白皮书

软件定义的网络 - 正确实现

目录

概述

软件定义的网络 (SDN) 是用于设计、构建和运营网络的革命性方法，旨在通过网络抽象化、虚拟化和编排来降低资本和运营成本，同时实现业务敏捷性。从概念上讲，SDN 使控制平面与数据平面分离，并在基于软件的控制器中以逻辑方式集中网络智能和控制，维护网络的全局视图。这样可以从应用程序中实现更精简的策略驱动型外部控制和自动化，从而最终增强网络可编程性并简化网络编排。因此，基于 SDN 的设计可实现具有高度弹性的网络，可以轻松地适应不断变化的业务需求。

第一波 SDN 部署专注于功能，但随着数据中心互连技术实现大量创新和增强，现在是时候以一个全新的角度重新审视更高效、更高性能的 SDN 部署选项了。

本白皮书重点讨论面向数据中心的 SDN 解决方案，这通常是构建云（无论是私有云还是公共云）的重要组成部分。其中包含以下主题：

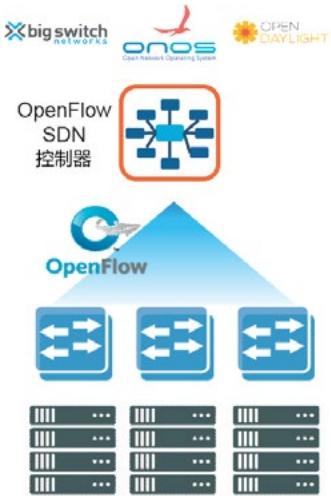
- 主要 SDN 部署模式的概述；
- Mellanox SDN 技术要点，其中介绍了使 SDN 部署更为高效的重要 Mellanox 功能和产品；
- 适用于 OpenFlow 和叠加 SDN 部署模式的 Mellanox SDN 解决方案。我们介绍了如何配合使用 Mellanox 产品和功能才能提供适用于各种 SDN 部署情况的总体解决方案，以及 Mellanox 解决方案所带来的主要优势。

SDN 部署模式

三种不同部署模式在当今的 SDN 环境中占据主导地位：

基于设备的 SDN 部署模式

在此模式中，SDN 控制器使用南向设备控制协议直接通信策略或将表信息转发给物理和虚拟交换和路由设备。OpenFlow 是最常用的协议，一些早期的 SDN 体系架构基于 OpenFlow 来使控制平面与网络设备分离。



基于此模式的 SDN 实现的示例包括 BigSwitch 的 Big Cloud Fabric、Open Networking Lab (ON.LAB) 的 ONOS，以及 Open Daylight (ODL)。除 OpenFlow 之外，ONOS 和 ODL 还支持使用其他南向协议（如 Netconf 和 SNMP）进行设备配置和管理。

基本上对于每个新流，该流遍历的所有设备都可能需要进行编程以处理适当的流操作。此模式要求网络设备支持 OpenFlow，当您用的是传统网络或将各代网络设备混用时，有时这会成为挑战。

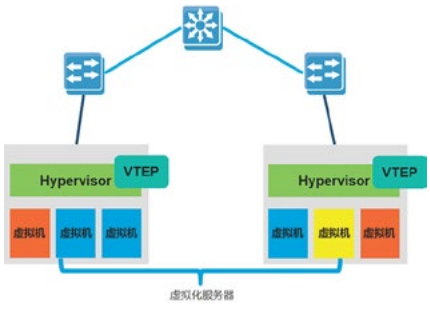
叠加 SDN 部署模式

许多客户已经安装了尚未启用 OpenFlow 的网络设备，可能无法将网络范围的升级纳入选择范围。叠加式 SDN 部署模式应运而生，为这些客户提供 SDN/网络虚拟化，而不需要进行可能产生高昂成本和中断业务服务的全面网络升级。叠加 SDN 已成为最常见的体系架构，诸如 VMware NSX、Nuage Networks（现在是诺基亚的一部分）VSP、PLUMGrid ONS、OpenContrail 和 Midokura MidoNet 之类的主流 SDN 解决方案都主要采用此模式。

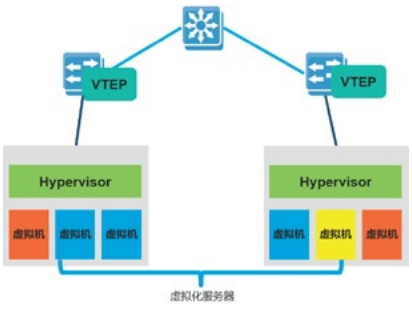
顾名思义，在叠加模式中，通过端点之间的隧道建立逻辑网络，而这些隧道叠加在现有物理网络上。关于多租户、网络和安全策略的智能被推向网络边缘。最常用的一些隧道协议包括虚拟可扩展 LAN (VXLAN)、使用 GRE 的网络虚拟化 (NVGRE) 和通用网络虚拟化封装 (GENEVE)。对于 VXLAN，隧道端点称为 VXLAN 隧道端点 (VTEP)。物理网络或底层成为"核心"网络，其功能可能被简化为在这些 VTEP 之间提供高性能 IP 连接。叠加 SDN 控制器将主要与 VTEP 通信，该控制器通常是驻留在服务器上的虚拟交换和路由设备。

可以部署叠加 SDN 来实现网络虚拟化和自动化，而不需要升级物理网络设备（具体来说是非 VTEP 网络设备）。尽管它有各种优点，叠加 SDN 在同时管理叠加层和底层以及在故障排除期间为来自两个层的信息建立联系方面也会引入额外的复杂性。

部署 VTEP 有两种常用的方法：虚拟交换机中的软件 VTEP，通常在服务器 Hypervisor 中运行；或架顶式 (ToR) 交换机中的硬件 VTEP，这两种方法之间存在权衡取舍。软件 VTEP 具有灵活和概念简单的特点，但会影响性能和增加边缘设备上的 CPU 开销，这是与相对较新的隧道协议相关的数据包处理所导致的，并非所有服务器网卡 (NIC) 都能从 CPU 卸载。当应用程序本身是网络功能虚拟化 (NFV) 部署中的虚拟化网络功能 (VNF) 时，这种情况会更加明显。硬件 VTEP 通常可以实现更高的性能，但增加了 ToR 交换机的复杂性，因为 ToR 交换机需要支持虚拟机、维护大型转发表，以及执行从虚拟机 Mac 地址或 VLAN 到 VXLAN 的转换。



软件 VTEP



硬件 VTEP

除了使用 VXLAN/NVGRE/GENEVE 的虚拟化环境外，通常还有裸机服务器 (BMS) 或传统网络，它们只能使用 VLAN，或发往 VPN 网络或互联网的南北向流量。在这些情况下，使用软件 VTEP 网关会增加额外跳转，还可能造成性能瓶颈，而最佳做法是使用 BMS 所连接的 ToR 作为硬件 VTEP。

Mellanox SDN 技术要点

专有 SDN 解决方案

市场上还有其他专有 SDN 解决方案，例如 Cisco Application Centric Infrastructure (ACI)、Plexxi 和 Pluribus。对于这些解决方案，SDN 控制器和 SDN 交换与路由元件通常紧密耦合。这类 SDN 解决方案不像上述两种解决方案那么开放，并且限制了生态系统供应商与它们进行整合。Mellanox 目前只能使用开放式 SDN 解决方案。

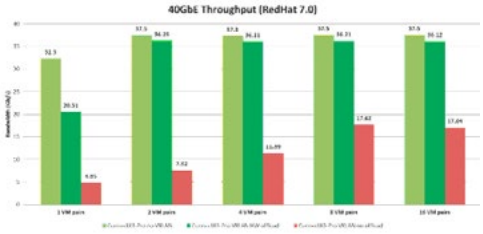
ConnectX 网卡上的 VXLAN 卸载

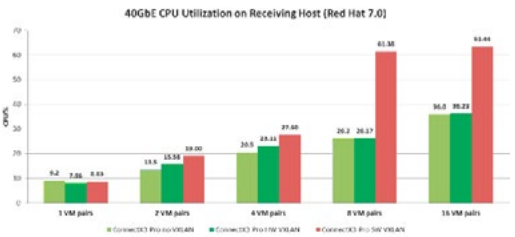
早期通过 VLAN 实现网络虚拟化时，在服务器主机上还可能实现线路速率性能，因为服务器能够将一些 CPU 密集型数据包处理操作（例如校验和、接收端转向 (RSS)、大型接收卸载 (LRO) 等）卸载到网卡硬件。这既提高了网络 I/O 性能，又降低了 CPU 开销，最终使基础架构更高效地运行。

如上文所述，对于叠加 SDN，引入了诸如 VXLAN、NVGRE 或 GENEVE 之类的隧道协议来封装原始有效载荷。对于不能识别这些新数据包头格式的网卡，即使最基本的卸载功能也会停止工作，导致所有数据包处理操作都由软件在 CPU 中完成。这可能导致网络 I/O 性能显著下降和过多的 CPU 开销，随着服务器 I/O 速度从 10Gb/s 发展到 25、40、50 甚至 100Gb/s，这一问题尤为明显。

从 ConnectX-3 Pro 系列网卡开始，Mellanox 支持 VXLAN 硬件卸载，包括无状态卸载，例如 VXLAN/NVGRE/GENEVE 数据包的校验和、RSS 和 LRO。对于 VXLAN 卸载，I/O 性能和 CPU 开销可恢复到与 VLAN 相似的水平。

下面两个图表显示了三种情形下的带宽和 CPU 开销比较：VLAN、不使用卸载的 VXLAN 和使用卸载的 VXLAN。VXLAN 卸载导致吞吐量提高了 2 倍多，并使 CPU 开销降低约 50%。



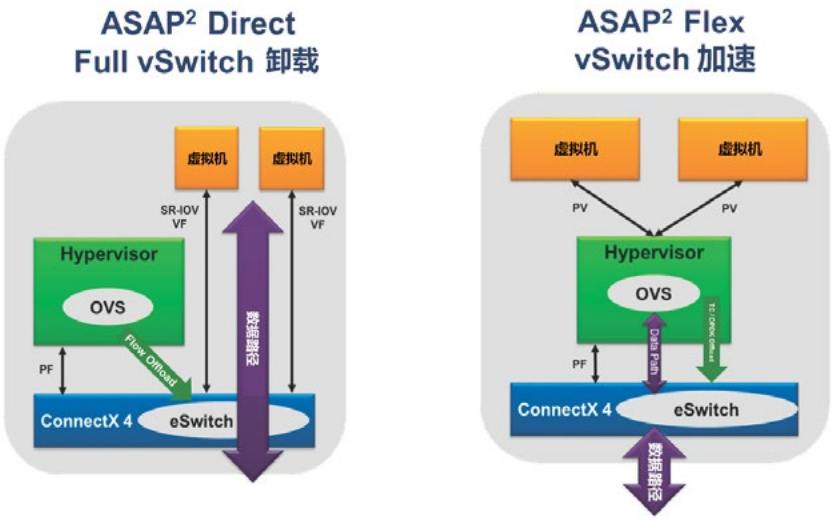


对于 Linux、Microsoft Hyper-V 和 VMWare ESXi，在操作系统/Hypervisor 内核级别上支持 VXLAN 卸载，并且不依赖于所使用的虚拟交换机或路由器的类型。

ConnectX-4 网卡上的 ASAP² （加速交换和数据包处理）

从 ConnectX-4 系列网卡开始，Mellanox 在服务器网卡硬件中通过 ASAP2 功能来支持 VTEP 功能。利用网卡中内置的基于管线的可编程 eSwitch，ConnectX-4 可以在硬件中处理大部分数据包处理操作。这些操作包括 VXLAN 封装/解封、基于一组通用 L2-L4 包头字段的数据包分类、QoS 和访问控制列表 (ACL)。ASAP2 功能基于这些增强的网卡硬件功能构建，提供可编程、高性能和高效的硬件转发平面，可与 SDN 控制平面无缝协作。它克服了与软件 VTEP 相关的性能降级问题，以及在使用硬件 VTEP 的情况下，服务器与 TOR 设备之间的协调复杂性问题。

有两种主要的 ASAP2 部署模式：ASAP2 Direct 和 ASAP2 Flex



ASAP² Direct

在此部署模式中，虚拟机通过 SR-IOV 虚拟功能 (VF) 建立对 Mellanox ConnectX-4 网卡硬件的直接访问，从而在虚拟化环境中实现最高的网络 I/O 性能。

与传统 SR-IOV 实现相关的问题之一是它会完全绕过 Hypervisor 和虚拟交换机，并且虚拟交换机在 SR-IOV 模式中不知道虚拟机的存在。因此，对于在服务器主机上使用 SR-IOV 的那些虚拟机，SDN 控制平面不能影响其转发平面。

ASAP2 Direct 通过在虚拟交换机与 ConnectX-4 eSwitch 转发平面之间启用规则卸载来克服此问题。在此情况下，我们使用开放虚拟交换机 (OVS) 作为示例，这是最常用的虚拟交换机之一。与相应 SDN 控制器通信的 OVS 所提供的 SDN 控制平面与网卡硬件转发平面相结合，提供了以下两方面的最佳性能：软件定义的灵活网络可编程性，以及适用于从 10G 到 25/40/50/100G 的最新速度的高网络 I/O 性能。通过让网卡硬件承担 CPU 的 I/O 处理负荷，CPU 资源可以专门用于应用程序处理，从而提高系统效率。

ASAP2 Direct 提供超越原始位吞吐量的优异小数据包性能。我们的基准测试显示，在具有 25G 接口的服务器上，ASAP2 Direct 实现了单流 33 百万数据包/秒 (MPPS) 的性能，并且消耗零 CPU 内核；而在 ConnectX-4 Lx eSwitch 中执行 VXLAN 封装/解封的 15,000 流实现了约 25 MPPS 的性能。

ASAP² Flex

在此部署模式中，虚拟机在半虚拟化模式下运行，并且仍然通过虚拟交换机满足其网络 I/O 需求。但是，通过一组开放 API（例如 Linux 流量控制 (TC) 或数据路径开发工具包 (DPDK)），虚拟交换机可以将一些 CPU 密集型数据包处理操作卸载到 Mellanox ConnectX-4 网卡硬件，其中包括 VXLAN 封装/解封和数据包分类。这是一项路线图功能，推出日期将在以后公布。

Spectrum 交换机上的 OpenFlow 支持

Spectrum 是 Mellanox 的 10/25/40/50 和 100Gb/s 以太网交换机解决方案，针对 SDN 进行了优化，可实现灵活高效的数据中心结构，具有领先的端口密度、低延迟、零丢包和无阻塞流量。

从底层开始，在交换机芯片级别上，Spectrum 设计为具有非常灵活的处理管线，使其可以适应可编程 OpenFlow 管线，允许将数据包发送到后续表以用于进一步处理，并允许在 OpenFlow 表之间传递元数据信息。此外，Spectrum 是一种 OpenFlow 混合交换机，同时支持 OpenFlow 操作和正常的以太网交换操作。用户可以在端口级别配置 OpenFlow，分配一些 Spectrum 端口来执行基于 OpenFlow 的数据包处理操作，分配其他端口来执行正常的以太网交换操作。此外，Spectrum 还提供了一种分类机制，将一个交换机端口内的流量引导到 OpenFlow 管线或正常的以太网处理管线。

下面简要列出了 Spectrum 所支持的 OpenFlow 功能：

- OpenFlow 1.3 控制数据包解析
  - » 将接口映射到 OpenFlow 混合端口
- 提供与 Open Daylight 和 ONOS 控制器的互操作性。
- 设置 OpenFlow 桥接 DATAPATH-ID
- 显示由控制器所配置的流
- 支持通过 ACL 在硬件中使用弹性规则
- 查询表功能
- 表操作 – 添加、删除、修改、优先级、硬超时
- 可配置的远程控制器：
  - » 选择性发送到控制器
- 按端口计数器
- 按端口状态：
  - » STP
  - » 操作
  - » 速度

Spectrum 交换机中的 VTEP 支持

- 第 2 层 VTEP 网关，在使用 VXLAN 的虚拟化网络与使用 VLAN 的非虚拟化网络之间（同一数据中心内或不同的数据中心之间）。
- 第 2 层 VTEP 网关，跨第 3 层网络提供到虚拟化服务器的高性能连接，并支持第 2 层功能，例如虚拟机实时迁移 (VMotion)。在网卡不具备 VTEP 功能并且软件 VTEP 无法满足网络 I/O 性能要求的虚拟化服务器主机上，可以在 Mellanox Spectrum ToR 上实现 VTEP。在某些情况下，在虚拟机中运行的应用程序可能希望使用高级网络功能（如远程直接内存访问 (RDMA)）进行虚拟机间通信或访问存储。在虚拟化服务器上，RDMA 需要在 SR-IOV 模式下运行，为了防备没有 Mellanox 网卡的情况，最好在 ToR 中实现 VTEP。
- 第 3 层 VTEP 网关，为不同 VXLAN 虚拟网络之间的流量提供 VXLAN 路由功能，或者为 VXLAN 网络与 VPN 网络或互联网之间的南北向流量提供 VXLAN 路由功能。在 Spectrum 硬件中支持此功能，而支持此功能的软件仍在开发中。

Spectrum 是开放式以太网交换机，可支持运行多种交换机操作系统。第 2 层 VTEP 网关功能将首先在 Cumulus Linux over Spectrum 中提供，随后在 MLNX-OS 中提供。

这样可以实现线路速率性能，同时节省计算和存储成本。

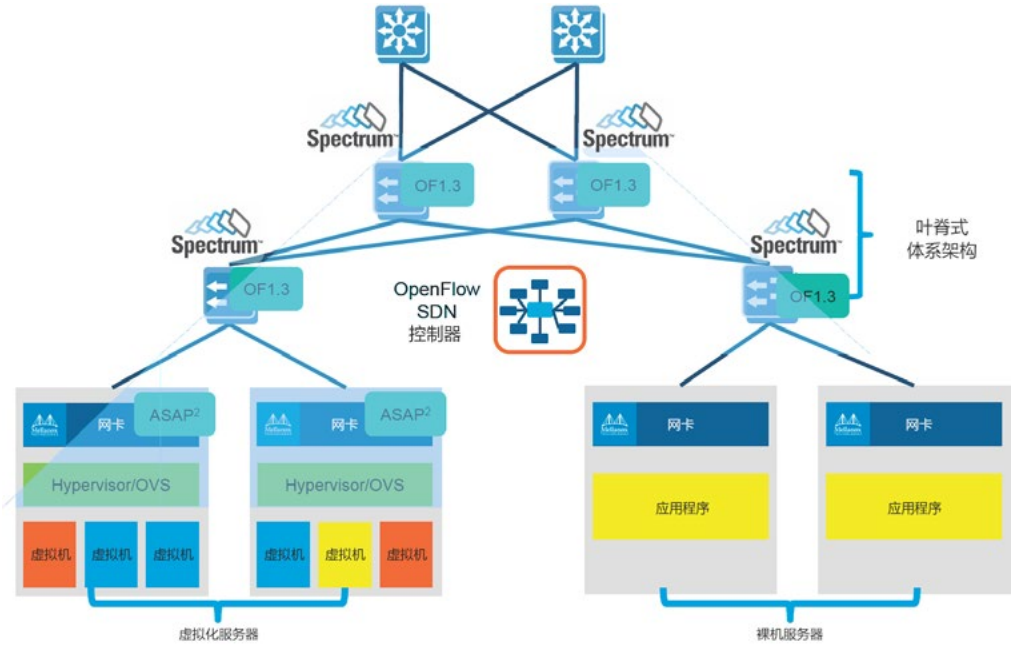
一些主要存储协议和平台现在都支持 RDMA。例如 iSER （基于 RoCE 的 iSCSI）、SMB Direct、OpenStack Cinder 中的 iSER 和 Ceph RDMA 等。

下面是 iSER 和 SMB Direct 的性能优势：

使用 Mellanox 互连技术构建最高效的 SDN 网络

在本节中，关于基于 OpenFlow 和叠加模式构建最高效的 SDN 网络，我们提供了最佳做法建议。

针对基于 OpenFlow 的 SDN 网络的建议部署



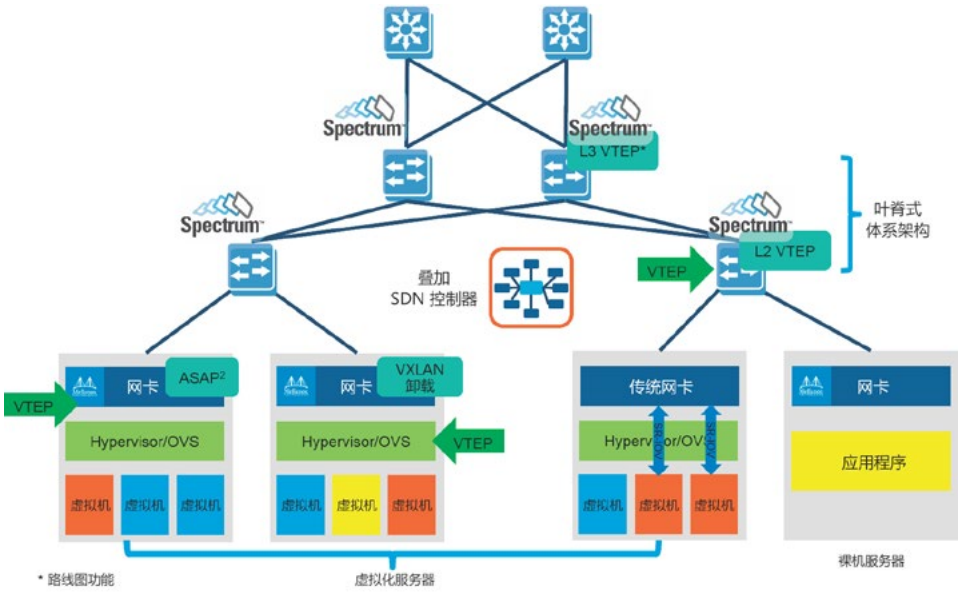
要点：

- 使用支持 OpenFlow 1.3 的 Spectrum 交换机构建的叶脊式体系架构
- 对于物理 + 虚拟结构，利用 ASAP2 将虚拟交换机数据平面卸载到 Mellanox ConnectX-4 网卡
- 利用 Spectrum 上的 SFLOW 功能的高级流量监控

主要优势：

- 高性能，提供从 10G 到 100Gb/s 之间的任意速度下的线路速率性能，在虚拟化服务器上也能实现
- 最灵活的 OpenFlow 交换机实现
- 对 OpenFlow 结构的深度可见性

针对叠加 SDN 网络的建议部署



要点：

- 多种虚拟化服务器部署方式：
  - » 虚拟交换机作为 VTEP + Mellanox VXLAN 无状态卸载
  - » Mellanox 网卡作为 VTEP（利用 ASAP2 将虚拟交换机数据平面卸载到 Mellanox ConnectX-4 网卡，同时在虚拟交换机中保留 SDN 控制平面操作）
  - » 对于需要 SR-IOV 的虚拟机以及不支持 ASAP2 的传统网卡，使用 Mellanox Spectrum ToR 作为 VTEP
- Mellanox Spectrum ToR 上的高性能硬件 VTEP，用于裸机服务器或存储调配
- （路线图功能）用于 VXLAN 路由的 Mellanox Spectrum 主干交换机上的高性能硬件第 3 层 VTEP。
- 利用 Spectrum 上的 SFLOW 功能的高级底层结构监控

主要优势：

- 高性能，提供从 10G 到 100Gb/s 之间的任意速度下的线路速率性能，在虚拟化服务器上也能实现；
- 面向未来的最先进 VTEP 实现，具有在网卡或交换机级别执行 VTEP 的灵活性，能够扩展到第 3 层硬件 VTEP，而无需全面升级；
- 对 SDN 底层结构的深度可见性便于关联来自两个层的统计信息，并实现轻松故障排除。

结论

SDN 是一项不断发展的技术，Mellanox 作为独一无二的数据中心网络连接解决方案供应商，能够利用其端到端互连技术及相关软件为 SDN 提供最全面、最灵活、最高效的支持。

微信

北京迈络思科技有限公司
咨询电话：+86-10-57892000
销售咨询：china_sales@mellanox.com
市场合作：marketing_cn@mellanox.com
*欲了解更多欢迎登陆www.mellanox.com

MLNX-15-5938