JUNIPER NETWORKS | Engineering Simplicity

参考架构

IP 交换矩阵 EVPN-VXLAN 参考架构

©2019, Juniper Networks, Inc.

目录

IP 交换矩阵 EVPN-VXLAN 解决方案概述 . . . 3
解决方案组件 . . . 3
IP 交换矩阵底层网络 . . . 3
网络虚拟化叠加 . . . 4
叠加服务 . . . 4
桥接叠加 . . . 4
集中路由叠加（又称主干路由） . . . 5
边缘路由叠加（又称叶路由） . . . 6
叠加设计和部署选项 . . . 6
边界设备 . . . 7
多宿主 . . . 8
针对以太网连接系统的多宿主支持 . . . 8
服务插入 . . . 8
数据中心互连 . . . 9
第 2 层 DCI . . . 9
第 3 层 DCI . . . 10
基于 EVPN 且已启用组播的数据中心 . . . 11
叠加环境中的组播 — 子网内组播转发 . . . 11
叠加环境中的组播 — 子网间组播转发 . . . 11
集中路由叠加架构中的组播 . . . 12
自动化数据中心 EVPN-VXLAN 交换矩阵 . . . 13
通过 Ansible/Saltstack 自动化 EVPN-VXLAN 交换矩阵 . . . 13
Contrail 企业多云 . . . 13
多云 . . . 13
遥测/分析 . . . 14
Contrail 见解 . . . 14
总结 . . . 16
关于瞻博网络 . . . 16

©2019, Juniper Networks, Inc.

简介

本参考架构介绍瞻博网络的以太网 VPN (EVPN) - 虚拟可扩展 LAN (VXLAN) 解决方案，这些解决方案可简化并自动执行与管理数据中心并将其连接扩展到其他数据中心或公共云服务相关的许多任务。

本文提供解决方案本身的功能描述，同时还介绍了组成解决方案的各个组件。

本文假定读者掌握了 IP 网络知识并且对数据中心技术有基本的了解。

IP 交换矩阵 EVPN-VXLAN 解决方案概述

传统上，数据中心使用第 2 层技术，如生成树协议 (STP) 和跨设备链路聚合组 (MC-LAG) 来连接计算和存储资源。随着这些数据中心横向扩展多租户网络，需要有一种新的数据中心架构能够实现底层（物理）网络与租户叠加网络的分离。

通过将第 3 层基于 IP 的底层与 EVPN-VXLAN 叠加相结合，数据中心、企业和云运营商可以部署比传统第 2 层基于以太网的架构更大型的网络。借助叠加，可以将服务器或虚拟机 (VM) 等端点放置在网络中的任何位置，并保持与同一逻辑第 2 层网络的连接，从而使虚拟拓扑与物理拓扑分离。

解决方案组件

IP 交换矩阵底层网络

在数据中心环境中，物理底层网络的作用是提供 IP 交换矩阵。IP 交换矩阵也称为 Clos 网络，负责提供从任意物理设备（服务器、存储设备、路由器或交换机）到任意其他物理设备的单播 IP 连接。理想的底层网络可以提供从网络中任意一点到网络中任意其他点的低延迟、无阻塞、高带宽连接。

IP 交换矩阵的大小和规模可以有所不同。典型的解决方案使用主干和叶两层架构来构建三级 IP 交换矩阵，其中每个叶设备都连接到每个主干设备，如图 1 所示。

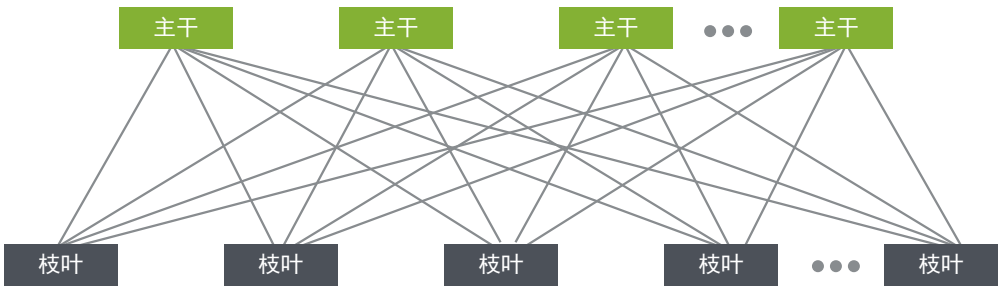

图 1：三级 IP 交换矩阵

©2019, Juniper Networks, Inc.

随着交换矩阵规模的扩大，有必要将其扩展为五级 IP 交换矩阵，如图 2 所示。该方案增加了一个交换矩阵层（或“超级主干”），用于提供 Pod 之间或数据中心之间的连接。

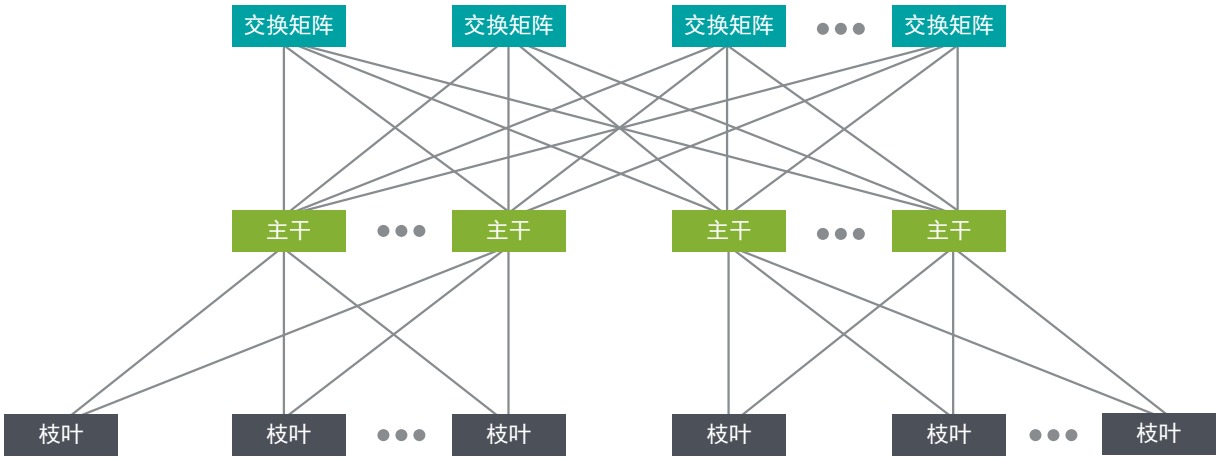

图 2：五级 *IP* 交换矩阵

基于 IP 的交换矩阵的主要优势就是它自身的天然弹性。它完全不需要诸如 MC-LAG 或瞻博网络的虚拟机箱技术等高可用性机制，因为 IP 交换矩阵在每一层和每个设备上都使用多个链路。物理网络基础架构本身即可提供弹性和冗余。

本文中描述的参考架构使用 EBGP 作为底层网络中的路由协议来获得依赖性和可扩展性。每个主干和叶设备都有自己的自主系统，使用唯一的 32 位自主系统编号来支持 EBGP。虽然数据中心底层网络中还可以使用 OSPF/IS-IS 等其他路由协议，但本文中我们不对此加以介绍。

网络虚拟化叠加

*网络虚拟化叠加*是在 IP 底层之上进行传输的一种虚拟网络，该功能构建块支持网络中存在多租户，允许多个租户共享单个物理网络，同时使每个租户的网络流量与其他租户隔离。

一个租户就是一个用户社区（如业务单位、部门、工作组或应用），其中会包含多个端点组。这些组可以与同一租用内的其他组通信，并且在网络策略许可的情况下，租户可以与其他租户通信。组通常表示为子网 (VLAN)，它可以与同一子网中的其他设备通信，通过虚拟路由和转发 (VRF) 实例可以到达外部组和端点。

叠加服务

IBGP 是一种路由协议，用于在 IP 网络上交换可达性信息。IBGP 与多协议 IBGP (MP-IBGP) 结合使用时，将允许 EVPN 与 VXLAN 虚拟隧道端点 (VTEP) 设备交换可达性信息。这是建立 VTEP 间 VXLAN 隧道并使用它们提供叠加连接服务所必需的。

以下部分将讨论作为 EVPN-VXLAN 参考架构的组成部分的各种叠加服务模型。

桥接叠加

在桥接叠加模型中（请参见图 4），在跨 VXLAN 隧道的叶设备之间扩展以太网 VLAN。这些叶到叶 VXLAN 隧道支持叶设备之间需要以太网连接、但 VLAN 之间不需要路由的数据中心网络。因此，主干设备只为叶设备提供基本底层和叠加连接，而不执行使用其他叠加方法显示出的路由或网关服务。Juniper Networks® MX 系列 5G 通用路由平台或 SRX 系列服务网关部署在 EVPN/VXLAN 交换矩阵外部，可用来执行必要的路由。

©2019, Juniper Networks, Inc.

叶设备会建立 VTEP 以便连接到其他叶设备。隧道使叶设备能够将 VLAN 流量发送到其他叶设备以及以太网连接数据中心内的终端系统。对于期待通过简单方式将 EVPN/VXLAN 引入基于以太网的现有数据中心网络的运营商而言，这种叠加服务的至简特性具有极大吸引力。

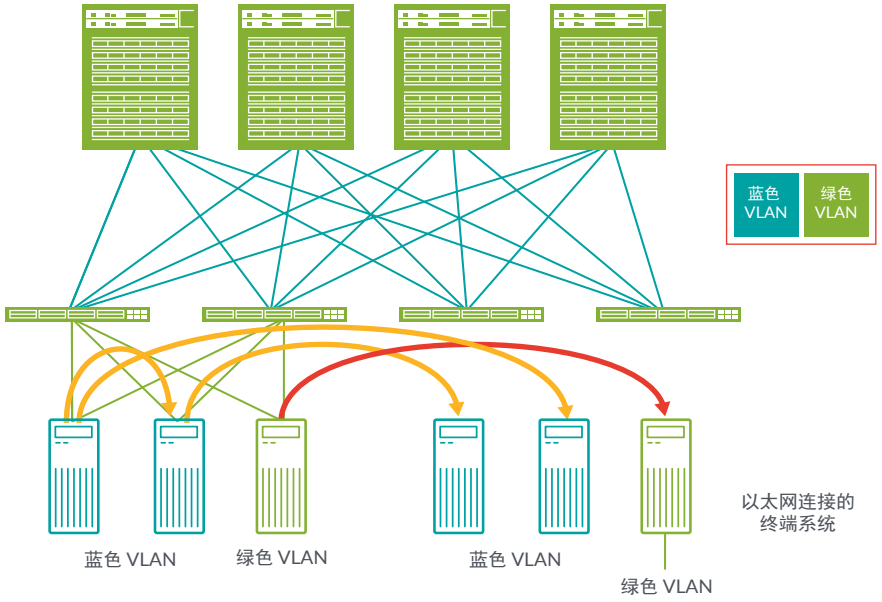

图 3：桥接叠加

集中路由叠加（又称主干路由）

集中路由桥接叠加的实质是，路由发生在数据中心网络的中央网关上（本例中为主干层），而不是发生在连接有终端系统的 VTEP 设备上（本例中为叶层）。当路由的流量需要通过集中式网关时，或者当边缘 VTEP 设备缺乏所需的路由能力时，您可以使用这种叠加模型。如图 4 所示，以太网连接终端系统发出的流量通过中继（多个 VLAN）或接入端口（单个 VLAN）转发至叶 VTEP 设备。该 VTEP 设备将流量转发至本地终端系统或远程 VTEP 设备上的终端系统。每个主干设备上的集成路由和桥接 (IRB) 接口在以太网虚拟网络之间路由流量。

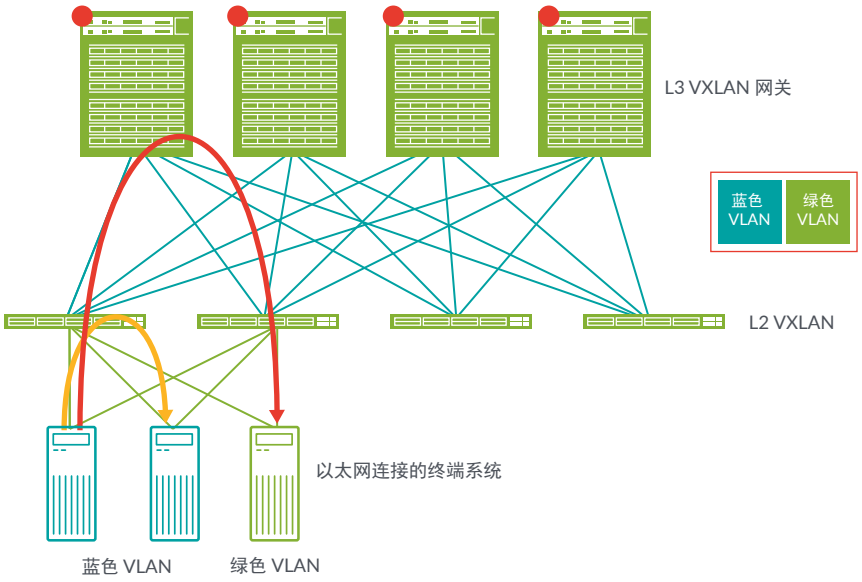

图 4：集中路由叠加

©2019, Juniper Networks, Inc.

边缘路由叠加（又称叶路由）

在这种以太网服务模型中，IRB 接口被移至位于叠加网络边缘的叶设备 VTEP 上，从而使 IP 路由更靠近终端系统。由于必须具备特殊的 ASIC 功能才能在一台设备中同时支持桥接、路由和 EVPN/VXLAN，因此边缘路由桥接叠加只能在某些特定的交换机上使用，例如，高性能、多用途的瞻博网络 QFX 10000 系列交换机。

此选项支持更快的服务器到服务器流量以及数据中心内的流量（也称为东西向流量），其中通信端系统连接至同一叶设备 VTEP。因此，与使用集中路由桥接叠加相比，这种路由更靠近终端系统。它还使整个网络更简单。

在这种模型中，主干设备被配置为只处理 IP 流量，因此无需将桥接叠加扩展到主干设备。

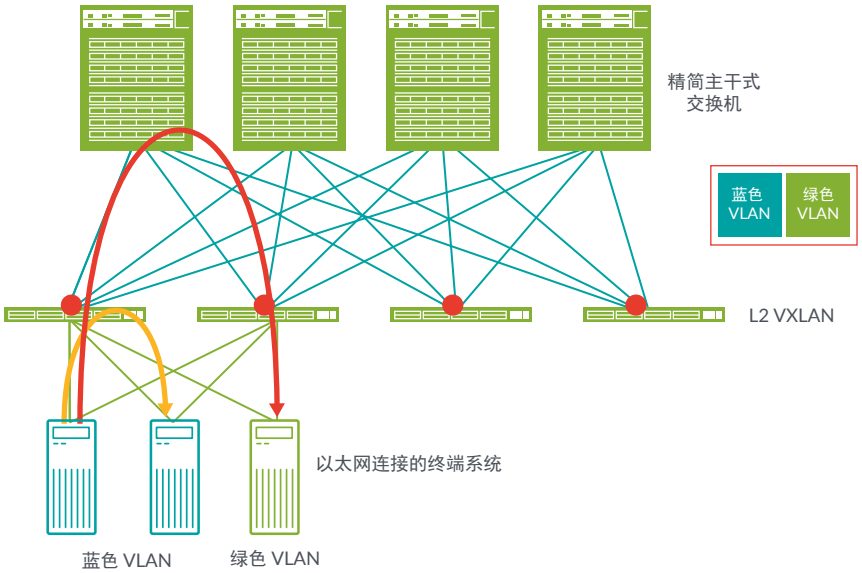

图 5：边缘路由叠加

叠加设计和部署选项

运营商可以通过多种方式来设计其叠加服务。表 1 显示了集中路由与边缘路由的部署注意事项。

表 1：集中路由与边缘路由注意事项

何时使用集中路由	何时使用边缘路由
当绝大部分为南北向流量时（比较典型的是在园区网络中）	当绝大部分为东西向流量而且 Pod 内存在大量分段时
当首选使用集中租户 IP/VRF 管理时	当规模较大且需要分配状态以缩短发生故障时的融合时间时
当叶设备不支持 VXLAN 路由时	当配置复杂性被控制器抽象化时

©2019, Juniper Networks, Inc.

边界设备

边界设备连接至外部设备（如 WAN 或交换矩阵外部）或连接至服务，包括（但不限于）防火墙、负载平衡器、域名系统 (DNS)、动态主机配置协议 (DHCP) 等。边界设备可以位于叶（图 6）或主干（图 7）上，完全独立于在叶或主干设备上的网关。

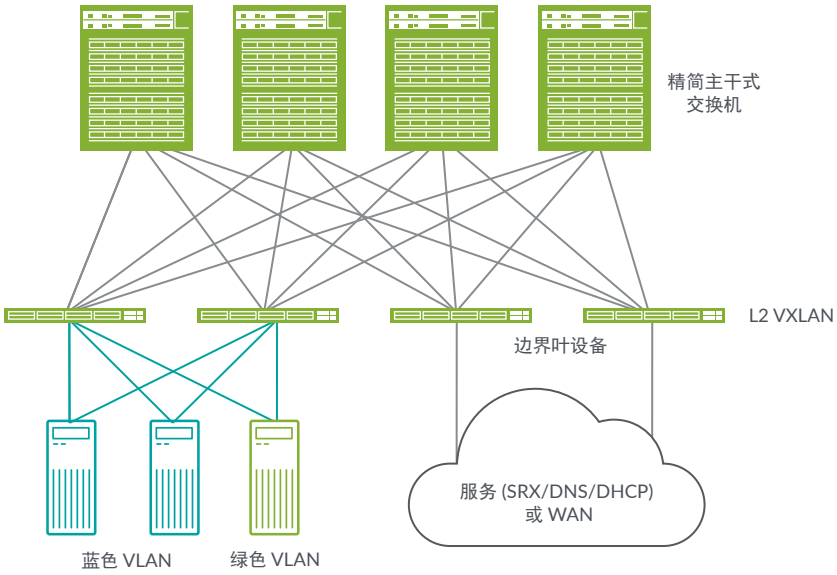

图 6：边界叶设备

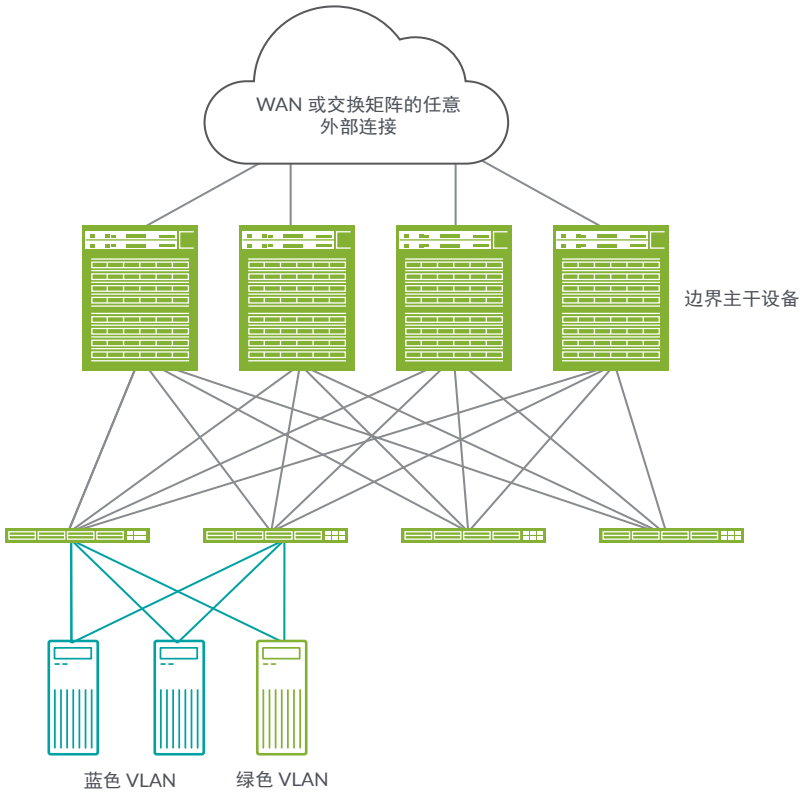

图 7：边界主干设备

©2019, Juniper Networks, Inc.

多宿主

针对以太网连接系统的多宿主支持

通过以太网连接多宿主技术，可以使以太网流量跨交换矩阵（位于连接到同一终端系统的不同叶设备上的 VTEP 之间）实现负载均衡。

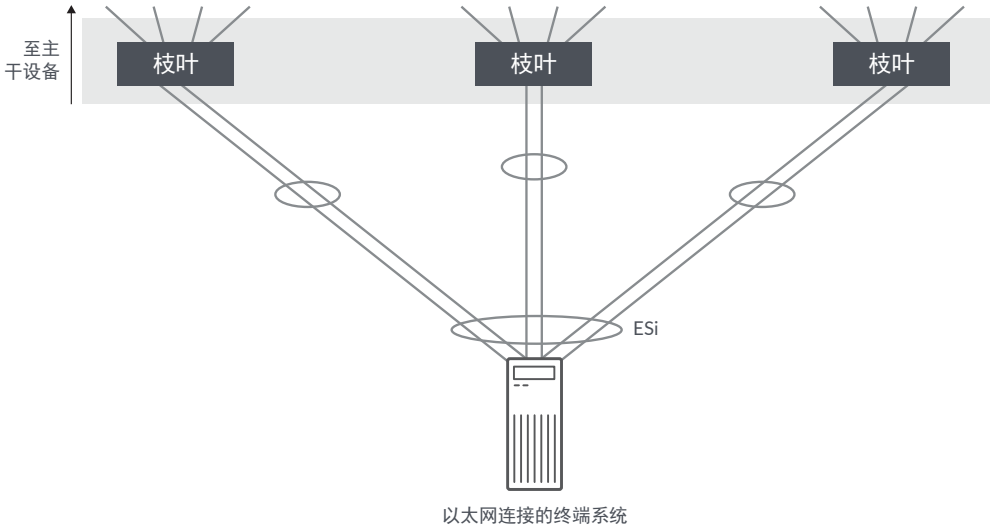

图 8：以太网连接终端系统多宿主

服务插入

每个数据中心都有与之相关的服务，包括（但不限于）防火墙、负载平衡器、DNS、DHCP 等。向网络中插入服务主要有两种方法。图 9 显示了一个服务块架构，其中的 SRX 系列群集直接连接至主干设备。

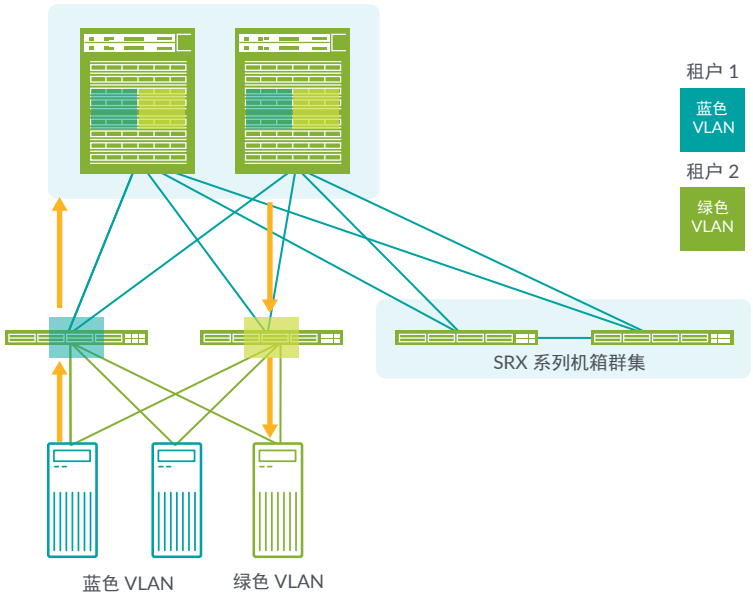

图 9：直接连接至主干设备的服务块

使用这种方法时，服务块直接连接至主干设备，集中度较低，但它仍然可以提供数据中心可能需要的任何服务。

©2019, Juniper Networks, Inc.

这种部署模型通常用于以下用例：

- 不介意其主干设备是否承担多种角色的运营商
- 使用分布式模型的运营商，其中的服务可能存在于数据中心的不同点上
- 主干设备端口可以腾出来连接服务设备的数据中心

图 10 显示了第二种插入方法：使用公共“服务块”将服务插入网络，该服务块通过一个或多个“服务叶”设备连接到网络。

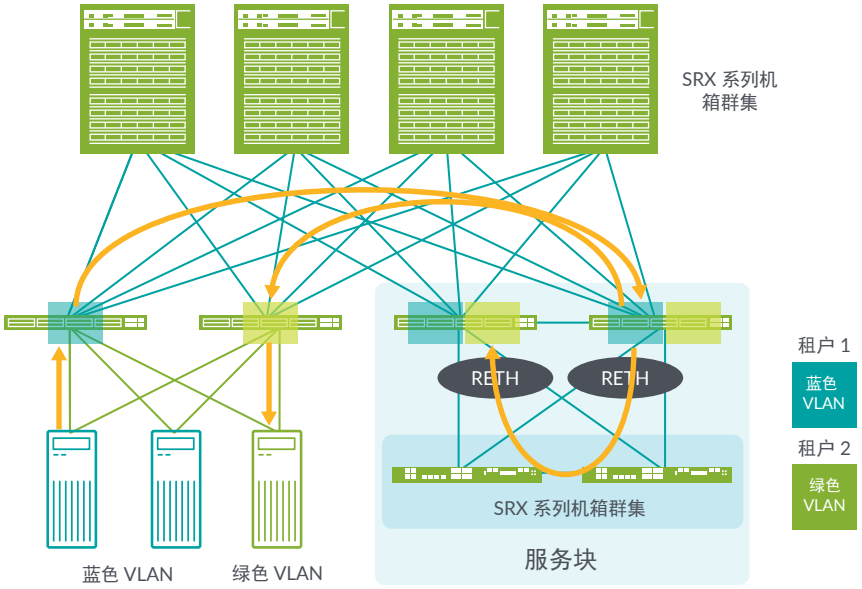

图 10：连接至服务叶设备的服务块

连接至服务叶设备的服务块架构通常用于将服务集中到数据中心的公共位置。服务块可用于提供任何服务，如防火墙、负载平衡器、DNS 服务器、DHCP 服务器和数据中心可能需要的任何其他类型的通用网络服务。

连接至服务叶设备的服务块通常部署于以下用例：

- 不希望其主干设备成为任何 EVPN-VXLAN 通信的端点的运营商
- 希望集中管理其服务的运营商
- 主干设备端口有限，必须保留端口以便连接叶设备的大型数据中心
- 需要更简单的路由策略的环境

数据中心互连

大多数企业现在都部署了多个数据中心。当使用多个数据中心时，通常需要将其互连起来。EVPN-VXLAN 使用数据中心内部使用的相同协议来创建 DCI 叠加，以简化数据中心互连 (DCI) 功能。

有两个 DCI 部署选项：第 2 层和第 3 层。请注意，虽然这些概念是在 DCI 上下文中提出的，但这些相同的概念也同样适用于数据中心中的 Pod 间连接。

第 2 层 DCI

在 L2 DCI（也称为“L2 扩展”）中，VLAN 及其相关的 IP 子网横跨两个或更多的数据中心。只有那些需要此功能的 VLAN 才应该被扩展。

©2019, Juniper Networks, Inc.

第 3 层 DCI

通过使用 EVPN Type 5 路由在 EVPN-VXLAN 环境中实现 L3 DCI。子网必须只存在于单个数据中心内；否则，它将成为一个 L2 DCI 连接，正如前面所解释的，只有那些需要此功能的 VLAN 才应该被扩展。EVPN Type-5 路由是实现 L3 DCI 的推荐方法。

图 11 显示了带有边界主干设备的 DCI 参考模型，图 12 显示了类似的参考模型，通过服务叶设备关闭 DCI 流量。

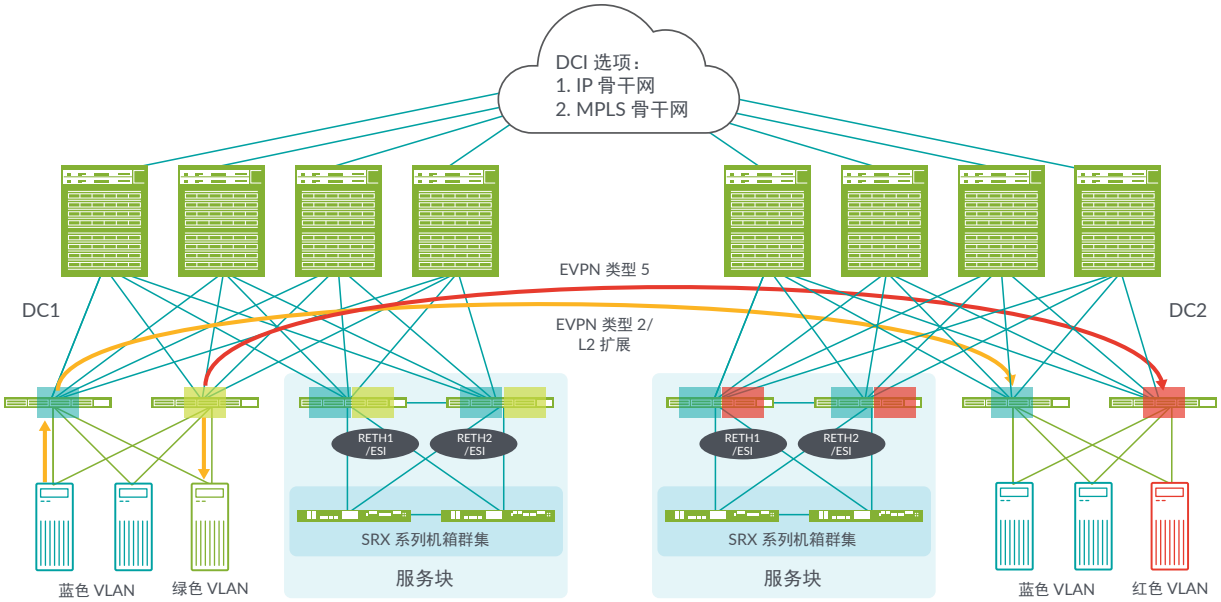

图 11：带有边界主干设备的 DCI 参考架构

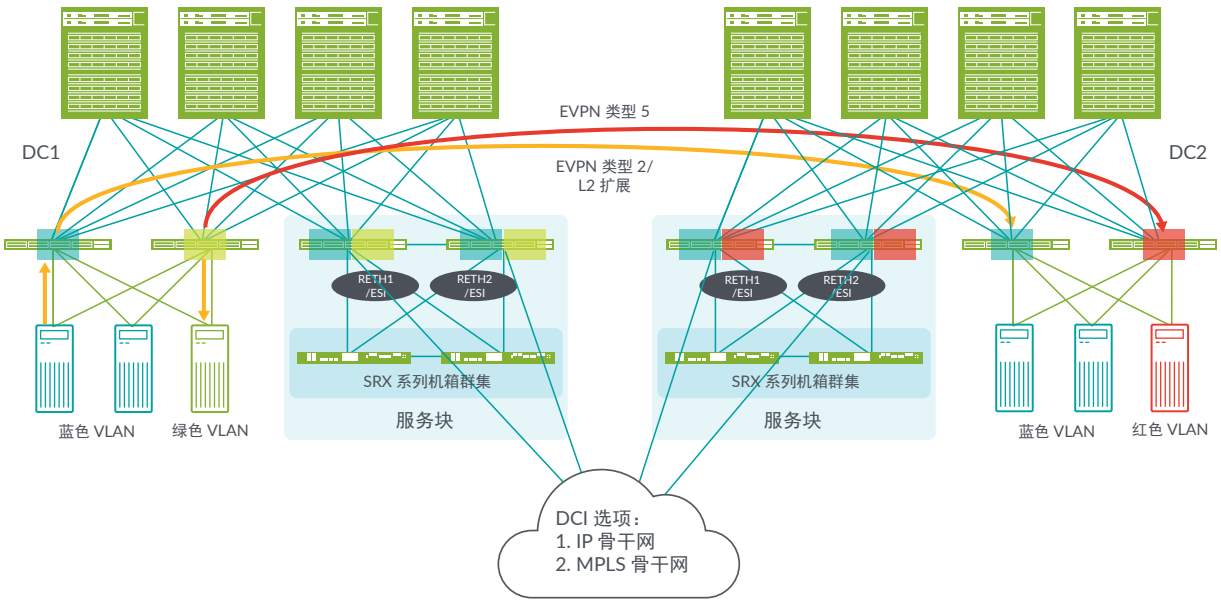

图 12：带有边界叶设备的 DCI 参考架构

©2019, Juniper Networks, Inc.

基于 EVPN 且已启用组播的数据中心

EVPN 环境中的组播相对简单，因为 EVPN-VXLAN 交换矩阵中的广播、未知单播和组播 (BUM) 流量将复制到属于同一虚拟网络标识符 (VNI) 的所有其他 VTEP（也称为入口复制或头端复制）。对于较小的部署来说，这是一种更简单的解决方案，因为底层网络中不需要任何组播协议。但是，对于采用支持组播的基础架构的大型部署，仅仅入口复制就可能导致可扩展性和性能问题。在叠加环境中，实施组播优化对于克服入口复制所固有的这些局限性至关重要。

叠加环境中的组播 – 子网内组播转发

如图 13 所示，在具有选择性组播 (SMET/EVPN Type 6) 优化的网络中，复制不会发送到没有活动接收器的叶 3。在缺少 SMET 优化的情况下，同一入口叶设备将承担复制多个副本的重担，而所有叶设备均将接收组播流量，这造成了交换矩阵带宽的浪费。此外，在将主干配置为辅助复制器 (AR) 的情况下，入口叶 1 会将复制任务委托给可感知 VNI 的主干。

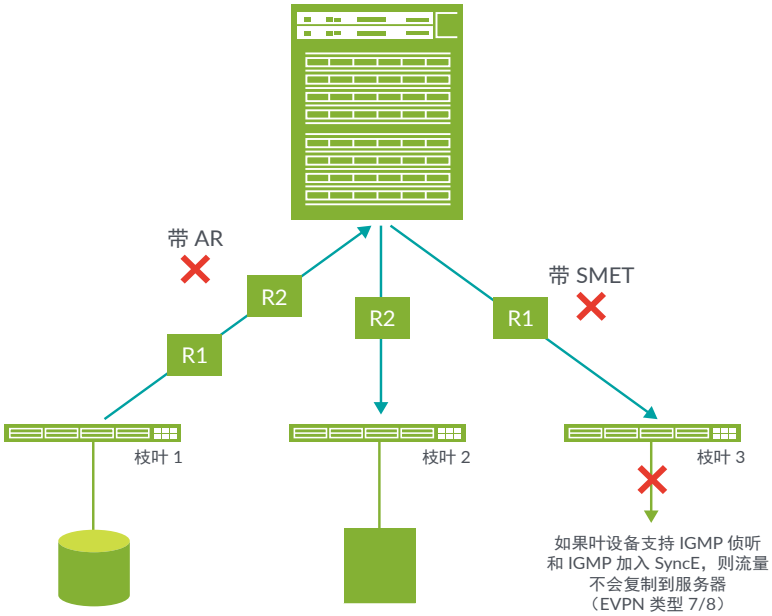

图 13：子网内叠加组播转发

叠加环境中的组播 – 子网间组播转发

无论采用集中路由还是边缘路由，VNI 间组播路由都是在主干上完成的。在图 14 中，叶 1 和叶 2 都有接收器，并且由主干发送多个副本，导致本地接收器（接收器 1）出现“发夹”现象。

©2019, Juniper Networks, Inc.

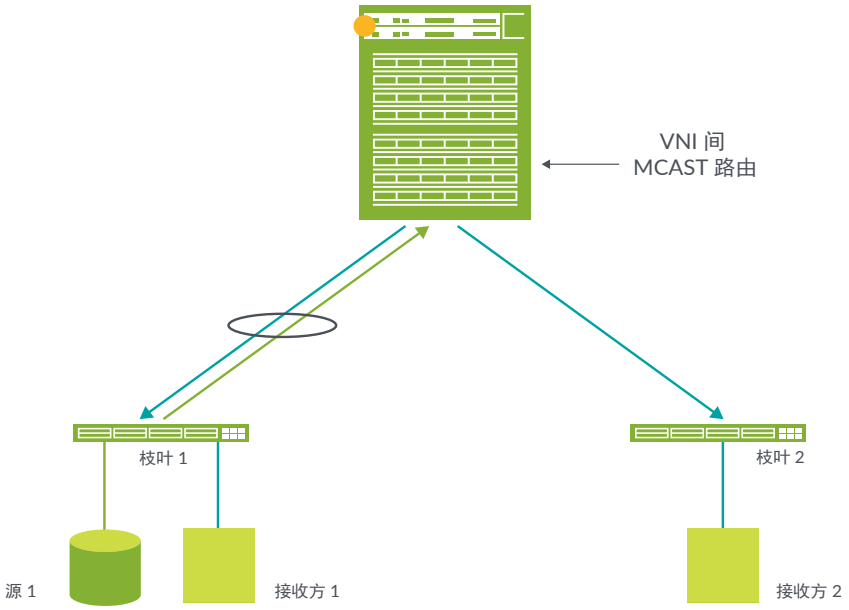

图 14：子网间组播转发

集中路由叠加架构中的组播

当已启用组播且采用集中路由的 EVPN-VXLAN 交换矩阵通过边界叶设备连接至远程数据中心或传统园区网络时，如果交换矩阵外有源和接收器，则组播网关通常配置在边界设备上（请参见图 15）。

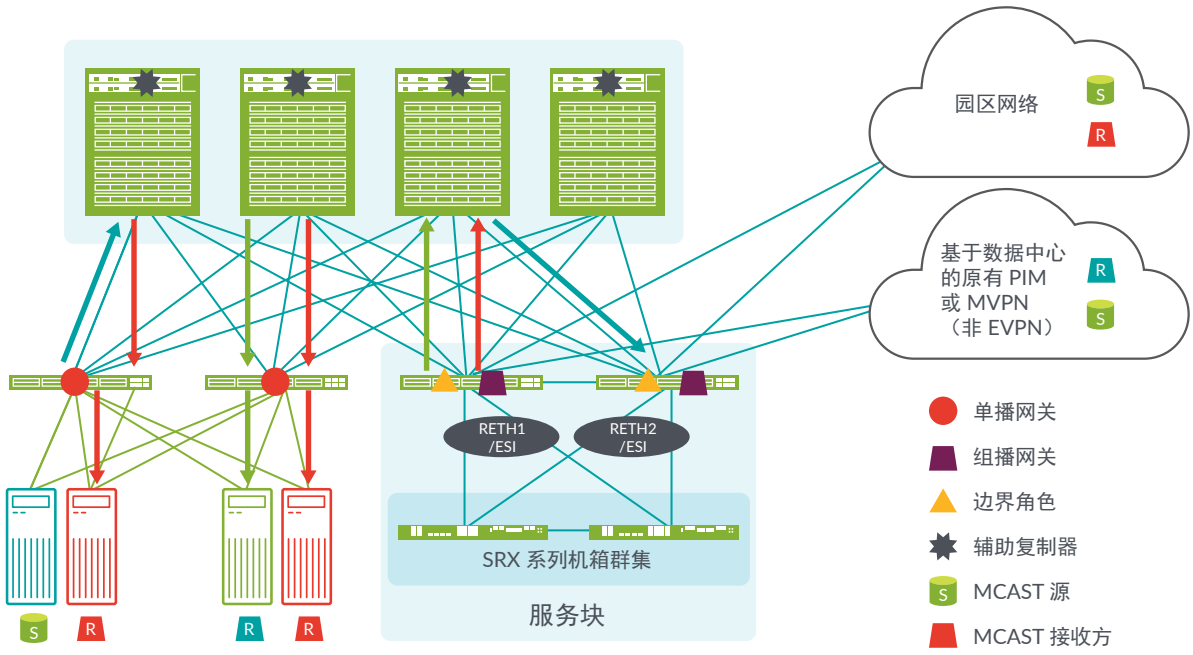

图 15：已启用组播且采用集中路由的叠加架构

©2019, Juniper Networks, Inc.

自动化数据中心 EVPN-VXLAN 交换矩阵

可通过 CLI，也可使用 Ansible/Saltstack 等自动化工具来配置 EVPN-VXLAN。您还可以通过 Juniper Contrail® 企业多云及其统一托管界面 Juniper Contrail Command 等多云编排器来编排底层和叠加设置。图 16 显示了可供客户自动化其 EVPN-VXLAN 交换矩阵的选项。

通过 Ansible/Saltstack 自动化 EVPN-VXLAN 交换矩阵

通过结合使用 Ansible/Saltstack 及瞻博网络的自动化和编排工具，将各个组织中有经验的 IT 解决方案开发人员、运维人员和管理人员的知识和专业技能汇聚在一起。联合解决方案解决了 IT 资源管理和调配中最常见的自动化和编配应用问题，包括配置自动化、测试驱动联网和持续合规性。

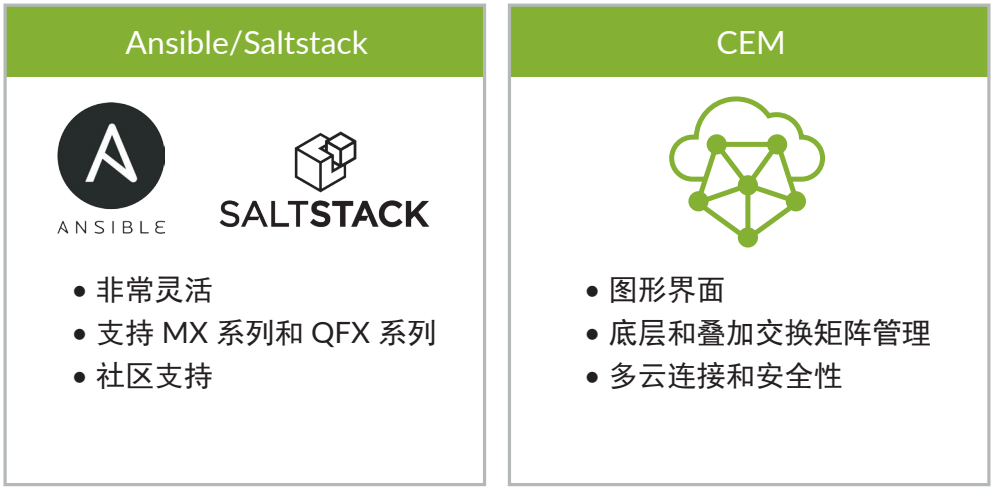

图 16：自动化 EVPN-VXLAN 交换矩阵

Contrail 企业多云

瞻博网络 Contrail 企业多云的作用是实现万物互联，从而在整个多云环境中实现自动化运维、可见性和系统集成。

为此，所有设备都必须是交换矩阵的组成部分，使得数据中心和私有云环境能够扩展到公共云，并跨不同的域创建统一架构。这就创建了一个同质管理环境，在这个环境中，不管资源位于哪里，都可以以相同方式使用。

多云

多云的目标是让您能够根据业务和职能需求（例如成本）将工作负载部署到任何位置。不应让用户察觉到工作负载由私有云还是公共云提供。

要做到这一点，网络必须最终实现与应用层（包括连接性和安全性）的集成，并控制部署和使用新应用程序及服务的方式。多云架构的这个顶层通过透明地将较低层抽象化到每个应用所需的服务集中，从而实现了基础架构与服务的分离。

©2019, Juniper Networks, Inc.

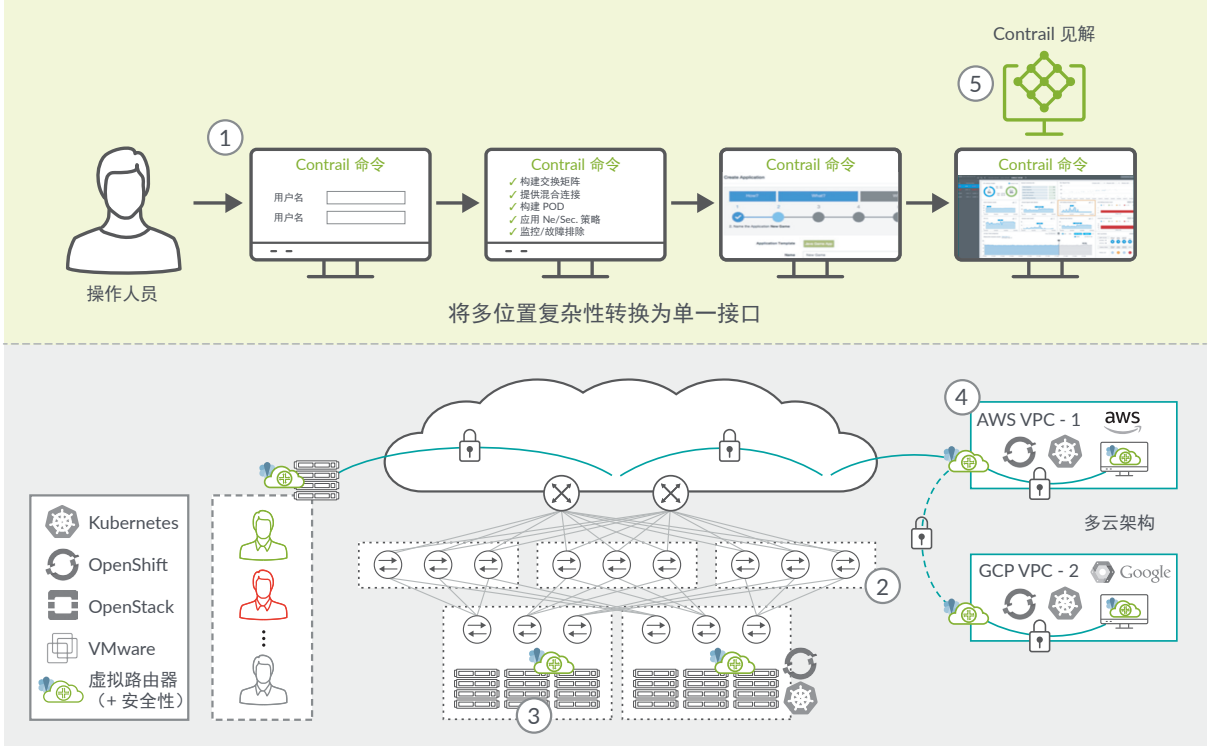

图 17：Contrail 企业多云参考架构

遥测/分析

诸如 SNMP 和 CLI 轮询模型等传统管理方法在可扩展性和效率方面存在一定的限制。Juniper Junos® Telemetry Interface (JTI) 解决方案通过采用异步推送模型等机制消除了传统轮询，从而克服了这些问题。高度可扩展的 JTI 可以监控瞻博网络的物理节点和虚拟节点中的数千个传感器。它还支持实时运行数据流传输，以同步外部控制器和分析平台的运行状态，因此，与传统方法相比，它可以帮助您更快速地制定决策。

Contrail 见解

瞻博网络 Contrail 见解可以提供多云环境中的端到端可见性（如图 18 和 19 所示），有助于消除任何潜在的网络问题，使操作更简单、更有效。它使您能够可视化并分析物理和虚拟环境，使用监控和基于意图的分析将各种资源的原始数据转换为可立即使用的格式。

Contrail 见解网络设备监控功能通过 JTI 和 OpenConfig，可以为数据中心和园区网络设备（包括 QFX 系列和 EX 系列交换机）提供实时基础架构性能监控 (IPM)。

云基础架构和应用监控功能针对 DevOps 和云原生环境提供实时监控和分析，支持混合使用私有云中的高性能裸机系统、OpenStack 中的虚拟机以及 Kubernetes 环境中的容器。Contrail 见解还可用来监控和分析在公共云上运行的应用，如 Amazon Web Services (AWS)、Microsoft Azure 和 Google Cloud Platform (GCP)。

©2019, Juniper Networks, Inc.

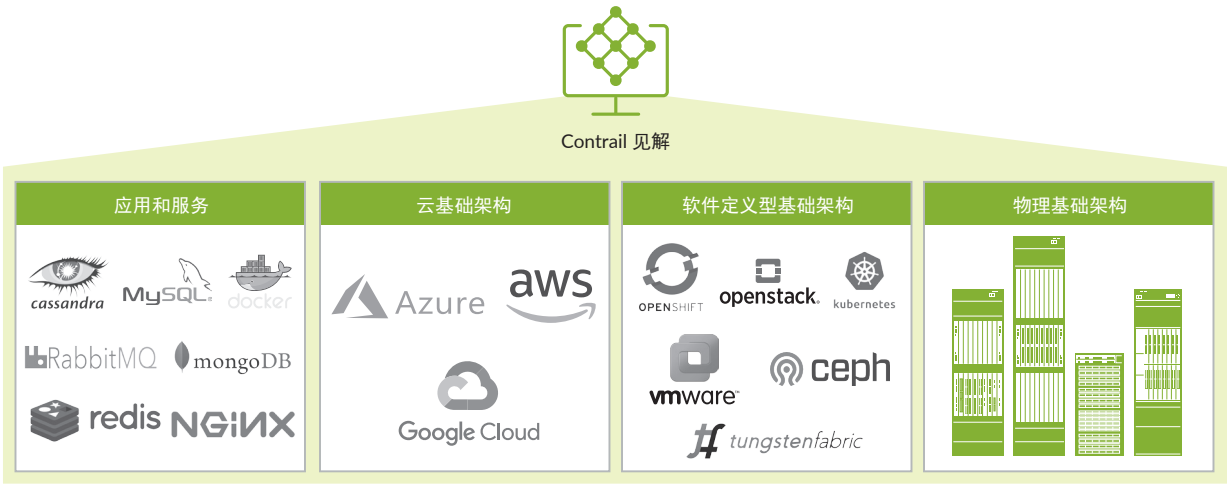

图 18：跨层可见性和分析

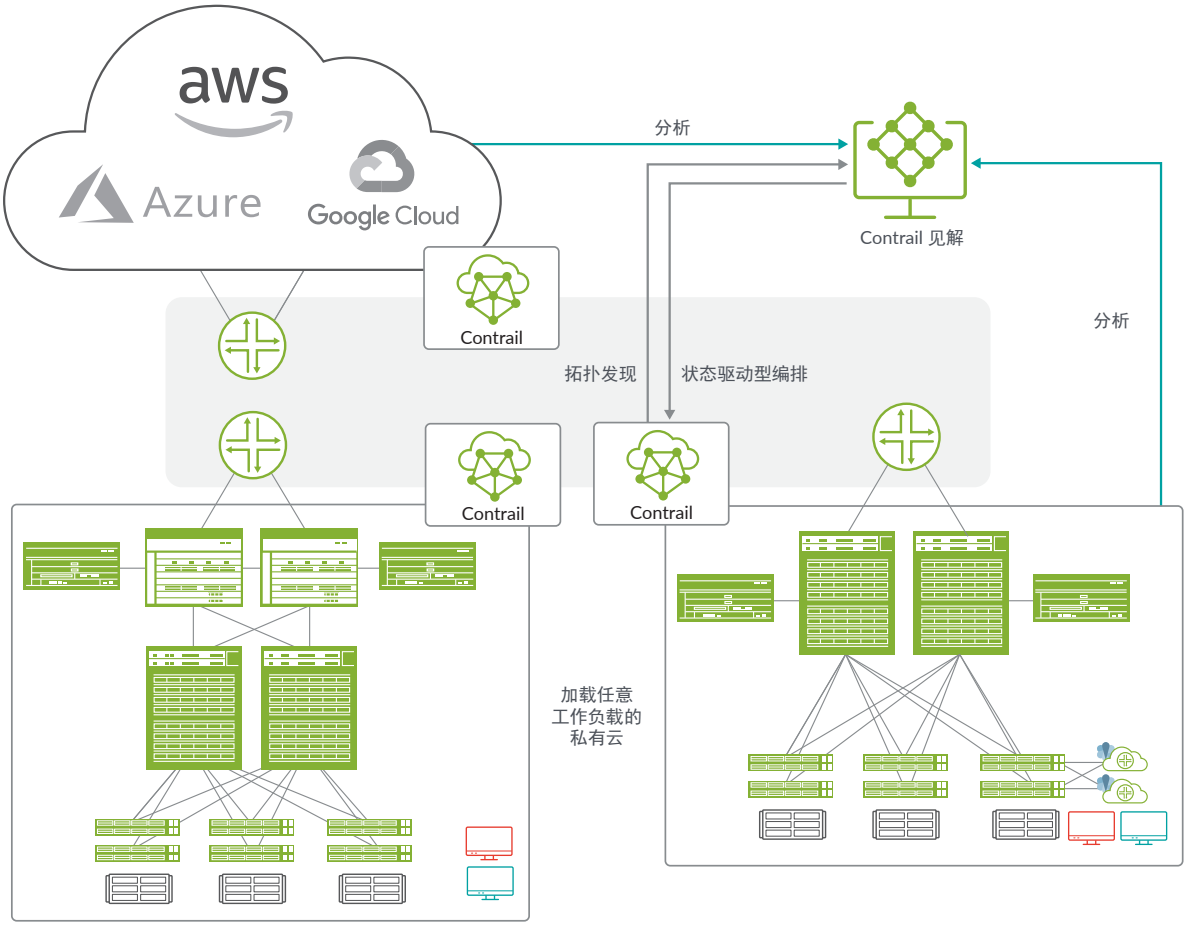

图 19：Contrail 企业多云环境中的分析

©2019, Juniper Networks, Inc.

总结

瞻博网络的 EVPN-VXLAN 部署代表一种现代的、开放的、基于标准的且支持原生自动化的控制平面，可以解决多种技术问题，帮助组织跨传统、本地和基于云的处理迁移和处理数据。瞻博网络参与了 EVPN 技术的设计、标准化和实施，这一方面展示了我们解决世界上最复杂问题的能力，另一方面也验证了我们推动精研至简的指导原则。

有关如何设计和部署叠加网络的更多信息，请参阅技术库中的云数据中心蓝图架构组件。

关于瞻博网络

瞻博网络将简单性融入到全球互联的产品、解决方案和服务之中。通过工程创新，我们消除了云时代网络的限制和复杂性，可应对我们的客户和合作伙伴日常面临的严苛挑战。在瞻博网络，我们坚信，网络是分享知识和实现人类进步的资源，它将改变这个世界。我们致力于开创具有突破性的方式，提供自动化、可扩展且安全的网络，以满足业务发展的需求。

公司和销售总部
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
电话：888.JUNIPER (888.586.4737)
或 +1.408.745.2000
传真：+1.408.745.2100
www.juniper.net

APAC 和 EMEA 总部
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
电话：+31.0.207.125.700
传真：+31.0.207.125.701

JUNIPER NETWORKS | Engineering Simplicity

版权所有 2019 Juniper Networks, Inc. 保留所有权利。Juniper Networks、Juniper Networks 徽标、Juniper 和 Junos 是 Juniper Networks, Inc. 在美国和其他国家/地区的注册商标。所有其他商标、服务标识、注册商标或注册服务标识均为其各自所有者的资产。瞻博网络对本文档中的任何不准确之处不承担任何责任。瞻博网络保留对本出版物进行变更、修改、转换或以其他方式修订的权利，恕不另行通知。