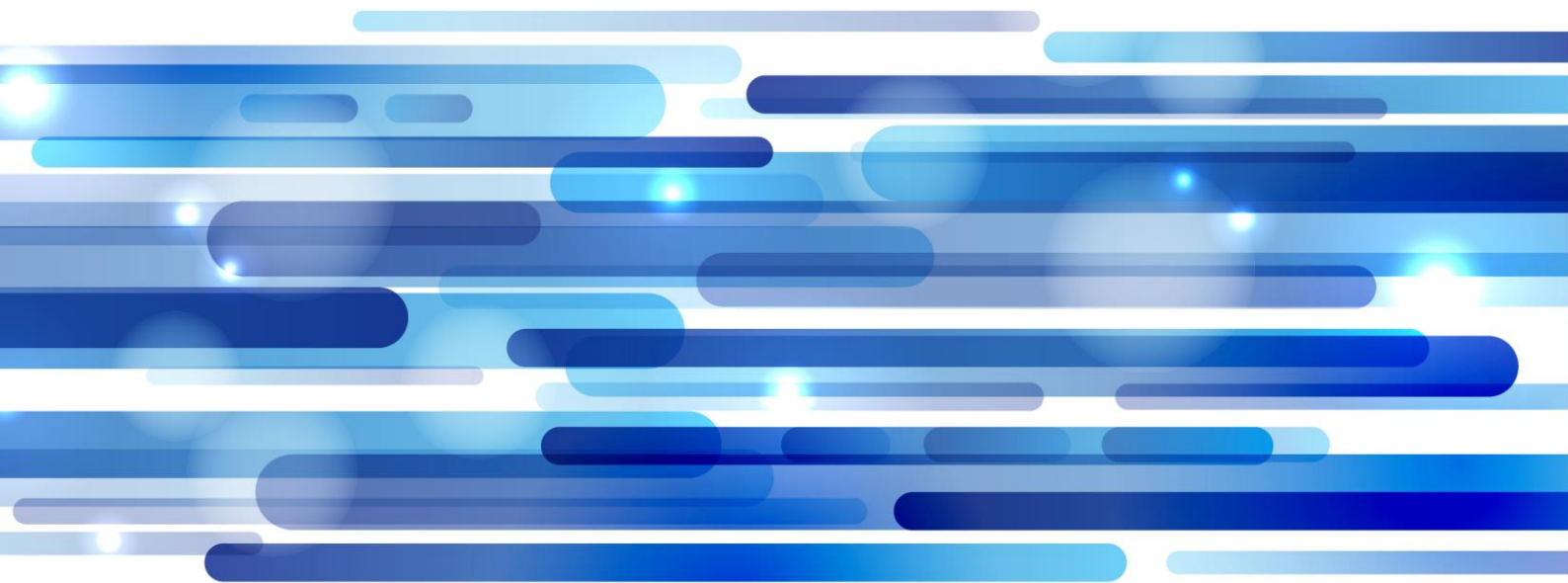




华为数据中心 网络设计指南



从零开始带您了解数据中心网络的基本知识。从交换机简介到布线方案,从收敛比的设计到Fabric网络技术的应用,从Overlay网络的起源到VXLAN技术讲解。阅读后,您将清楚如何使用CloudEngine系列交换机构建自己的数据中心。

版权所有 © 华为技术有限公司 2018。 保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

商标声明



HUAWEI 和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： <http://e.huawei.com>

前言

云计算、大数据、人工智能等技术的快速发展，对承载数据流量的数据中心网络提出了更高的要求，高吞吐量、高可靠性、低时延、适应服务器虚拟化等都是业务对数据中心网络提出的要求。为了满足业务对网络的要求，越来越多的企业选择构建自己的数据中心或者租用公用云来承载日益增长的业务流量。

作为网络流量汇聚转发的承担者，数据中心交换机在整个数据中心网络中扮演着举足轻重的角色。本书将重点为读者介绍如何设计并使用华为 CloudEngine 系列交换机（简称 CE 系列交换机）来构建高性能的数据中心网络。

作者介绍

本书作者均为华为公司 DCN 研发团队的技术文档开发工程师（TW， Technical Writer），在技术文档开发领域的工作年限均超过 7 年。他们在长期的研发和面向客户交付的过程中积累了丰富的经验。现在他们将这些经验撰写成书，希望与广大读者分享他们的经验，对大家有所帮助。

他们是来自 DCN 技术文档开发项目组的：

马腾腾、朱华莲、李双双、宗悦、周婷婷、高洋洋

特邀专家

本书在成书的过程中，还得到了华为公司 DCN 研发领域其他专家的大力支持。他们从各自专业的角度，在内容的广度、方案的成熟度、信息组织的专业性等方面对本书提供了大量宝贵的建议。编者对他们表示诚挚的谢意：

产品架构设计专家：王建兵、宗志刚、刘树名、谢莹

软件测试专家：陈冶、杨德华、涂霖、张永乐

高级资料专家：崔春、李学昭、韩翔

读者对象

本书适用于负责企业网络设计与部署的网络设计师/架构师、网络管理员、维护工程师以及任何想了解数据中心网络设计基本原理的读者。

本书以华为 CE 系列数据中心交换机为背景，介绍了数据中心网络的基本设计原则以及一些必要的背景知识，同时也适用于大多数的同类数据中心网络设计。

为了更好的理解本书所讲的内容，我们希望在阅读之前您对于 IP 网络协议及网络设计已经有了基本的了解。

内容介绍

通过阅读本书，您将了解以下内容：

- 第一章：介绍华为 CE 系列交换机的特点、优势以及在网络中的位置。您将看到，华为提供类型丰富、满足不同网络需求的多款数据中心交换机，其中肯定有您需要的那一款。
- 第二章：介绍交换机在机架或机柜的部署方式，或者也可以理解为服务器的接入方式。选择 TOR 还是 EOR/MOR，在这里您会找到答案。
- 第三章：介绍数据中心网络中的布线设计。在合适的场景下使用最优的线缆，是构建一张高速数据中心网络的必要条件。
- 第四章：介绍如何合理的设置收敛比。在网络设计过程中，收敛比是必然要考虑的要素。收敛比的设定取决于服务器的类型、带宽需求以及未来的网络扩展等多个因素。如果您想让网络达到线速，或者最大程度的满足东西向和南北向流量的最优化转发，那么您应该阅读下这一章的内容。
- 第五章：介绍华为提供的多种 Fabric 网络技术。从本章开始，专题的重心从物理架构设计转移到逻辑架构设计。这些 Fabric 技术可以帮助您实现设备虚拟化，简化网络管理，避免二层环路。
- 第六章：介绍 IP Fabric 和三层路由的设计原则。为什么 Spine-Leaf 架构能够成为目前大多数数据中心网络的选择？数据中心网络三层路由协议的选择依据是什么？这就是第六章要回答的问题。
- 第七章：介绍 Overlay 网络和技术。数据中心中有海量租户（VM），他们之间通常需要进行跨物理网络的二层或三层互通。Overlay 网络是构建在物理 Underlay 网络之上的逻辑网络，通过构建点到点或点到多点的隧道，实现租户之间的通信需求。本章重点讲述目前应用最广泛的 Overlay 技术——VXLAN。
- 第八章：介绍 VXLAN 控制面协议 BGP EVPN 的实现原理。有了 BGP EVPN，VXLAN 可以进行隧道自动建立以及表项学习，降低了网络中的泛洪流量。

本书聚焦数据中心网络设计中的重点概念，尽力通过简单易懂的描述将设计原理和实现原理描述清楚。阅读后，您将清楚如何使用 CE 系列交换机构建自己的数据中心。如果您想了解 CE 系列交换机详细的功能支持情况及配置和维护方法，请点击以下链接访问华为“企业用户技术支持网站”获取完整的产品文档：

<http://support.huawei.com/enterprise/zh/index.html>

版本说明

- 本书内容基于 CE 系列交换机 V200R002C50 版本的实现进行描述，如需获得 CE 系列交换机最新的、更详细的介绍，请访问以下网站获取产品最新的 Datasheet：

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches>

- 除了硬件形态的交换机，华为还提供软件形态的数据中心交换机 CE1800V。本书仅基于硬件形态交换机进行描述。

目 录

前言 ii

第一章 华为数据中心交换机简介 1

1.1 数据中心中的交换机 1

1.2 CE12800 系列 1

1.3 CE8800 系列 2

1.4 CE7800 系列 2

1.5 CE6800 系列 3

1.6 CE5800 系列 3

1.7 本章小结 4

第二章 交换机的部署位置 5

2.1 简介 5

2.2 TOR 5

2.3 EOR/MOR 9

2.4 本章小结 11

第三章 线缆连接 13

3.1 简介 13

3.2 网线 13

3.3 光纤 15

3.4 直连铜缆 19

3.5 本章小结 21

第四章 收敛比 23

4.1 什么是流量收敛 23

4.2 网络流量收敛设计 25

4.2.1 服务器接入 Leaf 29

4.2.2 Leaf 接入 Spine 32

4.2.3 Spine 接入 Border Leaf 35

4.3 本章小结 36

第五章 Fabric 网络 37

5.1 简介 37

5.2 堆叠组网方案.....	37
5.3 M-LAG 组网方案	40
5.4 本章小结	43
第六章 IP Fabric & 三层路由	44
6.1 IP Fabric	44
6.2 Spine+Leaf 网络架构起源	46
6.3 BGP	48
6.3.1 BGP 协议基础.....	48
6.3.2 BGP 网络设计	49
6.4 本章小结	54
第七章 Overlay 网络	55
7.1 Overlay 介绍	55
7.2 VXLAN	56
7.3 二层 MAC 学习及 BUM 报文转发	58
7.4 VXLAN 网关部署	60
7.5 双活网关	62
7.6 本章小结	64
第八章 BGP EVPN	65
8.1 EVPN 介绍	65
8.2 BGP EVPN 路由类型	65
8.3 Type2 类型路由	66
8.4 Type3 类型路由	67
8.5 Type5 类型路由	68
8.6 BGP EVPN 实现 DCI 互联.....	69
8.7 本章小结	70
总结	71
相关资料.....	72
术语与缩略语	73

第一章 华为数据中心交换机简介

1.1 数据中心中的交换机

在数据中心设计过程中，第一个要面临的问题通常就是“应该选择什么样的交换机？”

华为，针对不同客户的网络需求，生产既符合通用标准又满足特定网络需求的数据中心交换机。作为本书的第一章，我们首先来简要介绍一下华为 CE 系列交换机，以帮助您了解不同款型交换机在网络中的不同应用。

CE 系列交换机是华为公司面向数据中心推出的新一代高性能交换机，满足了用户对低时延、高性能、高端口密度的要求，灵活支持不同的网络结构，是数据中心交换机的理想选择。截止目前，CE 系列交换机包含 CE12800 系列、CE8800 系列、CE7800 系列、CE6800 系列、CE5800 系列五个系列的硬件形态的交换机。你可以访问以下网站查看 CE 系列交换机全家福照片：1.1 数据中心中的交换机

下面我们针对每个系列交换机进行简要介绍。

1.2 CE12800 系列

CE12800 系列交换机是华为公司面向数据中心和高端园区推出的新一代高性能框式核心交换机，目前包含 CE12804、CE12808、CE12812、CE12816、CE12804S、CE12808S、CE12804E、CE12808E、CE12816E 款型。CE12800 在提供稳定、可靠、安全的高性能 L2/L3 层交换服务基础上，助力用户构建弹性、虚拟和高品质的网络。其中，CE12804E/CE12808E/CE12816E 支持华为自研的以太网网络处理器，满足用户对业务灵活快速定制的需求，同时可实现对设备的精细化运维。

CE12800 采用先进的硬件架构设计，提供多种接口密度的 100GE/40GE/10GE/GE 线卡（部分线卡的接口可通过模式切换或接口拆分工作到 25G 速率），单机最大支持 576 个 100GE、576 个 40GE、2304 个 25GE、2304 个 10GE 或者 768 个 GE 线速接口，可支持大容量的高密服务器接入和 TOR 上行汇聚，确保数据中心网络对高性能、超大容量的要求。

CE12800 五大硬件（主控板、交换网板、监控板、电源、风扇）全部采用热备设计，整机可靠性高。主控板 1+1 热备份；交换网板 N+M 热备份；监控板 1+1 热备份；电源采用双路输入，N+N 备份，并自带散热系统；风扇框 1+1 备份，单风扇框内双风扇对旋设计，散热高效强劲。

在数据中心网络中，CE12800 系列交换机既可以作为网络的核心层/汇聚层交换机，也可以作为 Spine-Leaf 架构中的 Spine 交换机；能够支持单机、堆叠、M-LAG、VS 等多种组合形态，并支持 BGP EVPN VXLAN 等主流 Overlay 技术。

如需了解 CE12800 系列交换机更多内容介绍，请访问：

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches/ce12800>

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches/ce12800e>

1.3 CE8800 系列

CE8800 系列交换机是华为公司面向数据中心和高端园区推出的新一代高性能、高密度、低时延以太网交换机。CE8800 提供高密度的 100GE/40GE/25GE/10GE 端口，支持丰富的数据中心特性和高性能堆叠，帮助企业和运营商构建面向云计算时代的数据中心网络平台。

截止目前，CE8800 系列交换机包含 CE8860 和 CE8850 两个系列：

- CE8860：CE8860-4C-EI，2U 高插卡式交换机，通过插卡灵活组合及接口拆分，单机最高可支持 32 个 100GE、64 个 40GE、128 个 25GE 或 128 个 10GE 接口。
- CE8850：CE8850-32CQ-EI，1U 高 100GE 盒式交换机。单机最高支持 32 个 100GE、32 个 40GE、128 个 25GE 或 130 个 10GE 接口。

在数据中心网络中，CE8800 系列交换机可以作为高密接入 TOR/EOR 交换机，当服务器规模不大时，也可以作为网络的核心层/汇聚层交换机；可以作为 Spine-Leaf 架构中的 Spine 或 Leaf 交换机；能够支持单机、堆叠、M-LAG 等多种组合形态，并支持 BGP EVPN VXLAN 等主流 Overlay 技术。

如需了解 CE8800 系列交换机更多内容介绍，请访问：

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches/ce8800>

1.4 CE7800 系列

CE7800 系列交换机是华为公司面向数据中心和高端园区推出的新一代高性能、高密度、低时延 40GE 以太网盒式交换机。CE7800 交换机提供高密度的 40GE QSFP+ 端口，支持丰富的数据中心特性和高性能堆叠，帮助企业和运营商构建面向云计算时代的数据中心网络平台。

截止目前，CE7800 系列交换机包含 CE7850 和 CE7855 两个系列，均支持 32 个 40GE QSFP+ 以太网光接口，且接口可进一步拆分为 4 个 10GE 接口。

在数据中心网络中，CE7800 系列交换机可以作为网络的核心层/汇聚层交换机；也可以作为 Spine-Leaf 架构中的 Spine 或 Leaf 交换机；能够支持单机、堆叠、M-LAG、SVF 等多种组合形态，并支持 BGP EVPN VXLAN 等主流 Overlay 技术。

如需了解 CE7800 系列交换机更多内容介绍，请访问：

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches/ce7800>

1.5 CE6800 系列

CE6800 系列交换机是华为公司面向数据中心和高端园区推出的新一代高性能、高密度、低时延万兆以太网盒式交换机。CE6800 采用先进的硬件结构设计，提供高密度的 10GE/25GE 端口接入，支持 40GE/100GE 上行端口，支持丰富的数据中心特性和高性能堆叠。

截止目前，CE6800 系列交换机包含 CE6880、CE6870、CE6860、CE6850、CE6810 系列：

- **CE6880：**支持华为自主研发的以太网处理器，满足用户对业务灵活快速定制的需求，同时可实现对设备的精细化运维。单机下行最大提供 48 个 10GE SFP+以太网光接口或 48 个 10GBASE-T 以太网电接口；上行提供 2 个 40GE/100GE QSFP28 以太网光接口和 4 个 40GE QSFP+以太网光接口，其中 QSFP28 以太网光接口可进一步拆分成 4 个 10GE 接口。
- **CE6870：**提供 4GB 的大缓存，可轻松应对数据中心中视频、搜索等应用引起的流量浪涌。单机下行最大提供 48 个 10GE SFP+以太网光接口或 48 个 10GBASE-T 以太网电接口；上行最大提供 6 个 40GE/100GE QSFP28 以太网光接口，每个接口可进一步拆分成 4 个 10GE 或 25GE 接口。
- **CE6860：**单机下行最大提供 48 个 10GE/25GE SFP28 以太网光接口；上行最大提供 8 个 40GE/100GE QSFP28 以太网光接口，每个接口可进一步拆分成 4 个 10GE 或 25GE 接口。
- **CE6850：**包含 CE6856HI、CE6855HI、CE6850U-HI、CE6851HI、CE6850HI、CE6850EI 等细分款型。单机下行最大提供 48 个 10GE SFP+以太网光接口或 48 个 10GBASE-T 以太网电接口；上行最大提供 6 个 40GE QSFP+以太网光接口，每个接口可进一步拆分成 4 个 10GE 接口。
- **CE6810：**包含 CE6810EI 和 CE6810LI 两个系列。单机下行最大提供 48 个 10GE SFP+以太网光接口；上行最大提供 4 个 40GE QSFP+以太网光接口，每个接口可进一步拆分成 4 个 10GE 接口。其中，CE6810LI 通常作为二层交换机使用。

在数据中心网络中，CE6800 系列交换机主要定位于高密万兆接入交换机，也可以作为 Spine-Leaf 架构中的 Leaf 交换机；能够支持单机、堆叠、M-LAG、SVF 等多种组合形态。除 CE6810 系列和 CE6850EI 外，其余款型均支持 BGP EVPN VXLAN 等主流 Overlay 技术。

如需了解 CE6800 系列交换机更多内容介绍，请访问：

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches/ce6800>

1.6 CE5800 系列

CE5800 系列交换机是华为公司推出的支持 40GE 上行接口的新一代盒式千兆接入交换机。

截止目前，CE5800 系列交换机包含 CE5855、CE5850、CE5810 三个系列。CE5800 系列交换机单机下行提供 48 或 24 个 10/100/1000BASE-T 以太网电接口；上行最大提供 4 个 10GE SFP+以太网光接口和 2 个 40GE QSFP+以太网光接口，其中 QSFP+接口可进一步拆分成 4 个 10GE 接口。

在数据中心网络中，CE5800 系列交换机主要定位于高密千兆接入交换机，也可以作为 Spine-Leaf 架构中的 Leaf 交换机；能够支持单机、堆叠、M-LAG、SVF（仅支持作为 SVF 的叶子交换机）等多种组合形态。

如需了解 CE5800 系列交换机更多内容介绍，请访问：

<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches/ce5800>。

1.7 本章小结

以上对 CE 系列交换机做了简要介绍。可以看到，华为既提供采用商业芯片的数据中心交换机，也提供像 CE12800E 系列、CE6880EI 系列这种采用自研芯片的交换机。

一般而言：

- 使用商用芯片的交换机，与其他厂商使用同款芯片的产品，在特性支持及性能规格上基本一致。商用芯片通常可以提供比较高的吞吐量、高接口密度，可以提供大多数网络需要的软件能力；也正因为如此，此类交换机的功能特性往往会受限于芯片本身的能力，所以灵活性上相比自研芯片要差一些。
- 使用自研芯片的交换机，可以满足用户更多的定制需求，应用上更加灵活。这种类型的交换机通常在开放性、自动化、自主运维等方面能力更强。缺点是业务叠加越来越多后，转发性能可能会有所下降。

商用芯片及自研芯片各有其优劣势。当您进行网络规划与设计时，请根据具体业务需求及未来网络规划来选择合适的数据中心交换机。

第二章 交换机的部署位置

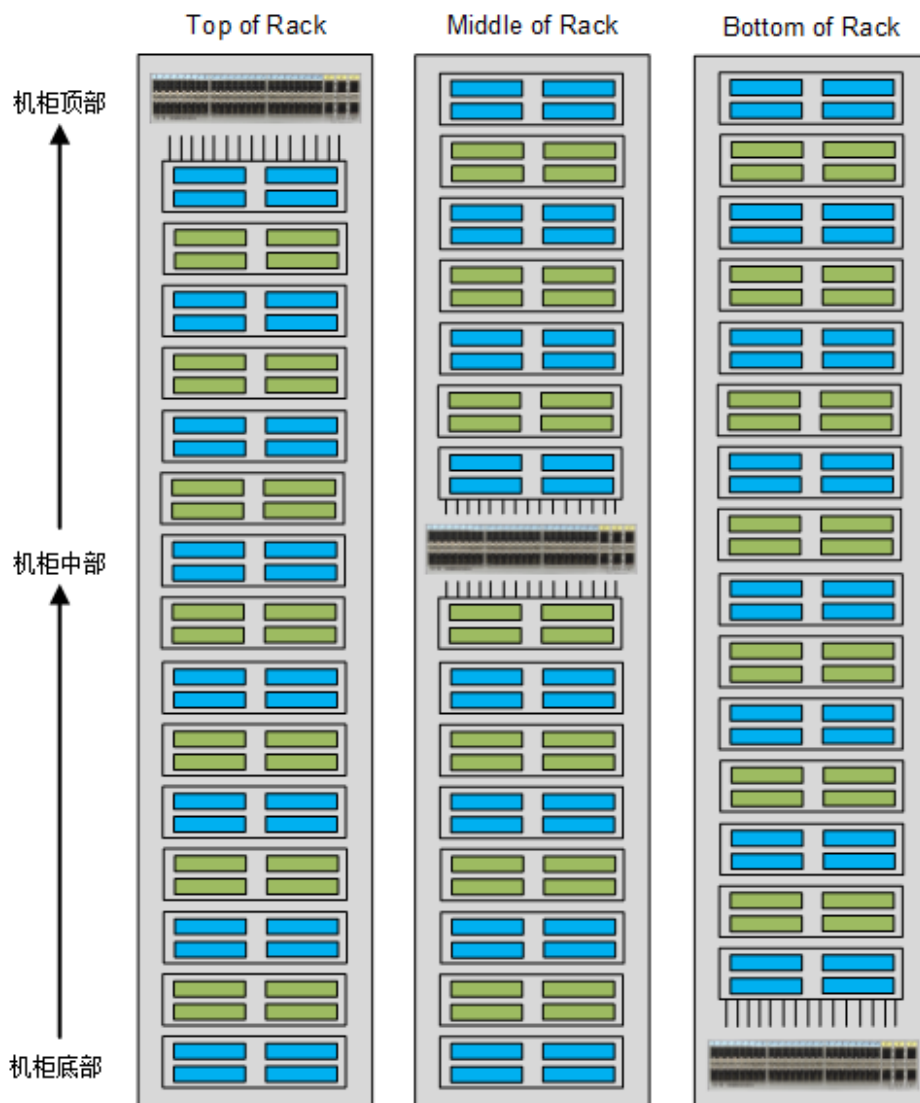
1.1 简介

经过第一章的介绍，您应该已经对华为 CloudEngine 系列数据中心交换机有了一个很好的了解。接下来我们来看一下，在数据中心机房中这些交换机应该部署在什么位置，在本章不妨称之为物理架构。目前，按照交换机在机柜上的部署位置（也可以理解为按照服务器的不同接入方式），其物理架构一般分为 TOR（Top of Rack）和 EOR（End of Row）/MOR（Middle of Row）两种。下面我们分别来了解下这两种架构。

1.2 TOR

TOR（Top of Rack）指的是在每个服务器机柜上部署 1~2 台交换机，服务器直接接入到本机柜的交换机上，实现服务器与交换机在机柜内的互联。虽然从字面上看，Top of Rack 指的是“机柜顶部”，但实际 TOR 的核心在于将交换机部署在服务器机柜内，既可以部署在机柜顶部，也可以部署在机柜的中部（Middle of Rack）或底部（Bottom of Rack），如图 1-1 所示。通常而言，将交换机部署在机柜顶部是最有利于走线的，因此这种架构应用最多。

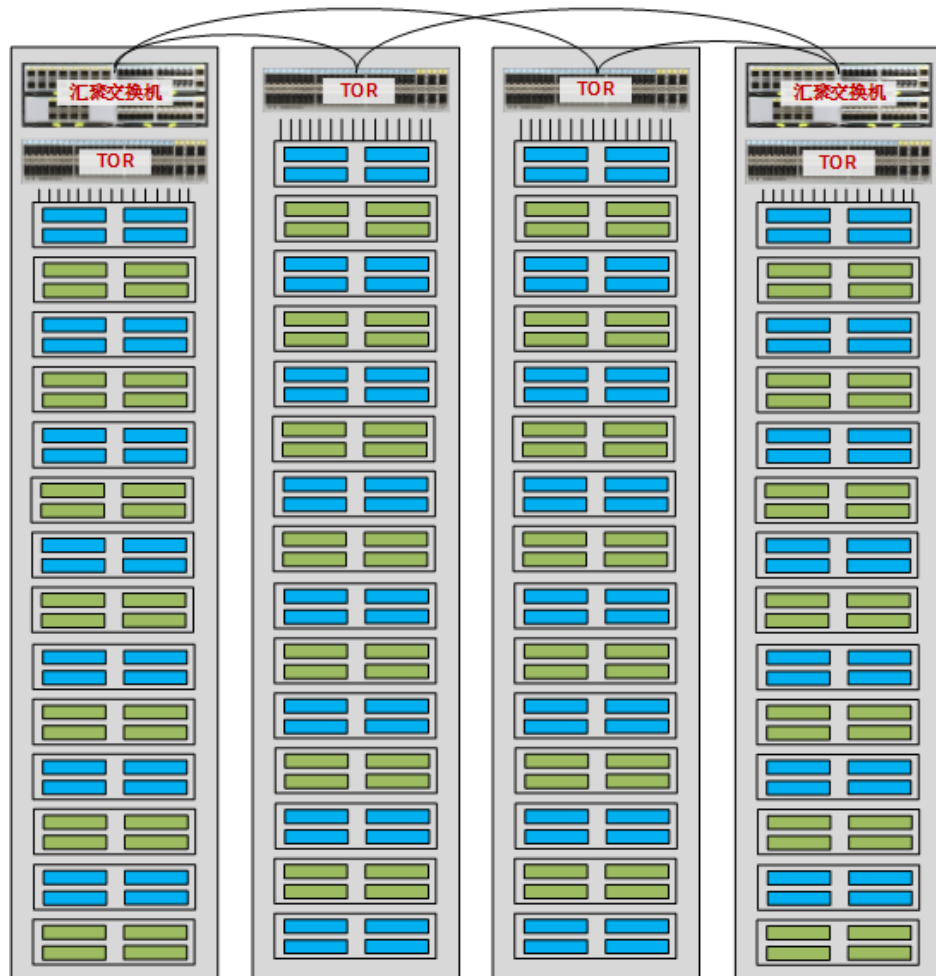
图1-1 TOR 交换机的部署位置



以 TOR 方式部署在服务器机柜上的交换机，我们称之为 TOR 交换机。TOR 交换机一般由高度在 1U~2U 的盒式交换机承担，比如华为的 CE5800 和 CE6800 系列交换机。

TOR 架构最大的优点就是简化了服务器到交换机之间的连线。机柜内服务器上的 GE/10GE/25GE 端口可以直接通过短跳线连接到 TOR 交换机上，再经由 10GE/40GE/100GE 光纤双上行连接至汇聚交换机，如图 1-2 所示。这样连线极大地缩短了线缆的使用距离，简化了线缆管理，降低了网络结构复杂性，更符合数据中心绿色和节能的趋势。当后续业务扩展需要更换线缆时，也更加便利。

图1-2 TOR 交换机连接至汇聚交换机



对于 TOR 架构，每个机柜可以被看做一个独立的管理实体。当服务器或交换机需要升级时，可以以机柜为单位逐一升级，升级过程中不会影响其他机柜的流量转发，将对业务的影响控制在最小范围内。

TOR 交换机上行链路通常会选择光纤，因为从长期的投资保护考虑，光纤比铜缆更有优势。光纤可以承载更高的带宽，当需要更换更高速率的链路时，光纤的选择也更加灵活。

因此在选择 TOR 交换机时，通常您需要同时考虑交换机下行连接服务器的端口数量和速率以及上行端口的灵活性。一般情况下：

- 当服务器端口为 GE 口时，您可以选择 CE5855EI 系列交换机。CE5855EI 系列交换机分为 CE5855-48T4S2Q-EI 和 CE5855-24T4S2Q-EI 两款，分别提供下行 48 和 24 个 10/100/1000BASE-T 以太网电接口；上行提供 2 个 40GE QSFP+以太网光接口和 4 个 10GE SFP+以太网光接口，同时每个 40GE 接口还支持拆分为 4 个 10GE 接口。
- 当服务器端口为 10GE 口时，您可以选择 CE6856HI 系列交换机。CE6856HI 系列交换机分为 CE6856-48S6Q-HI 和 CE6856-48T6Q-HI 两款，分别提供下行 48 个

10GE SFP+以太网光接口和 10GBASE-T 以太网电接口；上行均提供 6 个 40GE QSFP+以太网光接口，每个接口支持拆分为 4 个 10GE 接口。

- 如果您希望 TOR 交换机提供大缓存，则建议您选择 CE6870EI 系列交换机。按照下行口类型及数量的不同，CE6870 系列交换机分为 CE6870-48S6CQ-EI、CE6870-24S6CQ-EI 和 CE6870-48T6CQ-EI 三款。以图 1-3 所示的 CE6870-48S6CQ-EI 为例，该交换机上行支持 6 个 40GE/100GE QSFP28 以太网光接口，同时还支持拆分成 4 个 10GE 或 4 个 25GE 接口。CE6870EI 丰富的上行接口类型，为不同的业务需求提供了灵活的选择，从长期来看很好的保护了客户投资。CE6870 系列交换机提供 4GB 的大缓存，可轻松应对数据中心中视频、搜索等应用引起的流量浪涌。

图1-3 CE6870-48S6CQ-EI



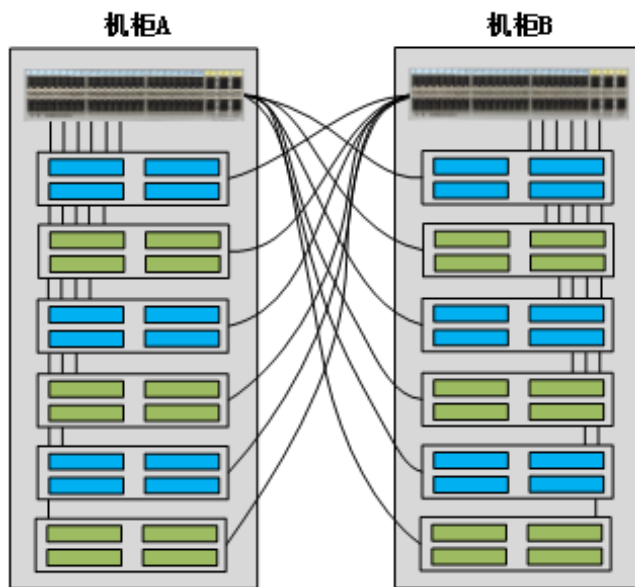
当然，TOR 架构也有其缺点。最明显的一个就是，TOR 架构扩大了整个数据中心机房的管理域。每个机柜上都部署交换机，也就意味着机房中交换机的数量会比较多，每一台交换机都需要进行配置、管理和维护。假设您的机房中有 10 排机柜，每排放置 10 台机柜，每台机柜上部署两台 TOR 交换机，那么您就需要管理维护 200 台 TOR 交换机。尽管这 200 台交换机的配置基本是相同的，但是依然需要花费大量的人力成本，并且也增加了设备误配置的概率。

针对上述问题，CE 系列交换机也提供了相应的解决方案。比如针对新出厂或空配置设备，您可以使用 ZTP（Zero Touch Provisioning）功能进行批量自动配置。CE 系列交换机默认开启 ZTP 功能。运行 ZTP 后，交换机可以从 U 盘或文件服务器获取版本文件（包括系统软件、配置文件、License 文件、补丁文件、自定义文件）并自动加载，实现设备的免现场配置。CE 系列交换机还支持丰富的设备虚拟化技术（比如堆叠），可有效简化设备的管理平面，进而降低人力成本，提升部署效率。在第五章将会对这些技术做详细介绍。

TOR 架构的另一个缺点是端口浪费。目前，大多数的 TOR 交换机都可以提供 48 个 GE/10GE/25GE 下行端口。以每机柜部署两台 TOR 交换机为例，则共有 96 个下行端口，那么您需要在机柜中放置大量的服务器才有可能全部利用好这些接口。

通过在相邻机柜之间交叉连线可以在一定程度上降低端口的浪费。如图 1-4 所示，两台机柜上分别部署一台 48 端口 TOR 交换机，每台交换机的 24 个端口提供本机柜的服务器接入，另外 24 个端口提供给相邻机柜的服务器接入。正如你所看到的，这种方案的代价就是需要增加在两机柜间的连线，因此也并非一个完美的方案。但是相比每个机柜部署两台 TOR 交换机造成的端口浪费，这种方案也不失为一种低成本并且行之有效的选择。

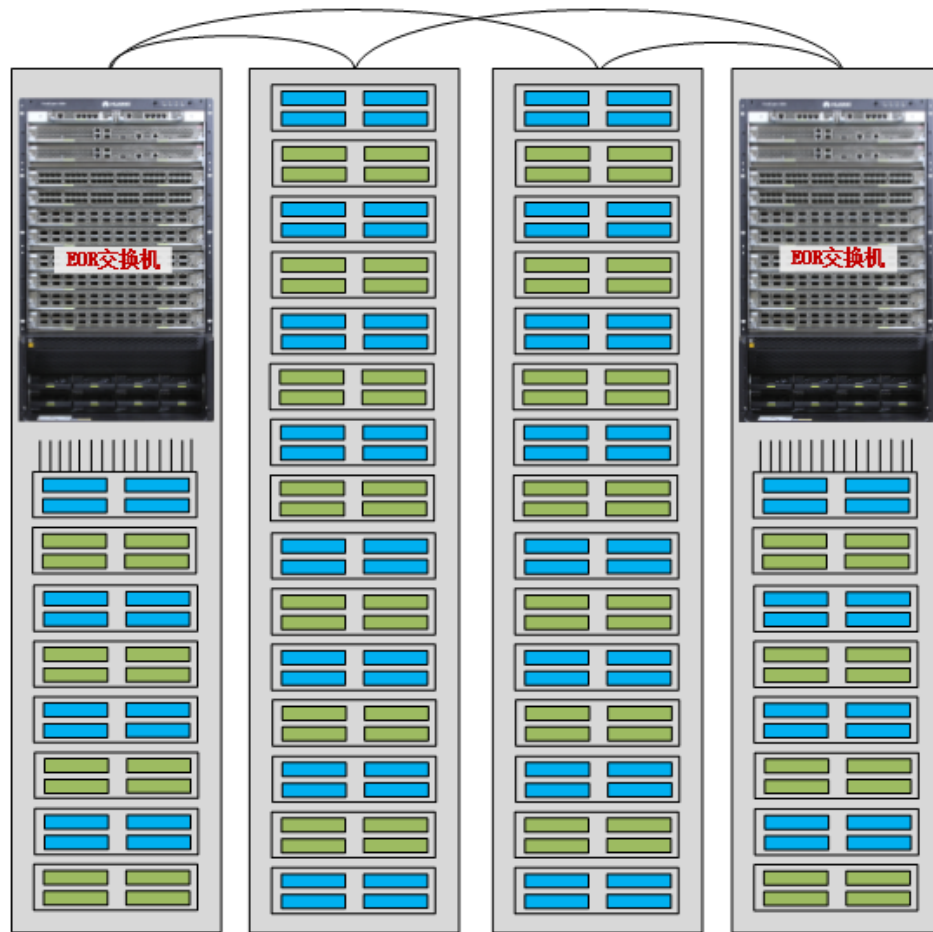
图1-4 相邻机柜间交叉连线



1.3 EOR/MOR

与 TOR 不同，EOR（End of Row）架构在每排机柜的末端提供统一的网络接入点。如图 1-5 所示，在每排机柜末端部署，供服务器统一接入网络的交换机，我们称之为 EOR 交换机。

图1-5 EOR 交换机的部署位置



出于可靠性考虑，每排机柜通常会配备两个网络机柜，分别位于这一排的头端和末端。服务器网卡使用相对较短的 RJ45 / DAC / 光纤跳线连接到同一机柜的配线架，配线架上的网线、光纤、铜缆等经捆绑后穿过架空线缆槽或地板连接到每排最边端的网络机柜。

EOR 架构将接入交换机集中放置在 1~2 个机柜中，方便了管理和维护，但同时也增加了服务器机柜到网络机柜之间的连线。服务器机柜离网络机柜越远，在机房中的布线距离越长，由此导致线缆维护工作量大，灵活性差。

MOR（Middle of Row）架构是对 EOR 的一种改进，也为服务器提供统一的网络接入机柜，但是 MOR 要求将网络机柜放在整排机柜的中部，在一定程度上缩短了服务器机柜到网络机柜的距离，简化了线缆的管理与维护。但是与 TOR 相比，无论是 EOR 还是 MOR，布线复杂、管理维护难度大依然是它们的最大缺点。若无特殊说明，下文中出现的 EOR 的相关描述同时也适用于 MOR。

EOR 交换机一般由框式交换机承担，比如华为的 CE12800 系列交换机。如果您的机房内服务器数量不多的话，也可以选择 CE8800 和 CE7800 系列。

相比盒式交换机，框式交换机在以下方面具有明显的优势：

- 提供更大数量、丰富类型的接入端口。通过在框式交换机上配置不同数量不同速率的接口板，您可以灵活控制接入端口的数量和速率。以 CE12800 系列交换机为

例，截止目前支持 36*100GE、36*40GE、48*10GE、48*GE 等接口板类型（以最高端口数量举例），接口还支持丰富的拆分类型，为不同数据中心服务器的接入提供了灵活的选择。

- 可靠性高。框式交换机提供冗余硬件，例如多交换网板、多电源模块、多风扇模块，整机可靠性更高。
- 保护客户投资。当数据中心需要比现在更高的接入速率，您只需要更换更高速率的接口板，而不用整台设备换掉。从整个生命周期来看，成本更低。

EOR 架构使得数据中心的管理领域大大减少，因为它是基于每排而不是每机架进行管理。但是，这同时也意味着一旦 EOR 交换机出现故障或者升级失败，它的影响范围也会是一整排的服务器。这也是为什么对 EOR 交换机要求更高的原因。

1.4 本章小结

如表 1-1 所示，我们对 TOR 和 EOR/MOR 两种架构做下总结：

表1-1 TOR 和 EOR/MOR 对比

物理架构	TOR	EOE/MOE
优点	<ul style="list-style-type: none">• 布线简单、线缆维护方便、扩展性好 <p>所有的服务器都连接到同一机柜的 TOR 交换机上，只有交换机上行链路会连接到机柜外部的汇聚交换机上，降低了布线的复杂性；当服务器需要升级时（比如从 10GE 切换至 25GE），线缆连接无需做大规模改动，具有良好的扩展性。</p> <ul style="list-style-type: none">• 基于机柜模块化管理，故障影响范围小 <p>TOR 架构可以基于机柜实现模块化管理，当设备故障或升级时能够将对业务的影响控制在最小的范围内。</p>	<ul style="list-style-type: none">• 管理简单，可靠性高 <p>需要管理维护的接入交换机数量少。且大部分 EOR 交换机为框式交换机，关键部件均为冗余设计，整机可靠性高。当需要业务扩展或升级时，可以仅增加或更换接口板，而不需要整机替换。</p> <ul style="list-style-type: none">• 端口利用率高 <p>服务器统一接入 EOR 交换机，交换机的端口能得到充分利用。</p>
不足	<ul style="list-style-type: none">• 端口浪费 <p>每个服务器机柜受电源输出功率限制，可部署的服务器数量有限，由此导致机柜内交换机的接入端口利用率不足。</p> <ul style="list-style-type: none">• TOR 交换机管理维护复杂 <p>TOR 方式的接入交换机数量多，网络设备管理维护工作量</p>	<ul style="list-style-type: none">• 布线复杂，维护难度大 <p>从服务器机柜到网络机柜的线缆多，且距网络机柜越远的服务器机柜的线缆在机房中的布线距离越长，从而导致线缆管理维护工作量大、灵活性差。</p> <ul style="list-style-type: none">• 故障影响范围大 <p>当 EOR 交换机出现故障时，</p>

物理架构	TOR	EOR/MOR
	大。	对交换机所在一排的服务器均会产生影响。

综上所述，TOR、EOR/MOR 均有各自的优势及不足，不同的架构适用于不同的场景。用户在进行选择时，一定要根据数据中心承载的业务类型、不同的业务特点以及不同数据中心的特定条件，包括投资和管理成本等进行综合考量后再决定采用哪种方式。当然，在业务数据量不特别大，对扩展性要求也不是特别高的传统用户的数据中心中，EOR/MOR 的方式仍然会受到很大程度上的青睐；而在采用分布式架构业务为代表的、对扩展性要求很高的用户数据中心，采用 TOR 将会是一种趋势。

第三章 线缆连接

1.1 简介

在上一章节中，我们详细介绍了数据中心机房中两种部署设备的方式：TOR 和 EOR 架构，从中我们可以看出数据中心物理布线设计的重要性。

在合适的场景下使用最优的线缆，是构建一张高速数据中心网络的必要条件。这其中不仅需要考虑信号的衰减、传输距离，设计线缆的走线方式、位置，还要考虑到成本、线缆的安装和拆除，以及未来网络的升级等。

大多数线缆供应商一般会提供三种类型的线缆：网线（本文中特指双绞线）、光纤和直连铜缆（Direct Attach Cable，DAC）。

接下来，我们将一起探讨进行数据中心网络设计时，如何选择最合适的线缆。

1.2 网线

根据频率和信噪比的不同，常见的网线包括五类线（CAT5）、超五类线（CAT5e）、六类线（CAT6），它们都是两端为 RJ45 连接器的双绞线，最大传输距离为 100 米。此外，网线还包括一类线（CAT1）、二类线（CAT2）、三类线（CAT3）、四类线（CAT4）、超六类线（CAT6A）、七类线（CAT7）等。一般来说，类型数字越大、版本越新，技术越先进、带宽也越宽，当然价格也越贵。

根据有无屏蔽层，网线又可分为屏蔽双绞线（Shielded Twisted Pair，STP）和非屏蔽双绞线（Unshielded Twisted Pair，UTP）。屏蔽双绞线可减少辐射，防止信息被窃听，也可阻止外部电磁干扰的进入，与同类的非屏蔽双绞线相比具有更高的传输速率，但是价格也相对更高，且安装时也更困难。非屏蔽双绞线的优点在于：成本低、重量轻、易弯曲等，且其性能对于一般网络来说影响不大，所以应用相对更为广泛。不过七类双绞线除外，因为要实现全双工 10Gbps 的速率传输，所以只能采用屏蔽双绞线，而没有非屏蔽的七类双绞线。

各类型网线的详细介绍如下：

- 一类线（CAT1）：主要用于传输语音（一类标准主要用于八十年代初之前的电话线缆），不用于数据传输。
- 二类线（CAT2）：传输频率为 1MHz，用于语音传输和最高传输速率 4Mbps 的数据传输，常见于使用 4Mbps 令牌传递协议的旧的令牌网。
- 三类线（CAT3）：传输频率 16MHz，用于语音传输及最高传输速率为 10Mbps 的数据传输，主要用于 10Base-T 网络，被 ANSI/TIA-568.C.2 作为最低使用等级。

- 四类线（CAT4）：传输频率为 20MHz，用于语音传输和最高传输速率 16Mbps 的数据传输，主要用于基于令牌的局域网和 10Base-T/100Base-T 网络。
- 五类线（CAT5）：传输频率为 100MHz，用于语音传输和最高传输速率为 100Mbps 的数据传输，主要用于 100Base-T 和 10Base-T 网络。这是最常用的以太网电缆，该类电缆增加了绕线密度，外套一种高质量的绝缘材料。
- 超五类线（CAT5e）：传输频率最大为 100MHz，主要用于千兆位以太网（1000Mbps）。具有衰减小，串扰少，并且具有更高的衰减与串扰的比值（ACR）和信噪比（SNR）、更小的时延误差，性能得到很大提高。
- 六类线（CAT6）：传输频率为 250MHz，适用于传输速率高于 1Gbps 的网络。六类双绞线在外形上和结构上与五类或超五类双绞线都有一定的差别，不仅增加了绝缘的十字骨架，将双绞线的四对线分别置于十字骨架的四个凹槽内，而且电缆的直径也更粗。
- 超六类线（CAT6A）：传输频率是 500MHz，是六类线的两倍，最大传输速度可达到 10Gbps，主要应用于万兆位网络中。超六类线是六类线的改进版，同样是 ANSI/EIA/TIA-568B.2 和 ISO 6 类/E 级标准中规定的一种非屏蔽双绞线电缆，在串扰、衰减和信噪比等方面有较大改善。
- 七类线（CAT7）：传输频率至少可达 600 MHz，传输速率可达 10 Gbps，它主要为了适应万兆位以太网技术的应用和发展。该线是 ISO 7 类/F 级标准中最新的一种屏蔽双绞线。另外，七类线的连接器类型也与其他类型网线不同，是 GigaGate45（CG45）。

表 1-2 介绍了几种常见网线的的基本参数。

表1-2 常见网线的的基本参数

网线类型	使用场景	传输频率	最大传输速率	传输距离
五类线（CAT5）	100Base-T 和 10Base-T 网络	1~100MHz	100Mbps	100m
超五类线（CAT5e）	1000Base-T 网络	1~100MHz	1000Mbps	100m
六类线（CAT6）	1000Base-T 网络	1~250MHz	1000Mbps/10Gbps	100m/37~55m
超六类线（CAT6A）	10GBase-T 网络	1~500MHz	10Gbps	100m
七类线（CAT7）	10GBase-T 网络	1~600MHz	10Gbps	100m

在 TOR 物理架构下，基本上每个机柜内都会用到网线，在 EOR 物理架构下，网线通常用来连接服务器和交换机。考虑到未来网络的升级需要，在使用时要注意服务器和交换机之间的距离和布线规模。

基于上述考虑，在 TOR 场景下，如果您需要使用网线来进行网络部署，可以选用带有电接口的华为 CE 系列交换机，例如 CE5855-48T4S2Q-EI 交换机，或者 CE6810-32T16S4Q-LI、CE6850-48T4Q-EI 交换机等。在 EOR 场景下，我们推荐您使用性能较

高的产品，例如 CE6870-48T6CQ-EI、CE6880-48T4Q2CQ-EI 交换机，或者使用 CE12800 系列交换机、并配置 CE-L48GT 系列、CE-L48XT 系列单板。

1.3 光纤

按光传输模式的不同，光纤可以分为多模光纤（MMF，Multi Mode Fiber）和单模光纤（SMF，Single Mode Fiber）。

下面我们详细介绍一下，多模光纤和单模光纤的差异。

多模光纤

多模光纤的纤芯较粗，能传输多种模式的光，但其模式色散较大，并且随着传输距离的增大模式色散会逐渐加重，因此常和多模光模块配合用于短距离低成本的通信传输。

根据光纤直径和模式带宽的不同，多模光纤可分为 OM1、OM2、OM3、OM4 几个等级，不同等级光纤的光纤直径和模式带宽如表 1-3 所示。常用的多模光纤为 G651 标准的光纤，可传输 800~900nm、1200~1350nm 波长的光。

外观上，多模光纤的外表上印有“MM”的字样，OM1、OM2 等级光纤一般为橘红色，OM3、OM4 等级的光纤一般为淡绿色。

图1-6 OM1/OM2 多模光纤



图1-7 OM3/OM4 多模光纤



表1-3 多模光纤的等级分类

光纤等级	光纤直径（ μm ）	模式带宽（MHz*km）
OM1	62.5	200
OM2	50	500
OM3	50	2000
OM4	50	4700

多模光纤的传输距离与接口类型、中心波长、使用的光纤等级相关，常用的多模光纤规格如表 1-4 所示。

表1-4 常用的多模光纤规格

应用类型	中心波长（nm）	光纤等级	最大传输距离（m）
1000BASE-SX	850	OM1	275
		OM2	550
10GBASE-SR	850	OM1	33
		OM2	82
		OM3	300
		OM4	450
10GBASE-LRM	1310	OM1	220
		OM2	220
		OM3	220
		OM4	220

单模光纤

单模光纤的纤芯较细，只能传输一种模式的光，因此模式色散很小，适用于远距离通信。

常用的单模光纤为 G652 标准的光纤，可传输 1260~1360nm、1530~1565nm 波长的光。

外观上，单模光纤的外表上印有“SM”的字样，单模光纤大多为黄色，如图 1-8 所示。

图1-8 单模光纤

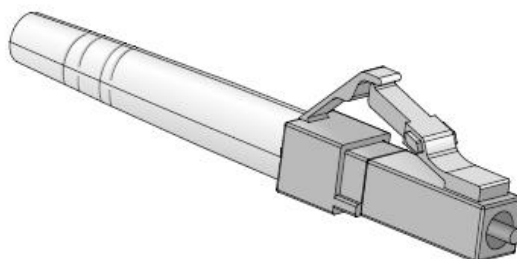


连接器类型

CE 系列交换机支持的华为光模块所采用的光纤连接器类型包括 LC 和 MPO 两种。

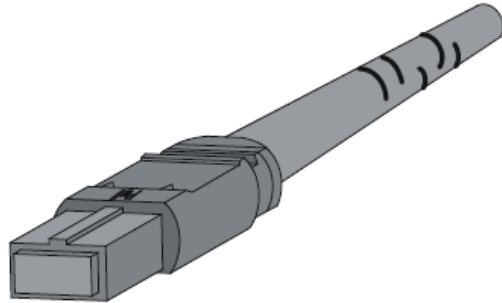
- LC (Lucent Connector or Local Connector)，外观如图 1-9 所示。

图1-9 LC 连接器



- MPO（Multi-fiber Push On）连接器，外观如图 1-10 所示。

图1-10 MPO 连接器



其中，华为光模块所采用的 MPO 连接器均为带有导向销的母头连接器。

在 TOR 场景下，多模光纤可用于机柜内部的设备连接。在传输距离不大于 400 米的场景下，多模光纤也可用于网络的汇聚层。在 EOR 场景下，多模光纤可以用来连接交换机和服务器。如果距离允许，多模光纤也可以用在网络的核心或汇聚层。

单模光纤一般用于远距离通信传输。

基于上述考虑，在 TOR 场景下，如果您选择使用光纤来进行网络部署，可以选用带有光接口的华为 CE 系列交换机，例如 CE6851-48S6Q-HI、CE6855-48S6Q-HI 交换机等。在 EOR 场景下，我们推荐您使用性能较高的产品，例如 CE6870-48S6CQ-EI、CE7855-32Q-EI 交换机，或者使用 CE12800 系列交换机、并配置带有光接口的单板。

另外，您也可以选择使用 AOC 光线缆（Active Optical Cables）。AOC 光线缆是光模块和光纤一体化的有源光线缆，您可以简单理解为，AOC 光线缆是将两只光模块和光纤封装在一起，但由于中间的传输介质是光纤，所以 AOC 光线缆两端的光模块中带有激光器件。

由于 AOC 光线缆是固定封装的，对用户而言，省去了制作过程中用到的一些光器件，因此成本较低。另外，AOC 光线缆对环境要求相对较低，没有光纤连接器的清洁问题，可靠性高。

但 AOC 光线缆的长度固定，配置灵活性上较差，一般适用于短距离传输的场景。

图1-11 AOC 光线缆



1.4 直连铜缆

直连铜缆（Direct Attach Cable，DAC），或称 Twinax 铜缆、高速线缆，是一种固定长度、两端有固定连接器的线缆组件，如图 1-12 所示。

图1-12 直连铜缆



DAC 铜缆包含有源（active）和无源（passive）两种，有源 DAC 铜缆内置了放大器和均衡器，可以提升信号质量，但相对成本较高。大多数情况下，当传输距离小于 5 米时，您可以选择使用无源 DAC 铜缆，而当传输距离大于 5 米时，选择有源 DAC 铜缆。

DAC 铜缆上的连接器与光模块相比，接口类型相同，但缺少了昂贵的光学激光器和其他电子元件，因此可以大大节省成本和功耗，广泛应用于数据中心网络中的短距离连接。

在 TOR 场景下，DAC 铜缆是进行机柜内短距离布线的最佳选择。在 EOR 场景下，如果传输距离小于 10 米，也可以选择使用 DAC 铜缆。

表 1-5 介绍了数据中心可能用到的不同类型 DAC 铜缆的基本信息，其中包括的 DAC 铜缆的弯曲半径。和其他很多线缆一样，DAC 铜缆对不良弯曲敏感，不良弯曲会影响线缆的传输速率。

表1-5 DAC 铜缆的基本信息

DAC 铜缆名称	最大传输距离 (m)	电气属性	连接器类型	最小出线空间 (mm)	最小弯曲半径 (mm)
SFP-10G-CU1M	1	无源	SFP+<->SFP+	60	35
SFP-10G-CU3M	3	无源	SFP+<->SFP+	60	35
SFP-10G-CU5M	5	无源	SFP+<->SFP+	60	35
SFP-10G-AC7M	7	有源	SFP+<->SFP+	60	35
SFP-10G-AC10M	10	有源	SFP+<->SFP+	60	35
QSFP-40G-CU1M	1	无源	QSFP+<->QSFP+	75	50
QSFP-40G-CU3M	3	无源	QSFP+<->QSFP+	75	50
QSFP-40G-CU5M	5	无源	QSFP+<->QSFP+	75	50
QSFP-4SFP10G-CU1M	1	无源	QSFP+<->4*SFP+	QSFP+端：100 SFP+端：60	QSFP+端：50 SFP+端：35
QSFP-4SFP10G-CU3M	3	无源	QSFP+<->4*SFP+	QSFP+端：100 SFP+端：60	QSFP+端：50 SFP+端：35
QSFP-4SFP10G-CU5M	5	无源	QSFP+<->4*SFP+	QSFP+端：100 SFP+端：60	QSFP+端：50 SFP+端：35
QSFP28-100G-CU1M	1	无源	QSFP28<->QSFP28	90	70
QSFP28-100G-CU3M	3	无源	QSFP28<->QSFP28	90	70
QSFP28-100G-CU5M	4	无源	QSFP28<->QSFP28	90	70

DAC 铜缆名称	最大传输距离（m）	电气属性	连接器类型	最小出线空间（mm）	最小弯曲半径（mm）
SFP-25G-CU1M	1	无源	SFP28<->SFP28	70	40
SFP-25G-CU3M	3	无源	SFP28<->SFP28	70	40
SFP-25G-CU3M-N	3	无源	SFP28<->SFP28	70	40
SFP-25G-CU5M	5	无源	SFP28<->SFP28	70	40
QSFP-4SFP25G-CU1M	1	无源	QSFP28<->4*SFP28	QSFP28 端：100 SFP28 端：70	QSFP28 端：50 SFP28 端：40
QSFP-4SFP25G-CU3M	3	无源	QSFP28<->4*SFP28	QSFP28 端：100 SFP28 端：70	QSFP28 端：50 SFP28 端：40
QSFP-4SFP25G-CU3M-N	3	无源	QSFP28<->4*SFP28	QSFP28 端：100 SFP28 端：70	QSFP28 端：50 SFP28 端：40
QSFP-4SFP25G-CU5M	5	无源	QSFP28<->4*SFP28	QSFP28 端：100 SFP28 端：70	QSFP28 端：50 SFP28 端：40



说明

SFP-25G-CU3M-N 铜缆与 SFP-25G-CU3M 铜缆的区别在于：

SFP-25G-CU3M-N 铜缆（26AWG）比 SFP-25G-CU3M（30AWG）更粗，所以传输损耗会更小，在使用时，端口可不用打开 FEC 功能。

QSFP-4SFP25G-CU3M-N 铜缆和 QSFP-4SFP25G-CU3M 也是同样的差异。

1.5 本章小结

经过上面的介绍，我们对网线、光纤、直连铜缆三种线缆的优劣势做如下总结，如表 1-6 所示。

表1-6 网线、光纤、直连铜缆的对比

线缆类型	优点	缺点	使用场景
------	----	----	------

线缆类型	优点	缺点	使用场景
网线	<ul style="list-style-type: none">• 价格低廉• 安装方式简单	<ul style="list-style-type: none">• 传输速率低• 传输距离短	低成本要求下，电口互连可以选用网线，且目前六类、超六类网线也能达到较高速率。
光纤	<ul style="list-style-type: none">• 传输容量大• 传输速率高• 传输距离远• 抗干扰性好	<ul style="list-style-type: none">• 部署成本高• 安装方式复杂• 对环境要求高	长距离、大容量、高可靠的数据传输要求下，可以选用光纤，并且全光网络是未来的趋势，可以支撑数据中心的未来演进。
直连铜缆	<ul style="list-style-type: none">• 部署成本低• 传输速率高• 抗干扰性好	<ul style="list-style-type: none">• 传输距离短• 长度固定，灵活性差	数据中心网络内部的短距离、高可靠的光接口互连，且部署成本比光纤更低。

如果您想要了解更详细的线缆信息，请您参考《[CloudEngine 12800 硬件描述](#)》或《[CloudEngine 8800&7800&6800&5800 硬件描述](#)》。

第四章 收敛比

1.1 什么是流量收敛

数据报文的流量收敛，是指数据报文在网络转发过程中由于架构、设备等非故障原因而不能实现线速无丢包转发。在流量收敛时，网络设备会有部分端口拥塞，进而丢弃部分报文。为了能够描述不同的收敛程度，我们通常用一个系统所有南向（下行）接口的总带宽比上这个系统所有北向（上行）接口总带宽的数值来表示，我们也将这个数值称为这个系统的收敛比。

举个例子，假设你有 10 台服务器，每台服务器通过 10GE 的接口连接到一个接入交换机，那我们一共就有 100G（ $10 \times 10G = 100G$ ）的南向带宽。假设这台交换机还有 2 个 40GE 的接口可以用于接入到更高一层的汇聚交换机，那我们一共就有 80G（ $2 \times 40G = 80G$ ）的北向带宽。此时，我们得到的收敛比则是 1.25: 1（ $100G \div 80G = 1.25$ ）。

需要说明的是，造成网络流量收敛的原因并不总是上述描述的这个例子，不过总的来说，我们可以将流量收敛的原因分为两类：

- 交换机不支持线速转发，在交换机内部可能形成流量收敛；
- 网络架构设计的原因，无论交换机是否线速，转发报文时也会存在流量收敛。

以下将分别以示例说明。



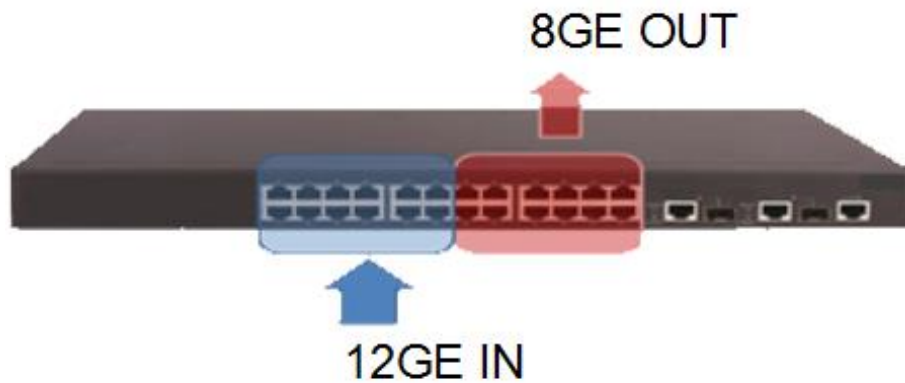
说明

文中对于传输报文速率、带宽收敛等的计算没有考虑到网络层协议等开销。

交换机非线速导致的收敛

某交换机只具有 8Gbps 线速转发的交换能力，某时刻从交换机前 12 个接口向后 12 个接口同时转发流量，当每个接口流量均达到 1Gbps 时，在交换机内部一定会有拥塞，此时便形成了转发的收敛（如图 1-13 所示）。实际每秒交换机接收流量为 12Gbps，但转发出去的报文只有 8Gbps，收敛比为输入带宽（12Gbps） \div 输出带宽（8Gbps）= 1.5: 1。

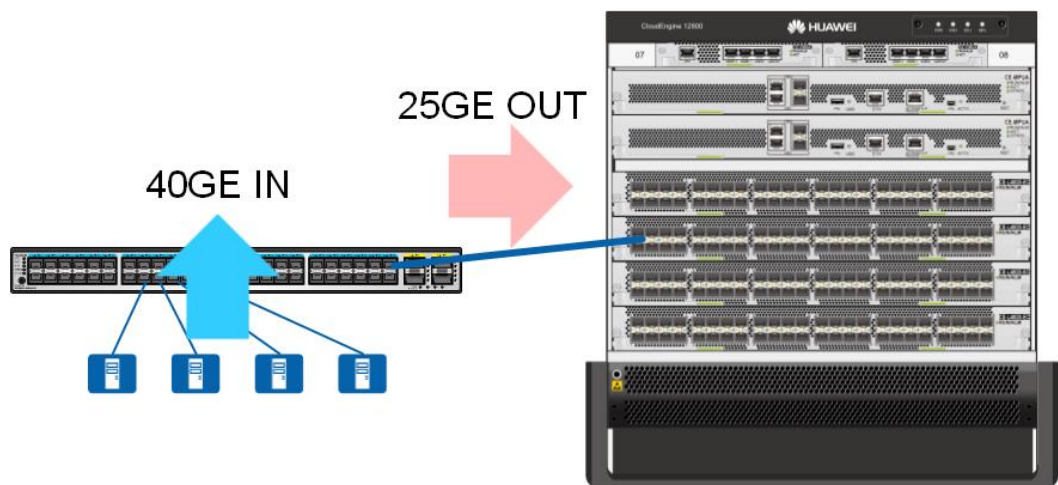
图1-13 交换机非线性速导致的收敛示意



网络设计导致的收敛

如图 1-14 所示，4 台服务器分别通过 10GE 链路连接接入交换机，接入交换机通过 1 条 25GE 链路连接核心交换机。即接入交换机的下行带宽为 40Gbps，接入交换机的上行带宽为 25Gbps。下上行链路收敛比为下行带宽（40Gbps）÷ 上行带宽（25Gbps）= 1.6: 1。

图1-14 网络设计导致的收敛示意



当然，最理想的收敛比是 1: 1。但是我们会注意到，低收敛比的设计意味着选用更高上行端口带宽的设备，这意味着更多的投入；如果在不计成本的情况下，1: 1 的收敛比是我们都期望能实现的。另外一方面，我们的服务器也不是每时每刻都工作在高负荷下，占用 100% 的带宽，这意味着即使不是 1: 1 的收敛比，也不是就一定会出现数据报文因拥塞丢包，业务仍可以正常运行。因此，找到这两者之间的平衡，找到最适合的收敛比，就显得十分有必要。

收敛比反映了一个网络线速转发流量的能力，因此通常我们会把收敛比作为衡量一个高性能网络的因素来考虑。一般在园区网，由于流量压力不大，园区网网络一般都会

存在较大的流量收敛；但在数据中心网络，由于其对性能要求高，流量收敛的设计就十分重要。

1.2 网络流量收敛设计

在进行网络流量收敛设计之前，我们需要了解网络中需要部署的业务应用及其特性，明确网络业务和流量模型。综合考虑东西、南北流量的大小、比例，来制定合适的收敛比和选择相应的设备。一般需要从以下几个方面考虑：

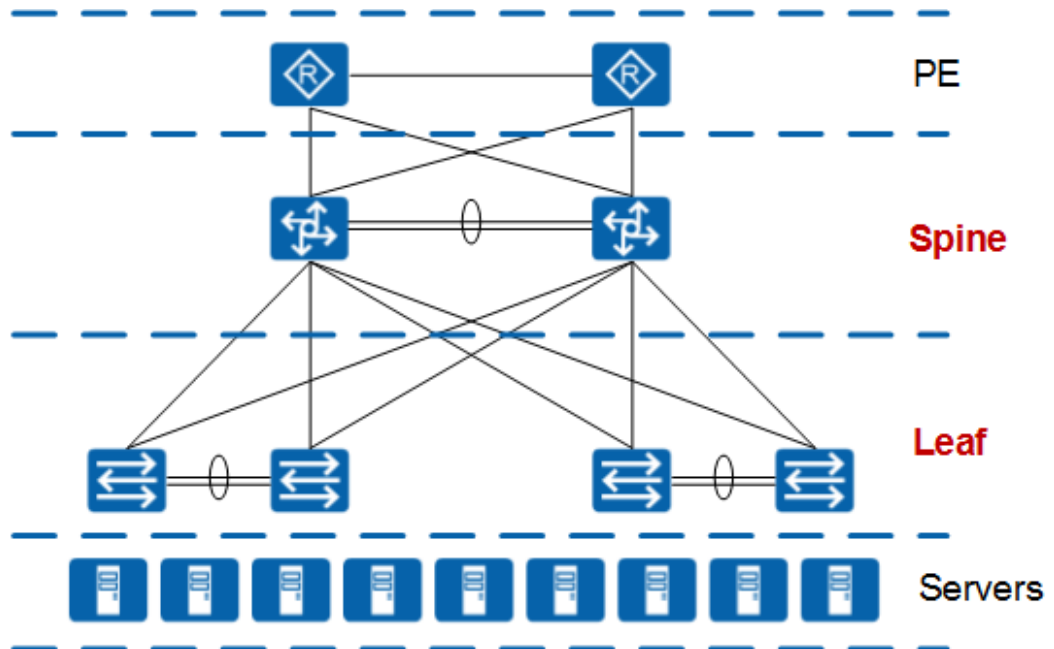
- 网络架构
- 可用链路设计
- 设备选型

在实际应用中，除非是对流量收敛比要求特别高的网络，我们也有一种简化的方法来考虑，即主要考虑设备的可用上行口的带宽来设计。

网络架构

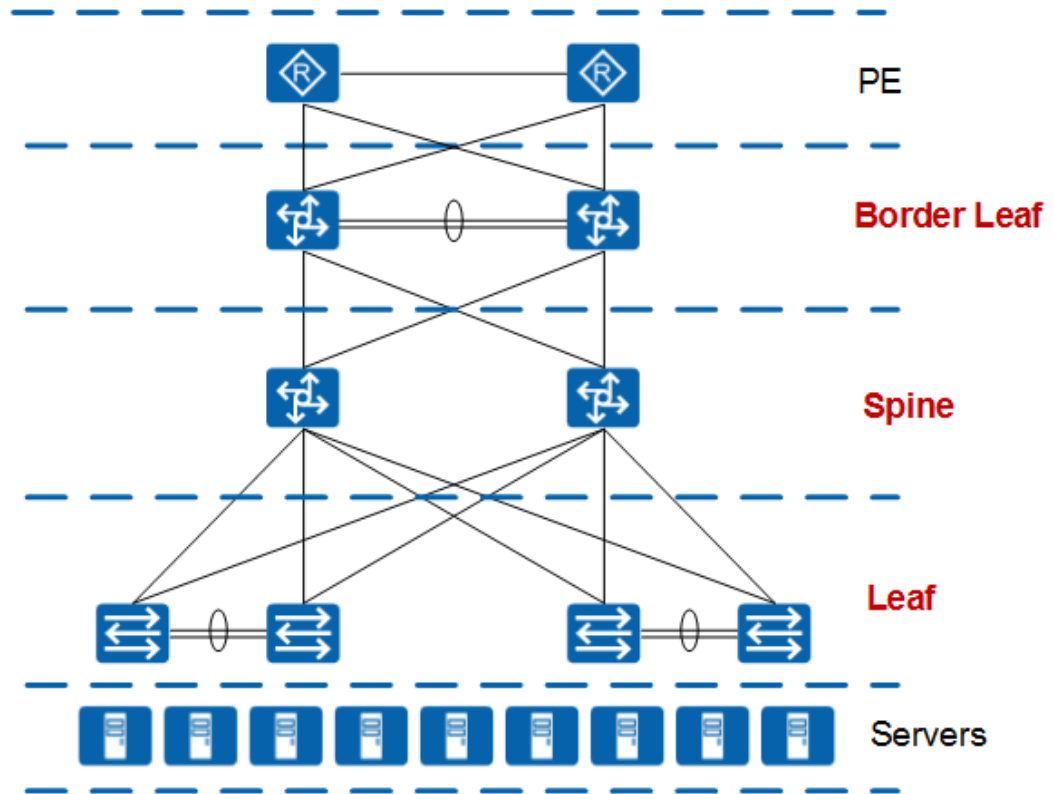
目前常见的网络结构一般采用“两层”结构。这里的“二层”是指 Spine+Leaf 两层设备的扁平化设计（如图 1-15 所示），二层架构比多层架构整体上具有更小的收敛比，在性能要求更高的数据中心等环境应考虑为二层扁平化架构设计。

图1-15 数据中心二层网络架构示意



现在，还有很多数据中心网络采用 Border Leaf、Spine、Leaf 的组网设计（如图 1-16 所示）。这种设计实际上也是二层的结构，因为 Border Leaf 和 Leaf 都是属于同一层的。在政府或金融或某些特殊领域等，由于业务架构或网络安全性的要求，需要 Spine 层设备和网关分离，部署安全隔离等，则会采用这种结构。

图1-16 数据中心 Border Leaf+Spine+Leaf 网络架构示意



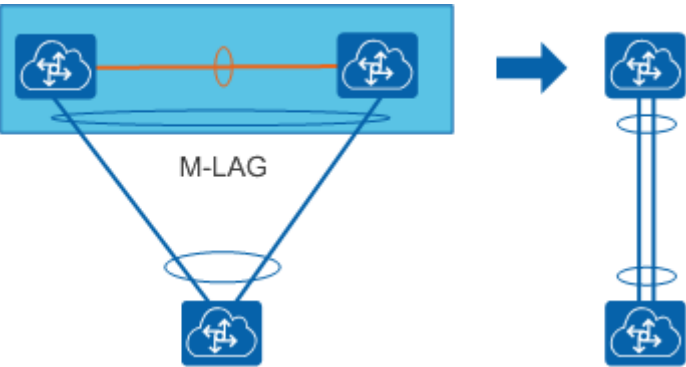
在实际设计中应该根据需要灵活选择，例如性能要求高的应用系统网络架构采用二层网络设计，而安全性要求高的应用系统网络架构采用三层或多层网络设计。

可用链路设计

在传统数据中心网络中，通常以生成树协议(STP)配合网关冗余协议(VRRP)提供服务器接入的可靠性。同时，服务器以多网卡连接网络以进一步提供冗余能力。但此种设计的冗余链路往往只能在主用链路故障时才发挥作用，链路及设备的利用率不高，也影响着网络的收敛比。

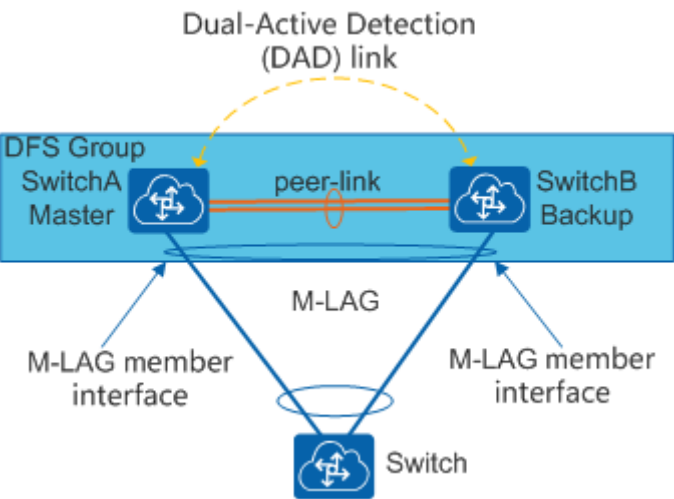
华为推荐采用跨设备链路聚合（M-LAG）技术构建数据中心网络。如图 1-17 所示，M-LAG 可以看做一种横向虚拟化技术，将双归接入的两台设备在逻辑上虚拟成一台设备。M-LAG 提供了一个没有环路的二层拓扑，同时 M-LAG 成员口所在链路均参与转发，不存在链路的浪费情况。

图1-17 M-LAG 示意图



需要注意的是，当采用 M-LAG 方式接入的时候，需要规划好端口的使用，因为 M-LAG 的 peer-link 链路和双主检测链路都需要预留端口，如图 1-18 所示。

图1-18 M-LAG 端口使用示意图



关于 M-LAG 链路接口的说明可参见表 1-7。

表1-7 M-LAG 链路接口说明

链路接口	说明
peer-link 链路	peer-link 链路是一条直连链路且必须做链路聚合，用于交换协商报文及传输部分流量。接口配置为 peer-link 接口后，该接口上不能再配置其它业务。 为了增加 peer-link 链路的可靠性，推荐采用多条链路做链路聚合。
双主检测链路	双主检测链路是一条三层互通链路，用于 M-LAG 主备设备间发送双主检测报

链路接口	说明
	文。
M-LAG 成员接口	M-LAG 主备设备上连接用户侧主机（或交换设备）的 Eth-Trunk 接口 为了增加可靠性，推荐链路聚合配置为 LACP 模式。

在服务器接入侧，也有类似的设计方式。若服务器双网卡为主备方式，则可设计为只有主用链路生效、备用链路在主用链路故障时启用；若服务器双网卡为负载分担方式，则全部上联链路均可以使用，配合链路收敛比设计可以提高网络中的实际可用带宽，提升网络转发性能。

设备选型

上述举例均假设所有交换机的所有端口可以线速转发，如果交换机不能线速转发，还需要考虑在交换机上的收敛。因此，为保证数据中心网络的高性能，最好选用具有全线速能力的交换机设备。


华为 CloudEngine 系列交换机，全系列均支持线速转发。如表 1-8 所示，下面我们以 **CE6870-48T6CQ-EI** 为例，给大家介绍一下如何初步判断一台交换机是否支持线速转发。

表1-8 CE6870-48T6CQ-EI 性能参数

项目	CE6870-48T6CQ-EI
10GE BASE-T 接口	48
10GE SFP+接口	NA
100GE QSFP28 接口	6
交换容量	2.16Tbps/19.44Tbps

如表 1-8 所示，查询 CE6870-48T6CQ-EI 的性能参数，我们可以看到 CE6870-48T6CQ-EI 共有 48 个 10GE 接口和 6 个 100GE 接口。我们首先计算 CE6870-48T6CQ-EI 上所有端口能提供的总带宽。计算公式为：端口数×相应端口速率×2（全双工模式）。如果得到的总带宽≤交换容量，则我们可认为该交换机在交换容量上可做到线速转发。下面我们来计算验证一下。

计算可得： $(48 \times 10G + 6 \times 100G) \times 2 = 2160G \leq 2.16T$ 。因此 CE6870-48T6CQ-EI 是线速转发的。至于 CE6870-48T6CQ-EI 还有一个交换容量 19.44Tbps，这个是指堆叠情况下的交换容量。CE6870-48T6CQ-EI 最多支持 9 台设备进行堆叠，因此 $2.16T \times 9 = 19.44T \leq 19.44T$ ，所以 CE6870-48T6CQ-EI 在堆叠情况下，也可以做到线速转发。

 **说明**
此处以盒式交换机进行了说明，实际上框式交换机的情况要复杂的多。框式交换机的业务板卡需要正确配合交换网板的选用才能做到设备的线速转发，可通过[转发性能评估工具](#)来搭配选择，确保选择的业务板卡和交换网板组合可以线速转发。

下面为了便于说明，我们假设数据中心网络整体采用 Border Leaf-Spine-Leaf 的三层结构，为大家介绍网络收敛比的设计。

1.2.1 服务器接入 Leaf

服务器的接入涉及到布线方式，以及接入交换机的选择。常见的布线方式有 TOR 以及 EOR/MOR 两类，参见图 1-19 和图 1-20。这两类方式可以组合出丰富的部署方式，在具体项目中，我们需要根据业务部署的要求，灵活选择不同的部署方式或组合。

图1-19 TOR 接入方式

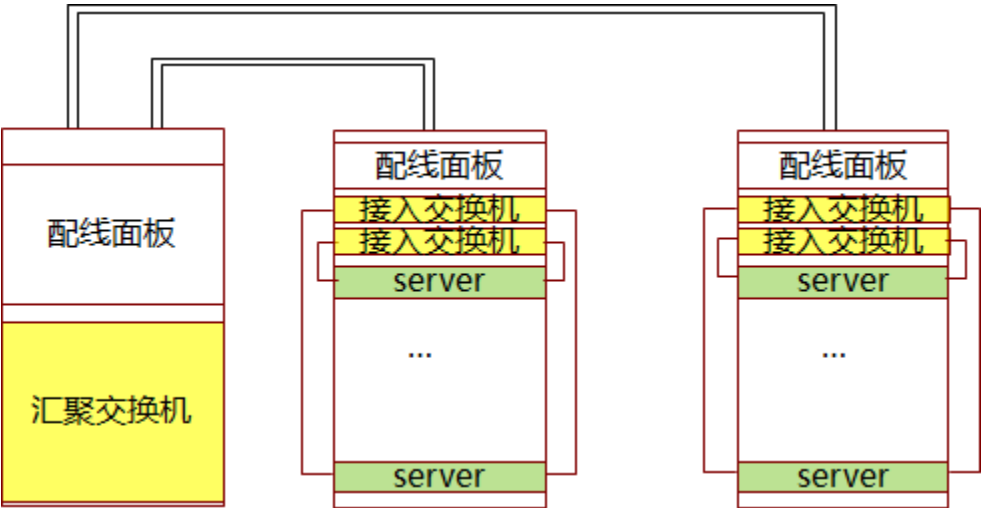
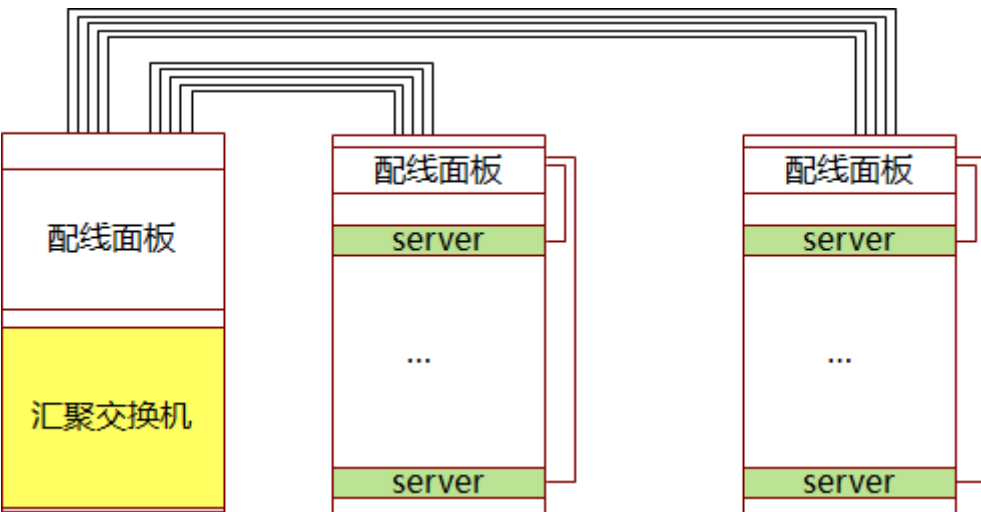


图1-20 EOR/MOR 接入方式



两者的对比如下，参见表 1-9。

表1-9 服务器接入方式对比

部署方式	TOR (Top Of Rack)	EOR (End Of Rack) /MOR (Middle Of Rack)
服务器类型	1U/2U/4U 机架服务器、刀片服务器(直通)	1U/2U/4U 机架服务器、刀片服务器(交换模块)
适用场景	高密度服务器机柜	低密度服务器机柜，高密刀片机柜
布线	简化服务器机柜与网络机柜间水平布线	布线复杂
维护	接入设备多，网络管理维护复杂； 电缆维护简单，扩展性好。	接入设备少，维护简单； 电缆维护复杂。

在数据中心网络的场景中，大部分属于高密度服务器的场景，因此采用 TOR 的方式比较常见。不同服务器和场景的部署思路可参考表 1-10。

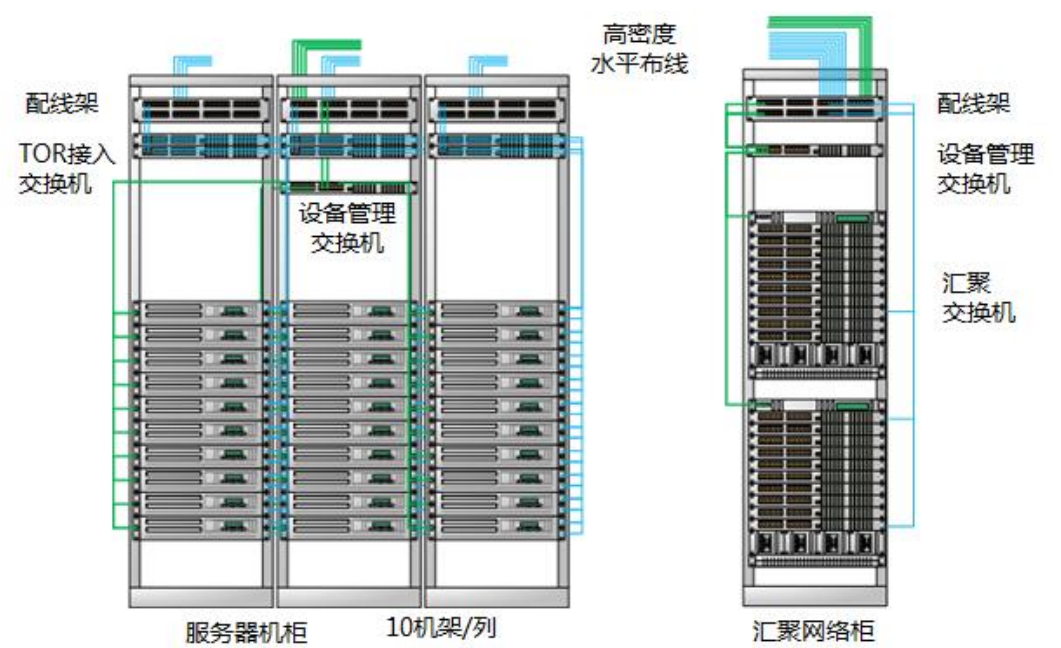
表1-10 服务器部署典型场景

服务器密度	服务器种类	I/O 模块配置	业务网口	存储网口	业务管理网口	建议的接入交换机部署
低密度分区	机架式服务器 2U 为例	固定 GE/10GE，并配置 FC 及 GE/10GE 网卡	2 个 GE/10GE	0~2 个 FC 或 GE/10GE	0~2 个 GE/10GE	EOR/MOR/TOR
高密度分区	机架式服务器 2U 为例	固定 GE/10GE，并配置 FC 及 GE/10GE 网卡	2 个 GE/10GE	0~2 个 FC 或 GE/10GE	0~2 个 GE/10GE	TOR
	刀片服务器 以 10 刀片/框为例 8U	配置直通模块 (每个网口与刀片通过背板直接连接)	20 个 GE 直通	0,10,20 个 FC 直通 或 0,10,20 GE 直通	0,20 个 GE	TOR 以太链路需小于 TOR 最大接入密度
		配置交换	2~4 个	0~6 个	0~2 个	EOR/MO

服务器密度	服务器种类	I/O 模块配置	业务网口	存储网口	业务管理网口	建议的接入交换机部署
		模块 (交换模块相当于一个接入交换机)	10GE	FC 或 10GE	GE/10GE	R

下面我们以高密机架服务器场景为例，介绍下具体的接入方案。如图 1-21 所示，每机架部署了 10 台机架式服务器（通常会部署 8~12 台服务器），每排共有 10 个机架（通常每排会部署 8~12 个机架）部署了服务器，构成高密度机架服务器部署。某数据中心共有 4 排这样的机架。

图1-21 高密机架服务器接入方案



每台服务器具有 2 个 10GE 业务网口和 1 个 FE 接入 BMC 管理口。为了保证可靠性每台服务器采用 M-LAG 的方式接入网络。相邻的两个机架组成一组 M-LAG 的系统，这样一个机架上面的 TOR 交换机需要接入 200G（10G×10×2=200G）带宽的流量。

根据经验值，我们在接入层的收敛比一般控制在 3: 1 左右，这主要取决于我们将为接入交换机设计多大的上行带宽。目前来说，接入交换机的单个上行接口可以达到 40G 的带宽，理论上通过 4 个 40G 的上行接口，我们就可以大致将收敛比做到 1: 1。但是此时我们至少需要为该 Spine 设置 4 台汇聚交换机，且每增加一个上行接口就需要增加一台汇聚交换机，因此这个开销还是十分可观的。在实际部署中，我们一般设置两台汇聚交换机，接入交换机通过 2 个 40GE 接口接入汇聚交换机，提供 80G（40G×

2=80G)的上行带宽。这样我们就可以得到 2.5: 1 ($200G \div 80G = 2.5$) 的收敛比, 很好的将其控制在 3:1 以内。

根据以上的分析, 我们推荐选择 **CE6870-24S6CQ-EI** 作为 TOR 接入交换机, 如图 1-22 所示, 该交换机具有 24 个 10GE SFP+以太网光接口和 6 个 40GE/100GE QSFP28 以太网光接口。其上行口还支持 100GE 的光模块, 在不增加汇聚交换机的情况下, 通过 100GE 接口也可以达到接近 1: 1 的收敛比。除此之外, CE6870 系列交换机提供 4GB 的大缓存, 可轻松应对视频、搜索等应用引起的流量浪涌。CE6870-24S6CQ-EI

图1-22 CE6870-24S6CQ-EI



CE6870 系列还提供 48 个 10GE 接口的款型, 可支持更高密度的服务器接入需求。如果不需要用到大缓存, 也可以选用 **CE6851-48S6Q-HI** 作为 TOR 接入交换机。

说明

此处没有考虑服务器和交换机管理网口的接入需求。由于管理网口不需要大的带宽保证, 从控制成本的角度出发, 一般会选择价格较便宜的接入交换机 (只要满足接口接入需求即可), 例如华为 S5700 系列交换机, 采用 MOR/EOR 的部署方式。

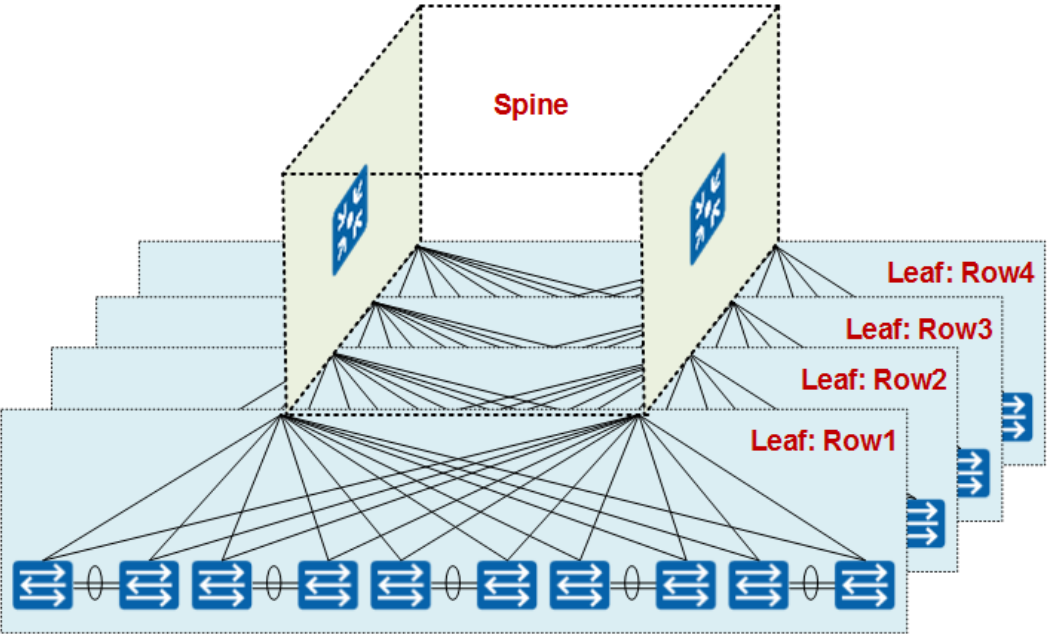
1.2.2 Leaf 接入 Spine

下面我们接着看 Spine 层交换机的选择。每排机架上共有 10 台 TOR 交换机需要连接到我们的汇聚交换机, 每台 TOR 交换机通过 2 个 40GE 接口接入到 Spine 层设备, 一共是 80 个 ($2 \times 10 \times 4 = 80$) 接口, 3200G ($80 \times 40G = 3200G$) 的带宽。

此时我们可以选择将这些接口分为几个 Spine 节点接入到 Spine 层设备。为了避免的设备单点故障引起网络问题, 我们每个 Spine 节点都至少有 2 台 Spine 层交换机。

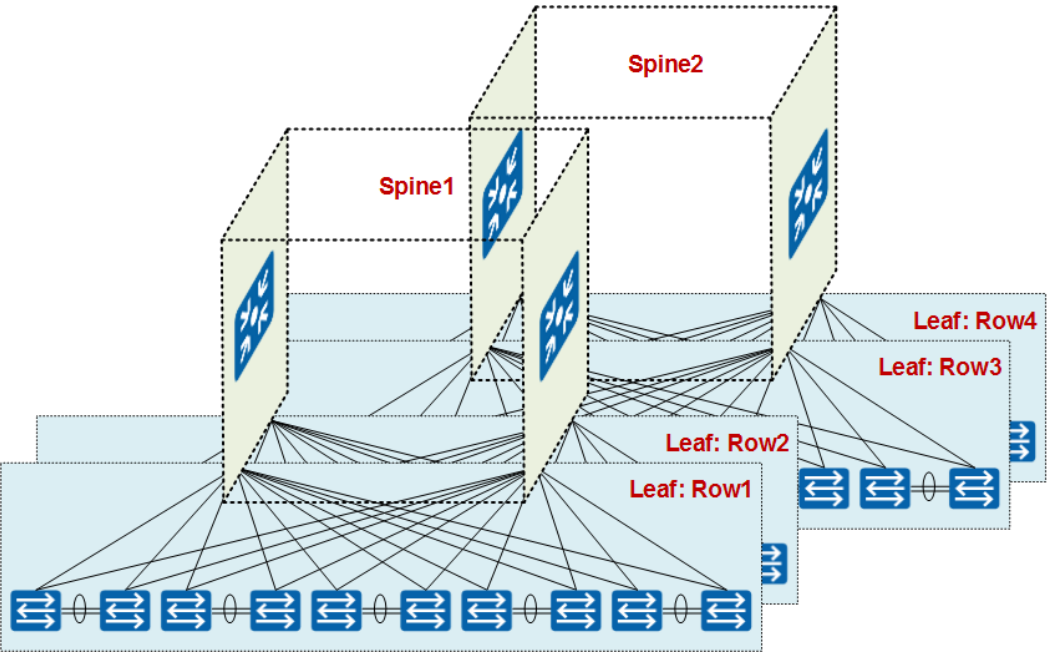
如果我们选择将这些 TOR 交换机作为一个 Spine 接入, 那么这个 Spine 接入的接口数为 80 个接口, 3200G 的带宽, 如图 1-23 所示。在这种情况下, 我们如果通过 100G 的上行链路链接到 Border Leaf 设备 (一般为两台, 北向连接到出口路由), 则可以提供 400G ($4 \times 100G = 400G$) 的带宽, 此时收敛比为 8:1 ($3200G \div 400G = 8$)。

图1-23 1 个 Spine 接入



如果我们选择将这些 TOR 交换机分为两个 Spine 接入，那么每个 Spine 接入的接口数为 40 个接口，1600G 的带宽，如图 1-24 所示。在这种情况下，我们如果通过 100G 的上行链路链接到 Border Leaf 设备，则可以提供 400G（ $4 \times 100\text{G} = 400\text{G}$ ）的带宽，此时收敛比为 4:1（ $1600\text{G} \div 400\text{G} = 4$ ）。

图1-24 2 个 Spine 接入



值得注意的是，在数据中心网络中 **Spine** 的划分，收敛比并不是唯一的依据，更主要的是根据业务和功能的分区来划分的。另外受限于交换机本身路由、ARP 等规格的限制，再加上现在虚拟机的大规模应用（虚拟比达到 1:30 或更高，对规格要求更高），一个 **Spine** 的规模也不会太大。

在这个场景下，我们采用了两个 **Spine** 接入的方式，即将 **Row1** 和 **Row2** 共用一个 **Spine**，**Row3** 和 **Row4** 共用一个 **Spine**。此时我们一般选用框式交换机或高性能的盒式灵活插卡交换机来作为 **Spine** 设备。

如图 1-25 所示，我们推荐选择 **CE12804** 作为 **Spine** 交换机。**CE12804** 具有 4 个业务板卡槽位（**CE12800** 系列交换机可以提供 4、8、12、16 个业务板卡槽位，可根据需要灵活选择），可选的板卡种类丰富，支持高密 40GE 板卡（最高可提供 144 个 40GE 接口， $36 \times 4 = 144$ ）和高密 100GE 板卡（最高可提供 144 个 100GE 接口， $36 \times 4 = 144$ ）。可根据实际需求灵活选择和搭配，也便于后期的调整或者扩容。

图1-25 CE12804



在上述的需求下，我们可以为每台 **CE12804** 交换机选择 2 块 24 接口的 40GE 单板（**CE-L24LQ-EC1**）用于下行连接和 1 块 4 接口的 100GE 单板（**CE-L04CF-EF**）用于上行连接。

实际上，还有一种更低成本的方案，即选用 **CE8860-4C-EI** 作为 **Spine** 交换机，如图 1-26 所示。**CE8860-4C-EI** 是具有 2U 高度的灵活插卡交换机，最多可以插入 4 个插卡。在以上需求下，我们可以选择 3 块 16 端口的 40GE 插卡（**CE88-D16Q**）和 1 块 8 端口的 40GE/100GE 插卡（**CE88-D8CQ**）。**CE88-D16Q** 用于下行连接；**CE88-D8CQ** 用于上行连接。

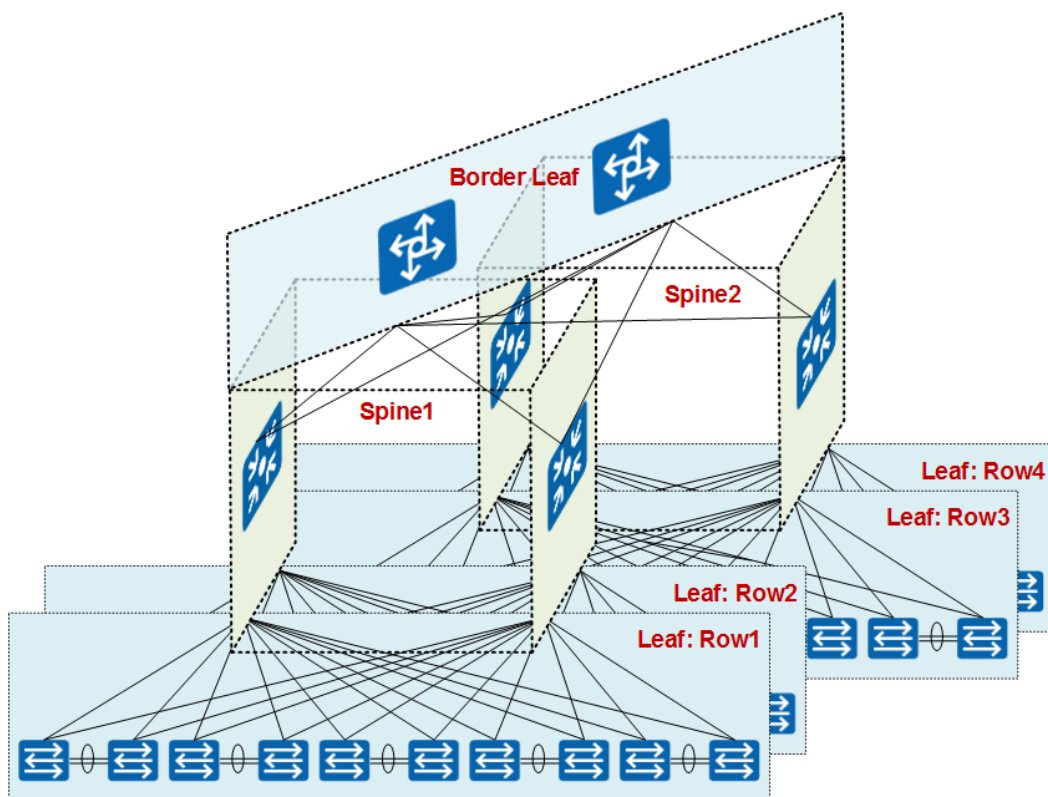
图1-26 CE8860-4C-EI



1.2.3 Spine 接入 Border Leaf

最后我们来看 Border Leaf 层的设计，如图 1-27 所示。Border Leaf 北向主要是连接出口路由器，南向连接不同的 Spine，承担 Spine 间东西向流量的转发。Border Leaf 的设计很重要的一个是考虑客户所购买的出口路由器的端口。这些端口相比较于我们下层的网络设备比较贵，一般情况下都是 10GE/40GE 的接口。这也意味着我们在 Border Leaf 的收敛比会比较大。

图1-27 Spine 接入 Border Leaf 示意

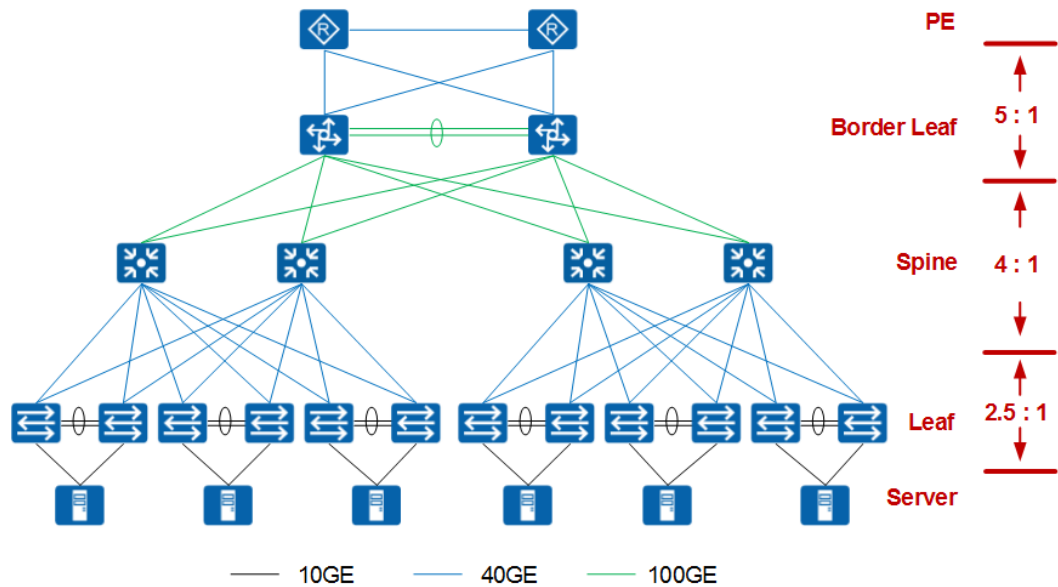


如果我们按照 4 个 40GE 的出口接口来计算，我们将有 160G 的带宽，此时收敛比是 5:1 ($800\text{G} \div 160\text{G} = 5$)。但是，根据目前统计约 75% 的流量都是发生在数据中心的内部，即东西向的流量。那么剩下的 25% 的流量，即南北向的流量大概只有 200G ($800\text{G} \times 25\% = 200\text{G}$)。如果按照这个来估算，我们的收敛比为 1.25:1 ($200\text{G} \div 160\text{G} = 1.25$)，属于可接受的范围。

由于 Border Leaf 的设备不需要很多接口，我们选用上面提到的 **CE6870-24S6CQ-EI** 作为 Border Leaf 设备就可以了。

最终我们得到网络收敛设计的逻辑图，如图 1-28 所示。

图1-28 收敛比设计逻辑示意图



1.3 本章小结

在设计数据中心网络收敛比时，我们需要根据网络业务和流量模型。综合考虑东西、南北流量的大小、比例，来制定合适的收敛比和选择相应的设备。根据经验值，可参考以下设计：

- 在服务器接入的 Leaf 层，南北向收敛比一般控制在 3:1 以下；
- 在 Spine 层，考虑和 Leaf 层的收敛比接近或更小；
- 在 Border Leaf 层，收敛比一般较大，根据客户的出口路由带宽灵活设计。

当采用二层的结构时，Spine 层除了南北向流量，东西向流量的压力也会更大，需要选用高性能的交换机，并尽量增加 Spine 间互联的带宽。

华为 CloudEngine 系列交换机，具有线速转发，高端口密度和大缓存等特点，是您构建低收敛比高性能网络的最佳伙伴。了解华为 CloudEngine 系列交换机更多转发性能和接口信息，请访问：<http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches>。

第五章 Fabric 网络

1.1 简介

在前面的几个章节里，我们详细讨论了在构造数据中心网络时位置如何摆放、机框如何布线以及流量收敛比的计算，相信你已经对华为 CE 系列数据中心交换机的物理架构有了一个很好的了解。接下来将要介绍的是 CE 系列交换机提供的不同的逻辑体系架构，说到逻辑架构，很多人肯定觉得，逻辑层面的设计不外乎就是那些环路协议算法、IP 寻址以及二三层转发机制（二层 VLAN+xSTP、三层路由），都是一些老生常谈的成熟技术了。所以在开始之前，我们首先得明确的一点就是，数据中心网络相较于园区网络有什么不同？

随着云计算的发展，在数据中心网络中服务器虚拟化技术得到广泛应用，但服务器在迁移时，为了保证迁移时业务不中断，就要求不仅虚拟机的 IP 地址不变，而且虚拟机的运行状态也必须保持原状（例如 TCP 会话状态），所以虚拟机的动态迁移只能在同一个二层域中进行，而不能跨二层域迁移，这就要我的二层网络足够大。而传统的二层技术，不论是通过缩小二层域的范围和规模来控制广播风暴的规模（VLAN, Virtual Local Area Network）亦或是阻塞掉冗余设备和链路来破坏（xSTP, Spanning Tree Protocol），网络中能够容纳的主机数量、收敛性能以及网络资源的带宽利用率对于数据中心网络而言是远远不够的。

此时，轮到华为 CE 系列交换机大显神通了，提供了多种解决二层扩展能力技术，包括了基于设备虚拟化的堆叠以及 M-LAG（Multichassis Link Aggregation Group），这些都是构造数据中心基础逻辑网络的关键，下面就让我们来讨论基于不同的场景每种技术的优缺点，哪些设计可能更适合。

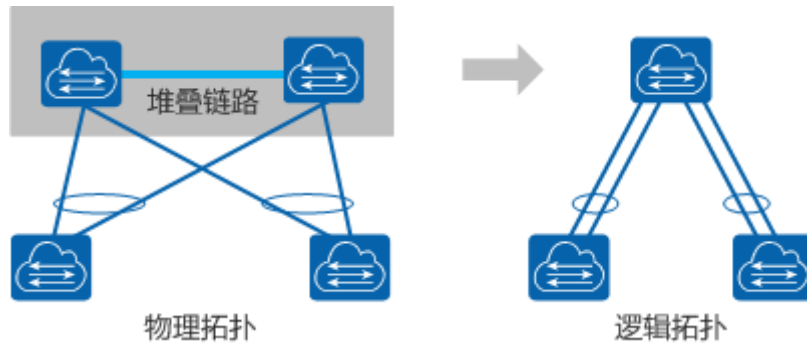
说明

如果你想要对数据中心的大二层网络技术有更多的了解，可以参考《闲话大二层网络》，请访问 <http://forum.huawei.com/enterprise/forum.php?mod=viewthread&tid=243089>。

1.2 堆叠组网方案

堆叠是指将多台交换机设备通过线缆连接后组合在一起，虚拟化成一台设备，作为一种横向虚拟化技术，将多台设备在逻辑上虚拟成一台设备，可以简化网络的配置和管理。同时，结合跨设备链路聚合技术，不仅可以实现设备及链路的高可靠性备份，而且可以避免二层环路。相对传统的 STP 环路保护，逻辑拓扑更加清晰、链路利用更加高效。

图1-29 堆叠示意图



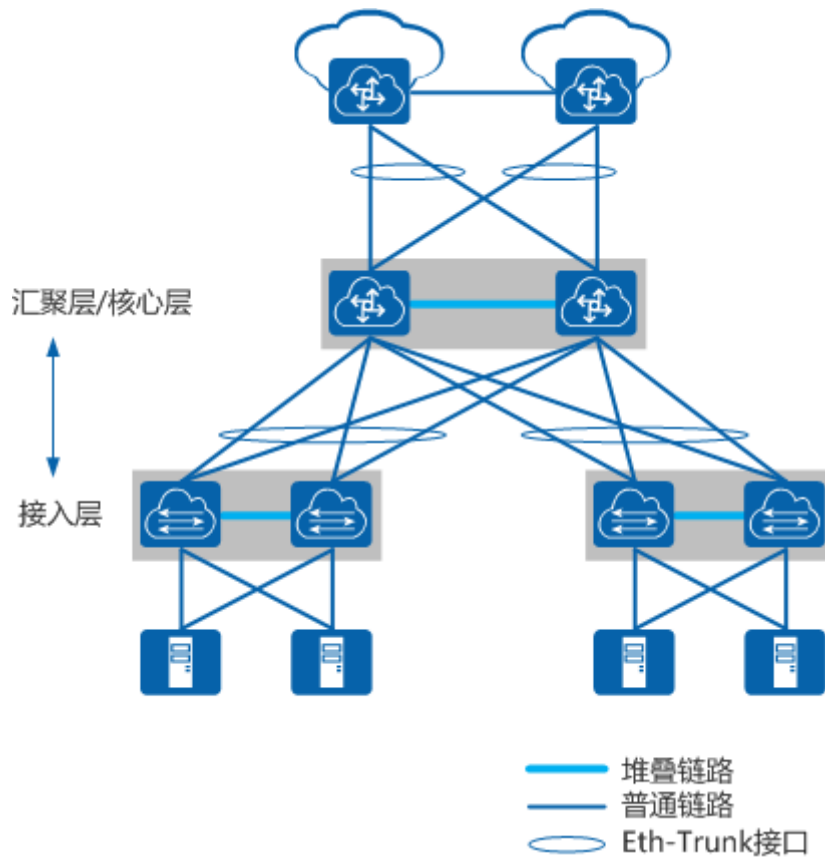
华为 CE 系列交换机提供的堆叠技术有 CSS 技术（框式堆叠）、iStack（盒式堆叠），主要典型特征有：

- 1、 交换机多虚一：CSS/iStack 对外表现为一台逻辑交换机，控制平面合一，统一管理。
- 2、 转发平面合一：CSS/iStack 内物理设备转发平面合一，转发信息共享并实时同步。
- 3、 跨设备链路聚合：跨 CSS/iStack 内物理设备的链路被聚合成一个 Eth-Trunk 端口，和下游设备实现互联。

如图 1-30 所示的数据中心网络中，接入层、汇聚层/核心层通过部署堆叠构造出一个逻辑简单、无环的网络。

- 接入层使用低成本的 CE8800&7800&6800&5800 系列交换机来部署堆叠。终端设备（服务器或其他网络设备）双归接入堆叠，可以保证接入链路的高可靠性。
- 汇聚层/核心层使用高性能 CE12800 系列交换机部署堆叠，与接入层之间通过跨框链路聚合连接，形成一个无环的网络。

图1-30 数据中心内堆叠组网图

**部署方案：**

- 通过堆叠（CSS/iStack）技术保证节点的可靠性：一台设备故障后，另外一台设备自动接管所有的业务。
- 通过 CSS+LAG+iStack 部署端到端可靠性架构，打造无间断数据中心，保证业务持续运营。
- 多台接入层 iStack 堆叠，2 台汇聚层 CSS 集群堆叠。
- 接入与汇聚间采用多条 10GE 或 40GE 链路全连接，保障链路高可靠。
- 汇聚与核心间采用高速 40GE 链路全连接，确保汇聚到核心无阻塞转发。

方案特色：

- 简化管理和配置

堆叠建立后，多物理设备虚拟成为一台设备，用户可以通过任何一台成员设备登录堆叠系统，对所有成员设备进行统一的配置和管理，使网络需要管理的设备节点减少一半以上。

其次，组网变得简洁，不再需要配置 xSTP、VRRP 等协议，简化了网络配置。

- 带宽利用率高

采用链路聚合的方式，带宽利用率可以达到 100%（STP 会阻塞链路），采用逐流方式负载均衡，支持多种负载分担方式。

- 快速的故障收敛
相对于 STP 秒级的故障收敛时间，链路聚合的故障收敛时间可控制在 ms 级内，大大降低了网络链路或节点故障对业务的影响。
- 扩容方便、保护投资
随着业务的增加，当用户进行网络升级时，只需要增加新设备即可，在不需要更改网络配置的情况下，平滑扩容，很好的保护了投资。

堆叠技术从低端盒式到高端框式都已经被广泛应用，具备了相当的成熟度和稳定度，估计会给部分读者造成堆叠技术简直堪称完美的错觉，但我们要知道任何技术都不是完美的，就堆叠技术而言，它的不足之处在于：

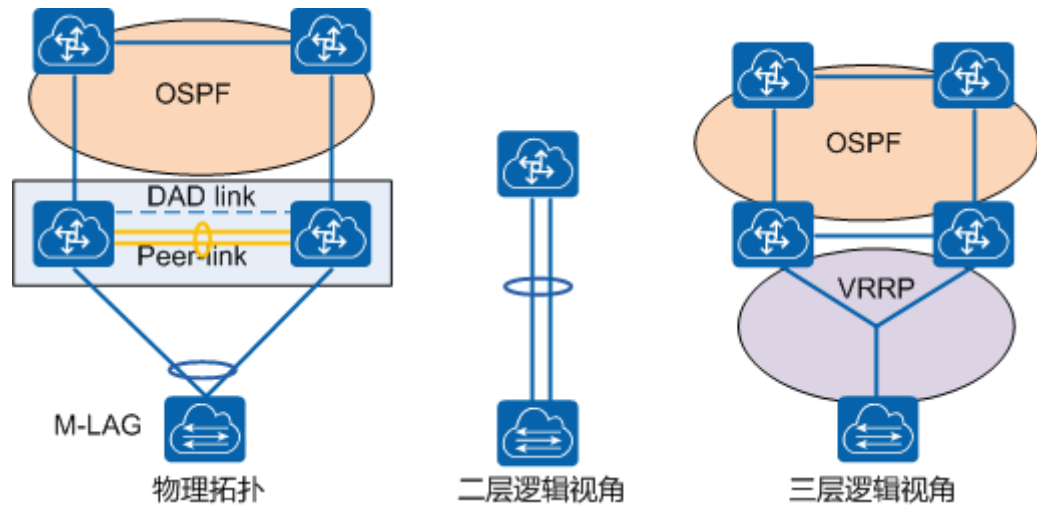
- 堆叠的虚拟化局限于单个层次
虽然横向虚拟化技术在一定程度上优化了网络结构、减少了管理节点，但是做的还不够彻底。一方面是横向虚拟化后依然没有减少网络的层级；另一方面是依然没有彻底解决管理节点较多的问题。大规模的数据中心都有高密度接入的特点，有大量的接入交换机，为了可靠性一般是多台接入交换机虚拟化（多是 2 台虚拟化），这样即使在做了横向虚拟化后管理节点的数量也是相当可观的。假设接入层有 40 台接入交换机，每两台交换机做虚拟化，那么依旧还是有多达 20 个管理节点。
- 堆叠虚拟化控制面多虚一
网络设备虚拟化之后，所有主控平面合一，但是这种合一只能采用主备备份的模式，即只有主设备的主控板正常工作，而其他主控板都处于备份状态。因此，整个系统的物理节点规模就受限于主控节点的处理能力，不是想做多大就做多大的。例如框式设备虚拟化一般为 2 台，盒式设备一般为 16 台。目前最大规模的虚拟化系统大概可以支持接入 1~2 万台主机，可以从容应付一般的中、小型数据中心，但对于一些超大型的数据中心来说，就显得力不从心了。

1.3 M-LAG 组网方案

M-LAG（Multichassis Link Aggregation Group）即跨设备链路聚合组，是一种实现跨设备链路聚合的机制，将一台设备与另外两台设备进行跨设备链路聚合，从而把链路可靠性从单板级提高到了设备级。

对二层来讲，可将 M-LAG 理解为一种横向虚拟化技术，将 M-LAG 的两台设备在逻辑上虚拟成一台设备，形成一个统一的二层逻辑节点。M-LAG 提供了一个没有环路的二层拓扑同时实现冗余备份，不再需要繁琐的生成树协议配置，极大的简化了组网及配置。这种设计相对传统的 xSTP 破坏保护，逻辑拓扑更加清晰、链路利用更加高效。

图1-31 M-LAG 物理与二三层逻辑结构示意图



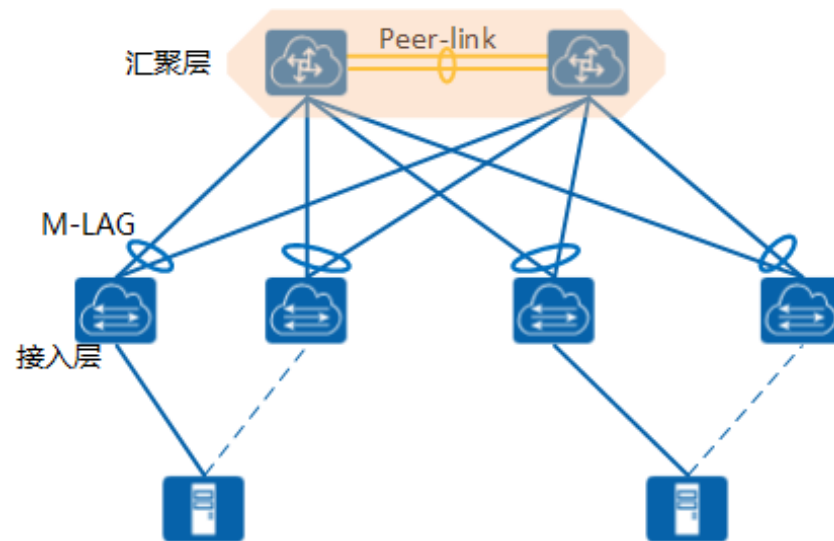
如图 1-31 所示，M-LAG 配对交换机对外提供 M-LAG 接口，用于接入二层业务；M-LAG 配对交换机之间部署 Peer-Link，用于 M-LAG 协议消息交互，以及设备间横向业务流量转发；从三层视角看，M-LAG 的配对设备又是两个独立的设备，可以支持独立的网管，并作为独立的 OSPF 路由节点。同时，M-LAG 支持本地优先转发，最大程度减少 M-LAG 配对设备之间的东西向流量。M-LAG 支持双主检测，由于两台配对设备为独立设备，因此通过带内或者带外的 IP 可达检测即可达到目的，不需要为此另外加线缆。

部署方案

- 组网方案一：汇聚层组建 M-LAG

通过跨设备端口虚拟化技术（M-LAG），实现汇聚层和接入层交换机之间的网络逻辑无环化，取代 STP。汇聚层两台交换机配对，汇聚交换机之间横向链路配置为 peer-link。两台汇聚交换机下行连接同一接入交换机的端口配置为跨框的 Eth-Trunk。

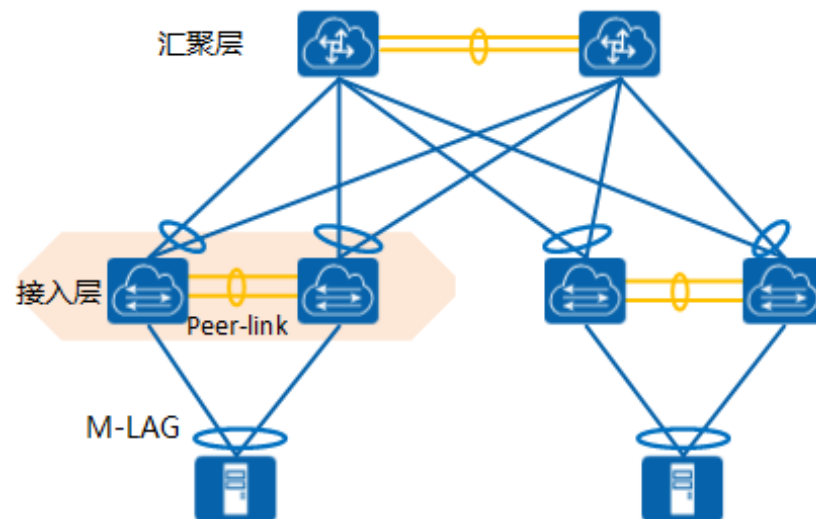
图1-32 汇聚层使用 M-LAG 示意图



这种设计相对传统的 STP 断点保护，逻辑拓扑更加清晰、链路利用更加高效。M-LAG 的配对设备，控制平面和管理平面独立，只有协议面的耦合，理论上可靠性相对堆叠更高，还提供设备独立升级的能力，带来维护的方便性。

- **组网方案二：接入层组建 M-LAG**

同样 M-LAG 技术适用于服务器双网卡要求双活接入的应用场景。服务器双活接入两网卡共享 MAC。双网卡实现基于流的负载分担策略。因此，通过 M-LAG 将服务器连接的端口配置为 Eth-Trunk，两个端口的 MAC 和 ARP 表项同步。接入层使用 M-LAG 示意图



方案特色

M-LAG 技术本质上还是控制平面虚拟化技术，但是和堆叠技术不同的是，由于 M-LAG 的目的仅仅是在链路聚合协商时，对外表现出同样的状态，所以不需要像堆叠那样同步设备上所有的信息，只需要同步接口和表项相关的一些内容。这样，控制面耦合程度相比堆叠来说，会小很多，而且堆叠技术的一些缺陷在 M-LAG 上也会缓解很多，比如上面我们说过的堆叠的三个主要的问题：

- 可靠性问题：M-LAG 需要同步的仅仅是协议面的一些内容，并不需要同步所有的设备状态，理论可靠性相对堆叠更加好。
- 维护问题：M-LAG 两台设备可以进行独立升级。仅协议面耦合，中断时间较短。
- 扩展性问题：M-LAG 技术主要目的是为了解决接入侧多路径问题，在数据中心网络中一般会配合路由或者一些大二层技术来实现网络侧的多路径转发问题。

1.4 本章小结

如表 1-11 所示，我们对虚拟化技术做以下总结：

表1-11 虚拟化技术优缺点比较

虚拟化技术	堆叠	M-LAG
优点	<ul style="list-style-type: none">● 简化管理和配置；● 带宽利用率高；● 快速的故障收敛；● 扩容方便、保护投资。	<ul style="list-style-type: none">● 跨设备链路聚合，可靠性高；● 流量负载分担转发，链路利用率高；● 控制平面与管理平面独立，升级简单。
不足	<ul style="list-style-type: none">● 虚拟化局限于单个层次；● 多虚一后控制面单一带来的可靠性、维护困难以及扩展难等问题。	无法解决网络扩展性问题，需要通过路由或者其他大二层技术来实现网络侧多路径转发。

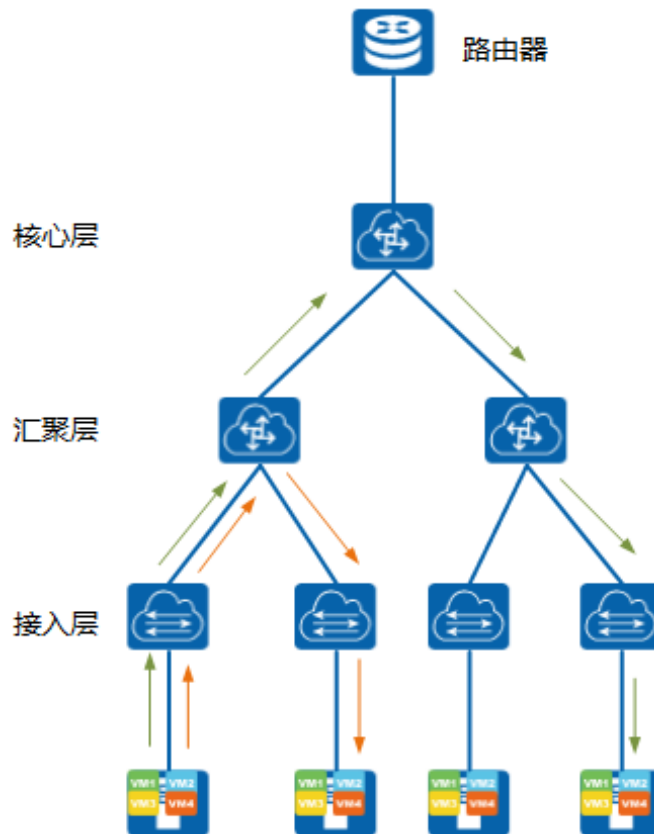
现在，读者应该对数据中心大规模二层网络的需求已经非常的清晰，CE 系列交换机提出了多种有针对性的技术和方案，满足当前要求和未来扩展需求。无论是堆叠技术，都是在控制平面将多台设备虚拟成一台的多虚一技术，如果需要将一台物理设备虚拟成多台逻辑设备，CE 系列交换机提供了 VS（Virtual System）技术，将一台物理设备 PS（Physical System）虚拟成多个相互隔离的逻辑系统。每个 VS 独立工作，在业务功能上等同于一台独立的传统物理设备。如果想要详细了解 VS 技术，可以参考《VS 分身有术》，请访问 <http://forum.huawei.com/enterprise/forum.php?mod=viewthread&tid=148461>。

第六章 IP Fabric & 三层路由

1.1 IP Fabric

直到几年前，大多数的数据中心网络还都是基于传统的三层架构，这些架构基本都是从园区网络设计中复制而来的。一个标准的传统三层的网络结构如图 1-33 所示：

图1-33 传统多层网络流量模型



对于大多数具有像园区网络这样的南/北配置的流量模型来说是很实用的，而且三层网络结构应用广泛而且技术成熟稳定。但随着技术的发展，它的瓶颈也不断涌现，那为什么三层网络结构存在短板？

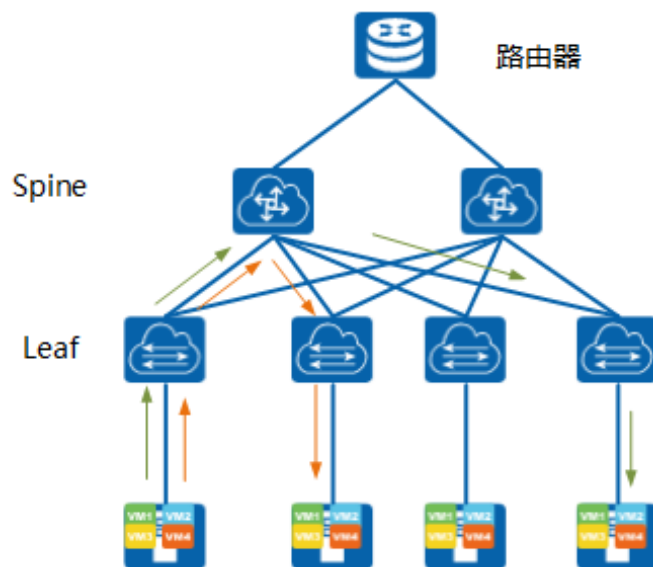
因为数据中心网络传输模式在不断的变化，大多数传统网络都是纵向(North-South)的传输模式---主机与网络中的其它非相同网段的主机通信都是设备-交换机-路由到达目的地。同时，在同一个网段的主机通常连接到同一个交换机，可以直接相互通讯。

然而，随着云计算的发展，横向(East-West)流量在数据中心中占据主导地位，涵盖几乎所有的云计算，虚拟化以及大数据。横向网络在纵向设计的网络拓扑中传输数据会带有传输的瓶颈，因为数据经过了许多不必要的节点(如路由和交换机等设备)。主机互访需要通过层层的上行口，带来明显的性能衰减，而三层网络的原始设计更会加剧这种性能衰减，这也就是为什么当前主流的三层网络拓扑结构越来越不能满足数据中心网络需求的原因。

在前面的第五章节，我们介绍了 Fabric 扁平化网络，整网在一个二层网络范围内，可以通过设备虚拟化技术（堆叠）以及跨设备链路聚合技术 M-LAG 来解决解决层网络环路以及多路径转发问题，也谈到了他们存在的规模限制，扩展性不足的问题。于是，就有了 IP Fabric 网络的概念。

什么是 IP Fabric 网络？IP Fabric 指的是在 IP 网络基础上建立起来的 Overlay 隧道技术。如图 1-34 所示，即为基于胖树的 Spine+Leaf 拓扑结构的 IP Fabric 组网图。

图1-34 IP Fabric 网络的两层架构



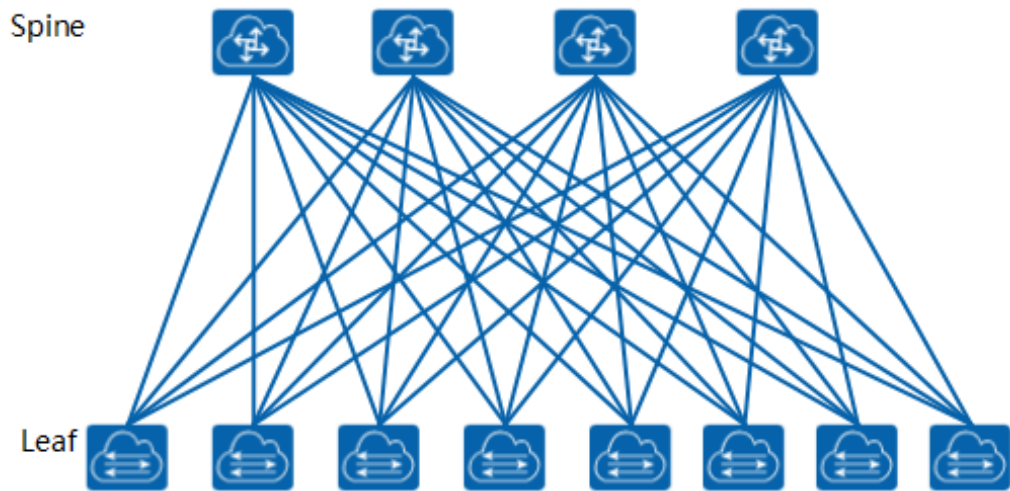
在这种组网方式中，任何两台服务器间的通信不超过 3 台设备，每个 Spine 和 Leaf 节点全互连，可以方便地通过扩展 Spine 节点来实现网络规模的弹性扩展。只要遍历一定数量的交换机，可以在几乎所有数据中心结构体系中的服务器节点之间传输流量，该架构由多条高带宽的直接路径组成，消除了网络瓶颈带来的潜在传输速度下降，从而实现极高的转发效率和低延迟。

根据不同的业务需要，Spine 和 Leaf 之间可以使用 IP 路由、VXLAN 或 TRILL 等技术。

- Spine 和 Leaf 之间使用 IP 路由
即三层到边缘，一般适用于协同计算业务，例如搜索。此类业务流量收敛比小（1:1~2:1），要求有一个高效的，无阻塞网络。
- Spine 和 Leaf 之间 VXLAN（Virtual eXtensible Local Area Network）或 TRILL（Transparent Interconnection of Lots of Links）
即大二层网络，适用于需要大范围资源共享或者虚拟机迁移的数据中心网络。

IP Fabric 网络允许简化扩展，仅受支持设备及其端口的数量限制，如图 1-35 所示。

图1-35 Spine+Leaf 网络架构



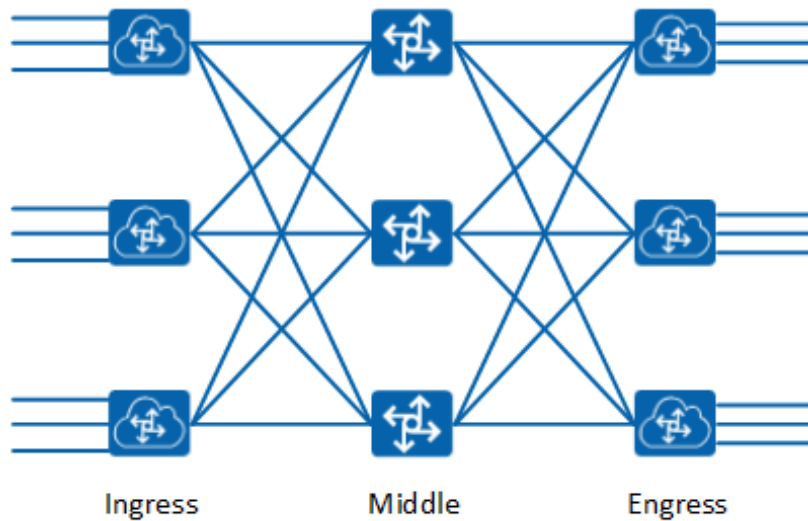
我们可以看到，Spine 层由四台设备组成。每台 Leaf 设备有四个上行链路连接到每个 Spine。此拓扑中支持的 Leaf 的最大数量由每个 Spine 设备的最大端口数决定。因此，如果我们的 Spine 交换机支持 40×40 GE 连接，叶子设备的最大数量将是 40（但考虑到存在上行连接，所以这边算成 36 更为合理）。

1.2 Spine+Leaf 网络架构起源

Spine+Leaf 两层设备的扁平化网络架构来源于 CLOS 网络，CLOS 网络以贝尔实验室的研究人员 Charles Clos 命名，他在 1952 年提出了这个模型，作为克服电话网络中使用的机电开关的性能和成本相关挑战的一种方法。Clos 用数学理论来证明，如果交换机按层次结构组织，在交换阵列（现在称为结构）中实现非阻塞性能是可行的，主要是通过组网来形成非常大规模的网络结构，本质是希望无阻塞。在此之前，要实现“无阻塞的架构”，只能采用 $N \times N$ 的 Cross-bar 方式。

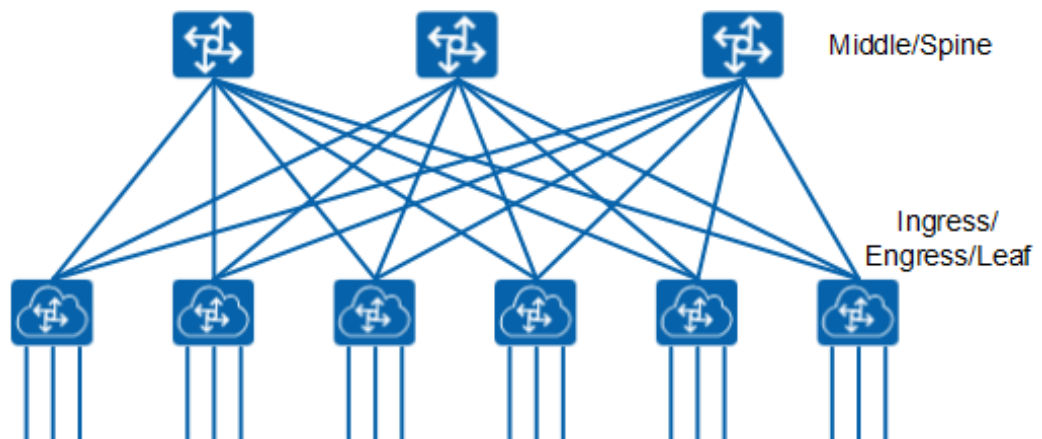
Charles Clos 提出的设计是他分为三层网络架构的 CLOS 模型。如图 1-36 所示，一个三层 CLOS 网络架构由一个 Ingress 节点，一个 Middle 节点和一个 Egress 节点组成。

图1-36 三层 CLOS 网络架构模型（Ingress、Middle&Egress）



现在假如我们将这种网络架构对折，统一放在一边，那么我们将得到与前面讨论过的 Spine+Leaf 相同的网络架构，如图 1-37 所示。

图1-37 对折三层 CLOS 网络架构模型



接入连接的数量仍然等于折叠后的三层 CLOS 网络架构的 Spine 与 Leaf 之间的连接数，正如本章后面将会讨论的，现在流量可以分布在所有可用的链接上，不用担心过载问题。随着更多的连接被接入到 Leaf 交换设备，我们的链路带宽收敛比将增加，可以通过增加 Spine 和 Leaf 设备间的链路带宽降低链路收敛比。

除了支持 Overlay 层面技术之外，Spine+Leaf 网络架构的另一个好处就是，它提供了更为可靠的组网连接，因为 Spine 层面与 Leaf 层面是全交叉连接，任一层中的单交换机故障都不会影响整个网络结构。因此，任一层中的一个交换机的故障都不会使整个结构失效。

1.3 BGP

在前面的 1.1 IP Fabric 章节，我们提到了 IP Fabric 组网根据不同的组网可以应用到的一些技术包括基本的三层路由协议以及大二层技术 TRILL 和 VXLAN。

TRILL 技术简而言之就是通过在二层报文前插入额外的帧头，并且采用路由计算的方式控制整网数据的转发，不仅可以在冗余链路下防止广播风暴，而且可以做 ECMP (Equal-Cost Multi-Path routing)。这样可以将二层网络的规模扩展到整张网络，而不会受核心交换机数量的限制。但由于 TRILL 技术在控制平面都引入了路由协议 IS-IS (Intermediate System to Intermediate System) 来进行网络拓扑的计算和同步，增加了网络的复杂度，另外对原始报文的封装/解封装也降低了整体的转发效率，并且 TRILL 协议的处理通常需要新的芯片才能支持，所以必须整体更换原来的设备，投资成本也是不小的负担。这也是 TRILL 技术没有得到广泛运用的原因，如果读者想要详细了解 TRILL 技术，可以参考《TRILL 路由破坏 移花接木》，请访问 <http://forum.huawei.com/enterprise/forum.php?mod=viewthread&tid=148461>。

VXLAN 技术作为 Overlay 网络技术的典型代表，将在下一章节做重点介绍，现在就将我们将目光聚焦于传统的三层路由协议。三层作为网络架构的控制平面，负责将路由信息分发至组网中的所有交换机，但众所周知，三层路由协议的选择可以有很多选择，最好的做法就是支持主流的三种开放标准协议中的任何一种：OSPF (Open Shortest Path First)，IS-IS 或者 BGP 协议。本质上每种路由协议都可以在网络中通告路由前缀，但每种协议在支持的组网规模与实现功能上都有所不同。

OSPF 和 IS-IS 都是用使用洪泛技术来发送更新报文以及其他路由信息。创建区域可以帮助限制洪泛的数量，但是这样一来也就开始失去 SPF 路由协议的好处。另一方面，边界网关协议 (BGP) 通过按组创建，支持大量的前缀和对等体。BGP 从多方面保证了网络的安全性、灵活性、稳定性、可靠性和高效性，互联网和大多数运营商都选择运行 BGP 协议来作为控制层面的路由协议。

1.3.1 BGP 协议基础

为方便管理规模不断扩大的网络，网络被分成了不同的自治系统。外部网关协议 EGP (Exterior Gateway Protocol) 被用于实现在 AS 之间动态交换路由信息，但是 EGP 设计得比较简单，只发布网络可达的路由信息，而不对路由信息进行优选，同时也没有考虑环路避免等问题，很快就无法满足网络管理的要求。

BGP 是为取代最初的 EGP 而设计的另一种外部网关协议。不同于最初的 EGP，BGP 能够进行路由优选、避免路由环路、更高效率的传递路由和维护大量的路由信息。虽然 BGP 用于在 AS 之间传递路由信息，但并不是所有 AS 之间传递路由信息都需要运行 BGP。比如在数据中心上行的连入 Internet 的出口上，为了避免 Internet 海量路由对数据中心内部网络的影响，设备采用静态路由代替 BGP 与外部网络通信。

BGP 作为事实上的 Internet 外部路由协议标准，被广泛应用于 ISP (Internet Service Provider) 之间。

BGP 协议具有如下特点：

- BGP 是一种外部网关协议 (EGP)，与 OSPF、RIP 等内部网关协议 (IGP) 不同，其着眼点不在于发现和计算路由，而在于在 AS 之间选择最佳路由和控制路由的传播。
- BGP 使用 TCP 作为其传输层协议，提高了协议的可靠性。

- BGP 进行域间的路由选择，对协议的稳定性要求非常高。因此用 TCP 协议的高可靠性来保证 BGP 协议的稳定性。
- BGP 的对等体之间必须在逻辑上连通，并进行 TCP 连接。目的端口号为 179，本地端口号任意。
- 路由更新时，BGP 只发送更新的路由，大大减少了 BGP 传播路由所占用的带宽，适用于在 Internet 上传播大量的路由信息。
- BGP 是一种距离矢量（Distance-Vector）路由协议，BGP 从设计上避免了环路的发生。
 - AS 之间：BGP 通过携带 AS 路径信息来标记途经的 AS，带有本地 AS 号的路由将被丢弃，从而避免了域间产生环路。
 - AS 内部：BGP 在 AS 内学到的路由不再通告给 AS 内的 BGP 邻居，避免了 AS 内产生环路。
- BGP 提供了丰富的路由策略，能够对路由实现灵活的过滤和选择。
- BGP 提供了防止路由振荡的机制，有效提高了 Internet 网络的稳定性。
- BGP 易于扩展，能够适应网络新的发展。

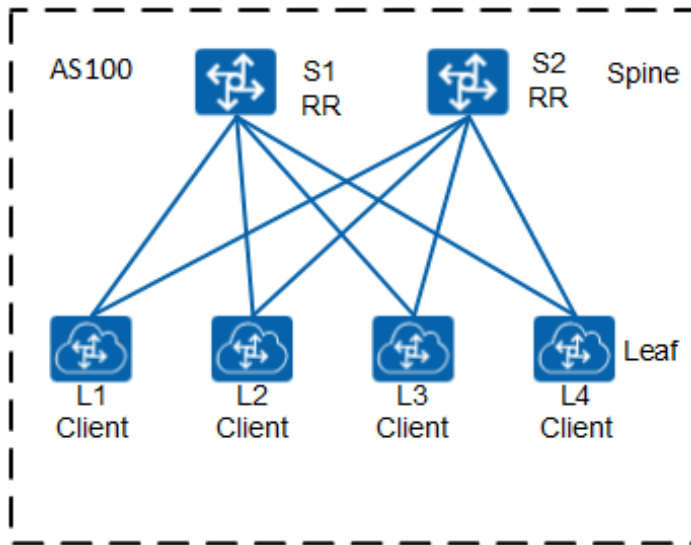
1.3.2 BGP 网络设计

当在规划部署 BGP 路由协议时，摆在我们面前的首要问题便是选择使用 IBGP 还是 EBGp 方式，虽然两者之间的差异可能看起来很小，但就是这些细微的差异可能导致数据中心在部署时的重大变化。IBGP 与 EBGp 两者之间的最大区别在于它们使用自治系统的方式，现在我们将通过比较 IBGP 与 EBGp 的差异来看每台交换机是如何分配路由前缀以及通告路由的。

IBGP

在 IBGP（Internal/Interior BGP）中，Spine 和 Leaf 的所有交换机位于单个 AS 之下，如图 1-38 所示。在 BGP 中，我们需要在 IBGP 对等体之间建立全连接（Full-mesh）关系来保证 IBGP 对等体之间的连通性。为什么 IBGP 要强调全连接概念呢，那是由于 IBGP 的防环机制导致的。IBGP 强制规定 ibgp speaker 不允许将从一个 IBGP 邻居学习到的前缀传递给其它 IBGP 邻居，因此 IBGP 要求逻辑全连接。但假设在一个 AS 内部有 n 台路由器，那么应该建立的 IBGP 连接数就为 $n(n-1)/2$ 。当 IBGP 对等体数目很多时，对网络资源和 CPU 资源的消耗都很大，在 IBGP 对等体间使用 BGP 联盟或者路由反射器都可以解决以上问题。

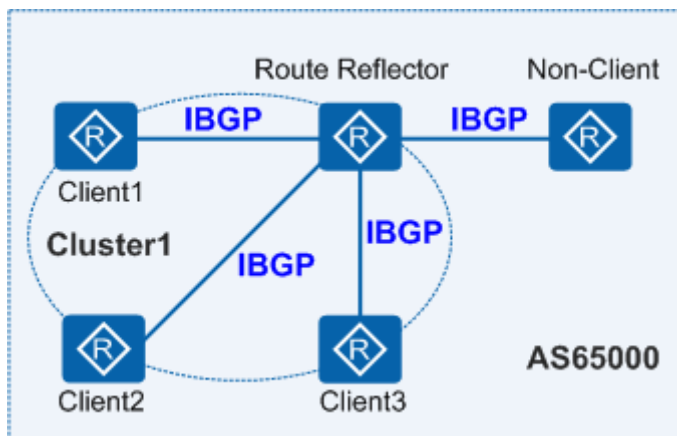
图1-38 IBGP 图示



联盟将一个 AS 划分为若干个子 AS。每个子 AS 内部建立 IBGP 全连接关系，子 AS 之间建立联盟 EBGP 连接关系，但联盟外部 AS 仍认为联盟是一个 AS。配置联盟后，原 AS 号将作为每个路由器的联盟 ID。这样有两个好处：一是可以保留原有的 IBGP 属性，包括 Local Preference 属性、MED 属性和 NEXT_HOP 属性等；二是联盟相关的属性在传出联盟时会自动被删除，即管理员无需在联盟的出口处配置过滤子 AS 号信息的操作。

路由反射器是 IBGP 路由器，它将重新向其他 IBGP 路由器通告路由。如图 1-39 所示，在一个 AS 内部关于路由反射器有以下几种角色：

图1-39 路由反射器示意图



- 路由反射器 RR（Route Reflector）：允许把从 IBGP 对等体学到的路由反射到其他 IBGP 对等体的 BGP 设备，类似 OSPF 网络中的 DR。
- 客户机（Client）：与 RR 形成反射邻居关系的 IBGP 设备。在 AS 内部客户机只需要与 RR 直连。

- 非客户机（Non-Client）：既不是 RR 也不是客户机的 IBGP 设备。在 AS 内部非客户机与 RR 之间，以及所有的非客户机之间仍然必须建立全连接关系。
- 始发者（Originator）：在 AS 内部始发路由的设备。Originator_ID 属性用于防止集群内产生路由环路。
- 集群（Cluster）：路由反射器及其客户机的集合。Cluster_List 属性用于防止集群间产生路由环路。

这可以通过创建 IBGP 路由器集群，并将其与反射器连接起来。同一集群内的客户机只需要与该集群的 RR 直接交换路由信息，因此客户机只需要与 RR 之间建立 IBGP 连接，不需要与其他客户机建立 IBGP 连接，从而减少了 IBGP 连接数量。但问题随之而来，反射器并不会发送每一条路线，它只会选择发送最优的路径给它的对等体。当你在 Spine 层面有多台交换机，且在 Spine 与 Leaf 之间存在多条链路时，链路冗余、利用率低的问题就出现了。为了解决这个问题，我们可以在 BGP 路由反射器上启用 BGP 负载分担功能，这样就可以向 Leaf 交换机通告存在四条等长的路由，可以通过等价多路径来分发流量。

我们从配置、设备连接和应用方面对 BGP 联盟和路由反射器进行了比较，如表 1-12 所示：

表1-12 路由反射器和 BGP 联盟比较

路由反射器	BGP 联盟
不需要更改现有的网络拓扑，兼容性好。	需要改变逻辑拓扑。
配置方便，只需要对作为反射器的设备进行配置，客户机并不需要知道自己是客户机。	所有设备需要重新进行配置。
集群与集群之间仍然需要全连接。	联盟的子 AS 之间是特殊的 EBGP 连接，不需要全连接。

EBGP

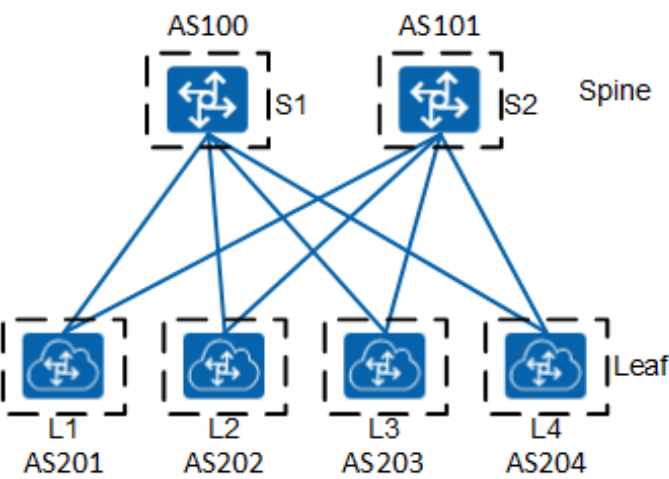
在 EBGP（External/Exterior BGP）中，Spine 和 Leaf 中的每个交换机都有自己的 AS，如图 1-40 所示。与 IBGP 的路由环路的避免措施不一样，EBGP 没有水平分割的概念，EBGP 对路由环路的避免是通过 AS_PATH 属性来实现的。AS_Path 属性按矢量顺序记录了某条路由从本地到目的地址所要经过的所有 AS 编号。在接收路由时，设备如果发现 AS_Path 列表中有本 AS 号，则不接收该路由，从而避免了 AS 间的路由环路。

由前面章节的讨论可知，IBGP 不需要 IBGP 邻居之间有物理连接，只需要逻辑连接即可，而 EBGP 在一般情况下都要求 EBGP 邻居之间存在物理连接。

唯一的问题是使用 IP Fabric 网络的 AS 数量，每台交换机都有自己的 BGP 自治系统号，BGP 的私有范围为 64512~65535，其中有 1023 个 BGP 自治系统号。如果您的 IP Fabric 网络大于 1023 台交换机，则需要考虑进入公共的 BGP 自治系统号码范围（不建议在数据中心内部使用），或移动到私有的四字节 AS 号规划。CE 系列交换机支持 4

字节私有 AS 号，4 字节私有 AS 号范围是 4200000000~4294967295（或者 64086.59904~65535.65535）。

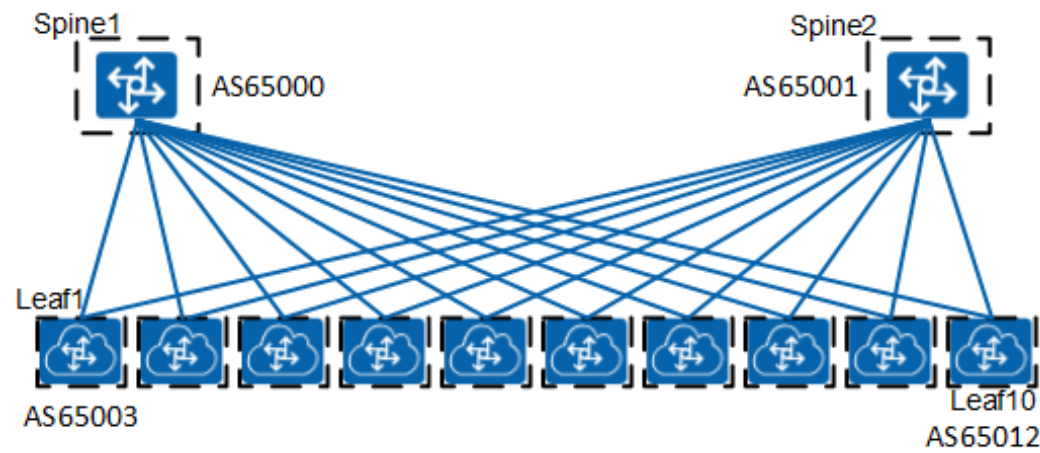
图1-40 EBGP 图示



BGP 在数据中心网络架构中的应用

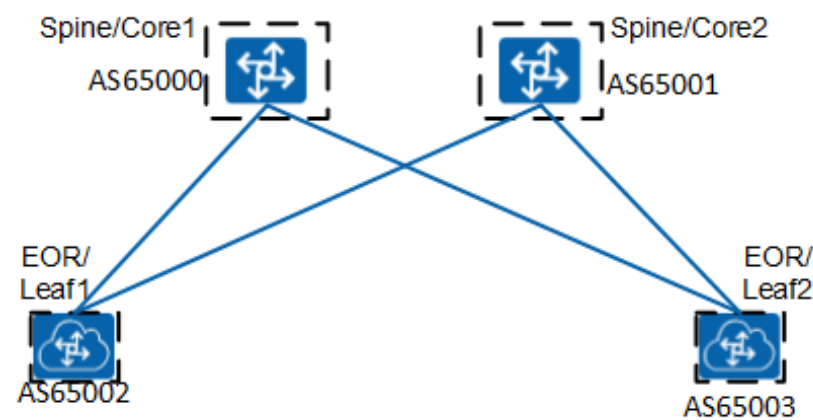
下面就让我们将 BGP 路由协议与具体的网络架构场景联系起来。在场景 DC1 中有五排机架，都采用 Spine+Leaf 网络架构。我们推荐使用 EBGP 建立 VXLAN 的 Underlay 网络，每行都使用 EBGP 的情况下设计将如图 1-41 所示，在 Spine 和 Leaf 层的每台交换机都有自己的 AS 号码。

图1-41 DC1 场景中运行 EBGP 图示



对于场景 DC2，我们采用的是 EOR 交换机部署，EBGP 的设计如图 1-42 所示。

图1-42 DC2 场景中运行 EBGP 图示



同样，如 DC1 场景图示所展现的一样，您可以将 AS65000 和 65001 用于第一行，然后为每个设备 AS 号递增加 1。如果您选择 IBGP 实施方法（通常在 DC 内用于 VXLAN overlay 路由的交换），设计将会变得非常简单，因为您只需将所有设备分配给同一 AS 号，覆盖整个 DC1，你可以对 DC2 采用相同的方法，如图 1-43 和图 1-44 所示。

图1-43 DC1 场景中运行 IBGP 图示

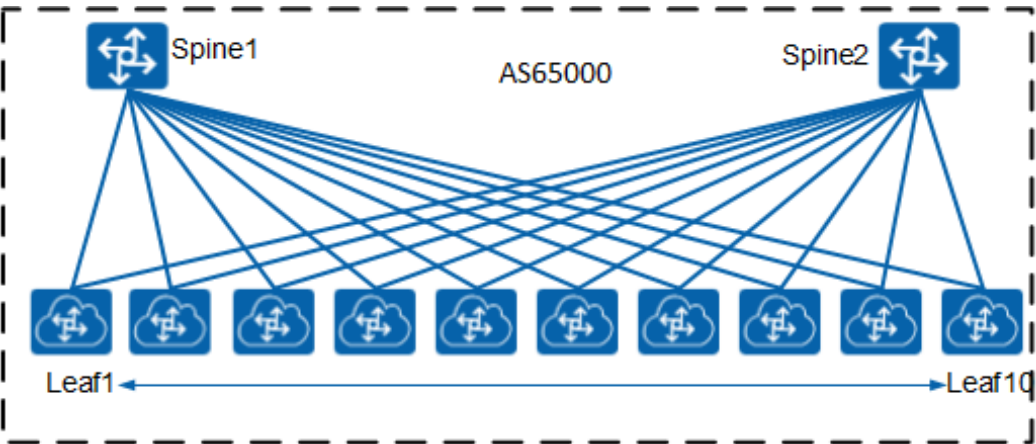
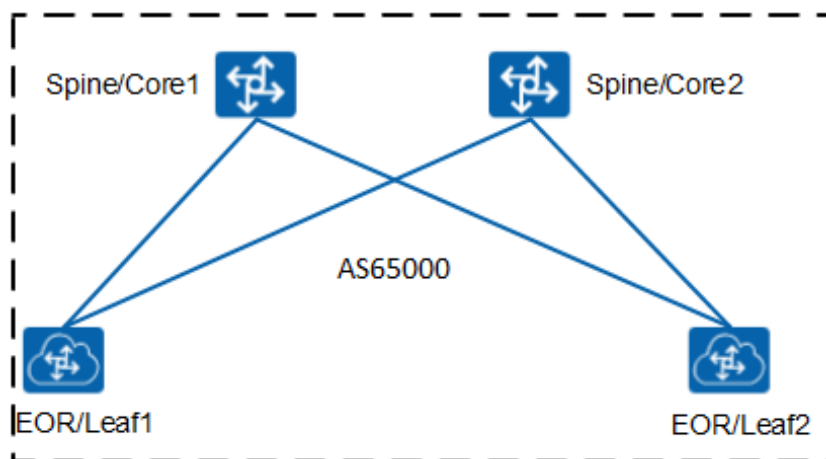


图1-44 DC2 场景中运行 IBGP 图示



1.4 本章小结

本章节介绍了 IP Fabric 网络的背景起源及三层转发路由协议的选择等内容。基于 Spine+Leaf 扁平化的网络架构，通过运行 BGP 路由协议，构建三层互通的 Underlay 网络，不仅运行稳定，更有着良好的扩展能力。

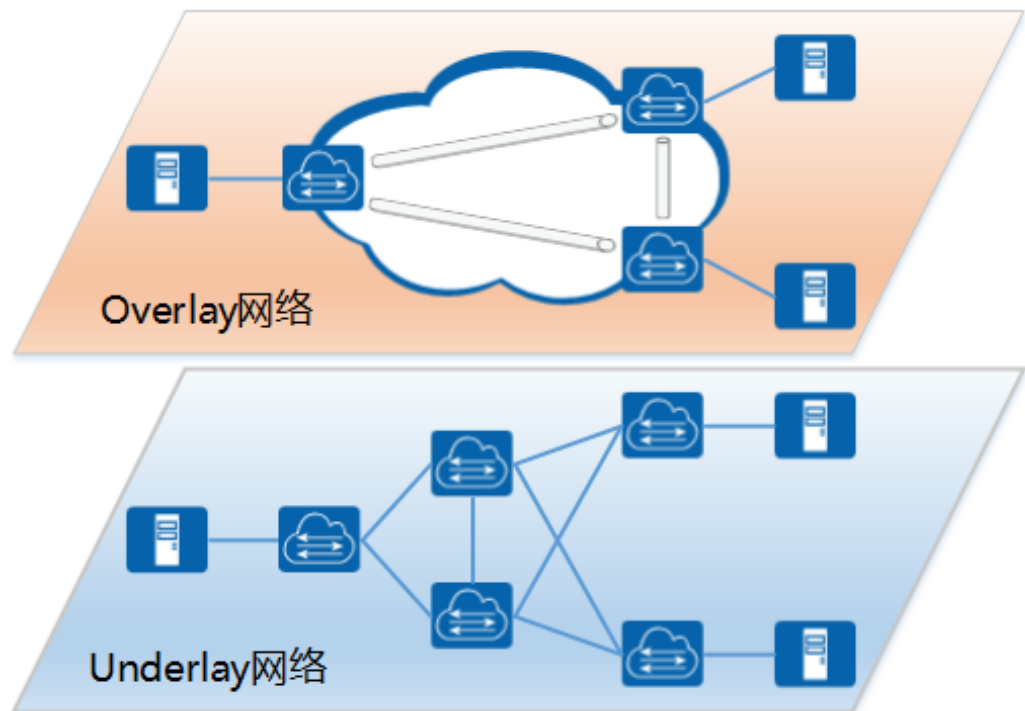
那基于 Underlay 的 IP 网络基础上构建的 Overlay 大二层网络又是怎么回事呢？下一章节会介绍什么是 Overlay 网络，为什么要有 Overlay 网络以及 CloudEngine 系列交换机实现 Overlay 的技术---VXLAN。

第七章 Overlay 网络

1.1 Overlay 介绍

如图 1-45 所示，Overlay 网络是将已有的物理网络（Underlay 网络）作为基础，在其上建立叠加的逻辑网络，实现网络资源的虚拟化。

图1-45 Overlay 网络概念图



Overlay 网络是建立在已有物理网络上的虚拟网络，具有独立的控制和转发平面，对于连接到 Overlay 的终端设备（例如服务器）来说，物理网络是透明的，从而可以实现承载网络和业务网络的分离。

为什么需要 Overlay 网络？

作为云计算核心技术之一的“服务器虚拟化”已经被数据中心普遍应用。随着企业业务的发展，虚拟机数量的快速增长和虚拟机迁移已成为一个常态性业务。由此也给传统网络带来了以下一些问题：

- 虚拟机规模受网络规格限制

在传统二层网络环境下，数据报文是通过查询 MAC 地址表进行二层转发，而网络设备 MAC 地址表的容量限制了虚拟机的数量。

- 网络隔离能力限制

当前主流的网络隔离技术是 VLAN，由于 IEEE 802.1Q 中定义的 VLAN ID 只有 12 比特，仅能表示 4096 个 VLAN，无法满足大二层网络中标识大量租户或租户群的需求。

- 虚拟机迁移范围受网络架构限制

为了保证虚拟机迁移过程中业务不中断，则需要保证虚拟机的 IP 地址、MAC 地址等参数保持不变，这就要求业务网络是一个二层网络，且要求网络本身具备多路径的冗余备份和可靠性。传统的 STP、设备虚拟化等技术只适用于中小规模的网络。

针对上述问题，为了满足云计算虚拟化的网络能力需求，逐步演化出了 Overlay 网络技术。

- 针对虚拟机规模受网络规格限制

虚拟机发出的数据包封装在 IP 数据包中，对网络只表现为封装后的网络参数。因此，极大降低了大二层网络对 MAC 地址规格的需求。

- 针对网络隔离能力限制

Overlay 技术扩展了隔离标识的位数（24 比特），极大扩展了隔离数量。

- 针对虚拟机迁移范围受网络架构限制

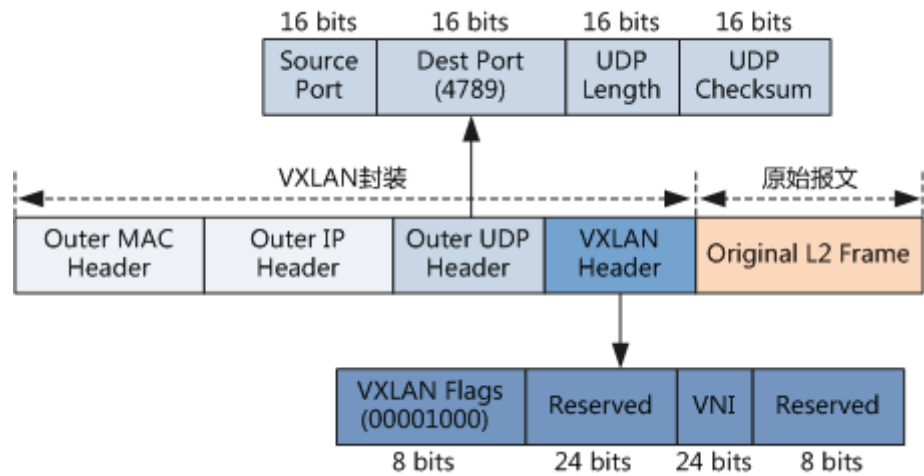
Overlay 将以太网报文封装在 IP 报文之上，通过路由在网络中传输。通过路由网络，虚拟机迁移不受网络架构限制。而且路由网络具备良好的扩展能力、故障自愈能力、负载均衡能力。

Overlay 技术有多种，例如 VXLAN、NVGRE、STT 等，其中 VXLAN 是目前获得最广泛支持的 Overlay 技术。

1.2 VXLAN

VXLAN（Virtual eXtensible Local Area Network，虚拟扩展局域网），是由 IETF 定义的 NVO3（Network Virtualization over Layer 3）标准技术之一，采用 MAC-in-UDP 的报文封装模式，如图 1-46 所示，原始报文在 VXLAN 接入点（被称为 VTEP）加上 VXLAN 帧头后再被封装在 UDP 报头中，并使用承载网络的 IP/MAC 地址作为外层头进行封装，承载网络只需要按照普通的二三层转发流程进行转发即可。

图1-46 VXLAN 报文格式



以 VXLAN 技术为基础的 Overlay 网络架构模型如所示。

图1-47 VXLAN 网络模型



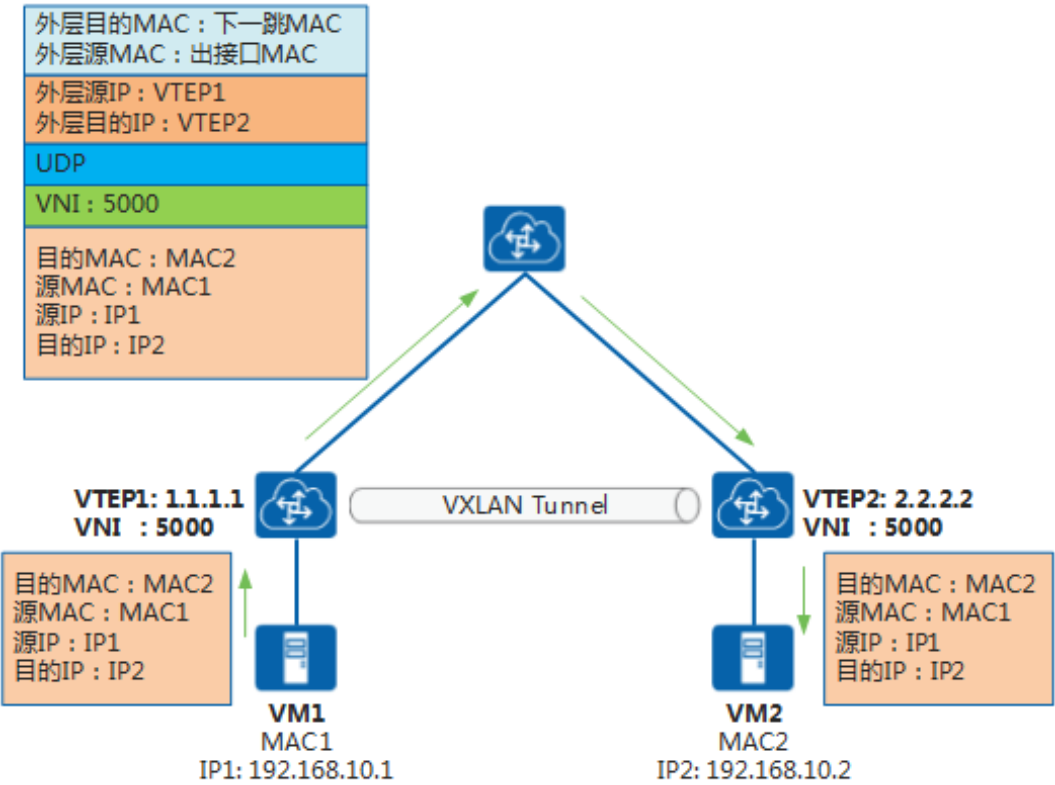
- **VTEP (VXLAN Tunnel Endpoints, VXLAN 隧道端点)**
VXLAN 网络的边缘设备，是 VXLAN 隧道的起点和终点，进行 VXLAN 报文的封装、解封装等处理。VTEP 既可以部署在网络设备上（网络接入交换机），也可以部署在 vSwitch 上（服务器上的虚拟交换机）。
- **VNI (VXLAN Network Identifier, VXLAN 网络标识符)**
VNI 是一种类似于 VLAN ID 的网络标识，用来标识 VXLAN 二层网络。一个 VNI 代表一个 VXLAN 段，不同 VXLAN 段的虚拟机不能直接二层相互通信。
- **VXLAN 隧道**

两个 VTEP 之间建立的逻辑隧道，用于传输 VXLAN 报文。业务报文在进入 VXLAN 隧道式进行 VXLAN、UDP、IP 头封装，然后通过三层转发透明地将报文转发给远端 VTEP，远端 VTEP 对报文进行解封装处理。

VXLAN 报文转发过程

下面以同网段的 VM 间相通简单介绍 VXLAN 网络中的报文转发过程。

图1-48 VXLAN 报文转发过程示意图



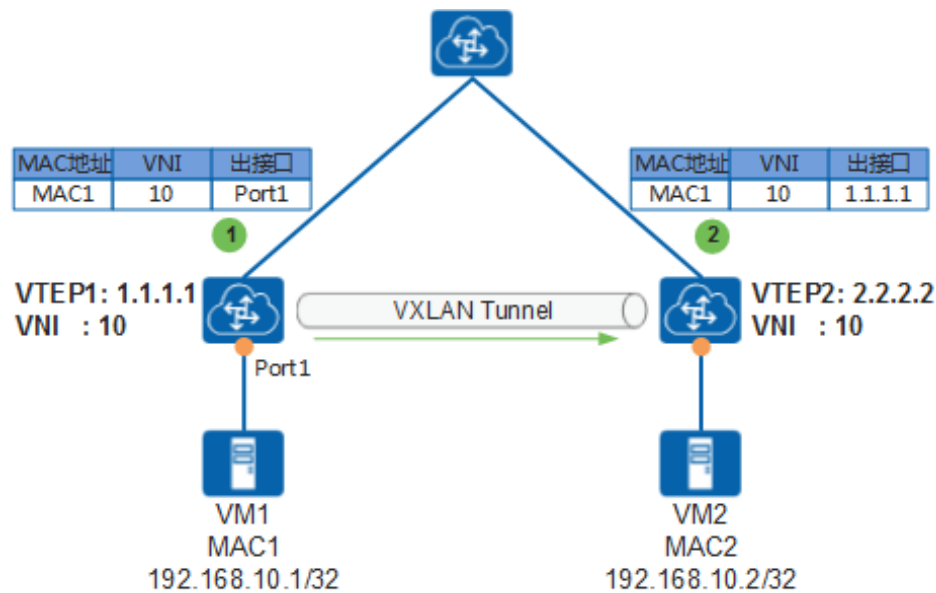
1. VM1 发送目的地址为 VM2 的报文。
2. VTEP1 收到该报文后进行 VXLAN 封装，封装的外层目的 IP 为 VTEP2。封装后的报文，根据外层 MAC 和 IP 信息，在 IP 网络中进行传输，直至到达对端 VTEP2。
3. VTEP2 收到报文后，对报文进行解封装，得到 VM1 发送的原始报文，然后将其转发至 VM2。

1.3 二层 MAC 学习及 BUM 报文转发

在 VXLAN 网络中，同子网虚拟机的互通是通过查找 MAC 表进行转发。如下图所示，VM1 给 VM2 发送报文时，经过 VTEP1 转发，VTEP1 上需要学习到 VM2 的 MAC 地址。

最初的 VXLAN 标准并没有定义控制平面，VTEP 之间无法传递学习到的主机 MAC 地址。但是 VXLAN 有着与传统以太网非常相似的 MAC 学习机制，当 VTEP 接收到 VXLAN 报文后，会记录源 VTEP 的 IP、虚拟机 MAC 和 VNI 到本地 MAC 表中，这样当 VTEP 接收到目的 MAC 为此虚拟机的 MAC 时，就可以进行 VXLAN 封装并转发。

图1-49 MAC 学习示意图



以 VTEP2 学习到 VM1 的 MAC 过程为例：

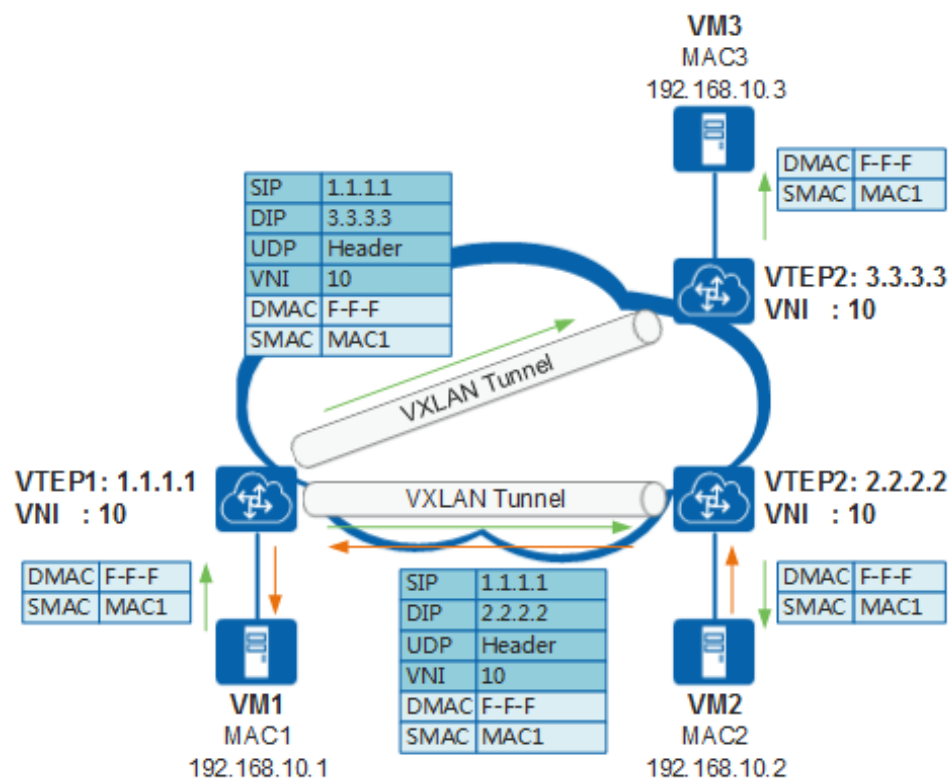
1. VM1 发送目的地址为 VM2 的报文。
2. VTEP1 接收到报文后，进行 VXLAN 封装，并将其转发至 VTEP2。同时，VTEP1 可以学习到 VM1 的 MAC 地址、VNI、入接口。
3. VTEP2 接收到报文后，对报文进行解封装。同时，VTEP2 可以学习到 VM1 的 MAC 地址、VNI、入接口（为 VTEP1）。

经过上述流程，VTEP1 和 VTEP2 可以学习到 VM1 的 MAC 地址。VTEP1 和 VTEP2 学习到 VM2 的 MAC 地址过程与之类似。

BUM 报文转发

前面描述的报文转发过程都是已知单播报文转发，如果 VTEP 收到一个未知地址的 BUM 报文（广播、组播、未知单播）如何处理呢。与传统以太网 BUM 报文转发类似，VTEP 会通过泛洪的方式转发流量。

图1-50 BUM 报文转发示意图



以上图中 VM1 想向 VM2 发送报文为例，因为 VM1 不知道 VM2 的 MAC 地址，所以会发送 ARP 广播报文请求 VM2 的 MAC 地址。

1. VM1 发送 ARP 广播请求，请求 VM2 的 MAC 地址。
2. VTEP1 收到 ARP 请求后，因为是广播报文，VTEP1 会在该 VNI 内查找所有的隧道列表，依据获取的隧道列表进行报文封装后，向所有隧道发送报文，从而将报文转发至同子网的 VTEP2 和 VTEP3。
3. VTEP2 和 VTEP3 接收到报文后，进行解封装，得到 VM1 发送的原始 ARP 报文，然后转发至 VM2 和 VM3。
4. VM2 和 VM3 接收到 ARP 请求后，比较报文中的目的 IP 地址是否为本机的 IP 地址。VM3 发现目的 IP 不是本机 IP，故将报文丢弃；VM2 发现目的 IP 是本机 IP，则对 ARP 请求做出应答。

由于此时 VM2 上已经学习到了 VM1 的 MAC 地址，所以 ARP 应答报文为已知单播报文，转发流程与前文描述的一致，此处不在赘述。

5. VM1 收到 VM2 的 ARP 应答后，就可以学习到 VM2 的 MAC 地址。后续的转发流程同已知单播转发流程一致。

1.4 VXLAN 网关部署

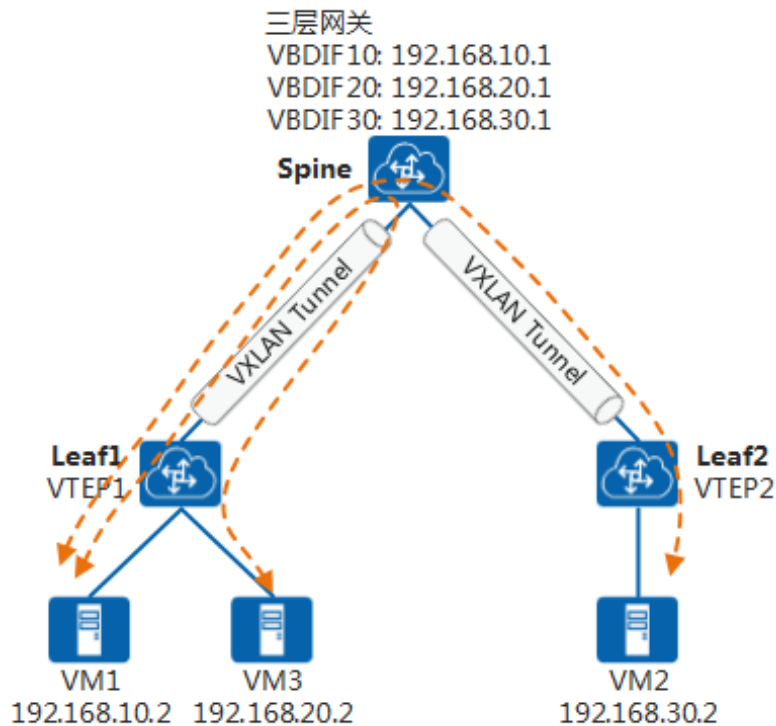
与不同 VLAN 需要通过三层网关互通一样，VXLAN 中不同 VNI 的互通也需要有三层网关。

在典型的“Spine-Leaf” VXLAN 组网结构下，根据三层网关的部署位置不同，VXLAN 三层网关可以分为集中式网关和分布式网关。

集中式网关部署

集中式网关是指将三层网关集中部署在 Spine 设备上，如下图所示，所有跨子网的流量都经过三层网关进行转发，实现流量的集中管理。

图1-51 集中式网关组网示意图

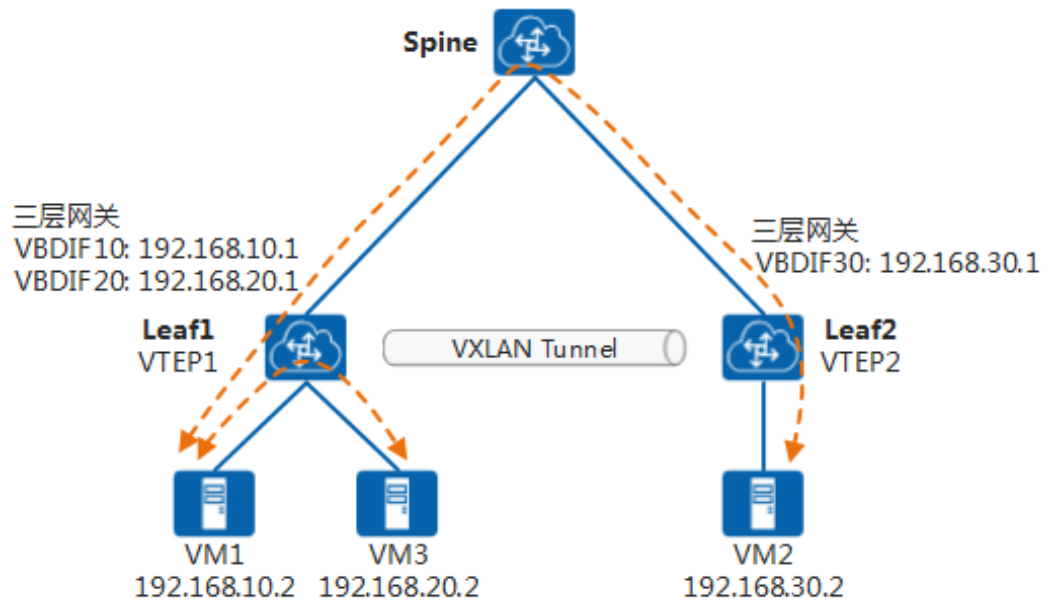


集中式网关部署方式可以对跨子网流量进行集中管理，网关的部署和管理比较简单，但是因为同 Leaf 下跨子网流量也需要经过 Spine 转发，所以流量转发路径不是最优。同时，所有通过三层转发的终端租户的表项都需要在 Spine 上生成。但是，Spine 的表项规格有限，当终端租户的数量越来越多时，容易成为网络瓶颈。

分布式网关部署

VXLAN 分布式网关是将 Leaf 节点作为 VXLAN 隧道端点 VTEP，每个 Leaf 节点都可作为 VXLAN 三层网关，Spine 节点不感知 VXLAN 隧道，只作为 VXLAN 报文的转发节点。

图1-52 分布式网关组网示意图



在 Leaf 上部署 VXLAN 三层网关，即可实现同 Leaf 下跨子网通信。此时，流量只需要在 Leaf 节点进行转发，不再需要经过 Spine 节点，从而节约了大量的带宽资源。同时，Leaf 节点只需要学习自身连接虚拟机的 ARP 表项，而不必像集中三层网关一样，需要学习所有虚拟机的 ARP 表项，解决了集中式三层网关带来的 ARP 表项瓶颈问题，网络规模扩展能力强。

对于分布式网关场景，因为需要在三层网关间传递主机路由才能保证虚拟机间互通，所以需要控制平面来进行路由的传递。下一章节就是描述 EVPN 作为控制平面技术在 VXLAN 网络里的应用。

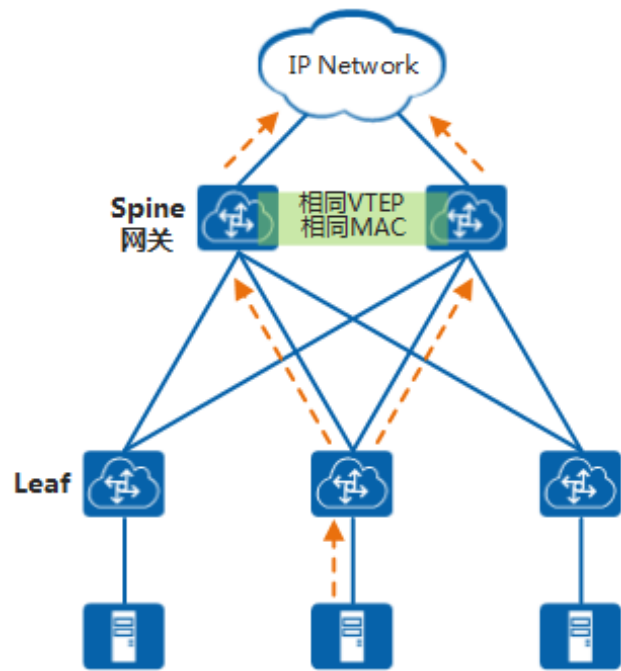
1.5 双活网关

在传统网络中，为了保证高可靠性，通常部署多个网关进行备份。与传统网络类似，VXLAN 网络也支持 Overlay 层面的双活网关。

集中式网关场景下双活网关

在典型的“Spine-Leaf”组网结构下，Leaf 作为二层网关，Spine 作为三层网关。多个 Spine 配置相同的 VTEP 地址、虚拟 MAC 地址，从而可以将多个 Spine 虚拟成一个 VXLAN 隧道端点。这样使得无论流量发到哪一个 Spine 设备，该设备都可以提供网关服务，将报文正确转发给下一跳设备。

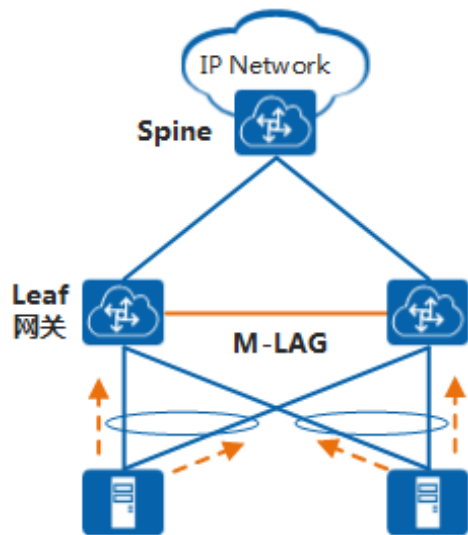
图1-53 集中式网关场景下双活网关组网示意图



分布式网关场景下双活网关

在分布式网关场景下，Spine 作为透传设备，Leaf 作为三层网关。通过在 Leaf 上部署 M-LAG，可以实现双活网关，即服务器可以双活接入到多个 Leaf。

图1-54 分布式网关场景下双活网关



1.6 本章小结

本章节介绍了 VXLAN 概念及报文转发流程等内容。VXLAN 可以基于已有的 IP 网络，通过三层网络构建出一个大二层网络。部署 VXLAN 功能的可以是物理交换机或服务器上的虚拟交换机（vSwitch），物理交换机作为 VTEP 的优势在于设备处理性能比较高，可以支持非虚拟化的物理服务器之间的互通，但是需要物理交换机支持 VXLAN 功能；vSwitch 作为 VTEP 的优势在于对网络要求低，不需要网络设备支持 VXLAN 功能，但是 vSwitch 处理性能不如物理交换机。

CloudEngine 系列交换机中 CE12800、CE8800、CE7800、CE6800（除 CE6810LI、CE6810EI、CE6850EI 外）交换机支持 VXLAN 功能，在典型的“Spine-Leaf”结构下，Spine 建议使用 CE12800、CE8800 或 CE7800 交换机，Leaf 建议使用 CE6800 交换机。

从前文的介绍可以知道 VXLAN 通过泛洪方式进行 MAC 学习，这是因为早期的 VXLAN 没有控制平面。下一章节会介绍作为 VXLAN 控制平面的技术——EVPN，以及如何通过 EVPN 实现 VXLAN 隧道自动建立和 MAC 路由学习。

第八章 BGP EVPN

1.1 EVPN 介绍

最初的 VXLAN 方案（RFC7348）中没有定义控制平面，是手工配置 VXLAN 隧道，然后通过流量泛洪的方式进行主机地址的学习。这种方式实现上较为简单，但是会导致网络中存在很多泛洪流量、网络扩展起来困难。

为了解决上述问题，VXLAN 引入了 EVPN（Ethernet VPN）作为 VXLAN 的控制平面。EVPN 参考了 BGP/MPLS IP VPN 的机制，通过扩展 BGP 协议新定义了几种 BGP EVPN 路由，通过在网络中发布路由来实现 VTEP 的自动发现、主机地址学习。

采用 EVPN 作为控制平面具有以下一些优势：

- 可实现 VTEP 自动发现、VXLAN 隧道自动建立，从而降低网络部署、扩展的难度。
- EVPN 可以同时发布二层 MAC 和三层路由信息。
- 可以减少网络中泛洪流量。

1.2 BGP EVPN 路由类型

传统的 BGP-4 使用 Update 报文在对等体之间交换路由信息。一条 Update 报文可以通告一类具有相同路径属性的可达路由，这些路由放在 NLRI（Network Layer Reachable Information，网络层可达信息）字段中。

因为 BGP-4 只能管理 IPv4 单播路由信息，为了提供对多种网络层协议的支持（例如 IPv6、组播），发展出了 MP-BGP（MultiProtocol BGP）。MP-BGP 在 BGP-4 基础上对 NLRI 作了新扩展。玄机就在于新扩展的 NLRI 上，扩展之后的 NLRI 增加了地址族的描述，可以用来区分不同的网络层协议，例如 IPv6 单播地址族、VPN 实例地址族等。

类似的，EVPN 在 L2VPN 地址族下定义了新的子地址族——EVPN 地址族，并新增了一种 NLRI，即 EVPN NLRI。EVPN NLRI 定义了以下几种 BGP EVPN 路由类型，通过在 EVPN 对等体之间发布这些路由，就可以实现 VXLAN 隧道的自动建立、主机地址的学习。

- Type2 路由——MAC/IP 路由：用来通告主机 MAC 地址、主机 ARP 和主机路由信息。
- Type3 路由——Inclusive Multicast 路由：用于 VTEP 的自动发现和 VXLAN 隧道的动态建立。

- Type5 路由——IP 前缀路由：用于通告引入的外部路由，也可以通告主机路由信息。

EVPN 路由在发布时，会携带 RD（Route Distinguisher，路由标识符）和 VPN Target（也称为 Route Target）。RD 用来区分不同的 VXLAN EVPN 路由。VPN Target 是一种 BGP 扩展团体属性，用于控制 EVPN 路由的发布与接收。也就是说，VPN Target 定义了本端的 EVPN 路由可以被哪些对端所接收，以及本端是否接收对端发来的 EVPN 路由。

VPN Target 属性分为两类：

- Export Target：本端发送 EVPN 路由时，将消息中携带的 VPN Target 属性设置为 Export Target。
- Import Target：本端在接收到对端的 EVPN 路由时，将消息中携带的 Export Target 与本端的 Import Target 进行比较，只有两者相等时才接收该路由，否则丢弃该路由。

1.3 Type2 类型路由

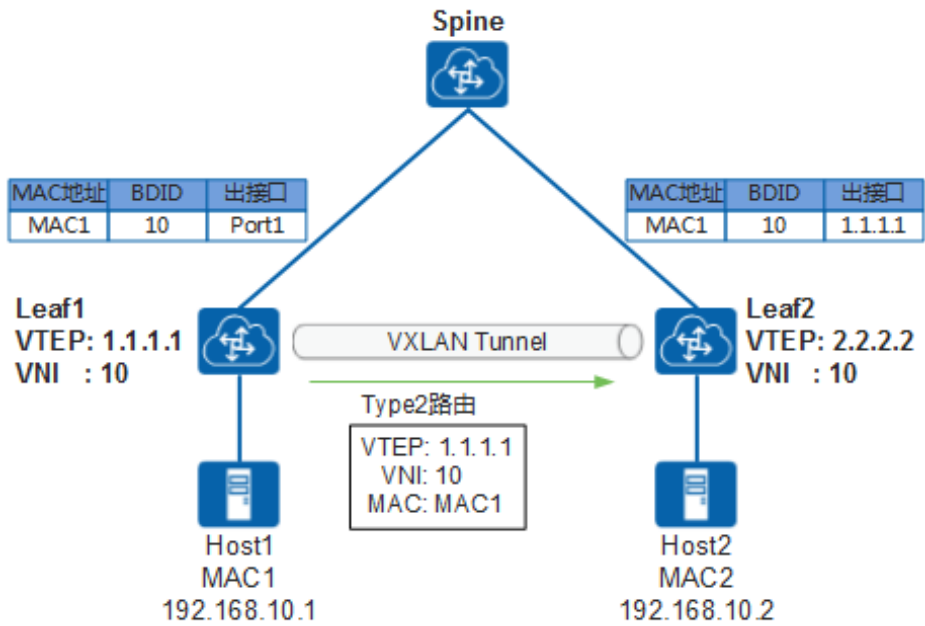
Type2 类型路由中 NLRI 格式如下：

图1-55 Type2 类型路由中 NLRI 格式

Route Distinguisher	路由RD值，由EVPN实例设置
Ethernet Segment Identifier	与对端连接的标识ESI
Ethernet Tag ID	VLAN ID
MAC Address Length	主机MAC地址的长度
MAC Address	主机MAC地址
IP Address Length	主机IP地址的掩码长度
IP Address	主机IP地址
MPLS Label1	二层VNI
MPLS Label2	三层VNI

由图 1-55 可以看出，Type2 类型路由中携带有主机 MAC、主机 IP 信息，因此 Type2 类型路由可以用于发布主机 MAC，还可以发布主机 IP 地址。

图1-56 Type2 类型路由发布 MAC 地址示意图



如图 1-56 所示，Leaf1 接收到 Host1 发送的报文后，会学习到 Host1 的 MAC 地址。Leaf1 学习到 Host1 的 MAC 后，会生成 Type2 类型的路由发送给 Leaf2，该路由会携带 EVPN 实例的 ERT、Host1 的 MAC 地址、Leaf1 的 VTEP IP 等信息。

Leaf2 收到 Leaf1 发送来的路由后，根据路由中的 ERT 是否与本端 EVPN 实例的 IRT 相同来决定是否接收该路由。如果相同，则接收该路由，Leaf2 可以学习到 Host1 的 MAC 地址；如果不同，则丢弃路由。

1.4 Type3 类型路由

Type3 类型路由中 NLRI 格式如下：

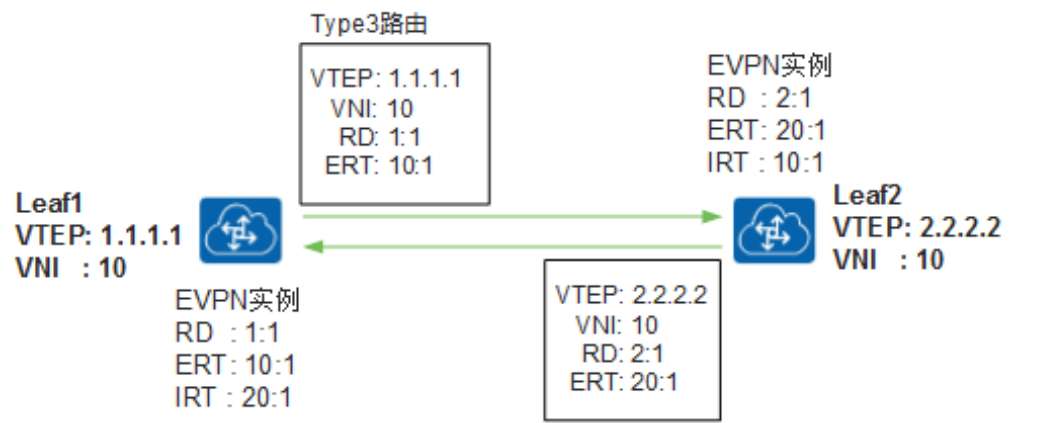
图1-57 Type3 类型路由中 NLRI 格式

前缀	
Route Distinguisher	RD值，由EVPN实例设置
Ethernet Tag ID	VLAN ID，此处为全0
IP Address Length	本端VTEP IP地址掩码长度
Originating Router's IP Address	本端VTEP IP地址

PMSI属性	
Flags	标志位，VXLAN中无实际意义
Tunnel Type	隧道类型，VXLAN为6
MPLS Label	二层VNI
Tunnel Identifier	隧道信息

Type3 类型路由中主要携带有 VTEP IP 信息，主要用于 VTEP 的自动发现和 VXLAN 隧道的动态建立。

图1-58 Type3 类型路由建立 VXLAN 隧道示意图



如图 1-58 所示，在 Leaf1 和 Leaf2 之间建立 BGP EVPN 对等体后，Leaf1 会生成 Type3 类型的路由发送给 Leaf2。该路由中会携带本端 VTEP IP 地址、VNI，EVPN 实例的 ERT 等信息。

Leaf2 收到 Leaf1 发送来的路由后，根据路由中的 ERT 是否与本端 EVPN 实例的 IRT 相同来决定是否接收该路由。如果相同，则接收该路由，建立一条到对端的 VXLAN 隧道。同时，如果对端 VNI 与本端相同，则创建一个头端复制表，用于后续广播、组播、未知单播报文的转发。

1.5 Type5 类型路由

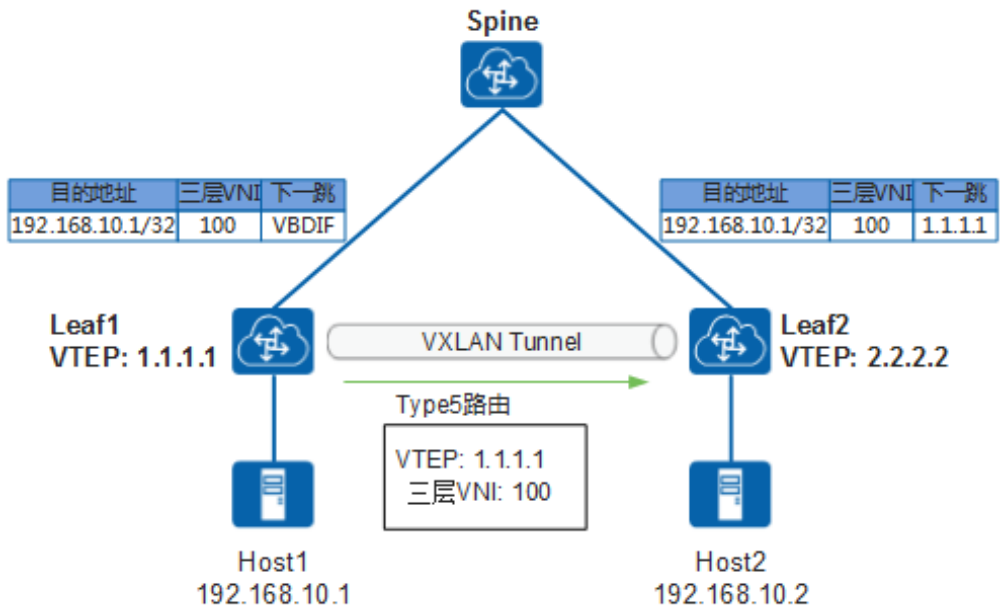
Type5 类型路由中 NLRI 格式如下：

图1-59 Type5 类型路由中 NLRI 格式

Route Distinguisher	路由RD值，由EVPN实例设置
Ethernet Segment Identifier	与对端链接的标识ESI
Ethernet Tag ID	VLAN ID
IP Prefix Length	IP前缀掩码长度
IP Prefix	IP前缀
GW IP Address	默认网关地址
MPLS Label	三层VNI

Type5 类型路由携带路由信息，主要用于发布路由。与 Type2 类型路由不同的是，Type5 类型路由既可以发布 32 位主机路由，也可以发布网段路由。

图1-60 Type5 类型路由发布路由示意图

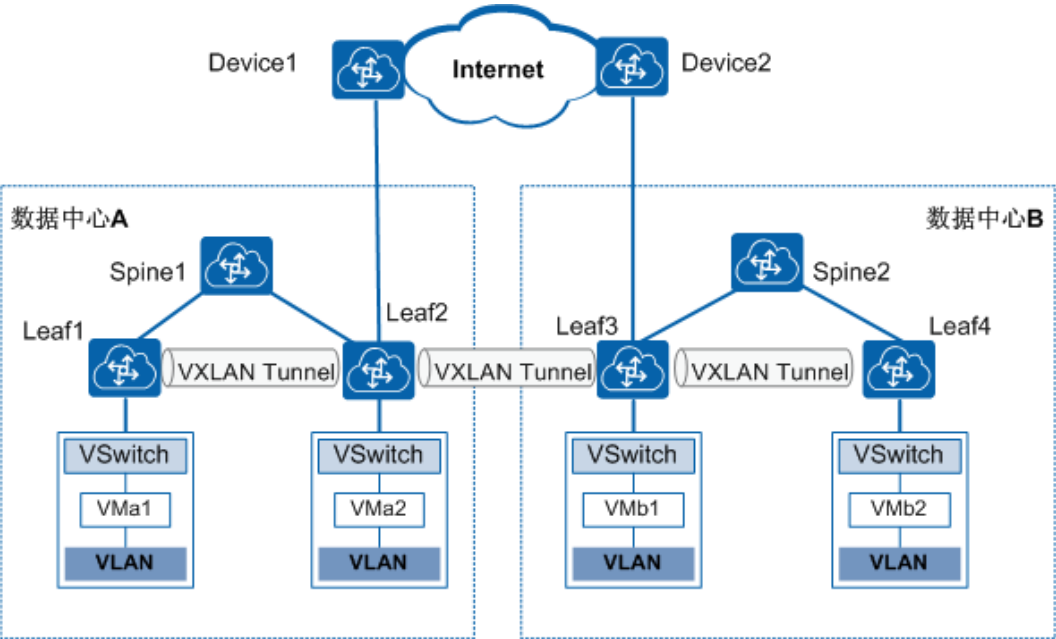


Type5 类型路由可以将本地其他协议的私网路由（例如静态路由、直连路由、其他路由协议路由）发布到其他 EVPN 网络中，在远端生成相应的主机/网段路由。所以 Type5 类型路由可以实现 VXLAN 网络主机访问外部非 VXLAN 网络。

1.6 BGP EVPN 实现 DCI 互联

通过 BGP EVPN 在两个数据中心内部各建立一段 VXLAN 隧道，数据中心之间再建立一段 VXLAN 隧道，可以实现数据中心互联。如图图 1-61 所示，分别在数据中心 A、数据中心 B 内配置 BGP EVPN 协议创建分布式网关 VXLAN 隧道，实现各数据中心内部 VM 之间的通信。Leaf2 和 Leaf3 是数据中心内连接骨干网的边缘设备，通过在 Leaf2 和 Leaf3 上配置 BGP EVPN 协议创建 VXLAN 隧道，将从一侧数据中心收到的 VXLAN 报文先解封装、然后再重新封装后发送到另一侧数据中心，实现对跨数据中心的报文端到端的 VXLAN 报文承载，保证跨数据中心 VM 之间的通信。

图1-61 过 BGP EVPN 实现数据中心互联组网图



1.7 本章小结

EVPN 是基于 BGP 协议的技术，需要部署在网络交换机上。这意味着网络交换机需要作为 VTEP 节点，进行 VXLAN 封装。服务器通过接口或 VLAN 接入网络交换机。这些接口或 VLAN 会映射到对应的广播域 BD，同时 BD 也会绑定一个 EVPN 实例，通过 EVPN 实例间路由的传递实现 VXLAN 隧道的建立、MAC 学习。

总结

本书讲述了数据中心网络设计中的重点概念。从 CE 系列交换机介绍到布线方案，从收敛比的设计到 Fabric 网络技术的应用，从 Overlay 网络的起源到 VXLAN 技术讲解。希望您阅读后，能够清楚的理解 CE 系列交换机的优势及能够帮助您解决的问题，了解如何使用 CE 系列交换机构建自己的数据中心。

数据中心网络技术的发展日新月异，文中提到的设计思路及技术实现也在不断的变化更新。如果您对这些技术的发展感兴趣，建议您关注华为企业用户技术支持网站 (<http://support.huawei.com/enterprise/zh/index.html>)，了解交换机最新版本功能的支持情况。

当前，随着 SDN 网络架构的逐步成熟与商用，作为 Fabric 网络流量承载主体的交换机，更多的将焦点放在了设备的开放性、低时延、精细化运维等能力上。为了帮助客户适应各种云业务的快速变化，华为公司也创新性地推出了云数据中心网络 SDN 解决方案，旨在为客户构筑简单、开放、弹性的云数据中心网络，加速企业数字化转型。如需了解更多”云数据中心网络 SDN 解决方案“的内容，请访问：
<http://support.huawei.com/onlinetoolsweb/NetSolution/DataCenterNetwork/zh/index.html>。

本书也不会止步于此，我们会着眼于最新的网络架构设计与技术应用，不断推出新的篇章，希望您能够持续关注。

相关资料

本书中涉及的相关资料如表 1-13 所示。

表1-13 相关资料

相关资料	网址
CE 系列交换机售前资料	http://e.huawei.com/cn/products/enterprise-networking/switches/data-center-switches
CE 系列交换机售后资料	http://support.huawei.com/enterprise/zh/index.html
CE 系列交换机全家福照片	http://e.huawei.com/cn/material/onLineView?materialid=21a1436f149c424db97519fc48d24800
数据中心网络技术红宝书	http://forum.huawei.com/enterprise/zh/thread-148461.html
CE12800 第“X”类接触	http://forum.huawei.com/enterprise/zh/thread-313577.html
CE 系列交换机转发性能评估工具	http://support.huawei.com/onlinetoolsweb/proforward_tool/cn/index.html

术语与缩略语

表1-14 术语与缩略语

术语与缩略语	解释
AS	Autonomous System，自治系统通常是指网络的一个组成部分，通常一个自治系统由一个组织控制，并运行一种路由协议。不同自治系统之间的路由通过域间协议完成。
AOC	Active Optical Cables，有源光缆是指光模块和光纤一体化的有源光线缆。
BD	Bridge Domain，VXLAN 网络中转发数据报文的二层广播域。
BGP	Border Gateway Protocol，边界网关协议是一种实现自治系统 AS 之间的路由可达，并选择最佳路由的距离矢量路由协议。IBGP（Internal/Interior BGP）是运行于同一 AS 内部的 BGP 协议，EBGP（External/Exterior BGP）是运行于不同 AS 之间的 BGP 协议。
DAC	Direct Attach Cable，直连铜缆是一种固定长度、两端有固定连接器的线缆组件。
ECMP	Equal-Cost Multi-Path routing，等价多路径路由实现了等价多路径负载均衡和链路备份的目的。
EGP	Exterior Gateway Protocol，外部网关协议 EGP 被用于实现在 AS 之间动态交换路由信息。
EOR	End of Row，行末交换机。EOR 交换机在每排机柜末端部署，供服务器统一接入网络。
EVPN	Ethernet Virtual Private Network，是一种用于二层网络互联的 VPN 技术。通过扩展 BGP，用于处在二层网络的不同站点之间的 MAC 地址学习和发布。
IGP	Interior Gateway Protocol，用于自治系统内部的一种路由协议。
IS-IS	Intermediate System to Intermediate System，中间系统到中间系统属于内部网关协议 IGP，用于自治系统内部。IS-IS 也是一种链路状态协议，使用最短路径优先 SPF（Shortest Path First）算法进行路由计算。
M-LAG	Multichassis Link Aggregation Group，跨设备链路聚合组，是一种实现跨设备链路聚合的机制。M-LAG 将一台设备与另外两台设备

术语与缩略语	解释
	进行跨设备链路聚合，从而把链路可靠性从单板级提高到了设备级，组成双活系统。
MMF	Multi Mode Fiber，多模光纤是指可传输多种模式光的光纤。
MOR	Middle of Row，列中模式，是对 EOR 的一种改进，也为服务器提供统一的网络接入机柜，但是 MOR 要求将网络机柜放在整排机柜的中部，在一定程度上缩短了服务器机柜到网络机柜的距离。
OSPF	Open Shortest Path First，开放式最短路径优先 OSPF 是 IETF 组织开发的一个基于链路状态的内部网关协议（Interior Gateway Protocol）。
SMF	Single Mode Fiber，单模光纤是指仅能传输一种模式光的光纤。
STP	Spanning Tree Protocol，用于局域网中消除环路的协议。运行该协议的设备通过彼此交互信息而发现网络中的环路，并适当对某些端口进行阻塞以消除环路。
SVF	Super Virtual Fabric，是一种纵向虚拟化技术，通过将一台低成本盒式设备作为远程接口板接入主设备，达到扩展端口密度和集中控制管理的目的，满足数据中心高密度接入和简化管理的需求。
TRILL	Transparent Interconnection of Lots of Links，是一种通过扩展 IS-IS 路由协议实现二层路由，把三层链路状态路由技术应用于二层网络的协议。
TOR	Top of Rack，架顶交换机。虽然从字面上看，Top of Rack 指的是“机柜顶部”，但实际 TOR 的核心在于将交换机部署在服务器机柜内，既可以部署在机柜顶部，也可以部署在机柜的中部（Middle of Rack）或底部（Bottom of Rack）。
VLAN	Virtual Local Area Network，即虚拟局域网，是将一个物理的 LAN 在逻辑上划分成多个广播域的通信技术。VLAN 内的主机间可以直接通信，而 VLAN 间不能直接互通，从而将广播报文限制在一个 VLAN 内。
VM	Virtual Machine，即虚拟机，与物理机一样是运行操作系统和应用程序的虚拟计算机。
VNI	VXLAN Network Identifier，VXLAN 网络标识，类似 VLAN ID，用于区分 VXLAN 段，不同 VXLAN 段的虚拟机不能直接二层相互通信。
VRRP	Virtual Router Redundancy Protocol，虚拟路由冗余协议通过把几台路由设备联合组成一台虚拟的路由设备，将虚拟路由设备的 IP 地址作为用户的默认网关实现与外部网络通信。当网关设备发生故障时，VRRP 机制能够选举新的网关设备承担数据流量，从而保障网络的可靠通信。
vSwitch	Virtualized Switch，虚拟交换机，通过软件方式实现物理交换机的

术语与缩略语	解释
	二层（或部分三层）网络功能。
VTEP	VXLAN Tunnel Endpoints，VXLAN 隧道端点，用于 VXLAN 报文的封装和解封装。一对 VTEP 地址就对应着一个 VXLAN 隧道。
VXLAN	Virtual eXtensible Local Area Network，虚拟扩展局域网，是一种网络虚拟化技术。