4 Datacenter as A Computer

数据中心

计算和存储正在从 PC 客户端转移到更小也更便携的设备，这些设备都配有强大的互联网服务（large Internet services）。

服务端计算的趋势源于两点，一是用户体验的需求（the need for user experience improvements），二是其给供应商带来的优势（the advantage it offers to vendors）。

对于供应商而言：

1）只需要考虑对自己的数据中心和硬件设备进行维护和更新，而不是数以百万计的客户端

2）数据中心使得应用服务在平均每个用户的成本上更低

3）数据中心的服务和存储要比桌面级的更加容易管理，因为他们是在一个统一的控制下

服务端计算和互联网服务的开发促使了新的一种计算系统的出现，即数据仓库级的计算机（warehouse-scale computers），称为 WSC。

WSC 中，程序会是一种互联网服务，其由数十上百个程序互相配合来提供（interact to implement）复杂的客户服务，比如邮件、搜索、地图。

硬件平台主要构成有：上千个独立的计算节点、网络和存储子系统、能源设备、调节设备、强大的冷却散热系统。

与传统数据中心的不同

1. WSCs 是一个统一的组织，使用相同的软硬件系统，共享同一个管理层，被视为一个单一的计算单元

2. 多数的应用、中间件和系统软件都是自建的（built in-house）

3. WSCs 只运行着几个应用或服务，但这些应用都非常庞大

技术挑战

1. 难以实验或仿真，因此需要发展新的技术来指导 WSC 的设计

2. 系统错误和能耗对 WSC 的设计有着巨大的影响

3. 相比多个独立的服务器，WSC 要多一层的复杂度

WSC 的架构概述

1. 低端服务器（low-end servers）之间通过局部以太网接口形成局域网，若干个低端服务器可放置在一个服务器机架上

2. 服务器机架（server rack）之间通过机架级交换机（rack-level switch）3. 集群（cluster），集群级交换机（cluster-level Ethernet switch）和若干个服务器机架相连，构成集群系统

【存储系统】

两种方式，一种是硬盘或闪存设备直接与每个独立的服务器相连，有一个全局分布式文件系统管理，如 GFS；一种是作为网络可获取的存储设备（Network Attached Storage，NAS），直接与集群级交换机互联。

前者倾向于针对每个服务器提供服务，相对更低的成本，更高的网络带宽，可深入挖掘数据的局部性，可靠性体现在及时服务器或整个机架宕机也仍然可以提供服务

后者倾向于针对每个应用提供服务，更简单的实现，提供更高的可用性和纠错能力，可以把存储和计算独立分开管理

【网络互联】

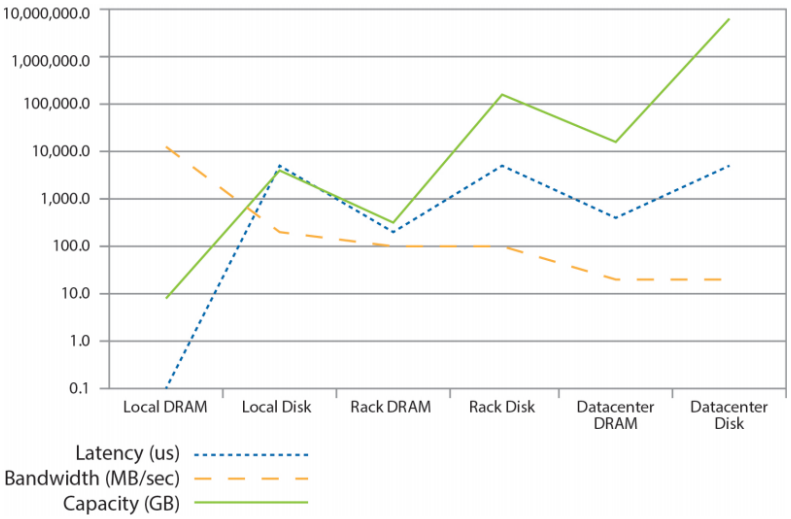
速度、规模和成本的权衡（trade-off between speed，scale and cost）

交换机（switch）提高 10 倍带宽通常会带来 100 倍甚至更多的成本提升

编程人员必须明白集群级交换机有限的带宽资源，因此需要发掘机架内的网络局部性（networking locality）、软件开发复杂度（complicating software development）和对资源利用的影响（impacting resource utilization）

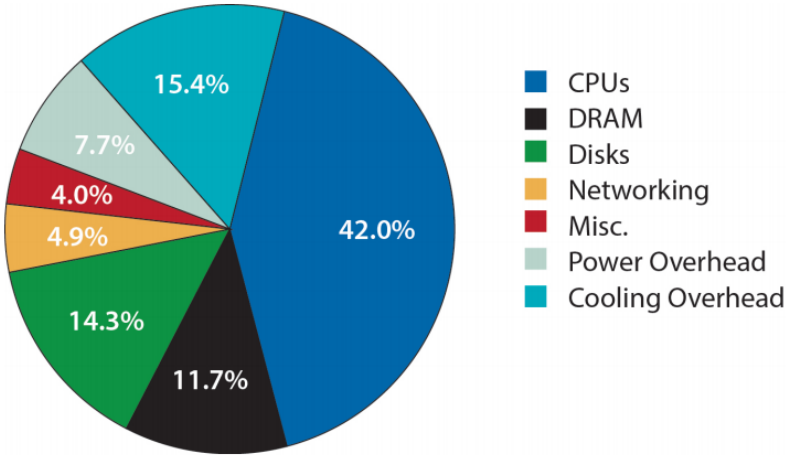
【延迟、带宽和容量的量化分析】

应用庞大到不足以在一个单一的机架内运行的时候，必须有效地处理这些延迟、带宽和容量的不平衡性



【能耗管理】

CPU 仍然是能耗的主要来源



硬件设施组成

主要包括：服务器硬件（server hardware）、网络互联（networking fabric）和存储层次结构（storage hierarchy somponents）

成本效益的服务器硬件

cost-efficient server hardware

WSCs 和 SMP 也有不同，其比 SMP 要多一层远程的访问，一个简单的模型如下，其中 f 是一个工作单元（1 ms）中远程访问的次数，结果如图 1 所示

$$Executiontime = 1ms + f \times [\frac{100ns}{nodes} + 100us \times (1 - \frac{1}{nodes})]$$

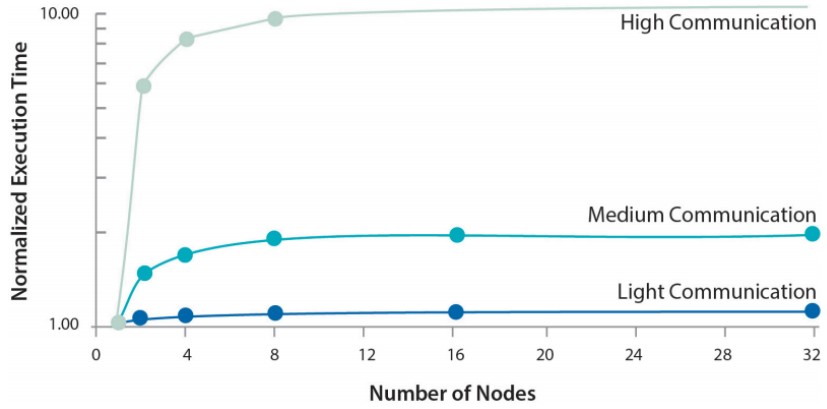


图 1. WSCs 的性能模型

1. 对于较低的通讯（light communication），WSCs 带来的性能下降很小

2. 对于较高或很大的通讯（medium- and high-communication），其带来的性能下降很严重，而且随着节点个数的增多变得更为剧烈

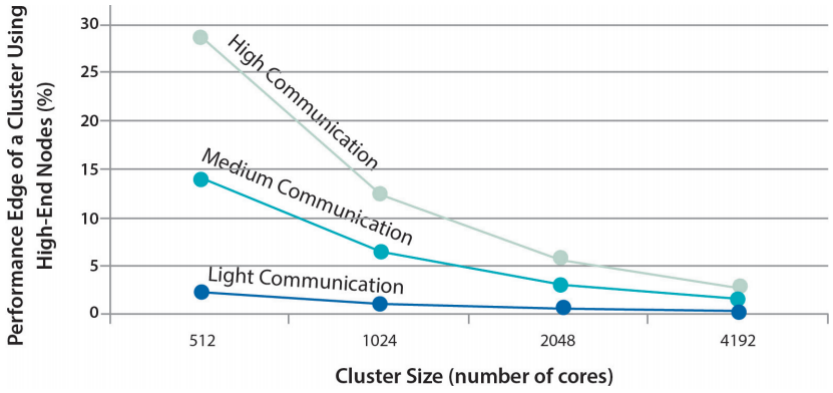


图 2. WSCs 相对于 SMP 的性能优势

随着集群规模的增大，其相对于 SMP 的优势也急剧下滑

即使一个应用需要两千以上个节点，而且是在很高的通讯需求下，WSCs 仍然要比 SMP 有着大约 5% 的性能优势，而 SMP 的高端节点带来的较高的成本，因此 WSCs 仍是一个更好的选择

【平衡的设计】

1）优秀的编程人员应当重构算法以适应廉价的设计选择

2）最具成本效益和平衡的设计，应当是针对多个具体的工作负载，满足其资源需求，而不必是完美的应多所有的工作负载

3）高效的利用可代替的资源

WSC 存储系统

分布式文件系统，如 GFS，专注于提供对上千个读写的高吞吐量，和高硬件错误率下的鲁棒的性能。

对于需要结构化数据的应用，WSC 也应当提供服务。Google 的 BigTable 和 Amazon 的 Dynamo，容许临时的数据不一致，将其处理交给软件。第二代如 MegaStore 和 Spanner 则牺牲了一些性能和便利来提供一个简单的编程接口。

WSC 的分布式存储系统也对网络产生了影响，由于网络和硬盘之间性能的巨大差距，Ananthanarayanan 指出数据中心内部（intra-datacenter）的局部性已经不再关系了

WSC 网络互联

网络互联并不存在简单直接的扩展方案（no straightforward horizontal scaling solution），如图 3 所示。对于 k 端口的交换机，使用 $\frac{5\times k^2}{4}$ 个交换机可以支持 $\frac{k^3}{4}$ 个服务器，来提供完全的吞吐量（full throughput），但也带来了巨大的成本

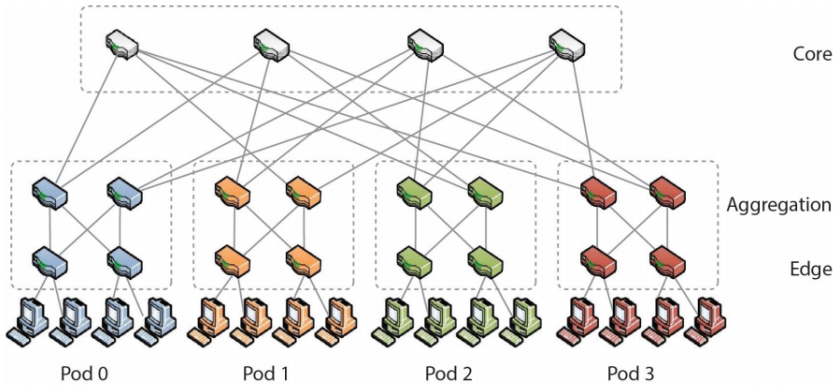


图 3. 3 层的树状拓扑结构

WSC 通过使用更多的顶层交换机（oversubscribing the network at the top-of-rack swith）来降低互联拓扑的规模，尽管其无法提供完全的吞吐量。另一个方法就是对于某些流量使用专门的网络，比如一个单独的网络来连接服务器和存储节点