# VMware® NSX for vSphere (NSX-V) 网络虚拟化设计指南

# 目标受众

本文档面向有兴趣在 vSphere 环境中部署 VMware® NSX 网络虚拟化解决方案的虚拟化和网络架构师。

# 概述

服务器虚拟化的直接成效显著，为 IT 组织带来了巨大的优势。服务器整合降低了物理复杂性，提高了运维效率，并且能够动态调整底层资源的用途以便快速且充分满足日趋动态化的业务应用的需求，这些仅仅是已实现功效的少数几个例子。

现在，VMware 软件定义的数据中心 (SDDC) 体系结构可在整个物理数据中心基础架构中扩展虚拟化技术。VMware NSX 是网络虚拟化平台，是 SDDC 体系结构中的关键产品。利用 VMware NSX，虚拟化现在可为网络连接提供它已经为计算和存储提供的虚拟技术。与服务器虚拟化以编程方式创建、删除和还原基于软件的虚拟机 (VM) 并为其拍摄快照大致相同，VMware NSX 网络虚拟化能够以编程方式创建、删除和还原基于软件的虚拟网络并为其拍摄快照。这最终形成了完全革新的网络连接方法，不仅使数据中心管理员能够将敏捷性和经济效益提高若干数量级，还为底层物理网络提供了大大简化的运维模式。NSX 是一款完全无中断的解决方案，能够部署在任何 IP 网络中，包括现有的传统网络连接模型，以及任何供应商提供的新一代结构架构。实际上，借助 NSX，您只需通过已有的物理网络基础架构即可部署软件定义的数据中心。

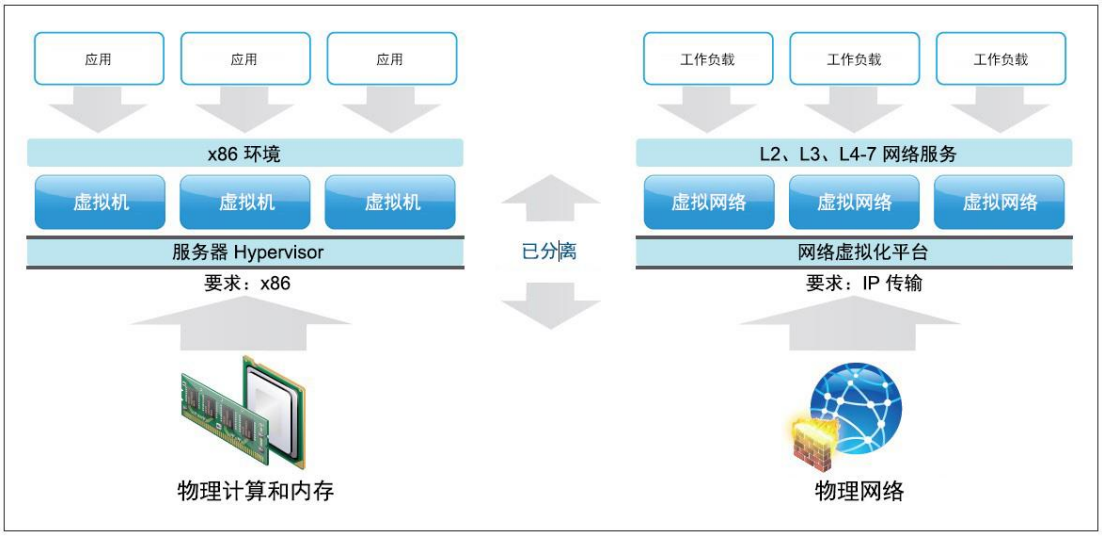



**图 1 - 服务器和网络虚拟化类比**

图 1 指出了计算与网络虚拟化之间的相似处。借助服务器虚拟化，软件抽象层（服务器 hypervisor）将在软件中重现熟悉的 x86 物理服务器属性（如 CPU、RAM、磁盘、网卡），允许它们以编程方式组建为任意组合，从而在数秒内生成唯一的虚拟机 (VM)。

借助网络虚拟化，与“网络 hypervisor”对等的功能可以软件形式重现第 2 至 7 层的全套网络连接服务（例如，交换、路由、防火墙和负载均衡）。因此，这些服务能够以编程方式组建为任意组合，只需要几秒钟时间即可生成唯一的、隔离的虚拟网络。

毫无疑问，这也会产生类似的优势。例如，正如虚拟机独立于底层 x86 平台并允许 IT 将物理主机视为计算容量池一样，虚拟网络也独立于底层 IP 网络硬件，并允许 IT 将物理网络视为一个可按需使用和调整用途的容量传输池。不同于传统体系结构，虚拟网络无需重新配置底层物理硬件或拓扑即可以编程方式进行调配、更改、存储、删除和还原。通过配合使用熟悉的服务器和存储虚拟化解决方案提供的功能和优势，这种革命性的网络连接方法能够释放软件定义的数据中心的全部潜力。

只需采用 VMware NSX，您就拥有了部署新一代软件定义的数据中心所需的网络。本白皮书将重点介绍 VMware NSX for vSphere (NSX-v) 体系结构提供的各种功能，以及您在充分利用现有网络投资并通过 VMware NSX-v 优化该投资时应考虑的设计因素。

**注意**：在本白皮书的其余章节，术语“NSX”和“NSX-v”交替使用，它们都是指 NSX with vSphere 部署实例。

# 网络虚拟化简介

## NSX-v 网络虚拟化解决方案概述

NSX-v 部署由数据平面、控制平面和管理平面组成，如图 2 所示。

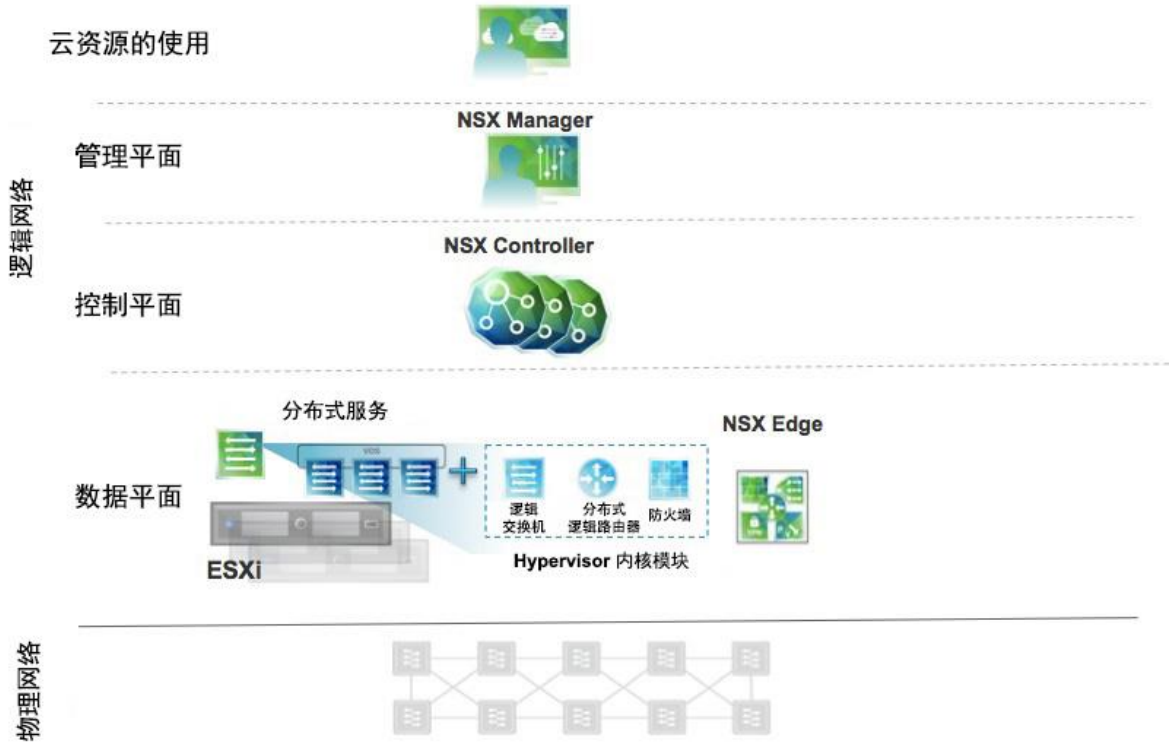



图 2 - NSX 组件

## 控制平面

NSX 控制平面在 NSX Controller 中运行。在采用 VDS 的 vSphere 优化环境中，Controller 支持无需多播 VXLAN 并可在控制平面对分布式逻辑路由 (DLR) 等元素进行编程。

在所有情况下，Controller 仅是控制平面的一部分，没有任何数据平面流量通过它。还会在具有奇数个成员的集群中部署 Controller 节点以实现高可用性和扩展。

## 数据平面

NSX 数据平面包含 NSX 虚拟交换机。NSX for vSphere 中的虚拟交换机以 vSphere Virtual Distributed Switch (VDS) 为基础，并添加了其他组件，因而可以提供丰富的服务。NSX 附加组件包括在 hypervisor 内核中运行的内核模块 (VIB)，可提供分布式路由和分布式防火墙等服务并支持 VXLAN 桥接功能。

NSX VDS 可以对物理网络进行抽象化处理并在 hypervisor 中提供访问级别的交换。NSX VDS 支持构建独立于物理结构（如 VLAN）的逻辑网络，因此是网络虚拟化的核心。NSX 虚拟交换机具有以下一些优势：

- 使用 VXLAN 协议和集中式网络配置为叠加网络连接提供支持。叠加的网络连接方式可实现下列功能：
  - 在现有物理基础架构的现有 IP 网络上创建一个灵活的第 2 层 (L2) 逻辑叠加网络，而无需重新构建任何数据中心网络。
  - 敏捷地调配通信流量（东西向和南北向），而租户之间仍保持相互隔离。
  - 应用工作负载和虚拟机根本察觉不到叠加网络的存在，就像连接到第 2 层物理网络一样运行。

- NSX 虚拟交换机可推动 hypervisor 的大规模扩展。
- 诸如端口镜像、NetFlow/IPFIX、配置备份和还原、网络运行状况检查、服务质量和 LACP 之类的多项功能可提供用于在虚拟网络中管理和监控流量以及排查故障的综合性工具包。

另外，数据平面还包括可提供从逻辑网络连接空间 (VXLAN) 到物理网络 (VLAN) 的通信的网关设备。可在第 2 层（NSX 桥接）或第 3 层（NSX 路由）使用此功能。

# 管理平面和使用平台

NSX 管理平面由 NSX Manager 构建。NSX Manager 在面向 NSX 的 vSphere 环境中提供单一配置点和 REST API 入口点。

可通过 NSX Manager UI 直接使用 NSX。在 vSphere 环境中，可通过 vSphere Web UI 本身使用 NSX。终端用户通常会将网络虚拟化与其云计算管理平台结合使用，以便部署应用。NSX 可以通过 REST API 在几乎任意 CMP 中提供一组丰富的集成功能。另外，用户目前还可以通过 VMware vCloud Automation Center (vCAC) 和 vCloud Director (vCD) 实现即时可用的集成。

# NSX 功能性服务

NSX 可在软件中如实重现网络和安全服务。


图 3 - NSX 逻辑网络服务

- **交换** – 支持在结构中的任意位置扩展第 2 层网段/IP 子网，不受物理网络设计的影响。
- **路由** – IP 子网之间的路由可以在逻辑空间内实现，数据不会向外流向物理路由器。此路由在 hypervisor 内核中执行，可最大限度减少 CPU/内存开销。此功能可以为虚拟基础架构中的路由流量（东西向通信）提供最佳数据路径。类似地，NSX Edge 可以提供一个与物理网络基础架构无缝集成的理想的集中位置，用于处理与外部网络之间的通信（南北向通信）。
- **分布式防火墙** – 安全性在内核和虚拟网卡级别实施。这样，系统便能够以高度可扩展的方式强制实施防火墙规则，且不会导致物理设备出现瓶颈。防火墙分布在内核中，因此可最大限度减少 CPU 开销，并且能够以线速运行。
- **逻辑负载均衡** – 借助 SSL 端接功能，为第 4 层至第 7 层负载均衡提供支持。
- **VPN**：通过 SSL VPN 服务启用第 2 层和第 3 层 VPN 服务。
- **连接至物理网络**：NSX-v 中支持第 2 层和第 3 层网关功能，可提供部署在逻辑和物理空间中的工作负载之间的通信。

在本文档随后两个章节中，我们将更加详细地介绍各种 NSX 组件之间的交互以及上面提到的每个逻辑服务。

# NSX-v 功能组件

NSX 平台创建的逻辑网络利用两种类型的访问层实体：Hypervisor 访问层和网关访问层。Hypervisor 访问层表示可连接到虚拟端点（虚拟机）的逻辑网络的连接点，而网关访问层为传统物理网络基础架构中的物理设备和端点提供到逻辑空间的第 2 层和第 3 层连接。

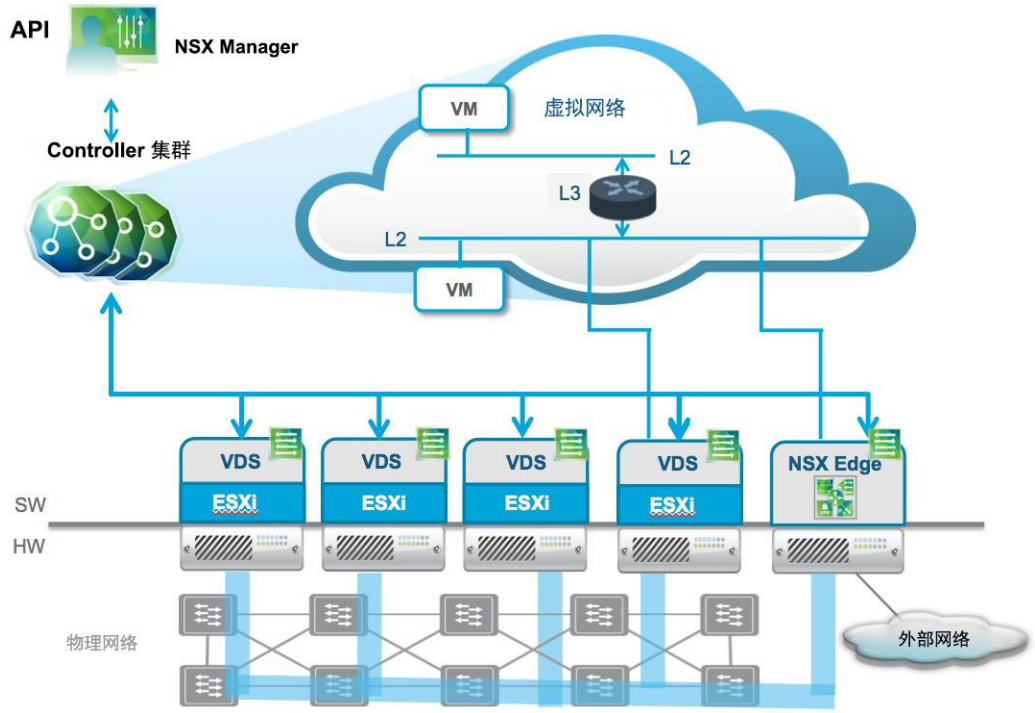



图 4 - NSX 功能组件

网络虚拟化的一个主要优势是能够将逻辑网络连接与底层物理网络结构分离开并提供逻辑网络抽象化和功能。如图 4 所示，有多个功能组件支持该功能并构成 NSX for vSphere 体系结构。正如在本文档其余章节中所讨论的，使用叠加协议 (VXLAN) 封装不同功能组件之间的虚拟机流量是逻辑/物理网络分离的关键功能。现在我们来更加详细地了解 NSX-v 平台的每个组件的功能。

## NSX Manager

NSX Manager 是管理平面虚拟设备，可帮助配置逻辑交换机并将虚拟机连接到这些逻辑交换机。它还提供了适用于 NSX 的管理 UI 和 API 入口点，可通过云计算管理平台帮助自动部署和管理逻辑网络。

在 NSX for vSphere 体系结构中，NSX Manager 与用于管理计算基础架构的 vCenter Server 紧密连接。事实上，NSX Manager 和 vCenter 之间是一对一关系，并且在安装后，NSX Manager 会向 vCenter 注册并向 vSphere Web Client 中插入插件以在 Web 管理平台中使用。

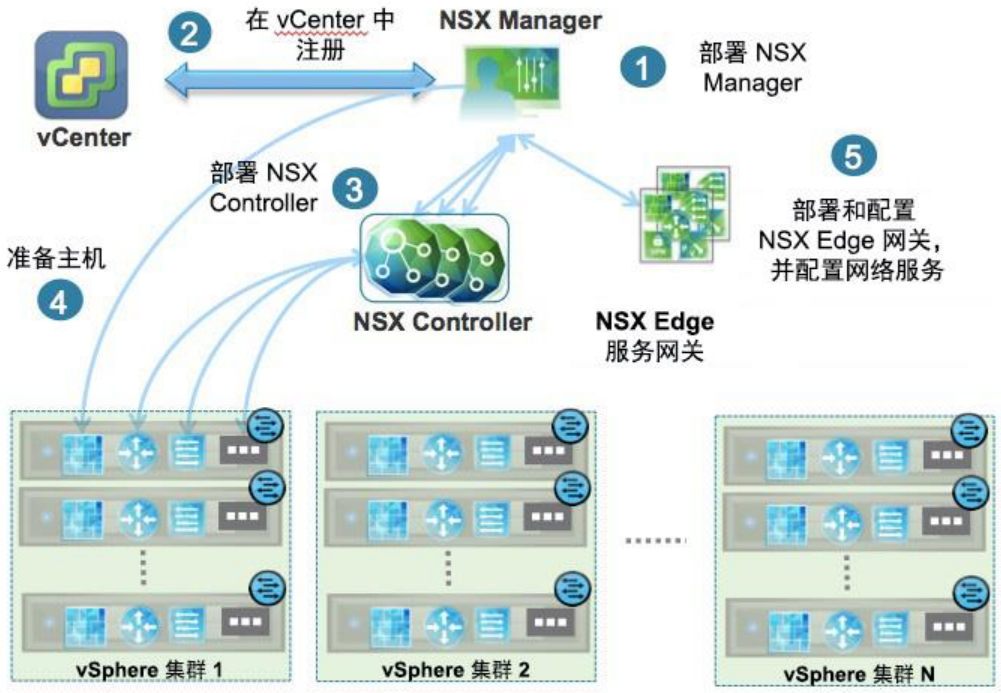



图 5 - vSphere Web Client 中的 NSX Manager 插件

正如图 5 中突出显示的，NSX Manager 负责部署控制器集群和准备 ESXi 主机，在主机上安装各种 vSphere 安装捆绑包 (VIB)，以便实现 VXLAN、分布式路由、分布式防火墙和用于在控制平面级别上通信的用户环境代理。NSX Manager 还负责部署和配置 NSX Edge 服务网关及相关网络服务（负载均衡、防火墙、NAT 等）。本白皮书后面的章节将更加详细地介绍这些功能。

NSX Manager 还为控制器集群的节点和应该允许加入 NSX 域的每个 ESXi 主机创建自签名证书，从而确保 NSX 体系结构的控制平面通信的安全性。NSX Manager 可以通过安全通道将这些证书安装到 ESXi 主机和 NSX 控制器上；之后，NSX 实体间的相互身份验证通过验证证书来进行。完成相互身份验证后，控制平面的通信就会被加密。

**注意**：在 NSX-v 软件版本 6.0 中，默认禁用 SSL。为了确保控制平面通信的保密性，建议通过 API 调用启用 SSL。从 6.1 版起，更改了默认值，启用了 SSL。

在恢复能力方面，由于 NSX Manager 是虚拟机，建议利用通常的 vSphere 功能作为 vSphere HA 以确保 NSX Manager 在所运行在的 ESXi 主机出现故障时可以动态移动。值得注意的是，此类故障情形只会临时影响 NSX 管理平面，而已部署的逻辑网络会继续无缝运行。

**注意**：NSX Manager 故障可能会影响特定功能（启用时），例如基于身份的防火墙（由于无法向 AD 组解析用户名）和流量监控收集。

最后，可通过从 NSX Manager GUI 执行按需备份并将其保存到必须可由 NSX Manager 访问的远程位置来随时备份 NSX Manager 数据，包括系统配置、事件和审核日志表（存储在内部数据库中）。还可以安排执行定期备份（每小时、每天或每周）。请注意，只能在可以访问以前备份的实例之一的新部署的 NSX Manager 设备上还原备份。

# Controller 集群

NSX 平台中的控制器集群是控制层组件，负责管理虚拟化管理程序中的交换和路由模块。控制器集群包含用于管理特定逻辑交换机的控制器节点。使用控制器集群来管理基于 VXLAN 的逻辑交换机，这样就不再要求物理网络基础架构支持多播。客户现在不必调配多播组 IP 地址，也无需启用物理交换机或路由器上的 PIM 路由或 IGMP 监听功能。

另外，NSX Controller 支持 ARP 抑制机制，减少连接虚拟机的第 2 层网络域中的 ARP 广播请求泛洪的需要。将在“逻辑交换”一节中更加详细地讨论不同的 VXLAN 复制模式和 ARP 抑制机制。

为了实现高恢复能力和高性能，生产部署必须在多个节点部署控制器集群。NSX Controller 集群表示横向扩展分布式系统，其中每个控制器节点分配有一组角色，这些角色定义了节点可以执行的任务类型。

为了提高 NSX 体系结构的可扩展性特征，系统会利用“切片”机制以确保所有控制器节点在任何给定时间都能保持活动状态。

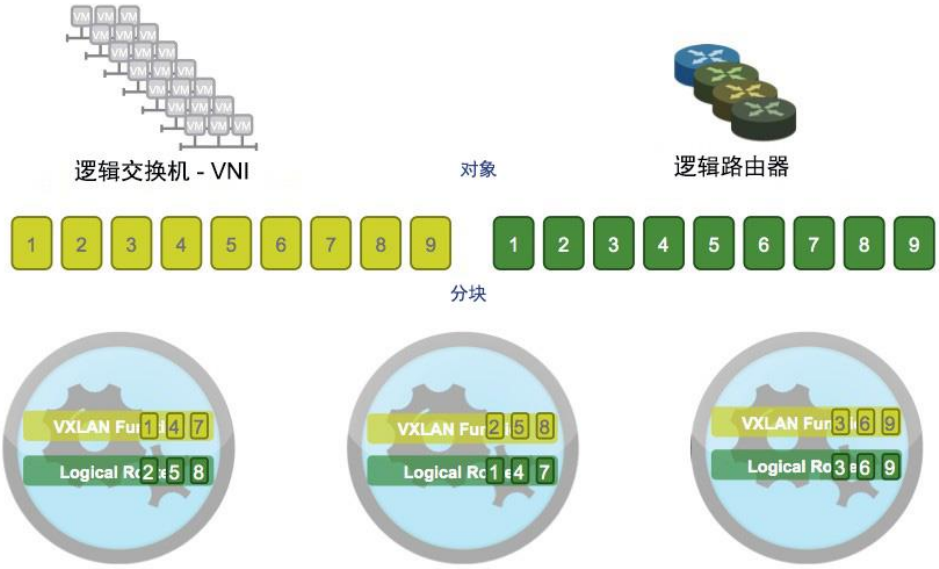



图 6 - 分割控制器集群节点角色

上面的图 6 说明了角色和责任如何完全分布在不同的集群节点之间。例如，这意味着不同的逻辑网络（或逻辑路由器）可能由不同的控制器节点管理：控制器集群中的每个节点由唯一的 IP 地址标识，并且当 ESXi 主机与集群的一个成员建立控制平面连接时，会将其他成员的 IP 地址完整列表传递给主机，以便能够与控制器集群的所有成员建立通信通道。这允许 ESXi 主机在任何给定时间知道哪个特定节点负责给定逻辑网络

当控制器节点发生故障时，发生故障的节点所负责的给定角色的细分块会重新分配给集群的剩余成员。为使此机制恢复能力强且具有确定性，选择控制器节点之一作为每个角色的“主节点”。主节点负责将细分块分配给各个控制器节点并确定节点何时发生故障，以便能够使用特定算法将细分块重新分配给其他节点。主节点还会通知 ESXi 主机集群节点发生故障，以便它们可以更新指定哪个节点负责各种逻辑网络细分块的内部信息。

为每个角色选择主节点需要集群中的所有活动和非活动节点进行过半数表决。这就是为什么总是必须利用奇数个节点部署控制器集群的主要原因。

| 集群中 NSX 节点的数量 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| 过半数 | 2 | 2 | 3 | 3 |
| 可能发生故障的节点的数量 | 0 | 1 | 1 | 2 |

图 7 - 控制器节点过半数

图 7 突出显示了不同的过半数情形，具体取决于控制器集群节点数。很容易看出为何部署 2 个节点（传统上视为冗余系统的示例）可增加控制器集群的可扩展性（因为在稳定状态，两个节点会并行工作），但不能提供任何其他恢复能力。这是因为使用 2 个节点，过半数为 2，这意味着如果两个节点中的一个节点发生故障，或它们之间无法通信（双活动情形），那么这两个节点将都无法保持运行（接受 API 调用等）。相同的注意事项也适用于具有 4 个节点的部署，这种部署无法提供比具有 3 个元素的集群更强的恢复能力（即使提供更好的性能）。

**注意**：NSX 目前（自软件版本 6.1 起）只支持具有 3 个节点的集群。上面提供的具有不同节点数量的各种示例只是为了说明过半数表决机制的工作原理。

从 NSX Manager UI 将 NSX Controller 节点作为虚拟设备进行部署。每个设备都有一个用于所有控制平面交互的 IP 地址以及目前无法修改的特定设置（4 个虚拟 CPU、4 GB 的 RAM）。

为确保控制器集群的可靠性，最好将集群节点的部署分散在不同的 ESXi 主机上，以确保单个主机的故障不会导致失去集群中的过半数节点。NSX 目前没有提供任何嵌入式功能来确保这一点，因此建议利用本机 vSphere 反关联性规则来避免在同一 ESXi 服务器上部署多个控制器节点。有关如何创建虚拟机间反关联性规则的更多信息，请参阅以下知识库文章：

http://pubs.vmware.com/vsphere-55/index.jsp#com.vmware.vsphere.resmgmt.doc/GUID-7297C302-378F-4AF2-9BD6-6EDB1E0A850A.html

# VXLAN 入门

叠加技术的部署能够将逻辑空间中的连接设置与物理网络基础架构分离开，因此得到了非常广泛的应用。连接到逻辑网络的设备可以利用前面在图 3 中突出显示的全部网络功能，不依赖底层物理基础架构的配置方式。物理网络有效地成为了用于传输叠加流量的底板。

此分离效果有助于解决传统数据中心部署现今面临的许多挑战，例如：

- 敏捷快速的应用部署：传统的网络连接设计成为一种瓶颈，降低了按企业所需的步伐推出新应用的速度。调配网络基础架构以支持新应用所需的时间通常按天计算（如果不是按周计算）。
- 工作负载移动化：利用计算虚拟化，可以跨连接到数据中心网络的不同物理服务器实现虚拟工作负载的移动化。在传统的数据中心设计中，这需要在整个数据中心网络基础架构中扩展第 2 层域 (VLAN)，从而影响总体可扩展性并可能影响设计的总体恢复能力。
- 大规模多租户：使用 VLAN 作为创建隔离网络的方法将可支持的租户数上限限制为 4094。此数字对通常的企业级部署来说可能似乎很大，但成为大多数云服务提供商的严重瓶颈。

虚拟可延展局域网 (VXLAN) 已成为名副其实的标准叠加技术，已被许多供应商采用（它由 VMware 与 Arista、Broadcom、Cisco、Citrix、Red Hat 及其他公司联合开发）。要能够构建在工作负载间提供第 2 层邻接的逻辑网络并且确保不出现传统第 2 层技术中的难题和可扩展性问题，部署 VXLAN 是一项关键工作。

如图 8 所示，VXLAN 是一种叠加技术，用于封装由连接到同一第 2 层逻辑网段（通常名为逻辑交换机 (LS)）的工作负载（虚拟或物理）生成的原始以太网帧。

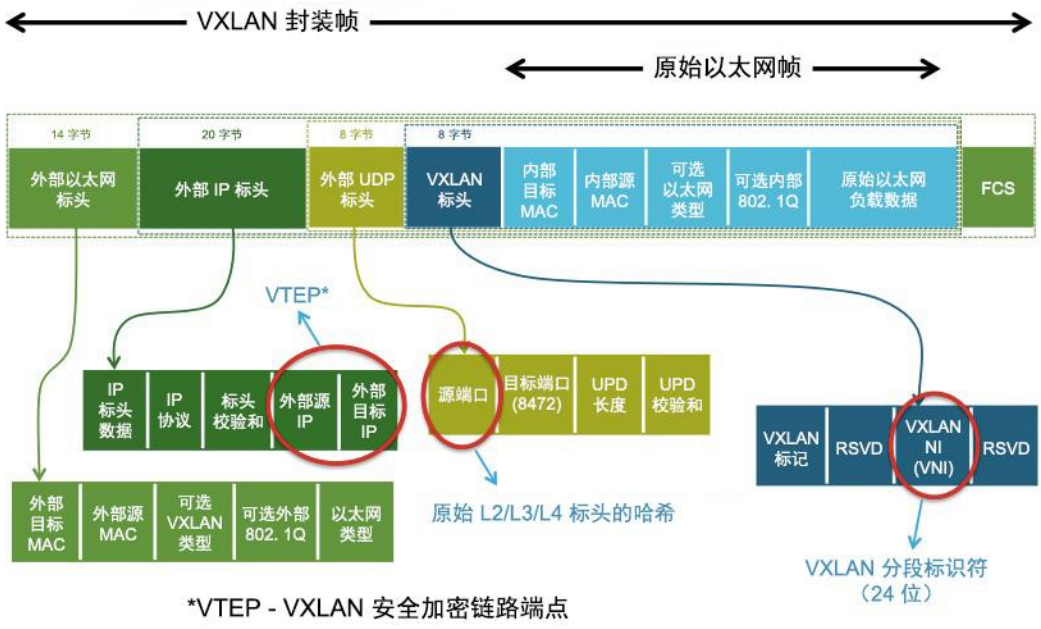



图 8 - VXLAN 封装

上图所示有关 VXLAN 数据包格式的几点注意事项：

- VXLAN 是一种封装技术，能够在第 3 层网络的基础上建立第 2 层网络 (L2oL3)：可通过外部 VXLAN、UDP、IP 和以太网标头封装工作负载生成的原始以太网帧，以确保其可跨与 VXLAN 端点（ESXi 主机）互连的网络基础架构传输。
- 扩展为超过传统交换机上 4094 个 VLAN 限制的问题已通过一个 24 位标识符得以解决，该标识符名为 VXLAN 网络标识符 (VNI)，与逻辑空间中创建的每个第 2 层网段关联。在 VXLAN 标头中传输此值，并且此值通常与 IP 子网关联，类似于传统上 VLAN 的情况。在连接到同一虚拟网络（逻辑交换机）的设备之间进行 IP 子网内通信。

  **注意**：术语“VXLAN 网段”、“虚拟网络”(VN) 和“逻辑交换机”均指在逻辑网络空间中创建的第 2 层逻辑域，并在本文档中交替使用。

- 在此，我们执行原始以太网帧中存在的第 2 层/第 3 层/第 4 层标头的哈希运算以得出外部 UDP 标头的源端口值。这对确保跨传输网络基础架构中潜在可用的等价路径的 VXLAN 流量负载均衡非常重要。
- 到当前版本 6.1，NSX 使用 8472 作为外部 UDP 标头的目标端口值。这不同于 IANA 为 VXLAN 分配的数字 4789，如下面的链接中所述：

  http://tools.ietf.org/html/rfc7348#page-19

  不过，可以通过 REST API 调用在 NSX 部署中修改此值。不过，在执行此操作之前，为了避免数据平面出现故障，确保所有 ESXi 主机都运行 NSX-v 版本（不与运行 vCNS 和 vSphere 5.1 的主机向后兼容）且物理网络中的配置（访问列表、防火墙策略等）已正确更新非常重要。
- 外部 IP 标头中使用的源和目标 IP 地址唯一标识发起和终止 VXLAN 帧封装的 ESXi 主机。这些通常称为 VXLAN 安全加密链路端点 (VTEP)。
- 将原始以太网帧封装到 UDP 数据包中显然会增加 IP 数据包的大小。因此，建议将物理基础架构中将传输帧的所有接口的总体 MTU 大小增加至最小 1600 字节。在（从 NSX Manager UI）为 VXLAN 准备主机时，会自动增加执行 VXLAN 封装的 ESXi 主机的 VDS 上行链路的 MTU，本白皮书的后面将对此进行详细讨论。

下面的图 9 概述了利用 VXLAN 叠加功能在连接到不同 ESXi 主机的虚拟机之间建立第 2 层通信所需的步骤。

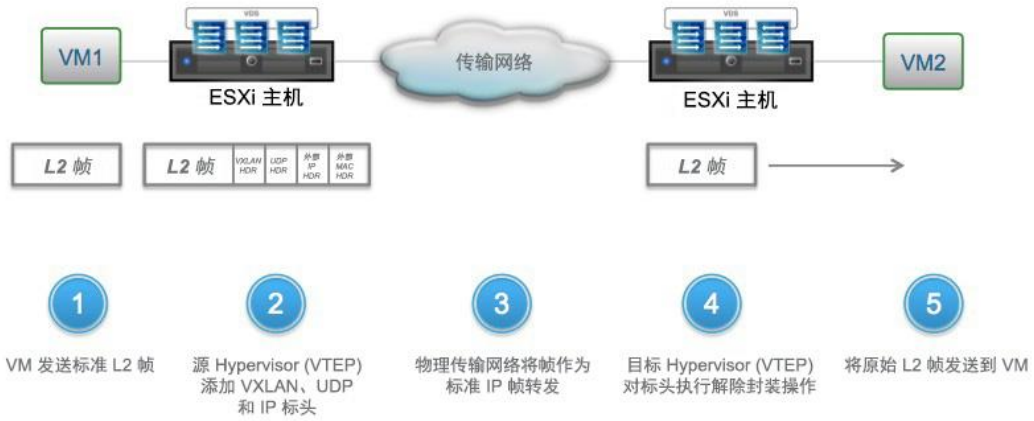



**图 9 - 利用 VXLAN 封装建立第 2 层通信**

1. 虚拟机 1 发出一个帧，去往同一第 2 层逻辑网段（IP 子网）的虚拟机 2 部分。
2. 源 ESXi 主机会确定虚拟机 2 连接到的 ESXi 主机 (VTEP) 并在将此帧发送到传输网络之前对它进行封装。
3. 仅当在源和目标 VTEP 之间启用 IP 通信时才需要传输网络。
4. 目标 ESXi 主机会收到 VXLAN 帧，对帧解除封装并确定它所属的第 2 层网段（利用源 ESXi 主机插入 VXLAN 标头的 VNI 值）。
5. 该帧传输到虚拟机 2。

将在“逻辑交换”一节中提供有关利用 VXLAN 时如何实施第 2 层通信的更加详细的信息，以及有关 NSX-v 提供的特定 VXLAN 控制和数据平面增强功能的讨论。

# ESXi Hypervisor 与 VDS

正如在本白皮书的简介章节中已提到的，Virtual Distributed Switch (VDS) 是总体 NSX-v 体系结构的构造块。VDS 可在所有 VMware ESXi hypervisor 上使用，因此现在阐明它们在 NSX 体系结构中的控制和数据平面交互非常重要。

**注意**：有关 VDS 虚拟交换机和它在 vSphere 环境中的部署最佳实践的更多信息，请参阅以下白皮书：
http://www.vmware.com/files/pdf/techpaper/vsphere-distributed-switch-best-practices.pdf

## 用户空间和内核空间

如图 10 所示，每个 ESXi 主机均具有用户空间和内核空间。

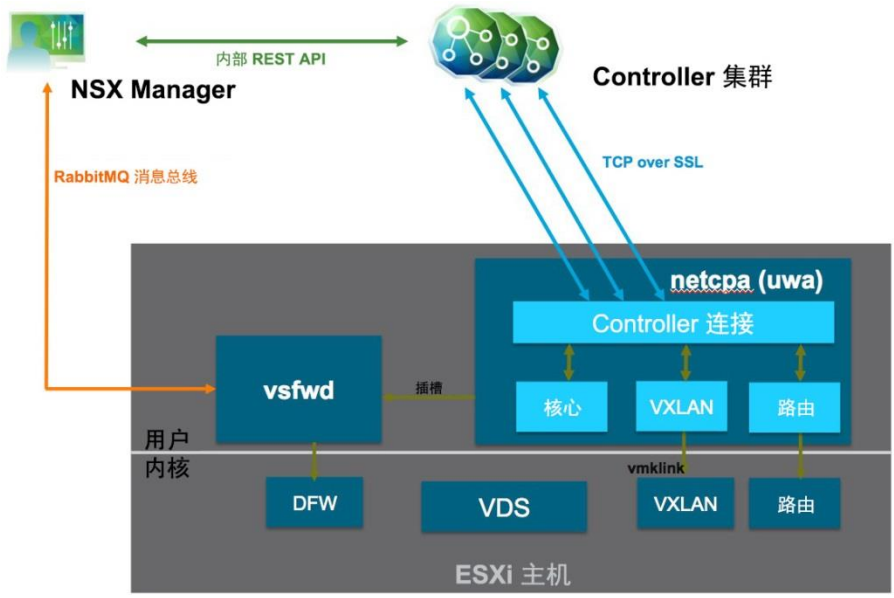



图 10 - ESXi 主机用户空间和内核空间图表

VMware 安装捆绑包 (VIB) 是安装在 ESXi hypervisor 的内核空间中以启用 NSX-v 体系结构中请求的功能的文件集：分布式交换和路由、分布式防火墙保护和 VXLAN 封装/解除封装。

用户空间中是向 NSX Manager 和控制器集群节点提供控制平面通信路径的软件组件：

- 对 vsfwd（RMQ 客户端）和托管在 NSX Manager 上的 RMQ 服务器进程之间的通信使用 RabbitMQ 消息总线。此消息总线由 NSX Manager 用于将各种信息发送给 ESXi 主机：需要在内核中的分布式防火墙上编程的策略规则、控制器节点 IP 地址、用于对主机和控制器之间的通信进行身份验证的私钥和主机证书以及对创建/删除分布式逻辑路由器实例的请求。
- 用户环境代理进程 (netcpa) 会建立到控制器集群节点的 TCP over SSL 通信通道。通过将此控制平面通道与 ESXi hypervisor 结合使用，控制器节点会填充本地表（MAC 地址表、ARP 表和 VTEP 表）以跟踪已部署逻辑网络中连接的工作负载。

# NSX Edge 服务网关

NSX Edge 就像“瑞士军刀”，因为它可以向 NSX-v 体系结构提供多项服务。

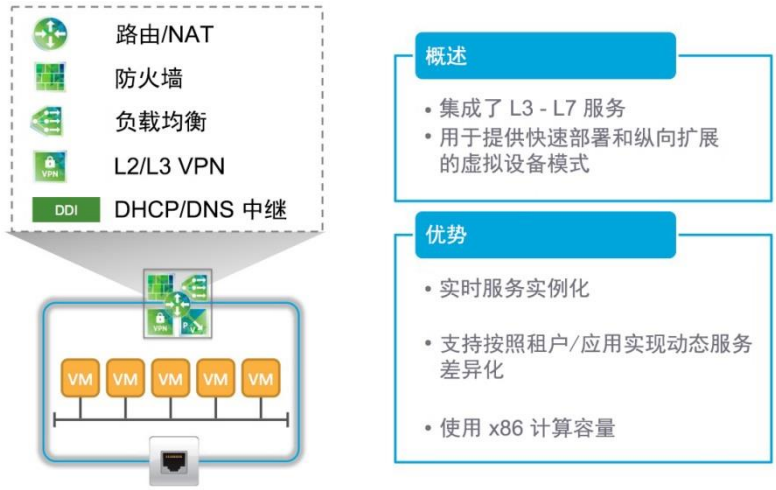



图 11 - NSX Edge 提供的服务

图 11 突出显示了 NSX Edge 提供的逻辑服务：

- 路由和网络地址转换 (NAT)：NSX Edge 可以在 NSX 域和外部物理网络基础架构中部署的逻辑网络之间提供集中式入口路由和出口路由。NSX Edge 支持各种路由协议（OSPF、iBGP 和 eBGP），并且还可以利用静态路由进行通信。可以对流经 Edge 的流量执行 NAT，并且源 NAT 和目标 NAT 均受支持。
- 防火墙：NSX Edge 支持有状态防火墙保护功能，可以对 ESXi 主机的内核中启用的分布式防火墙 (DFW) 进行补充。DFW 主要用于对连接到逻辑网络的工作负载间的通信（所谓的东西向通信）强制实施安全策略，而 NSX Edge 上的防火墙通常用于过滤逻辑空间和外部物理网络之间的通信（南北向流量）。
- 负载均衡：NSX Edge 可以为逻辑空间内部署的工作负载服务器群执行负载均衡服务。Edge 原生支持的负载均衡功能可满足真实部署中的大多数常见要求。
- 第 2 层和第 3 层虚拟专用网络 (VPN)：NSX Edge 还可以在第 2 层和第 3 层上提供 VPN 功能。第 2 层 VPN 通常用于扩展分散在不同地理位置的独立数据中心站点之间的第 2 层域，以支持计算资源需求激增或混合云部署等使用情形。部署第 3 层 VPN 可允许在两个 NSX Edge 或其他供应商 VPN 终端之间进行 IPSec 站点间连接，或允许远程用户利用 SSL VPN 连接来访问 NSX Edge 后面的专用网络。
- DHCP、DNS 和 IP 地址管理 (DDI)：最后，NSX Edge 还可以支持 DNS 中继功能，且可充当 DHCP 服务器，能够为连接到逻辑网络的工作负载提供 IP 地址、默认网关、DNS 服务器和搜索域信息。

  **注意**：NSX 版本 6.1 引入了对 NSX Edge 上的 DHCP 中继的支持，以便可以使用集中部署和远程部署的 DHCP 服务器为连接到逻辑网络的工作负载提供 IP 地址。

从部署角度看，NSX Manager 提供了根据需要启用的功能使用的不同规格，如图 12 所示。

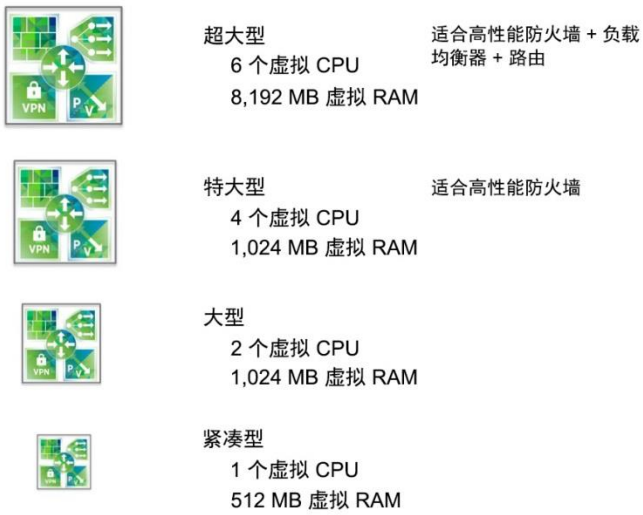



**图 12 - NSX Edge 规格**

请注意，可以从 NSX Manager（通过 UI 或 API 调用）更改规格，即使在初始部署之后也可以。不过，在该情形下，预期会出现短暂的服务中断。

最后，在考虑恢复能力和高可用性时，NSX Edge 服务网关可部署为一对活动/备用单元。

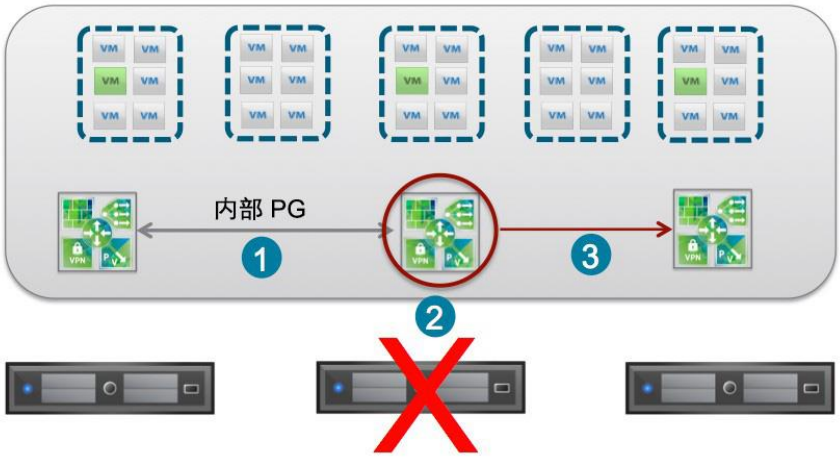



图 13 - NSX Edge 活动/备用部署

NSX Edge HA 基于图 13 中突出显示的行为：

1. NSX Manager 在不同的主机上部署一对 NSX Edge（反关联性规则）。每秒在活动和备用 Edge 实例之间交换一次检测信号保活以监控相互的运行状况。保活是通过内部端口组发送的第 2 层探测（利用 NSX Edge 上部署的第一个虚拟网卡），除非用户明确选择使用不同的接口。请注意，VXLAN 第 2 层网段可用于交换保活，以便可以通过路由的物理网络基础架构进行交换。
2. 如果托管活动 NSX Edge 的 ESXi 服务器出现故障，则在“Declare Dead Time”（声明失效时间）计时器过期后，备用服务器会接管活动职责。此计时器的默认值为 15 秒，但可调短（通过 UI 或 API 调用）为 6 秒。
3. NSX Manager 还会监控已部署的 NSX Edge 的运行状况，以确保在另一个 ESXi 主机上重新启动出现故障的 NSX Edge。

**注意**：在活动和备用 NSX Edge 之间通过内部端口组交换的消息还用于同步逻辑服务（路由、NAT、防火墙保护等）所需的状态信息，以这种方式提供有状态的服务。

可以在“NSX-v 设计注意事项”章节中找到 NSX Edge 的更多部署注意事项。

# 传输域

从最简单的意义上讲，传输域界定了一组可以跨物理网络基础架构进行互相通信的 ESXi 主机。正如前面提到的，利用每个 ESXi 主机和指定的 VXLAN 安全加密链路端点 (VTEP) 上定义的一个（或多个）特定接口进行此通信。

传输域可以跨一个或多个 ESXi 集群进行扩展，并且可以不严谨地定义逻辑交换机的跨度。要更好地理解这点，阐明逻辑交换机、VDS 和传输域之间存在的关系非常重要。

VDS 可以跨越特定数量的 ESXi 主机，因为可以从特定 VDS 添加/删除单个 ESXi 主机。在实际的 NSX 部署中，很可能在给定的 NSX 域中定义多个 VDS。在图 14 显示的场景中，一个“计算-VDS”跨越属于计算集群的一部分的所有 ESXi 主机，一个单独的“Edge-VDS”跨越 Edge 集群中的 ESXi 主机进行扩展。

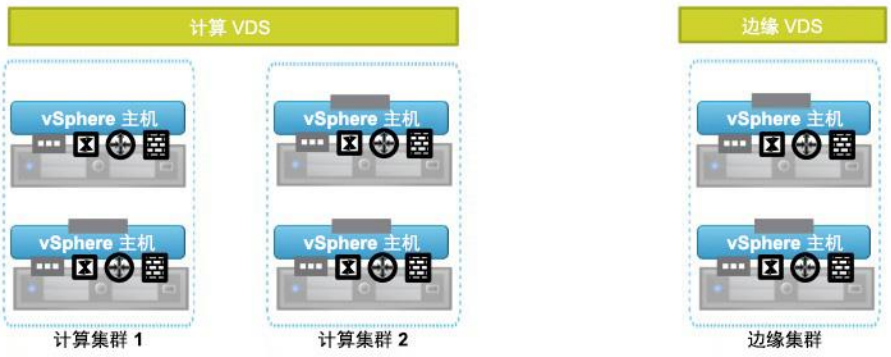



图 14 - 跨越 ESXi 集群的单独 VDS

逻辑交换机可跨多个 VDS 进行扩展。参考前面的示例，给定逻辑交换机可以为连接到计算集群或 Edge 集群的虚拟机提供连接能力。逻辑交换机始终作为特定传输域的一部分来创建。这表明在通常情况下，传输域可以跨越所有 ESXi 集群进行扩展并能定义逻辑交换机的跨度，如图 15 中所突出显示的。

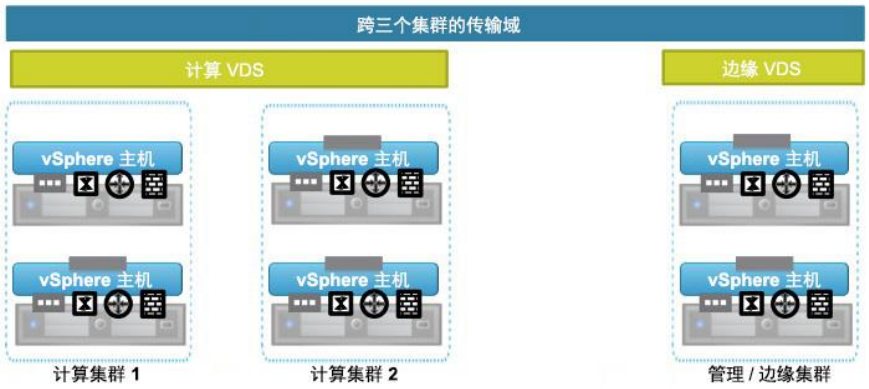



**图 15 - 跨越 VDS 和 ESXi 集群的传输域**

**注意**：可以在“NSX-v 域中的 VDS 设计”一节中找到有关 NSX-v 域中的 VDS 设计的更多注意事项和建议。

# NSX 分布式防火墙 (DFW)

NSX DFW 为 NSX 环境中的任何工作负载提供第 2 层至第 4 层有状态防火墙服务。DFW 运行在内核空间中并因此执行近线速网络流量保护。可通过添加新的 ESXi 主机来线性扩展 DFW 性能和吞吐量，最高达到 NSX 支持的最大限制。

在主机准备过程完成后立即激活 DFW。如果某个虚拟机根本不需要 DFW 服务，可以将该虚拟机添加到功能排除列表中（默认情况下，会自动从 DFW 功能中排除 NSX Manager、NSX Controller 和 Edge 服务网关）。

为每个虚拟机虚拟网卡创建一个 DFW 实例（例如，如果创建了具有 3 个虚拟网卡的新虚拟机，则会向此虚拟机分配 3 个 DFW 实例）。根据“apply to”（应用于）应用，这些 DFW 实例的配置可以相同或完全不同：在创建 DFW 规则后，用户可以为此规则选择实施点 (PEP)：逻辑交换机和虚拟网卡的选项不同。默认情况下，没有选择“apply to”（应用于），这意味着 DFW 规则会向下传播到所有 DFW 实例。

可以 2 种方式编写 DFW 策略规则：使用第 2 层规则（以太网）或第 3 层/第 4 层规则（常规）。

- 第 2 层规则映射到第 2 层 OSI 模型：只能在源和目标字段中使用 MAC 地址，并且只能在服务字段中使用第 2 层协议（例如 ARP）。
- 第 3 层/第 4 层规则映射到第 3 层/第 4 层 OSI 模型：可以使用 IP 地址和 TCP/UDP 端口编写策略规则。请务必记住始终在实施第 3 层/第 4 层规则之前实施第 2 层规则。举一个具体的例子，如果第 2 层默认策略规则修改为“block”（阻止），则 DFW 也会阻止所有第 3 层/第 4 层流量（例如，ping 会停止工作）。

DFW 是旨在保护工作负载间网络流量（虚拟到虚拟或虚拟到物理）的 NSX 组件。换句话说，DFW 主要目标是保护东西向流量；这就是说，由于 DFW 策略实施应用于虚拟机的虚拟网卡，因此它还可用于阻止虚拟机和外部物理网络基础架构之间的通信。DFW 通过可提供集中式防火墙功能的 NSX Edge 服务网关进行了充分补充。ESG 通常用于保护南北向流量（虚拟到物理），因此是软件定义的数据中心的第一个入口点。

下图描述了 DFW 和 ESG 的不同定位角色：

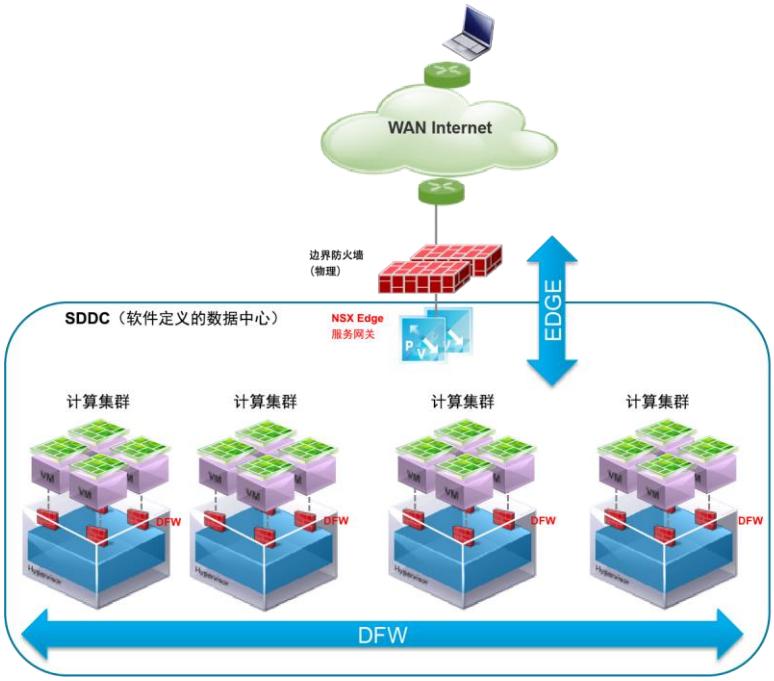



图 16 - DFW 和 Edge 服务网关流量保护

NSX DFW 在虚拟机虚拟网卡级别上运行，这意味着虚拟机始终受到保护，无论采用什么方式连接到逻辑网络：虚拟机可连接到支持 VLAN 的 VDS 端口组或逻辑交换机（支持 VXLAN 的端口组）。所有这些连接模式都受到全面支持。ESG 防火墙还可用于保护物理服务器和设备（例如 NAS）上的工作负载。

DFW 系统体系结构基于 3 个不同的实体，其中每个实体扮演明确定义的角色：

- **vCenter Server** 是解决方案的管理平面。通过 vSphere Web Client 创建 DFW 策略规则。策略规则的源/目标字段中可以使用任何 vCenter 容器：集群、VDS 端口组、逻辑交换机、虚拟机、虚拟网卡、资源池等。
- **NSX Manager** 是解决方案的控制平面。它从 vCenter Server 接收规则并将其存储在中央数据库中。NSX Manager 然后将 DFW 规则推送到已做好实施准备的所有 ESXi 主机。每次修改并发布 DFW 安全策略规则表后即会对该表进行备份。如果使用 CMP 实现安全性自动化，NSX Manager 还可以直接从 REST API 调用接收 DFW 规则。
- **ESXi 主机**是解决方案的数据平面。从 NSX Manager 接收 DFW 规则，然后转换为内核空间以实时执行。检查虚拟机网络流量并在每个 ESXi 主机上实施。我们选择此情形：虚拟机 1 位于 ESXi 主机 1 上并将数据包发送给位于 ESXi 主机 2 上的虚拟机 2。在 ESXi 主机 1 上为传出流量完成策略实施（在数据包离开虚拟机 1 时），然后在 ESXi 主机 2 上为传入流量完成策略实施（在数据包到达虚拟机 2 时）。

  **注意：**图 17 中描述的 SSH 客户端可以访问 ESXi 主机 CLI 以执行与 DFW 相关的特定故障排除。

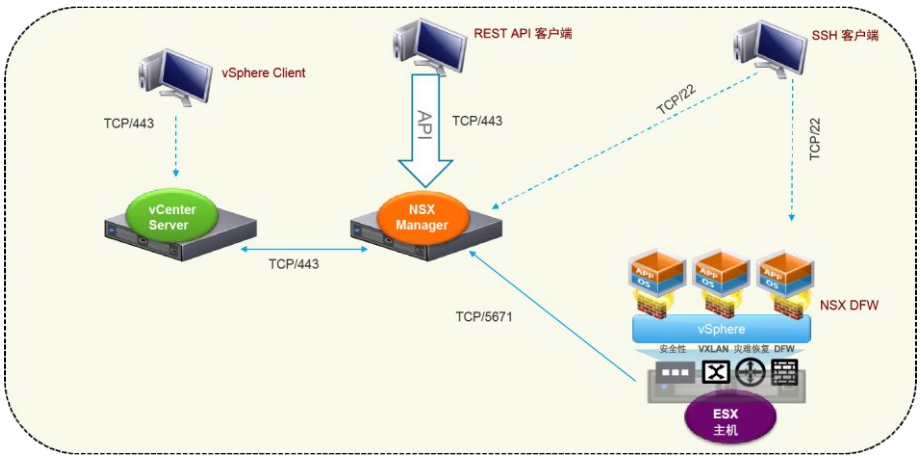



图 17 - NSX DFW 系统体系结构和通信通道

请务必记住，在 DFW 策略规则中使用 vCenter 容器时，VMtools 必须安装在客户虚拟机上。VMtools 知道虚拟机的 IP 地址（来自 DHCP 的动态 IP 地址或管理员在虚拟机上手动配置的静态 IP 地址），然后将此信息提供给基于 MAC、IP 和 TCP/UDP 端口字段运行的 DFW 引擎。

**注意**：如果 DFW 策略规则只使用主机 IP 或子网 IP 或 IP 集之类的 IP 信息，则明显不再需要客户虚拟机上存在 VMtools。

正如前面提到的，在主机准备好后激活 DFW 功能。在此操作期间，将内核 VIB 加载到 hypervisor 中：VMware 互连服务插入平台，即 **VSIP**。

VSIP 负责所有数据平面流量保护（它本身是 DFW 引擎）并以近线速运行。为每个虚拟机虚拟网卡创建 DFW 实例，并且此实例位于虚拟机和虚拟交换机（支持 VLAN 的 VDS 端口组或逻辑交换机）之间：为 DFW 实例分配 DVfilter 插槽 2。到达/来自此虚拟机的所有传入和传出数据包必须通过 DFW，绝对没有任何方法可绕过它。

名为 vsfwd 的一组守护程序永久地运行在 ESXi 主机上并执行多项任务：

- 与 NSX Manager 交互以检索 DFW 策略规则。
- 收集 DFW 统计信息并将其发送给 NSX Manager。
- 将审核日志信息发送给 NSX Manager。

vCenter Server 和 ESXi 主机之间的通信路径（使用 ESXi 主机上的 vpxa 进程）只用于与 vSphere 相关的目的（例如创建虚拟机或修改存储）以及使用 NSX Manager 的 IP 地址对主机进行编程。此通信不用于任何 DFW 操作。

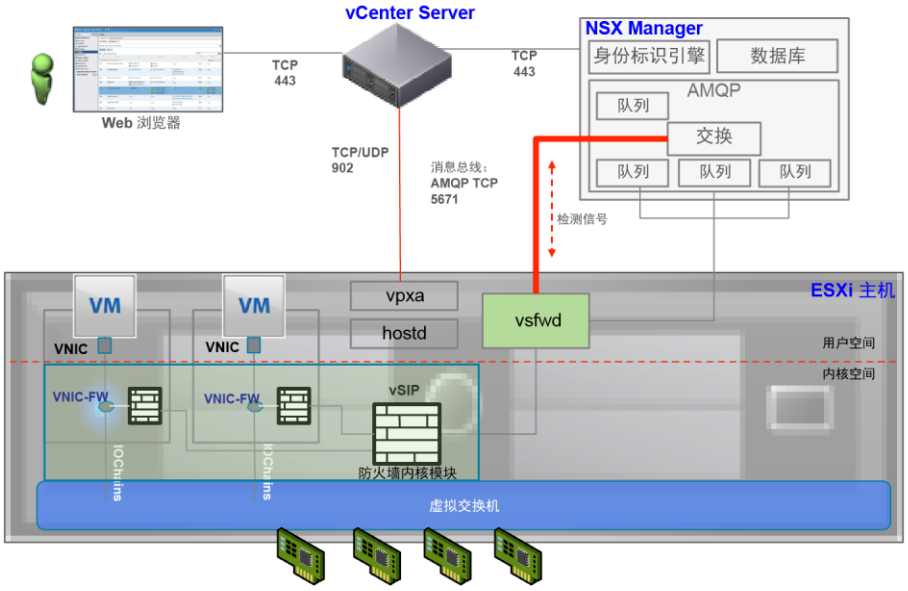



图 18 - NSX DFW 组件详细信息

VSIP 内核模块事实上不仅仅提供了 DFW 功能。此服务插入平台添加了补充性服务，例如，Spoofguard 以及将流量重定向到第三方合作伙伴服务（Palo Alto Networks 或 Trend 或 Symantec 等）。

Spoofguard 通过维护包含虚拟机名称和 IP 地址（在虚拟机首次启动时从 VMtools 检索此信息）的参考表来防御 IP 欺骗。Spoofguard 默认处于非活动状态，并且必须明确为每个逻辑交换机或 VDS 端口组启用才能运行。在检查到虚拟机 IP 地址发生更改后，DFW 会阻止来自/到达此虚拟机的流量，直到 NSX 管理员批准此新 IP 地址。

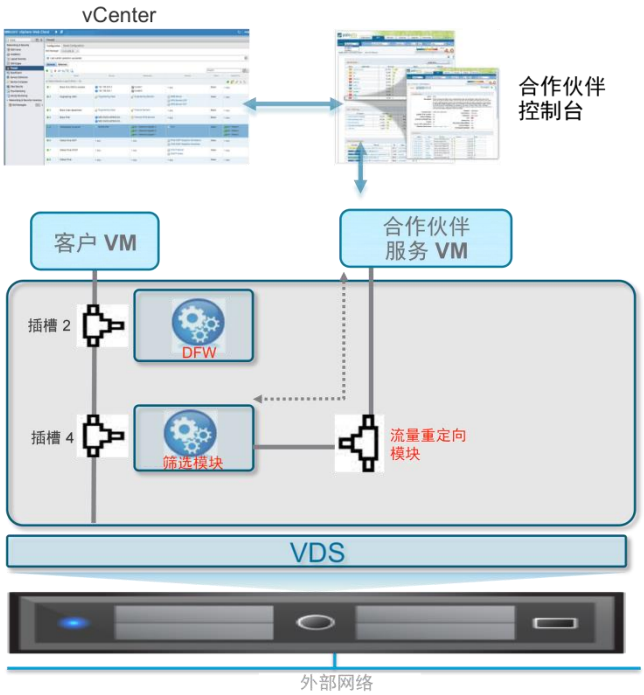



**图 19 - 将流量重定向到第三方合作伙伴服务**

将流量重定向到第三方供应商可以引导特定类型的流量到达所选合作伙伴服务虚拟机。例如，从 Internet 到一组 APACHE 或 TOMCAT 服务器的 Web 流量可重定向到第 4 层至第 7 层深层数据包检测防火墙以进行高级保护。

在“Service Composer”/“Security Policy”（安全策略）下（NSX 版本 6.0）或在“DFW”菜单的“Partner Security Services”（合作伙伴安全服务）选项卡下（NSX 版本 6.1）定义流量重定向。

“Partner Security Services”（合作伙伴安全服务）选项卡提供了功能强大且易于使用的用户界面，可定义需要将什么类型的流量重定向到哪些合作伙伴服务。它遵循与 DFW 相同的策略定义结构（即，源字段、目标字段和服务字段的选项相同），唯一的区别在操作字段中：不是“Block”（阻止）/“Allow”（允许）/“Reject”（拒绝），用户可以在重定向/不重定向之间选择，然后选择合作伙伴列表（已向 NSX 注册且已成功部署在平台上的任何合作伙伴）。最后，可以为此流量重定向规则启用日志选项。

ESXi 主机上的 DFW 实例（每个虚拟机虚拟网卡 1 个实例）包含 2 个单独表：

- **规则表**：用于存储所有策略规则。
- **连接跟踪器表**：用于缓存允许的操作所对应的规则的流条目。

  **注意**：特定流是通过 5 元组信息“源 IP 地址/目标 IP 地址/协议/第 4 层源端口/第 4 层目标端口”确定的。请注意，默认情况下，DFW 不在第 4 层源端口上执行查找，但可以配置为通过定义特定策略规则来执行此操作。

在了解这 2 个表的使用情形之前，我们首先了解如何实施 DFW 规则：

1. DFW 规则是按照从上到下的顺序实施的。
2. 按从上到下的顺序逐个根据规则表中的规则检查每个数据包。
3. 将实施表中第一条符合流量参数的规则

鉴于此行为，编写 DFW 规则时，我们建议始终将最细化的策略置于规则表的最上方。这是确保在实施任何其他规则之前实施它们的最佳方法。

DFW 默认策略规则（位于规则表底部的规则）是“全部接收”规则：默认规则将实施不符合默认规则之上的任何规则的数据包。在主机准备操作完成后，DFW 默认规则设置为“allow”（允许）操作。主要是因为 VMware 不希望在生产前调试或迁移阶段中中断任何虚拟机间通信。不过，最佳实践是将默认规则更改为“block”（阻止）操作并通过积极控制模式实施访问控制（仅允许防火墙策略中定义的流量进入网络中）。

我们现在来看一下策略规则查找和数据包流：

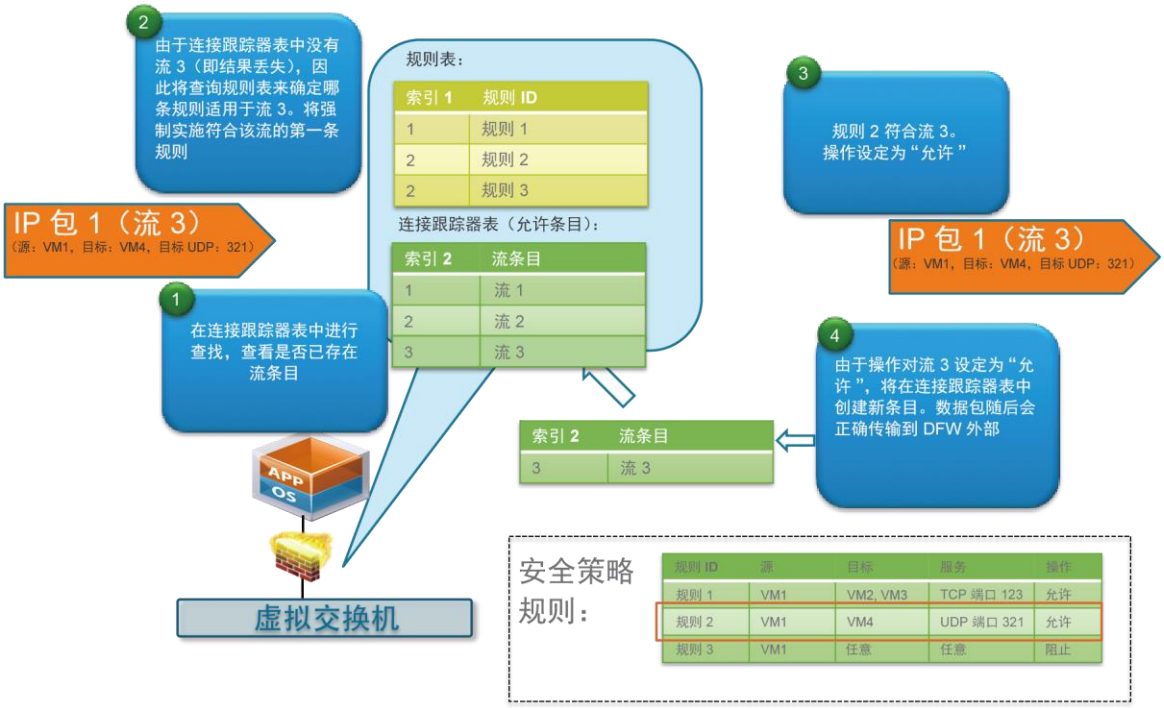



图 20 - DFW 策略规则查找和数据包 - 第一个数据包

虚拟机发送符合第 2 条规则的 IP 数据包（第一个数据包 - pkt1）。操作顺序如下：

1. 在连接跟踪器表中进行查找，以查看是否已存在相应流的条目。
2. 由于连接跟踪器表中没有流 3（即结果丢失），因此将在规则表中进行查找来确定哪条规则适用于流 3。将实施适用于该流的第一条规则。
3. 规则 2 符合流 3。将操作设置为“Allow”（允许）。
4. 由于针对流 3 将操作设置为“Allow”（允许），因此将在连接跟踪器表中创建一个新条目。数据包随后会正确传输到 DFW 外部。

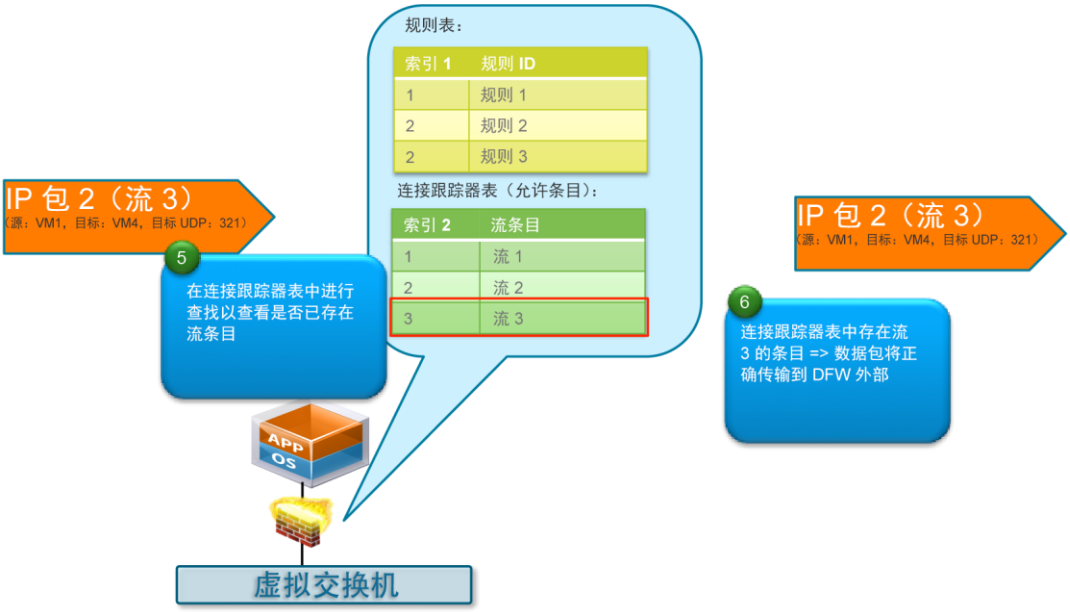



图 21 - DFW 策略规则查找和数据包 - 后续数据包。

按此顺序处理后续数据包：

1. 在连接跟踪器表中进行查找，以查看是否已存在相应流的条目。

2. 连接跟踪器表中存在流 3 的条目 => 数据包正确传输到 DFW 外部 要强调的一个重要方面是 DFW 完全支持 vMotion（采用 DRS 的自动 vMotion 或手动 vMotion）。在 vMotion 操作过程中，规则表和连接跟踪器表始终随虚拟机而发生变化。积极效果是在工作负载移动期间没有中断流量，并且在 vMotion 之前启动的连接在 vMotion 完成之后保持完好。DFW 使虚拟机可以自由移动，同时确保持续的网络流量保护。

**注意**：此功能不依赖于控制器，也无需具有可供使用的 NSX Manager。

NSX DFW 引入了以前不可能实现的模式转变：安全服务不再依赖网络拓扑。利用 DFW，安全性与逻辑网络拓扑完全分离开。

在传统环境中，要为一台服务器或一组服务器提供安全服务，必须使用 VLAN 梳理方法或第 3 层路由操作将来自/到达这些服务器的流量重定向到防火墙：流量必须通过此专用防火墙才能保护网络流量。

利用 NSX DFW，无需再这样做，因为直接将防火墙功能引入了虚拟机中。DFW 将系统地处理此虚拟机发送或接收的所有流量。因此，如果虚拟机位于同一逻辑交换机（或支持 VLAN 的 VDS 端口组）上或不同的逻辑交换机上，则可以实施虚拟机间（工作负载间）的流量保护。

这一主要的概念是将在本白皮书的下一章中介绍的**微分段**使用情形的基础。

# NSX-v 功能性服务

在 NSX 体系结构环境中，逻辑网络连接功能（交换、路由、安全性等）可被视为具有数据包转发管道（该管道可实施多项功能，包括第 2 层/第 3 层转发、安全筛选和数据包排队）。

不同于物理交换机或路由器，NSX-v 不依赖单个传输设备即可实施整个转发管道。NSX Controller 会在所有相关 ESXi 主机上创建数据包处理规则来模拟单个逻辑设备处理数据的方式。因此，这些 hypervisor 可被视为实施逻辑网络功能管道的一部分的“线路卡”。连接 ESXi 主机的物理网络充当将数据包从一个“线路卡”传送到另一个“线路卡”的“底板”。

## 多层应用部署示例

如何部署上一章介绍的 NSX-v 组件对敏捷灵活地创建应用及应用需要的网络连接和服务极为重要。创建图 22 中突出显示的多层应用就是一个典型示例。

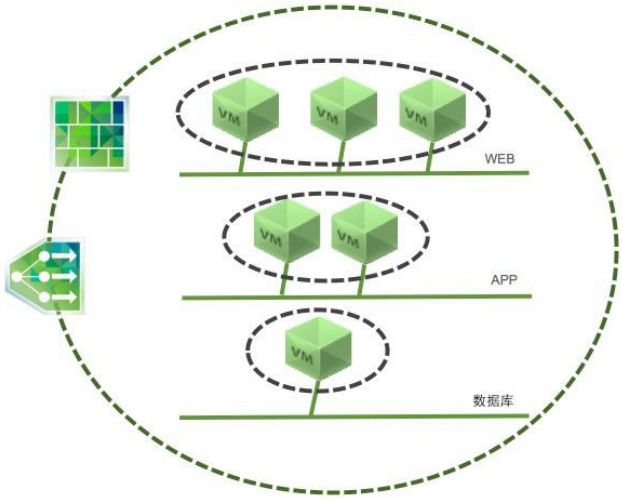



**图 22 - 多层应用的部署**

以下各小节将讨论动态创建多层应用所需的、通过部署 NSX-v 组件启用的功能（逻辑交换、路由和其他服务）。

# 逻辑交换

利用 NSX-v 平台中的逻辑交换功能，客户可以像搭建虚拟机一样灵活、敏捷地搭建隔离的第 2 层逻辑网络。然后，虚拟和物理端点都可以连接到这些逻辑网段并建立连接，而与它们在数据中心网络中的特定部署位置不相关。这是由于网络基础架构与 NSX 网络虚拟化所提供的逻辑网络（即，底层和叠加网络）之间实现了分离。

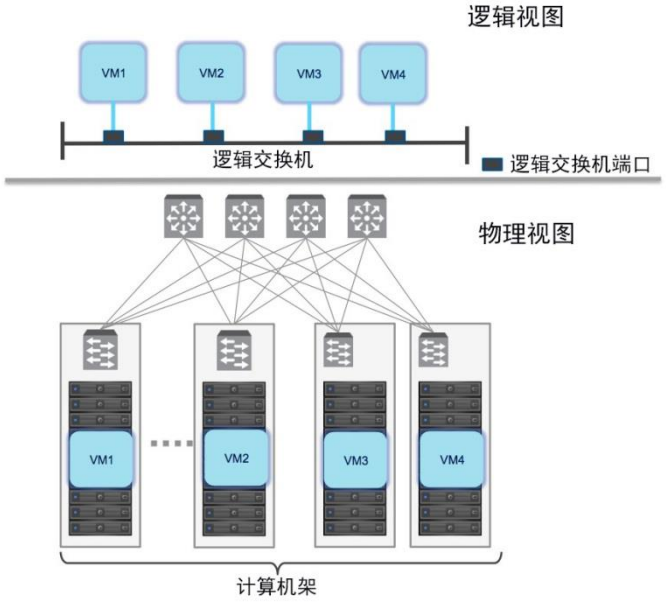



**图 23 - 逻辑交换（逻辑和物理网络视图）**

图 23 显示了利用 VXLAN 叠加技术部署逻辑交换时的逻辑和物理网络视图，该叠加技术允许跨多个服务器机架延伸第 2 层域（逻辑交换机），且独立于底层机架间连接（第 2 层或第 3 层）。

参考前面讨论的多层应用的部署示例，逻辑交换功能允许创建映射到连接各种工作负载（虚拟机或物理主机）的不同层的不同的第 2 层网段。

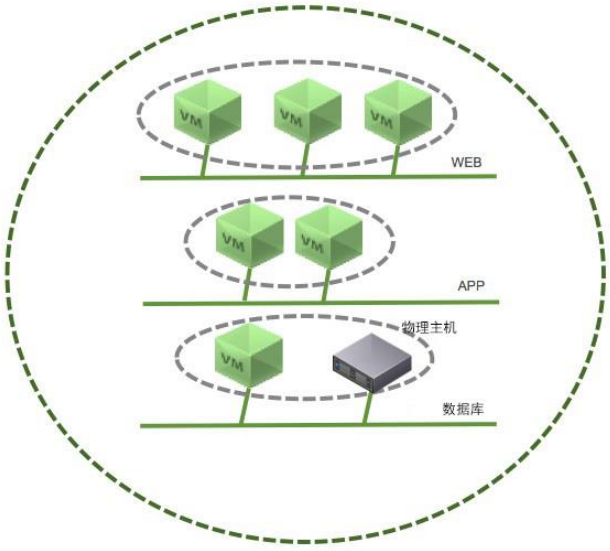



**图 24 - 为应用层创建逻辑交换机**

值得注意的是，逻辑交换功能必须在每个网段中启用虚拟到虚拟和虚拟到物理通信，并且还需要使用 NSX VXLAN 到 VLAN 桥接才能允许连接到物理节点的逻辑空间，DB 层通常就是如此。

# 多目标流量的复制模式

当连接到不同 ESXi 主机的两个虚拟机需要相互直接通信时，在与两个 hypervisor 关联的 VTEP IP 地址之间交换单播 VXLAN 封装流量。有时，虚拟机发出的流量需要发送给属于同一逻辑交换机的所有其他虚拟机。以下三种类型的第 2 层流量尤其如此：

- 广播
- 未知单播
- 多播

**注意**：通常我们使用缩写 BUM（广播、未知单播、多播）来代指这几种多目标流量类型。

在上述三种场景的任何一种场景下，来源于指定 ESXi 主机的流量都需要复制到多个远程主机（托管同一逻辑网络中的其他虚拟机）。NSX 支持三种不同的复制模式，以便在支持 VXLAN 的逻辑交换机上进行多目标通信，即多播、单播和混合模式。默认情况下，逻辑交换机从传输域继承复制模式。不过也可能会根据指定逻辑交换机的情况覆盖此模式。

在进一步详细论述这 3 种复制模式之前，有必要介绍一下“VTEP 网段”的概念。

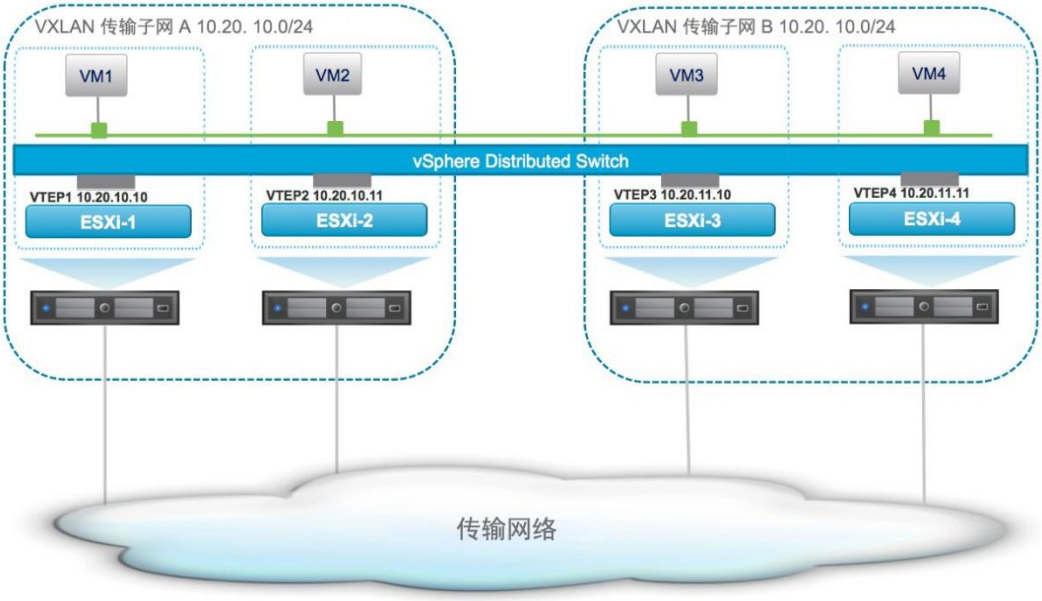



图 25 - VTEP 网段

在图 25 所示的示例中，有四个 ESXi 主机，分属于两个独立的“VTEP 网段”。ESXi-1 和 ESXi-2 的 VTEP 接口属于同一传输子网 A (10.20.10.0/24)，而 ESXi-3 和 ESXi-4 的 VTEP 接口在独立的传输子网 B (10.20.11.0/24) 中定义。VTEP 网段的定义非常重要，有助于清楚地理解以下部分所述的复制模式。

## 多播模式

如果为指定逻辑交换机选择多播复制模式，NSX 将依靠数据中心物理网络的第 2 层和第 3 层多播功能来确保封装 VXLAN 的多目标流量能够发送至所有 VTEP。多播模式是处理 VXLAN IETF 草案所指定 BUM 流量的方式，不会利用因 NSX 引入控制器集群而带来的任何增强功能（实际上，在这种模式下，根本不会用到控制器）。这种行为不允许利用逻辑和物理网络连接之间的松耦合效应，因为逻辑空间中的通信依赖于物理网络基础架构中所需的多播配置。

在此模式下，多播 IP 地址必须与每个定义的 VXLAN L2 网段（逻辑交换机）相关联。利用 L2 多播功能可以将流量复制到本地网段中所有 VTEP 上（各 VTEP IP 地址属于同一 IP 子网），而且应当在物理交换机上配置 IGMP 监听功能，以便优化 L2 多播流量的传输。若要确保多播流量也可以从源 VTEP 传输到不同子网中的 VTEP 上，网络管理员必须配置 PIM 并启用 L3 多播路由功能。

在下面的 Figure 26 示例中，VXLAN 网段 5001 已与多播组 239.1.1.1 相关联。所以，只要第一个虚拟机连接到该逻辑交换机，托管该虚拟机的 ESXi hypervisor 就会立即生成一条 IGMP Join 消息，来通知物理基础架构它想要接收发送至这一特定组的多播流量。

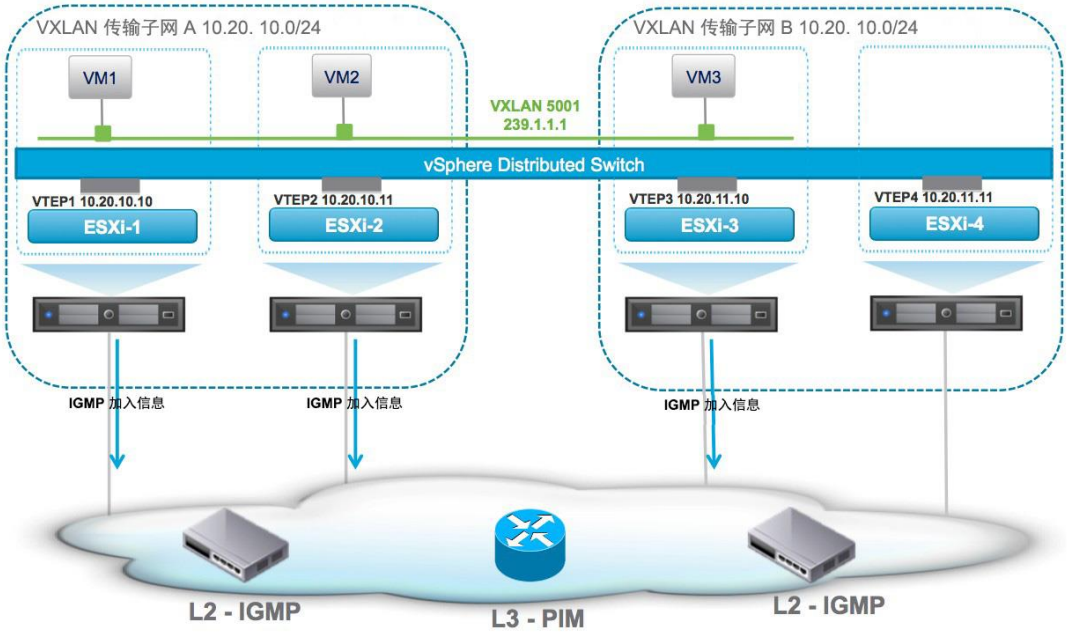



图 26 - IGMP Join

由于 ESXi1-ESXi-3 发送了 IGMP Join 消息，导致多播状态将内置于物理网络之中，从而确保发送到 239.1.1.1 目标的多播帧顺利传输。注意，ESXi-4 没有发送 IGMP Join 消息，因为它并未托管任何处于活动状态的接收器（连接到 VXLAN 5001 网段的虚拟机）。

图 27 描绘了在多播模式下传输 BUM 帧需要依循的事件顺序。

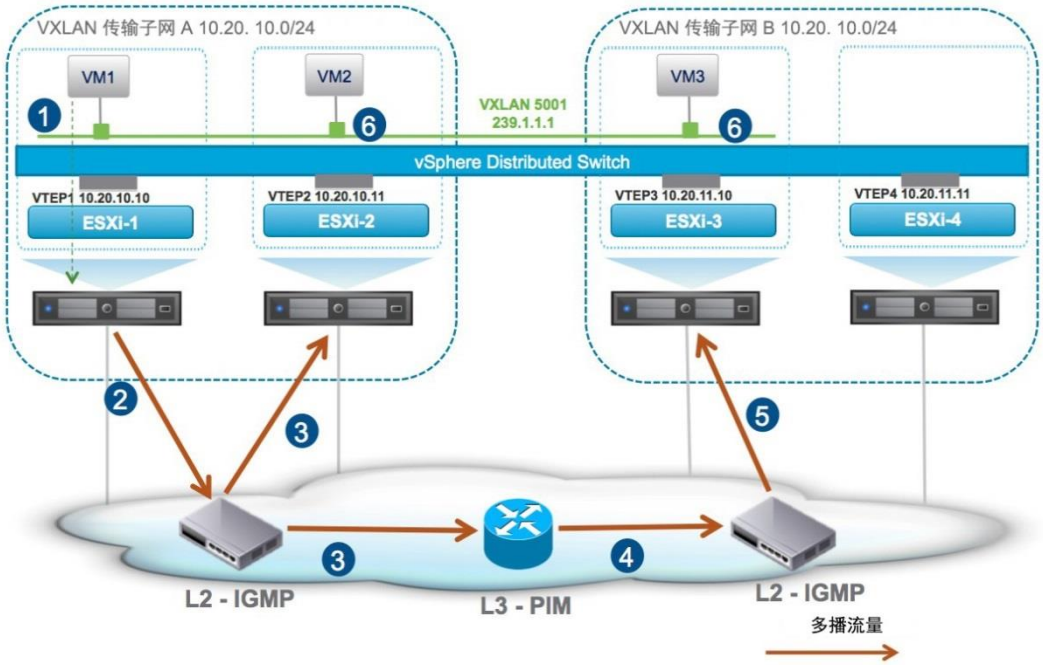



图 27 - 多播模式复制机制

1. 虚拟机 1 生成 BUM 帧。
2. ESXi-1 VXLAN 封装原始帧。外部 IP 包头中的目标 IP 地址设置为 239.1.1.1，多播数据包将发送到物理网络中。换句话说，ESXi-1 将充当多播流 239.1.1.1 的源。
3. 接收多播帧的 L2 交换机执行复制：如果该交换机上配置了 IGMP 监听功能，则只能将该帧复制到与 ESXi-2 和 L3 路由器连接的相关接口。如果未启用 IGMP 监听功能或该功能不受支持，则 L2 交换机会将此帧视为 L2 广播数据包，并将它复制到属于同一 VLAN 端口的所有接口，该端口是之前接收数据包的端口。
4. L3 路由器执行 L3 多播复制，并将数据包发送到传输子网 B 中。

5. L2 交换机会执行与步骤 3 中所述类似的操作，负责复制该帧。

6. ESXi-2 和 ESXI-3 对已接收的 VXLAN 数据包解除封装，从而公开传输到虚拟机 2 和虚拟机 3 的原始以太网帧。

当配置多播模式时，要进行一次重要选择，即如何执行 VXLAN 网段和多播组之间的映射：

- 第一种选择是执行 1:1 映射：这样做的优势在于可以非常精确地传输多播流量，因为只有在至少一台本地虚拟机连接到相应多播组时，指定的 ESXi 主机才会为该多播组接收流量。不足之处在于，选择这种映射可能会显著增加需要内置于物理网络设备中的多播状态量，因此，必须要确定这些平台能够支持的组的最大数量。
- 从另一个角度来说，可以选择将单个多播组用于所有定义的 VXLAN 网段。这将大大降低传输基础架构中的多播量，但可能会导致 ESXi 主机接收不必要的流量。仍参考图 27 中的示例，只要某台本地虚拟机连接到任一逻辑网段（非 VXLAN 5001 亦可），ESXi-4 就会立即接收属于 VXLAN 5001 的流量。
- 在大多数情况下，通常采用 m:n 映射比作为介于上述两种选择之间的折中方案。经过这样的配置后，每当新的 VXLAN 网段进行实例化时（最大指定值为“m”），都将以循环的方式映射到指定范围的多播组部分。这意味着，VXLAN 网段“1”和“n+1”将使用相同的组，其他网段就此范围以此类推。

总之，本文对多播模式的介绍主要测重于基础知识，有助于更好地理解和领会 NSX-v 通过其他两个选项（单播模式和混合模式）实现的增强功能。

## 单播模式

单播模式是一种与多播模式完全相反的方法，在该模式下，逻辑和物理网络之间完全实现了松耦合。在单播模式下，NSX 域中的 ESXi 主机根据其 VTEP 接口所属的 IP 子网，划分为几个独立的组（VTEP 网段）。每个 VTEP 网段中的 ESXi 主机都被选来充当单播安全加密链路终端 (UTEP)。UTEP 负责将通过 ESXi hypervisor（托管用于获取流量来源的虚拟机，并且属于不同的 VTEP 网段）接收的多目标流量复制到其网段（网段的 VTEP 属于 UTEP 的 VTEP 接口的同一子网）中的所有 ESXi 主机。

为了优化复制行为，每一个 UTEP 都只是将流量复制到本地网段上的 ESXi 主机，这些主机至少有一台虚拟机有效地连接到用于接收多目标流量的逻辑网络。同样，流量也只能由源 ESXi 发送到远程 UTEP，前提是至少有一台处于活动状态的虚拟机连接到该远程网段中的某个 ESXi 主机。

下方的图 28 阐明了单播模式复制机制。

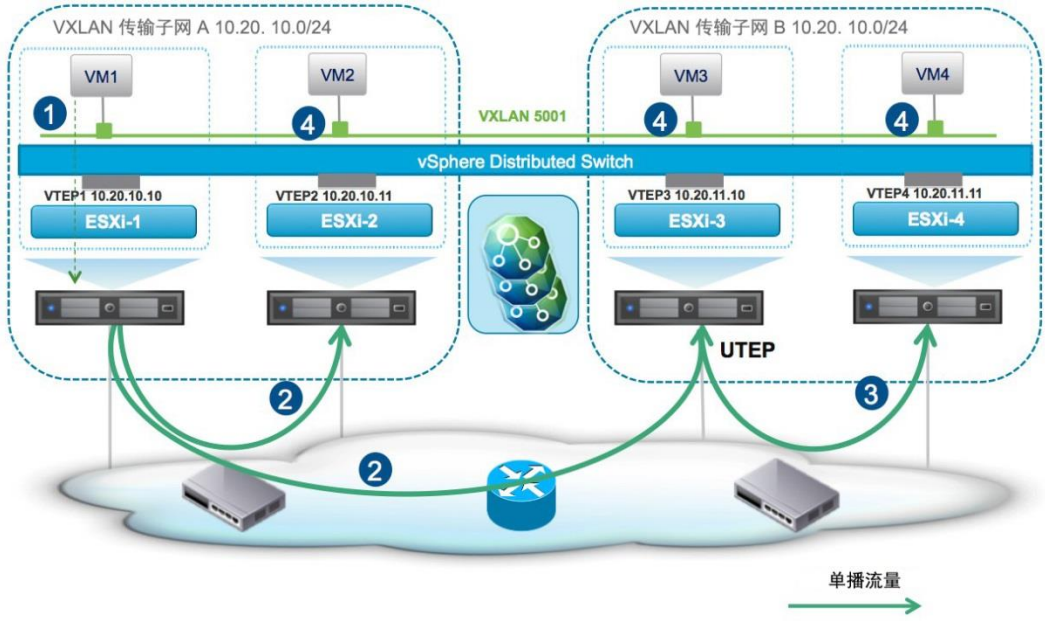



图 28 - 单播模式

1. 在这个示例中，有 4 个虚拟机连接到逻辑交换机 5001，虚拟机 1 生成要发送到所有虚拟机的 BUM 帧。注意，在这种情况下，不需要指定与此 VXLAN 网段相关联的多播组。

2. 由于本地 VTEP 表填充了通过控制平面与控制器节点的通信接收的信息，ESXi1 将查看此表，并确定是否需要仅将数据包复制到其他属于本地网段 (ESXi2) 的 VTEP 和远程网段（即 ESXi3，因为本示例中仅显示了一个远程 VTEP 网段）的 UTEP 中。对于发送到 UTEP 的单播复制，它的特点是在 VXLAN 包头中拥有一个特定的位组（即“REPLICATE_LOCALLY”位），用于向 UTEP 指示此帧来自于远程 VTEP 网段，并且可能需要进行本地复制。

3. UTEP 将接收该帧、查看本地 VTEP 表并将它复制到本地 VTEP 网段中的所有 ESXi 主机上，前提是至少要有一台虚拟机连接到 VXLAN 5001（在本示例中仅指 ESXi-4）。

仍参考以上示例，如果是虚拟机 2 生成 BUM 帧，单播复制将由 ESXi-2 按照与上述相同的步骤来执行。ESXi-2 可能会选择不同的主机（例如 ESXi-4）作为远程 UTEP，因为这是由每台 ESXi 主机在本地独立做出的决定，目的是确保复制任务可以由多个 VTEP 共同负责，即使复制属于同一逻辑网段的流量也是如此。

单播模式复制模式不需要在物理网络上进行任何明确的配置，就可以支持多目标 VXLAN 流量的分发。这种模式非常适合规模较小的部署，其中每个网段都具有很少的 VTEP，并且物理网段也很少。但是在大型环境中，尤其是在虚拟机可以生成大量 BUM 流量的环境中，并不建议使用此模式，因为复制的开销随网段数量的增加而增加。

## 混合模式

混合模式可提供类似于单播模式的操作简易性（在物理网络中不需要进行任何 IP 多播路由配置），同时还可以利用物理交换机的第 2 层多播功能。

这一点在图 29 的示例中进行了说明，在该图中还可以看到负责向同一子网中的其他 VTEP 执行本地复制的特定 VTEP 现已命名为“MTEP”。因为在混合模式下，[M]TEP 使用 L2 多播对 BUM 帧进行本地复制，而 [U]TEP 则利用单播帧来执行这一操作。

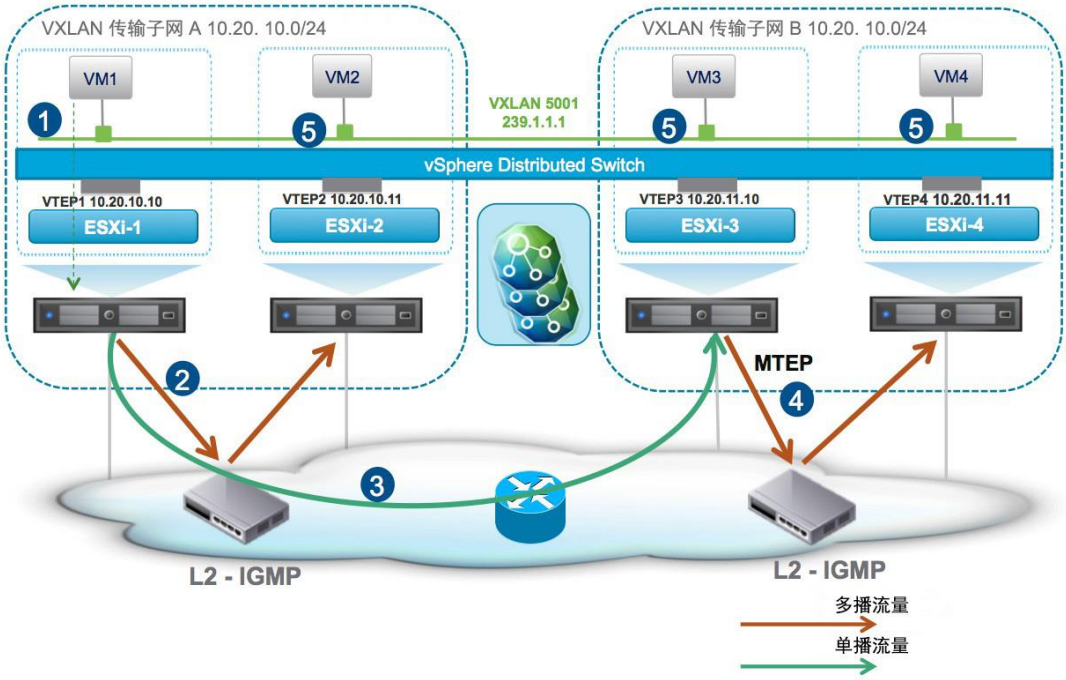



图 29 - 混合模式逻辑交换机

1. 虚拟机 1 生成需要复制到 VXLAN 5001 中所有其他虚拟机的 BUM 帧。当对本地流量复制执行多播封装时，多播组 239.1.1.1 必须与 VXLAN 网段相关联。

2. ESXi1 将该帧封装在发送到 239.1.1.1 组的多播数据包中。利用物理网络中的第 2 层多播配置可确保 VXLAN 帧传输到本地 VTEP 网段中的所有 VTEP；在混合模式下，当存在有意接收多目标流量的本地虚拟机时，ESXi 主机将发送一条 IGMP Join 消息（与前面的图 26 所示类似）。由于不需要使用 PIM（所以通常不进行配置），因此强烈建议为每个 VLAN 定义一个“IGMP 查询器”，以便确保 L2 多播传输成功，避免不确定的行为。

   **注意**：当启用 IGMP 监听功能但没有定义“IGMP 查询器”时，某些类型的以太网交换机将在 VLAN 中实施多播流量泛洪，而其他交换机则丢弃流量。请参阅您选择的供应商提供的文档。

3. 与此同时，ESXi-1 将查看本地 VTEP 表并确定是否需要将数据包复制到远程网段（即 ESXi3，因为本示例中仅显示了一个远程 VTEP 网段）的 MTEP 部分。通过 VXLAN 包头中的位组发送到 MTEP 的单播复制，可用来向 MTEP 指示此帧来自于远程 VTEP 网段，而且需要从本地重新注入网络。

4. MTEP 将创建一个多播数据包并将其发送至物理网络，以便本地 L2 交换基础架构复制该数据包。

**注意**：与之前针对单播模式的论述类似，如果是虚拟机 2 生成 BUM 帧，ESXi-2 主机将向远程 MTEP 执行单播复制。同样，每个 ESXi 主机都在本地确定哪些属于远程 VTEP 网段的 ESXi 主机将充当 MTEP：ESXi-1 可能会使用 ESXi-3 作为远程 MTEP（如上图所示），而 ESXi-2 可能会使用 ESXi-4 作为远程 MTEP。

# 填充 Controller 表

在介绍完如何在 NSX-v 部署过程中处理多目标流量后，我们现在来看看第 2 层单播通信。此处将重点介绍充分利用 NSX Controller 节点的 NSX-v 部署；因此，首先要讲述的是 ESXi 主机与控制器集群之间用于填充控制器节点上 VTEP、MAC 和 ARP 表的控制平面通信。

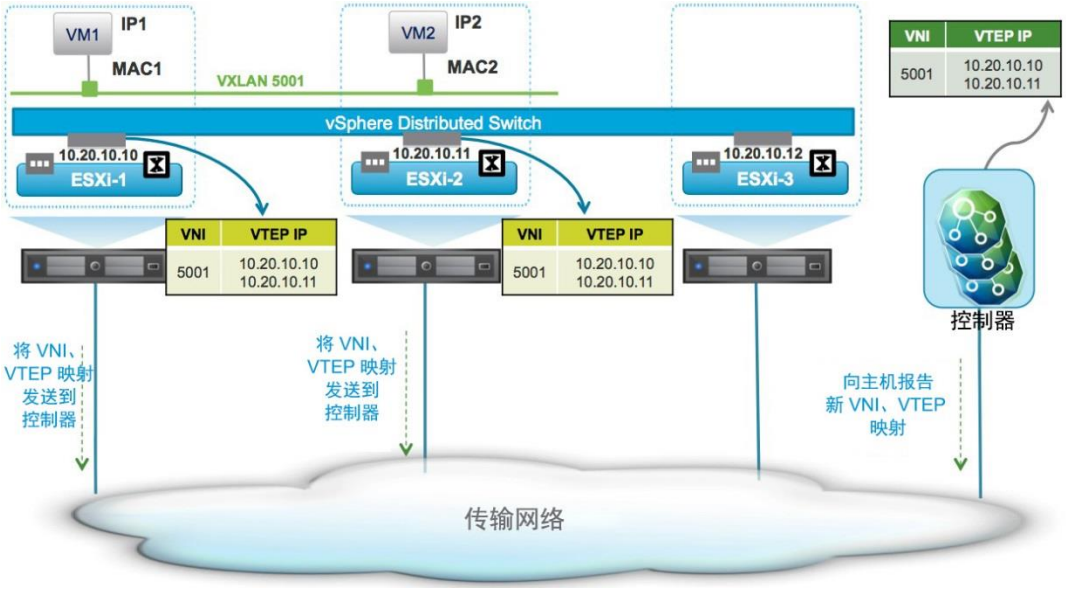



图 30 - VNI-VTEP 报告

如图 30 所示，当第一个虚拟机连接到 VXLAN 网段（本例中为 ESXi1 上的虚拟机 1，ESXi-2 上的 2 号虚拟机）时，ESXi 主机将生成一条控制平面报文，与 VNI/VTEP 映射信息一起发送到控制器（或者最好发送到控制该特定逻辑交换机细分块的特定控制器节点）。该控制器节点将以此信息填充其本地 VTEP 表，并向所有 ESXi hypervisor（托管有效地连接到同一 VXLAN 网段的虚拟机）发送一条报告消息（这就是我们的示例中未将该消息发送至 ESXi-3 的原因）。
然后这些 ESXi 主机就可以填充它们的本地 VTEP 表；例如，正如前文所述，可以利用此信息来确定多目标流量所要复制到的 VTEP 列表。

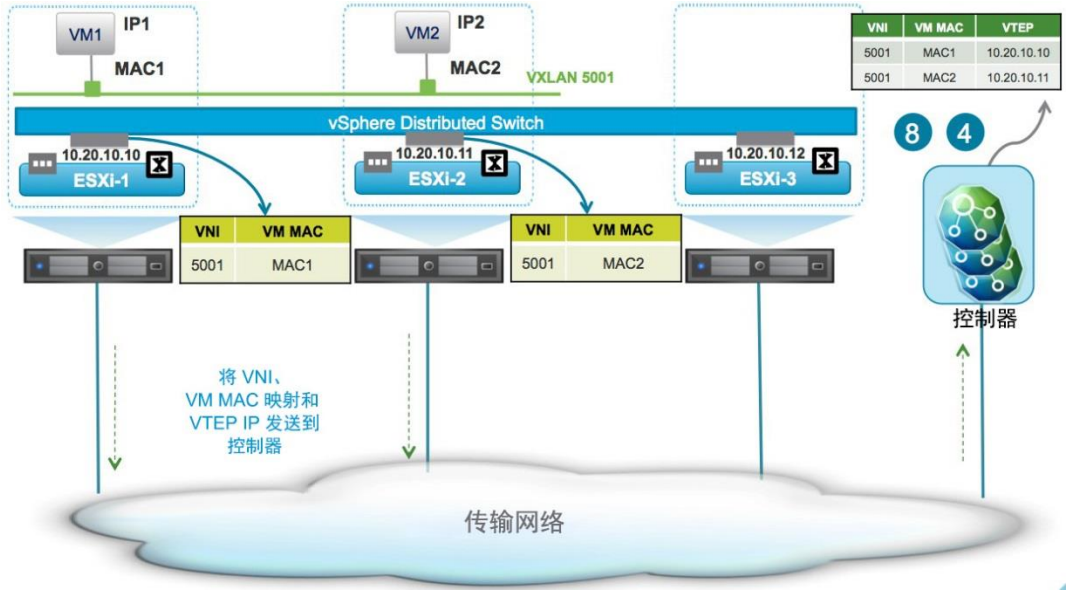



图 31 - VNI-MAC 地址报告

这些 ESXi 主机向控制器发送的第二条信息是已在本地连接到特定 VNI 的虚拟机的 MAC 地址。控制器将使用这些报告来填充其本地 MAC 表，但与 VNI-VTEP 报告不同的是，它不会将此信息发送到所有 ESXi 主机。这就是为什么这些 ESXi 主机目前只能获知本地连接的 MAC 地址，如上方图 31 所示。

最后，控制器需要的最后一条信息是这些虚拟机的 IP 地址，以便填充用于执行 ARP 抑制功能的本地 ARP 表，在下一节中讨论单播通信的时，将对该功能进行介绍。

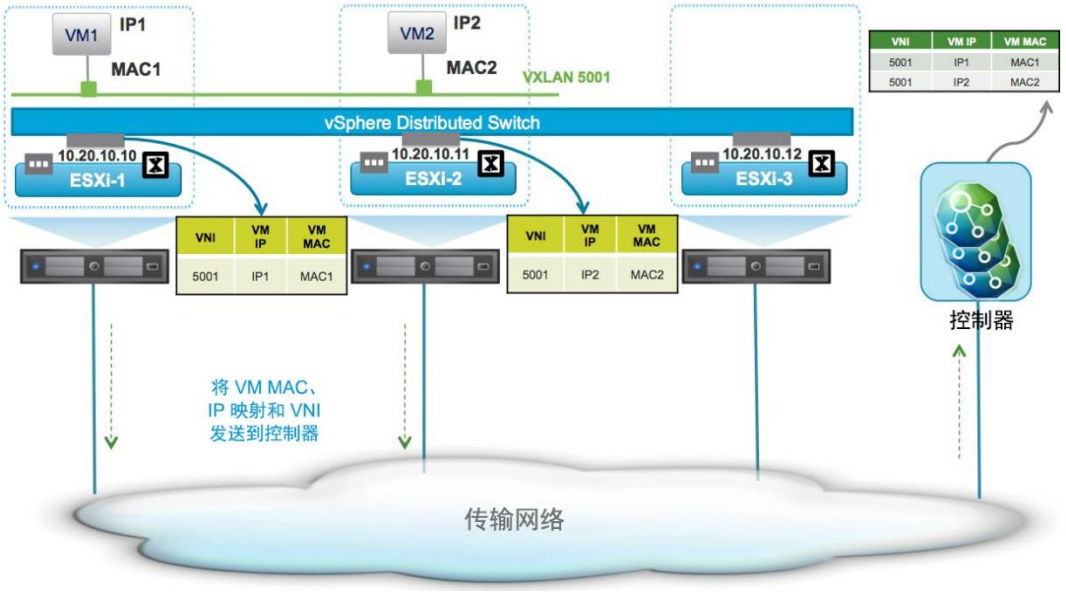



图 32 - VNI-IP 地址报告

这些 ESXi 主机主要通过以下两种方法获取本地连接的虚拟机的 IP 地址：

1. 对于通过 DHCP 获得 IP 地址的虚拟机，ESXi 主机能够通过监听 DHCP 服务器发送的去往本地虚拟机的 DHCP 回复来获取 IP 地址。
2. 对于具有静态地址的虚拟机，将使用由这些虚拟机发出的 ARP 请求来获取它们的 IP 地址。

获取虚拟机的 IP 地址后，ESXi 主机将立即向控制器发送 MAC/IP/VNI 信息，以便使用此信息填充本地 ARP 表。

# 单播流量（虚拟到虚拟通信）

利用上述表中的信息，NSX Controller 可以执行“ARP 抑制”功能，该功能不需要在连接了虚拟机的 L2 域（VXLAN 网段）中实施 ARP 流量泛洪。ARP 请求代表着网络中发现的绝大多数 L2 广播流量，删除该请求将大大有益于整个网络基础架构的稳定性和可扩展性。

图 33 展示了需要在属于同一逻辑交换机（VXLAN 网段）的虚拟机之间建立单播通信时，如何利用带有 NSX Controller 的控制平面来执行 ARP 解析。

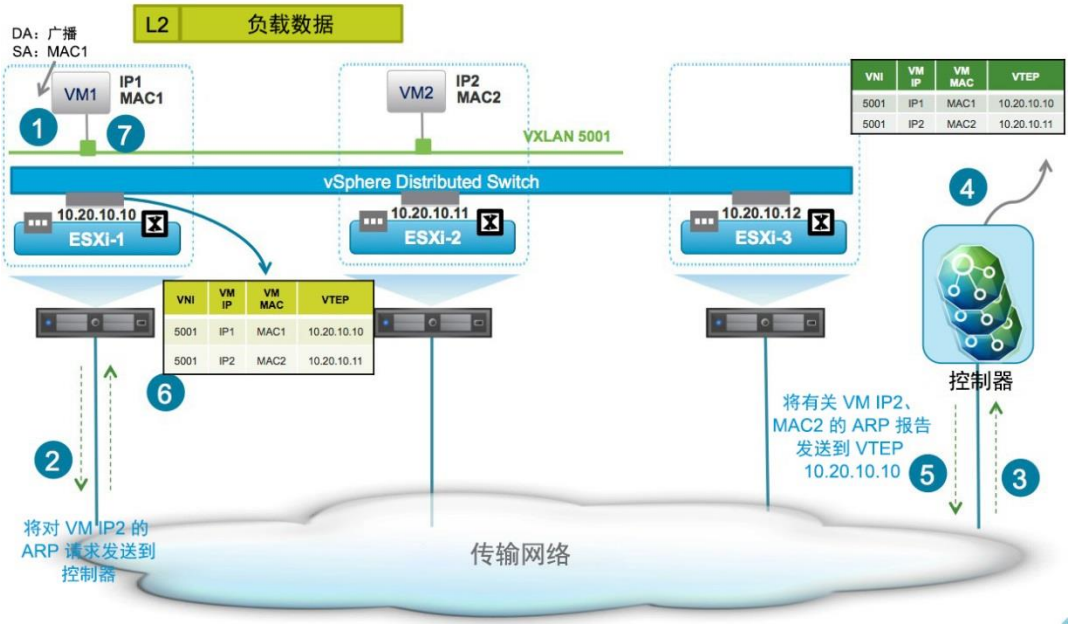



图 33 - 通过 NSX Controller 执行 ARP 解析

1. 虚拟机 1 生成 ARP 请求（L2 广播数据包）来确定虚拟机 2 的 MAC/IP 映射信息。
2. ESXi-1 拦截该请求并生成控制平面请求，用于向控制器请求 MAC/IP 映射信息。
3. 控制器接收控制平面请求。
4. 控制器查看其本地 ARP 表中是否存在所需的映射信息。
5. 映射信息与控制平面 ARP 报告消息一起发送至 ESXi-1。
6. ESXi-1 接收控制平面消息并使用此映射信息更新其本地表，因此，目前已知连接虚拟机 2 的 VTEP (10.20.10.11)。
7. ESXi-1 以虚拟机 2（帧中的源 MAC 地址是 MAC2）的名义生成 ARP 响应，并将它传送至虚拟机 1。这一点至关重要，因为虚拟机 1 不需要知道 ARP 回复得到了怎样的处理，从自身的角度来看，它直接来自虚拟机 2。

**注意**：如果控制器没有映射信息，它将通知 ESXi-1，然后在 VXLAN 5001 网段内实施 ARP 帧泛洪，以便发现虚拟机 2。执行 ARP 请求泛洪的方式取决于已配置的逻辑交换机复制模式，如“适用于多目标流量的复制模式”部分中所述。

在虚拟机 1 填充完它的 ARP 缓存后，它可以向虚拟机 2 发送数据流量，如图 34 所示。

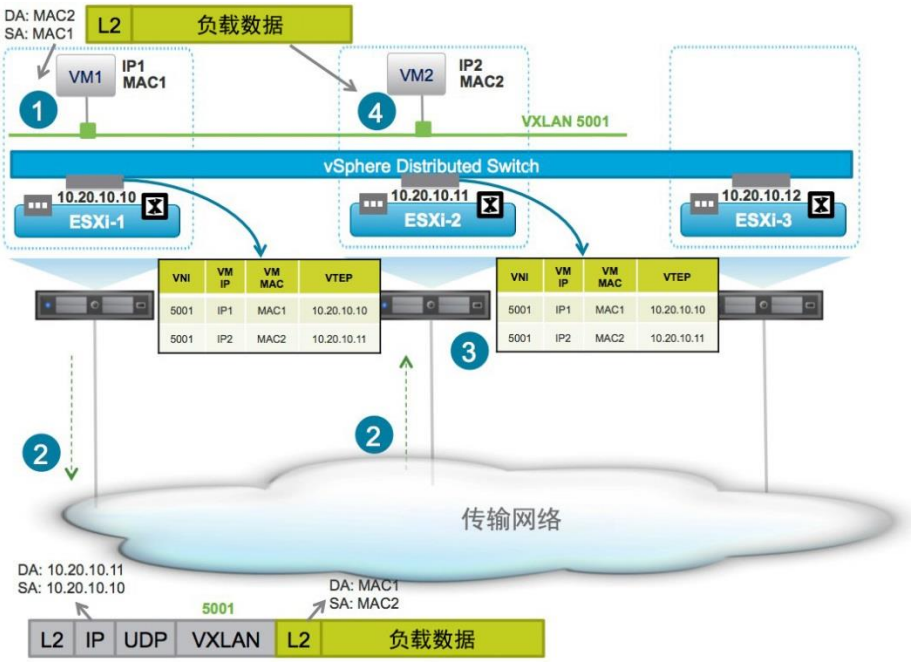



图 34 - 逻辑交换机内单播通信

1. 虚拟机 1 生成定向到虚拟机 2 的数据包。
2. ESXi-1 通过从控制器接收的 ARP 报告获得虚拟机 2 的位置（即控制平面获得了此位置信息），因此它会将虚拟机 1 发出的数据包封装在目标为 ESXi2 的 VTEP (10.20.10.11) 的 VXLAN 数据包中。
3. ESXi-2 接收该数据包，并在对其解除封装时，利用外部 IP 包头中的信息来虚拟机 1 的位置（使虚拟机 1 的 MAC 和 IP 地址与 ESXi-1 的 VTEP 相关联）。本示例演示了数据平面获取虚拟机特定位置的过程。
4. 该帧传输到虚拟机 2。

**注意**：由于 ESXi-1 和 ESXi-2 都填充了它们的本地表，所以流量可以双向流动。

# 单播流量（虚拟到物理通信）

在有些情况下，可能需要在虚拟和无论工作负载之间建立 L2 通信。此处列出了几种典型情况（并非详尽无遗）：

- 部署多层应用：在某些情况下，Web、应用和数据库层可以作为同一 IP 子网的一部分进行部署。Web 和应用层通常利用虚拟工作负载，而数据库层则与此不同，对于后者，通常部署裸机服务器。因此，可能还需要在应用和数据库层之间建立子网内（即 L2 域内）通信。
- 物理到虚拟 (P-to-V) 迁移：许多客户都将运行于裸机服务器的应用虚拟化，但在物理到虚拟迁移的过程中，需要支持同一 IP 子网上的虚拟和物理混合节点。

- 利用外部物理设备作为默认网关：在此类情况下，可能要部署物理网络设备来充当已连接到逻辑交换机的虚拟工作负载的默认网关，并且需要利用 L2 网关功能建立与该网关的连接。
- 部署物理设备（防火墙、负载均衡器等）。

若要满足以上所列的具体要求，可以部署能够执行“桥接”功能的设备，为“虚拟环境”（逻辑交换机）和“物理环境”（连接传统 VLAN 的非虚拟化工作负载和网络设备）之间的通信提供支持。

NSX 通过部署 NSX L2 桥接在软件中提供这一功能，从而使虚拟机可以在第 2 层连接到物理网络（VXLAN 到 VLAN ID 映射），即使运行虚拟机的 hypervisor 并未以物理方式连接到 L2 物理网络也是如此。

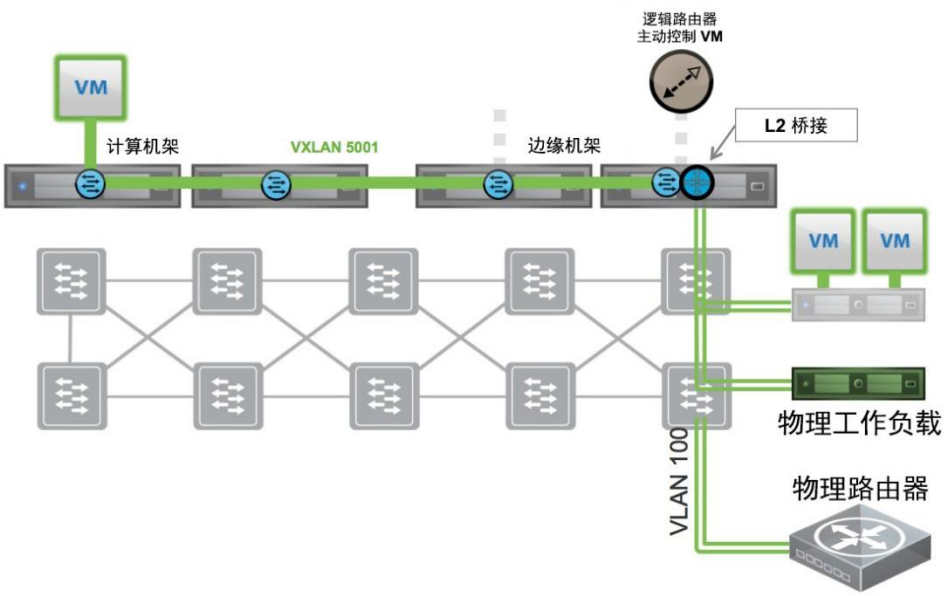



**图 35 - NSX L2 桥接**

图 35 展示了 L2 桥接的示例，其中在逻辑空间中连接到 VXLAN 网段 5001 的虚拟机需要与部署于同一 IP 子网但连接到（VLAN 100 中）物理网络基础架构的物理设备进行通信。在当前的 NSX-v 实施中，VXLAN-VLAN 桥接配置是分布式路由器配置的一部分；因此，执行 L2 桥接功能的特定 ESXi 主机就是该分布式路由器的控制虚拟机运行时所在的主机。如果该 ESXi 主机出现故障，托管备用控制虚拟机（一旦检测到活动控制虚拟机出现故障，备用控制虚拟机就会立即激活）的 ESXi 将起到 L2 桥接的作用。

**注意**：若要了解有关控制虚拟机的分布式路由及其作用的更多信息，请参阅下面的“逻辑路由”部分。

独立于特定的实施细节，以下是一些有关 NSX L2 桥接功能的重要的部署注意事项：

- VXLAN-VLAN 映射始终按 1:1 执行。这意味着指定 VXLAN 的流量只能桥接到特定的 VLAN，反之亦然。
- 指定桥接实例（针对特定的 VXLAN-VLAN 对）仅在特定的 ESXi 主机上始终处于活动状态。
- 但是，可以通过配置创建多个桥接实例（针对不同的 VXLAN-VLAN 对）并且可以确保这些实例遍及各个独立的 ESXi 主机。这将改进 L2 桥接功能的整体可扩展性。
- NSX 第 2 层桥接数据路径完全在 ESXi 内核中执行，而不是在用户空间中执行。再次说明，控制虚拟机仅用于确定指定桥接实例处于活动状态的 ESXi 主机，而不是用来执行桥接功能。

图 36 和图 37 展示了虚拟机与裸机服务器之间的 ARP 交换，这是为了能够提供虚拟到物理单播通信所需的第一步。

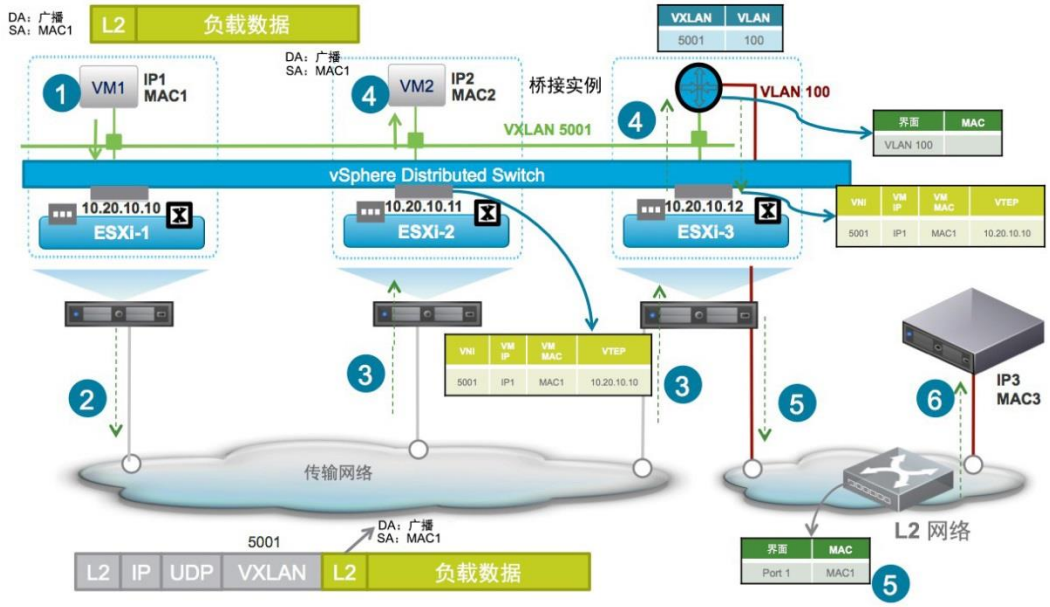



图 36 - 虚拟机发出的 ARP 请求

1. 虚拟机 1 生成 ARP 请求，用来检索裸机服务器的 MAC/IP 映射信息。
2. 正如前文所述，ESXi 主机将拦截该 ARP 请求并生成定向到控制器的控制平面请求，以便检索映射信息。假设在此场景中，控制器并没有此信息（例如，因为裸机服务器刚刚连接到网络，尚未生成流量）。在这种情况下，ESXi-1 必须利用“多目标流量的复制模式”部分中介绍的任一种方法，在 VXLAN 5001 网段中实施 ARP 请求泛洪。
3. ARP 请求将发送到 ESXi-2 和 ESXi-3，因为二者所具有的工作负载有效地连接到 VXLAN 5001（ESXi-2 上的虚拟机 2 和 ESXi-3 上的桥接实例）。ESXi 主机通过接收此数据包获得虚拟机 1 的位置信息。
4. 虚拟机 2 接收并放弃该请求。而桥接实例将 L2 广播数据包转发到物理 L2 网络。
5. 该帧在 VLAN 100 中实施泛洪，并且所有物理 L2 交换机都将执行数据平面获取虚拟机 1 MAC 地址 (MAC1) 的常规过程。
6. ARP 请求到达裸机服务器。

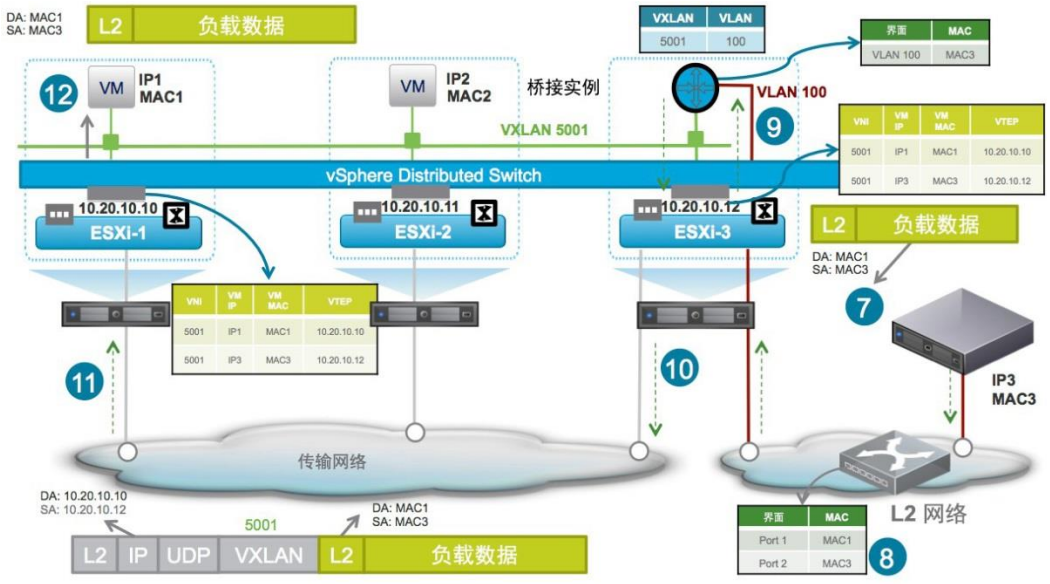



图 37 - 裸机服务器发出的 ARP 响应

7. 裸机服务器生成去往虚拟机 1 的单播 ARP 回复。
8. 该 ARP 回复将在物理网络中进行切换，并且物理 L2 交换机将执行数据平面来获取裸机服务器的 MAC 地址 (MAC3)。
9. ESXi-3 上的活动 NSX L2 桥接也将接收 ARP 回复并获得 MAC3。
10. 通过在前面的步骤 3 中执行的操作，ESXi-3 获得了 MAC2 的位置，所以它将对该帧进行封装，然后将其发送到 ESXi-1 的 VTEP (10.20.10.10)。
11. ESXi-1 在接收到该帧后，会对它进行封装解除并将 MAC3 与 ESXi3 的 VTEP 相关联的信息添加到自己的本地表，其中 ESXi3 是运行 VXLAN 5001 的活动桥接实例的主机。
12. ESXi-1 以裸机服务器（帧中的源 MAC 地址是 MAC3）的名义生成 ARP 响应，并将它传送至虚拟机 1。

此时，虚拟机 1 与裸机主机便可以开始进行数据平面通信，方式类似于之前介绍的单播虚拟到虚拟 L2 通信。

# 逻辑路由

利用 NSX 平台中的逻辑路由功能，客户可以使部署于不同逻辑 L2 网络中的终端（虚拟和物理）相互连接。通过部署网络虚拟化在网络基础架构和逻辑网络之间实现松耦合，即可做到这一点。

图 38 展示了使两个逻辑交换机互相连接的路由拓扑逻辑视图和相应的物理视图。

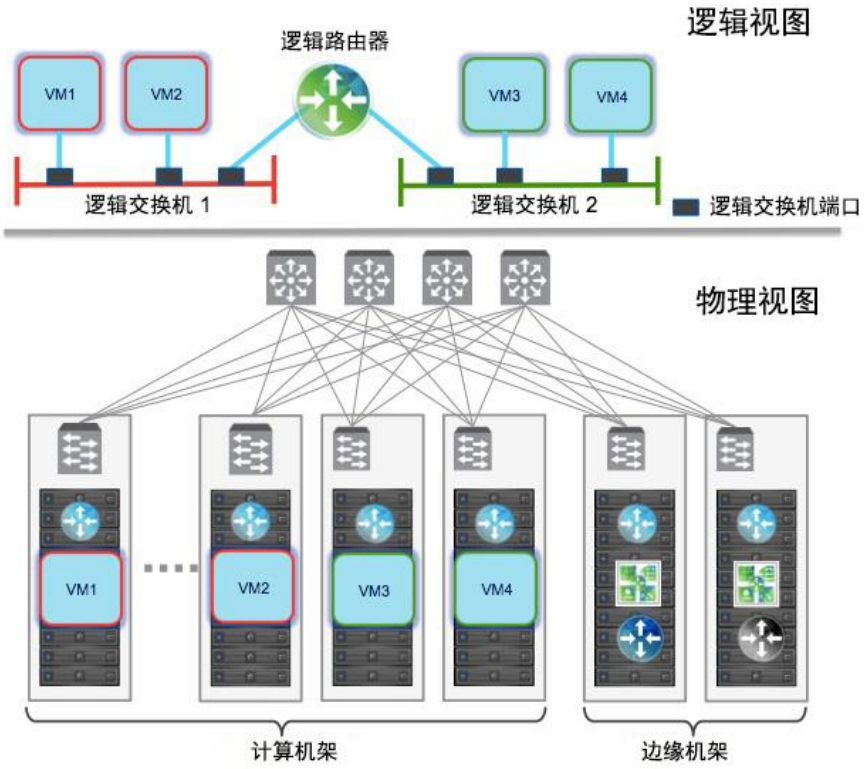



**图 38 - 逻辑路由（逻辑和物理网络视图）**

部署逻辑路由的目的有两个，即分别使以下两种终端相互连接：属于独立的逻辑 L2 域（如上图所示）的终端（逻辑或物理），或者属于设备部署在外部 L3 物理基础架构中的逻辑 L2 域的终端。第一种通信，通常局限于数据中心内部，称为东西向通信；第二种则被称为南北向通信，可以使外部物理环境（WAN、Internet 或内联网）连接到数据中心。

重新回到部署多层应用的示例，逻辑路由是使不同应用层（东西向通信）之间相互连接所必需的功能，也是从外部 L3 域（南北向通信）访问 Web 层所必需的功能，如下图 39 所示。

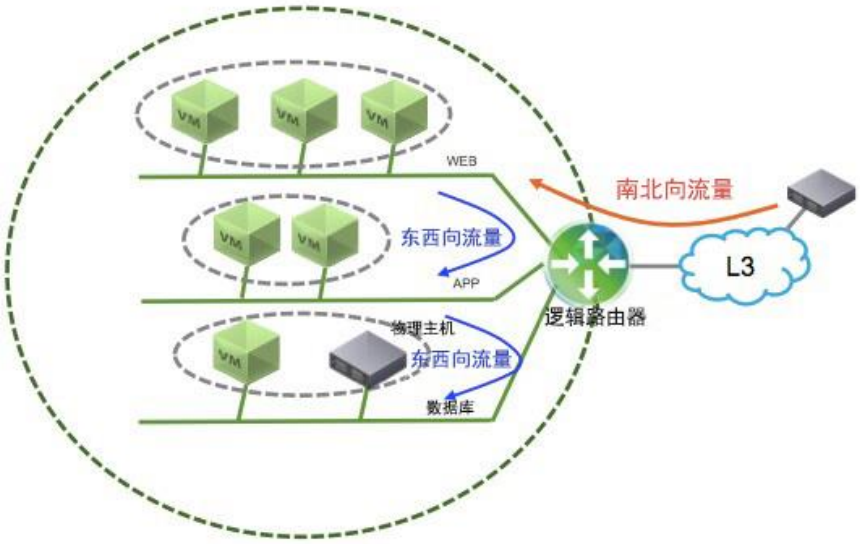



图 39 - 多层应用的逻辑路由

# 逻辑路由组件

上述两种逻辑路由通信通常利用两种不同的功能来实现：集中式路由和分布式路由。

集中式路由是指允许在逻辑网络空间与外部第 3 层物理基础架构之间进行通信的入口/出口功能。

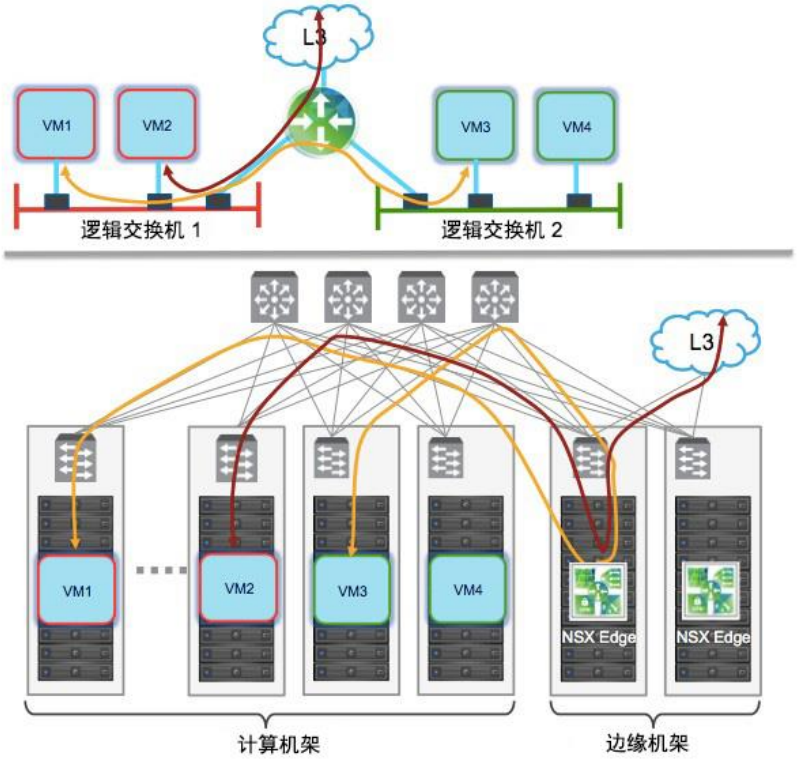



图 40 - 集中式路由

图 40 突出展示了 NSX Edge 服务网关如何在 NSX-v 平台中提供传统的集中式路由支持。除了路由服务之外，NSX Edge 还支持其他网络服务，如 DHCP、网络地址转换、防火墙、负载均衡和 VPN。

注意，集中式路由部署可同时用于东西向和南北向通信。但是，在集中式路由部署中无法最大限度地利用以东西向路由的流量，因为流量始终以“发夹”传输方式从计算机架流向部署活动 NSX Edge 的边缘机架。即使属于相互独立的逻辑交换机的两台虚拟机部署在同一 hypervisor 内也是如此。

如图 41 所示，部署分布式路由将通过提供 hypervisor 级路由功能来阻止虚拟机与虚拟机之间出现“发夹式”路由通信。每个 hypervisor 都在内核中装有特定流信息（从 NSX Controller 接收的信息），用于确保即使当这些终端属于独立的逻辑交换机（IP 子网）时，仍存在直接通信路径。

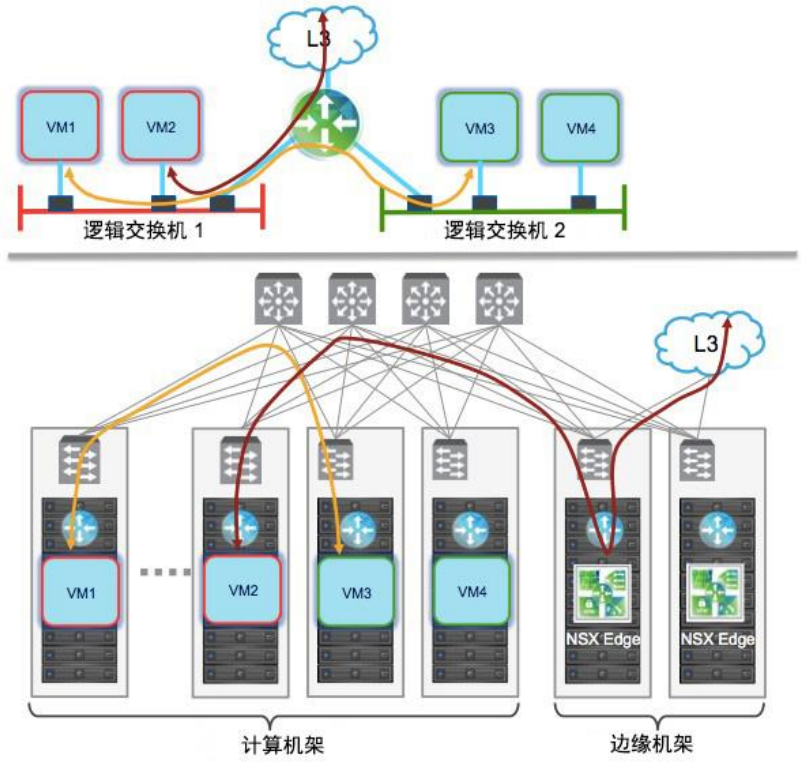



**图 41 - 分布式路由**

图 42 展示了分布式路由和集中式路由所涉及的 NSX 组件。


我们已经探讨了 NSX Edge 在处理集中式路由和其他逻辑网络服务方面起到的作用。逻辑路由通过一个称为“分布式逻辑路由器（DLR）”的逻辑元件提供，该路由器具有以下两个主要组件：

- DLR 控制平面由 DLR 控制虚拟机提供：该虚拟机可支持动态路由协议（BGP、OSPF）、与下一个第 3 层跃点设备（通常为 NSX Edge）交换路由更新，并且可与 NSX Manager 和 Controller 集群进行通信。DLR 控制虚拟机的高可用性通过“活动-备用”配置提供支持：可提供一对在活动/备用模式下运行的虚拟机，与之前针对 NSX Edge 的论述类似。

- 在数据平面级别上，DLR 内核模块 (VIB) 安装在属于 NSX 域的 ESXi 主机上。内核模块类似于支持第 3 层路由的模块化机架中的线路卡。这些内核模块拥有一个通过控制器集群递送的路由信息库 (RIB)。路由查找、ARP 条目查找等所有数据平面功能均由这些内核模块执行。这些内核模块配备了逻辑接口（称为 LIF），用于连接不同的逻辑交换机（和/或支持 VLAN 的端口组）。每个 LIF 都分配了一个 IP 地址，用于表示它所连接的逻辑 L2 网段的默认 IP 网关，还分配了一个虚拟 MAC 地址。很显然，每个 LIF 都有一个唯一的 IP 地址，而所有定义的 LIF 都分配了相同的虚拟 MAC 地址。

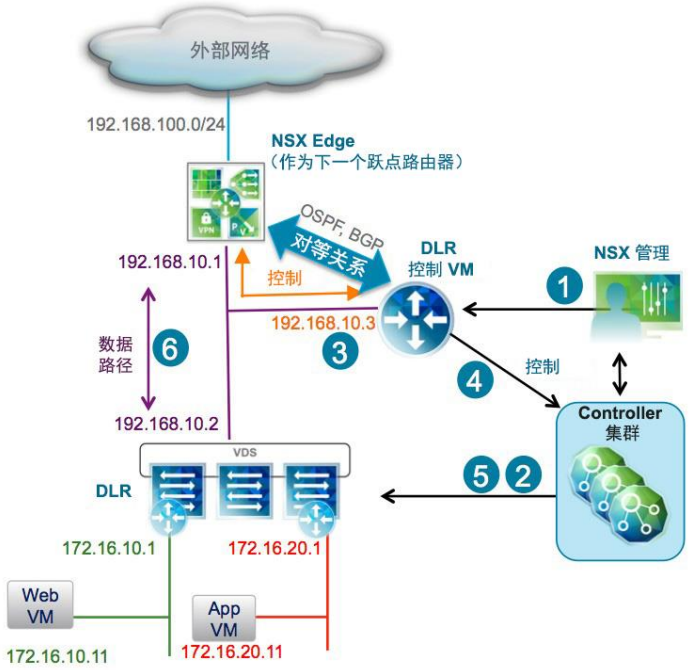



图 42 - 逻辑路由组件

图 42 展示了所有逻辑路由组件之间用于启用分布式路由功能的交互。

1. 利用所选的协议（OSPF 或 BGP），从 NSX Manager UI（或通过调用 API）创建 DLR 实例并启用路由功能。
2. Controller 利用控制平面与 ESXi 主机来递送新的 DLR 配置，包括 LIF 及其相关联的 IP 和虚拟 MAC 地址。
3. 假设在下一个跃点层设备（在本例中为 NSX Edge）上还支持采用路由协议，并且在 NSX Edge 与 DLR 控制虚拟机之间建立了 OSPF/BGP 对等关系。然后，NSX Edge 和 DLR 可以交换路由信息：
   - DLR 控制虚拟机可以配置为将所有已连接的逻辑网络（本示例中为 172.16.10.0/24 和 172.16.20.0/24）的 IP 前缀重新分发到 OSPF。因此，它随后会将这些路由通告递送至 NSX Edge。注意，这些前缀的下一个跃点不是已分配给控制虚拟机的 IP 地址 (192.168.10.3)，而是用于识别 DLR 数据平面组件的 IP 地址 (192.168.10.2)。前者称为 DLR“协议地址”，而后者则称为“转发地址”。
   - NSX Edge 可以将这些前缀递送至控制虚拟机，以便连接外部网络中的 IP 网络；在大多数情况下，单个默认路由可能由 NSX Edge 发送，因为它表示退出到物理网络基础架构的单个退出点。
4. DLR 控制虚拟机会将这些从 NSX Edge 获取的 IP 路由递送至 Controller 集群。
5. Controller 集群负责将从 DLR 控制虚拟机获取的路由数据分发到各个 hypervisor 中。集群中的每个控制器节点均负责分发特定逻辑路由器实例的信息。如果在部署过程中部署了多个逻辑路由器实例，负载将分发至各个控制器节点上。单独的逻辑路由器实例通常与每个部署的租户相关联。
6. 主机上的 DLR 路由内核模块处理数据路径流量，以便通过 NSX Edge 与外部网络进行通信。

图 43 展示了在两台已连接到独立逻辑网段的虚拟机之间建立路由通信所需的步骤。

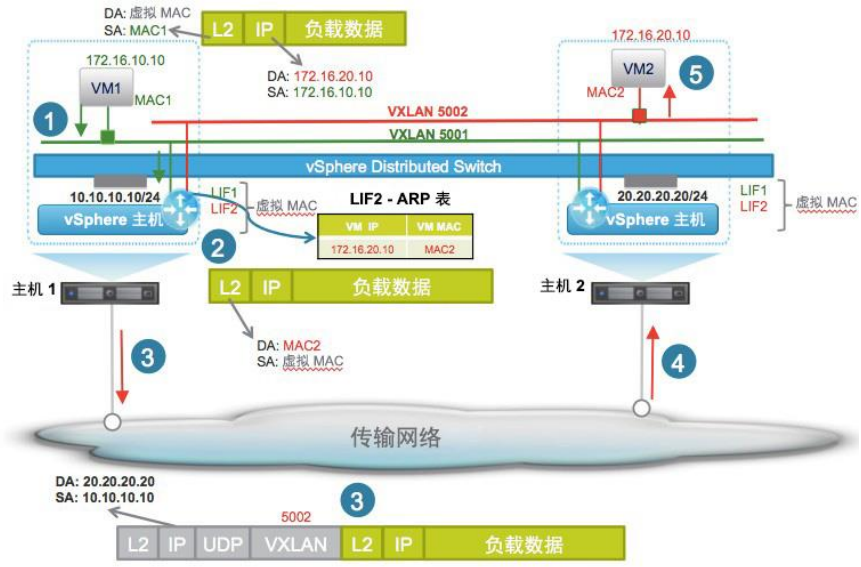



图 43 - 虚拟机之间的路由通信

1. 虚拟机 1 想要向已连接到不同 VXLAN 网段的虚拟机 2 发送数据包，所以数据包将发送到位于本地 DLR 上的虚拟机 1 默认网关接口 (172.16.10.1)。
2. 在本地 DLR 上执行路由查找，确定目标子网是否已直接连接到 DLR LIF2。在 LIF2 ARP 表中执行查找，确定与虚拟机 2 的 IP 地址相关联的 MAC 地址。

   **注意**：如果 ARP 信息不可用，DLR 将在 VXLAN 5002 上生成 ARP 请求，以便确定所需的映射信息。
3. 在本地 MAC 表中执行 L2 查找，确定如何连接到虚拟机 2，然后原始数据包将进行 VXLAN 封装，并发送至 ESXi2 的 VTEP (20.20.20.20)。
4. ESXi-2 负责对该数据包解除封装，并在与 VXLAN 5002 网段相关联的本地 MAC 表中执行 L2 查找。
5. 数据包传送至目标虚拟机 2。

注意，当虚拟机 2 回复虚拟机 1 时，两个逻辑网段之间的路由将在虚拟机 2 所连接的 ESXi-2 上执行。这是 NSX DLR 的正常行为，目的是为了始终在运行于 ESXi 内核的 DLR 实例上执行本地路由，其中 ESXi 托管了发起通信的工作负载。

图 44 展示了以下过程的所需步骤：在外部网络与连接了分布式逻辑路由器的逻辑网段之间建立传入通信。

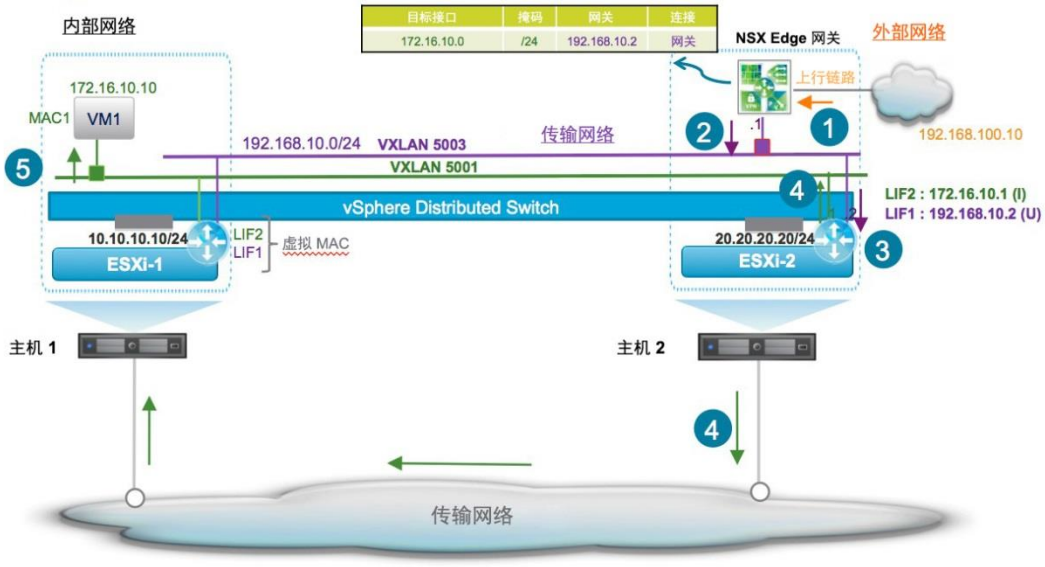



图 44 - DLR：来自外部网络的传入流量

1. 外部网络上的设备 (192.168.100.10) 想要与 VXLAN 5001 网段上的虚拟机 1 (172.16.10.10) 进行通信。
2. 数据包将从物理网络传输至托管 NSX Edge 的 ESXi 服务器。NSX Edge 接收数据包并执行路由查找，即通过传输网络上的 DLR IP 地址查找定向到 172.16.10.0/24 子网的路由。之前已通过与 DLR 控制虚拟机进行路由交换获得 IP 前缀，下一个跃点 (192.168.10.2) 表示数据平面上 DLR 的 IP 地址（转发地址）。
3. 此外，ESXi-2 的内核中也安装了 DLR，活动 NSX Edge 就部署在此处。这意味着，NSX Edge 级别和 DLR 级别的两种路由管道均在 ESXi-2 上本地执行。
4. 目标 IP 子网 (172.16.10.0/24) 直接连接到 DLR，所以数据包将从传输网络路由至 VXLAN 5001 网段中。在完成 L2 查找后，数据包将封装在 VXLAN 包头中，并发送至虚拟机 1 驻留的 VTEP 地址 (10.10.10.10)。
5. ESXi-1 负责对该数据包解除封装，并将它传送至目标。

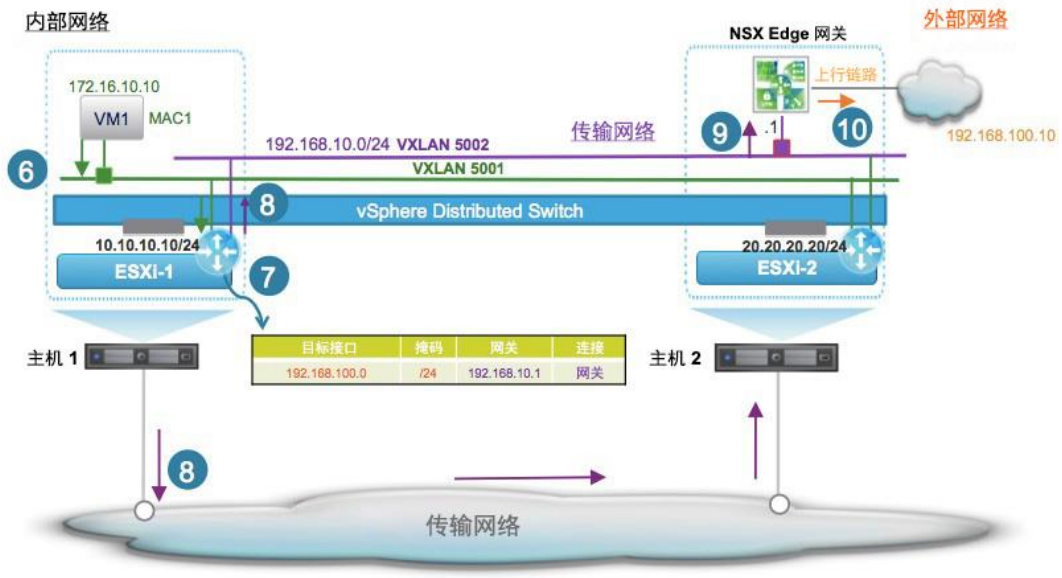



图 45 - DLR：去往外部网络的传出流量

6. 虚拟机 1 想要回复外部目标 192.168.100.10，所以数据包将发送到位于本地 DLR 上的虚拟机 1 默认网关接口 (172.16.10.1)。
7. 在本地 DLR 上执行路由查找，确定指向目标的下一个跃点是否为传输网络上的 NSX Edge 接口 (192.168.10.1)。此信息已通过 NSX Edge 自身接收在 DLR 控制虚拟机上，并已由 Controller 递送至内核 DLR 模块。
8. 执行 L2 查找，确定如何连接到传输网络上的 NSX Edge 接口，然后数据包将进行 VXLAN 封装，并发送至 ESXi2 的 VTEP (20.20.20.20)。
9. ESXi-2 对该数据包解除封装，并将它发送至目标（NSX Edge）。
10. NSX Edge 执行路由查找并将数据包发送到物理网络中，然后再发送至通往目标的路径上的下一个 L3 跃点。然后，数据包将由物理基础架构传送至最终目标 192.168.100.10。

# 逻辑路由部署选项

根据客户的不同要求，可以利用 NSX 平台的逻辑交换和逻辑路由功能构建各种各样的拓扑。在本部分中，我们将介绍使用分布式和集中式逻辑路由功能的两种路由拓扑：

1) 物理路由器作为下一个跃点
2) Edge 服务网关作为下一个跃点

## 物理路由器作为下一个跃点

如下图所示，某个组织托管了多种应用，并且希望在不同应用层之间建立连接，同时连接到外部网络。

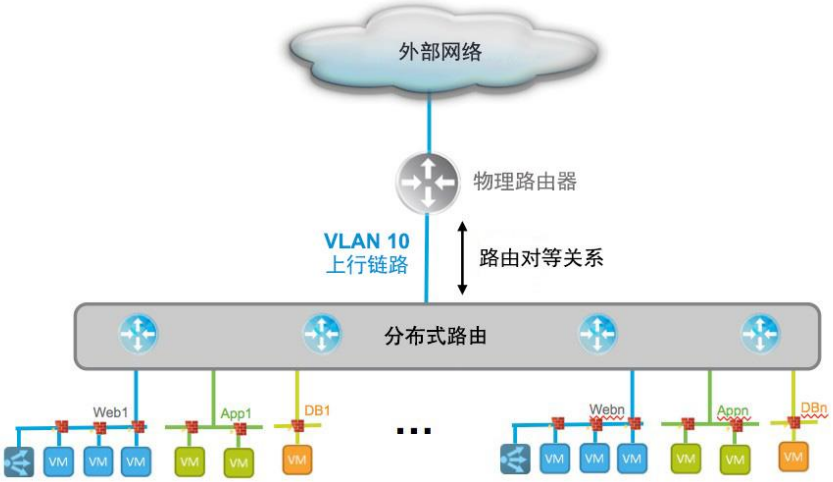



图 46 - 物理路由器作为下一个跃点

在此拓扑中，多个逻辑交换机分别为特定层中的虚拟机提供了第 2 层网络连接，东西向和南北向的路由决策在 hypervisor 级别上以分发的方式执行。分布式逻辑路由配置允许两个不同层上的虚拟机互相通信。同样，逻辑路由器上的动态路由协议支持允许与下一个跃点物理路由器进行路由交换。这反而使得外部用户可以访问已连接到数据中心内逻辑交换机的应用。

这种设计的缺点在于，DLR 用来向物理路由发送流量的 VLAN 必须提供给托管计算资源的所有 ESXi，但是否能够做到这一点，要取决于所选的特定数据中心结构设计。关于这方面的更多注意事项，请参阅本文的“NSX-v 部署注意事项”部分。

## Edge 服务网关作为下一个跃点

在 DLR 和物理路由器之间引入 NSX Edge 服务网关（如下方图 47 所示），可以解决上述问题。

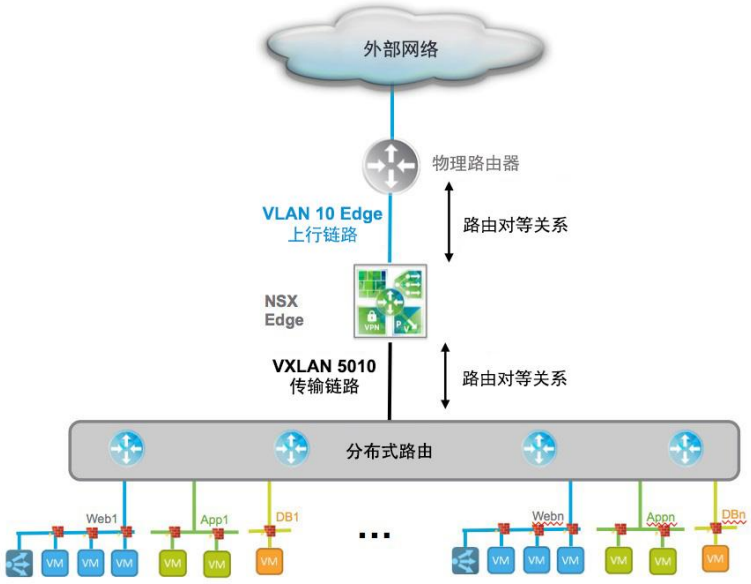



图 47 - DLR 和物理路由器之间的 NSX Edge

在此部署模式中，DLR 控制虚拟机与 VXLAN 传输链路上的 NSX Edge 建立对等关系，而 NSX Edge 使用其上行链路接口进行连接并与物理网络基础架构建立对等关系。这种方法的主要优势包括：

- 利用 DLR 与 NSX Edge 之间的 VXLAN 可以针对计算 ESXi 主机（针对出站流量执行 DLR 路由的主机）和 NSX Edge 之间的连接性相关问题，删除物理网络中的依赖关系，因为流量将在数据平面中进行 VXLAN 封装。

- 可以在初始配置和调配阶段建立逻辑空间与物理网络之间的路由对等关系。在逻辑空间中部署其他 DLR 不需要进一步修改物理网络，也不需要更改通过物理路由器建立的路由邻接关系的数量，从而在逻辑和物理网络之间完全实现 NSX 所承诺的松耦合。

使用此选项可能出现的问题是，所有南北向流量都需要通过 NSX Edge 进行传输，这表示带宽受到了限制。使用 NSX Edge 中的活动/活动 ECMP 功能（从 SW 版本 6.1 开始引入）可以为 NSX-Edge 提供的入口/出口功能横向扩展适用的带宽。

**注意**：若要了解有关活动/活动 ECMP NSX Edge 部署的更多信息，请参阅“边缘机架设计”部分。

在服务提供商环境中，有多个租户，并且每个租户对隔离逻辑网络和 LB、防火墙以及 VPN 等其他网络服务的数量要求都有所不同。在此类部署中，NSX Edge 服务网关可以提供网络服务功能以及动态路由协议支持。

如下方图 48 所示，两个租户都通过 NSX Edge 连接到外部网络。每个租户都有各自的逻辑路由器实例，用于在租户内提供路由服务。另外，租户逻辑路由器与 NSX Edge 之间的动态路由协议配置使租户虚拟机可以连接到外部网络。

在此拓扑中，东西向流量路由通过 hypervisor 中的分布式路由器处理，南北向流量则流经 NSX Edge 服务网关。

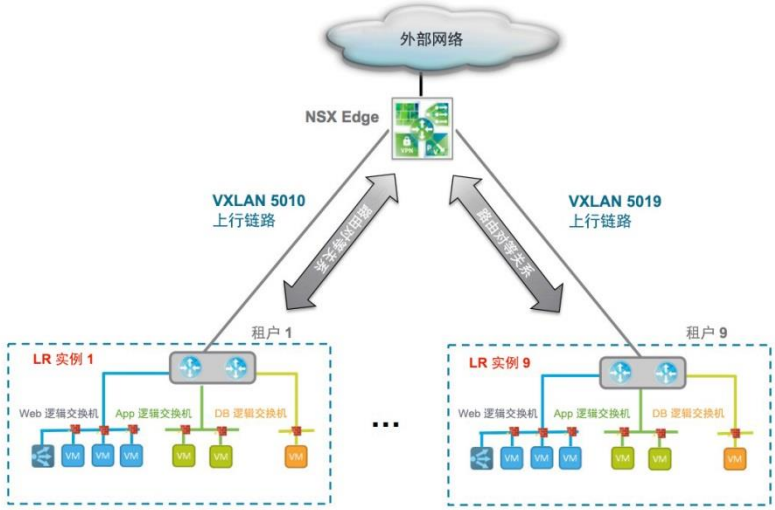



图 48 - NSX Edge 服务网关作为下一个跃点（并提供网络服务）

如上所示，此部署模式最多只能有 9 个租户可以连接到同一 NSX Edge。之所以有这样的限制，是因为 NSX Edge 像任何其他虚拟机一样，配备了 10 个虚拟接口（虚拟网卡），其中一个接口用作朝向物理网络的上行链路。另外，在这种情况下，连接到同一 NSX Edge 的 9 个租户之间不一定存在重叠的 IP 地址，除非这些重叠的前缀未在 DLR 和 NSX Edge 之间公布，并且仅在每个租户环境中单独使用。

## 可扩展的拓扑

前文中所述的服务提供商拓扑可进行横向扩展，如图 49 所示。该图左侧展示了 NSX Edge 为其提供服务的九个租户，右侧展示了 Edge 为其提供服务的另外九个租户。服务提供商可以轻松调配另一个 NSX Edge 来为更多的租户提供服务。

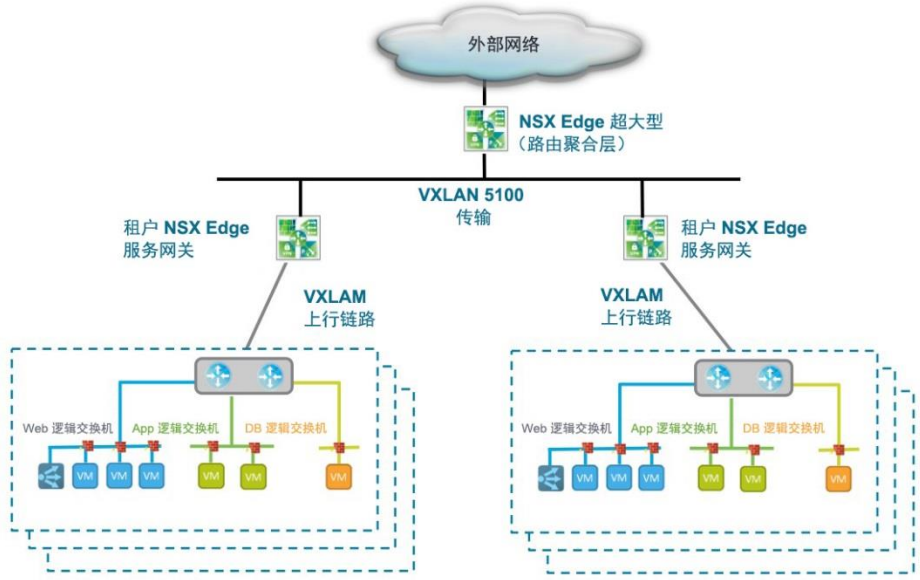



图 49 - 可扩展的拓扑

除了提供可扩展性之外，此模式的优势还在于，带有重叠 IP 地址的租户可以部署在连接到不同 NSX Edge 的独立组中。为了支持这一情形，必须在第一层 NSX Edge 上执行网络地址转换，以便确保当使用重叠 IP 地址的租户的流量流向聚合 NSX Edge 后，可以区分这些租户。

# 逻辑防火墙

建立逻辑防火墙是要为我们的多层部署示例添加的另一个功能要素，如下方图 50 所示。

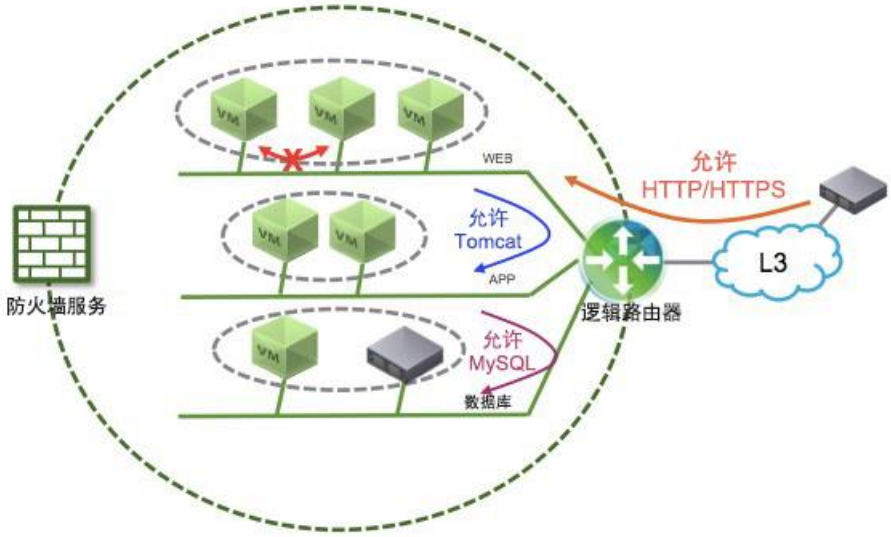



图 50 - 控制多层应用中的通信

VMware NSX 平台包含两种防火墙功能：一种是由 NSX Edge 服务网关提供的集中式防火墙服务，另一种是可在给定 NSX 域中的所有主机上以 VIB 软件包形式启用的分布式内核（类似于前文介绍的分布式路由功能）。第一章论述的分布式防火墙 (DFW) 提供的防火墙功能具有接近线速的性能、虚拟化功能、身份识别和活动监控功能以及网络虚拟化自带的其他网络安全功能。

# 网络隔离

隔离是大多数网络安全机制的基础，无论是为了确保合规性、控制力，还是仅仅用于防止开发、测试和生产环境之间发生交互。虽然过去使用物理设备上手动配置和维护的路由、ACL 和/或防火墙规则来建立和强制实施隔离，但多租户是网络虚拟化的固有功能。

虚拟网络（使用 VXLAN 技术）彼此隔离，并且默认情况下与底层物理网络隔离，从而实现最少特权安全原则。除非专门将虚拟网络连接在一起，否则它们以隔离方式创建并且始终保持隔离状态。无需使用任何物理子网、VLAN、ACL 以及防火墙规则，即可实现这种隔离。

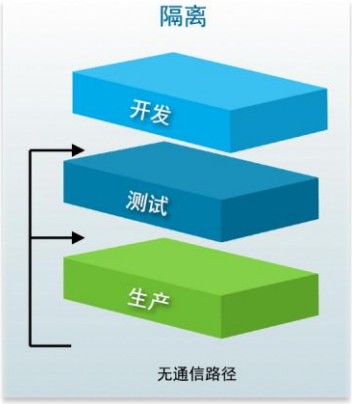



图 51 - 网络隔离

任何隔离的虚拟网络均可以由分布在数据中心任何位置的工作负载组成。同一虚拟网络中的工作负载可以驻留在相同或不同 hypervisor 中。此外，多个隔离的虚拟网络中的工作负载也可以驻留在相同 hypervisor 中。相关案例：虚拟网络之间的隔离允许重叠的 IP 地址，从而能够实现相互隔离的开发、测试和生产虚拟网络，每个网络都包含不同的应用版本，但却拥有相同的 IP 地址，所有网络均可同时运行，并且全部位于同一底层物理基础架构中。

虚拟网络还将与底层物理网络隔离。由于 hypervisor 之间的流量经过封装，因此运行物理网络设备的地址空间与工作负载用于连接到虚拟网络的地址空间完全不同。例如，虚拟网络可以基于 IPv4 物理网络支持 IPv6 应用工作负载。这种隔离可以帮助底层物理基础架构抵御由任何虚拟网络中的工作负载发起的任何可能的攻击。再次说明，这一切都不依赖于过去创建这种隔离所需的任何 VLAN、ACL 或防火墙规则。

# 网络分段

分段与隔离相关，但通常在多层虚拟网络中采用。过去，网络分段是物理防火墙或路由器的一项功能，旨在允许或拒绝流量在各网段或各层之间的传输。例如，对 Web 层、应用层和数据库层之间的流量进行网络分段。传统的分段定义和配置过程非常耗时，而且极易出现人为错误，从而会导致大量安全违规情况。实施过程需要用户在设备配置语法、网络寻址、应用端口和协议方面具有深入的专业技能。

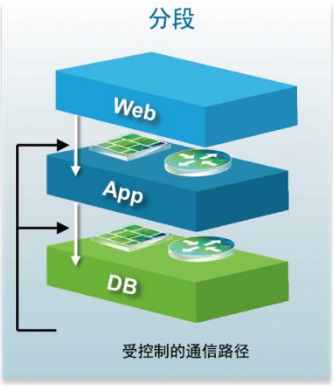



图 52 - 网络分段

与隔离一样，网络分段也是 VMware NSX 网络虚拟化的一项核心功能。虚拟网络可以支持多层网络环境，这意味着带有 L3 分段的多个 L2 网段或工作负载全部连接到单个 L2 网段的单层网络环境可以使用分布式防火墙规则。这两种方案都可以实现同一目标：对虚拟网络进行微分段，从而提供工作负载到工作负载的流量保护（也称为东西向保护）。

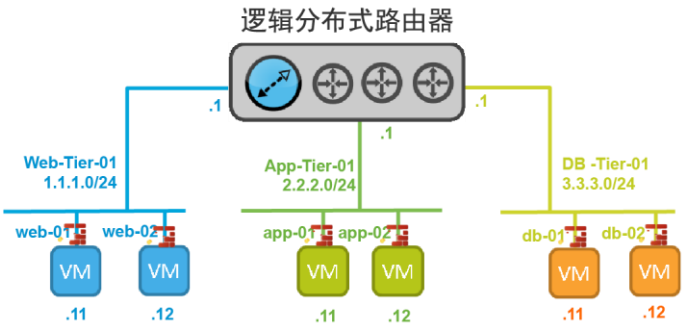



图 53 - 使用 NSX DFW 进行微分段 - 多层网络设计

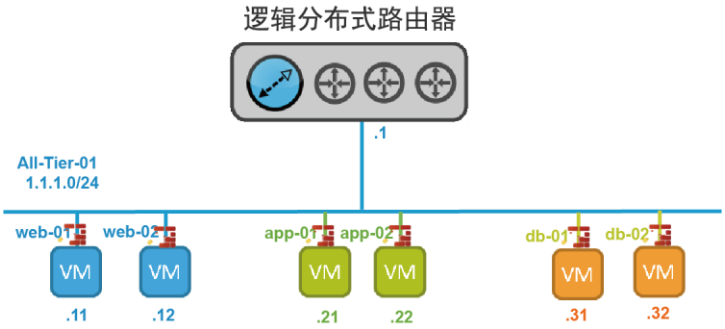



图 54 - 使用 NSX DFW 进行微分段 - 单层网络设计

物理防火墙和访问控制列表过去可提供成熟的、受网络安全团队和合规性审核员信任的分段功能。不过，人们对在云数据中心中采用这种方法的信心已经动摇，因为过时的手动网络安全调配和变更管理流程中的人为错误导致了越来越多的攻击和漏洞以及越来越长的停机时间。

在虚拟网络中，配备有工作负载的网络服务（L2、L3、防火墙等）以编程方式创建并分发到 hypervisor 虚拟交换机。包括 L3 分段和防火墙在内的网络服务均在虚拟机的虚拟接口处强制实施。

虚拟网络内的通信绝不会离开虚拟环境，因此无需在物理网络或防火墙中配置和维护网络分段。

## 充分利用抽象化

过去，要确保网络安全，安全团队需要深入了解网络寻址、应用端口、协议、与网络硬件绑定的所有项、工作负载位置和拓扑。网络虚拟化可对物理网络硬件和拓扑中的应用工作负载通信进行抽象化设置，从而使网络安全能够打破这些物理限制，并根据用户、应用和业务环境应用网络安全机制。

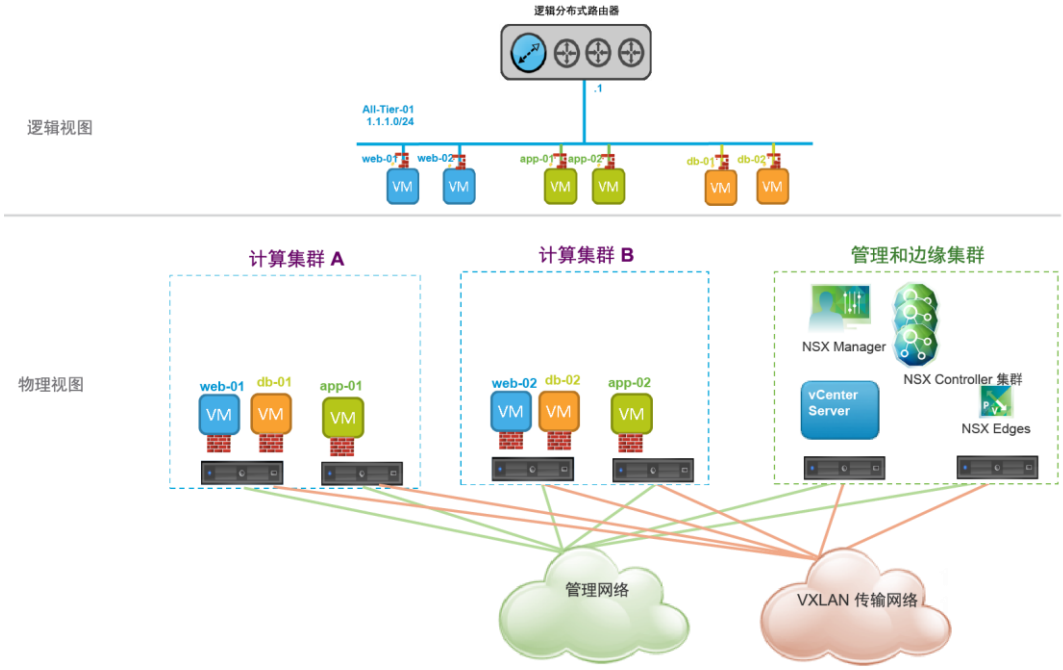



图 55 - 安全策略不再与物理网络拓扑紧密相关！

## 高级安全服务注入、串联和引导

NSX 网络虚拟化平台可提供有状态的 L2-L4 防火墙保护功能，以便在虚拟网络内实现分段。在某些环境中，需要更高级的网络安全功能。在这些情况下，客户可以利用 VMware NSX 在虚拟化网络环境中分发、启用和强制实施高级网络安全服务。NSX 将网络服务分发到虚拟网卡环境中，以便形成一个应用于虚拟网络流量的逻辑服务管道。可以将第三方网络服务注入此逻辑管道中，从而允许在逻辑管道中使用物理或虚拟服务。

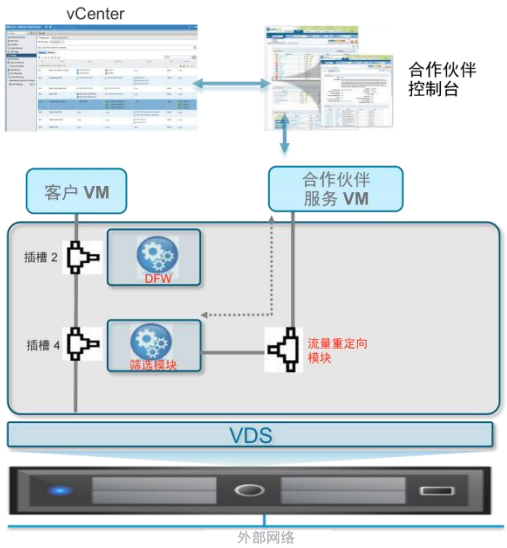



图 56 - 服务注入、串联和引导

在客户虚拟机与逻辑网络（逻辑交换机或支持 VLAN 的 DVS 端口组）之间，有一个在虚拟网卡环境中实现的服务空间。插槽 ID 可以具体表示与虚拟机的服务连接。如上方图 56 所示，插槽 2 分配到 DFW，插槽 4 分配到 Palo Alto Networks VM 系列 FW。另一组插槽可用于接入更多第三方服务。

从客户虚拟机传出的流量始终遵循插槽 ID 编号不断增大的路径（即，数据包先重定向到插槽 2，然后再到插槽 4）。传入客户虚拟机的流量则遵循插槽 ID 编号不断减小的路径传输（先到插槽 4，再到插槽 2）。

每个安全团队都会使用各种网络安全产品的独特组合来满足其环境的需求。目前，VMware 的整个安全解决方案提供商生态系统都在使用 VMware NSX 平台。网络安全团队经常面临着需要协调多个供应商提供的网络安全服务之间的关系这一挑战。NSX 方法的另一个强大优势是它能够构建策略来利用 NSX 服务的注入、串联和流量引导功能，以便基于其他服务的结果推动服务在逻辑服务管道中执行，从而能够协调多个供应商提供的本来毫不相关的网络安全服务。


**图 57 - 通过第三方供应商提供的高级服务实现的网络分段。**

例如，我们与 Palo Alto Networks 的集成将利用 VMware NSX 平台来分发 Palo Alto Networks 虚拟机系列的新一代防火墙，从而在每个 hypervisor 上以本地方式提供高级功能。那些为调配或转移到 hypervisor 的应用工作负载定义的网络安全策略，将插入虚拟网络的逻辑管道中。在运行时，服务注入功能将利用本地提供的 Palo Alto Networks 新一代防火墙功能集，在工作负载虚拟接口交付和强制实施基于应用、用户以及应用环境的控制策略。

## 跨物理和虚拟基础架构的一致的可见性和安全模式

VMware NSX 支持自动化调配，以及在各个虚拟和物理安全平台之间共享应用环境。通过在虚拟接口结合使用流量引导和策略实施功能，过去部署在物理网络环境中的合作伙伴服务现在可以在虚拟网络环境中轻松进行调配和实施，从而使 VMware NSX 可以跨驻留在物理或虚拟工作负载上的各个应用，为客户提供一致的可见性和安全模式。

1. **充分利用现有工具和流程。**显著提高了调配速度、运维效率和服务质量，同时确保了服务器、网络和安全团队的职责分离。
2. **加强对应用的控制，但不会产生不良影响。**过去，这种级别的网络安全性使网络和安全性团队不得不在性能和功能之间做出选择。通过在虚拟接口分发和强制实施高级功能集，就可以做到两全其美。
3. **减少因人为错误而导致的问题。**基础架构保留了相应的策略，使工作负载能够安置在数据中心内的任何位置并随意移动，且无需任何人工干预。用户可以采用编程方式应用预先批准的应用安全策略，从而让复杂的网络安全服务也可以实现自助部署。

# 通过 NSX DFW 实现的微分段用户场景和实施

现在，我们来探讨一下如何利用 DFW 实现三层应用（Web/应用/数据库）的微分段，其中多个组织（财务和人力资源）共享相同的逻辑网络拓扑。

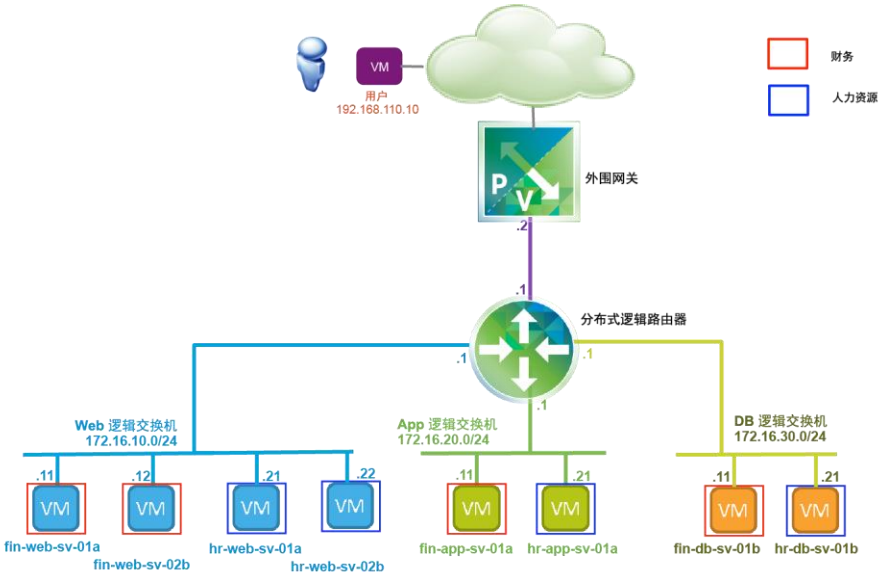



图 58 - 微分段用户场景

在这一网络设计中，3 个逻辑交换机（Web LS、App LS 和 DB LS）通过分布式逻辑路由器互连。在 DLR 上定义的 IP 地址是连接到任一逻辑交换机的虚拟机的默认网关。

DLR 所连接的 Edge 服务网关与物理环境相连接。终端用户（连接到外部网络）可以访问 Web 逻辑交换机子网（在 DLR 和 ESG 上启用的动态路由协议向 Web LS 子网告知相关实例）。

在本例中，财务和人力资源团队各拥有 2 台 Web 服务器、1 台应用服务器和 1 台数据库服务器。

这种微分段设计的特殊之处在于，角色相同的服务器连接到同一逻辑交换机，与它们所属的团队无关。人力资源和财务组织仅用作示例，更多其他组织也可以托管在这种逻辑网络基础架构之上。

借助 DFW，如果财务和人力资源工作负载连接到不同的逻辑网络，仍然可以实现相同的安全性级别。请注意，DFW 的强大功能在于，网络拓扑不再是安全机制实施的障碍：它支持任何类型的拓扑，并且对于其中任何一种，都可以实现相同级别的流量访问控制。

为了将角色相同以及组织相同的虚拟机分为一组，我们将利用 Service Composer 功能。

Service Composer 内的一个有用属性是安全组 (SG) 结构。SG 支持以动态方式或静态方式将对象包含在容器中。该容器将用作 DFW 安全策略规则的源或目标字段。

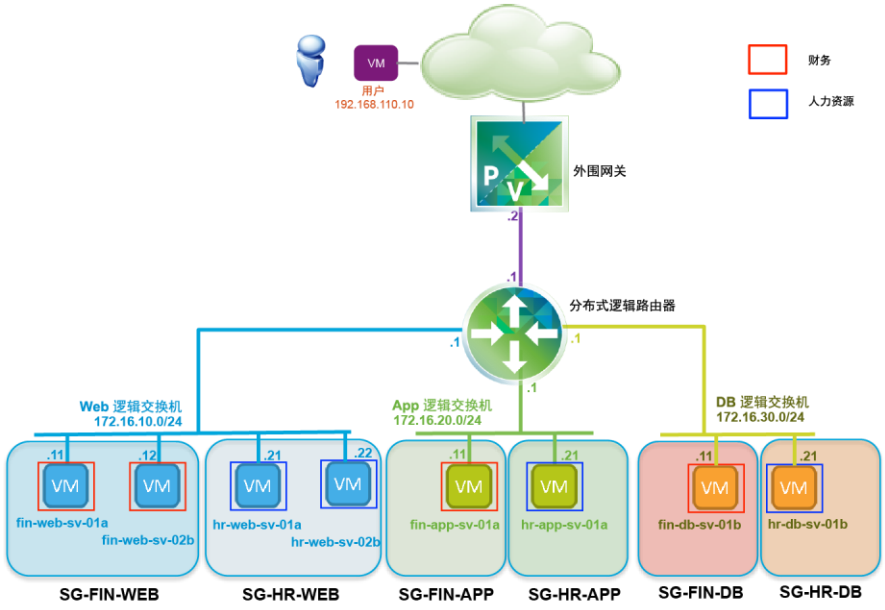



图 59 - 将虚拟机归入 Service Composer/安全组中

如上图所示，来自财务组织的 Web 服务器归入称为 SG-FIN-WEB 的安全组中。比如，为了构建此 SG，我们可以使用基于虚拟机名称或安全标记的动态包含：名称包含“fin-web”语法的虚拟机将自动包含在 SG-FIN-WEB 中，或者分配有安全标记“FIN-WEB-TAG”的虚拟机将自动添加到 SG-FIN-WEB 中。如果用户希望使用静态包含，可以手动将 fin-web-sv-01a 和 fin-web-sv-02b 添加到 SG-FIN-WEB 中。

在本设计中，总共有 6 个安全组：3 个用于财务组织，3 个用于人力资源组织（每层 1 个 SG）。

将虚拟机归入合适的容器是实现微分段的基础。完成后，现在就能够轻松实施流量策略，如下所示：

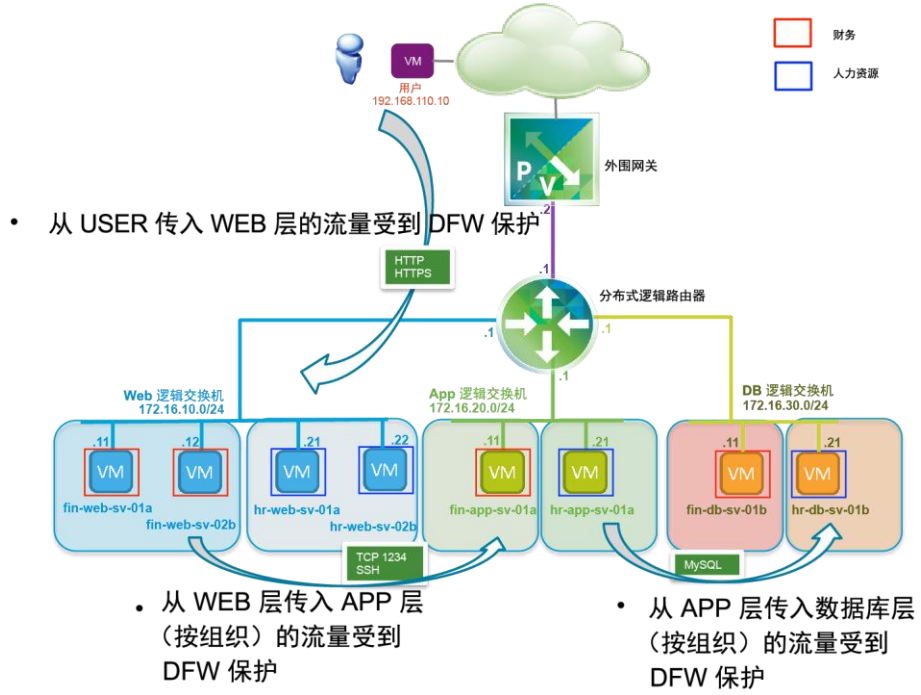



图 60 - 层间网络流量安全策略

层间网络流量安全策略实施方式如下：

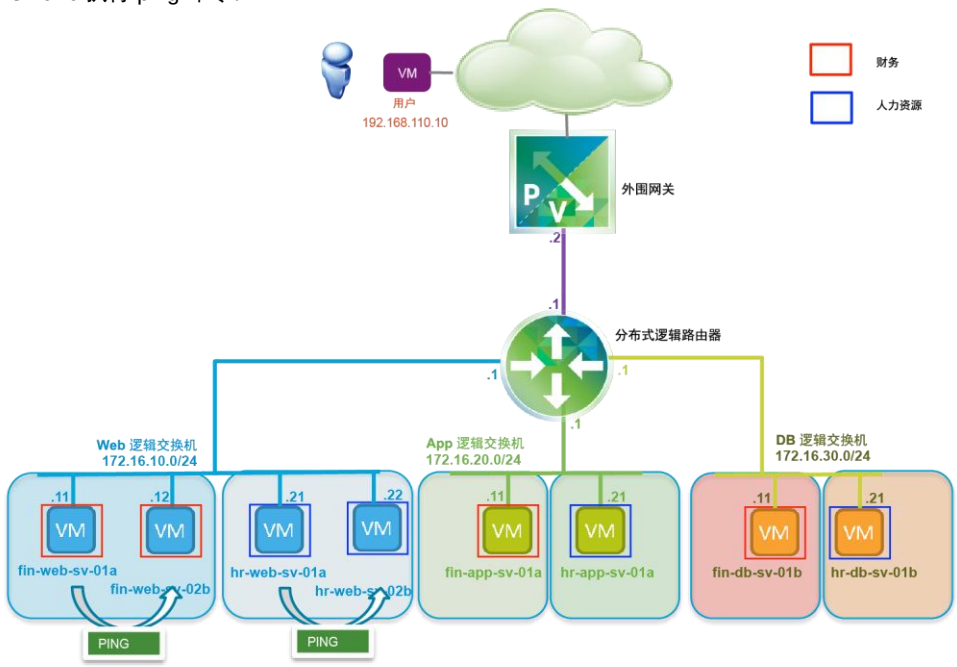
- 从 Internet 到 Web 服务器，允许 HTTP 和 HTTPS。丢弃其他流量。
- 从 Web 层到应用层，允许 TCP 1234 和 SSH（这些服务是简单示例）。丢弃其他流量。
- 从应用层到数据库层，允许 MySLQ。丢弃其他流量。

值得注意的重要一点是，就层间网络流量而言，财务组织并不能与人力资源组织通信。这一设计完全禁止两个组织在层间通信。

层内网络流量安全策略实施方式如下：

- 如果同一组织的服务器属于同一层，则可以互相执行 ping 命令。

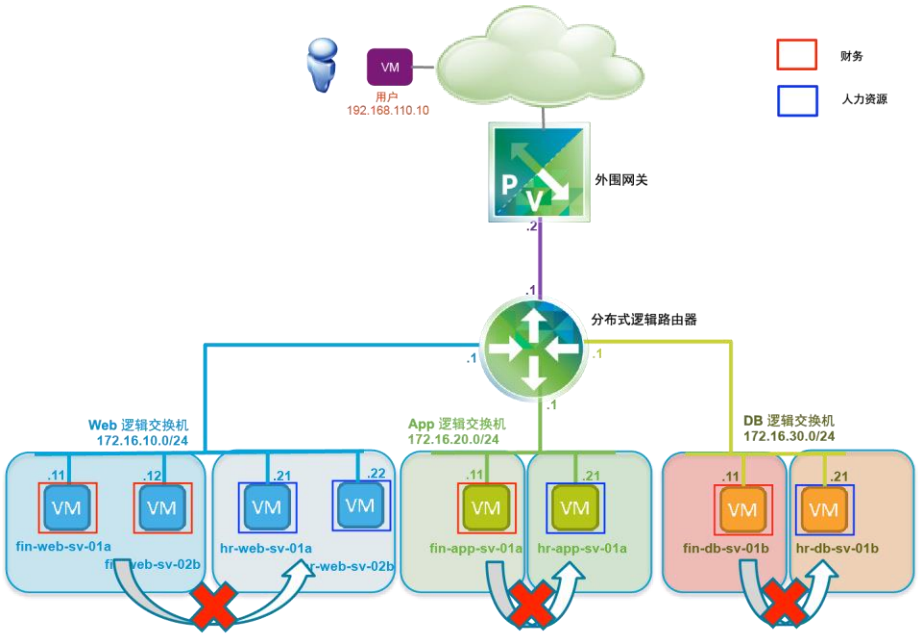
例如，fin-web-sv-01a 可以对 fin-web-sv-02b 执行 ping 命令。但是，fin-web-sv-01a 无法对 hr-web-sv-01a 或 hr- web-sv-02b 执行 ping 命令。




- 每个组织中允许执行 PING 操作的层内流量

**图 61 - 层内网络流量安全策略（同一组织）**

同样，位于同一层但来自不同组织的工作负载之间完全不得通信，如下面的图 62 所示。




- 跨组织阻止的层内流量

图 62 - 层内网络流量安全策略（跨组织）

为实现这一微分段用户场景，首先要执行的操作是将 DFW 默认策略规则从“Action = Allow”改为“Action = Block”。这是因为我们将使用积极安全模式，在此模式中，只有允许的流量才能在安全规则表中定义，其他通信则一概拒绝。

| 名称 | 源 | 目标 | 服务 | 操作 |
|---|---|---|---|---|
| Default rule | ANY | ANY | Any | Block |

图 63 - DFW 默认策略规则

**注意**：如果 vCenter Server 托管在已为 VXLAN 准备好的 ESXi hypervisor 上，那么对于这种部署，DFW 默认策略规则同样适用于 vCenter 的虚拟网卡。为了避免丢失与 vCenter 的连接，建议将 VC 加入 DFW 排除列表（用于避免针对特定虚拟机强制实施 DFW 策略）。
此外，还可以创建专用 DFW 规则来指定允许与 vCenter 通信的 IP 子网。请注意，相同的注意事项仍适用于其他管理产品（vCAC、vCOPS 等），它们将以每个虚拟网卡为基础获取 1 个 DFW 实例（只有 NSX Manager 和 NSX Controller 自动从 DFW 策略实施中排除）。

因此，控制层间流量的安全策略规则可以按照图 64 所示进行定义。

| 名称 | 源 | 目标 | 服务 | 操作 |
|---|---|---|---|---|
| FIN - INTERNET to WEB Tier | ANY | SG-FIN-WEB | HTTP<br>HTTPS | Allow |
| HR - INTERNET to WEB Tier | ANY | SG-HR-WEB | HTTP<br>HTTPS | Allow |
| FIN - WEB Tier to APP Tier | SG-FIN-WEB | SG-FIN-APP | TCP-1234<br>SSH | Allow |
| HR - WEB Tier to APP Tier | SG-HR-WEB | SG-HR-APP | TCP-1234<br>SSH | Allow |
| FIN - APP Tier to DB Tier | SG-FIN-APP | SG-FIN-DB | MYSQL | Allow |
| HR - APP Tier to DB Tier | SG-HR-APP | SG-HR-DB | MYSQL | Allow |

图 64 - DFW 层间策略规则

最后，对于层内通信，我们将实施如图 65 所示的安全策略规则。

| 名称 | 源 | 目标 | 服务 | 操作 |
|---|---|---|---|---|
| FIN – WEB Tier to WEB Tier | SG-FIN-WEB | SG-FIN-WEB | ICMP echo<br>ICMP echo reply | Allow |
| HR – WEB Tier to WEB Tier | SG-HR-WEB | SG-HR-WEB | ICMP echo<br>ICMP echo reply | Allow |
| FIN – APP Tier to APP Tier | SG-FIN-APP | SG-FIN-APP | ICMP echo<br>ICMP echo reply | Allow |
| HR – APP Tier to APP Tier | SG-HR-APP | SG-HR-APP | ICMP echo<br>ICMP echo reply | Allow |
| FIN – DB Tier to DB Tier | SG-FIN-DB | SG-FIN-DB | ICMP echo<br>ICMP echo reply | Allow |
| HR – DB Tier to DB Tier | SG-HR-DB | SG-HR-DB | ICMP echo<br>ICMP echo reply | Allow |

图 65 - DFW 层内策略规则

# 逻辑负载均衡

负载均衡是 NSX 内的另一项网络服务，可在 NSX Edge 设备本机上启用。部署负载均衡器的两个主要推动因素包括扩展应用（通过在多个服务器之间分配工作负载），以及改善高可用性特征。

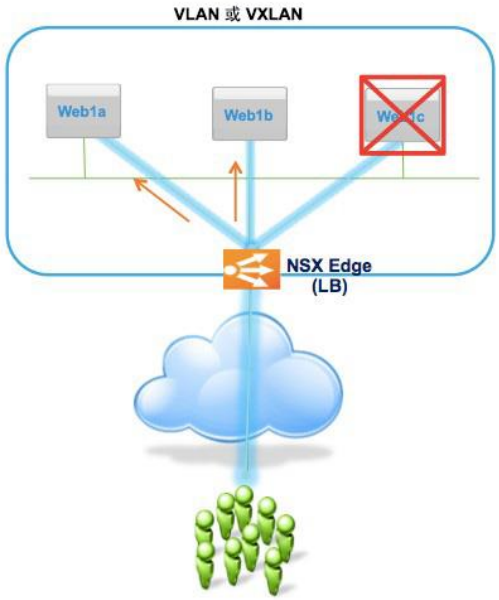



图 66 - NSX 负载均衡

NSX 负载均衡服务专为具有以下特征的云环境设计：

- 可完全通过 API 编程
- 与其他 NSX 网络服务相同的单一管理/监控中心点

NSX Edge 本身提供的负载均衡服务可满足大多数应用部署的需求。这是因为 NSX Edge 提供了大量功能：

- 支持任何 TCP 应用，包括但不限于 LDAP、FTP、HTTP、HTTPS
- NSX SW 版本 6.1 及更高版本支持 UDP 应用。
- 提供多种负载均衡分配算法：轮询调度、最少连接数、源 IP 哈希、URI
- 多重运行状况检查：TCP、HTTP、HTTPS（包含内容检测）
- 持久性：源 IP、MSRDP、Cookie、SSL 会话 ID
- 连接限制：最大连接数和每秒连接数
- L7 操控，包括但不限于 URL 阻止、URL 重写、内容重写
- 通过支持 SSL 负载分流实现的优化功能

**注意**：NSX 平台还可以集成由第三方供应商提供的负载均衡服务。此类集成不在本设计指南的讨论范围之内。

就部署而言，NSX Edge 为两种模式提供支持：

- 单路并联模式（称为代理模式）：此场景如图 67 所示，涉及到 NSX Edge 的一种部署情况，即 NSX Edge 直接连接到要为其提供负载均衡服务的逻辑网络。

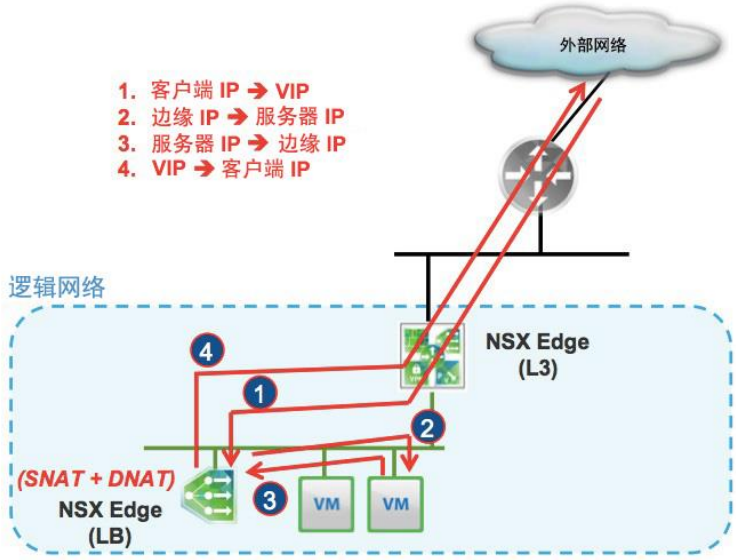



图 67 - 单路并联模式负载均衡服务

单路并联负载均衡器功能如上图所示：

1. 外部客户端将流量发送至负载均衡器提供的虚拟 IP 地址 (VIP)。
2. 负载均衡器对从客户端接收的原始数据包执行两个地址转换操作：目标 NAT (D-NAT) 用于将 VIP 替换为服务器群中部署的服务器之一的 IP 地址，源 NAT (S-NAT) 用于将客户端 IP 地址替换为标识负载均衡器自身的 IP 地址。S-NAT 是必需的，从而强制服务器群的返回流量先经过负载均衡器，再到达客户端。
3. 服务器群中的服务器通过向负载均衡器发送流量进行回复（由于前述 S-NAT 功能）。
4. 负载均衡器将其 VIP 用作源 IP 地址，再次执行源和目标 NAT 服务以将流量发送至外部客户端。

此模式的优势在于部署过程较为简单且灵活，因为它允许直接在需要负载均衡器服务（NSX Edge 设备）的逻辑分段上部署这些服务，而无需对向物理网络提供路由通信的集中式 NSX Edge 进行任何修改。此模式的缺点在于它要求调配更多的 NSX Edge 实例，还要求对禁止数据中心内服务器洞悉原始客户端 IP 地址的源 NAT 进行部署。

**注意**：负载均衡器可以先将客户端的原始 IP 地址插入到 HTTP 标头中，然后再执行 S-NAT（这项功能称为“Insert X-Forwarded-For HTTP 标头”）。这样服务器便可洞悉客户端 IP 地址，但这仅限于 HTTP 流量。

- 串联模式（称为透明模式）要求部署与传输到服务器群的流量串联的 NSX Edge。此类工作方式如图 68 所示。

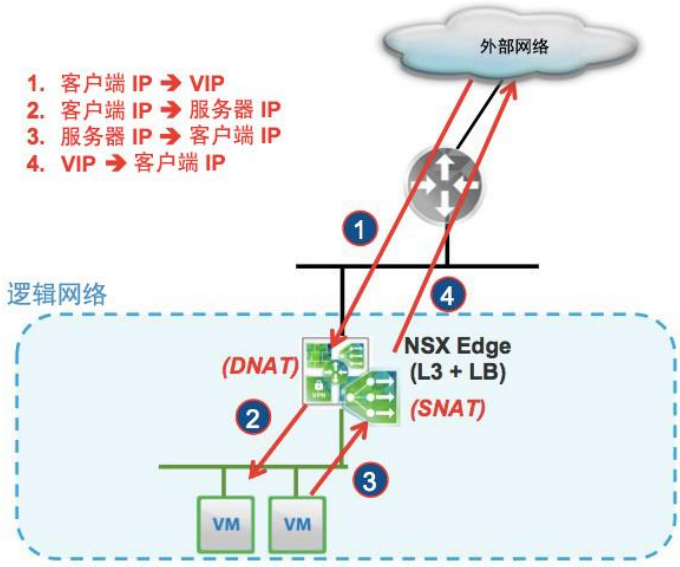



图 68 - 双路串联模式负载均衡服务

1. 外部客户端将流量发送至负载均衡器提供的虚拟 IP 地址 (VIP)。
2. 负载均衡器（集中式 NSX Edge）仅执行目标 NAT (D-NAT)，将 VIP 替换为服务器群中部署的服务器之一的 IP 地址。
3. 服务器群中的服务回复原始客户端 IP 地址，负载均衡器再接收一次流量，因为它以串联方式部署（并且通常作为服务器群的默认网关）。
4. 负载均衡器将其 VIP 作为源 IP 地址，执行源 NAT 以将流量发送至外部客户端。

此类部署模式也很简单，并且允许服务器全面洞悉原始客户端 IP 地址。同时，从设计的角度来看，此模式不太灵活，这是因为，它通常强制使用负载均衡器作为部署服务器群的逻辑分段的默认网关，这表明必须为这些分段采用集中式（而非分布式）路由。还要务必注意的是，在这种情况下，负载均衡器是添加至 NSX Edge 的另一逻辑服务，后者已在逻辑网络与物理网络间提供路由。因此，建议将 NSX Edge 的规格增至超大型，然后再启动负载均衡服务。

就可扩展性和吞吐量而言，每个 NSX Edge 提供的 NSX 负载均衡服务都可以纵向扩展到以下目标（在最佳情况下）：

- 吞吐量：9 Gbps
- 并发连接数：100 万个
- 每秒新连接数：13.1 万个

图 69 中展示了一些租户部署示例，这些租户具有不同的应用和不同的负载均衡需求。注意，其中的每个应用通过 NSX 提供的网络服务托管在相同的云环境中。

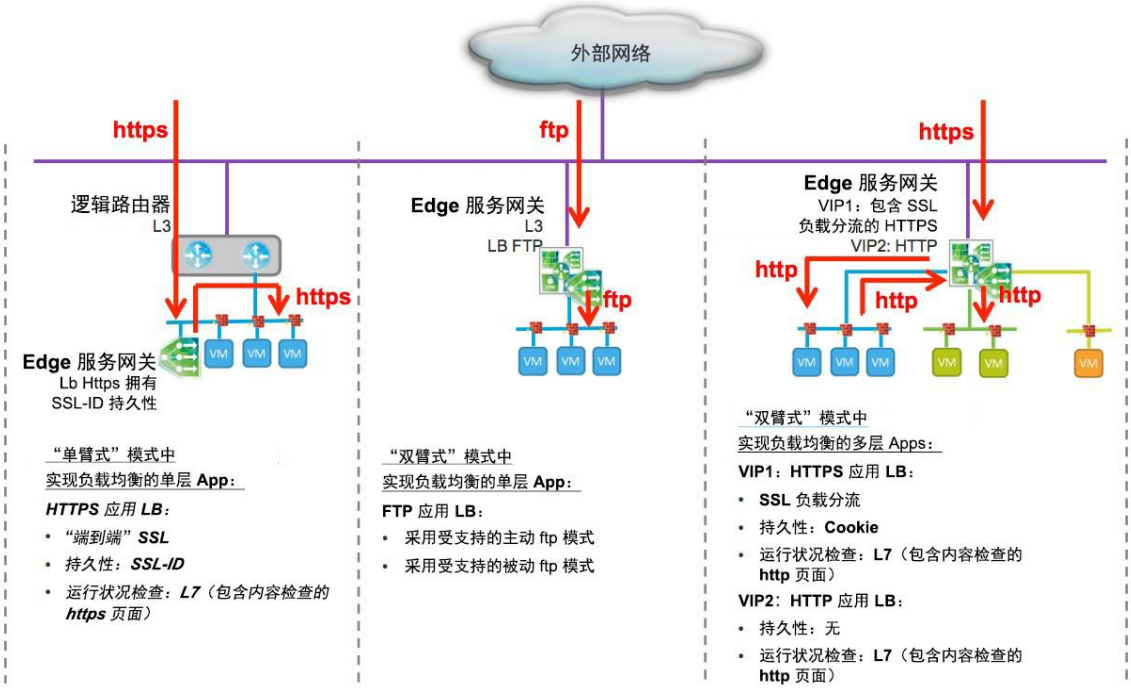



图 69 - NSX 负载均衡部署示例

最后需要关注的两个重点：

- 负载均衡服务可以完全分布在多个租户中。这会带来以下好处：
  - 每个租户都有自己的负载均衡器。
  - 每个租户配置的更改都不会影响其他租户。
  - 一个租户负载均衡器上的负载增加不会影响其他租户负载均衡器的规模。
  - 每个租户负载均衡服务都可以纵向扩展到上文所述的上限。
- 其他网络服务仍然完全可用
  - 同一租户可以将其负载均衡服务与路由、防火墙、VPN 等其他网络服务组合在一起使用。

# 虚拟专用网络 (VPN) 服务

本设计指南讨论的最后两个 NSX 逻辑网络功能都属于虚拟专用网络 (VPN) 服务。分为以下两个类别：L2 VPN 和 L3 VPN。

# L2 VPN

通过部署 L2 VPN 服务，可以跨两个单独的数据中心位置延展 L2 连接。

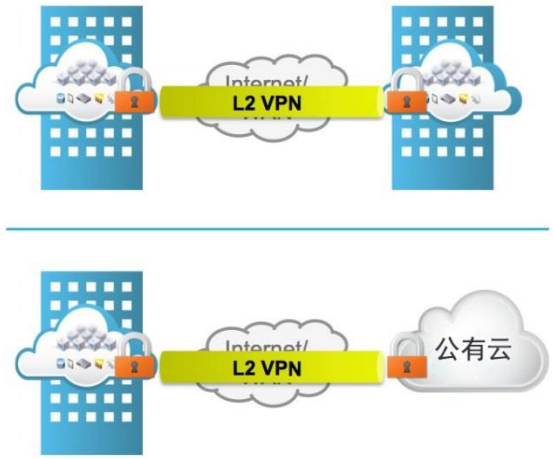



图 70 - NSX L2 VPN 用户场景

无论是对于企业部署还是 SP 部署，许多用户场景都可以从此功能中受益：

- 企业工作负载迁移/数据中心整合
- 服务提供商租户加入系统
- 云突发（混合云）
- 延伸应用层（混合云）

图 71 重点介绍了如何在独立数据中心站点内部署的一对 NSX Edge 设备之间创建 L2 VPN 安全加密链路。

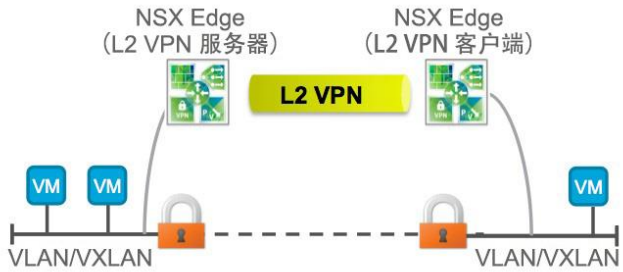



图 71 - 面向 L2 VPN 服务的 NSX Edge

对于这一特定部署，下面列出了一些注意事项：

- L2 VPN 连接是将两个位置的独立网络连接起来的 SSL 安全加密链路。正是因为可以将建立连接的独立网络可连接至同一地址空间（IP 子网）这一特点，此服务被称为 L2 VPN 服务。
- 本地网络可为任何性质，如 VLAN 或 VXLAN；L2 VPN 服务还可以互连不同性质的网络（一个站点是 VLAN，另一个站点是 VXLAN）。
- L2 VPN 当前只是可在两个位置间建立的点对点服务。部署在一个数据中心站点的 NSX Edge 充当 L2 VPN 服务器，而另一个站点的 NSX Edge 是用于发起与服务器的连接的 L2 VPN 客户端。
- NSX L2 VPN 通常部署在互连站点的网络基础架构上，后者由服务供应商提供或为企业所有。独立于网络的所有者和管理者，无论是延迟和带宽方面，还是 MTU 方面，都没有任何特定要求。构建 NSX L2 VPN 解决方案时加强了稳定性，以便轻松应对任何可用网络连接。

NSX 6.1 软件版本为 L2 VPN 解决方案带来许多改进。其中最相关的改进包括：

- 在 6.0 版本中，必须在需要建立连接的两个站点中部署两个独立的 NSX 域。

这表明需要在每个位置中部署独立的 vCenter、NSX Manager 和 NSX Controller 集群，而这可能会成为一个问题，尤其对于服务提供商部署（例如，对于混合云用户场景）。从 NSX 6.1 软件版本开始，允许进行远程 NSX Edge 部署（充当 L2 VPN 客户端），无需在远程站点使用 NSX，从而在根本上允许将解决方案的范围扩展到仅限 vSphere 的客户。

- 除上行链路接口和内部接口之外，NSX 6.1 版本还引入了第三种类型的 NSX Edge 接口，称为中继接口。利用中继接口可以在每个站点上部署的多个网络（支持 VLAN 或 VXLAN 的端口组）之间延伸 L2 连接（在 6.0 中，网络的扩展范围限制为每个 NSX Edge 仅一个 VLAN/VXLAN）。
- 从 6.1 开始，为部署为 L2 VPN 服务器或客户端的 NSX Edge 提供全面的高可用性支持。因此，可以在每个站点中部署一对配置为“活动/备用”模式的 NSX Edge。

## L3 VPN

最后，L3 VPN 服务用于提供从远程位置到数据中心网络的安全 L3 连接。

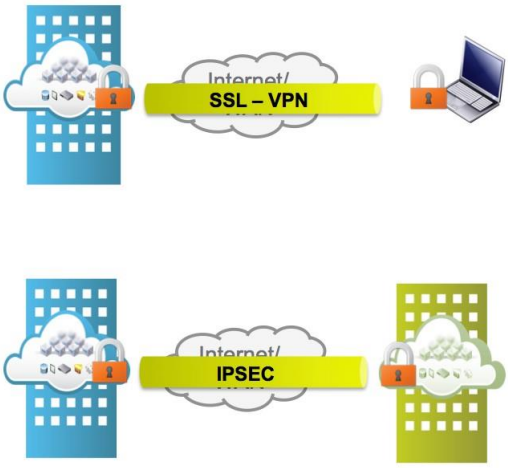



图 72 - NSX L3 VPN 服务

如图 72 所示，使用 SSL 安全加密链路的远程客户端可借助 L3 VPN 服务安全连接至 NSX Edge 网关（充当数据中心内的 L3 VPN 服务器）后面的专用网络。此服务通常被称为“SSL VPN-Plus”。

此外，还可以对 NSX Edge 进行部署，使它能够使用标准化 IPSec 协议设置与所有主要的物理 VPN 供应商设备实现互操作，然后建立站点到站点的安全 L3 连接。可将多个远程 IP 子网连接至 NSX Edge 背后的内部网络。不同于前面所述的 L2 VPN 服务，这种情况下的连接可进行路由，因为远程子网与本地子网属于不同的地址空间。还有一点要注意，在当前的实施中，各 IPSec 连接之间支持仅限单播的通信，因此无法使用动态路由协议，而需要进行静态路由配置。

# NSX-v 设计注意事项

在这一部分中，我们将讨论如何基于常用数据中心网络拓扑，部署 NSX-v 使用 VXLAN 叠加技术实现的逻辑网络。首先解决物理网络的需求，然后考虑关于部署 NSX-v 网络虚拟化解决方案的设计注意事项。

## 数据中心网络设计的演变

物理网络应具备哪些属性才能为网络虚拟化部署提供稳定的 IP 传输？在讨论此问题之前，有必要先简要说明一下数据中心网络的当前趋势和设计演变。就在几年以前，大部分数据中心网络还专门采用模块化形式构建，如图 73 所示。

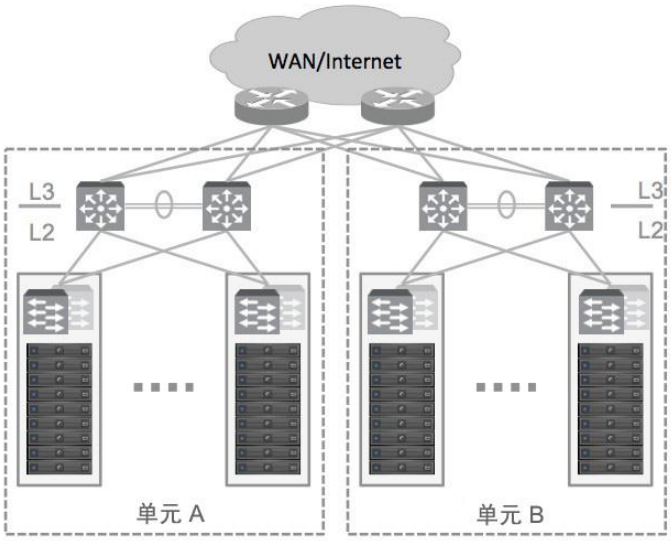



图 73 - 经典访问模式/聚合/核心数据中心网络

上面所示的模块化设计展示了一种可扩展方法，其中多个单元可以通过 L3 数据中心核心互连。每个单元都表示一个聚合块，限制数据中心结构内 VLAN 的扩展。每个单元的聚合层设备实际上是 L2 和 L3 网络域之间的分界线，这样做是为了限制 L2 广播和故障域范围的扩大。这种方法的直接后果是所有需要 L2 邻接关系的应用（应用集群、虚拟机实时迁移等）都必须在给定单元内部署。此外，如果大部分流量以南北向交互，并且仅限于单元间通信，这一经典设计通常也具有非常高的效率。

以上所述的经典模块化方法在过去几年中演化为“分支-主干”设计（图 74）。

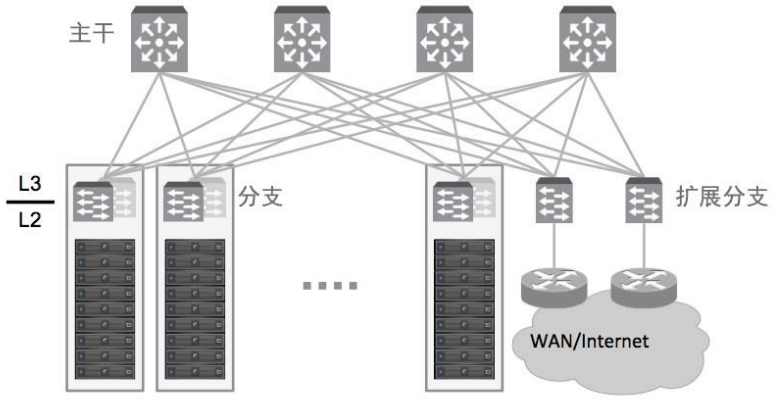



图 74 - “分支-主干”结构设计

以上所示的网络通常采用“folded Clos”设计，一般由具有以下需求的客户进行部署：数据中心结构能够具有高度可扩展性特征，但功能数量要有所减少。Clos 结构中的每台设备都采用分支或主干的名称，具体取决于它在哪层部署：

**分支层**：分支层可确定超额预定，因此可能影响“主干”所需的规模。在本层中的设备通常采用较为简单的设计，即“移动部分”较少，以便高效转发数据包，同时准确学习网络图。

**主干层**：主干层负责互连所有“分支”。“主干”内的各节点通常并未互相连接，也未在彼此之间形成任何路由协议邻接关系。主干设备的特征通常在于简单且极少的配置，因为它们负责执行快速流量交换，在本质上充当结构“底板”。

**边界分支层**：在某些场景中，主干设备可能用于直接连接到外部环境，但它们通常用于将传输到外部的流量转发到专门充当边界分支的一对分支节点。边界分支可能注入结构默认路由以吸引针对外部目标的流量。

这一设计演变由两个主要需求推动：不断增多的东西向通信；希望能够在数据中心结构中部署应用，但不再受到每个单元内 L2 域扩展上限的限制。

解决第一个需求的方法是，增加每个分支所连接主干设备的数量：如上图所示，为每个分支提供多个等价路径，以便增加可用于东西向通信的带宽，并且可预测和确定这些流量出现的延迟。

解决第二个需求的方法是，创建跨整个数据中心结构的更大 L2 域。不同交换机供应商提出多种替代解决方案建议来构建大型 L2 结构网络，不再将生成树用作控制层协议，例如 Cisco 的专有 FabricPath 技术（与之类似的有当前作为 IETF 草案的 Transparent Interconnection of Lots of Links (TRILL)），或 Juniper 的 QFabric 解决方案。

但是，越来越多的客户都转向“分支-主干”路由结构网络部署，其中 L2 和 L3 之间的边界下移到分支（或架顶式）设备层（如前面的图 74 所示）。

这种路由结构设计同样提供了较低的超额预定、较高的东西向带宽可用性，以及可预测性和配置一致性。此外，还加强了设备互操作性（因为不同供应商的交换机都可以互相连接在一起）和整体恢复能力。考虑到 L2 域的扩展限制在每台分支交换机之后，路由设计的主要挑战在于无法支持以下应用：要求在不同节点之间具有 L2 邻接关系的应用。这样一来，NSX 和网络虚拟化便有了用武之地：逻辑网络和物理网络之间的松耦合可在逻辑空间内提供任何类型的连接，而不依赖于底层网络的配置方式。

因为预计路由结构网络的部署将成为客户的首要选择，所以本文剩余部分的重点在于数据中心路由访问设计，其中分支/访问节点执行完全的第 3 层功能。

另一方面，我们还预计客户需要慢慢从原始多单元部署迁移到纯分支/主干体系结构。得益于虚拟网络和物理网络的松耦合，NSX 让迁移过程变得轻松无比，它能够无缝创建 IT 资源的公用池，并且在多种不同类型的网络之间建立一致的网络服务模型（图 75）。

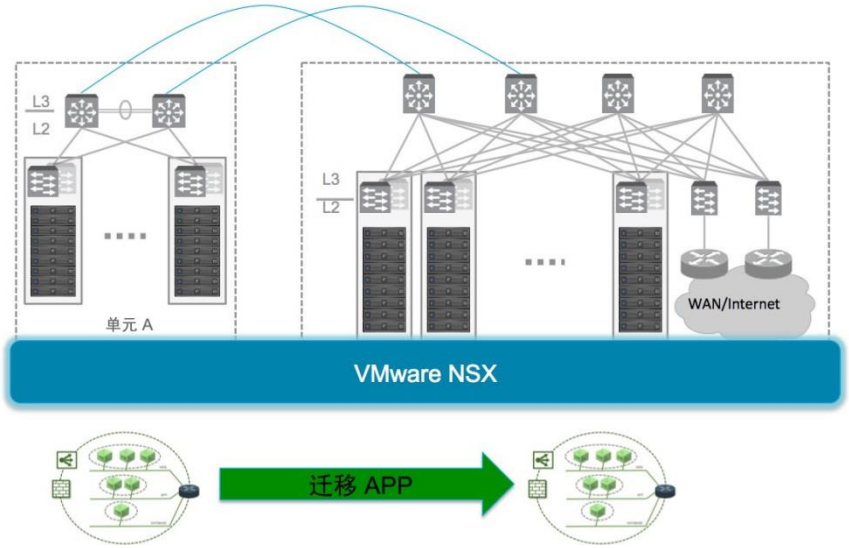



图 75 - 跨数据中心单元迁移应用

# 物理网络要求

NSX 的主要目标之一及其网络虚拟化功能是实现虚拟到物理网络抽象化和松耦合。但是，为了成功部署，物理结构必须具备以下属性才能提供稳定的 IP 传输：

- 简易性
- 可扩展性
- 高带宽

- 容错
- 服务质量 (QoS)

以下部分详细介绍了这些参数。

**注意：**在以下讨论中，术语访问层交换机、架顶式 (ToR) 交换机和分支交换机可以互换使用。分支交换机（或者更确切地是，成对的冗余分支交换机）通常位于机架内，并且支持通过网络访问连接到该机架的服务器。术语聚合和主干层（有效提供机架之间的连接）是指聚合所有访问交换机的网络中的位置。

# 简易性

在数据中心内组成整体结构的交换机的配置必须非常简单。常规或全局配置（如 AAA、SNMP、SYSLOG、NTP 等）应当几乎逐行复制，与交换机的位置无关。

## 分支交换机

面向机架内服务器的端口应采用最低配置。随着在 ESXi 主机上采用 10 GE 接口，使用 801.Q 中继接口传输少量 VLAN 流量这一做法变得更为流行；例如 VXLAN 流量、管理流量、存储流量和 VMware vSphere vMotion 流量。分支交换机终止，并为每个 VLAN 分别提供默认网关功能；即，它为每个 VLAN 都提供一个交换机虚拟接口 (SVI)。

成对的分支交换机通常部署在机架内，以便为服务器提供冗余连接点。

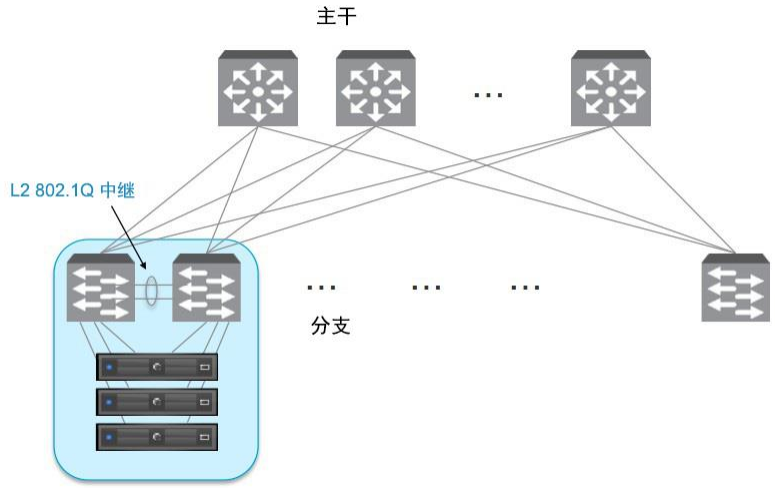



图 76 - 部署成对的冗余分支节点

如图 76 所示，此对分支设备之间需要直接连接，通常配置为 802.1Q L2 中继接口，用于传输机架中的服务器所使用的不同的 VLAN。对此 L2 中继接口的需求与特定服务器的上行链路配置无关（如“NSX-v 域中的 VDS 上行链路连接”部分中所述）。

从分支交换机到聚合或主干层的上行链路是点对点路由的链路，如图 77 所示。

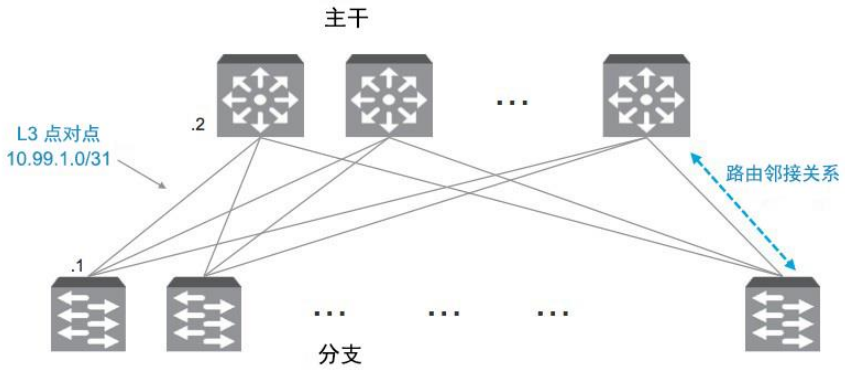



图 77 - 分支交换机和主干交换机之间的 L3 连接

不允许在分支和主干之间的链路上设置 VLAN 中继接口，甚至对单个 VLAN 也不允许。在分支和主干交换机之间配置动态路由协议，例如 OSPF 或 eBGP。机架中的每个 ToR 交换机都会公布一小组前缀，通常每个 VLAN 或它所在的子网各对应一个前缀。相应地，它将计算从其他 ToR 交换机接收的前缀的等价路径。

**注意**：在利用封装技术时，增加在 ESXi 主机上和部署在物理网络中的设备的所有接口上受支持的 MTU 很重要。VXLAN 向每个原始帧添加 50 字节，因此通常建议将 MTU 大小增加到至少 1600 字节的值。在从 NSX Manager UI 中将 VXLAN 调配到主机本身时，ESXi 服务器上的 MTU（用于内部逻辑交换机和 VDS 上行链路）将自动进行调整（默认为 1600 字节）。

### 主干交换机

主干交换机只有连接到分支交换机的接口；所有接口都配置为点对点路由的链路，有效充当分支交换机的点对点上行链路的“另一端”。

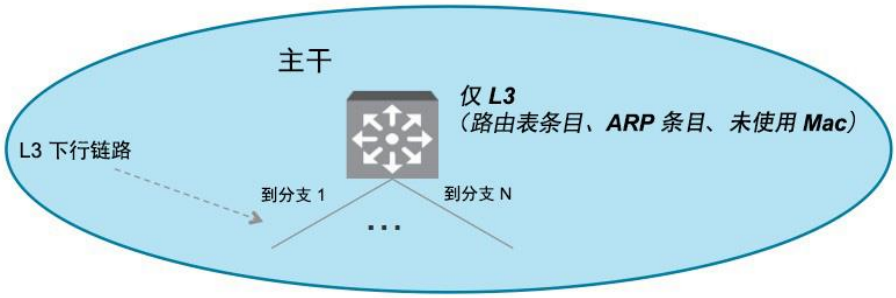



图 78 - 主干交换机接口

不需要在主干交换机之间设置链路。如果主干交换机与分支交换机之间出现链路故障，路由协议会确保受影响机架的流量不会流入与该机架断开连接的主干交换机。

## 可扩展性

关于规模的因素包括结构中支持的机架数量、数据中心内任两个机架之间的带宽数量、分支交换机在与另一个机架通信时可选择的路径数量等。

所有主干交换机上的可用端口总数指示了结构中受支持的机架数量和可接受的超额预定。“高带宽”部分中提供了更多详细信息。

不同的机架可能会托管不同类型的基础架构。可以识别三种类型的机架功能（在“NSX-v 部署注意事项”部分中有更详细的讨论）：计算机架、边缘机架和基础架构机架。

使设计保持模块化可确保向各种机架提供的网络连接具有灵活性。这很重要，因为不同类型的机架可能具有不同的带宽要求（例如存储文件管理器驱动更多流量传入和传出基础架构机架）。因此，为了满足不同的带宽需求，链路速度以及链路数量可能也各异。无需牺牲任何主干交换机或分支交换机的体系结构，即可为每个机架完成上述设置。

给定分支节点和主干交换机之间的链路数量体现了来源于此机架的流量可以采用的物理路径数量。由于任意两个机架之间的跳数是一致的，所以可使用等价多路径 (ECMP)。假设来自服务器的流量携带 TCP 或 UDP 标头，则每个流都可能发生流量喷射。正如之前在“VXLAN 入门”中提到的，这也适用于 VXLAN 流量，将计算 UDP 源端口值，该值提供 L2/L3/L4 标头的哈希，该哈希最终呈现在得到封装的原始 L2 帧中。这意味着即使有一些 ESXi hypervisor 曾交换来源于单个 VTEP IP 地址并从该地址接收的 VXLAN 流量，也可以基于在连接到逻辑网络的工作负载之间交换的原始流，在整个结构中对此流量进行负载均衡。

## 高带宽

在主干-分支交换机拓扑中，通常会在分支交换机这个点上发生超额预定（如果发生）。计算很简单：与给定分支交换机连接的所有服务器可用的带宽总量除以上行链路的带宽总量便可得出超额预定。

例如，具有单个 10 GB 以太网 (10 GbE) 端口的 20 台服务器，每台可创建最多每秒 200 GB 的带宽。假设主干有八个 10 GbE 上行链路（总共每秒 80 GB），这将得出超额预定系数为 2.5:1。

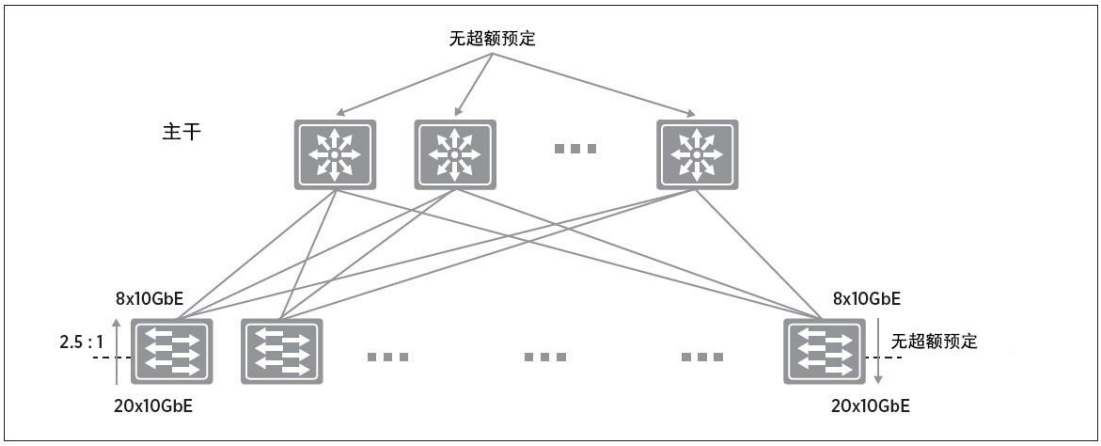



图 79 - 分支-主干拓扑的超额预定示例

如之前部分中所述，根据机架的功能，可以通过调配更多或更少的上行链路来增加或减少机架可用的带宽。换句话说，每个机架的超额预定水平可能有所不同。

从体系结构的观点来说，必须遵守一条规则：从一个分支交换机连接到每个主干交换机的上行链路的数量必须相同。这意味着，主干交换机有两条上行链路而主干交换机 B、C 和 D 只有一条上行链路是糟糕的设计，因为将有“更多”的流量通过主干交换机 A 发送到分支交换机，从而有可能创建一个热点。

## 容错

环境越大、构成整体结构的交换机越多，数据中心交换结构中的单个组件发生故障的可能性就越大。构建恢复能力强的结构的理由是，它可以承受个别链路或机盒故障，而不会产生大面积影响。

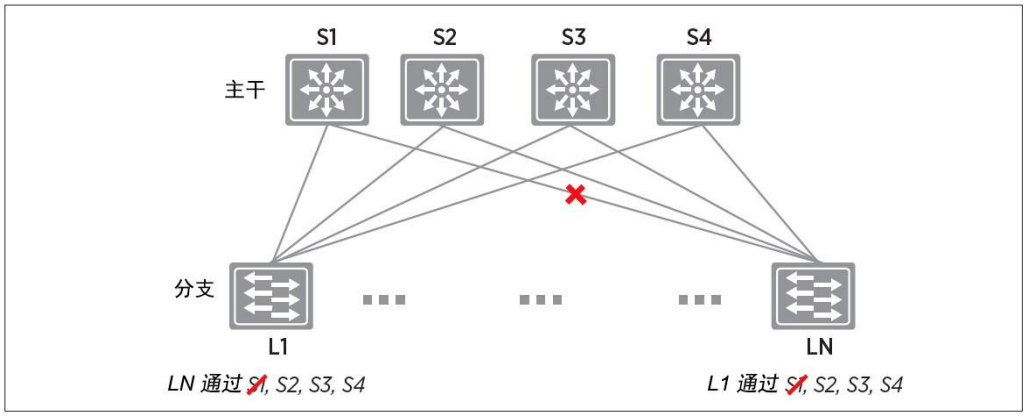



图 80 - 分支-主干拓扑中的链路故障场景

举例来说，如果一条链路或主干交换机之一发生故障，那么机架之间的流量会继续通过其余主干交换机在 L3 结构中路由。对于 L3 结构，路由协议可确保仅选择剩余路径。并且由于可以安装两台以上主干交换机，可降低主干交换机故障所造成的影响。

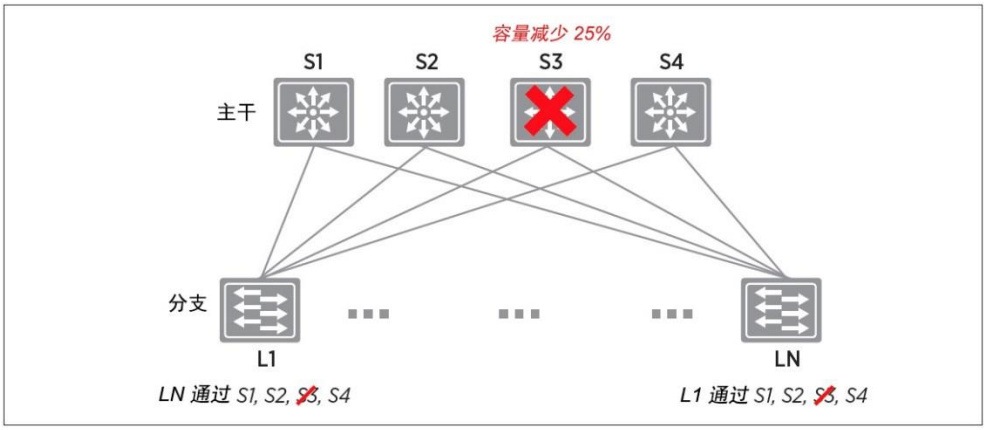



图 81 - 主干交换机故障场景和对带宽的影响

支持多路径的结构可处理机盒或链路故障，从而减少手动网络维护和操作的需要。如果必须在结构交换机中执行软件升级，那么可以通过更改路由协议指标来逐渐使节点停止运行；很快，通过该交换机的流量将消失，从而使它有空闲进行维护。根据主干的宽度（即，聚合或主干层中有多少交换机），其余交换机必须承担的额外负载不会像聚合层中只有两台交换机时那么大。

## 差异化服务 – 服务质量

虚拟化环境必须在交换基础架构中承载各种类型的流量，包括租户流量、存储流量和管理流量。每种流量类型都有不同的特征，并对物理交换基础架构提出不同的要求。尽管管理流量通常量很小，但它可能对控制物理和虚拟网络状态至关重要。IP 存储流量通常量很大，并且通常保留在数据中心内。云运营商可能为租户提供各种级别的服务。不同租户的流量在结构中传输不同的服务质量 (QoS) 值。

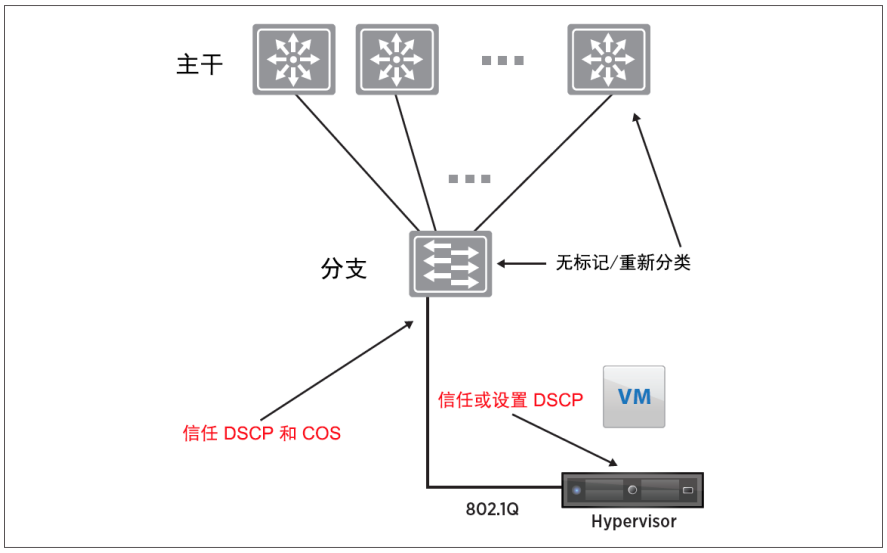



图 82 - 服务质量 (QoS) 标记

对于虚拟化环境，hypervisor 呈现受信任的边界，这表示它为不同的流量类型设置各自的服务质量值。在这种情况下，物理交换基础架构应“信任”这些值。不需要在面向服务器的分支交换机端口处进行重新分类。如果物理交换基础架构中存在拥塞点，那么将查看服务质量值来确定应如何为流量排序（并可能丢弃）或设置优先级。

物理交换基础结构中有两种类型的服务质量配置受支持；一种在 L2 上处理，另一种在 L3 或 IP 层上处理。L2 服务质量有时称为“服务等级”(CoS)，而 L3 服务质量有时称为“DSCP 标记”。

NSX-v 允许信任最初由虚拟机应用的 DSCP 标记，或在逻辑交换机级别上明确修改和设置 DSCP 值。在这两种情况下，DSCP 值随后都会传播到 VXLAN 封装帧的外部 IP 包头。这样，外部物理网络可以根据外部标头上的 DSCP 设置来设置流量的优先级。

# NSX-v 部署注意事项

本部分讨论如何将 VMware NSX-v 提供的网络虚拟化放置在之前部分中展示的可扩展 L3 网络结构的顶部。网络虚拟化由三个主要方面组成：分离、重现和自动化。所有这三个方面对于实现所需的效率都至关重要。本部分侧重于分离，这是简化和扩展物理基础结构的关键。

能够在逻辑网络级别上提供 L2 连接并且独立于底层网络基础架构的特征是实现松耦合效果的基本属性。在 L3 数据中心结构的示例中，L2/L3 边界位于分支层，VLAN 无法跨越交换基础架构内的单个机架，但这不会妨碍在连接到逻辑网络的工作负载之间调配 L2 邻接关系。

在构建新环境时，选择为将来增长留有余地的体系结构很重要。此处所述的方法适用于从小规模开始增长为大规模的部署，同时保留相同的整体体系结构。

此类部署的指导原则是，网络虚拟化解决方案不暗示任何 VLAN 跨越单个机架。尽管这看起来是很简单的要求，但它对物理交换基础架构的构建方式和扩展方式都有着广泛的影响。

在启用了 NSX 的数据中心内，为提供特定功能（如计算、管理和边缘服务）的 ESXi 主机实现逻辑分离和分组是可取的。此分离实质上是逻辑分离，可为数据中心架构提供以下优势：

- 针对特定功能扩展和缩减资源的灵活性。
- 隔离和制定控制范围的能力。
- 针对特定功能管理某些资源的生命周期（更好的硬件 – CPU、内存和网卡，以及升级、迁移等）。
- 基于功能连接需求的高可用性，以及 DRS、FT、Edge P/V 和 HA 等
- 对需要进行频繁更改的领域或功能的自动化控制（应用层、安全标记、策略、负载均衡器等）

图 83 重点展示了不同类型的机架中的 ESXi 主机的聚合：计算机架、基础架构机架和边缘机架。

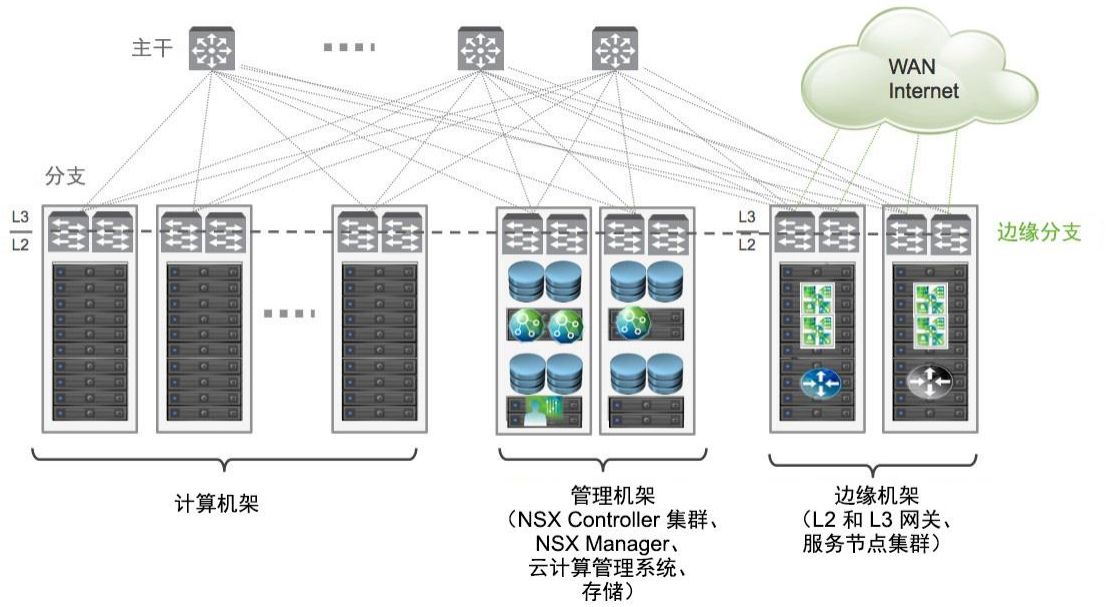



**图 83 - 计算、边缘和管理机架**

**注意**：针对每种类型的机架分离功能可能是逻辑上的，而不一定是物理上的，例如，正如本章节稍后将讨论到的，这在较小规模的部署中可能很常见，以便将边缘和基础架构功能整合在同一个物理机架中。

大致来说，下面是不同类型的机架的主要特征：

- **计算机架**：它们组成了基础架构中托管租户虚拟机的部分。它们应具有以下设计特征：
  - 对于新部署或重新设计：
    - 不需要针对虚拟机的 VALN。
    - 不需要基础架构 VLAN（VXLAN、vMotion、存储、管理）扩展到计算机架之外。
  - 提供可重复的机架设计，可轻松按需扩展。

- **边缘机架**：提供与物理基础架构的更紧密的集成，同时将叠加的架构与物理基础架构桥接起来。以下是边缘机架提供的主要功能：
  - 提供与物理网络的入口和出口连接（NSX-Edge 虚拟设备上的南北向 L3 路由）。
  - 允许与连接到物理网络中的 VLAN 的物理设备进行通信（NSX L2 桥接）。
  - 托管集中式逻辑或物理服务（防火墙、负载均衡器等）。
- **管理机架**：它们托管管理组件，包括 vCenter Server、NSX Manager、NSX Controller、Cloud Management Systems (CMS) 以及其他共享 IP 存储相关组件。关键是基础架构的此部分中没有任何特定于租户的寻址。如果将占用大量带宽的基础架构服务（例如基于 IP 的存储）放在这些机架中，可以动态扩展这些机架的带宽，如之前的“高带宽”部分中所述。

# NSX-v 域中的 ESXi 集群设计

集群设计涵盖以下关键属性：

- 计算、边缘和管理机架的 vSphere 集群规模和范围。
- 计算、边缘和管理机架的 VDS 上行链路连接注意事项
- NSX-v 域中的 VDS 设计

## 计算、边缘和管理集群

NSX 参考设计建议将用于计算的 ESXi 主机分组为分散在专用机架中的单独集群。边缘和管理集群则可以组合在单个机架或专用机架中，具体取决于设计规模。集群设计以及基于服务的机架设计显示在图 84 中。

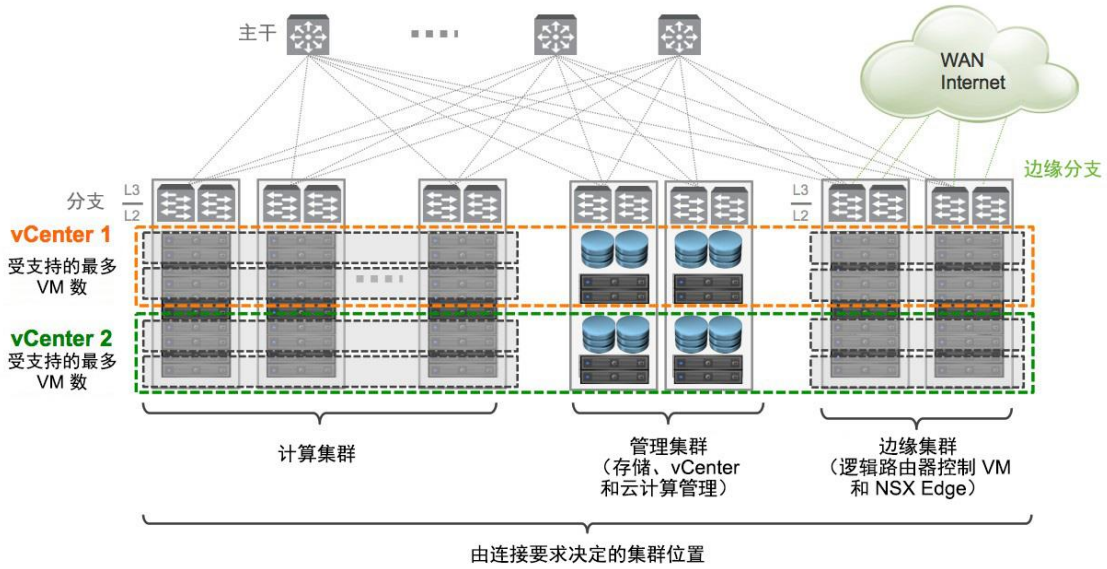



图 84 - 集群设计

根据用途来布局计算、管理和边缘集群。集群大小可以调整到 vSphere 平台的上限并跨越机架，以便某个交换机或机架的故障只会影响部分集群容量。

已启用 NSX 的数据中心的 vCenter 设计大致分为两类：小型/中型与大规模数据中心。在这两种场景中，以下注意事项在部署 NSX 时均有效。

- vCenter 与 NSX Manager 具有一对一的映射关系。换句话说，每个给定的 vCenter 只对应一个 NSX Manager（NSX 域）。这意味着，由 vCenter 的规模限制管理整体 NSX 部署的规模。
- NSX Controller 作为安装过程的一部分，必须部署到 NSX Manager 所连接到的同一个 vCenter Server 中。
- 小型/中型与大规模设计之间的差异通常在于部署的 vCenters 数量和它们映射到 NSX Manager 的方式。

在小型/中型数据中心部署中：

- 通常部署单个 vCenter 用于管理所有 NSX 组件，如图 85 中所示。


图 85 -管理所有 NSX 组件的单个 vCenter

- 建议仍然为计算资源分配单独的集群和机架组。
- 还建议部署单独的边缘和管理集群以适应将来的增长。与图 83 和图 84 中所示的不同，边缘和管理机架通常整合起来，对应的 ESXi 主机集群共享架顶式连接。

大多数企业部署将专用的 vCenter 用于管理集群。甚至在体系结构中引入 NSX 平台以前，通常就已经部署了此 vCenter。当发生这种情况时，通常会引入管理集群的一个或多个专用 vCenter Server 部分来管理 NSX 域的资源（边缘和计算集群），如下面的图 86 所示。


图 86 - 用于管理 NSX 资源的专用 vCenter Server

采用此类方法具有若干优势：

- 避免循环依赖 – 管理集群应始终在它所管理的域之外。
- 管理集群的可移动性，可进行远程数据中心操作。
- 与现有 vCenter 集成。
- 能够部署多个 NSX 域。
- 主 vCenter 的升级不会影响 NSX 域。
- 可进行 VMware Site Recovery Manager 和其他明确的状态管理。

此外，大规模设计采用一个或多个专用机架（管理机架）和 ToR 交换机来托管管理集群。

关于 NSX 域中的 ESXi 主机集群部署的一些其他注意事项：

- 计算容量可通过两种方式进行扩展，即通过添加更多计算机架的横向扩展和通过添加 vCenter Server 的纵向扩展。
- 管理集群的 ESXi 主机部分通常不需要针对 VXLAN 进行调配。如果它们部署在两个机架中（以抵御机架故障场景），它们通常需要针对管理工作负载（如 vCenter Server、NSX Controller、NSX Manager 和 IP 存储）将 L2 连接 (VLAN) 延展到这些机架。图 87 重点介绍了跨管理机架提供 L2 连接的推荐（即使不是唯一的）方法。

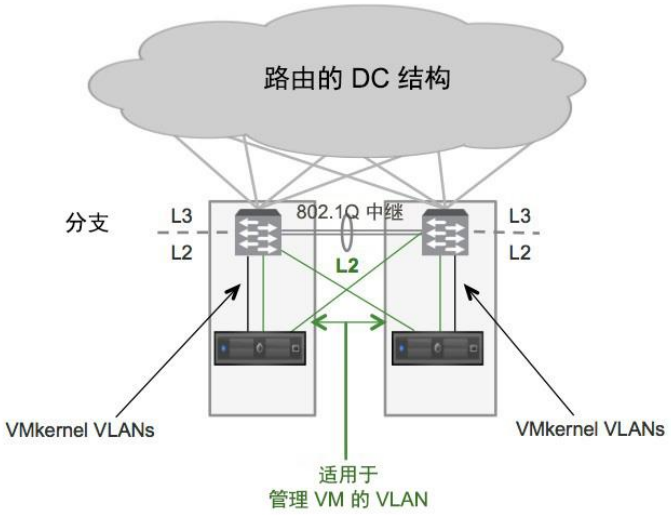



**图 87 - 跨管理机架的 L2 连接要求**

成对的分支交换机部署在两个管理机架中（一个物理分支用于机架），并且使用专用的跨机架 L2 电缆使 ToR 设备互连（L2 802.1q 中继接口连接）。此外，每个服务器对于这两个 ToR 交换机都是双宿的，利用部署了服务器的机架本地的物理上行链路和一个跨机架物理链路来到达第二个机架中的分支（上图中的绿色连接）。假设每台服务器都为所有类型的流量使用两个物理上行链路，则本地上行链路可以传输 VMkernel VLAN（用于 VXLAN、vMotion、存储等基础架构流量）和用于管理虚拟机的 VLAN，而只有后者将在连接到第二个机架中的分支的物理上行链路上传输。此外，不同类型的物理上行链路可以专用于这些类型的流量。

- 边缘机架通常是连接到外部物理网络的唯一机架（除了连接到公用数据中心结构）。这使外部 VLAN 可以包含在一小组机架中，而不是扩展到整个环境中。NSX Edge 随后提供外部网络和计算工作负载之间的连接。
- 边缘集群可以利用成对的 ToR 交换机部署在单个机架中，或分散在两个机架中，每个机架中各有一对 ToR 交换机，以应对机架故障。为了将边缘集群分散到单独的机架，可能需要延展这些机架之间的 L2 连接，在“边缘机架设计”部分中将对此进行更详细的讨论。尽管可以使用图 87 所示的相同方法实现这一目的，但使用单独的 L2 网络基础架构更常见。

## NSX-v 域中的 VDS 上行链路连接

本部分介绍了将 ESXi 主机连接到每种类型的机架中存在的 ToR 交换机（VDS 上行链路连接）的一般设计选项。用于连接主机的设计标准如下所示：

- 传输的流量类型 – VXLAN、vMotion、管理、存储。此情况下的具体侧重点在于 VXLAN 流量，因为它是 NSX-v 部署中找到的特定附加流量类型。
- 基于流量服务级别协议所需的隔离类型 – 专用上行链路（例如用于 vMotion/管理）与共享上行链路。
- 集群类型 – 计算工作负载、边缘和管理，无论是否具有存储等。
- VXLAN 流量需要的带宽量，可决定部署单个还是多个 VTEP。

VDS 将称为 DVUplink 的特殊端口组用于上行链路连接。下面的图 88 显示了在 ESXi 集群上部署 VXLAN 时动态创建的特定端口组支持的绑定选项。该端口组的绑定选项必须在 VXLAN 调配过程中指定。

| 绑定和故障转移模式 | NSX 支持 | 多 VTEP 支持 | 上行链路行为<br>2x10G |
|---|---|---|---|
| 基于源端口的路由 | ✓ | ✓ | 两个均处于活动状态 |
| 基于源 **MAC** 哈希的路由 | ✓ | ✓ | 两个均处于活动状态 |
| **LACP** | ✓ | × | 基于流，<br>两个均处于活动状态 |
| 基于 **IP** 哈希的路由<br>（静态以太网通道） | ✓ | × | 基于流，<br>两个均处于活动状态 |
| **Explicit Failover Order** | ✓ | × | 只有一个处于活动状态 |
| 基于物理网卡负载路由 **(LBT)** | × | × | × |

**图 88 - NSX 中的上行链路连接选项的绑定选项**

如上所示，VXLAN 流量唯一不支持的绑定选项是基于负载的绑定 (LBT)。所有其他选项都可供使用，并且所作的具体选择将对 VXLAN 封装流量在 VDS 上行链路上的发送和接收方式以及所创建的 VMkernel (VTEP) 接口数量产生影响。

根据为 VXLAN 端口组选择的绑定模式，仍可以将 LBT 用于其他端口组（即，不同类型的流量）。下面是一些其他注意事项，与选定上行链路连接如何影响创建端口组的灵活性以及如何在整个集群或传输域中继承它有关。

- 在考虑上行链路连接时，VDS 提供了出色的灵活性。定义为 VDS 一部分的每个端口组原则上都可以使用不同的绑定选项，具体取决于与该特定端口组关联的流量的要求。
- 与给定端口组关联的绑定选项对于连接到该特定 VDS 的所有 ESXi 主机必须相同（即使属于单独的集群）。在此仍然引用 VXLAN 传输端口组的案例，如果为计算集群 1 的 ESXi 主机部分选择了 LACP 绑定选项，则必须将此相同选项应用到连接到相同 VDS 的所有其他计算集群。不接受为不同的集群选择不同的绑定选项，这样做会在 VXLAN 调配时触发错误消息。
- 如果为属于给定 VDS 的集群选择 LACP 或静态以太网通道作为 VXLAN 流量的绑定选项，那么应为该 VDS 上定义的所有其他端口组（流量类型）使用相同的选项。原因是，选择以太网通道绑定意味着在物理网络端也要求配置端口通道。一旦在该物理交换机（或多个交换机）上的端口通道中捆绑物理 ESXi 上行链路后，在主机端上使用不同的绑定方法可能会导致无法预测的结果（并且经常处于断开通信状态）。
- VTEP 设计会影响上行链路选择，因为 NSX 支持单个或多个 VTEP 配置。
  - 单 VTEP 选项首先可以简化操作：如果要求是在每台给定主机上支持少于 10 G 的 VXLAN 流量，那么明确的故障转移顺序选项是有效的选项，因为它允许在物理上将重叠流量与所有其他类型的通信分开（一条上行链路用于 VXLAN，另一条上行链路用于其他流量类型）。使用明确的故障转移顺序还可以为应用提供一致的质量和体验，不受任何故障影响。使一对物理上行链路专用于 VXLAN 流量并使它们保持活动/备用状态事实上可以保证，当物理链路或交换机发生故障时，应用仍然可以访问同一个 10 G 带宽管道。显而易见，这样做的代价是部署更多的物理上行链路，并且仅主动使用一半的可用带宽。

    如果要求是支持多于 10 G 的 VXLAN 流量，那么可以部署静态以太网通道或 LACP，这样只需轻松部署单个 VTEP，还能增加带宽。如前所示，这样做的代价是在物理交换机上需要进行额外的配置和协调，并且经常要求特定功能 (vPC/MLAG) 在架顶式设备上受支持。另外，在此情况下必须考虑物理故障将对应用产生的影响。

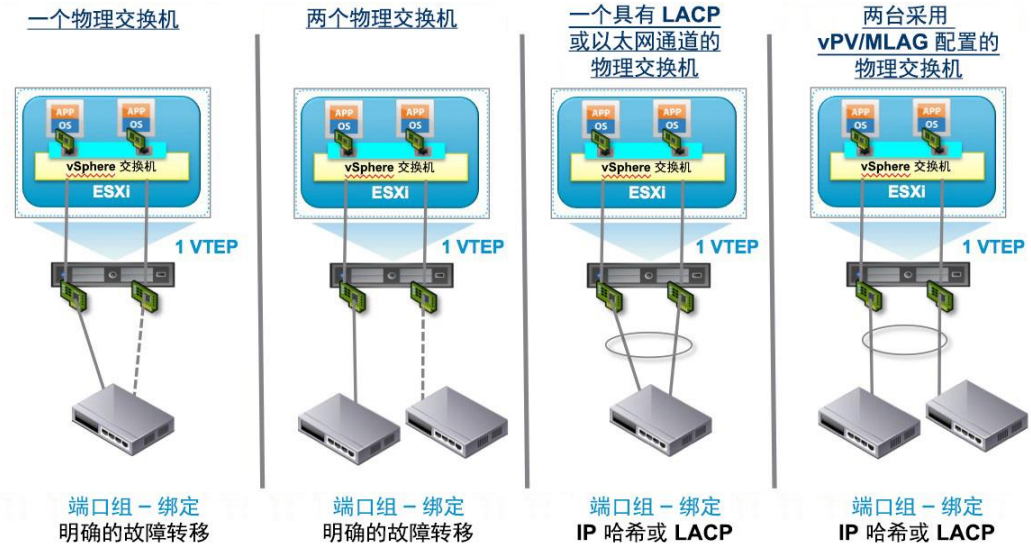



**图 89 - 单 VTEP 上行链路选项**

图 89 显示了单 VTEP 选项。注意，尽管存在两个活动的 VDS 上行链路，右边的端口通道场景中同样部署单 VTEP。这是因为来源于 VTEP 的流量可以在这两个物理路径上逐个流进行哈希处理，所以端口通道被视为单个逻辑上行链路。

- 当无法使用端口通道（例如，部署了刀片服务器）或希望使 VDS 上行链路配置保持分离状态并且独立于物理交换机配置时，可配置多个 VTEP 来增加 VXLAN 流量可用的带宽。此选项需要付出的代价是增加了操作复杂性，在某些情况下可能超过它的好处。


**图 90 - 多 VTEP 上行链路选项**

**注意**：调配的 VTEP 数量始终与物理 VDS 上行链路的数量匹配。在图 92 中所示的用户界面的“VMKNic Teaming Policy”部分中选择 SRC-ID 或 SRC-MAC 选项后，将自动执行此操作。

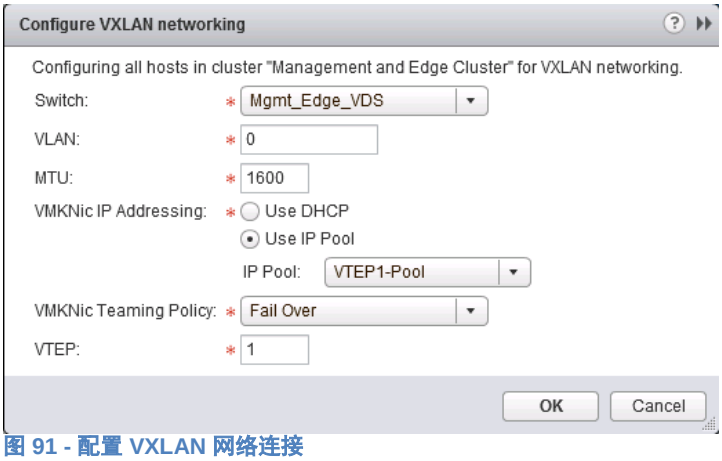



图 91 - 配置 VXLAN 网络连接

针对要部署的上行链路配置的类型，可以根据以下注意事项制定选择标准：

- 配置简易性 - 单个 VTEP 与物理交换机配置对比。
- 每种流量所需的带宽。
- 融合要求。
- 集群具体内容 - 计算、边缘和管理。
- 上行链路利用因素 - 基于流与基于 MAC 对比。

对于计算集群中 ESXi 主机的 VXLAN 流量，推荐的绑定模式是 LACP。它可以充分利用两种链路并减少故障转移时间。另外，该模式可以简化 VTEP 配置和故障排除（无需进行额外配置），并且还可以与物理交换机配合工作。很显然，只有在部署的物理交换机支持一种多机架以太网通道功能的情况下，如支持 vPC (Cisco) 或 MLAG（Arista、Juniper 等），才能实现 ESXi 连接的 ToR 多样化。

而于边缘集群中的 ESXi 主机，则不建议选择 LACP 或静态以太网通道。如前文所述，为 VXLAN 流量选择 LACP 意味着必须对同一 VDS 中的所有其他端口组（流量类型）使用同一绑定选项。边缘机架的主要功能之一就是提供与物理网络基础架构的连接，这一点通常可以使用支持 VLAN 的专用端口组来实现，因为在此类端口组中，NSX Edge（处理南北向路由通信）与下一个跃点 L3 设备建立了路由邻接关系。当 ToR 交换机作为 L3 设备来使用时，为支持 VLAN 的端口组选择 LACP 或静态以太网通道将使 NSX Edge 与 ToR 设备之间的交互复杂化。

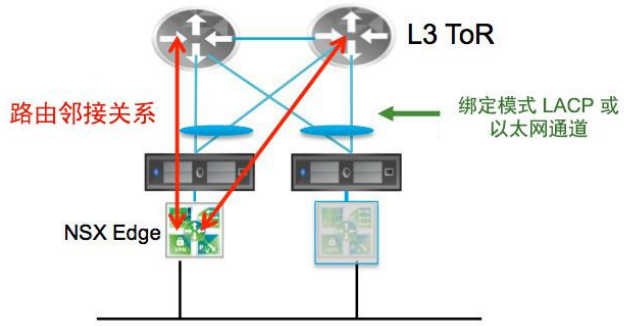



图 92 - 基于多机架以太网通道的路由对等关系

如图 92 所示，ToR 交换机不仅需要支持多机架以太网通道功能（vPC 或 MLAG），而且还要能够与此逻辑连接上的 NSX Edge 建立路由邻接关系。这将增强所选物理网络设备的依赖关系，甚至在许多种情况下，这可能是一种不受支持的配置。因此，对于边缘集群，根据前面所列的对优缺点的评估，建议选择“明确的故障转移顺序”或“SRC-ID/SRC-MAC 哈希”作为 VXLAN 流量的绑定选项。

## NSX-v 域中的 VDS 设计

vSphere Virtual Distributed Switch (VDS) 是创建 VXLAN 逻辑网段所需的基本组件。VXLAN 逻辑交换机（第 2 层广播域）的范围取决于传输域的定义；每个逻辑交换机都有可能跨越整个传输域上多个独立的 VDS 实例。

图 93 描绘了 VDS、ESXi 集群和传输域之间的关系，并展示了推荐的设计，即对计算和边缘集群使用独立的 VDS 交换机。

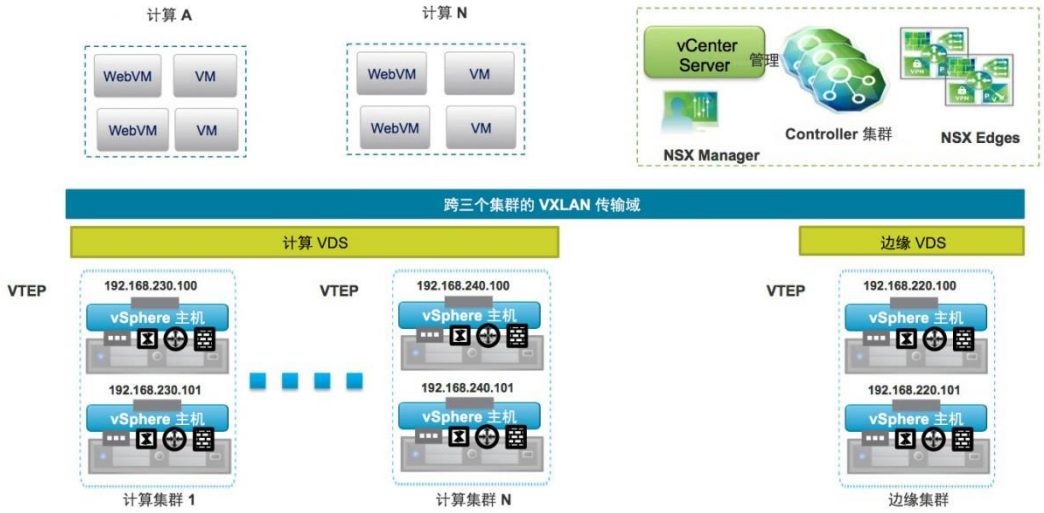



**图 93 - 集群、VDS 和传输区域**

尽管可以采用同时跨越计算和边缘集群的单个 VDS 设计，但对计算和边缘集群分别使用独立的 VDS 具有以下几项优势：

- 灵活的运维控制范围：通常，计算/虚拟基础架构管理员和网络管理员是独立的实体，因此每个域都可以管理特定于集群的任务。这些优势已成为设计特定服务的专用集群和机架时要考虑的一项因素，然后进一步通过 VDS 设计选择具体实现。
- 灵活管理计算和边缘集群上的上行链路连接：例如，请查看前文中有关上行链路设计的论述以及对计算和边缘集群使用不同绑定选项的建议。
- 通常，VDS 边界与传输域一致，因此连接到逻辑交换机的虚拟机可以跨越传输域。但是，vMotion 边界始终受到 VDS 扩展的限制，所以对计算资源使用独立的 VDS 可确保这些工作负载始终不会移动（无论出于无意还是出于选择）到专用于其他服务的 ESXi 主机上。
- 灵活管理 VTEP 配置：请查看前文中有关 VTEP 设计选择的论述。
- 避免向计算机架公开由部署于边缘机架（NSX L3 路由和 NSX L2 桥接）中的服务所使用的支持 VLAN 的端口组。

**注意**：用于基础架构服务的 VDS 不属于 VXLAN，因为通常情况下，管理集群（管理机架的一部分）代表一个独立的实体，可以提供远不止是为给定 NSX 域提供服务的功能（如前文所述）。

# ESXi 主机流量类型

部署于 NSX 域中的 ESXi hypervisor 通常可以获取不同类型流量的来源。在接下来的部分中，我们将介绍叠加、管理、vSphere vMotion 和存储流量。叠加流量是一种新的流量类型，可以携带所有虚拟机通信并将这些通信封装在 UDP (VXLAN) 帧中。正如前文所述，VXLAN 流量通常来源于计算和边缘集群中的 ESXi 主机，而不是管理集群中的 ESXi 主机。其他类型的流量通常应用于整个服务器基础架构。

传统设计需要使用不同的 1G 网卡接口才能将不同类型的流量传入和传出 ESXi hypervisor。随着数据中心中 10G 接口的采用率不断增加，将一对通用上行链路上不同类型的流量整合在一起的情况越来越普遍。

在这种情形下，不同的流量类型可以使用 VLAN 进行隔离，从 IP 寻址的角度来看，即实现了明确的分离。在 vSphere 体系结构中，特定内部接口在每台 ESXi 主机上都进行了定义，用于获取不同类型流量的来源。这些接口称为“VMkernel 接口”，它们全都分配了独立的 VLAN 和 IP 地址。

在部署路由数据中心结构时，这些 VLAN 中的每一个都将终止于分支交换机（架顶式设备），所以分支交换机将为每个 VLAN 都提供一个 L3 接口。此类接口也称为 SVI 或 RVI（如图 94 所示）。

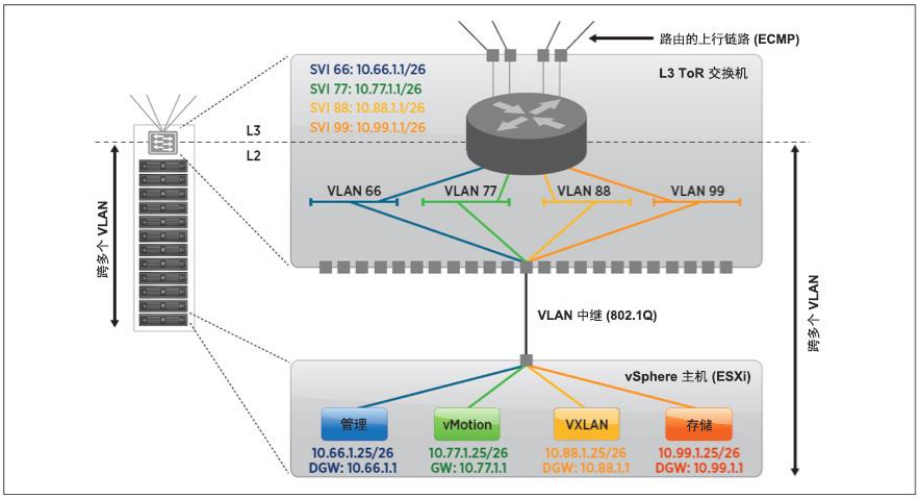



图 94 - 机架中的主机和分支交换机配置示例

接下来，我们将进一步介绍不同类型的主机流量。

## VXLAN 流量

在 vSphere 主机准备好使用 VXLAN 进行网络虚拟化后，这些主机便可以使用一种新的流量类型。连接到基于 VXLAN 的逻辑 L2 网络之一的虚拟机将使用此流量类型进行通信。来自虚拟机的流量将进行封装，并作为 VXLAN 流量发出。外部物理结构永不会检测该虚拟机的 IP 和 MAC 地址。

VXLAN 安全加密链路终端 (VTEP) IP 地址（与 VMkernel 接口相关联）用于在整个结构中传输帧。就 VXLAN 而言，这些安全加密链路均由 VTEP 启动和终止。在位于同一数据中心的虚拟机之间流动的流量通常称为东西向流量。对于这种类型的流量，源 VTEP 和目标 VTEP 都位于计算和边缘机架内的 hypervisor 中。例如，离开数据中心的流量将在租户虚拟机与 NSX Edge 之间流动。这种流量称为南北向流量。

当通过路由结构基础架构部署 NSX 时，已连接不同计算和边缘机架的主机 VTEP 接口必须作为不同 IP 子网（即所谓的“VXLAN 传输子网”）一部分进行部署。如果结合这一点考虑对各独立机架中计算和边缘集群进行条带化的建议，则会面临一项很有意思的部署挑战。在 ESXi 主机上调配 VXLAN 实际上在集群级别上完成，并且在此过程中，用户必须指定如何处理这些属于该集群的主机 VTEP 接口（请参见前面的图 91）。

这一难题包括将不同子网中的 IP 地址分配给这些 VTEP 接口。可以通过以下两种方式做到这一点：

1. 利用“使用 DHCP”选项，这样，每个 VTEP 都将在适当的 IP 子网中获得一个地址，具体取决于它所连接的机架。在生产部署场景中，建议采用此方法。
2. 通过 ESXi CLI 或 vSphere Web Client UI 以静态方式将 IP 地址分配给每台主机的 VTEP。在此情况下，如果要将 VXLAN 调配到 ESXi 集群，建议选择 DHCP 选项、等待 DHCP 过程超时，然后以静态方式分配 VTEP 地址。很显然，这种方法仅适用于规模较小的部署，并且应仅限于实验室环境。

## 管理流量

管理流量源自于主机上的管理 VMkernel 接口，并终止于该接口，这种流量还包含 vCenter Server 和各主机之间的通信以及与 NSX Manager 等其他管理工具进行的通信。

一个 VDS 可以横跨多个部署在一台分支交换机之外的 hypervisor。由于没有任何 VLAN 可以延展到分支交换机之外，参与通用 VDS 并连接到独立分支交换机的 hypervisor 的管理接口位于独立的各 IP 子网中。

## vSphere vMotion 流量

在 vSphere vMotion 迁移过程中，运行中的虚拟机状态通过网络传输到另一台主机。每台主机上的 vSphere vMotion VMkernel 接口均用于移动此虚拟机状态。主机上的每个 vSphere vMotion VMkernel 接口都分配了一个 IP 地址。根据物理网卡的速度，可以确定同时迁移虚拟机 vSphere vMotion 的数量。在 10GbE 网卡上，可同时迁移八个 vSphere vMotion。

VMware 技术支持团队过去曾一直建议要将所有用于 vMotion 的 VMkernel 接口作为通用 IP 子网的一部分进行部署。但是当使用访问层中的 L3 针对网络虚拟化设计网络时，这明显行不通，因为此时需要为这些 VMkernel 接口强制选择不同机架中的不同子网。在此设计得到 VMware 的完全和正式支持之前，建议用户先完成 RPQ 流程，以便 VMware 可以根据具体情况对设计进行验证。

## 存储流量

VMkernel 接口的用途是提供诸如共享存储或非直接连接存储的功能。通常，我们是指可以通过 IP 连接（例如 NAS 或 iSCSI，而不是 FC 或 FCoE）进行连接的存储。从 IP 寻址的角度来说，适用于管理流量的规则同样适用于存储 VMkernel 接口。机架内部服务器的存储 VMkernel 接口，即连接到分支交换机的存储 VMkernel 接口，属于同一 IP 子网。但是，该子网的范围无法超出这台分支交换机。因此，在不同的机架中，主机的存储 VMkernel 接口 IP 位于不同的子网中。

## 适用于 VMkernel 接口 IP 寻址的主机配置文件

在本部分中，我们将讨论如何使用 vSphere 主机配置文件方法，将 IP 地址自动调配到每种流量类型的 VMkernel 网卡上。利用主机配置文件功能，用户可以创建一个其属性在整个部署中进行共享的引用主机。在识别此引用主机并执行所需的示例配置后，即可从这台主机创建一个主机配置文件，并且在部署过程中可以将该配置文件应用于其他主机。使用这种方法，用户可以快速配置大量主机。

我们先来看看机架中的主机所需的示例配置类型，然后再介绍如何在配置整个机架的服务器过程中使用主机配置文件方法。如前文所述，每个机架（仅限于计算和边缘机架的 VXLAN 机架）通常都提供了相同的 VLAN 组，一组有四个 VLAN，即存储、vSphere vMotion、VXLAN 和管理。与这些 VLAN 相关联的 VLAN ID 同样在各机架之间通用，因为 VDS 跨越了经过横向条带化的主机集群。这意味着，使用 L3 结构设计时，必须更改与不同机架中相同 VLAN ID 相关联的 IP 子网，如表 1 所示。

| IP 地址管理和 VLANs[1] | | |
|---|---|---|
| 功能 | 全局 VLAN ID | IP 地址 |
| 存储 | 66 | 10.66.R_id.x/26 |
| vMotion | 77 | 10.77.R_id.x/26 |
| VXLAN/VTEP | 88 | 10.88.R_id.x/26 |
| 管理 | 99 | 10.99.R_id.x/26 |

[1] VLAN、IP 地址和掩码的值都是示例（并非设计规范）

**表 1 - IP 地址管理和 VLAN ID**

以下内容是每台主机所需的部分配置：

1) 各个 IP 子网中每种流量类型的 VMkernel 网卡 (vmknic) IP 配置。

2) 每个子网的静态路由配置，用于处理路由到各自网关的相应流量。

之所以需要静态路由，是因为 ESXi 仅支持两个 TCP/IP 堆栈：

- VXLAN：此堆栈专用于源自 VMkernel VTEP 接口的流量。对于每个指向在本地 ToR 上所部署网关的 ESXi，都可以在此堆栈上配置一个专用的默认路由 0.0.0.0/0，这样便能够与部署在不同传输子网中的所有远程 VTEP 进行通信。
- 默认：此堆栈用于所有其他流量类型（vMotion、管理、存储）。通常利用默认路由来进行管理（因为与 vmk0 管理接口之间的连接可能起始于多个远程 IP 子网）。这说明，对于其他类型的流量，要支持子网间通信需要进行静态路由配置。

回到我们的示例，机架 1（请参阅表 1，R_id 为 1）中给定的主机 1 具有以下 vmknic 配置：

- 存储 vmknic 的 IP 地址为 10.66.1.10。
- vSphere vMotion vmknic 的 IP 地址为 10.77.1.10。
- 管理 vmknic 的 IP 地址为 10.99.1.10。

主机 1 上用于默认 TCP/IP 堆栈的默认网关配置位于管理 vmknic 子网 10.99.1.0/26 中。为了支持其他子网的正确路由，以下静态路由将作为主机 1 准备工作的一部分进行配置：

- 存储网络路由：esxcli network ip route ipv4 add -n 10.66.0.0/16 -g 10.66.1.1
- vSphere vMotion 网络路由：esxcli network ip route ipv4 add -n 10.77.0.0/16 -g 10.77.1.1

机架 1 的主机 1 配置完成后，主机配置文件便已创建，然后应用到该机架中的其他主机。当将配置文件应用到这些主机时，新的 vmknics 将完成创建，并且静态路由也添加完毕，从而简化了部署过程。

**注意**：VXLAN vmknic 将获得 IP 地址 10.88.1.10，并且可以利用该子网中专用的默认网关配置（TCP/IP VXLAN 堆栈的一部分）。正如前面提到的，在将 VXLAN 配置调配到计算/边缘集群时，通常会利用 DHCP 分配 VTEP IP 地址。

如果在 vSphere Auto Deploy 环境中，预启动执行环境启动基础架构将连同 Auto Deploy 服务器和 vCenter Server 一起支持主机启动流程，并帮助自动部署和升级 ESXi 主机。

# 边缘机架设计

边缘机架托管的服务允许在逻辑和物理网络之间进行通信。通信可能在第 3 层（NSX 逻辑路由）或第 2 层（NSX 桥接）进行；这些功能将在本部分中进行详细深入的介绍。

为了恢复边缘机架中提供的服务，部署边缘机架的冗余对是当前的最佳做法，目的是防止整个机架出现灾难性故障。部署 L2 和 L3 网络服务强制要求在各边缘机架之间扩展一定数量的 VLAN，本章稍后将对这一点展开详述。这些 VLAN 代表用于支持逻辑网络与外部物理基础架构之间通信（在 L2 或 L3）的接口。

图 95 展示了允许在边缘机架之间扩展 VLAN 的建议部署模式。

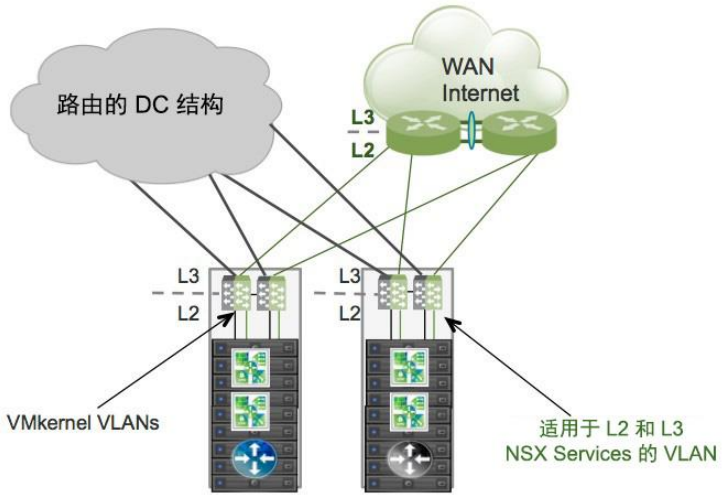



**图 95 - 边缘机架到物理网络的连接**

在这种部署模式下，架顶式 (ToR) 交换机的冗余对在每个边缘机架中均已连接起来，发挥着双重作用：

- 充当第 3 层分支节点，通过 L3 点对点链路连接到结构主干设备。这样一来，ToR 交换机就可以为各服务器（Mgmt、VXLAN、vMotion、存储等服务器）使用的所有本地 VMkernel VLAN 提供 L2/L3 边界和默认网关。这些 VLAN 的跨度仅限于每个本地边缘机架（与计算和管理机架的情况类似）。
- ToR 交换机还表示数据中心结构与外部物理网络基础架构之间的接口，有时它们也称为“边界分支”。在这种情况下，它们只能为用于与物理网络空间（L2 和 L3 NSX 服务）进行互连的 VLAN 提供 L2 传输功能。与这些 VLAN 连接的设备的默认网关通常调配到一对专用核心路由器。每个部署的租户都将使用一个单独的 VLAN（或 VLAN 组）。

## NSX Edge 部署注意事项

在“NSX Edge”部分中，已讨论如何部署一对活动/备用 Edge 服务网关来恢复这些 NSX 组件提供的所有服务（路由、防火墙等）。如果我们深入了解逻辑和物理网络之间的路由功能，便可以发现实际上可以部署三种 HA 模式： 活动/备用模式、独立模式和 ECMP 模式，后者代表从 NSX SW 版本 6.1 开始引入的较新功能。

请注意，图 96 展示了参考拓扑，这种拓扑将用于描述各种 HA 模式，并且可以利用 DLR 和物理网络之间的 NSX Edge（如前文“逻辑路由部署”部分中所述）。

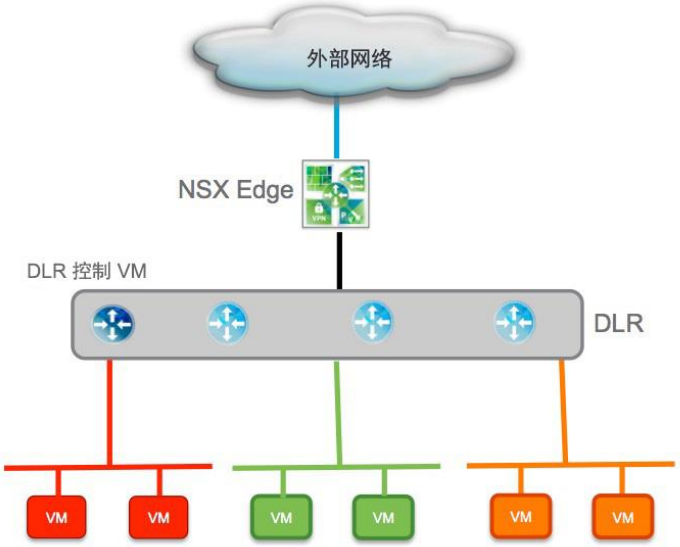



图 96 - NSX 逻辑路由参考拓扑

接下来的三个部分将详细分析上述 HA 模式，特别要着重介绍如何为 NSX Edge 和 DLR 控制虚拟机功能组件提供恢复能力，以及这些组件发生故障时对南北向数据平面通信产生的影响。

### 有状态的活动/备用 HA 模式

这是一种冗余模式，该模式为每个租户部署一对 NSX Edge 服务网关；一个 Edge 在活动模式下运行（即主动转发流量并提供其他逻辑网络服务），而第二个单元则处于备用模式，等到活动 Edge 发生故障时接管其职责。

利用内部通信协议，各种逻辑网络服务的运行状况和状态信息可以在活动和备用 NSX Edge 之间进行交换。默认情况下，使用部署在 Edge 上第一个“内部”类型的虚拟网卡接口来建立这种通信，但用户也可以明确指定使用哪个 Edge 内部接口。对要使用的接口拥有这种程度的控制力非常重要，因为这样用户就可以选择是将支持 VLAN 的端口组还是将逻辑交换机（支持 VXLAN 的端口组）作为通信通道。选择前者要求所使用的 NSX Edge 虚拟网卡要邻接于 L2（即连接到同一 VLAN），而前者可以提供更大的拓扑灵活性。

**注意**：必须在 NSX Edge 上至少配置一个内部接口才能在活动和备用单元之间交换运行状况。删除最后一个内部接口将打破这种 HA 模式。

图 97 从控制和数据平面两个角度展现了活动 NSX Edge 如何保持活动状态。正因为如此，它才能够与物理路由器（在“外部 VLAN”公用网段上）以及与 DLR 控制虚拟机（在“传输 VXLAN”公用网段上）建立路由邻接关系。与 DLR 和物理基础架构连接的逻辑网段之间的流量始终只流经活动 NSX Edge 设备。

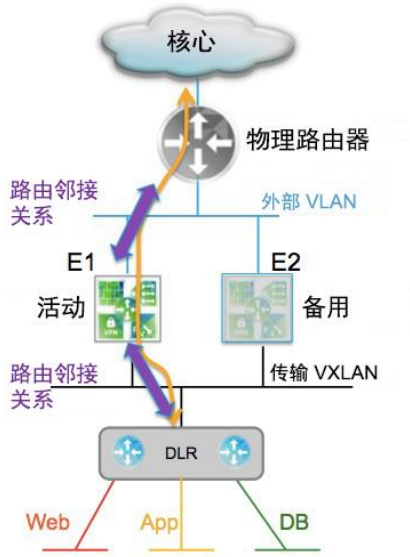



图 97 - NSX Edge 活动/备用 HA 模式

如果活动 NSX Edge 发生故障（例如由于 ESXi 主机出现故障），控制平面和数据平面都必须在接管当前职责的“备用”单元上激活（如图 98 所示）。

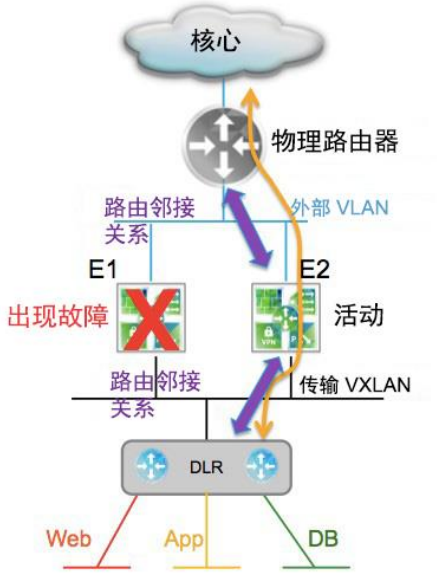



图 98 - 在活动 NSX Edge 发生故障后恢复流量

下面是一些与此 HA 模式行为相关的设计和部署重要注意事项：

- 备用 NSX Edge 可通过判断“Declare Dead Time”（声明失效时间）计时器是否结束计时来检测活动 NSX Edge 是否出现故障。该计时器默认配置为 15 秒，最低可以调整到 6 秒（通过 UI 或 API 调用），它是影响此 HA 模式所遭遇流量中断问题的主要因素。
- 备用单元激活后，它将启动在出现故障的 Edge 上运行的所有服务（路由、防火墙等）。在这些服务重新启动期间，通过利用在活动和备用 Edge 单元之间保持同步的 NSX Edge 转发表中的信息，流量仍可以进行路由。对其他逻辑服务也同样适用，因为状态进行了同步，并且还可以用于防火墙、负载均衡、网络地址转换等。

  **注意**：对于负载均衡服务，只有持久性状态在活动和备用单元之间同步。

- 为了成功进行南北向通信，DLR（在 NSX Edge 南端）和物理路由器（在北端）应开始向新激活的 Edge 发送流量。为此，需要满足以下三个条件：

  1. 在控制平面（路由协议）级别上，DLR 活动控制虚拟机和物理路由器应忽略切换已发生的事实，仍保持之前建立的路由邻接关系处于活动状态。若要做到这一点，可以为 Hello/保持时间计时器设置一个足够长的时间值，确保在 DLR 和物理路由器上的保持时间计时器停止之前，新激活的 NSX Edge 有足够的时间来重新启动路由协议。如果保持时间过期，DLR 和物理路由器将中断路由邻接关系，从而在转发表中删除从 NSX Edge 获取的路由信息。这将导致第二次流量中断，直到新的邻接关系重新建立并且路由信息再次进行交换为止。因此，我们建议在这些设备之间配置 OPSF 或 BGP Hello/保持时间计时器，时间值设置为 40/120 秒。

     **注意**：需要重点强调的一点在于，这些计时器并不是流量中断时长的决定性因素，因此建议值相当高，以便包含最糟糕的情况，即重新启动 NSX Edge 可能会导致激活网络服务的时间更长。

  2. 在数据平面级别上，DLR 内核模块和物理路由器将继续向相同的下一个跃点 IP 地址（即 NSX Edge）发送流量。但是，新激活的 NSX Edge 所使用的 MAC 地址与出现故障的单元所使用的 MAC 地址并不相同。为了避免出现流量黑洞，新 NSX Edge 必须同时在外部 VLAN 网段和传输 VXLAN 网段上发送无故 ARP 请求，才能更新有关这些设备的映射信息。

  3. 在新激活的 NSX Edge 重新启动其路由控制平面后，即可开始向 DLR 控制虚拟机和物理路由器发送 Hello 消息。这就是 NSX Edge 的正常重新启动 (GR) 功能起到的作用。利用 GR 功能，NSX Edge 可以在请求物理路由器和 DLR 控制虚拟机继续使用旧邻接关系的同时，开始与它们建立新的邻接关系。如果没有 GR 功能，这些邻接关系将断开，并在从 Edge 接收到第一个 Hello 消息后进行重新协商，而这最终会再次导致流量中断。

- 除了活动/备用 HA 功能之外，还建议启用 vSphere HA 来保护 NSX Edge 服务网关虚拟机驻留的边缘集群。在该边缘集群中，至少需要部署三个 ESXi 主机才能确保出现故障的 NSX Edge 能够还原到不同的 ESXi 主机（根据反关联性规则），并成为新的备用单元。

- 最后，还请务必考虑 DLR 活动控制虚拟机出现故障的情况，以便确定这一事件在流量中断方面将产生怎样的后果。

  DLR 控制虚拟机可以利用上面针对 NSX Edge 服务网关进行介绍的相同活动/备用 HA 模式。主要不同之处在于，控制虚拟机仅运行 DLR 控制平面，而数据平面完全分布在 NSX 域中的 ESXi 主机内核中。

  如前文中所述，为 DLR 和 NSX Edge 之间用于应对 NSX Edge 故障情况的路由计时器设置较高的时间值（40/120 秒），也可以应对控制虚拟机故障情况，从而实现零秒中断。实际上，当备用控制虚拟机检测到活动控制虚拟机出现故障并重新启动路由控制平面时：

  - 南北向流量将继续流动，因为系统仍在使用从 Edge 接收到的原始路由信息对 ESXi 主机内核中的转发表进行编程。
  - 南北向流量将保持流动，因为 NSX Edge 不会中断与 DLR 控制虚拟机之间的路由邻接关系（针对“协议地址”IP 地址建立），并继续向用于识别 DLR 数据平面组件的“转发地址”IP 地址发送数据平面流量。对控制和数据平面运维使用不同的 IP 地址还意味着，不需要从新激活的控制虚拟机生成 GARP 请求。
  - 在控制虚拟机准备好开始发送路由协议 Hello 消息之后，将使用 DLR 一侧的“正常重新启动”功能来确保 NSX Edge 可以继续维持邻接关系，并继续转发流量。

## 独立 HA 模式（NSX 6.0.x 版本）

独立 HA 模式在 DLR 和物理网络之间插入了两个独立的 NSX Edge 设备（如图 99 所示），并且 NSX 6.0.x SW 各版本都支持该模式。

**注意**：从 NSX SW 版本 6.1 开始，可以采用经过演化的 ECMP HA 模式，该模式将在下一部分中进行介绍。

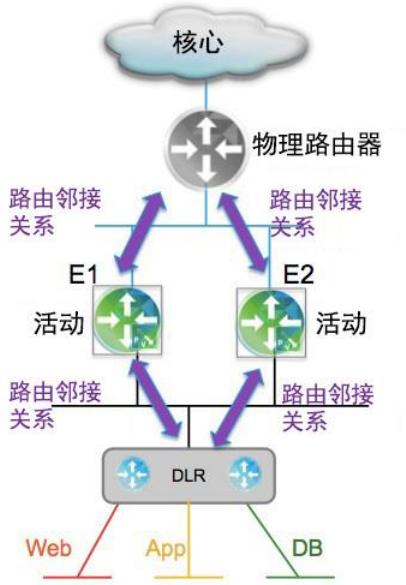



图 99 - NSX Edge 独立 HA 模式

在本例中，从数据平面和控制平面两个角度来看，两个 NSX Edge 设备均处于活动状态，并且可以与物理路由器和 DLR 控制虚拟机建立邻接关系。但是，在所有 6.0.x NSX SW 版本中，DLR 都无法支持等价多路径。因此，即使要为物理网络中存在的 IP 前缀从两个 NSX Edge 接收路由信息，DLR 也仅在它的转发表中安装一个可能的下一跃点（活动路径）。这意味着，所有南北向流量都将只流经一个 NSX Edge，而无法同时利用两个设备。南北方向可能会进行流量负载均衡，因为大多数物理路由器和交换机都默认支持 ECMP。

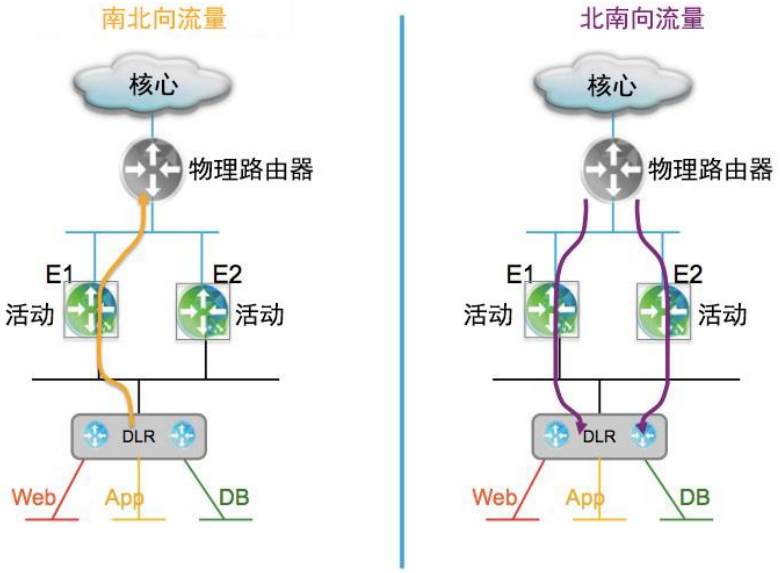



图 100 - 独立 HA 模式下的流量走向

如果在 NSX Edge 服务网关上启用了防火墙服务，上图中由方向相反的流量走向展现出的非对称行为可能会阻止通信。因此，在部署这种 NSX Edge 冗余模式时，建议禁用集中式防火墙。或者，可以调整 Edge 接口的开销来确保从北到南和从南到北两个方向的流量走向对称，即使这可能会增加解决方案的整体运维复杂度。

现在，在 NSX Edge 发生故障时出现的流量中断取决于物理路由器和 DLR 能够以多快的速度检测到故障，并因此而更新它们的转发表，以便开始指向其他（之前未使用的）NSX Edge（如图 101 所示）。

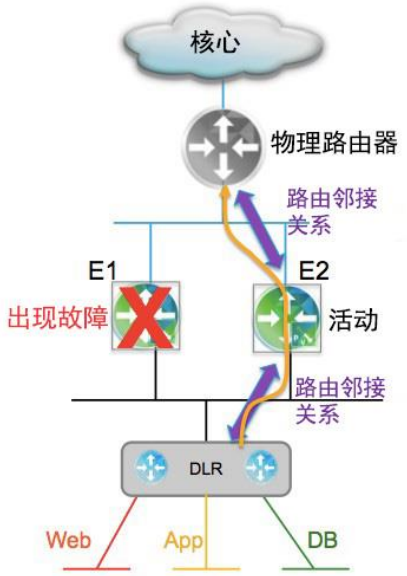



图 101 - 在独立 NSX Edge 出现故障后恢复流量

下方列出了一些关于独立 HA 模式的重要注意事项：

- 在此模式下，建议积极调整 NSX Edge、DLR 控制虚拟机和物理路由器上的路由 Hello/保持时间计时器（支持的最小值为 1/3 秒）。之所以需要这样做，是因为在这种情况下，路由协议保持时间是影响流量中断是否发生的主要因素。当 E1 发生故障时，物理路由器和 DLR 将继续向出现故障的单元发送流量，直到保持时间计时器停止计时为止。此时，它们将断开邻接关系并更新自己的转发表，以便开始使用 E2 作为下一个跃点（如果物理路由器已经对南北向流量使用 E2，只需将 E1 作为可行的下一个跃点删除即可）。
- 建议在此 HA 模式下部署两个独立的活动 HA Edge 网关，摆脱对活动/备用功能的依赖。原因在于，为了将流量中断的可能性降到最低，最好丢掉一个 Edge，而不是等待“备用”单元来接管。
- 也就是说，由于 Edge 网关仍是一台虚拟机，因此还建议利用 vSphere HA 来确保出现故障的 Edge 可以在不同的 ESXi 主机上重新启动。这样，物理路由器可以将它用于南北向通信，并且当 E2 在某个时刻发生故障时，DLR 也可以随时使用它（用于南北向流量）。
- 只有不需要在 Edge 上部署除路由以外的任何逻辑服务（防火墙、负载均衡、网络地址转换等）时，才应当采用这种独立 HA 模式。这是因为在这两个部署的 Edge 网关之间不会同步任何信息，所以就无法实现这些逻辑服务的有状态性。仍可以利用 DFW 功能以分发的方式部署防火墙保护，但是如前文所述，建议在这些 NSX Edge 上禁用防火墙逻辑功能，以便能够容许图 100 中示例所展现的非对称流量路径行为。

  **注意**：在 NSX SW 版本 6.0.x 中，只有当部署超大规格的 NSX Edge 时，才可以禁用逻辑防火墙功能（通过 UI 或 API）。从 SW 版本 6.1 开始，可以针对任何规格禁用该功能。
- 在处理活动控制虚拟机出现的特定故障情况时，主动设置这种独立 HA 模式所需的路由协议计时器具有重要意义。此故障将导致两个 NSX Edge 都（在 3 秒之内）断开之前与控制虚拟机建立的路由邻接关系。这意味着，这两个 NSX Edge 将通过删除最初从 DLR 获取的所有前缀来刷新它们的转发表，从而导致南北向通信停止。

  针对这一问题可以采取的解决方法是，在这两个 NSX Edge 的路由表中安装静态路由信息，如图 102 所示。

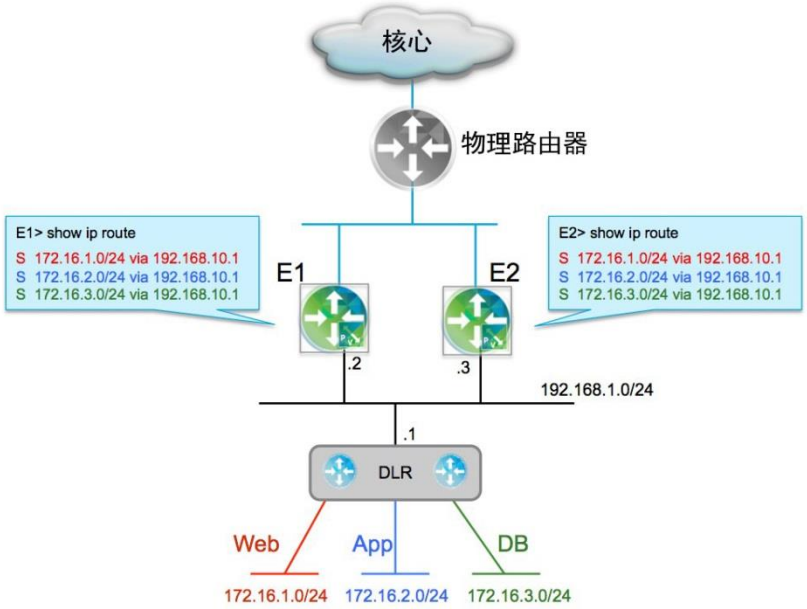



图 102 - 在活动 NSX Edge 上配置静态路由

这样，当邻接关系中断以及动态获取的路由（与逻辑交换机相关联）删除时，从北到南的流量将根据静态路由信息继续流动。为实现此目的，这两个 NSX Edge 还必须将静态路由重新分发到 OSPF 中，以便向物理路由器通告这些静态路由。

**注意**：不需要在 DLR 一侧配置静态路由信息。在活动控制虚拟机发生故障，并且备用控制虚拟机接管并开始重新启动路由服务后，其转发表（以及 ESXi 主机内核中的相应路由表）中的路由信息将保持不变，无论 DLR 控制虚拟机与 ESG 设备之间的邻接关系是否已断开。这意味着，所有定向到北向 IP 前缀的流量都将在通过 NSX Edge 提供的各个等价路径中继续进行哈希处理。当控制虚拟机重新启动其路由服务后，它与 NSX Edge 之间的邻接关系将重新建立，IP 前缀也将动态地与它们进行交换。此时，从各 Edge（如果存在）中接收的新路由信息将推送至控制器，并相应地推送至 ESXi 主机的内核。

在 Edge 上安装静态路由的任务可以自动执行，因此每当新的逻辑交换机已创建后，两个 NSX Edge 网关上都将添加相应的静态路由，反之，如果逻辑交换机已删除或从 DLR 断开连接，则删除静态路由。

## ECMP HA 模式（从 NSX 6.1 版本开始）

NSX 软件版本 6.1 支持新的双活 ECMP HA 模式，可以将其视为之前介绍的独立 HA 模式的改进和演化版。

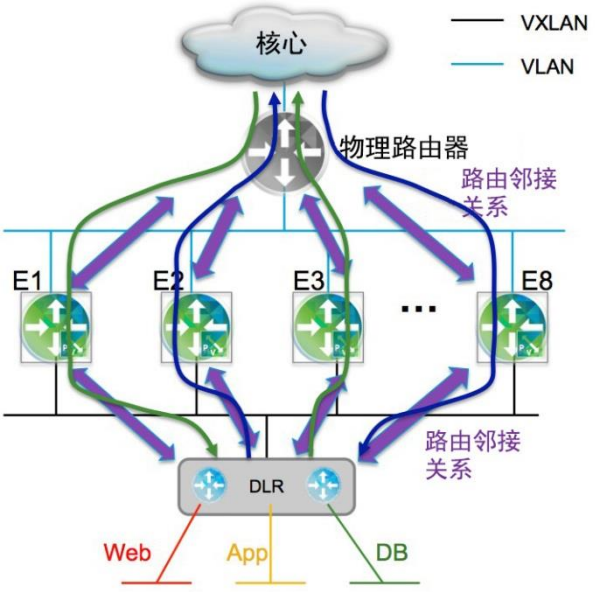



图 103 - NSX Edge ECMP HA 模式

在 ECMP 模式下，DLR 和 NSX Edge 功能得到了增强，可以在它们的转发表中支持多达 8 个等价路径。现在将关注点转移到 DLR 的 ECMP 功能上，利用这些功能，最多可以同时部署 8 个活动 NSX Edge，并且所有可用的控制和数据平面都将得到充分利用，如图 103 所示。

这一 HA 模式具有以下两大优势：

1. 增大南北向通信的可用带宽（每个租户高达 80 Gbps）。
2. 减少因 NSX Edge 故障导致流量中断（使用受影响流量的百分比表示）的情况。

从上图中的图示可以看出，流量很有可能按照非对称路径流动，其中同一通信从北到南和从南到北方向的流量分支由不同 NSX Edge 网关处理。DLR 先对来源于逻辑空间中工作负载的原始数据包源 IP 地址和目标 IP 地址进行哈希处理，在此基础上将南北向流量分发在各个等价路径中。而物理路由器分发从北到南方向的流量的方式取决于设备的特定硬件功能。

特定 NSX Edge 出现故障之后的流量恢复将以类似于在前文独立 HA 模式部分中所描述的方式进行，因为 DLR 和物理路由器将需要快速中断与故障单元之间的邻接关系，并利用剩余的活动 NSX Edge 网关重新对流量进行哈希处理（如图 104 所示）。

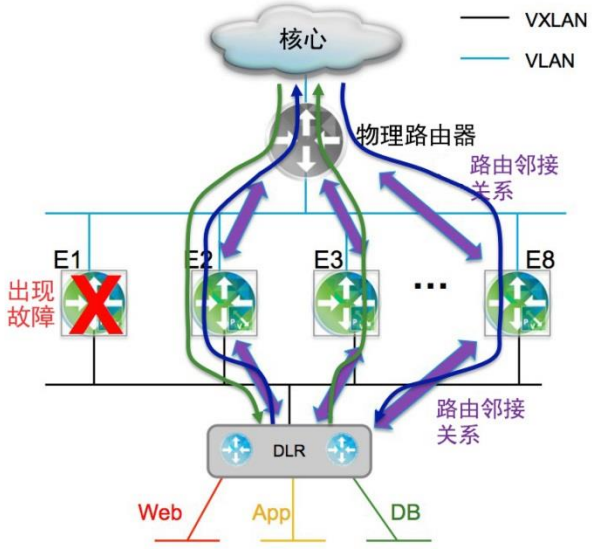



**图 104 - 在 ECMP NSX Edge 出现故障后恢复流量**

下面列出一些特定于此 HA 模式的部署注意事项：

- 再次说明，流量中断的持续时间取决于物理路由器和 DLR 控制虚拟机能够以多快的速度中断与故障 Edge 网关之间的邻接关系。因此，建议积极调整 Hello/保持时间计时器（1/3 秒）。
- 但是，与 HA 独立模式不同，现在一个 NSX Edge 出现故障只影响南北向流量（由出现故障的单元进行处理的流量）的一个子网。得益于这项重要因素，此 ECMP HA 模式拥有更出色的全面恢复功能。
- 当部署多个活动 NSX Edge 时，尚未对 NSX Edge 服务网关上提供的各种逻辑服务提供状态同步化支持。因此，建议使用 DFW 和单路并联式负载均衡器部署方式，并且只将路由服务部署在不同的 NSX Edge 网关上。
- 图 103 中的图示展示了 NSX Edge 网关北侧物理网络基础架构的简化视图。物理路由器的冗余对预计将始终部署在生产环境中，而这样会带来一个问题，即如何以最佳方式将 NSX Edge 连接到路由器（从逻辑角度）。

  首先要确定的是每个 NSX Edge 上应部署多少逻辑上行链路：建议的设计选择是逻辑上行链路数量要始终与托管 NSX Edge 的 ESXi 服务器上可用的 VDS 上行链路数量一一对应。

  在下方图 105 示例中，所有已部署的活动 NSX Edge 网关（从 E1 到 E8）都在通过两个上行链路连接到架顶交换机的 ESXi 主机上运行。因此，建议在每个 NSX Edge 上部署两个逻辑上行链路。由于 NSX Edge 逻辑上行链路与支持 VLAN 的端口组相连，需要使用两个外部 VLAN 网段来连接物理路由器，并建立路由协议邻接关系。

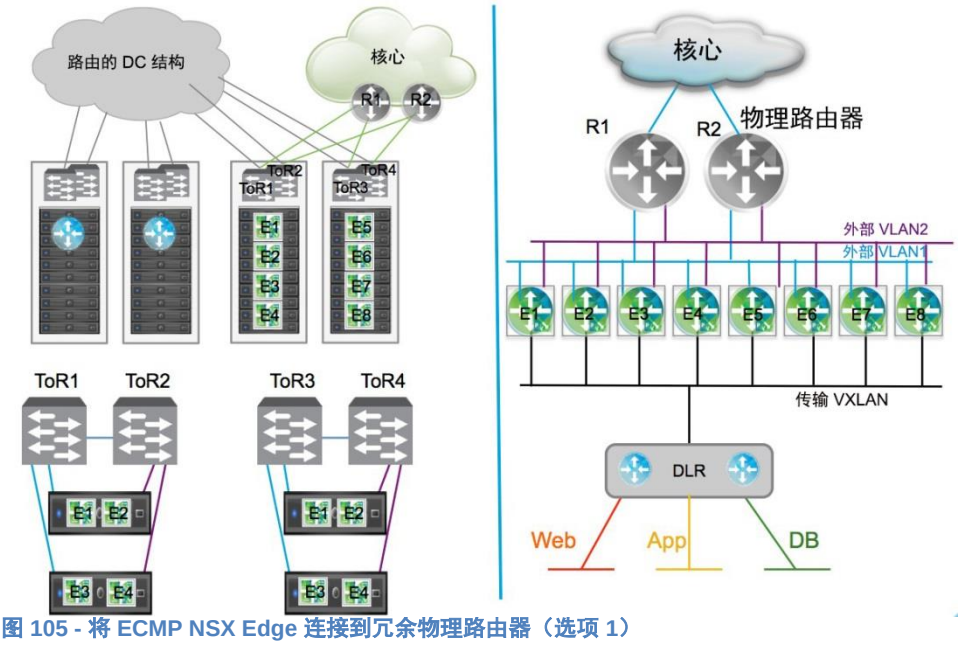



图 105 - 将 ECMP NSX Edge 连接到冗余物理路由器（选项 1）

然后，每个外部 VLAN 应仅连接在一个 ESXi 上行链路上（在上述示例中，“外部 VLAN1”连接在指向 ToR1/ToR3 的上行链路上，“外部 VLAN2”连接在指向 ToR2/ToR4 的上行链路上）。这样做的目的是，在正常情况下，两个 ESXi 上行链路可以同时用于发送和接收南北向流量，甚至不需要在 ESXi 主机和 ToR 设备之间创建端口通道。另外，在此模式下，ESXi 网卡出现物理故障相当于在该主机内部运行的 NSX Edge 出现逻辑上行链路故障，并且 Edge 将利用第二个逻辑上行链路（即第二个物理 ESXi 网卡接口）继续发送和接收流量。

为了打造一个恢复能力强、能够容许边缘机架完全中断的设计，我们还建议在两个独立的边缘机架中部署两组（四个为一组）Edge 网关。这表示需要扩展两个边缘机架之间的外部 VLAN；如前文所述，可以利用外部 L2 物理网络实现这一目的。

如果采用图 106 所示的第二种部署选项，则不需要跨越这些位于边缘机架之间的外部 VLAN。

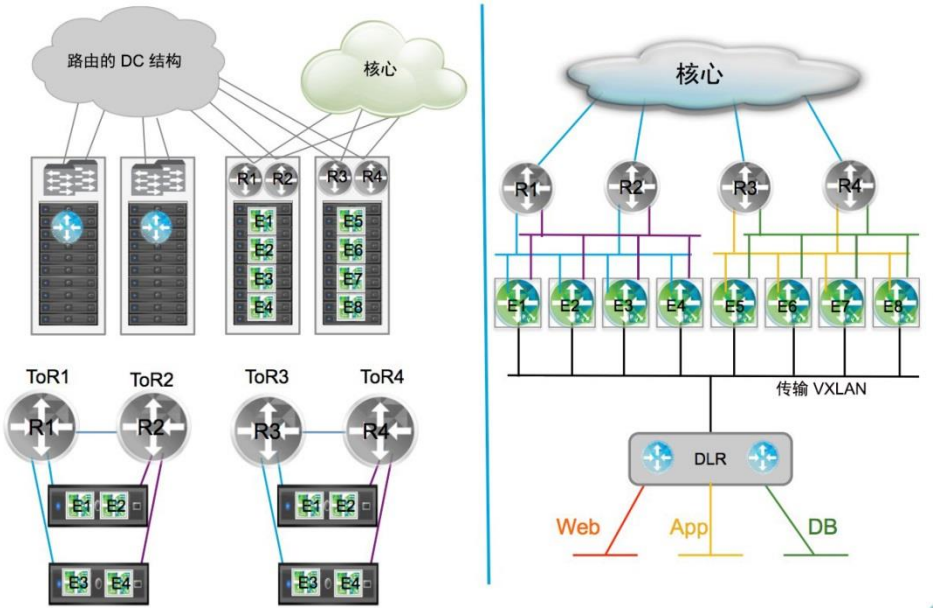



图 106 - 将 ECMP NSX Edge 连接到冗余物理路由器（选项 2）

此设计方案的相关要点如下：

- 两个单独的 VLAN 要在每个边缘机架中定义后才能将本地 NSX Edge 连接到物理路由器。
- 这些 VLAN 的范围仅限于每个机架。这表示，现在可以部署两对独立的物理路由器，并将它们定位为架顶式设备（而不是将边缘机架连接到外部 L2 基础架构）。部署于边缘机架 1 中的 NSX Edge 可与 R1 和 R2 建立路由对等关系，而部署于边缘机架 2 中的 NSX Edge 可与 R3 和 R4 建立对等关系。

- 与各自边缘机架中对等网段相关联的 VLAN ID 实际上可以完全相同（即 VLAN 10 和 VLAN 11 都可以用于边缘机架 1 和边缘机架 2），但是它们将与各自的 IP 子网相关联。
- 由于我们不需要扩展机架之间的 VLAN，因此建议确保 NSX Edge 虚拟机可以在属于同一边缘机架（绝不是跨机架）的 ESXi 主机之间动态移动（例如，通过 vSphere HA）。

- 最后，要针对独立 HA 模式考虑的有关处理控制虚拟机故障的注意事项，同样也适用于 ECMP HA 模式。所以，同样需要在所有已部署的 Edge 上，为所有在逻辑空间中创建并且需要从外部网络可以访问的 IP 子网配置静态路由。

## NSX 第 2 层桥接部署注意事项

利用 NSX L2 桥接功能，连接到物理基础架构中所定义 VLAN 的物理设备便可以访问逻辑网络连接。NSX L2 桥接必须能够终止 VXLAN 流量（源自给定逻辑交换机中的虚拟机）并对该流量进行封装解除，还要能够在特定 VLAN 上对它进行桥接。

NSX 桥接功能在“单播流量（虚拟到物理通信）”部分中有所介绍。现在重点讨论的是，部署 NSX L2 桥接功能对网络设计有何意义。

首先让我们回顾一下，给定 VXLAN-VLAN 对的 L2 桥接功能始终在一个特定的 ESXi 主机上执行。由于 L2 桥接实例配置始终与特定 DLR 的配置相关联，活动桥接功能将由特定 ESXi 主机执行，因为该 DLR 的活动控制虚拟机在这些主机中运行。

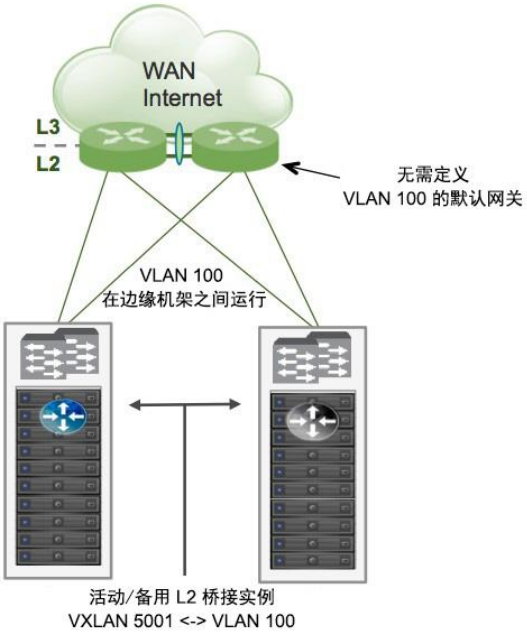



**图 107 - 活动/备用 NSX L2 桥接实例**

如图 107 所示，在给定 DLR 上定义的所有 L2 桥接实例在边缘机架 1 中的 ESXi 主机上都处于活动状态，相应的 DLR 活动控制虚拟机也在该主机中运行。也就是说，可以创建多组桥接实例并将它们与不同的 DLR 相关联，以便能够将桥接负载扩散到不同 ESXi 主机中。

以下是一些重要的部署注意事项：

- 如果部署了两个边缘机架来提升解决方案的恢复能力，建议将这些独立机架中的活动和备用控制虚拟机连接起来。这意味着，VXLAN 流量要桥接到的 VLAN 必须在这两个边缘机架中进行扩展，才能涵盖活动控制虚拟机从边缘机架 1 移动到边缘机架 2 的情况。可以选择使用外部 L2 物理基础架构扩展边缘机架之间的 VLAN。
- 在可能的情况下，建议将需要连接到逻辑网络的裸机服务器与边缘机架连接起来。如果要将它们所属的桥接 VLAN 的扩展仅限于这些机架，而不是连接到不同 ToR 交换机的其他计算机架，这一点非常重要。

  **注意**：当构建上图所示的独立 L2 物理基础架构时，可以选择部署一组裸机服务器的专用机架，并将它们的架顶式交换机直接连接到交换物理网络。
- 作为流量桥接目标的 VLAN（在上述示例中为 VLAN 100）将连接在前文介绍的相同交换基础架构上，并且将用于 L3 南北向通信。裸机服务器的默认网关可以采用两种方式部署，如图 108 所示。

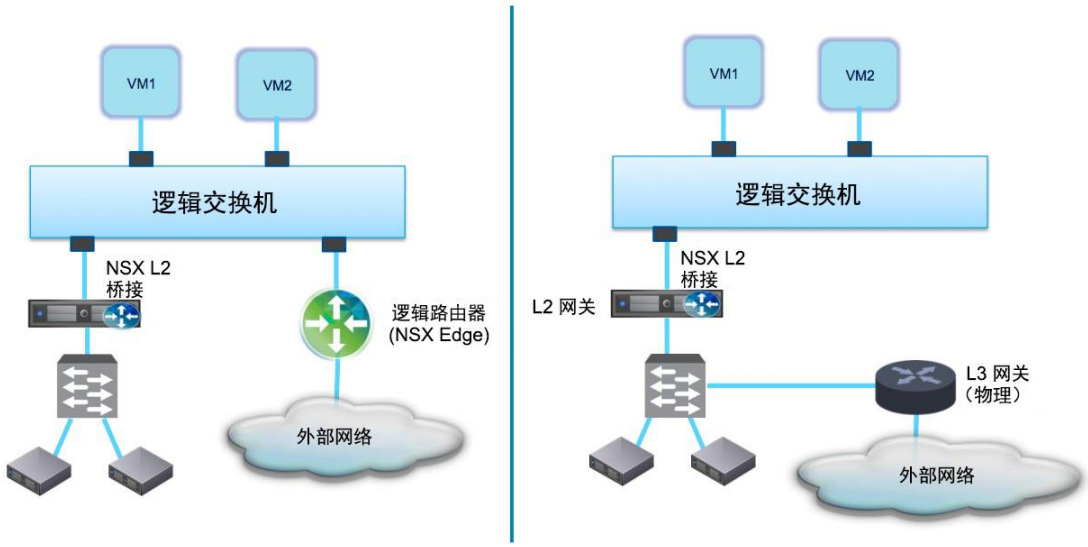



图 108 - 物理服务器的默认网关部署选项

1. 默认网关部署在逻辑空间中，物理服务器可以通过 NSX L2 桥接功能访问该网关。在这种情况下，需要指明一点：当前不支持在同时连接到分布式逻辑路由器的逻辑交换机上创建一个活动 L2 桥接。因此，与逻辑交换机和裸机服务器所连接虚拟机的默认网关需要放置在 NSX Edge 服务网关上。
2. 默认网关部署在物理空间中，连接到逻辑交换机的虚拟机可以通过 L2 网关服务访问该网关。

- 活动和备用控制虚拟机之间的运行状况消息可以利用 VXLAN 进行交换，因而不需要为此扩展其他 VLAN。或者，也可以在一个已扩展到各边缘机架的桥接 VLAN 上交换这些消息。

  **注意**：如前文所述，运行状况消息只能在内部 Edge 接口（而非上行链路）上进行交换。

# 总结

VMware NSX-v 网络虚拟化解决方案可通过物理网络基础架构应对当前挑战，并通过基于 VXLAN 的逻辑网络提供灵活性、敏捷性和扩展能力。除了能够使用 VXLAN 创建按需逻辑网络外，NSX Edge 服务网关还可帮助用户在这些网络上部署各种逻辑网络服务，例如防火墙、DHCP、网络地址转换和负载均衡。这源于它能够将虚拟网络与物理网络分离开，然后在虚拟环境中重现相应的属性和服务。

NSX-v 可以在逻辑空间中再现过去由物理基础架构提供的典型网络服务，如交换、路由、安全、负载均衡、虚拟专用网络连接等，并且 NSX-v 允许将连接性扩展到逻辑空间中，进而扩展到与外部网络连接的物理设备（如图 109 所示）。


**图 109 - NSX-v 提供的逻辑网络服务**

参考资源

[1]《VMware vSphere 5.5 中的新功能特性》(What’s New in VMware vSphere 5.5)
http://www.vmware.com/files/pdf/vsphere/VMware-vSphere-Platform-Whats-New.pdf

[2]《vSphere 5.5 最高配置》(vSphere 5.5 Configuration Maximums)
http://www.vmware.com/pdf/vsphere5/r55/vsphere-55-configuration-maximums.pdf