Hadoop Virtualization Extensions

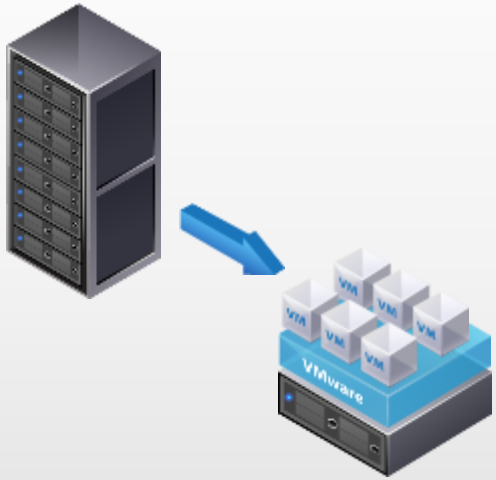
Junping Du

Sr.MTS, VMware, Inc

Cloud: Big Shifts in Simplification and Optimization

1. Reduce the Complexity

to simplify operations and maintenance



2. Dramatically Lower Costs

to redirect investment into value-add opportunities

3. Enable Flexible, Agile IT Service Delivery

to meet and anticipate the needs of the business

A Unified Analytics Cloud Significantly Simplifies

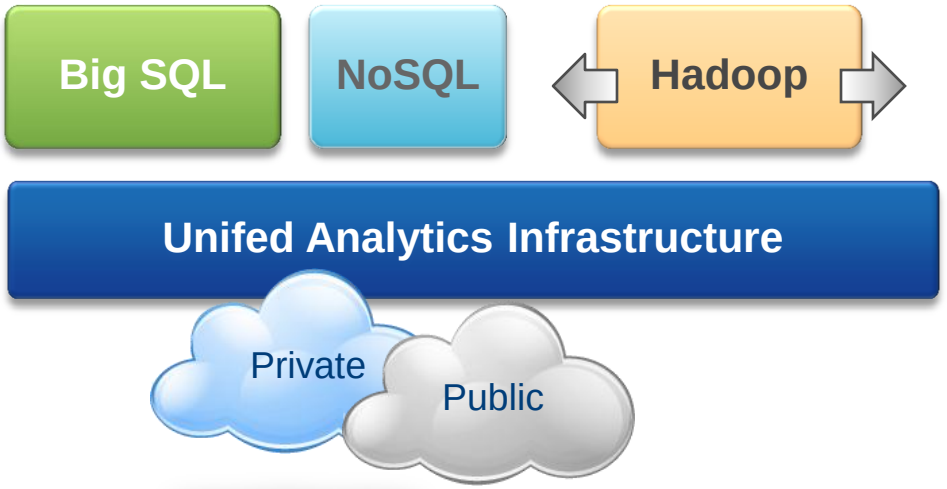
SQLCluster

NoSQL Cluster

Hadoop Cluster

Decision Support Cluster

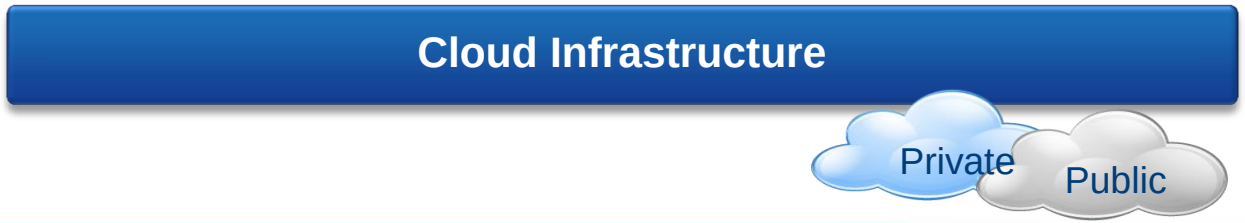
- **Simplify**
  - Single Hardware Infrastructure
  - Faster/Easier provisioning



- **Optimize**
  - Shared Resources = higher utilization
  - Elastic resources = faster on-demand access

Unifying the Big Data Platform using Virtualization

- **Goals**
  - Make it fast and easy to provision new data clusters on demand
  - Allow Mixing of Workloads
  - Leverage virtual machines to provide isolation (esp. for Multi-tenant)
  - Optimize data performance based on virtual topologies
  - Make the system reliable based on virtual topologies
- **Leveraging Virtualization**
  - Elastic scale
  - Use high-availability to protect key services, e.g., Hadoop's namenode/job tracker
  - Resource controls and sharing: re-use underutilized memory, cpu
  - Prioritize Workloads: limit or guarantee resource usage in a mixed environment



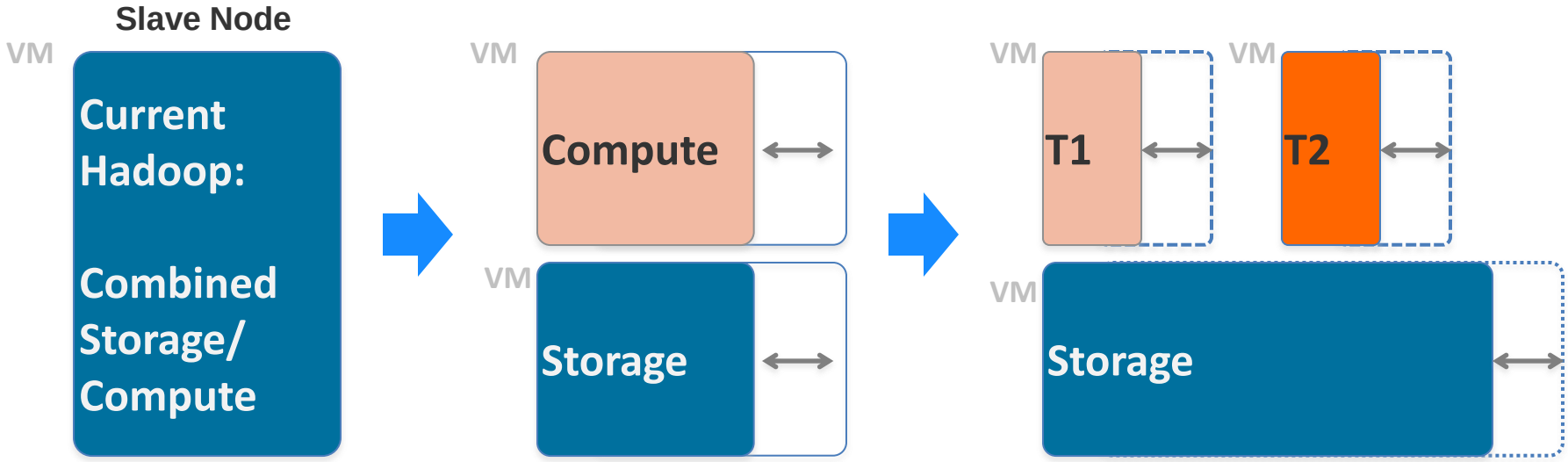
VMware is committed to be the Best Virtual platform for Hadoop

- **Performance Studies and Best Practices**
  - Studies through 2010-2011 of Hadoop 0.20 on vSphere 5
  - White paper, including detailed configurations and recommendations
- **Making Hadoop run well on vSphere**
  - Performance optimizations in vSphere releases
  - VMware engagement in Hadoop Community effort
  - Supporting key partners with their distributions on vSphere
  - Contributing enhancements to Hadoop
  - Automate Hadoop deployment on vSphere

Serengeti

- **Hadoop Framework Integration**
  - Spring for Hadoop: Enabling Spring to simplify Map-Reduce Jobs
  - Spring Batch: Sophisticated batch management

Evolution of Hadoop on VMs



Hadoop in VM

- VM lifecycle determined by Datanode
- Limited elasticity
- Limited to Hadoop Multi-Tenancy

Separate Storage

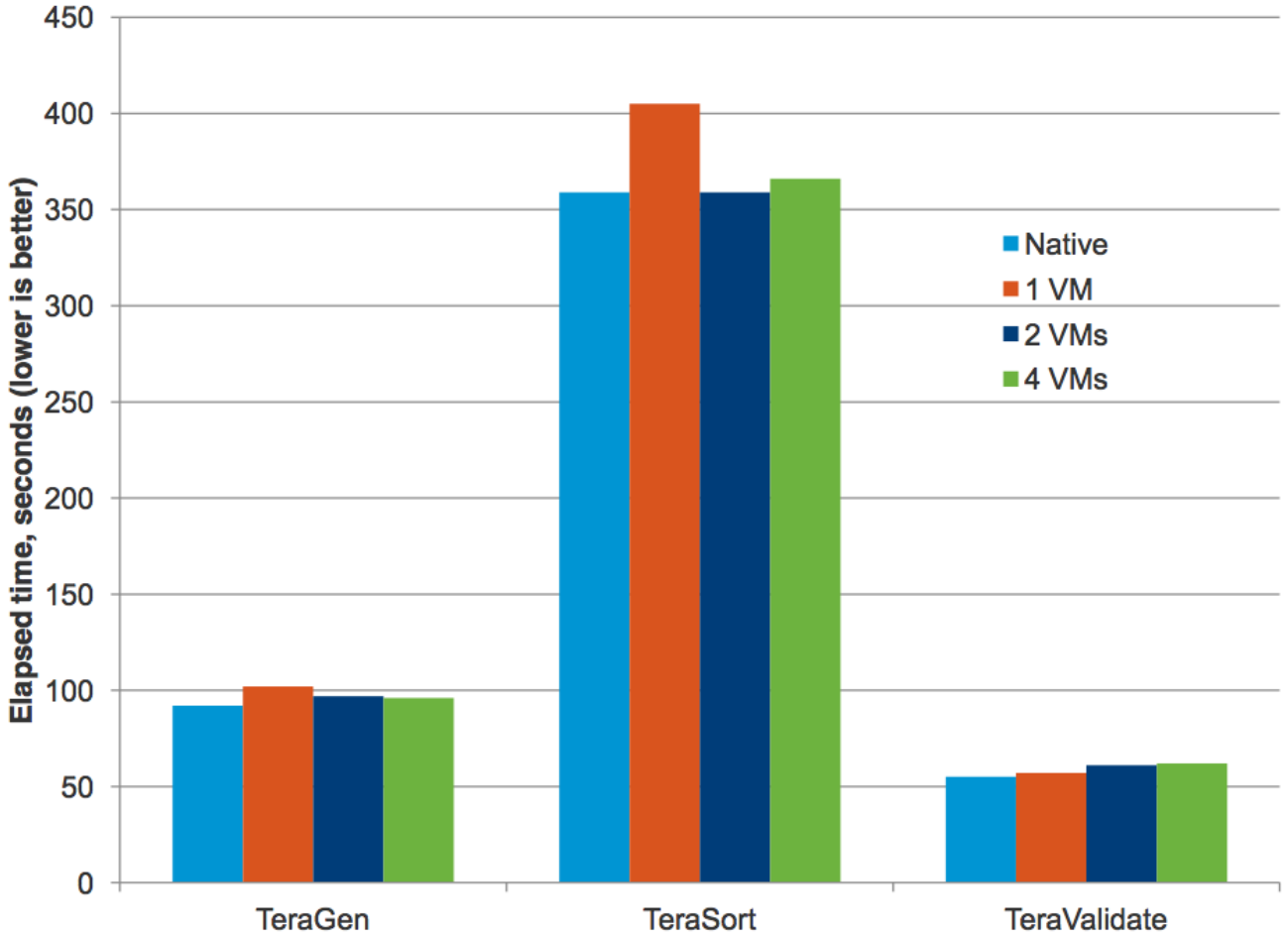
- Separate compute from data
- Elastic compute
- Enable shared workloads
- Raise utilization

Separate Compute Clusters

- Separate virtual clusters per tenant
- Stronger VM-grade security and resource isolation
- Enable deployment of multiple Hadoop runtime versions

Performance Analysis of Hadoop on Virtualization

Ratio of time taken – Lower is Better



Source: http://www.vmware.com/files/pdf/techpaper/VMW-Hadoop-Performance-vSphere5.pdf

Project Serengeti

- Open source project launched in June, 2012
- Toolkit that leverage virtualization to simplify Hadoop deployment and operations
- To learn more, projectserengeti.org

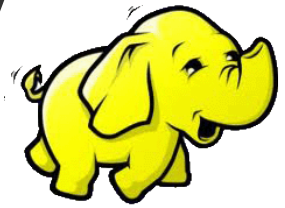
- Deploy a Hadoop cluster in 10 Minutes
- Customize Hadoop cluster
- Use Your Favorite Hadoop Distribution
- One stop command center

Project HVE (Hadoop Virtualization Extensions)

- **Open Source project on Hadoop code base**
  - Deliver patches to Apache Open Source community
  - Work with Hadoop distro vendors

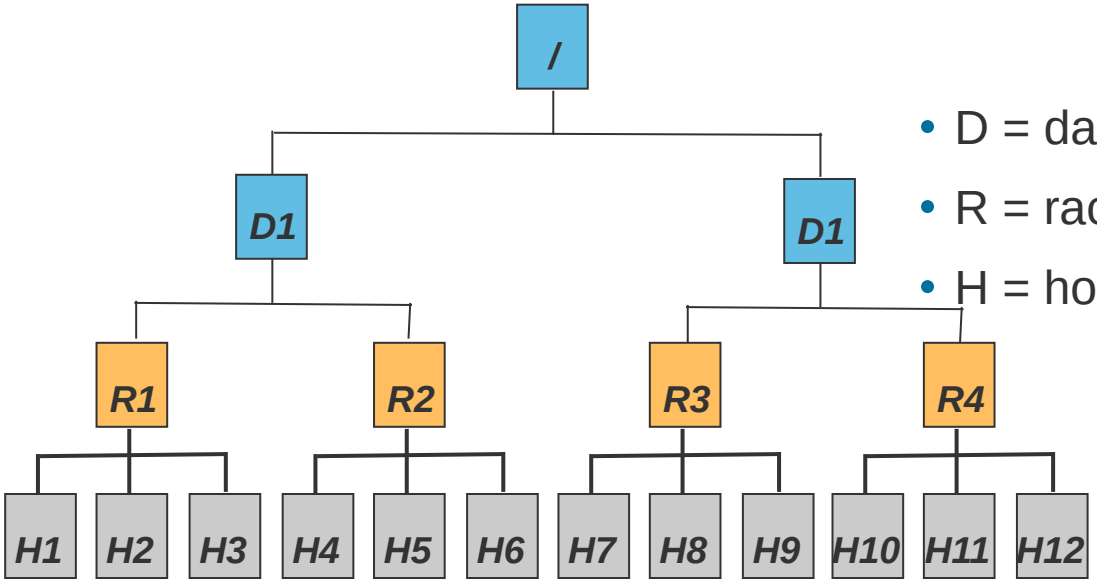
- **Refine Hadoop for running on virtualized infrastructure**
  - Enable multiple-layer network topology
  - Enable resource sharing/over-commitment
  - Enable compute/data node separation without losing locality

- **100% Contribution back to Apache Hadoop Community**
  - http://www.vmware.com/hadoop
  - Umbrella JIRA: HADOOP-8468
  - Sub JIRAs: HADOOP-8469, HADOOP-8470, HADOOP-8817, HDFS-3495, HDFS-3498, HDFS-3461, YARN-18, YARN-19, etc.



vmware

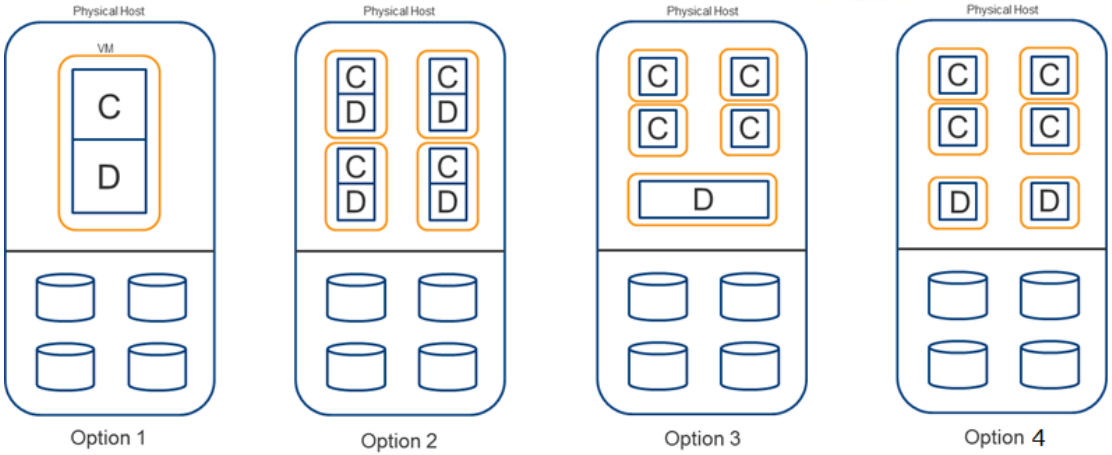
Current Network Topology



- D = data center
- R = rack
- H = host

However, you have more choices on virtualized infrastructure

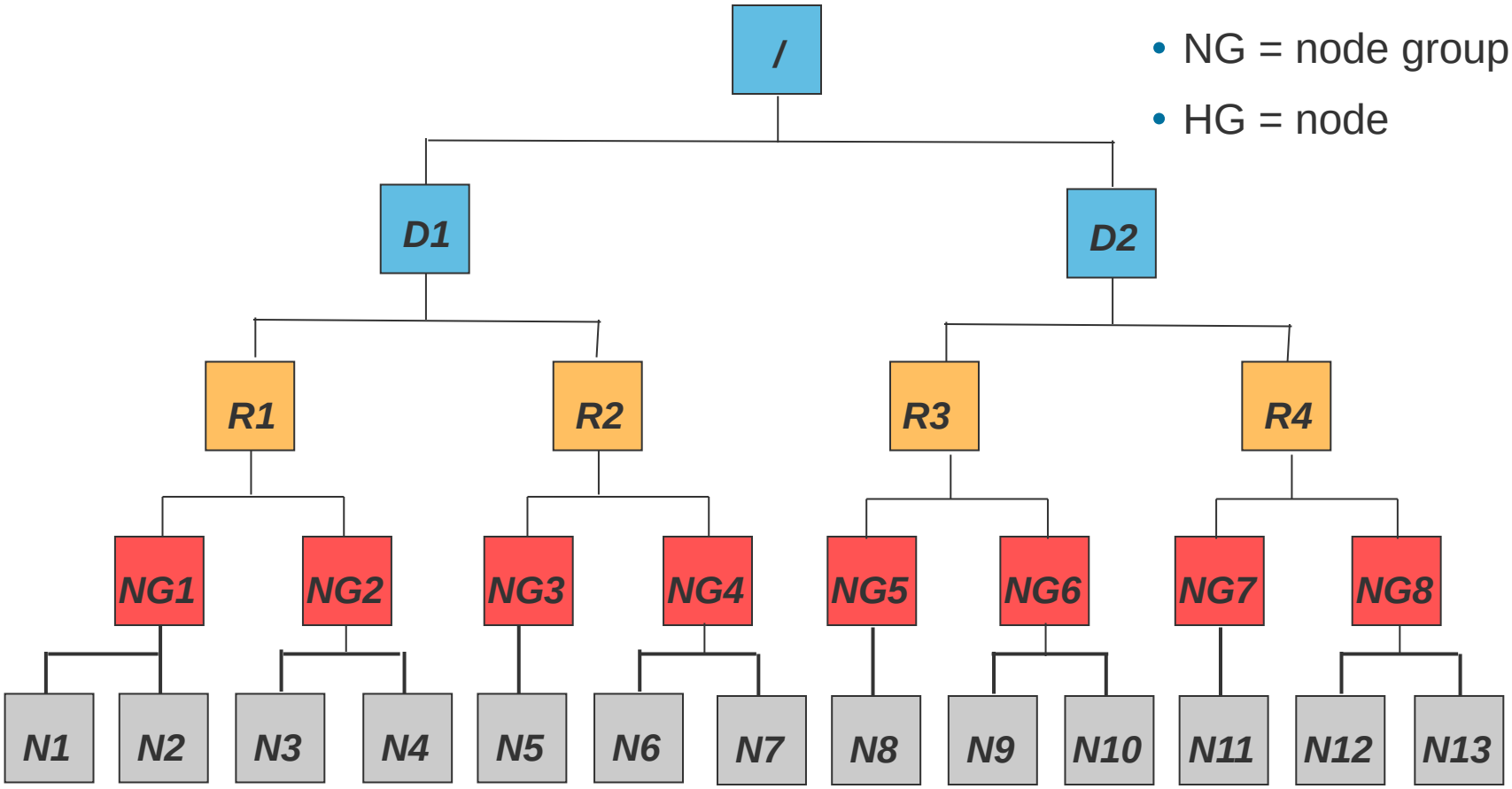
- C = compute node (TaskTracker)
- D = data node



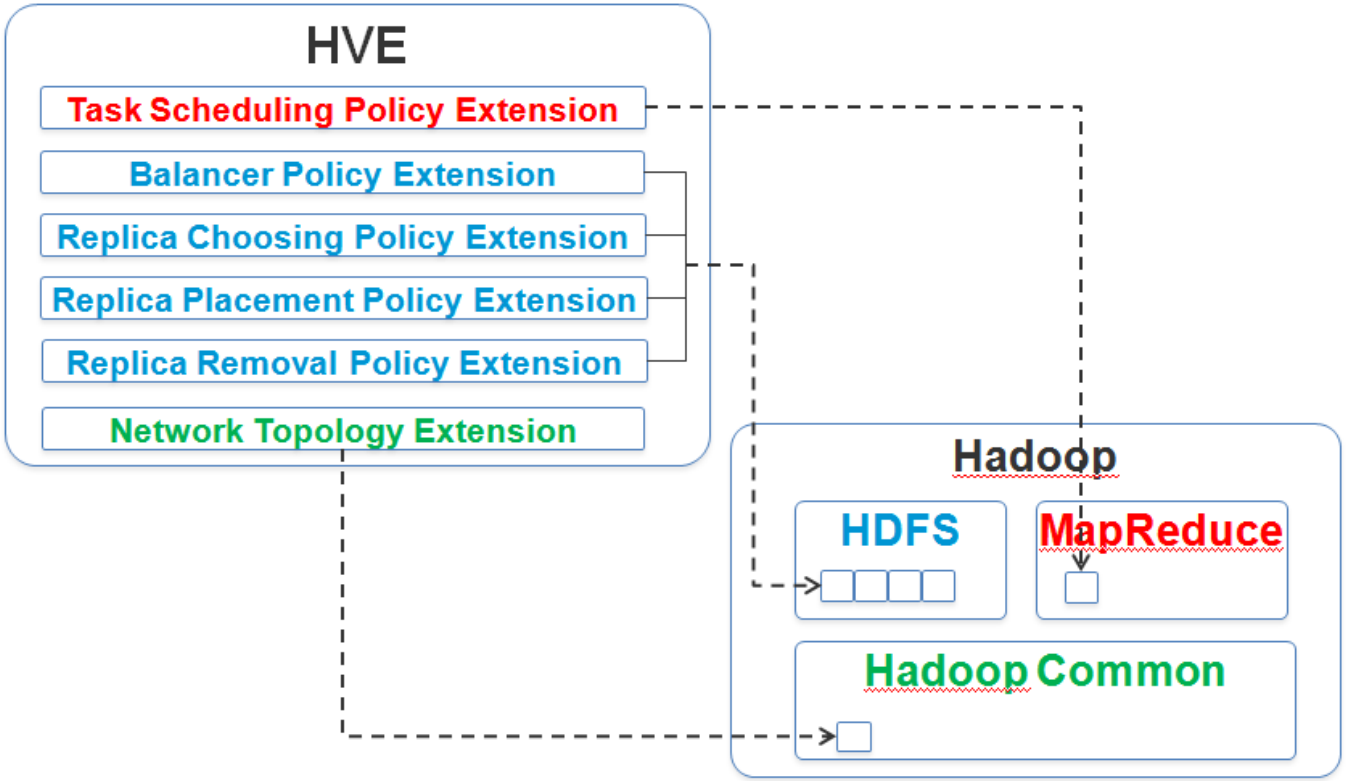
vmware

Additional network topology layer to aware virtualization

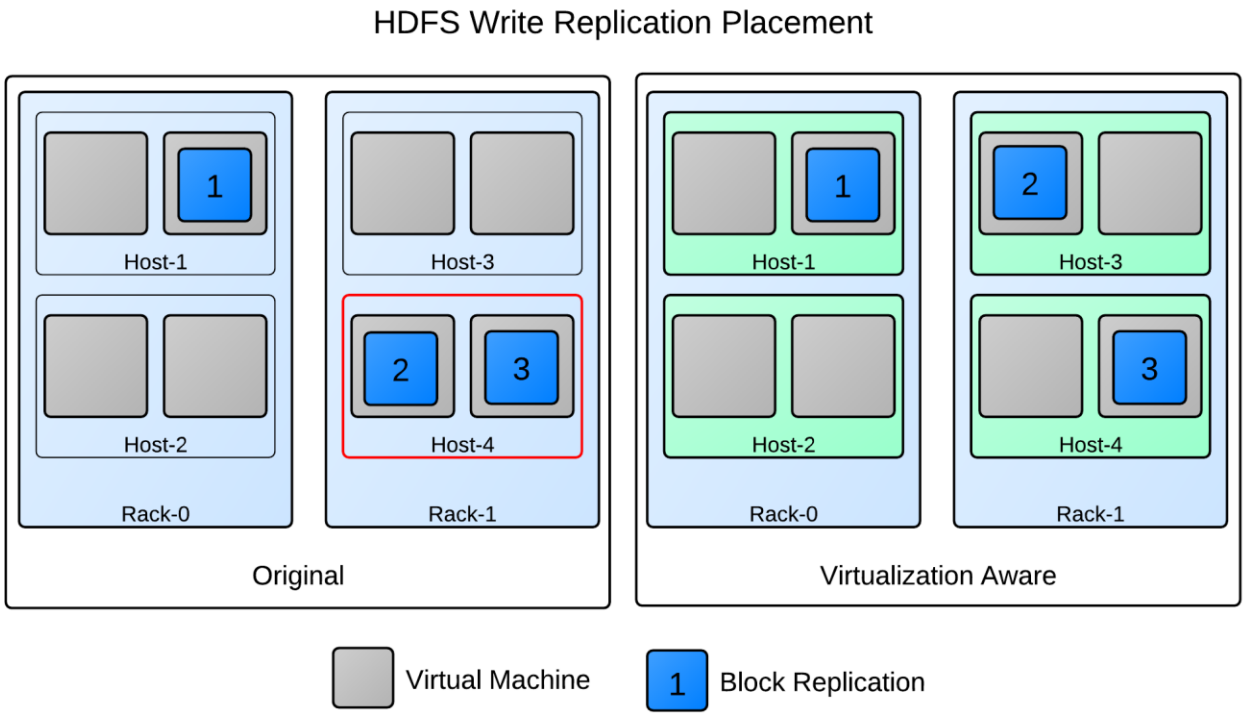
- D = data center
- R = rack
- NG = node group
- HG = node



High Level View on HVE changes



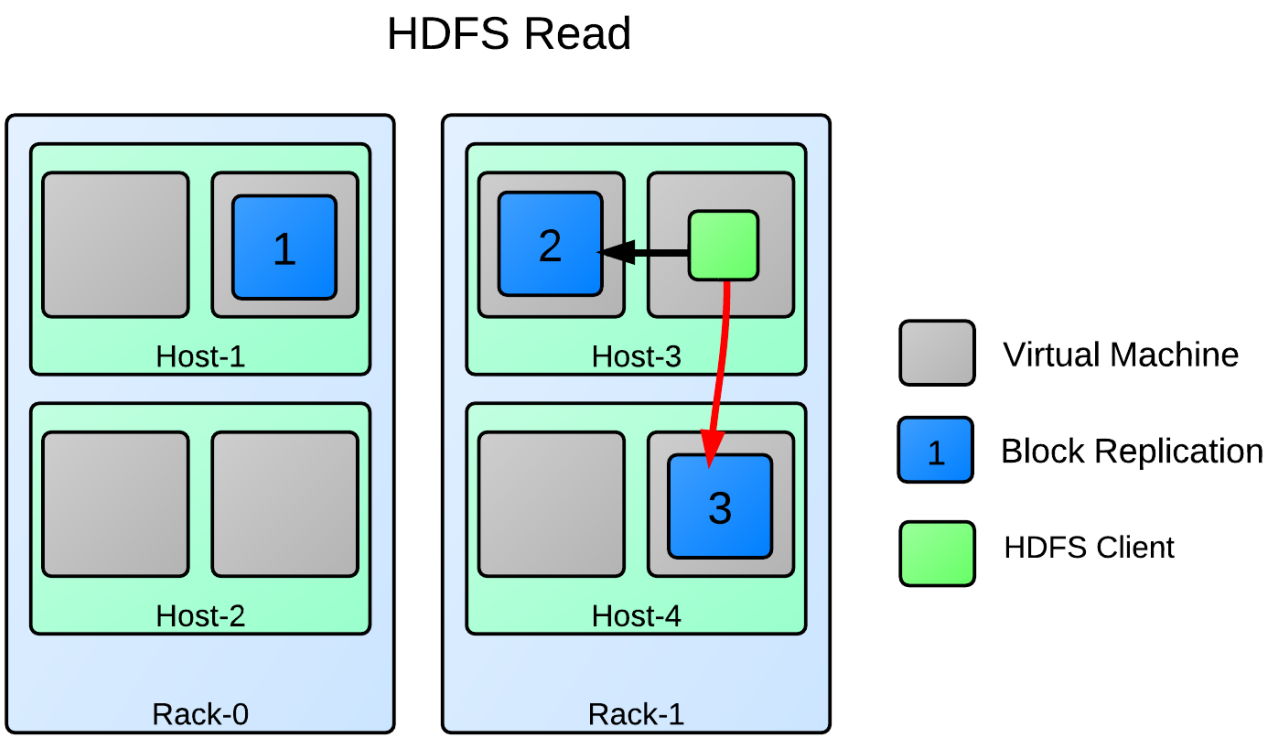
“Virtualization Aware” Replica Placement Policy



Updated Policies:

- No replicas are placed on the same node or **nodes under the same node group**
- 1st replica is on the local node or **one of nodes under the same node group** of the writer
- 2nd replica is on a remote rack of the 1st replica
- 3rd replica is on the same rack as the 2nd replica
- Remaining replicas are placed randomly across rack to meet minimum restriction.

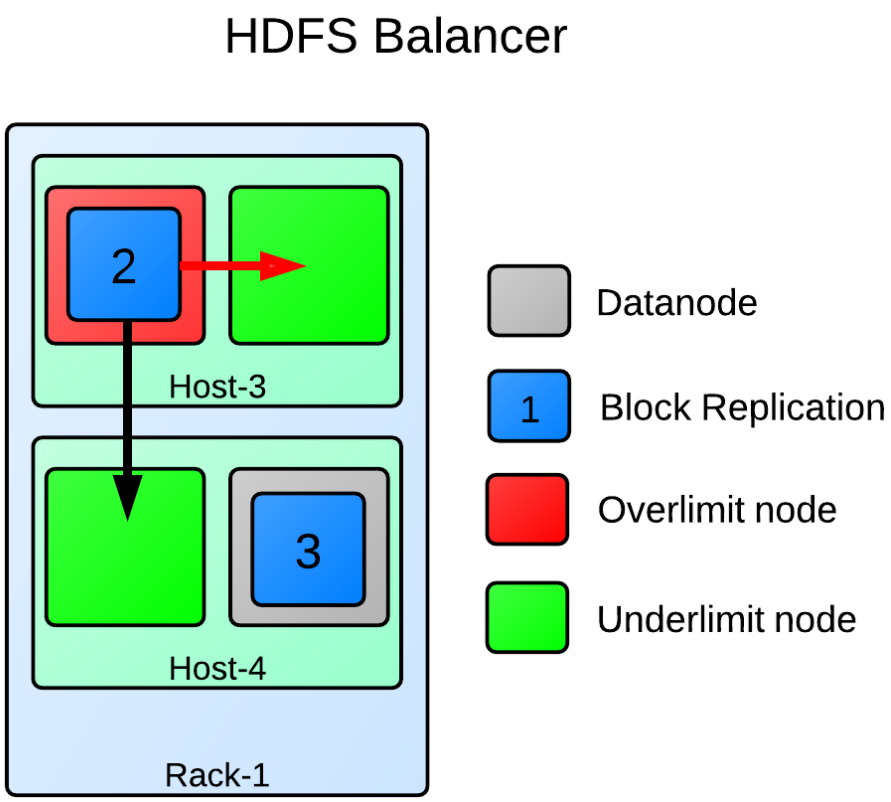
“Virtualization Aware” Replica Choosing Policy



Distances for data locality:

- Node local (0)
- **Node group local (2)**
- Rack local **(4)**
- Off rack **(6)**

“Virtualization Aware” Balancer Policy

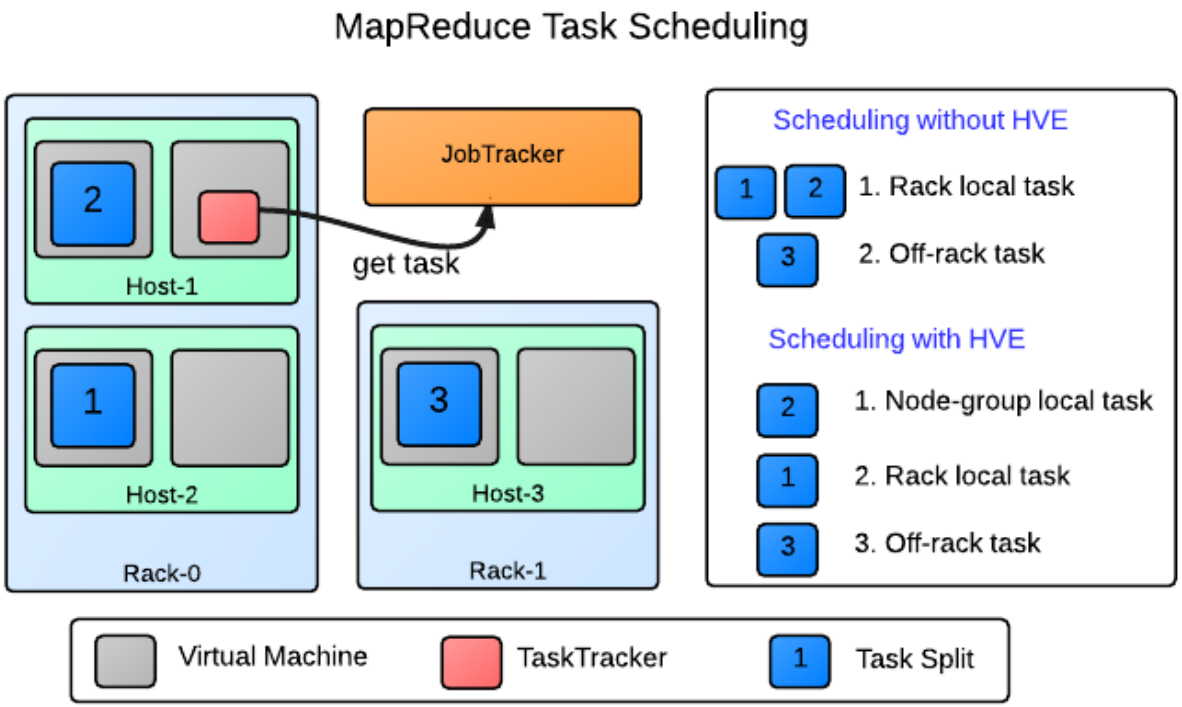


- Balancer policies contains two levels choosing policy

  - choosing node pairs of source and target, in sequence of: **local node group**, local rack, off rack

  - choosing blocks to move within node pair, a replica block is not a good candidate if another replica is on the target node **or on the same node group** of the target node

“Virtualization Aware” Task Scheduling Policy



Get task split for TaskTracker or NodeManager in following sequences:

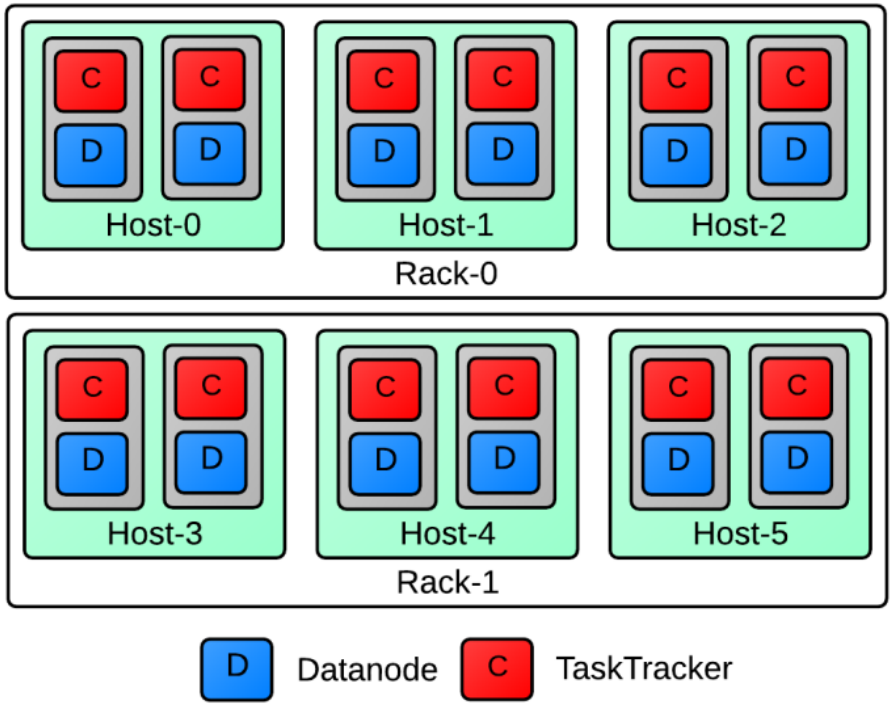
- Node local
- **Node group local**
- Rack local
- Off rack

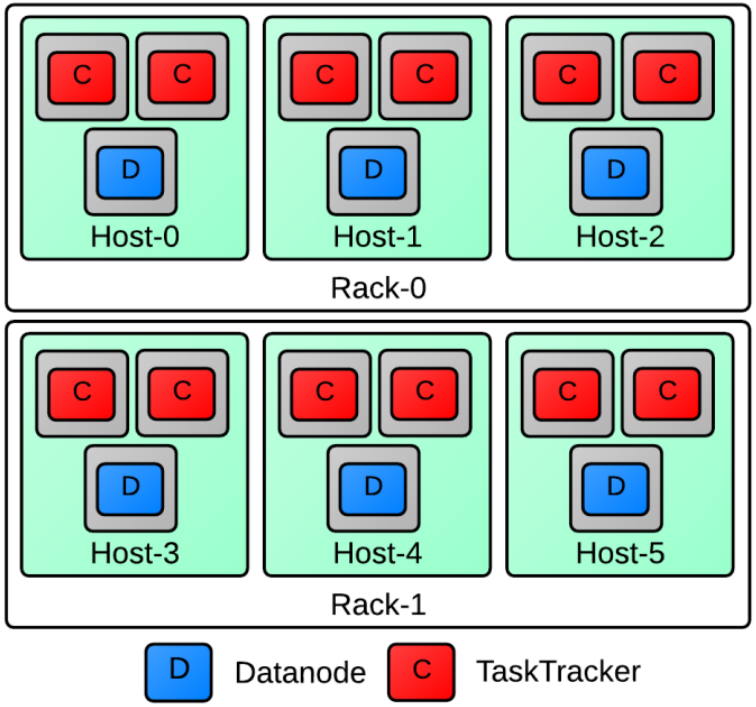
It works well with

- FifoScheduler
- FairScheduler
- Capacity scheduler

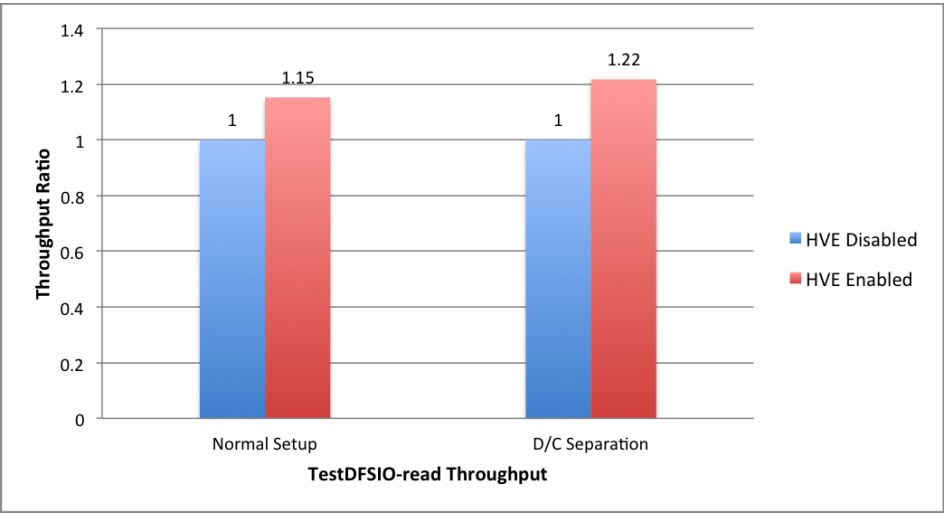
HVE Topology Benchmark Result

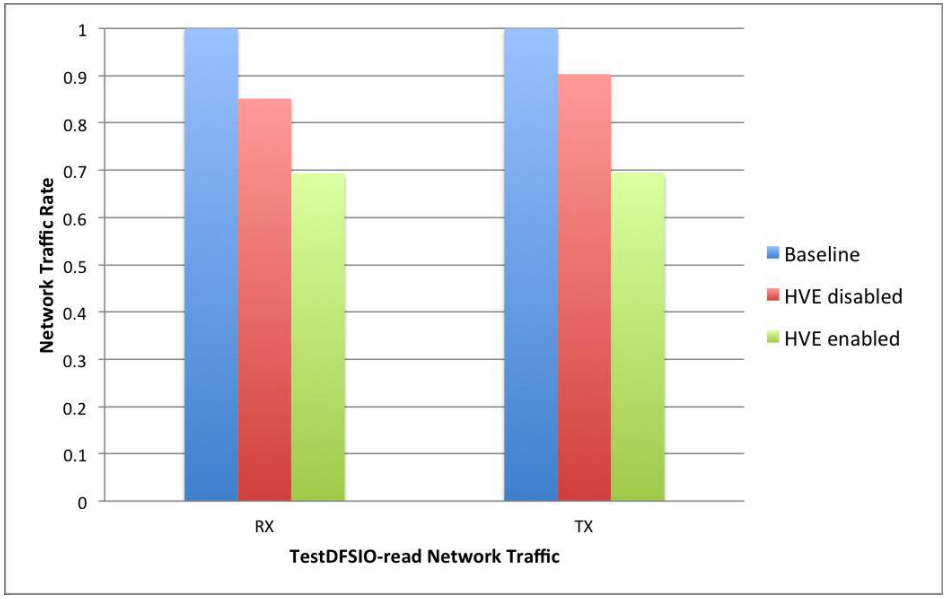
- **Integrate HVE with Apache Hadoop 1.0.3**
- **Cluster Deployments**
  - 6 physical nodes
  - 12 virtual nodes (combined case), 18 virtual nodes (d/c separation case)

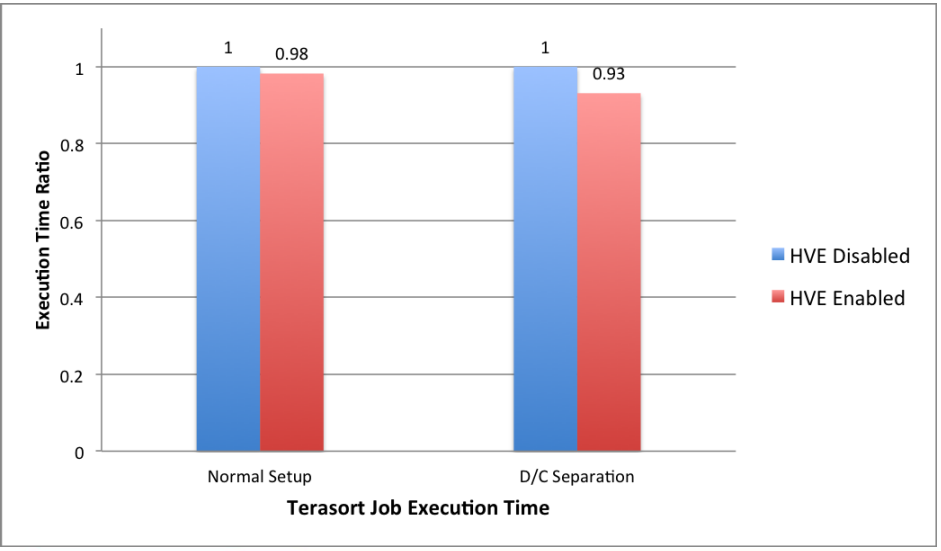




HVE Topology Benchmark Result



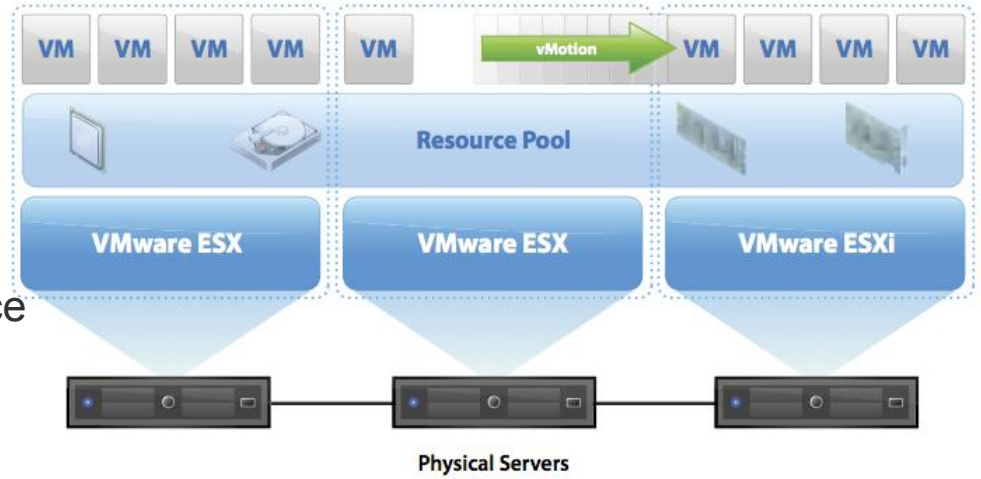




Terasort locality	Data Local	Node-group Local	Rack Local
Normal	392	-	8
Normal with HVE	397	2	1
D/C separation	0	-	400
D/C separation with HVE	0	400	0

HVE for Resource Elasticity

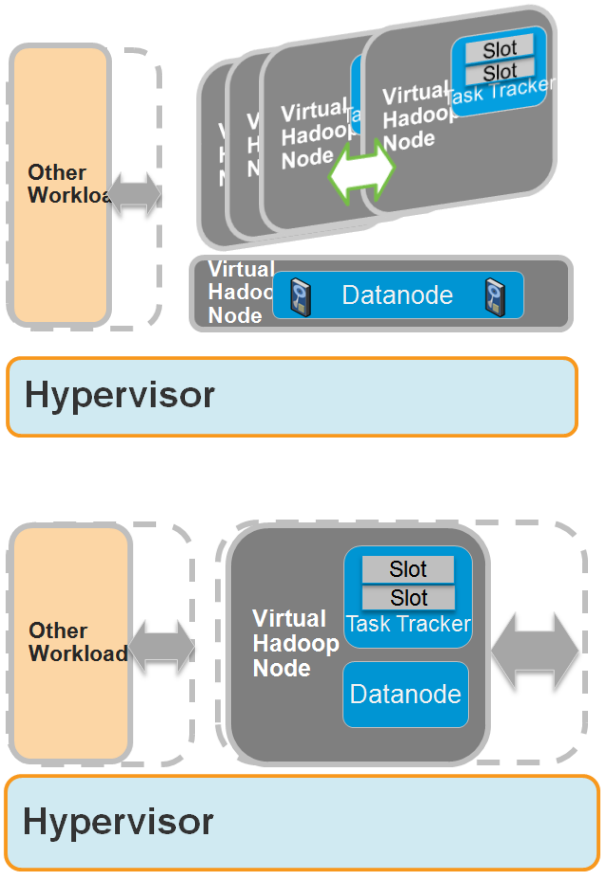
- **Resource Elasticity in cloud scenario**
  - Resource sharing environment
  - Different types of workloads: cpu-bound, I/O-bound, etc.
  - Different peak time for Apps
  - It is a perfect chance to achieve high resource utilization
- **How could we achieve this?**
  - Art of scheduling
  - Schedule Apps (VMs) to Resources
    - DRS, based on vMotion
  - Schedule resources to Apps(VMs)
    - Scale up/down per node(VM)'s resource
    - Add more VMs



Elastic Resource on Virtualization

- **Schedule resources to Apps**
  - A policy-based Cloud Apps Resource Manager can monitor resource usage for each App
  - Trigger on-demand resource movement among Apps

- **Elastic Hadoop cluster**
  - Horizonal scaling: scale in and out (node number)
    - Data/compute node separation
    - Bring up/down compute nodes
  - Vertical scaling: scale up and down (node size)
    - Resource over-commitment
  - Mixed



Summary

- **Big Data application going to Cloud is under way**
  - Get simplified and optimized
- **Hadoop on Virtualization**
  - Proven performance
  - Cloud/Virtualization values apparent for Hadoop use
  - Project Serengeti – Simplify Hadoop deployment and operations
- **Project HVE (Hadoop Virtualization Extensions)**
  - Enhance Hadoop running on Virtualization by bring more virtualization awareness to Hadoop
    - Virtualization-aware Network Topology
    - Virtualization-aware Resource Scheduling
    - More in future

References

- **Hadoop at VMware**

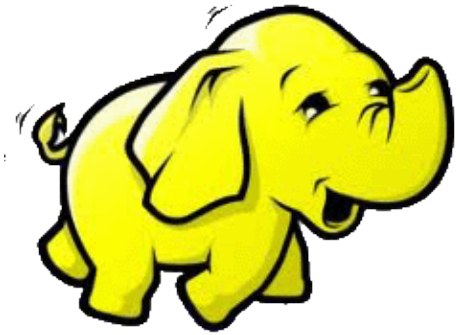
  www.vmware.com/hadoop

- **Project Serengeti**

  projectserengeti.org

- **Project HVE**
  - HVE Whitepaper: http://serengeti.cloudfoundry.com/pdf/Hadoop%20Virtualization%20Extensions%20WP.pdf
  - Umbrella Jira:

    https://issues.apache.org/jira/browse/HADOOP-8468



- **Hadoop on vSphere**
  - Talks @ Hadoop World, Hadoop Summit
  - Performance Paper

    http://www.vmware.com/files/pdf/techpaper/VMW-Hadoop-Performance-vSphere5.pdf

Q & A

Thank you!