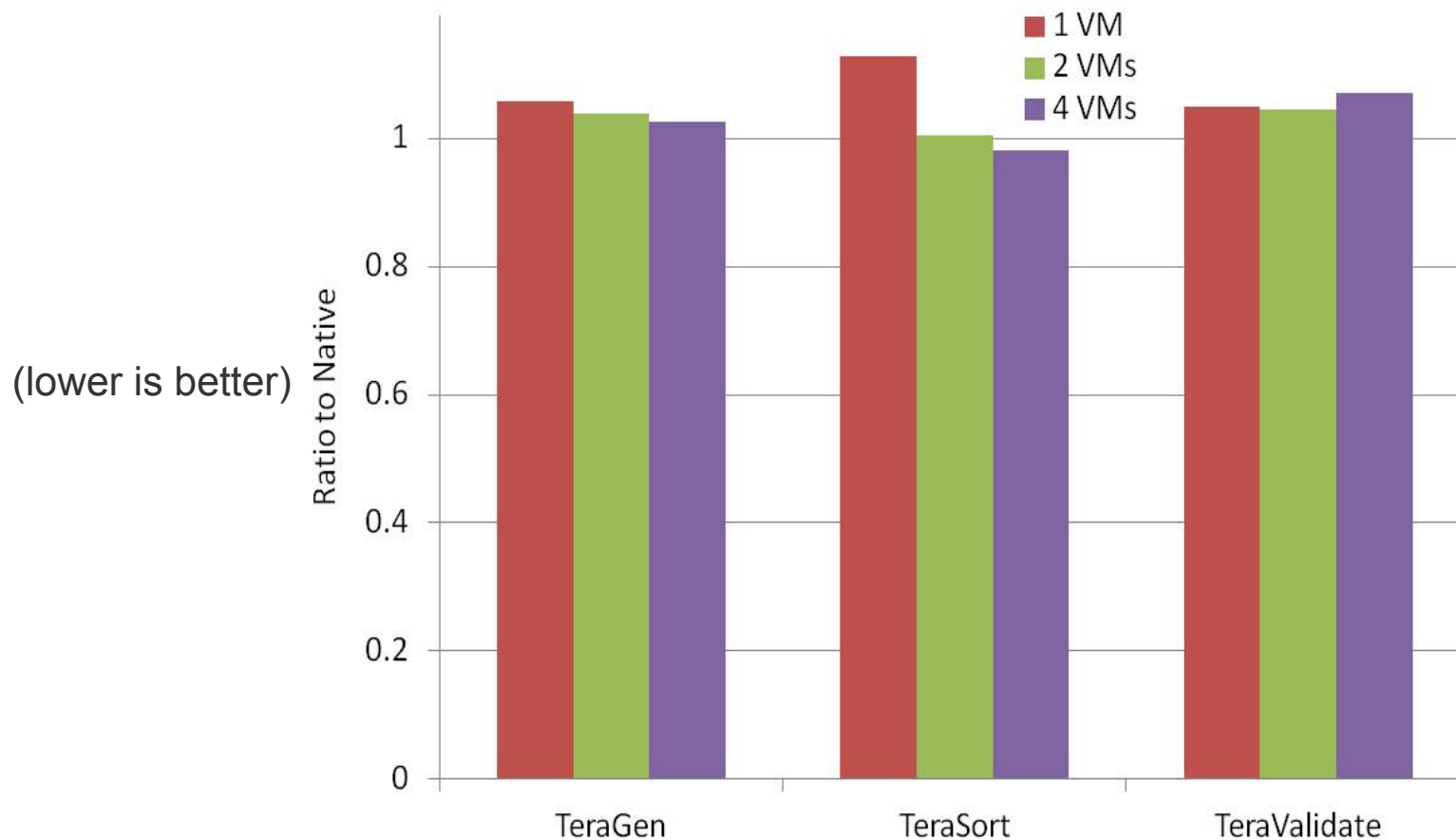


# Cloud Enabled Auto-elastic Hadoop



# Hadoop Performance with Virtualization

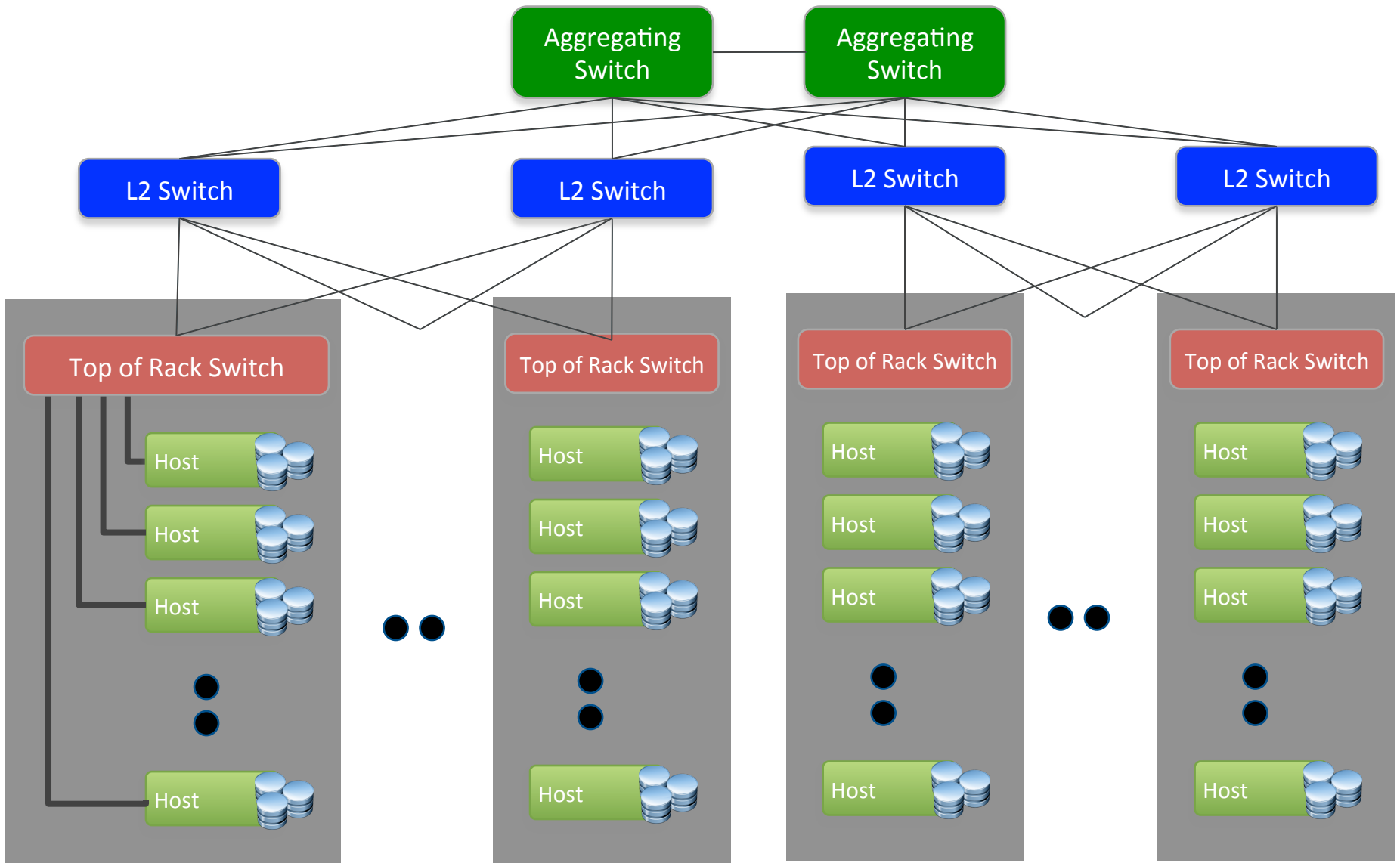
32 hosts/3.6GHz 8 cores/15K RPM 146GB SAS disks/10GbE/72-96GB RAM



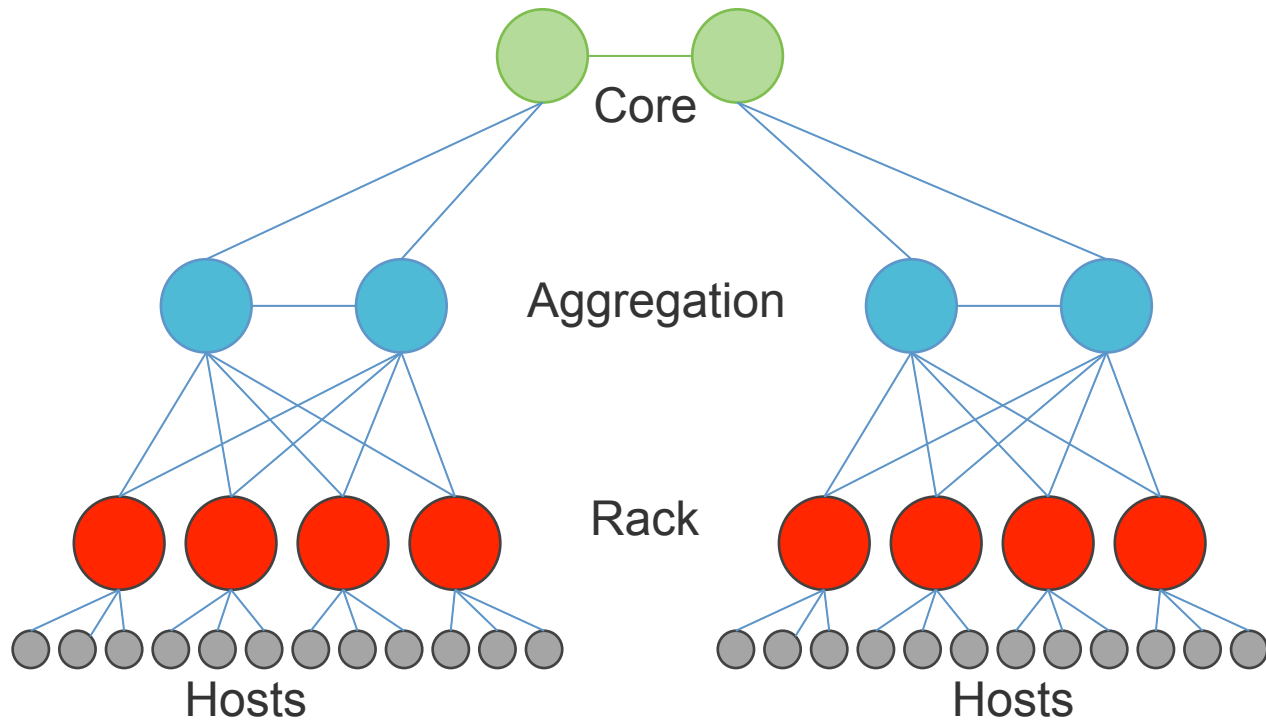
[<http://www.vmware.com/resources/techresources/10360>, Jeff Buell, Apr 2013]



# A Typical Network Architecture



# Traditional Networks: Core Switch is the Choke Point



Network Topology



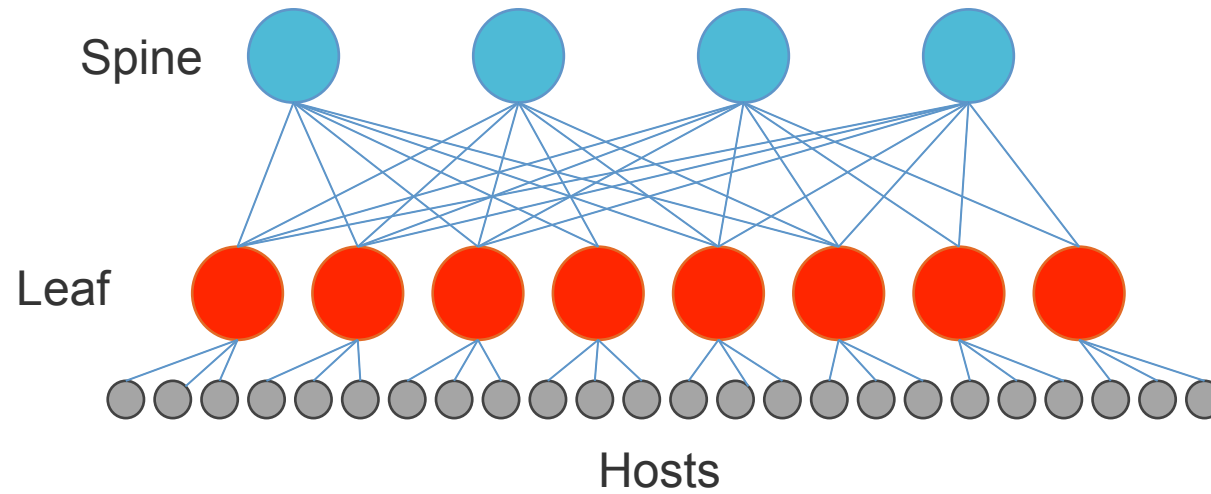
Modeled Bandwidth

100s of Gbits



# Modern Networks: Great for Big Data

---

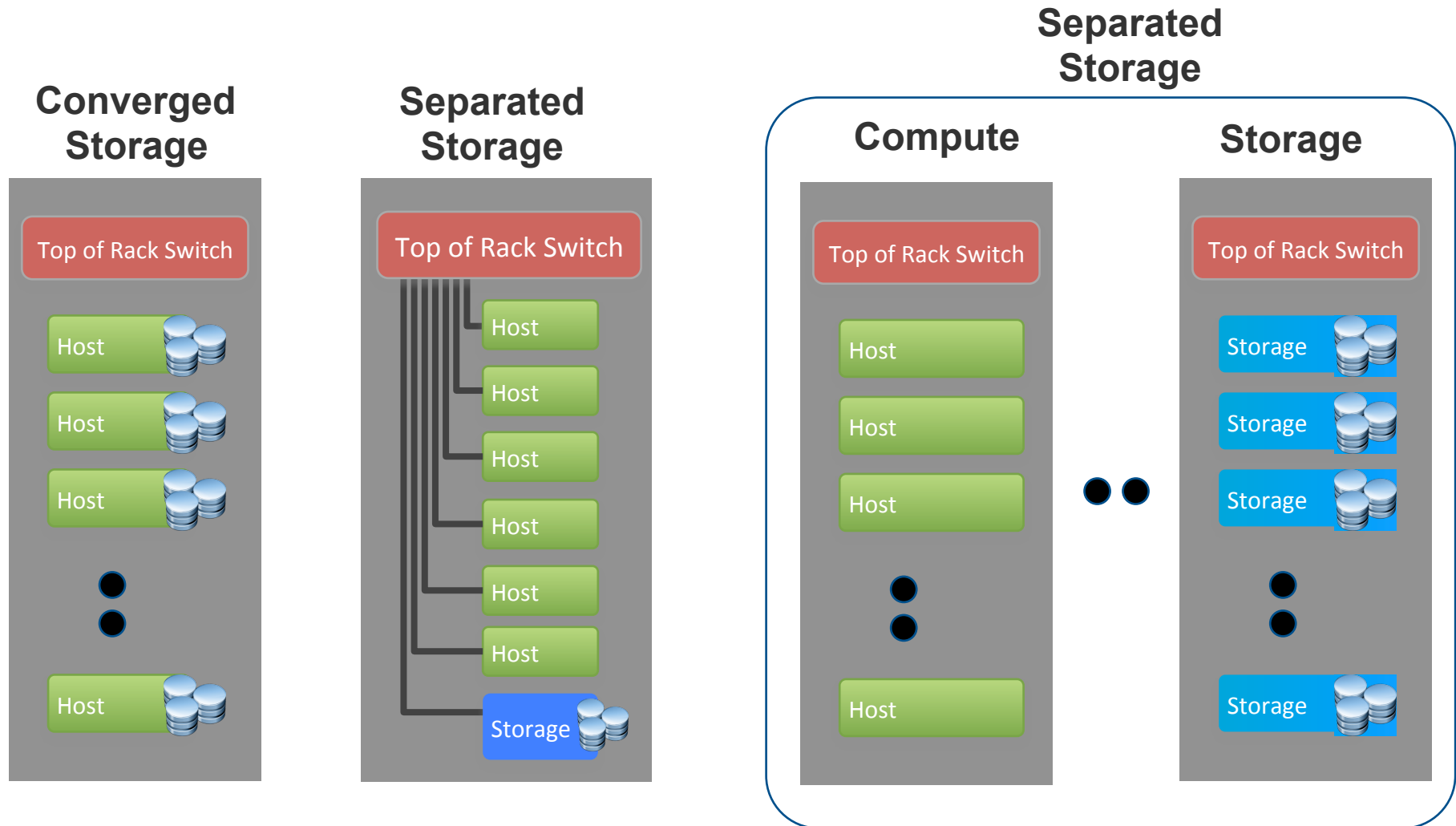


Network  
Topology



Modeled  
Bandwidth

# Flat Networks Allow for New Infrastructure Models





# Use Local Disk where it's Needed



**SAN Storage**

\$2 - \$10/Gigabyte

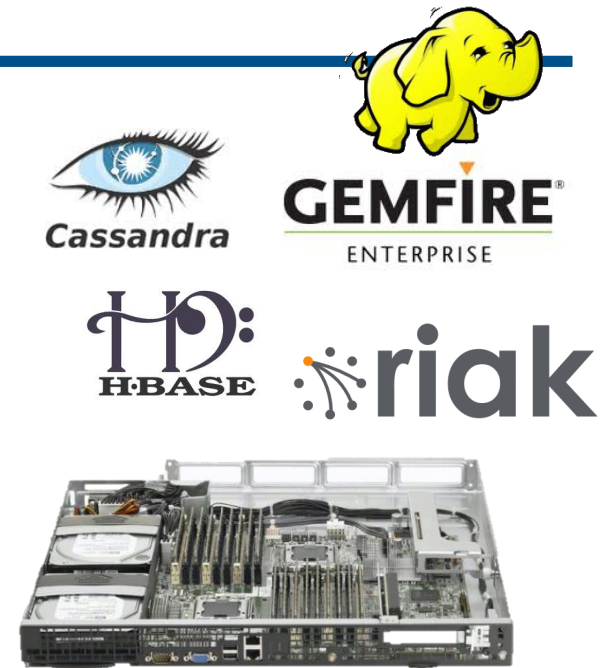
\$1M gets:  
0.5Petabytes  
200,000 IOPS  
8Gbyte/sec



**NAS Filers**

\$1 - \$5/Gigabyte

\$1M gets:  
1 Petabyte  
200,000 IOPS  
10Gbyte/sec

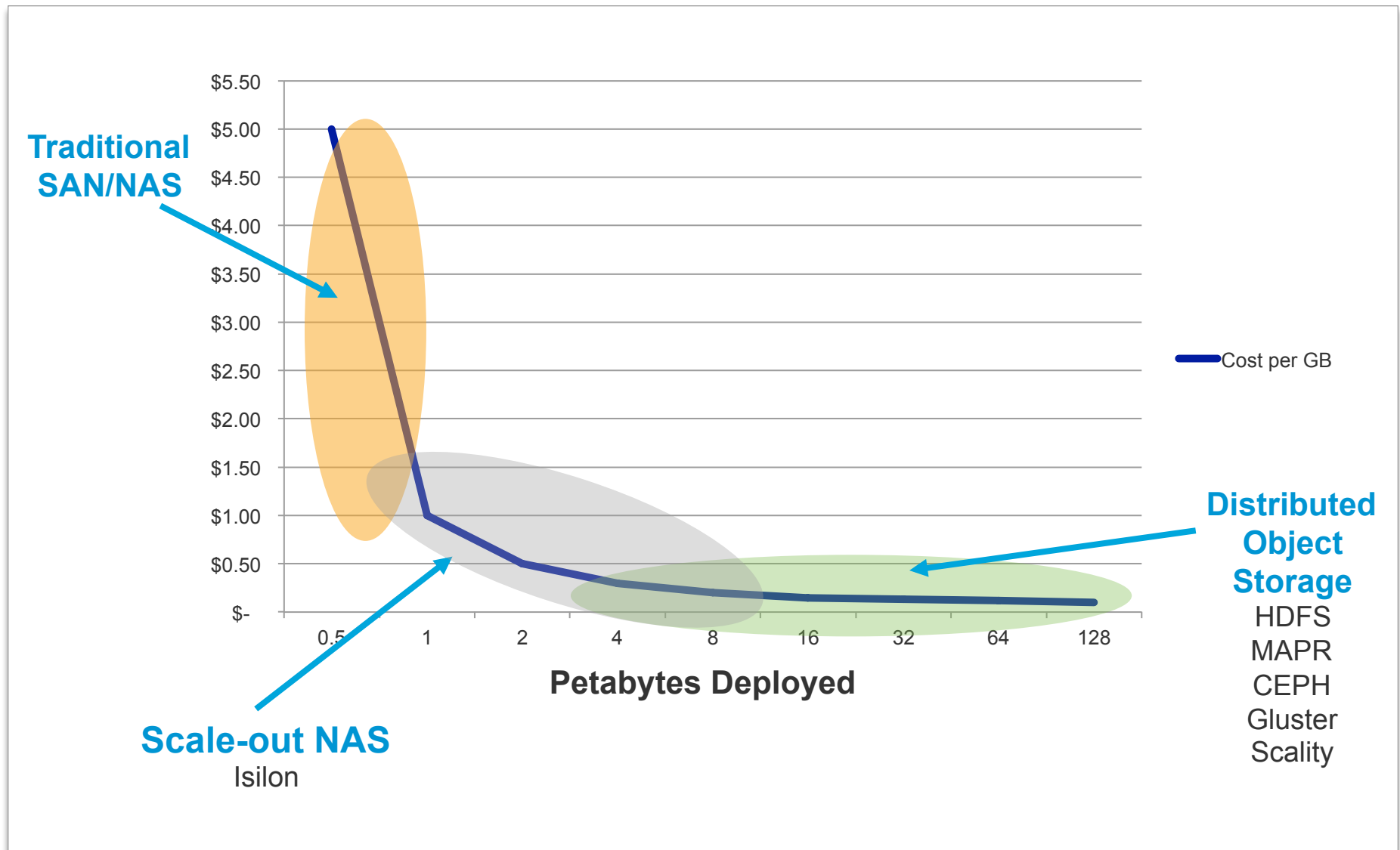


**Local Storage**

\$0.05/Gigabyte

\$1M gets:  
10 Petabytes  
400,000 IOPS  
250 Gbytes/sec

# Storage Economics: Traditional vs. Scale-out



# Big Data Storage Architectures

**External SAN  
With HDFS**



**Local Disks  
With HDFS or  
Other**



**HDFS,  
CEPH,  
MAPR,  
Gluster,  
Scality,  
...**

**External  
Scale-out NAS**



## Features from New Storage Solutions

---

Snapshots Universal File Store

Clones

Geo Replication

Erasure Coding

NFS Access

QoS Controls

Posix Support

SSD Capability





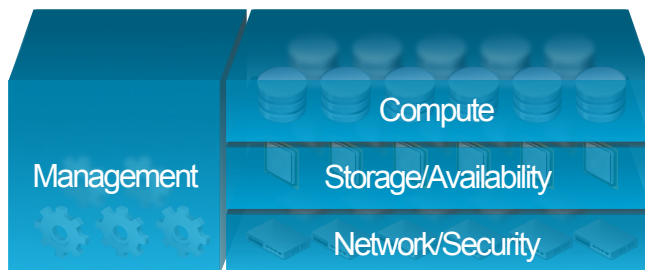
# Customers Winning from Consolidated Big Data Platforms

*“Dedicated hardware makes no sense”*

*“Software-defined Datacenter enables rapid deployment multiple tenants and labs”*

*“Our mixed workloads include Hadoop, Database, ETL and App-servers”*

*“Any performance penalties are minor”*



# Cloud Infrastructure is Ready for Big Data – Are you?

---



Cloud Infrastructure



