

按需启用高可用，弹性，多租户的Hadoop

董波, 产品线经理

dbo@vmware.com

VMware Inc.

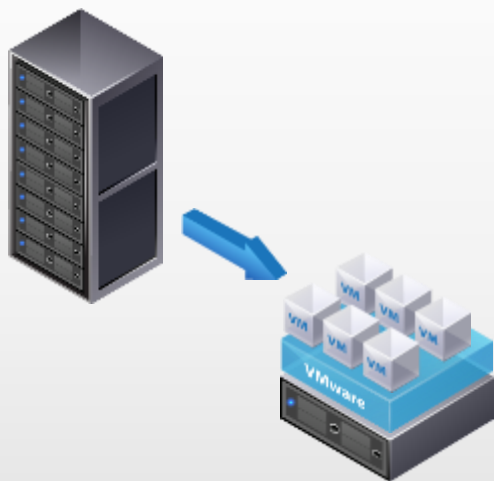
议程

- 云计算的好处
- 消除误解
- 为何要虚拟化
- 总结
- Q & A

云计算:带来简便、优化的重大变革

1. 降低复杂性

简化运维



2. 显著降低成本

资金更多投入到增值业务中



3. 启动灵活敏捷的IT服务

满足业务需求



议程

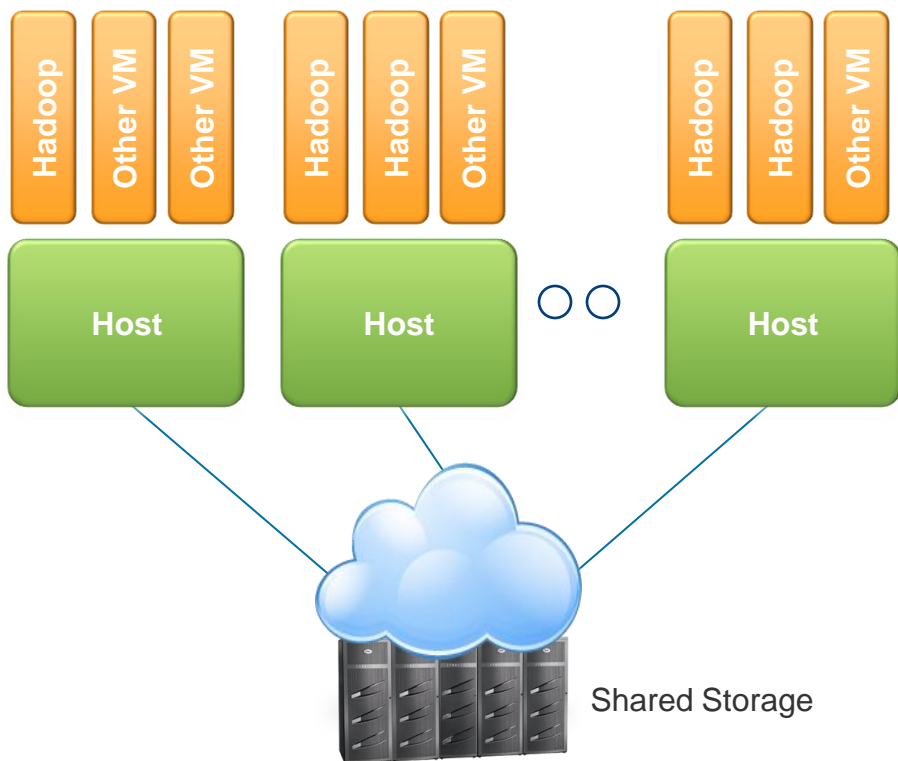
- 云计算的好处
- 消除误解
- 为何要虚拟化
- 总结
- Q & A

- 虚拟机运行在SAN上，Hadoop却是运行在本地磁盘上的
- 虚拟机会带来额外负载，使得Hadoop性能降低很多

包含本地磁盘的虚拟存储架构

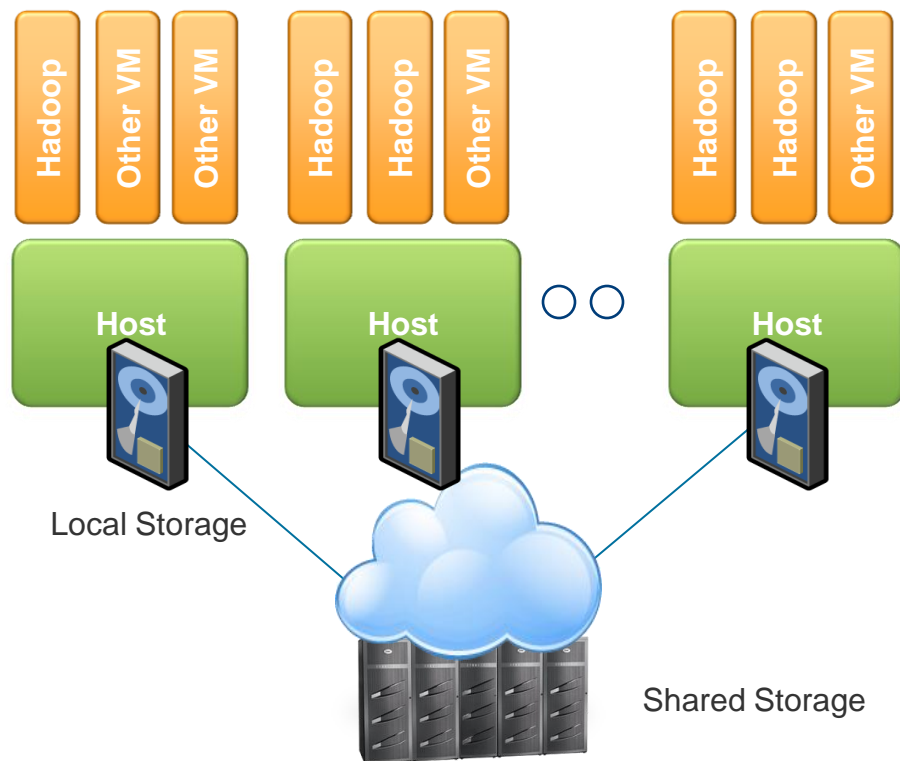
■ 共享存储: SAN 或 NAS

- 部署方便简单
- 集群的自动平衡
- 利用 vMotion/HA/FT技术

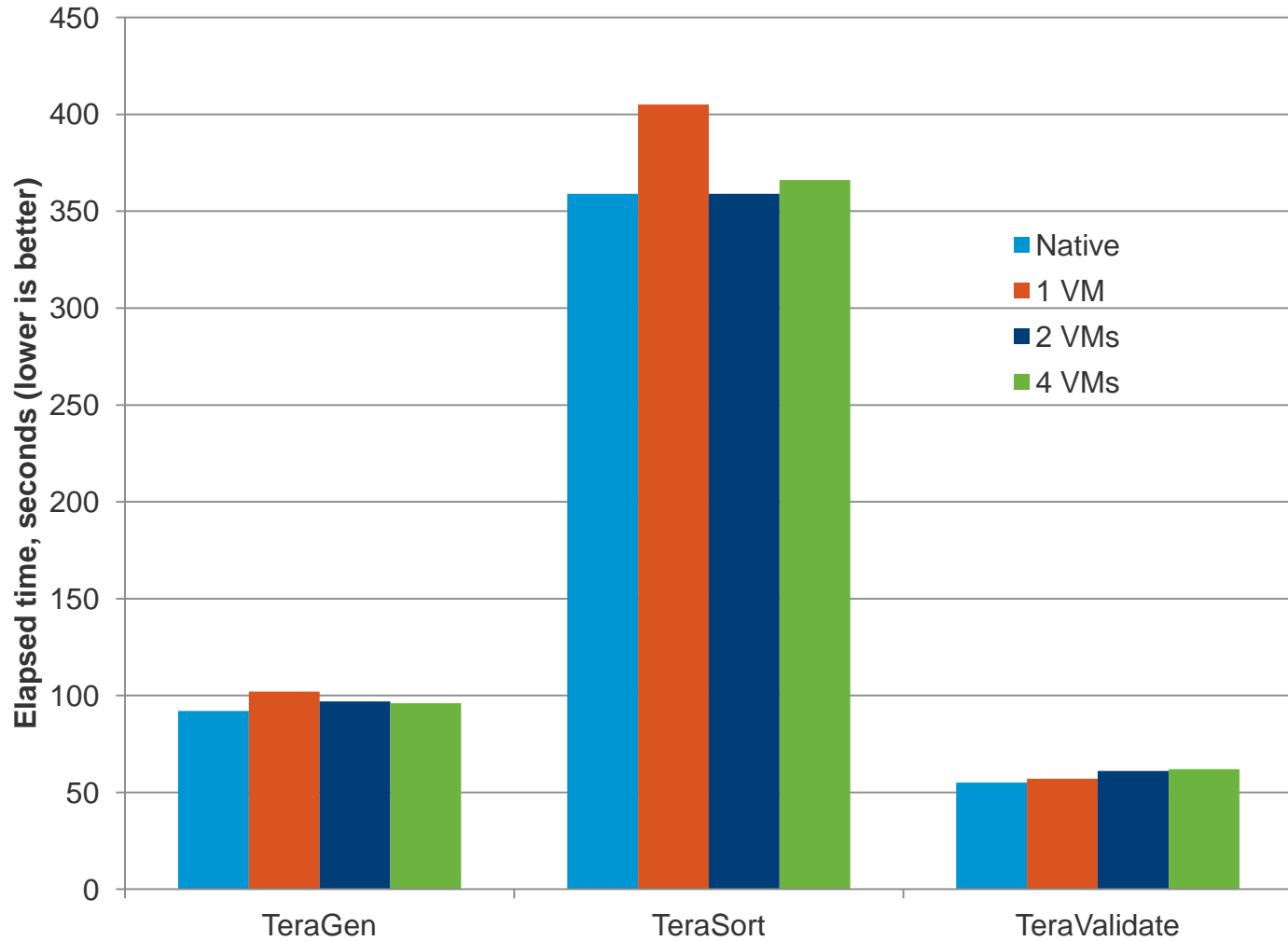


■ 本地存储: 本地磁盘

- 为Hadoop使用本地磁盘
- 易扩展的带宽，每GB更低费用



Hadoop在虚拟化平台上运行良好



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

议程

- 云计算的好处
- 消除误解
- **为何要虚拟化**
- 总结
- Q & A

企业使用Hadoop所面临的挑战

■ 部署

- 部署费时费力
- 系统调优困难

■ 单点失效

- Name Node和Job tracker存在单点失效问题
- 相关非核心Hadoop模块如Hive, HCatalog, 等没有HA保障

■ 利用率低

- 专用Hadoop集群CPU使用率低
- Hadoop和非Hadoop负载不能直接共享资源
- 缺乏资源管控

■ 需要多租户隔离

- 不同用户间缺乏足够的性能和安全隔离机制
- 无法实现配置隔离

Hadoop在企业内部的使用

Integrated

Standalone

0 node

20

300

Scale

阶段一: 试点POC

- ✓ 经常从业务线开始
- ✓ 使用1-2个用例验证Hadoop价值
- ✓ 典型应用一般在20个节点以下

Stage 2: Hadoop 生产应用

- ✓ 为一些部门服务
- ✓ 更多使用用例
- ✓ 核心Hadoop和其他相关软件
- ✓ 几十个到数百个节点的典型规模

Stage 3: 大数据生产应用

- ✓ 为许多部门服务
- ✓ 经常支持一部分关键任务流程
- ✓ 与其他大数据服务整合
如MPP DB, NoSQL等

Stage 1: 试点POC

Stage1: 试点POC

- ✓ 经常从业务线开始
- ✓ 使用1-2个用例验证Hadoop价值
- ✓ 典型应用一般在20个节点以下
- ✓ 数据组或是基础架构组主导

■ 需求:

✓ 快速

- ✓ 不希望等待数周甚至数月
- ✓ 快速得到Hadoop集群

✓ 便捷

- ✓ 能方便地访问数据
- ✓ 可以方便的使用不同算法和数据集

为何要在此阶段进行虚拟化?

- 利用vSphere基础架构和硬件→无需等待
- 应用空闲机器为Hadoop试点项目提供资源→无需购买额外的软硬件
- 使用资源池和DRS技术为Hadoop试点项目提供计算资源→对当前负载无影响
- 共享存储上已经存在有数据→无需迁移数据

→此阶段利用虚拟化技术是不明智的

Serengeti项目

- 2012年6月启动的开源项目，定期发布更新（三个月一个周期）
- 利用虚拟化技术来简化Hadoop部署的管理的工具包
- 了解更多信息,请访问 projectserengeti.org



Serengeti

10分钟之内部署一个Hadoop集群

可定制Hadoop集群

使用您最青睐的Hadoop发行版

一站式命令操作中心

Demo : 使用Serengeti 10分钟布署一个Hadoop集群

Computer



```
Serengeti CLI 0.5.0
[serengeti@localhost ~]$ serengeti
-----
*
* Serengeti *
*
*
*
*
*
*
-----
Version:0.5.0
Welcome to Serengeti CLI
serengeti>cluster create --name demo1
```



自定义Hadoop集群

```
...
"distro": "apache",
"groups": [
  { "name": "master",
    "roles": [
      "hadoop_namenode",
      "hadoop_jobtracker" ],
    "storage": {
      "type": "SHARED",
      "sizeGB": 20},
    "instance_type": MEDIUM,
    "instance_num": 1,
    "ha": true},
  { "name": "worker",
    "roles": [
      "hadoop_datanode",
      "hadoop_tasktracker" ],
    "instance_type": SMALL,
    "instance_num": 5,
    "ha": false
  }
]
...
```

- 选择发布版
- 设定存储
 - 可使用共享存储或本地硬盘
- 设定资源
- 高可用
- 节点数

加速使用Serengeti

■ Serengeti 作为一站式命令中心

■ 部署和管理Hadoop 集群

```
> cluster create --name <clustername>
```

■ 上传和下载数据

```
> fs ls /tmp  
> fs put --from /tmp/local.data --to /tmp/hdfs.data
```

■ 从Serengeti CLI使用 MapReduce/Pig/Hive 任务

```
> cluster target --name myHadoop  
> mr jar --jarfile /opt/serengeti/cli/lib/hadoop-examples-1.0.1.jar  
--mainclass org.apache.hadoop.examples.PiEstimator --args "100 1000000000"
```

■ 为ODBC/JDBC 服务部署Hive Server

```
"name": "client",  
  "roles": [  
    "hadoop_client",  
    "hive",  
    "hive_server",  
    "pig"  
  ], ...
```


阶段2: Hadoop 生产应用

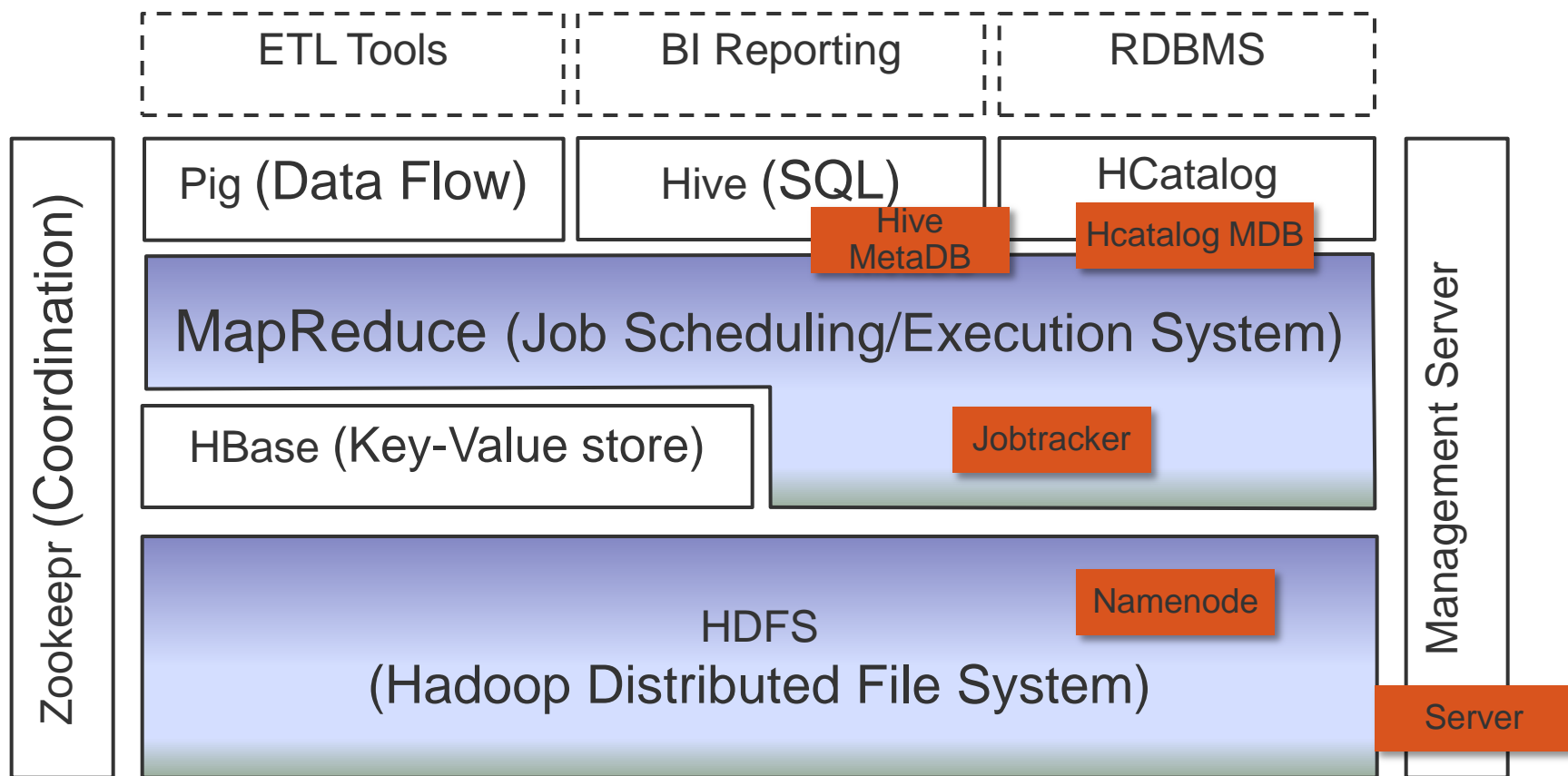
阶段 2: Hadoop 产 品化

- ✓ 为一些部门服务
- ✓ 更多使用用例
- ✓ 核心Hadoop和其他相关非核心软件
- ✓ 成百上千个节点的典型规模
- ✓ 专用的Hadoop管理员

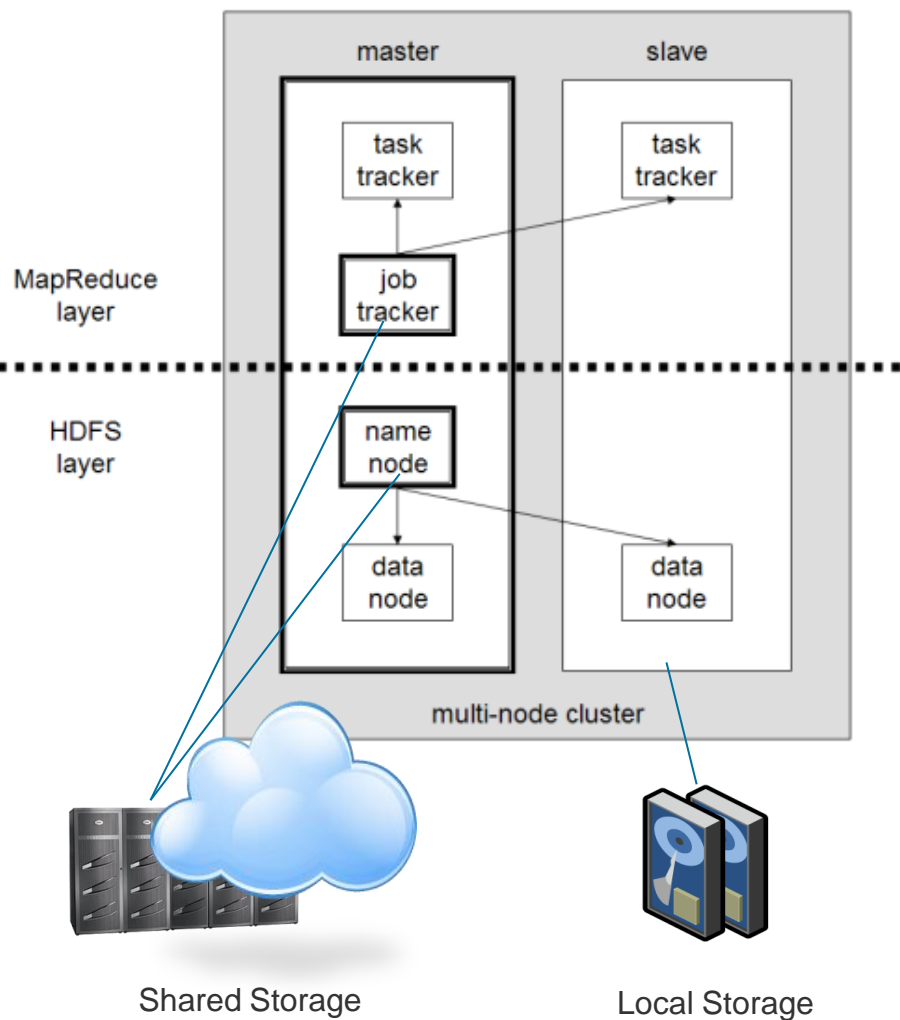
需求:

- ✓ 高可用性
 - ✓ 我们是产品环境，需要一定的产品服务等级
 - ✓ 为Hadoop应用提供整体的高可用性解决方案
- ✓ 敏捷性
 - ✓ 我们一直在搜集Hadoop使用的最新需求，我希望能够很容易地扩展Hadoop集群
 - ✓ 我们需要经常配置Hadoop集群
- ✓ 差异化的服务等级
 - ✓ 我们要运行产品线上的Hadoop任务，需要保证高优先级
 - ✓ 我们也需要满足随机的Hadoop任务请求

HA for Hadoop stack 不仅仅是Name node HA



混合存储模式提供最好的存储算法



- Master nodes:
 - 在共享存储上搭建Name node, job tracker etc.
 - 利用 vSphere vMotion, HA 和FT
- Slave nodes
 - 在本地存储上搭建Task tracker/data node
 - 低成本，可扩展的带宽

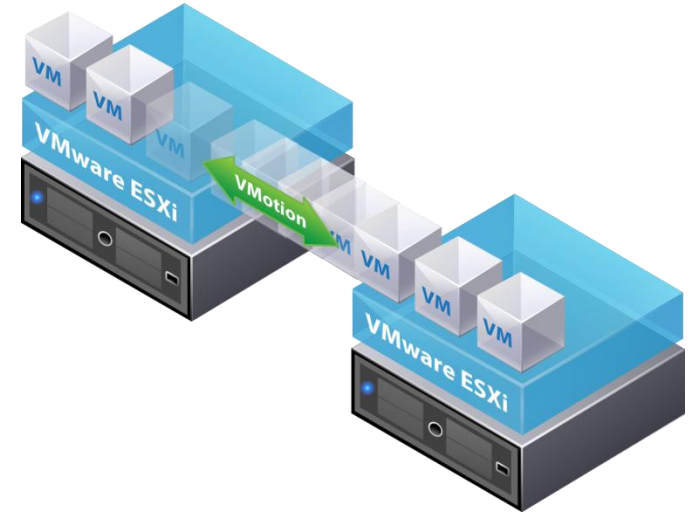
vMotion 降低计划宕机时间

描述:

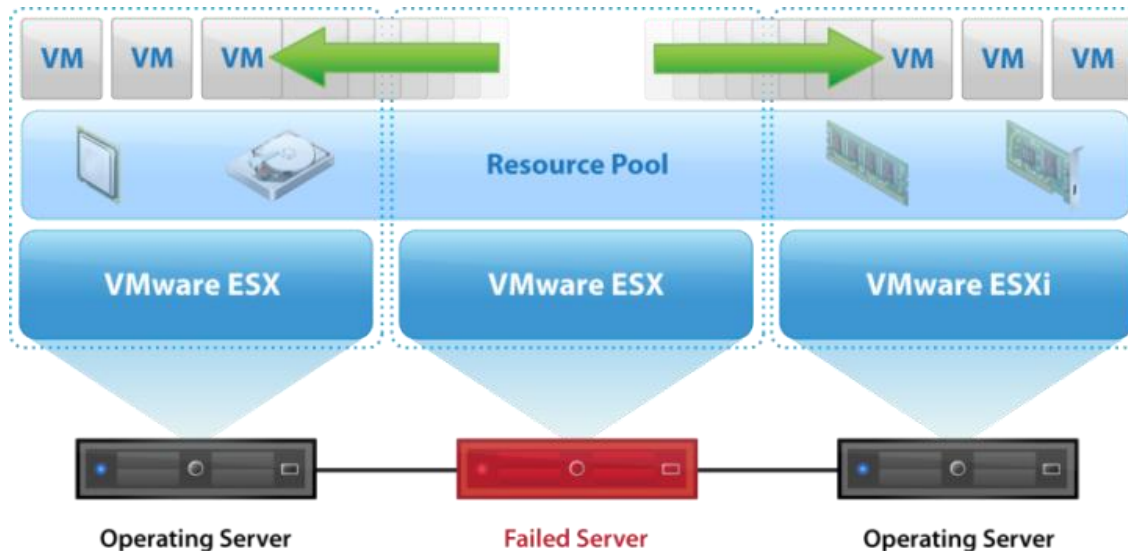
为虚拟机开启host之间的热迁移，提供持续性服务

优势:

- 虚拟机自动迁移的革命性技术满足了服务级别和性能目标



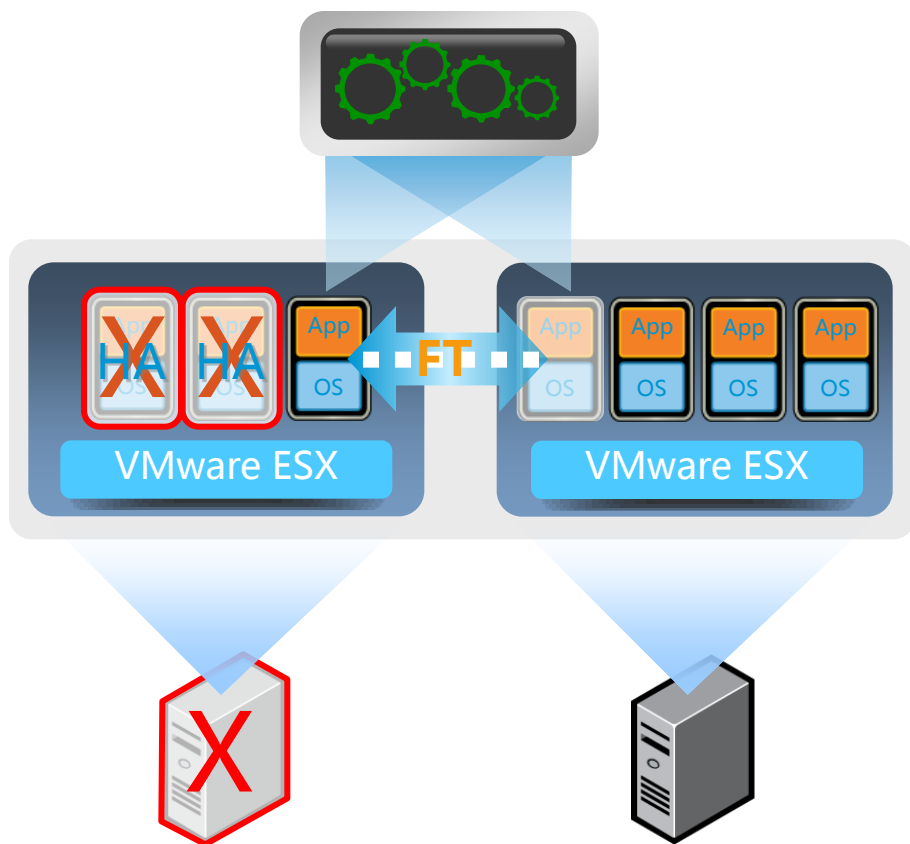
Hadoop HA-减少非计划宕机时间



Overview

- 监控物理主机或虚拟机故障
- 应用级别的HA：防止NameNode和Jobtracker失败
- 故障检测并于数分钟内自动重启虚拟机
- 不需要复杂的设定
- Hadoop任务恢复

vSphere Fault Tolerance 提供持续性保护



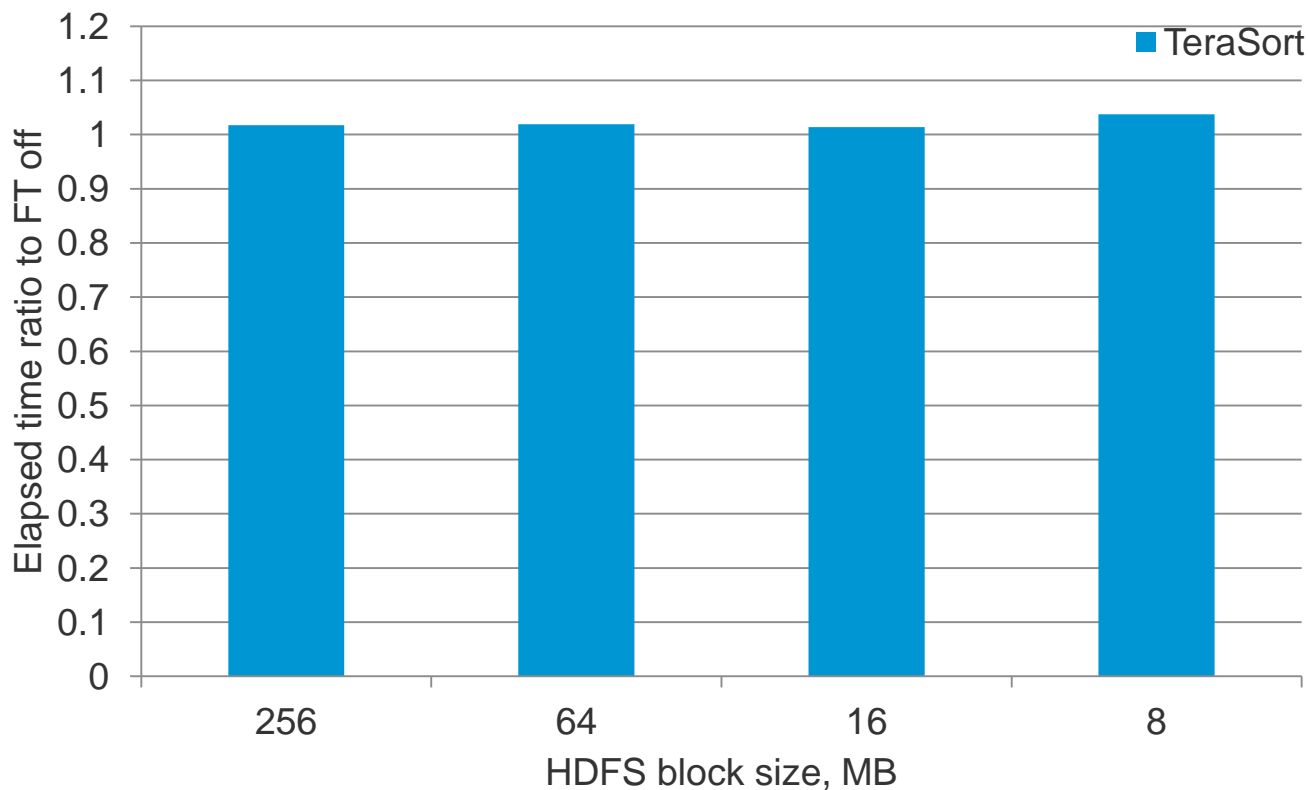
Overview

- 一个虚机的多个实例在不同的服务器上同步运行
- 硬件故障发生时系统**无宕机，无数据丢失**
- 与VMware HA/DRS集成
- 无需复杂的集群或特殊设备
- 使用同一的机制保护所有的应用和操作系统

无宕机系统

Hadoop Master节点开启FT的性能分析

- NameNode和JobTracker分布在两个VM中
- 额外开销很小：对于TeraSort开启FT降低2-4%的性能
- 单节点的NameNode可以支持大于200个Hadoop节点



NameNode HA – 故障恢复时间

■ NameNode使用vSphere和Linux HA的故障恢复时间

- 故障发现 – 0.5至2分钟
- 操作系统启动 – 10-20秒
- Namenode 启动(从safemode退出)
 - 中小型集群 – 1到2分钟
 - 大型集群 – 5至15分钟

■ NameNode 启动时间测量

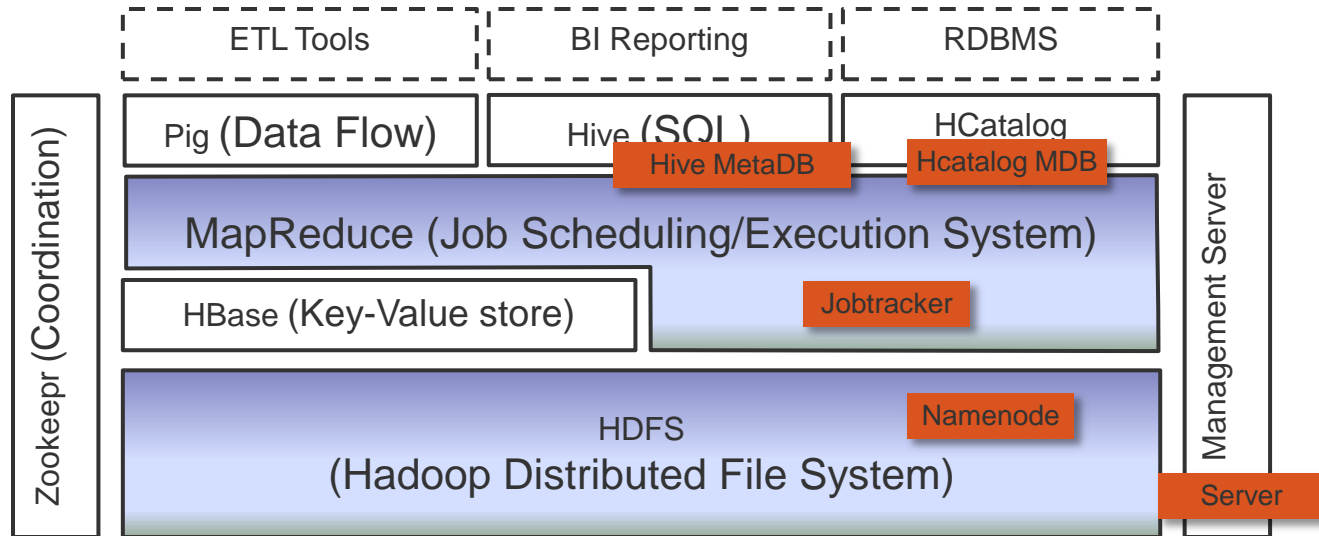
- 60节点, 60K文件, 6 百万blocks, 300 TB数据 – 40 秒
- 180 节点, 200K 文件, 18 million blocks, 900TB数据 – 120 秒

冷故障恢复可以很好满足中小型集群

故障检测和自动故障恢复

保护所有Hadoop Stack

- 实战验证的高可用性技术
- 对于所有Hadoop生态系统使用同一机制启动HA
- 一键式启动HA/FT



- HDFS 2 HA与之对比
 - 只覆盖NameNode – 其他更多的5个master服务怎么办？
 - Apache Hadoop 1不可用
 - 不像vSphere HA/FT经过实战验证
 - 安装和管理更复杂

利用Serengeti一个简单命令扩展集群

- `>cluster resize -name <clustername> --nodegroup worker -instanceNum <#>`



```
Serengeti CLI 0.5.0

SUCCESS 100%

node group: master, instance number: 1
roles:[hadoop_namenode, hadoop_jobtracker]
NAME          IP           STATUS      TASK
-----
demol-master-0 10.111.90.103 Service Ready

node group: worker, instance number: 3
roles:[hadoop_datanode, hadoop_tasktracker]
NAME          IP           STATUS      TASK
-----
demol-worker-0 10.111.90.89  Service Ready
demol-worker-1 10.111.90.114 Service Ready
demol-worker-2 10.111.90.100 Service Ready

node group: client, instance number: 1
roles:[hive, hadoop_client, pig]
NAME          IP           STATUS      TASK
-----
demol-client-0 10.111.90.115 Service Ready

cluster demol created
serengeti>
```

使用Serengeti简化配置Hadoop

■ 使用Serengeti修改Hadoop集群的配置

- 在json spec 文件中使用 “configuration” 字段
- 指定core-site.xml, hdfs-site.xml, mapred-site.xml, hadoop-env.sh, log4j.properties等文件中的Hadoop的属性

```
"configuration": {
  "hadoop": {
    "core-site.xml": {
      // check for all settings at http://hadoop.apache.org/common/docs/r1.0.0/core-default.html
    },
    "hdfs-site.xml": {
      // check for all settings at http://hadoop.apache.org/common/docs/r1.0.0/hdfs-default.html
    },
    "mapred-site.xml": {
      // check for all settings at http://hadoop.apache.org/common/docs/r1.0.0/mapred-default.html
      "io.sort.mb": "300"
    },
    "hadoop-env.sh": {
      // "HADOOP_HEAPSIZE": "",
      // "HADOOP_NAMENODE_OPTS": "",
      // "HADOOP_DATANODE_OPTS": "",
    },
    ...
  }
}
```

- 使用编辑Spec文件命令应用新的Hadoop配置

```
> cluster config --name myHadoop --specFile /home/serengeti/myHadoop.json
```

Stage 3: 大数据生产应用

Stage 3: Big Data Production

- ✓ Offer other big data components
- ✓ 为许多部门服务
- ✓ 经常支持一部分关键任务流程
- ✓ 支持其他大数据服务如MPP DB, NoSQL等更多非核心组件

需求:

✓ 多租户

- ✓ 在集群上我们又很多租户，需要在多租户之间保障资源隔离性，配置隔离性

✓ 可扩展性

- ✓ 系统拥有越来越多的用户和任务，我们需要保证Hadoop集群是可扩展的，可以按需调整

✓ 与大数据产品整合

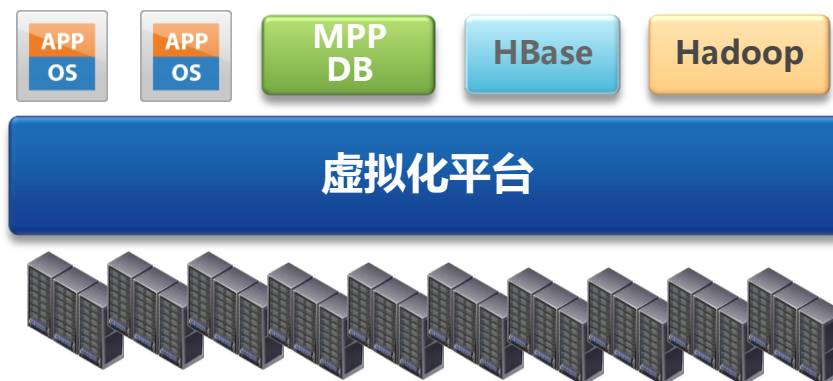
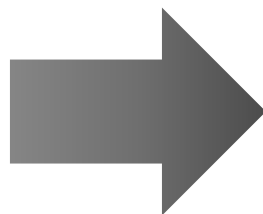
- ✓ 不仅仅是Hadoop自身，Hadoop是整个大数据分析流程的关键部分。

大数据的通用基础设施



集群扩展

使用单一集群为各种业务程序服务导致集群扩张的需求



集群整合

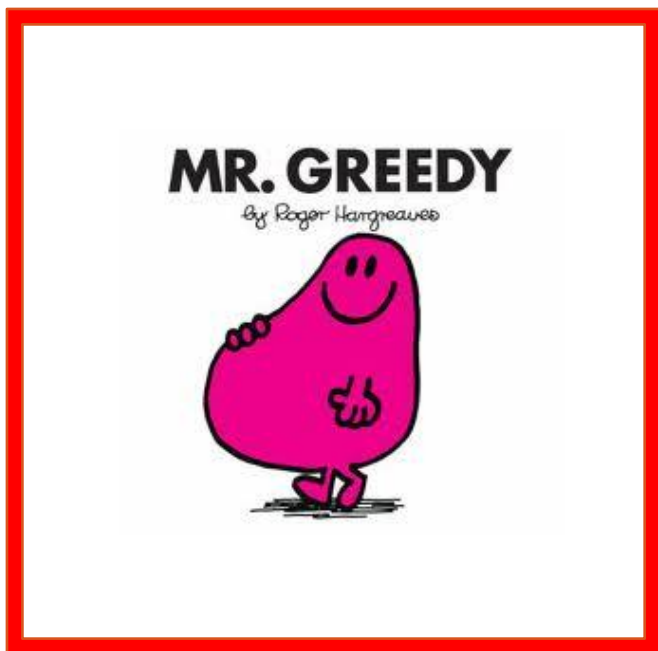
- 简化
 - 一套硬件设施
 - 统一操作
- 优化
 - 共享资源, 提高利用率
 - 弹性资源, 加快按需访问速度

容器隔离是经过验证的解决方案



分布式资源管理系统和文件系统





■ 资源隔离

- 控制高资源消耗任务
- 保证重要工作的资源

■ 版本隔离

- 允许多种OS, 应用等的不同版本共存

■ 安全隔离

- 保证不同用户和组的安全
- 数据和运行时的安全

分布式资源管理系统和文件系统

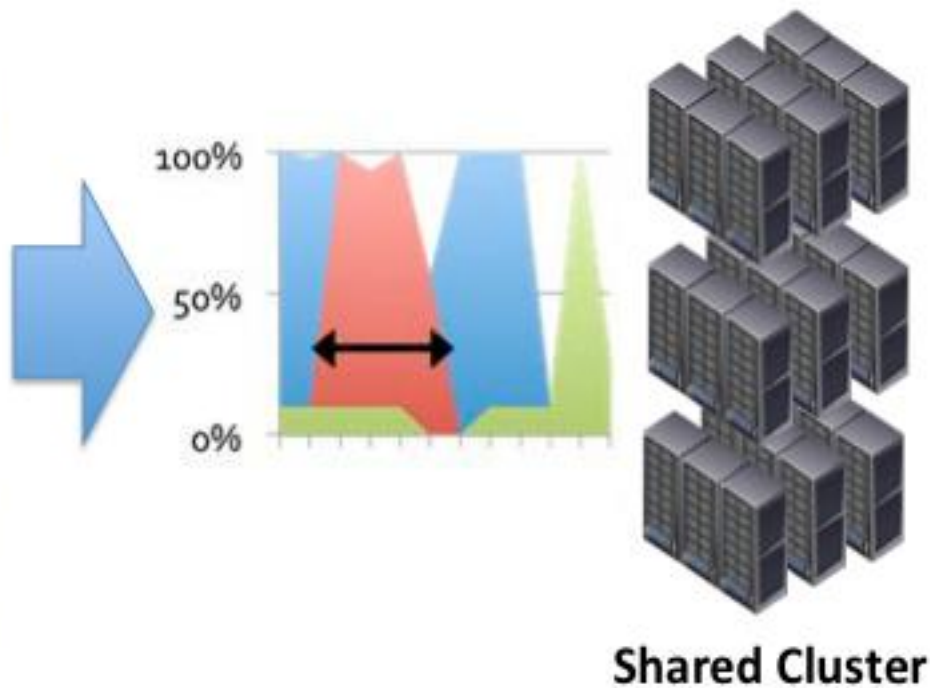


通过弹性伸缩可以共享资源

Today: App Silos

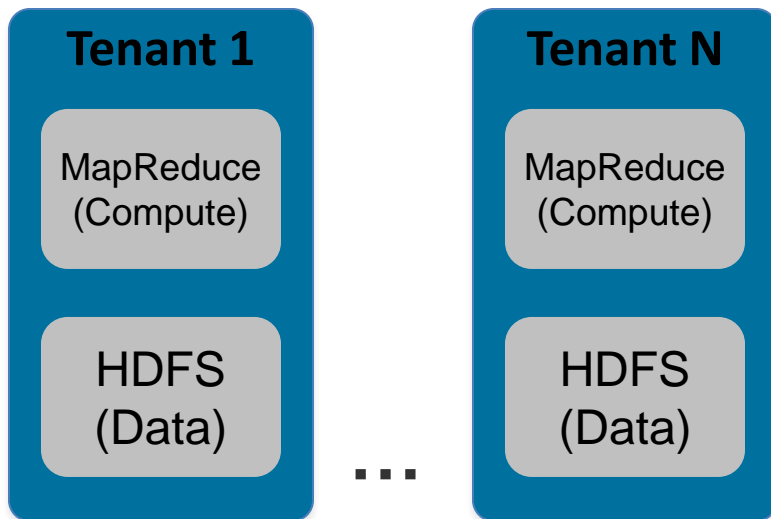


Mixed Workloads



多租户模型1: 专用的集群

Model 1: Dedicated cluster



- 优势:
 - 为每个租户指定一个专用的Hadoop集群→ 在多租户之间进行隔离
- 缺点:
 - 多租户之前没有数据共享
 - 缺乏资源共享-资源低利用率
 - 需要很多时间管理大量的Hadoop集群

通过虚拟化:

- 快速部署新集群
- 同时轻松管理大量集群
- 通过共享底层硬件集群, 更有效得在多租户之间利用资源, 提高资源利用率

多租户 模型 2:

Hadoop Cluster

Tenant 1
MR

Tenant 2
MR

...

Tenant 2
MR

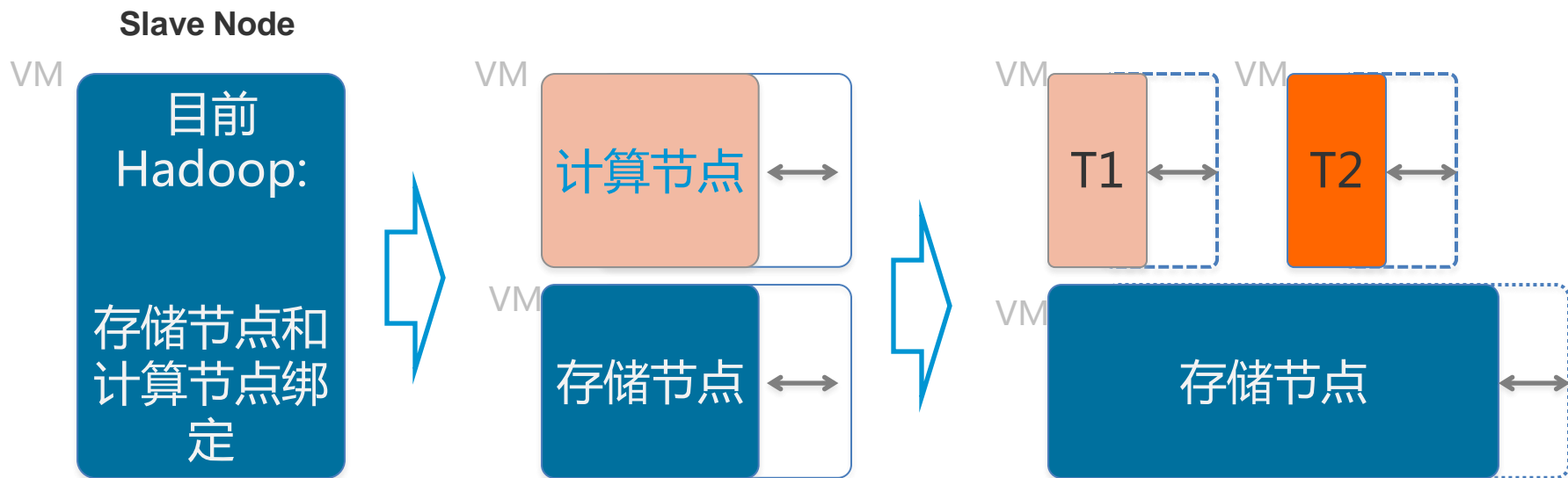
HDFS
(Data)

- 优点:
 - 各个租户之间共享相同的Hadoop集群
 - 更充分地使用资源
 - 不同租户之间能共享数据
- 缺点:
 - 使用现在的Hadoop技术，在各个租户之间隔离有限的资源，不强制资源使用

使用虚拟化:

- 虚拟机在不同的租户之间提供很好的隔离性
- 不仅仅是物理集群共享，Hadoop集群也是共享的 → 在多租户之间更有效地共享资源

vSphere之上的Hadoop弹性伸缩及多租户



1. 虚拟机中的Hadoop

- 受限使用多租户
- 资源固定，存储节点决定VM的生命周期
- 受限的扩展性

2. 分离存储节点

- 从存储节点分离出计算节点
- 计算节点弹性可扩展
- 可共享负载
- 提高易用性

3. 分离计算集群

- 多租户
- 虚拟机级别的安全性和资源隔离
- 计算节点弹性可扩展
- 同时支持多个Hadoop运行时版本

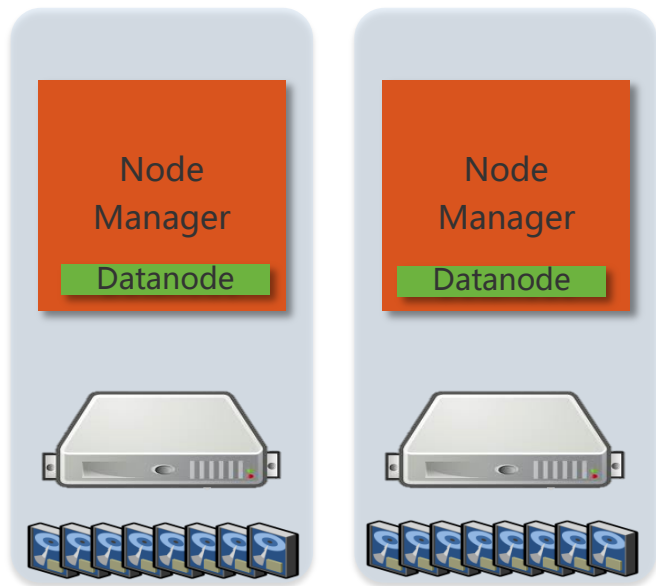
控制资源消耗，满足SLA（服务等级协议）

- `>cluster limit --name <clustername> --activeComputeNodeNum <#>`

计算和存储分离的性能分析

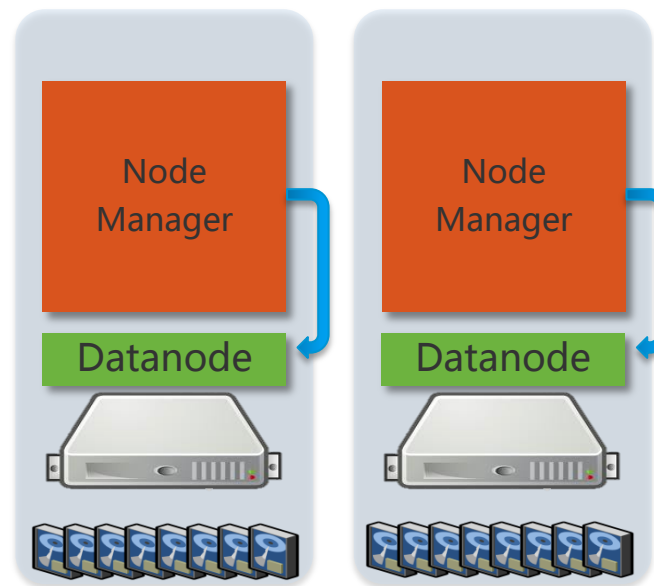
结合模式

每台硬件主机上有1个绑定计算/存储的节点



分离模式

每台硬件主机上有1个存储虚拟机, 1个计算虚拟机

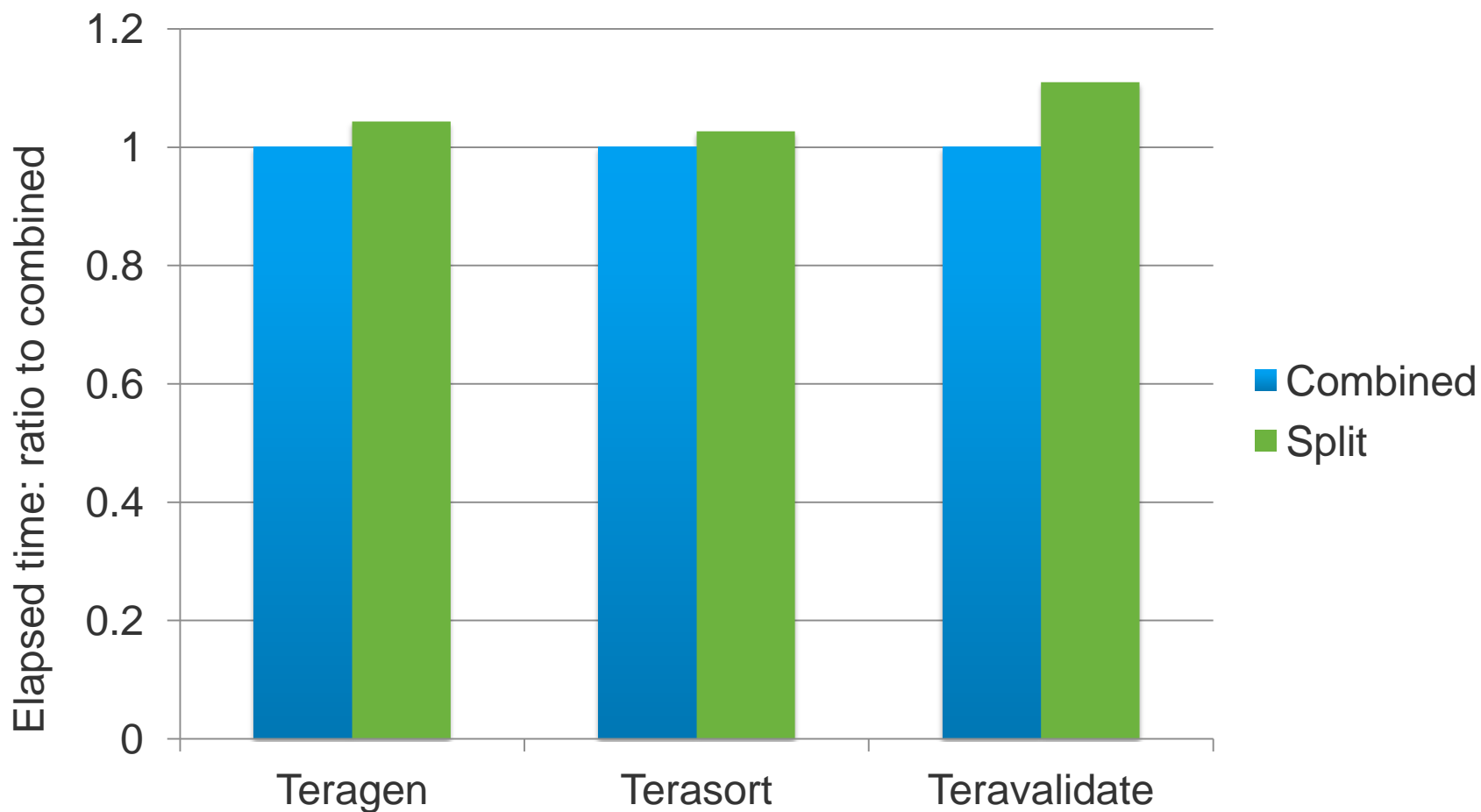


负载: Teragen, Terasort, Teravalidate

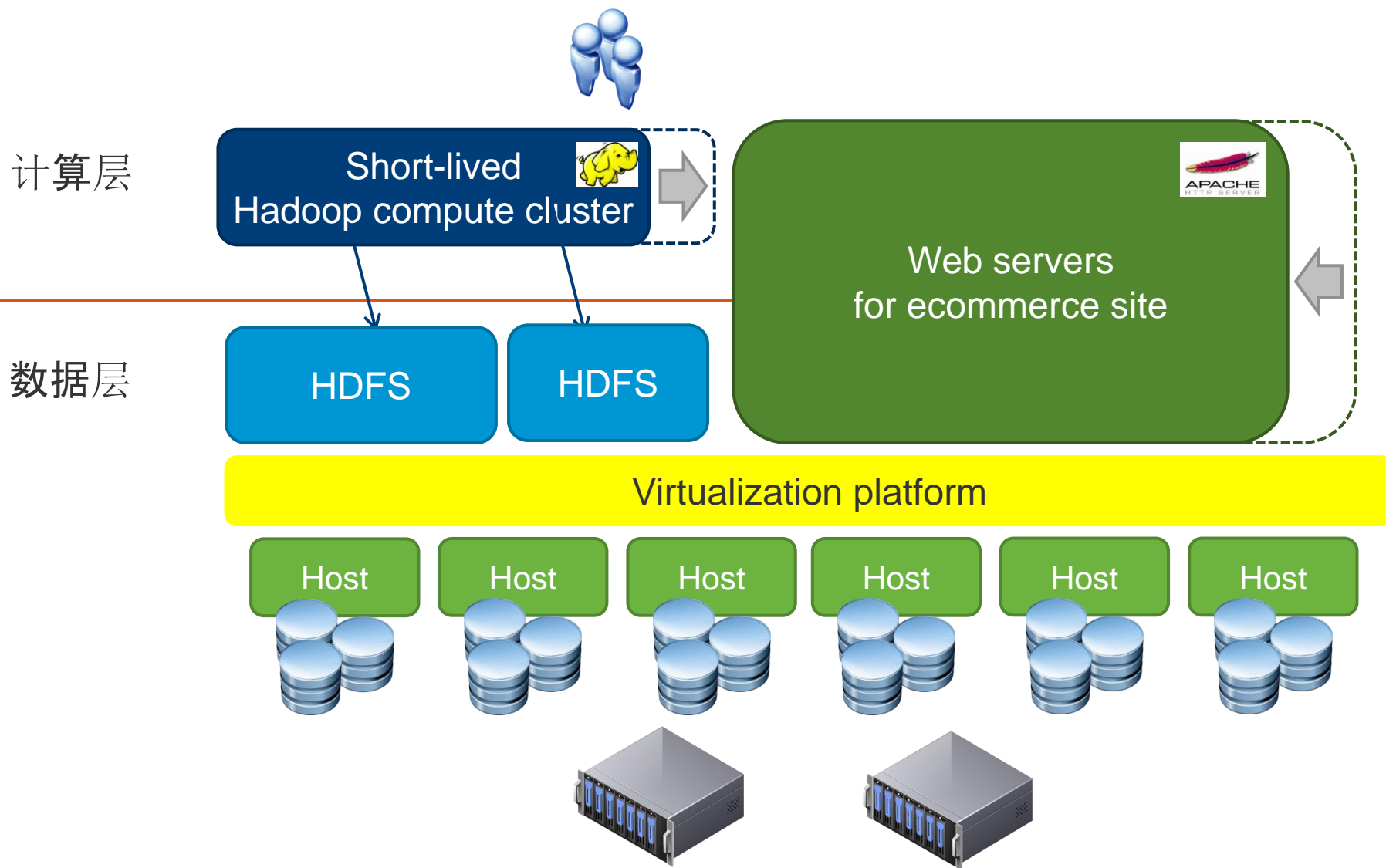
硬件配置: 8 cores, 96GB RAM, 16 disks per host x 2 nodes

计算和存储分离的性能分析

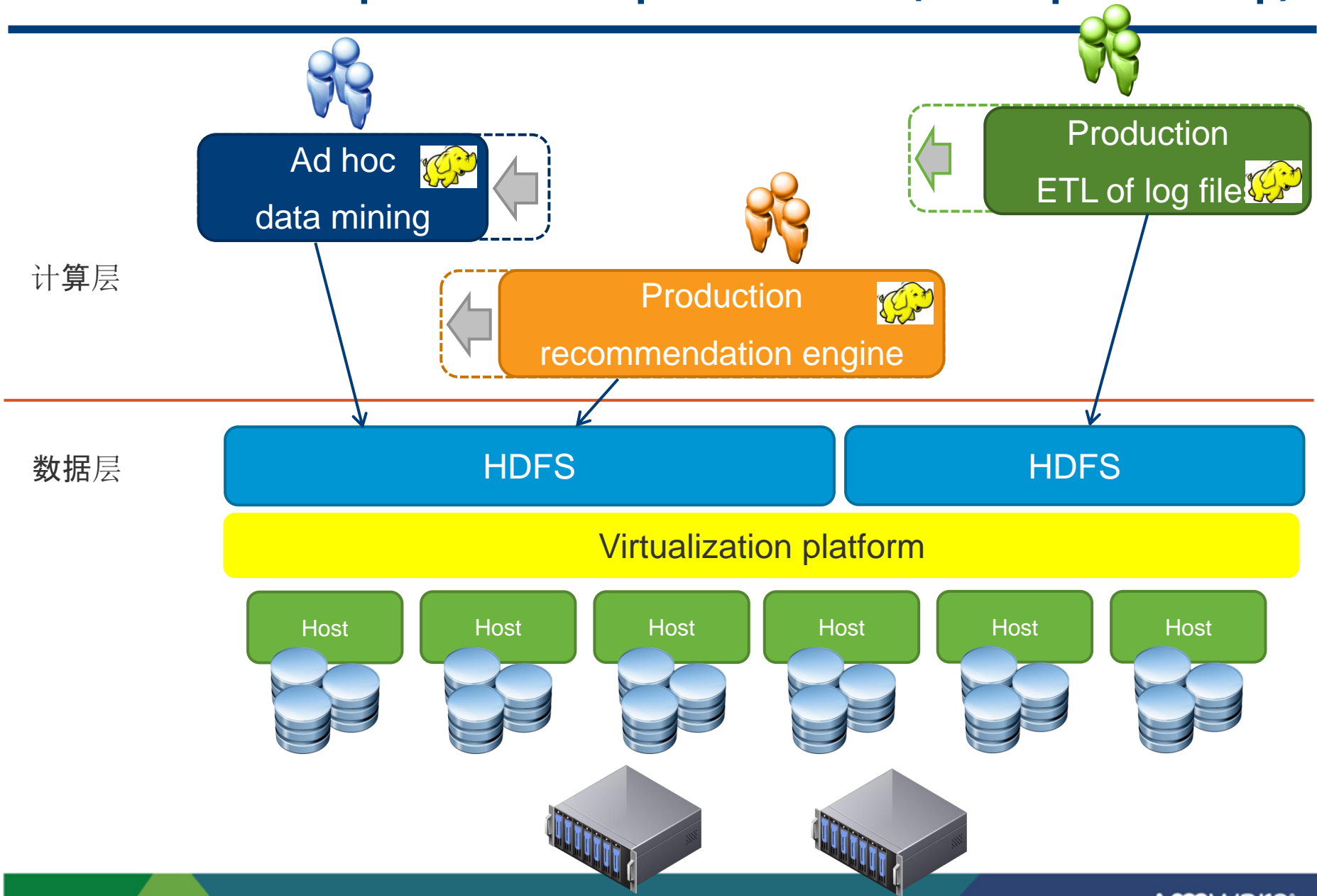
计算和存储分离的最小性能影响



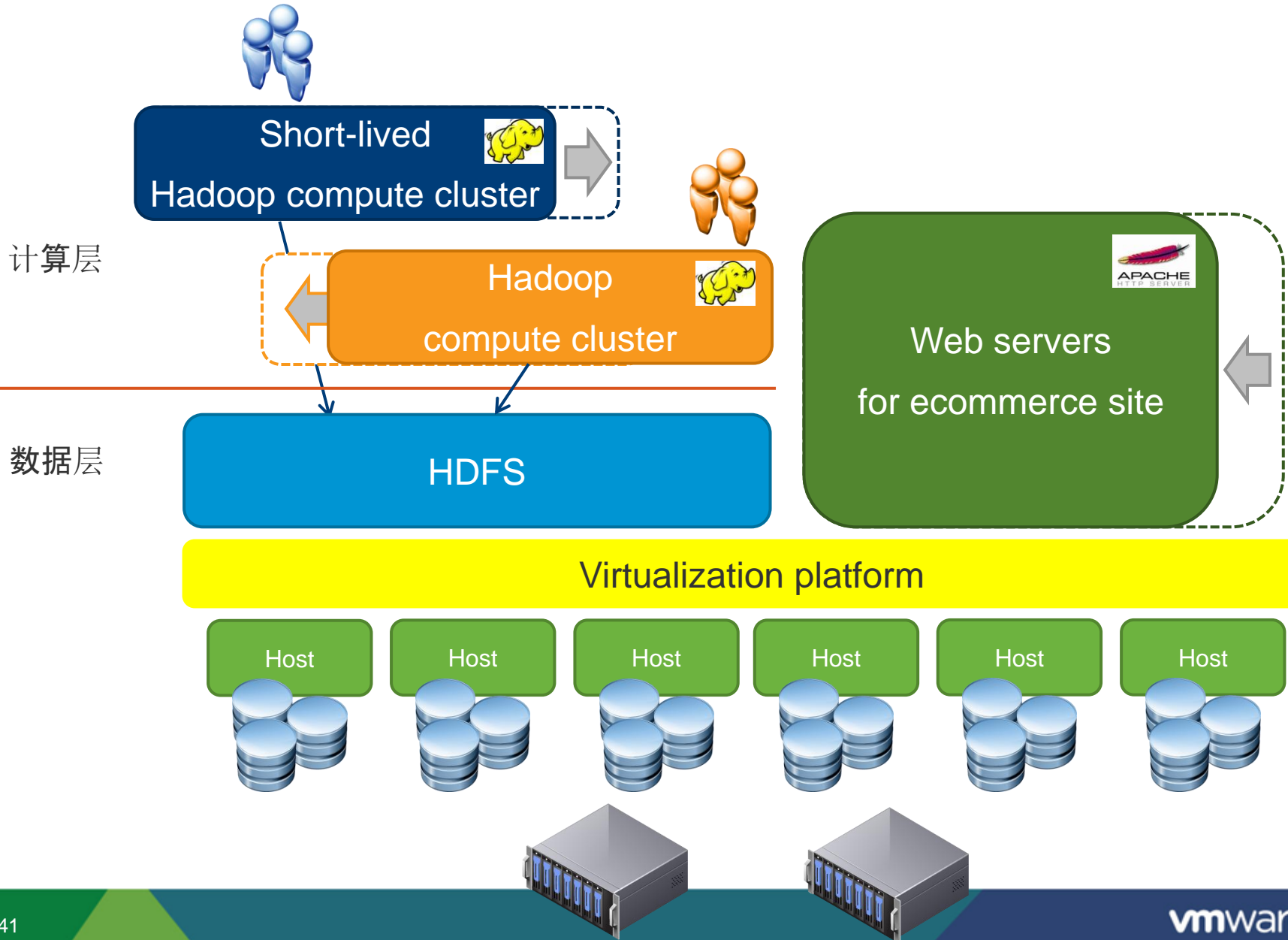
利用现有资源进行PoC (Hadoop + other workloads)



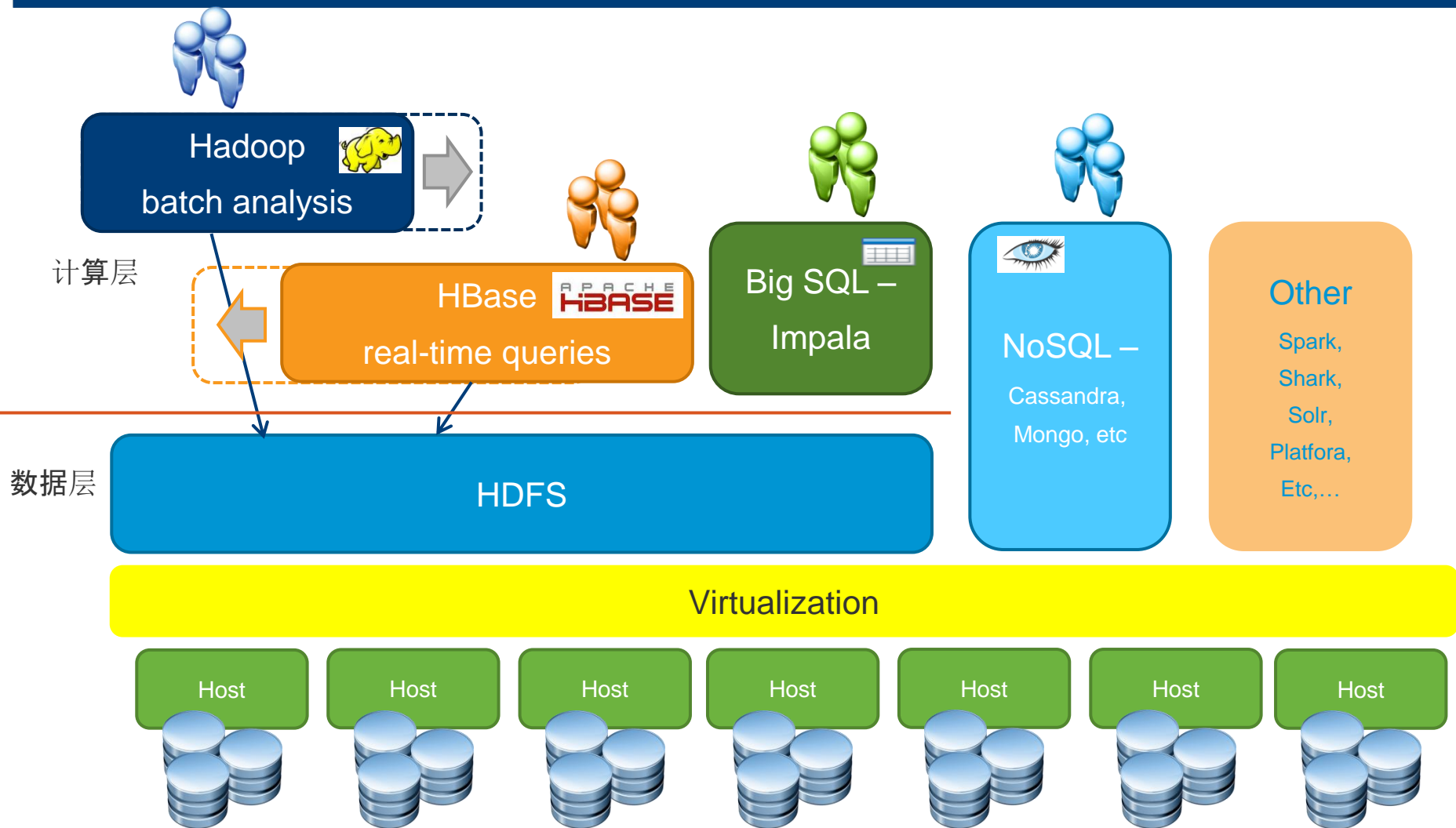
在企业内实现Hadoop即服务 “Enterprise EMR” – (Hadoop + Hadoop)



整合Hadoop和Web应用 – (Hadoop + 其它负载应用)



整合大数据系统 – (Hadoop + 其它大数据产品)



议程

- 云计算的好处
- 消除误解
- 为什么使用虚拟化
- **总结**
- Q & A

Hadoop在企业内部的使用

Integrated

Standalone

0 node

20

300

Scale

阶段一: 试点POC

- ✓ 经常从业务线开始
- ✓ 使用1-2个用例验证Hadoop价值
- ✓ 典型应用一般在20个节点以下

Stage 2: Hadoop 生产应用

- ✓ 为一些部门服务
- ✓ 更多使用用例
- ✓ 核心Hadoop和其他相关软件
- ✓ 几十个到数百个节点的典型规模

Stage 3: 大数据生产应用

- ✓ 为许多部门服务
- ✓ 经常支持一部分关键任务流程
- ✓ 与其他大数据服务整合
如MPP DB, NoSQL等

按需启用高可用，弹性，多租户的Hadoop

简单

- 快速部署
- 一站式管理使用
- 容易定制

高可用

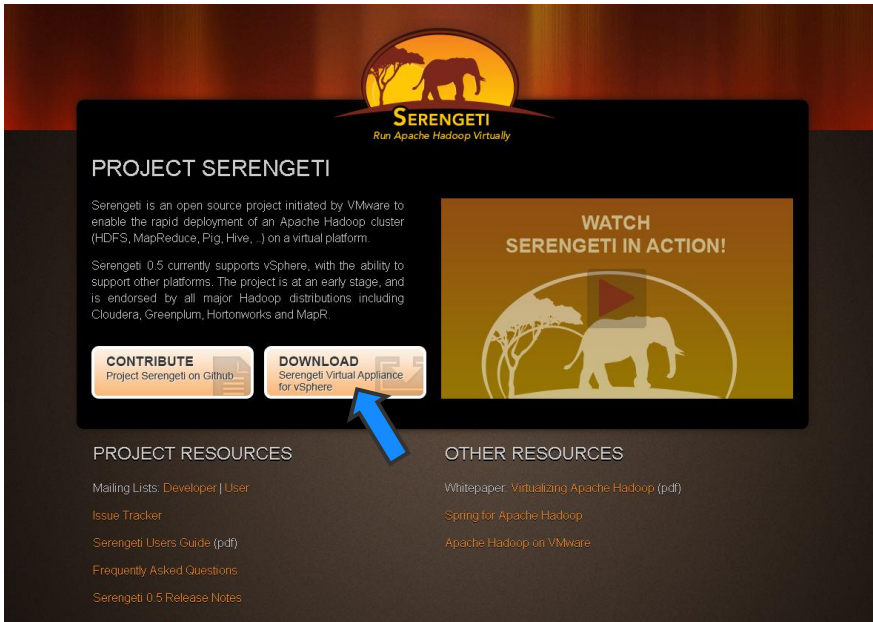
- 高可用Name Node和Job Tracker
- 设定简单
- 方案成熟

弹性伸缩 多租户

- 集群规模按需伸缩
- 计算和存储分离
- 租户完全隔离

■ 下载和试用Serengeti

- projectserengeti.org



SERENGETI
Run Apache Hadoop Virtually

PROJECT SERENGETI

Serengeti is an open source project initiated by VMware to enable the rapid deployment of an Apache Hadoop cluster (HDFS, MapReduce, Pig, Hive, ...) on a virtual platform.

Serengeti 0.5 currently supports vSphere, with the ability to support other platforms. The project is at an early stage, and is endorsed by all major Hadoop distributions including Cloudera, Greenplum, Hortonworks and MapR.

CONTRIBUTE
Project Serengeti on Github

DOWNLOAD
Serengeti Virtual Appliance for vSphere

WATCH SERENGETI IN ACTION!

PROJECT RESOURCES

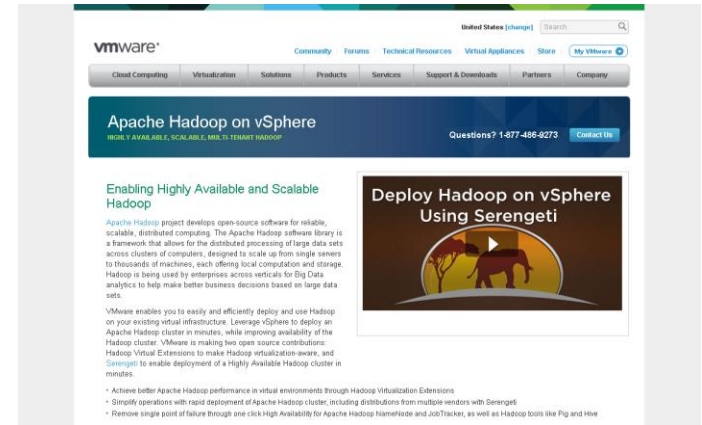
- Mailing Lists: [Developer](#) | [User](#)
- Issue Tracker
- Serengeti Users Guide (pdf)
- Frequently Asked Questions
- Serengeti 0.5 Release Notes

OTHER RESOURCES

- Whitepaper: [Virtualizing Apache Hadoop](#) (pdf)
- Spring for Apache Hadoop
- Apache Hadoop on VMware

■ VMware Hadoop 网站

- vmware.com/hadoop



vmware

United States (change) Search

Community Forums Technical Resources Virtual Appliances Store My VMware

Cloud Computing Virtualization Solutions Products Services Support & Downloads Partners Company

Apache Hadoop on vSphere

HIGHLY AVAILABLE, SCALABLE, MULTI-TENANT HADOOP

Questions? 1-877-486-9273 [Contact Us](#)

Enabling Highly Available and Scalable Hadoop

Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers, designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop is being used by enterprises across verticals for Big Data analytics to help make better business decisions based on large data sets.

VMware enables you to easily and efficiently deploy and use Hadoop on your existing virtual infrastructure. Leverage vSphere to deploy an Apache Hadoop cluster in minutes, while improving availability of the Hadoop cluster. VMware is making two open source contributions: Hadoop Virtual Extensions to make Hadoop virtualization-aware, and Serengeti to enable deployment of a Highly Available Hadoop cluster in minutes.

- Achieve better Apache Hadoop performance in virtual environments through Hadoop Virtualization Extensions
- Simplify operations with rapid deployment of Apache Hadoop cluster, including distributions from multiple vendors with Serengeti
- Remove single point of failure through one click High Availability for Apache Hadoop Heartbeats and JobTracker, as well as Hadoop tools like Pig and Hive

Deploy Hadoop on vSphere Using Serengeti

■ Hadoop 在vSphere上的性能

- vmware.com/files/pdf/VMW-Hadoop-Performance-vSphere5.pdf

■ Hadoop 高可用性解决方案

- vmware.com/files/pdf/Apache-Hadoop-VMware-HA-solution.pdf

Q & A

非常感谢