

云设计简介

IaaS 的四个设计原则

什么是云

什么是云.....	1
为什么选择 Mellanox 构建云	2
构建 IaaS 云的设计注意事项	2
总结.....	4

云计算是用于抽象化计算机硬件预配置和管理的技术和实践的集合。其目标是简化用户体验，使他们可以享受按需获取计算资源的好处；或以云计算语言表述为“即服务”。

构成单个云的资源通常使用符合三个抽象级别之一的接口提供给最终用户。从最细粒度到最抽象的接口分别为：基础设施即服务 (IaaS)、平台即服务 (PaaS) 和软件即服务 (SaaS)。

通过 Web 浏览器或瘦客户端提供给最终用户的应用程序被称为软件即服务 (SaaS)，几乎完全在云中进行存储、管理和更新。使用 SaaS 模式部署应用程序是社交网络、协作、媒体和内容管理类应用程序的典型做法。然而，我们看到更多传统桌面应用程序正在转移到这种模式。这类例子很广泛，涉及 CRM（如 salesforce.com）、电子邮件（如 Gmail）和网页游戏（如 FarmVille）。

平台即服务 (PaaS) 是 SaaS 之下的抽象级别。平台是定制的环境，旨在加速应用程序开发。平台包含用于托管应用程序某些方面的特定规格的计算能力，以及一组软件，如 Web 服务器、数据库、负载均衡等。这些预构建的平台组件通过将应用程序开发人员的职责转移给云维护者，简化了这些资源的部署和维护。PaaS 提供常见平台元素的一站式部署，如 Apache Web 服务器、SQL 数据库、负载均衡等。

基础架构即服务 (IaaS) 是对虚拟硬件资源的创建，包括虚拟机、虚拟网络和虚拟化存储。IaaS 与虚拟化概念耦合得非常紧密，并且通常构成 PaaS 和 SaaS 更高级抽象的基础。

虽然存储是 IaaS 的逻辑组成部分，但通常单独考虑它。随着云存储的使用越来越独立于 IaaS 的其他部分，这种情况尤为明显。例如，Amazon S3 和 Rackspace Cloud Files，以及诸如 Dropbox 和 box.com 之类的 SaaS 应用。

本文概述了 IaaS 云部署的四个主要设计注意事项，并探讨了 Mellanox 面向 IaaS 的以太网和 InfiniBand 互连的优势。

为什么选择 Mellanox 构建云

Mellanox Technologies 是 RDMA 技术的全球领导者。RDMA（远程动态内存访问）是一种网络适配器功能，允许在通过网络连接的商用 x86 系统之间进行内存到内存传输。Mellanox RDMA 允许在虚拟机之间移动网络和存储数据，在市场上的所有互连技术中具有最高的带宽、最低的延迟和消耗最少的 CPU 周期。Mellanox 互连技术的主要受益者是需要构建可扩展和低成本 IaaS 云的云提供商。本文仅重点介绍对 IaaS 的好处，但由于 SaaS 和 PaaS 是建立在 IaaS 基础之上的层，因此也能受益。

构建 IaaS 云的设计注意事项

通过大量客户互动、构建数据中心以及与 IaaS 架构师和管理者紧密合作，Mellanox 观察到世界上所有最出色云数据中心所共有的一些事实。这些相似之处被划分为以下四项设计原则：

原则 1 — 可扩展的物理设计

最出色的 IaaS 云能够以快速而简单的方式部署物理机，然后开始将其用作虚拟基础架构的一部分。大多数公共云和私有云将从小规模开始，随着更多用户开始将其应用程序迁入云端，需要增加额外的容量。

主要目标是能够使用高密度的物理外形规格。第二目标是采用高度模块化的设计。这两个目标彼此关联，因为它们允许以精细化的方式将新资源添加到云中，从而使预配置变得简单。密集的外形规格通常在散热和功耗方面更有效率，从而获得更佳的 TCO。通常，高密度节点的采购价格与等效的多机架设备相当或更便宜。妨碍 IaaS 架构师部署高密度解决方案的限制通常有两个：这些模型的存储容量和 IO 可扩展性。

在将新的物理系统部署到现有云的过程中，要防止在现有系统中引入停机。导致这种情况的原因是新资源的物理安装的复杂性。在设计云集群时，一个重要的注意事项是在安装物理资源后避免因重新平衡导致性能下降。

从网络角度来看，最好选择允许使用密集外形规格并最大限度地降低物理安装难度的技术和拓扑。使用高带宽互连解决方案可实现通过最少的物理连接最大程度地增加虚拟基础架构设备数量的设计。此外，它通过单一线路提供了不同流量类型（即管理、存储和网络）的融合，从而简化了物理安装。

假设某个给定的虚拟基础架构设备需要大约 1Gb/s 的连接，包括网络、存储和管理流量，这可能是保守估计。如果 IaaS 旨在将存储资源（原则 3）从虚拟机主机剥离，则此数字尤为保守。基于这些假设，大约需要 n gbps 的带宽，其中 n 表示在任何给定的虚拟机主机上运行的虚拟机数量。除非 IaaS 旨在使用 Nehalem 或更新的组件运行科学或渲染类型的计算作业，否则预计典型的云部署将在每台物理基础架构服务器上的 15-30 个虚拟机上运行。在单个端口上允许这种带宽尺度的唯一技术是 40Gb/s 以太网或 40-56Gb/s 的 InfiniBand。

拥有在单个端口中提供所需容量的解决方案很重要，因为这允许密集设计。如前所述，这对于资本支出和运营支出都有益处。此外，由于不再需要用于聚合和冗余的绑定链路，简化了配置。

这种密集网络的理想外形规格是将网卡直接插入主板 (LOM) 或专用附加卡 (Mezz 或 ALOM) 中。例如，惠普 SL 390 的主板上配备 40Gb InfiniBand，Dell C6100 具有夹层附加选项，用于 Mellanox 40Gb/s 以太网或 InfiniBand。在这两个例子中，尽管具有非常密集的外形规格，但系统中的 PCIe 插槽都保持开放，允许进行扩展。

原则 2 — 简单的预配置规则

最出色的 IaaS 云能够以快速而简单的方式部署新虚拟机，并将其连接到虚拟基础架构的其余部分。此任务通常包括确定合适的 Hypervisor 来创建虚拟机，预配置虚拟机所需的存储卷，然后预配置虚拟机所需的网络连接。调度是将虚拟基础架构映射到物理资源的行为。这也是云设计中最具挑战性的方面。拥有简单的预配置方法意味着让调度选择受到限制、变得不重要或实现自动化。

在云中自动或手动执行预配置的决定在很大程度上取决于部署规模、客户群大小以及分配的动态或静态程度。最出色的解决方案能够使用良好集成的中央管理系统以任意方式工作。

Mellanox 的统一架构管理 (UFM) 软件提供了这种类型的解决方案。UFM 使管理员能够控制节点的所有方面，包括预配置、QoS 和 vNIC 管理，以及远程管理网络的固件和软件修订版的能力。UFM 还为常见任务提供了简化的自动化。当网络中发生更改（如虚拟机迁移）时，UFM 将自动预配置物理和虚拟网络元素，以保持连接和策略的完整性。UFM 负责网络调度，并集成从 MOAB 到 OpenStack 的一系列行业标准的自动调度程序，以处理非网络调度。

当今的大多数 IaaS 解决方案都有能力对 CPU 和内存等组件执行 SLA。然而，只有最出色的 IaaS 部署才能考虑并计划提供涵盖云中网络的 SLA。UFM 具有独特的性能监控引擎，可以提供有关性能卷和潜在问题（如流量瓶颈）的近实时信息。该信息自动与表示云应用程序和服务的 UFM 逻辑模型相关联。这样用户最后可以看到有一个位置能够查看网络上的云服务的服务级别、带宽消耗量和遇到的任何流量问题，以及调整和配置功能 — 因此完成了完整的操作周期。

InfiniBand 技术还提供了云供应商正在利用的一些关键优势。InfiniBand 是一种自我修复式架构，这意味着中央管理不断自动监控网络中的链路故障和拥塞链路。该技术采用基于标准的方法，可以根据实时信息动态地平衡或纠正网络中的路由。其次，InfiniBand 具有扩展优势，单个子网通常可以增长到包含 1,000 个节点，而以太网只能支持数百个节点。这简化了网络，并且可以减少与路由器配置相关的复杂性和错误。最后，InfiniBand 提供无损通信，这对于支持基于网络协议的存储和避免网络拥塞至关重要。

原则 3 — 弹性设计

IaaS 系统应该能够适应不断变化的负载要求。IaaS 应该能够在负载增加时进行纵向扩展，并在负载减少时进行合并。工作负载也可能动态变化，因此能够重新平衡很重要。从 IaaS 的角度来看，这意味着能够重新调整虚拟机或存储卷的物理位置，以平衡存储子系统上的负载。

云基础架构经理可以做出的最重要的设计决策是将存储节点与运行虚拟机的节点分离。这会带来三大优势。最重要的是可以高效地分配存储。当存储形成集群时，不会碎片化。其次，这意味着调度运行中的虚拟机映像不再局限于节点的本地存储子系统。这使分配变得简单。最后，它支持更大的弹性，因为实时迁移只涉及运行中的映像，永不包括存储卷。

提供分离式存储所面临的挑战在于网络性能。运行远程文件系统或基于块的协议（如 NFS 或 iSCSI）将需要为每个虚拟机提供额外带宽。然而光有带宽是不够的，因为存储位于远程，访问时间将受到影响，除非使用低延迟网络。分布式文件系统在这项任务中也受欢迎，因为它们将带宽需求分散在大量节点上。然而，额外的通信将在网络上产生

额外负载。为了减轻这种影响而不创建完全独立的存储网络，您需要在设计中采用具有强大的流量隔离和 QoS 规则机制的网络。

将存储与虚拟机基础架构节点分离时面临的最终风险是提高处理存储流量的 CPU 利用率。使用提供卸载或更理想 RDMA 的网络适配器来移动存储流量，可以解决此问题。

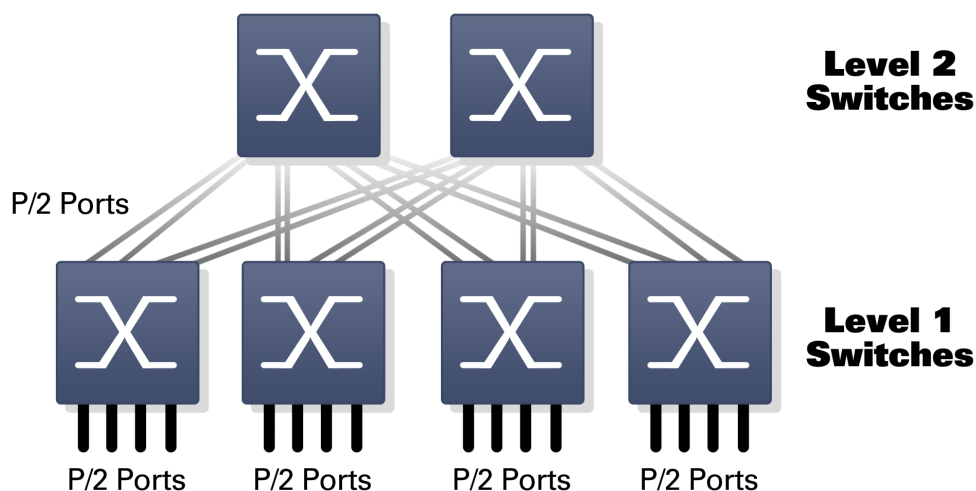
一旦存储与实时迁移过程脱钩，在基础架构节点之间重新分配负载方面，云就会变得更加灵活。最后一个挑战是最大程度地缩短迁移时间，这是一项占用大量 CPU 和网络带宽的操作。为此，采用低延迟、高带宽互连（如 40Gb/s 以太网或 56Gb/s InfiniBand）可使虚拟机实时迁移速度比传统的 10GbE 网络快 4 倍。（Beck，2011）

原则 4 — 支持东西向流量的容量

南北向流量被定义为从 Hypervisor 依次通过架顶式交换机和聚合交换机，向外发送到互联网的流量。这通常是最终用户或 WAN 生成的流量。然而，研究表明云中的大部分流量是东西向流量，这是指云中虚拟服务器之间的流量。这主要是由于 SaaS 应用程序的标准构成，它们利用许多虚拟基础架构和平台实体来构建能够在预期的响应时间内提供动态内容的应用程序。诸如 memcached 服务器、负载均衡器和数据库之类的平台元素必须从基础架构实体（如存储和云文件）检索数据，并且处理和汇总此信息。（Morgan，2011）在云内部发生的所有这些处理会产生非常庞大的东西向流量（相对于最终响应 WAN 的北向流量）。

当云基础架构不平衡以及资源零散地分布在物理基础架构的各处时，通常会做出关于东西向流量的“错误”调度决策。在预先不知道哪些虚拟基础架构设备将需要交互的情况下，保证东西向流量正确运作的唯一方法是使用保证恒定的横截面带宽和延迟的网络拓扑（如“胖树”体系架构）。

胖树体系架构与传统的三层数据中心体系架构的不同之处在于，架顶 (TOR) 和聚合层都采用完全互连的设计方式。该树将为无阻塞或具有低阻塞率。对于超大规模部署，聚合式胖树可以绑定到 POD 中并聚合到第三层。但是在这种配置中，客户端的应用程序应该绑定到一个 POD 中，以确保它们可以充分利用胖树。



总结

通过在 IaaS 设计中应用上文建议的四个原则，可以实现高度优化和高效的云部署。该方法通过实现高度动态的 IaaS 环境，可以实现最佳的硬件利用。Mellanox 的硬件和软件产品本身支持这些原则，可提供最佳的 ROI 和 TCO：

原则 1 — 可扩展的物理设计

- 选择具有板载网络（或 Mezz）的模块化密集外形规格
- 通过单一线路融合网络、存储和管理 I/O

原则 2 — 简单的预配置规则

- 使用允许自动化和手动预配置的工具
- 考虑 InfiniBand 提供动态路由并且具有自我修复能力

原则 3 — 弹性设计

- 将存储节点与基础架构节点分离
- 使用高带宽互连将实时迁移时间减少 4 倍

原则 4 — 规划容量以支持东西向流量

- 使用胖树拓扑来避免东西向流量拥塞
- 考虑 InfiniBand，它可在单个二层中扩展到数千个节点

功能特性	Mellanox FDR (56Gb/s) InfiniBand	Mellanox 10/40Gb/s 以太网	传统 1/10Gb/s 以太网
低延迟	最低	低 – VMA 或 RoCE	高
无损运行模式	原生	是，使用 DCB	是，使用 DCB
集中式管理	是	是	否
支持自动调度程序	是	是	可能是
支持 RDMA 存储	是	是	可能是
硬件强制段隔离	是	是	否
自我修复	是	否	否
动态负载平衡	是	否	否

作品引用

Beck, M. (2011). VM Migration Acceleration over 40GbE (基于 40GbE 的虚拟机迁移加速)。来自 <http://www.mellanox.com/pdf/PPT/VM%20Migration%20over%2040GigE.pdf>

Morgan, T. P. (2011 年 11 月 29 日)。The Register。来自 The Register：
[w://www.theregister.co.uk/2011/11/29/cisco_cloud_data_center_traffic_index/](http://www.theregister.co.uk/2011/11/29/cisco_cloud_data_center_traffic_index/)



北京市朝阳区望京东园七区保利国际广场 T1 15 层
Tel: 010-5789 2000
www.mellanox.com