Linux 中的虚拟网络

NICs，交换机，网络和设备

随着平台虚拟化的迅速发展，对公司生态系统的其他部分进行虚拟化也并不稀奇。最近的之一就是虚拟化网络。平台虚拟化的早期实现创建了虚拟 NICs，但是今天，网络中更大的部分正在被虚拟化，例如支持在同一个服务器上或者分布在服务器间的 VM 间通信的交换机。专注于 NIC 和交换机虚拟化，探索虚拟网络背后的创意。

M. Tim Jones 是一位嵌入式固件架构师，同时还是 Artificial Intelligence：A Systems Approach，GNU/Linux Application Programming（第二版），AI Application Programming（第二版）和 BSD Sockets Programming from a Multilanguage Perspective 的作者。他的工程背景包括地球同步航天器内核开发，嵌入式系统架构和网络协议开发等。Tim 还是科罗拉多州朗蒙特市 Emulex Corp. 的顾问工程师。

2010 年 12 月 06 日

现在计算又重新兴盛起来。虽然虚拟化出现是在几十年前，但通过商品硬件的使用，它真正的潜力现在才被认识到。虚拟化加强了服务器负载的效率，但服务器生态系统的其他部分也成为了未来加强的选项。许多人视虚拟化为 CPU，内存和存储的巩固，但是这样太过简单化解决方案了。网络是虚拟化的一个关键方面，代表虚拟化设置中第一等的元素。

联系 Tim

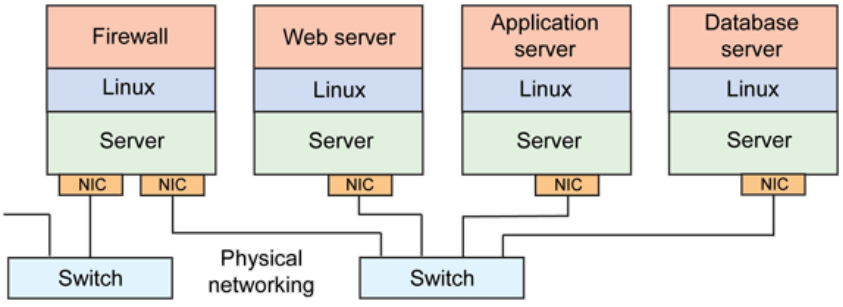
是我们最受欢迎、最多产的作者之一。 浏览 developerWorks 上 Tim 的所有文章。查看 Tim 的个人信息，并在 My developerWorks 中与 Tim、其他作者和各位读者联系。

虚拟化网络

我们从问题的高层次开始探索，然后深入到 Linux® 构建和支持的网络虚拟化各种方法。

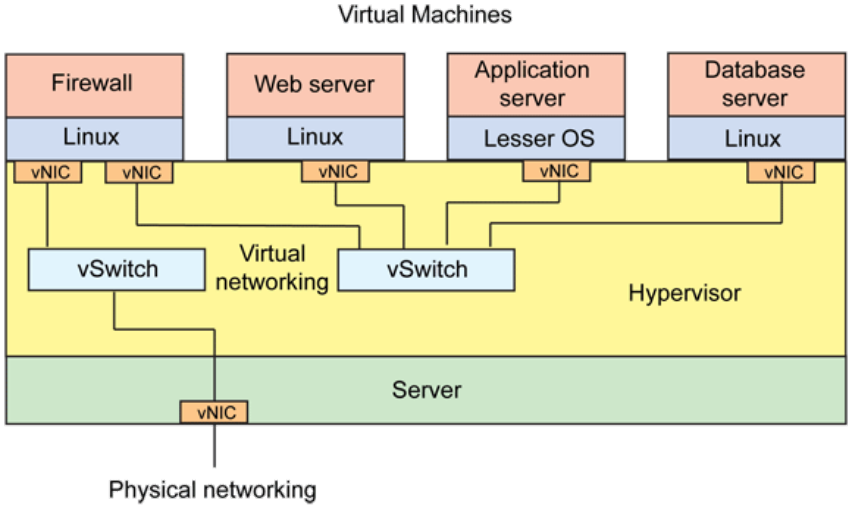
在传统环境中（见图 1），一系列物理服务器支持所需的应用程序设置。为了实现服务器间的通信，每个服务器都包含一个或者多个网络接口卡（NICs），它们连接到一个外部网络设施上。带有网络软件栈的 NIC 通过网络设施支持端点间的通信。正如图 1 所示，这个在功能上表示为一个交换机，它支持参与其中的端点间的高效数据包通信。

图 **1.** 传统网络基础架构



服务器合并背后的关键改革是物理硬件的抽象，允许多操作系统和应用程序共享硬件（见图 2）。这一改革名为 *hypervisor*（或者 *virtual machine [VM] monitor*）。每个 VM（一个操作系统和应用程序设置）视底层硬件为非共享的，一个完整机器，即使它们部分可能并不存在，或者被多个 VM 共享。虚拟的 NIC（vNIC）就是一个例子。管理程序为每个 VM 创建一个或者多个 vNICs。这些 NICs 对 VM 可以作为物理 NICs，但是它们实际上只表示 NIC 的接口。管理程序也允许虚拟网络的动态构建，由虚拟交换机完成，支持可配置的 VM 端点间的通信。最后，管理程序还允许和物理网络基础架构的通信，通过将服务器的物理 NICs 连接到管理程序的逻辑设施，允许管理程序中 VMs 间高效的通信，以及和外部网络的高效通信。在 参考资料 部分，您将会找到更多关于 Linux 管理程序信息的链接（开源操作系统的丰富区域）。

图 **2.** 虚拟的网络设施



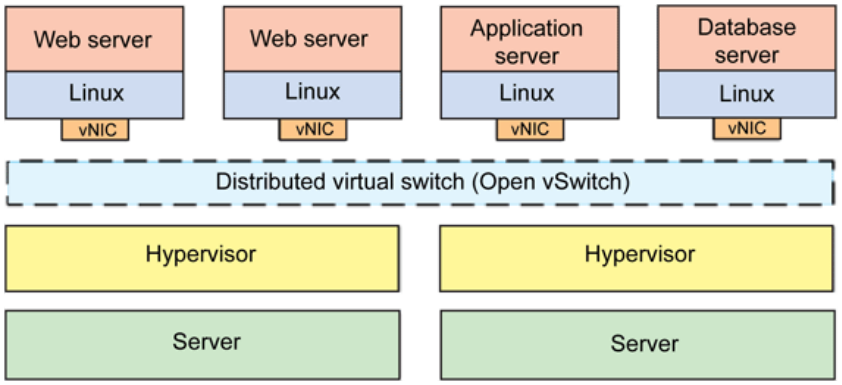
虚拟网络设施还支持其他有趣的革新，比如虚拟设备。除了虚拟网络的元素以外，我们还关注这些内容，作为该探索的一部分。

虚拟交换机

虚拟网络设施的关键开发之一就是虚拟交换机的开发。虚拟交换机连接 vNICs 到服务器的物理 NICs，并且 —更重要的是—它将 vNICs 连接到服务器中的其他 vNICs，进行本地通信。这之所以有趣是因为在一个虚拟交换机中，所受限制和网络速度无关，而是和内存带宽有关，它允许本地 VMs 间的高效通信，并且最小化网络设施的开销。这个节省是源自物理网络只用于服务器间的通信，服务期间的跨 VM 通信被隔离。

但是，因为 Linux 已经在内核中包含一个 2 层交换机，所以有人可能会问，为什么会需要虚拟交换机？答案包括多个属性，但是最重要之一的是由这些交换机类型的新分类定义的。新的类名为分布式虚拟交换机，它采用使底层服务器架构更透明的方法，支持跨服务器桥接。一个服务器中的虚拟交换机能够透明地和其他服务器中的虚拟交换机连接（见 图 3），使服务器间（以及它们的虚拟接口）的 VM 迁移更简单，因为它们可以连接到另一个服务器的分布式虚拟交换机，并且透明地连接到它的虚拟交换网络。

图 **3.** 分布式虚拟交换机



在这期间最重要的项目之一名为 *Open vSwitch*， 接下来本文会探讨这部分内容。

在服务器中隔离本地流量的一个问题就是流量不是外部可视的（例如，对网络分析员）。实现通过各种计划解决了这一问题，例如 OpenFlow，NetFlow 和 sFlow，它们还用于输出远程访问来控制和监控流量。

Open vSwitch

分布式虚拟交换机的早期实现已经结束，并且受限于管理程序专有设置的操作。但是在今天的云环境中，支持多管理程序共存的异构环境是很理想的。

Open vSwitch 是一个多层的虚拟交换机，在 Apache 2.0 许可下可用作为开放资源。截止 2010 年 5 月，Open vSwitch 已有版本 1.0.1 可用，并且支持一系列有用的功能。Open vSwitch 支持领先的开源管理程序解决方案，包括基于内核的 VM（KVM），VirtualBox，Xen 和 XenServer。它还是当前 Linux 桥模块的下拉替换。

Open vSwitch 由交换机守护，管理基于流的交换机的配套内核模块组成。还存在各种其他的守护程序和实体，用于管理交换机（特别是从 OpenFlow 方面）。您可以在用户空间完全运行 Open vSwitch，但是这么做会导致性能的下降。

除了为 VM 环境提供一个生产品质的交换机，Open vSwitch 还有令人印象深刻的功能路线图，和其他相似的、专有的解决方案竞争。

网络设备虚拟化

NIC 硬件的虚拟化以各种形式已经存在了一段时间 — 在虚拟交换机出现之前。本节将说明实现和硬件加速的部分内容，它们可用于改善网络虚拟化的速度。

QEMU

虽然 QEMU 是一个平台模拟器，但它还提供各种硬件设备的软件模拟，包括 NICs。此外，QEMU 还提供了用于 IP 地址分配的内部 Dynamic Host Configuration Protocol 服务器。QEMU 和 KVM 一起运作，提供平台模拟和独立的设备模拟，为基于 KVM 的虚拟化提供平台。您可以在 参考资料 部分了解更多关于 QEMU 的内容。

virtio

virtio 是一个 Linux 的输入/输出（I/O）准虚拟化框架，它简化并加快了 VM 到管理程序的 I/O 通信。virtio 创建了 VM 和用于虚拟块设备，通用的外围组件互连（PCI）设备，网络设备等的管理程序间 I/O 的标准化传输机制。您可以在 参考资料 部分了解更多 virtio 的内容。

TAP 和 TUN

虚拟化在网络栈中实现已经有一段时间了，允许 VM 访客网络栈访问主机网络栈。计划之二就是 TAP 和 TUN。TAP 是一个虚拟网络内核驱动，该驱动实现 Ethernet 设备，并在 Ethernet 框架级别操作。TAP 驱动提供了 Ethernet “tap”，访客 Ethernet 框架能够通过它进行通信。TUN（或者网络“通道”）模拟网络层设备，并且在 IP 数据包的较高层进行通信，这些数据包提供一些优化，因为底层 Ethernet 设备能够管理 TUN 的 IP 数据包的 2 层框架。

I/O 虚拟化

I/O 虚拟化来自在硬件层上支持加速虚拟化的 PCI-Special Interest Group（SIG）的标准化计划。特别是，Single-root IOV（SR-IOV）提供一个接口，通过它独立的 PCI Express（PCIe）卡能够作为多 PCIe 卡出现在众多用户面前，允许多个独立的驱动连接到 PCIe 卡，无需相互了解。SR-IOV 通过将虚拟功能扩展到各种用户来实现，这是作为 PCIe 空间的物理功能，但是在卡中作为共享功能表示。

SR-IOV 带给网络虚拟化的好处就是性能。比起实现物理 NIC 共享的管理程序，卡自身实现复合，允许从访客 VM I/O 接口直接到卡的通路。

Linux 今天包含对 SR-IOV 的支持，这对 KVM 管理程序很有好处。Xen 也包括对 SR-IOV 的支持，允许它高效地向访客 VMs 显示 vNIC。对 SR-IOV 的支持在 Open vSwitch 的路线图上。

虚拟 LANs

虽然相关，但是虚拟 LANs（VLANs）是网络虚拟化的物理方法。VLANs 提供创建跨分布网络的虚拟网络的能力，这样就会出现不同的主机（在独立的网络上），如果它们是相同广播域的一部分。VLANs 通过使用 VLAN 信息标记框架完成这个，用来识别特定 LAN（按照 Institute of Electrical and Electronics Engineers [IEEE] 802.1Q 标准）的成员关系。主机和 VLAN 交换机一起运作，进行物理网络虚拟化。然而，虽然 VLANs 提供独立网络的假象，但它们共享同一个网络以及可用带宽，影响阻塞带来的结果。

硬件加速

许多针对 I/O 的虚拟化加速开始出现，寻址 NICs 和其他设备。Intel® Virtualization Technology for Directed I/O（VT-d）提供隔离 I/O 资源的功能来获得改进的可靠性和安全性，它包括重映射直接内存访问（使用多级页表）和设备相关的中断重映射，支持未修正的和虚拟化感知的访客。Intel Virtual Machine Device Queues（VMDq）还通过硬件中的嵌入排序和智能排序，加速了在虚拟化设置中的网络通信流，实现了管

理程序较低的 CPU 利用率和总体系统性能的更大程度改善。Linux 包含对两者的支持。

网络虚拟设备

目前为止，本文探讨了 NIC 设备和交换机的虚拟化，当前实现的部分内容，通过硬件加速虚拟化的部分方法。现在，我们将这个讨论扩大到通常的网络服务。

虚拟化范围内的有趣革新之一就是从服务器整合演化而来的生态系统。比起将应用程序投入到特定的硬件版本，服务器的一部分和服务器内扩展服务的强大 VM 相隔离。这些 VMs 被称为虚拟设备，因为它们关注一个特定的应用程序，被部署用于虚拟化设置。

虚拟设备通常连接到管理程序 — 或者有管理程序的良好网络设置— 来扩展特定的服务。这个之所以独特是因为，在合并服务器中，处理功能的部分（例如核）和 I/O 带宽能够为虚拟设备动态地配置。这个功能使它更成本有效（因为一个独立的服务器并不会为它而被隔离），并且您能够根据在服务器上运行的其他应用程序的需求，动态地改变它的功能。虚拟设备还能更易于管理，因为应用程序被绑定在操作系统中（在 VM 内）。无需特殊配置，因为 VM 是作为整体进行预配置的。这对于虚拟设备来说是个值得考虑的好处，这也是今天它一直发展的原因。

虚拟设备已经为许多企业软件进行了开发，并且包括 WAN 优化，路由器，虚拟专用网，防火墙，防止/检测入侵的系统，邮件分类和管理等等。除了网络服务以外，虚拟设备还用于存储，安全，应用程序框架以及内容管理。

结束语

曾几何时一切都可管理还是物理上可实现的。但是今天，在我们不断虚拟化的世界中，物理设备和服务已经消失不见。物理网络被虚拟化地分割，允许通信隔离和跨地理实体的虚拟网络的构建。应用程序消失在虚拟设备中，这些设备在强大服务器的核之间被分割，虽然为管理者添加了很多复杂性，但是也提供了更好的灵活性，改善了可管理能力。当然，Linux 就走在前沿。

参考资料

学习

- Linux 代表了良好的操作系统和虚拟化解决方案的平台。您可以在 虚拟 Linux（developerWorks，2006 年 12 月）和 剖析 Linux hypervisor（developerWorks，2009 年 5 月）了解到更多关于 Linux 和虚拟化的内容。
- Linux 实现了名为 `virtio` 的 I/O 虚拟化框架（由 KVM 使用）。`virtio` 为高效的准虚拟化驱动开发提供了通用框架。您可以在 Virtio：针对 Linux 的 I/O 虚拟化框架

（developerWorks，2010 年 1 月）了解更多关于 `virtio` 及其内在的内容。

- SR-IOV 提供了对被多个访客 VM 使用的物理适配器进行虚拟化的方法。您可以在 Linux 虚拟化和 PCI 透传技术（developerWorks，2009 年 10 月）读到更多关于设备模拟和 I/O 虚拟化的内容。
- SR-IOV 允许多访客操作系统共享 PCIe 设备。您可以从 Intel 硬件设计网站 了解更多关于 SR-IOV 的内容。PCI-SIG 提供了各种 IOV 技术的规格。
- 虚拟设备是软件设备传输的相对较新形式。虚拟设备中的重要目标就是共享的能力，而不是一个管理程序的最大便携性。在这个方向上的一个方法就是 Open Virtualization Format（OVF），它定义了虚拟设备元数据的格式。您可以在虚拟设备和 Open Virtualization Format（developerWorks，2009 年 10 月）了解更多关于虚拟设备和 OVF 的内容。
- QEMU 是全计算机系统的开源模拟器，还提供完全的虚拟化解决方案（通过模拟）。您可以在 使用 QEMU 进行系统仿真（developerWorks，2007 年 9 月）了解更多关于 QEMU 的内容。
- IEEE 802.1Q 标准提供了 VLAN 标记的网络标准，定义了设备虚拟隔离中位于 MAC 层的 VLAN 的概念。
- 在 developerWorks Linux 专区 寻找为 Linux 开发人员（包括 Linux 新手入门）准备的更多参考资料，查阅我们 最受欢迎的文章和教程。
- 在 developerWorks 上查阅所有 Linux 技巧 和 Linux 教程。
- 随时关注 developerWorks 技术活动和网络广播。
- 观看 developerWorks 演示中心，包括面向初学者的产品安装和设置演示，以及为经验丰富的开发人员提供的高级功能。

获得产品和技术

- OpenSolaris 实现了作为名为 *Crossbow* 项目一部分的 NIC 和交换机虚拟化。Project Crossbow 将虚拟化和带宽资源控制带入到网络栈中，最小化复杂性和开销。
- Open vSwitch 是第一个开源多层交换机，用于服务虚拟化的生态系统。在最近发布的版本 1.0 中，Open vSwitch 提供良好的功能清单，支持许多开源管理程序（包括 KVM，Xen，XenServer 和 VirtualBox）。
- Xen Cloud Platform 是结合了作为其部分栈的 Open vSwitch 虚拟交换机数据包的虚拟化设施。

- 以最适合您的方式 IBM 产品评估试用版软件：下载产品试用版，在线试用产品，在云环境下试用产品，或者在 IBM SOA Sandbox for People 中花费几个小时来学习如何高效实现面向服务架构。

讨论

- 欢迎加入 My developerWorks 中文社区。
