硬件建议

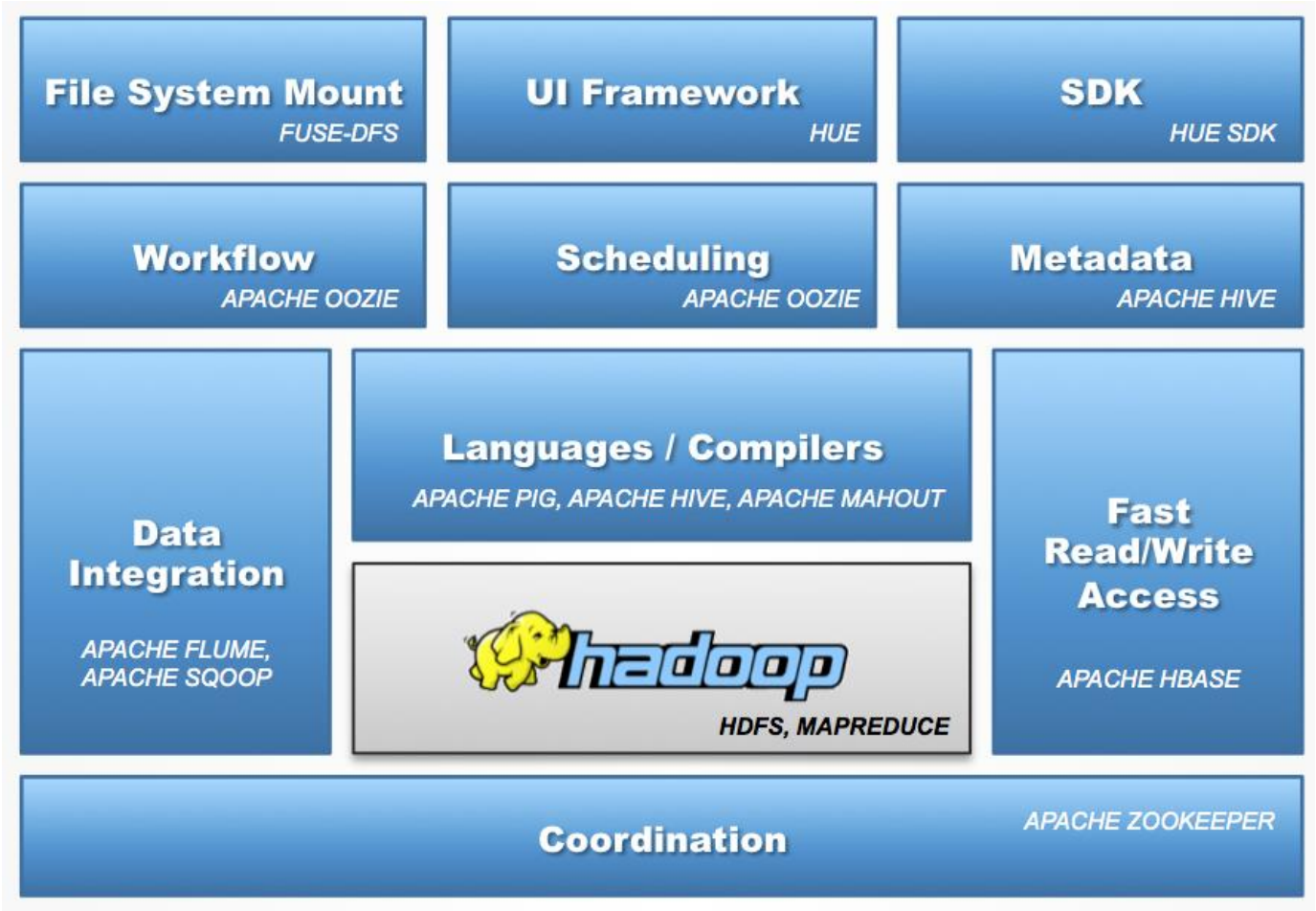
Hadoop和其他系统的不同

- 不同应用，对Hadoop的配置、规划以及硬件要求都不一样
  - 虽然可以将其分为主要的几大类应用
  - 考虑批处理系统和低延时处理系统
  - 考虑存储密集型系统和处理密集型系统
- “One size not fit all”

硬件选择的考虑因素

- 要使得Hadoop集群能够充分发挥作用，需要足够好的硬件，以及足够好的软件
- 虽然台式机硬件也能够运行Hadoop环境，但是在性能上有差距，解决问题的规模有限
- 合理选择硬件需要对自己所需要处理的问题有全面地了解，这样才能够投资合理的硬件
  - 计算密集型应用
    - 机器学习
    - 数据挖掘
  - IO密集型应用
    - 索引，检索
    - 统计，聚类
    - 数据解码与解压缩

Hadoop生态系统


File System Mount
FUSE-DFS
UI Framework
HUE
SDK
HUE SDK
Workflow
APACHE OOZIE
Scheduling
APACHE OOZIE
Metadata
APACHE HIVE
Data Integration
APACHE FLUME, APACHE SQOOP
Languages / Compilers
APACHE PIG, APACHE HIVE, APACHE MAHOUT
hadoop
HDFS, MAPREDUCE
Fast Read/Write Access
APACHE HBASE
Coordination
APACHE ZOOKEEPER

Hadoop主要核心组件

- Hadoop Distributed File System (HDFS)
  - 可靠的存储PB级别数据
  - 文件设计为批处理优化，如大量数据块(Block)的顺序读写
    - HDFS中文件按块(Block)分割存储及处理，缺省64MB
  - 可配置的每文件副本数，缺省3份
  - 支持机架(rack)感知的数据块放置策略

Hadoop主要核心组件

- MapReduce
  - 批量处理框架
  - 从HDFS读取海量数据
  - 大量上层应用框架，如Hive以及Pig
- HBase
  - 提供低延时随机读写
  - 使用HDFS作为底层可靠存储
  - 基于Hadoop核心(HDFS/MapReduce) 提供服务

HDFS

特点：

- 并行磁盘访问
- 节点磁盘容错
- 节点失效会导致数据块副本重新复制
- 流水线副本复制
- 副本数3或者10无太大性能差异

性能要求：

- **主要对于网络带宽以及存储容量要求高**

硬件偏好：

- **硬盘 > 网络 > …**

MapReduce

特点：

- 通常需要读取整个数据集
- 数据写数量因应用不同而不同
  - ETL为读写密集型应用
  - 机器学习为读密集型应用
- Shuffle过程对网络要求通常极大
  - 是Map和Reduce任务之间的M:M数据传输对应
  - 可能导致网络风暴

性能要求：

- **CPU能力直接影响并行能力(slot数目)**

硬件偏好：

- **CPU > 网络带宽 > …**
  - 内存要求视具体应用

HBase

特点：

- 高性能数据随机写
  - 通过Memstore缓存数据写入再flush，并做compaction
  - 顺序写WAL(write-ahead log)文件以避免磁盘寻址操作
- 高性能数据随机读
  - 使用BlockCache避免过多的磁盘IO操作

性能要求：

- **延时，内存大小以及Cache命中率直接影响数据读写性能**

硬件偏好：

- **内存 > 网络带宽 > …**

Hadoop哲学

- Hadoop集群构造
  - 使用大量的大众化，scale-out和share-nothing的硬件架构
  - 无特别厂家或供应商硬件特性要求
  - 本地化策略：存储及其计算部署与同一节点
    - 实际部署中由于业务隔离、资源隔离、利旧、成本以及差异化服务器硬件要求(HDFS/MapReduce等不同功能对服务器硬件要求不同)等因素，可能使用不同节点甚至部署不同的集群

大众化硬件(Commodity Hardware)

- 什么是Commodity Hardware？
  - 主流标准化商用市场产品
  - 价格20K-70K人民币
- 为何使用Commodity Hardware？
  - 成本：
    - 价格实惠
    - 单位计算价格以及存储价格低
  - 应用：
    - 实现不依赖于任何特别的硬件特性
    - 方便数据以及应用扩展和迁移
  - 运维：
    - 硬件即插即用，部署简单
    - 故障部件维护及更换方便

总体架构

- 管理节点(Head/Master Node)
  - 提供关键的、集中的、无替代的集群管理服务
  - 若该管理服务停止，则对应集群Hadoop服务停止
  - 如NameNode，JobTracker，以及HBase Master
  - **需要可靠性高的硬件设备**
- 数据节点(Data/Worker/Slave Node)
  - 处理实际任务，如数据存储，子任务执行等
  - 若该服务停止，则由其他节点自动代替服务
  - **硬件各部件皆可能损坏，但能方便的替换**
- 边缘节点(Edge Node)
  - 对外提供Hadoop服务代理以及包装
  - 作为客户端访问实际Hadoop服务
  - **需要可靠性高的硬件设备**

管理节点硬件要求

- 管理节点角色主要包括NameNode，Secondary NameNode，JobTracker等
  - Hive Meta Server以及Hive Server通常部署与其他管理节点服务器上
  - Zookeeper Server以及HMaster通常选取数据节点服务器，由于一般负载有限，对节点无太大特殊要求
  - 所有HA候选服务器(Active以及Standby)使用相同配置
- 通常对**内存**要求高但对存储要求低
- 建议使用高端PC服务器甚至小型机服务器，以提高性能和可靠性
  - 双电源、冗余风扇、网卡聚合、RAID…
  - 系统盘使用RAID1
  - 由于管理节点数目很少且重要性高，高配置一般不是问题
- 建议管理网络与数据网络隔离，使用单独网络同集群管理系统服务及终端连接

管理节点硬件标准配置建议

- 硬盘
  - >1TB的SAS硬盘作为系统盘，并使用RAID10
    - 日志在其他盘保存以提高性能
  - 至少2个SATA2硬盘作为数据盘JBOD配置
- 内存
  - 小型集群(<20节点)至少24GB DDR3内存；中型集群(<300节点)集群至少48GB;大型集群(>300节点)至少96GB
  - 若非使用单独服务器，而是重叠部署于其他服务器(如数据节点)，内存可参照一下：
    - JobTracker建议至少分配4GB内存
    - Zookeeper Server建议至少分配4GB内存
    - HMaster建议至少分配2GB内存
- CPU: 两路4核2.6Ghz处理器
- 网卡至少千兆网卡并配置双网卡聚合

NameNode服务器配置

主要需求为内存容量大小以及持久化存储可靠性

- 硬盘
  - 容量至少1TB
    - 须大于内存(FSImage)大小加上日志
  - 系统盘使用RAID10
  - 数据盘(NameDir)使用RAID或者JBOD多硬盘配置
  - 配合使用NFS作为数据存储
- 内存
  - NameNode角色内存最低配置至少48GB，一般集群建议配置96GB并
    - NameNode的FSImage完整保存在内存中
  - 内存需求决定于集群大小，集群规模变大需增加内存
    - 大约一百万block需要1GB内存
- Standby/Secondary NameNode使用Active NameNode相同配置

其他管理服务器配置

JobTracker

- 对硬件主要要求为内存大小
- 用于存储历史Job的元信息，包括task状态，计数器(counter)，进度信息等，供web UI等访问
- 内存容量需求决定于需保留的Job数及其规模，与具体应用类型以及频率关系密切，可能和集群大小无简单必然联系
  - 缺省为100个Job。若对此有特殊要求，可适当调整JobTracker内存。

Zookeeper Server

- 可以使用SSD等快速存储提高性能
  - Zookeeper数据持久化在硬盘上，空间要求小，但对时延敏感

数据节点配置策略建议

- 数量少但单点性能高的集群 vs. 数量多但单点性能低的集群
  - 一般而言，**使用更多的机器**而不是升级服务器配置
    - 采购主流的最"合算"配置的服务器，可以降低整体成本
    - 数据多分布可获得更好的scale-out并行性能以及可靠性
    - 需要考虑物理空间、网络规模以及其他配套设备等综合因素来考虑集群服务器数目

数据节点配置通常建议

- 基本配置
  - 4x1TB或者2TB硬盘
    - 为降低每单位数据的成本，在满足业务应用性能的前提下，尽量增加每台服务器的存储空间
  - 2x4核CPU
  - 24至32GB内存
  - 双万兆网卡
- **通常按1块硬盘+2个CPU核+6至8GB内存的比例配置升级硬件可以满足多数应用的需求**
  - 尤其是IO密集型应用

Hard Disk

硬 盘

硬盘配置建议

- 主流硬盘为+1TB容量的SATA 7200RPM
  - 比SAS硬盘更大更便宜
  - IOPS能力缺陷能被架构补偿，没必要使用15000RPM硬盘
  - Nearline SATA bridges the gap
- 3.5寸硬盘还是 2.5寸(SFF)硬盘
  - 一般而言皆可， 2.5寸硬盘在部署密度上有优势，而3.5寸硬盘在单盘容量、成本、以及速度上比2.5寸略有优势，在无其他考虑因素影响外，可优先选用3.5寸硬盘
  - 据观察，2.5寸硬盘的故障率比3.5寸硬盘高

容量考虑

- 通常建议使用更多数目的硬盘
  - 获得更好的并行能力
    - 不同的任务可以访问不同的磁盘
  - 8个1.5TB的硬盘性能好于6个2TB的硬盘
- 除去数据永久存储需求外，一般建议预留20%至30%的空间用于存储临时数据
  - MapReduce任务中间数据
- 实际部署中每服务器配备12个硬盘非常常见
- BP：单节点存储容量最大值不超过24TB
  - 12x2TB
  - 太大的单点容量会造成节点失效后大量的数据副本复制

JBOD vs. RAID

- 数据节点数据盘使用JBOD(Just a Bunch Of Disks)
  - 不要使用RAID
    - RAID将HDFS并行流式读写操作变成随机读写，降低了性能
    - RAID的读写性能受制于阵列中速度最慢的磁盘
    - JBOD各磁盘操作独立，因此平均性能 好于性能最差的磁盘
    - HDFS的流式并行读取使得RAID0没有意义
      - Yahoo测试表明JBOD性能比RAID0快10%到30%
    - HDFS的副本冗余存储策略降低了单机数据可靠性的重要性，使得RAID1-6没有意义
  - 在RAID无法被移除的情况下，每一个物理硬盘可以被设为一个单独的RAID 0
- 数据节点系统盘可使用RAID 10

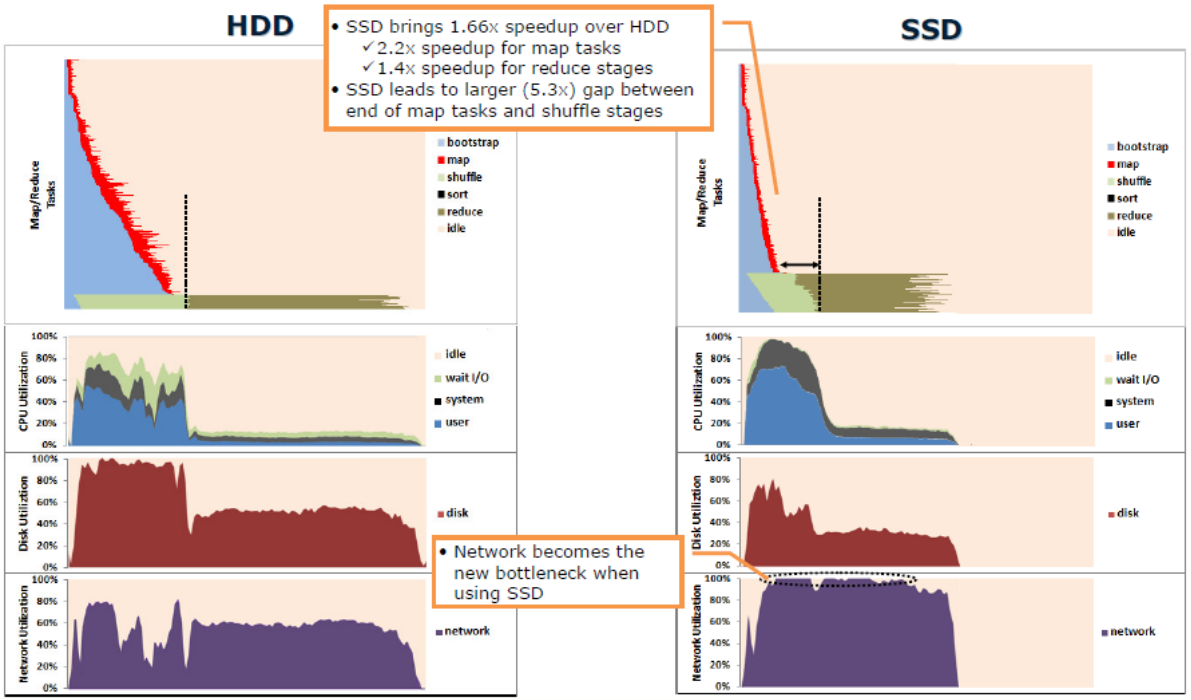
SSD

- SSD特点及优势
  - 读写性能好，尤其随机读写性能好，其中随机读性能最为突出，比传统硬盘快近10倍
  - 环保省电
- 劣势
  - 价格昂贵
  - 容量小
  - 寿命短

SSD与Hadoop

【来源】

- 由于Hadoop的磁盘IO多为顺序读写，因此不能完全发挥SSD的性能优势，大约比HDD快1倍



混合SSD与HDD架构

- 总则：使用同样投入采购更多低性能设备，通过更多的并发可以获得更高的性能
  - Hadoop的构建哲学：scale-out
- 将SSD作为HDD缓存，对随机反复读写应用有一定帮助，但总体帮助不大且性价比不高
- Hadoop未来的分级存储机制可以支持将数据写入不同类型的介质，充分利用不同类型存储的优点
  - 将WAL，MapReduce临时文件等写入高速介质以提高性能

第三方缓存产品

- 有很多硬件以及软件缓存产品，能减少对硬盘的直接读写，从而提高性能
  - 但由于缺乏对Hadoop内部机制的理解，性能提升非常有限
- 通常对随机读写占多数的RDBMS类产品性能提升明显
- 对于流式读写的Hadoop应用无意义
  - 尤其是MapReduce任务
- 对于有数据热点的系统(数据写入后会被短时间多次读取)，缓存系统仍然能有20%到30%以上的性能提升
  - 一般数据时效性决定了其冷热程度
- 应用场景非常有限，再加入成本考虑，通常不建议使用
  - 使用相同成本扩展硬件(服务器个数)很可能效果更好

共享存储系统(SAN/NAS/NFS)特点

- 存储与计算分开
- 集中式存储声称比分散本地存储有如下优点：
  - 数据自动修复(RAID)
  - 更简单的管理维护工作
  - 使用共享的预留空间，更有效的空间利用
  - 通过增加磁盘进行容量扩展
  - 更好的性能(无需Hadoop中的副本复制)

共享存储系统和Hadoop

- 共享存储系统在Hadoop的缺点
  - 内部一般使用RAID技术，参见前页与JBOD比较
  - 通常使用oversubscribe策略，IO带宽比使用本地存储并发小，不合适MapReduce流量风暴(输入，输出，shuffle)
  - 存储系统成为单点，性能以及可靠性瓶颈
- 不建议使用共享存储系统作为Hadoop数据存储
  - 共享存储系统可以作为管理节点元数据存储介质，提供高可靠性保障

CPU

处理器

CPU

- 4(Quad)核CPU现在已经是服务器标配
  - 6核和8核CPU也正逐渐慢慢成为数据中心主流配置
  - Hadoop应用很少是CPU密集型的，通常是磁盘或者网络IO密集型的
  - 因此，通常没必要使用顶级CPU
- 使用超线程(Hyper-Threading) 以及快速通道互联(QPI)技术
  - Intel技术
  - 测试表明这些特性都对Hadoop有利
- 两路CPU主板现在也是服务器标配
- “虚拟“总CPU核数至少16个
  - CPU核数直接影响系统并行处理能力，建议MapReduce的slot数与CPU核数相当
  - 在系统配置中确认CPU配置(个数、频率等)
  - 能支持一般Hadoop应用，但仍视具体应用相关，尤其是MapReduce任务并行计算需求

Memory

内存

内存配置

- 最低内存要求为24GB内存
- 典型的数据节点内存配置为48GB至96GB
- 物理内存配置大小满足应用最大需求
  - 确保RAM大小满足所有任务，以及DataNode以及TaskTracker守护进程和操作系统的总要求，**不应该使用到虚拟内存**
  - 用户可以通过简单叠加相应服务需要的内存要求来计算推荐的内存(其他内存需求见后)
    - DataNode 2GB，TaskTracker 2GB，操作系统 4GB
- 提示：
  - 分配内存时要利用到硬件偏向，如三通道或四通道配置

数据节点服务内存需求

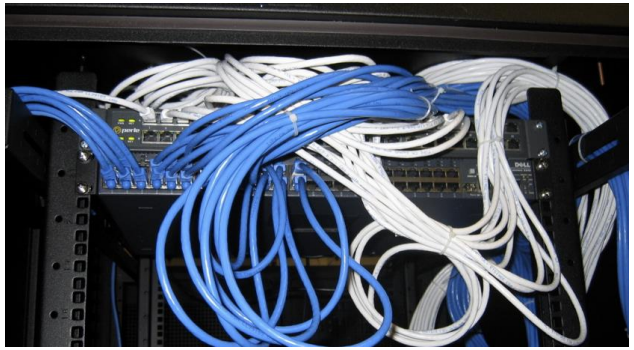
- MapReduce任务
  - 内存需求计算：任务数*单任务内存需求
  - 典型的Map和Reduce任务需要1GB到4GB内存
  - 通常并发的任务（slot）数配置为CPU核数的1.5倍
    - 具体slot配置视应用类型而定
  - 通常需要8GB至16GB
- HBase RegionServer
  - 内存需求计算：Region数目*Memstore大小*BlockCache大小
  - Region数目一般每台服务器100至150个
  - Memstore影响数据写入性能，缺省64MB
  - BlockCache影响数据随机读性能
  - 至少需要16GB内存，越多越好

NIC

网卡

网络的重要性

- 网络可能是集群里最昂贵的部分
- 网络部署完成后的改造或替换代价可能很高，部署需要一定的前瞻性
- 网络因素可能引起的问题不仅类型繁多，影响系统的几乎所有组件及服务，而且往往不易被定位或解决
  - 性能问题
  - 原因“不明”的任务执行失败
  - 数据冲突
  - …



网络配置

- 网络使用以太网络
- 最低要求使用千兆网络连接
- 中型及以上集群，或者大IO任务(如ETL)需要至少万兆网络
  - 单节点+12个100MB/s的硬盘至少需要万兆带宽
- 建议配置双网卡链路聚合，并把工作模式设为6(通过循环平衡负载，且网络适配器可以在没有配置交换器的情况下正常工作)
  - 提高网卡可靠性
  - 增加一倍网络带宽
  - 更多的网卡聚合会增加网络端口成本，宁可直接升级网速
- 即使带宽足够，使用更高带宽的网络(如Infiniband)也可以提升性能，尤其是低延时组件，如HBase
  - 相同配置下，使用40Gb IB网络的HBase读写性能比10Gb网络快10%至20%

交换机

- 为了能够使得Hadoop的处理能力能够得到充分的释放，交换机对于系统运行的性能起到了决定性的左右，建议在可能的情况下尽量选择高端的交换机
  - 高端交换机的每一个接口都能够达到线速(line rate)（网线能够达到什么速度，交换机就能够提供什么速度，没有性能损失）
- 千兆以太网接口是最基本的要求，更重要的是交换机的背板带宽，是决定数据传输的关键因素
- 小型集群可使用单层交换机架构，必要时增加背板扩展
- 中型或大型集群建议使用双层交换机架构，TOR(Top-Of-Rack)和aggregation交换机
  - TOR交换机采用万兆或4万兆交换机，aggregation交换机采用4万兆交换机
- 如有可能，交换机采用双机冗余

其他服务器技术

- 随着硬件的发展，很多新兴数据中心技术层出不穷
  - 更高效的冷却技术
  - …
- 同时由于Hadoop的流行，很多厂商也推出了针对Hadoop更加友好的硬件，比如：
  - 密度更高服务器
  - 能配备更多硬盘(48块)的服务器
  - …
- 这些对Hadoop都很有利，但并不是所有的技术都适合Hadoop...

刀片机(Blade)

- 优点
  - 节省服务器安装空间
  - 节省电力等资源消耗
  - 适用于计算密集型的应用场景，如HPC
- 缺点
  - 可能的部署需要附加存储设备(参见JBOD)：
    - 附加存储刀片机到计算刀片机
    - 使用共享存储(NAS/SAN)
  - 共享电源、背板(backplane)等使得刀片机共享IO以及网络连接等功能，不仅有SPOF问题，而且通常oversubscribed
  - 没有本地存储或很少
  - 成本高
- 因此，**通常不建议使用刀片机搭建Hadoop**

一体机(Highly Integrated Rack)

- 经过厂家验证的推荐配置以及部署
  - 硬件：服务器、网络、存储配置以及位置部署
  - 软件：操作系统、JDK以及Hadoop验证版本
  - Hadoop推荐参数设置
- 一般有定制的部署以及服务器管理软件或服务支持
- 售后服务一体化
- 现有一体机方案一般仅限于配置参照，尚无特殊的定制功能
  - 即与单独购买相同配置的产品并自主部署相比，无任何差别
  - 尚无针对某一体机硬件配置的Hadoop功能
  - 尚无针对Hadoop的特有一体机
    - 有针对Hadoop进行有调教的服务器产品以及针对底层硬件的Hadoo优化，但都非一体机级别

虚拟化技术(Virtualization)

- 虚拟化的好处
  - 解决Hadoop关键服务的HA，如NameNode
    - 使用现有虚拟化技术，使用硬件或软件部署冗余且透明的NameNode
    - Hadoop 2.0已经有HA解决方案
  - 更好的利用物理资源
    - 现有Hadoop的资源管理框架不够完善，可以使用虚拟机实现更好的MT(Multi-Tenant)资源共享
  - 对于计算密集型的应用可能可以接受
    - 虚拟化+共享存储

虚拟化技术与Hadoop

- 虚拟化技术构建Hadoop的弱点
  - 对于物理层的抽象将硬盘并行流式读写变成随机读写，即使现在虚拟化厂商采用了很多技术改进，对于IO密集型的应用至少有10%至20%以上的性能下降
  - 虚拟化对于数据节点没有意义
    - 比起使用一个可以支持多个虚拟机的高配置节点，使用更多更便宜的节点更好
  - 虚拟机的属主机器自身的SPOF问题
  - 再加上虚拟机相关软件，整体成本高昂
- 因此，不建议使用虚拟机搭建Hadoop服务

云服务

- 将数据以及计算放到云上，在公有云计算服务至少搭建Hadoop集群
  - 比如，亚马逊EC2和亚马逊EMR
- 对于PoC(Proof of Concept)项目或原型类方案有用，适合起步阶段Hadoop应用
  - 集群以及数据规模增大后，使用云服务成本将大于私有Hadoop部署成本，届时可将应用以及数据迁回
- 对于Hadoop核心服务(HDFS/MapReduce)可以接受，但很难满足HBase性能要求
  - 请仔细查看云服务商SLA以及可靠性保证

建议-管理节点

- 小型集群使用一个管理节点
  - 可增加一个HA节点
- 中型集群使用3至5个管理节点
  - 使用HA节点
- 如果可能，为管理节点选择高配置服务器
  - 双电源、冗余风扇、大内存、网卡聚合、RAID…
- 如果条件允许，根据管理节点角色分别配置
  - 内存
  - 系统盘RAID、NFS
- 一般需要额外一台服务器作为集群管理服务器

建议-网络

- Start with workgroup-class, managed switches and bonded 1GbE NICs in a **traditional tree** topology (ToR + distribution switches)
- 使用高端非阻塞线速交换机Transition to higher end, non-blocking, line rate switches
- Then move to different network topologies, such as **spine fabric**
- Along the way ensure that aggregation switches can cope at any time – or live with the net effect

建议

- 低端配置
  - 入门级集群，解决较小规模问题
  - 4至10个节点

处理器CPU	双路四核2.6GHz服务器处理器
内存	32G或者以上内存，DDR3，ECC
磁盘接口	SAS 6GB/s
磁盘	6x或者12x SATA 1T 7200RPM监控级硬盘
网络	两个千兆以太网口

建议

- 主流配置
  - 承载多种应用，海量存储，解决中等规模问题，实际上能够满足大多数中小企业的需求
  - 10个以上节点

处理器CPU	双路六核2.9GHz服务器处理器，处理器缓存15MB
内存	64GB及以上内存，DDR3-1600，ECC
磁盘接口	SAS 6GB/s
磁盘	6x或者12x SATA 1T或者3T 7200RPM监控级硬盘（依据数据规模而定）
网络	两个千兆以太网口

From: Hadoop Operations by Eric Sammer

建议

- 高端配置
  - 大内存，高速网络
  - 可以考虑使用Infiniband网络

处理器CPU	双路六核2.9GHz服务器处理器，处理器缓存15MB，依据应用可以选用更高端的处理器
内存	96G或者以上内存，DDR3-1600，ECC
磁盘接口	2xSAS 6GB/s
磁盘	24x 1TB SFF高速SAS 7200RPM硬盘
网络	10Gb以太网口

From: Hadoop Operations by Eric Sammer

资源

- Dell | Cloudera Solution Reference Architecture
- Hadoop Operations, Eric Sammer