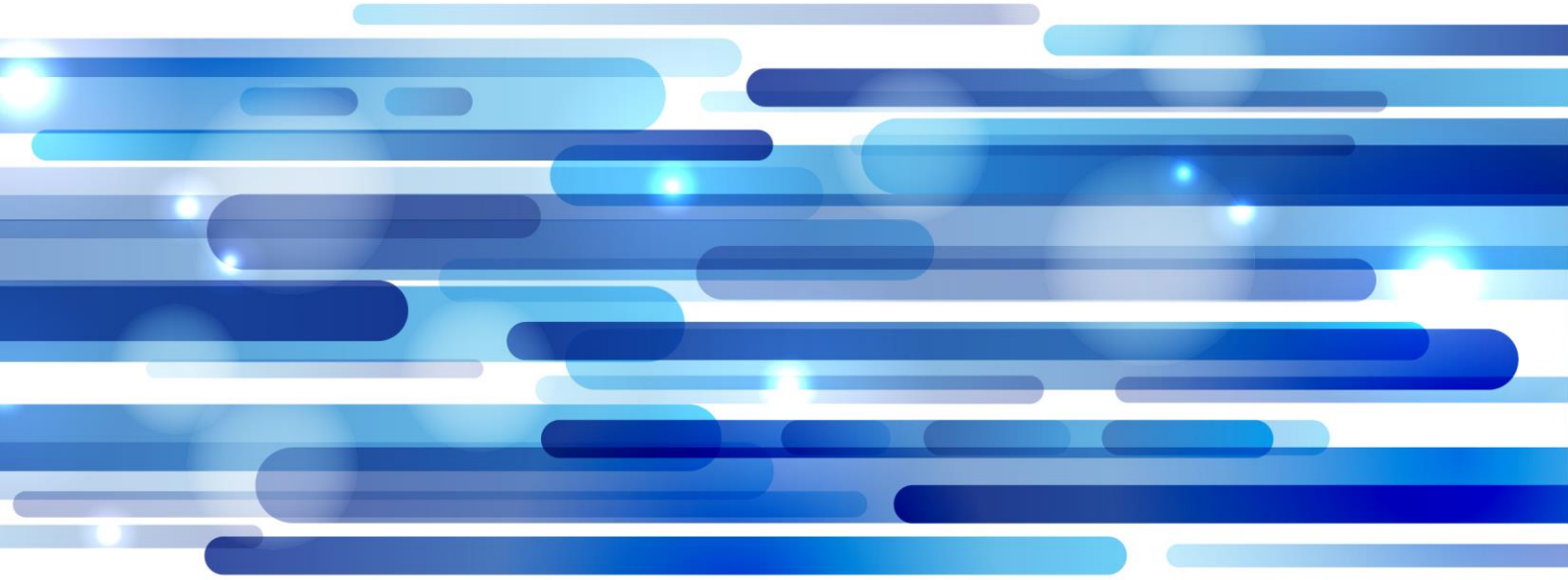
Huawei Data Center

Network Design Guide



Without any prior knowledge, this book can help you understand all the basics of data center networks. Starting with an overview of switches and cabling solutions, the book goes on to explain oversubscription ratio design and the application of fabric network technologies. It also clearly describes the origin of overlay networking and VXLAN technology. By the end of this book, you will be able to build your own data center using CloudEngine series switches.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://e.huawei.com

About This Book

The rapid development of technologies such as cloud computing, Big Data, and artificial intelligence creates higher requirements for data center networks that carry data traffic. Service requirements for data center networks include high throughput, high reliability, low latency, and adaptation to server virtualization. In order to meet the network requirements of services, increasing numbers of enterprises choose to construct their own data centers or rent public clouds to transmit their growing service traffic.

As the core of traffic aggregation and forwarding, data center switches play an important role in the entire data center network. This book describes how to build a data center network using Huawei CloudEngine series switches, abbreviated as CE series switches.

About the Author

The authors of this book are technical writers from Huawei's DCN R&D team and have worked in the technical document development field for more than seven years. They accumulate extensive experience during long-term R&D and customer-oriented delivery. Now they summarize the experience into this book to share it with readers and help them construct data center networks.

The authors from the DCN technical document development team are as follows:

Ma Tengteng, Zhu Hualian, Li Shuangshuang, Zong Yue, Zhou Tingting, and Gao Yangyang

Invited Experts

During development of this book, we also received strong support from other experts in Huawei's DCN R&D field. From their professional perspectives, they provided a large number of valuable suggestions on this book in terms of content, solution maturity, and professional information organization.

Our sincere gratitude goes to:

Product architecture design experts: Wang Jianbing, Zong Zhigang, Liu Shuming, and Xie Ying

Software test experts: Chen Ye, Yang Dehua, Tu Lin, and Zhang Yongle

Senior documentation experts: Cui Chun, Li Xuezhao, and Han Xiang

Intended Audience

This book is intended for network designers, network architects, network administrators, and maintenance engineers responsible for enterprise network design and deployment, and readers who want to understand basic principles of data center network design.

Based on Huawei CE series data center switches, this book describes basic design principles and some necessary background knowledge of data center networks. It also applies to design of most data center networks of the same type.

To better understand the contents of this book, we hope that you have a basic understanding of IP network protocols and network design before reading this book.

Overview

This book contains the following chapters:

- Chapter 1 describes the features, advantages, and network locations of Huawei CE series switches. Huawei provides various types of data center switches that meet a comprehensive variety of different network requirements.
- Chapter 2 describes how to deploy switches in racks and cabinets, providing information on how to connect servers to switches. This chapter provides information for choosing between TOR and EOR/MOR deployment modes.
- Chapter 3 describes the cabling design of a data center network. Using optimal cables in proper scenarios is a prerequisite for constructing a high-speed data center network.
- Chapter 4 describes how to set a proper oversubscription ratio. The oversubscription ratio must be considered during network design. The oversubscription ratio depends on the server type, bandwidth requirement, and plans for future network expansion. If you want to configure network devices to transmit traffic at the line rate or use the optimal links to forward east-west traffic and north-south traffic, you should read this chapter.
- Chapter 5 describes various fabric technologies provided by Huawei. From this chapter on, the focus of this book shifts from physical architecture design to logical architecture design. These fabric technologies support the implementation of device virtualization, simplify network management, and prevent Layer 2 loops.
- Chapter 6 describes the design principles of IP fabric and Layer 3 routes. This chapter answers questions including: Why is the spine-leaf architecture chosen for most data center networks? What is the basis for the selection of a Layer 3 routing protocol for a data center network?
- Chapter 7 describes the overlay network and technologies. There are a lot of tenants (VMs) in a data center. These tenants need to communicate at Layer 2 or Layer 3 across the physical network. An overlay network is a logical network built on a physical underlay network. Point-to-point or point-to-multipoint tunnels are established on the overlay network to allow tenants to communicate. This chapter focuses on VXLAN, which is the most widely used overlay technology.
- Chapter 8 describes the implementation principles of the VXLAN control plane protocol BGP EVPN. With BGP EVPN, VXLAN-enabled devices can automatically establish tunnels and learn entries, reducing flooding traffic on the network.

This book focuses on key concepts in data center network design, and provides simple, easily comprehensible descriptions of design and implementation principles. After reading this book, you will know how to use CE series switches to build your own data center network. For details about the feature support and configuration and maintenance methods of CE series

switches, visit the Huawei enterprise technical support website and obtain complete product documentation at the following URL:

http://support.huawei.com/enterprise/en/index.html

Software Version

- This book describes the implementation on CE series switches running V200R002C50. To obtain the most current and detailed information about CE series switches, obtain the latest datasheets at the following URL:

  http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches
- In addition to hardware switches, Huawei also provides the CE1800V software data center switch. This book describes only hardware switches.

Contents

Chapter 1: Introduction to Huawei Data Center Switches

1.1 Switches in a Data Center

Switch selection is generally considered first during data center design.

To meet the needs of different customers, Huawei produces data center switches that comply with general standards and meet specific network requirements. This chapter briefly describes Huawei CE series switches and the applications of different switch models on the network.

CE series switches are next-generation high-performance switches designed for data centers. They meet customers' requirements for low latency, high performance, and high interface density, and flexibly support a variety of network structures. This makes them ideal for data center networks. Currently, CE series switches include CE12800 series, CE8800 series, CE7800 series, CE6800 series, and CE5800 series hardware switches. To view the CE switches portfolio, visit http://e.huawei.com/en/material/onLineView?MaterialID=fe45e4fdd09e4dbd920b40cf35c757c1.

The following describes each series of switches briefly.

1.2 CE12800 Series

CE12800 series switches are next-generation, high-performance, core modular switches designed for data centers and high-end campus networks. Currently, CE12800 series switches include the CE12804, CE12808, CE12812, CE12816, CE12804S, CE12808S, CE12804E, CE12808E, and CE12816E. CE12800 series switches help build elastic, virtualized, and high-quality networks by providing high-performance Layer 2/Layer 3 switching capabilities that are stable, reliable, and secure. The CE12804E, CE12808E, and CE12816E use the Huawei-developed Ethernet network processor to meet customers' requirements for flexible and fast service customization and implement refined O&M.

CE12800 switches use an advanced hardware architecture design and provide 100GE, 40GE, 10GE, and GE line cards with various interface densities. Interfaces on some line cards can work at the rate of 25 Gbit/s through mode switching or interface split. A CE12800 switch supports a maximum of 576 100GE, 576 40GE, 2304 25GE, 2304 10GE, or 768 GE interfaces that support line-rate forwarding. CE12800 switches support high-density access of high-speed servers and aggregation of TOR switches, providing high performance and large capacity for data center networks.

CE12800 switches provide high system reliability with hot standby of five hardware modules: Main Processing Units (MPUs), Switch Fabric Units (SFUs), Centralized Monitoring Units

(CMUs), power modules, and fan trays. MPUs work in 1+1 hot standby mode. SFUs work in N+M hot standby mode. CMUs work in 1+1 hot standby mode. Power modules support dual inputs and N+N backup, and have their own fans. Fan trays work in 1+1 backup mode. Each fan tray has two counter-rotating fans working in 1+1 backup mode, ensuring efficient heat dissipation.

On data center networks, CE12800 series switches can function as core and aggregation switches, or as spine switches in the spine-leaf architecture. CE12800 series switches can be deployed as standalone switches, in a stack or an M-LAG, or configured with VS. The switches support mainstream overlay technologies such as BGP EVPN VXLAN.

For more information about CE12800 series switches, visit:

http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce12800

http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce12800e

1.3 CE8800 Series

CE8800 series switches are next-generation Ethernet switches designed for data centers and high-end campus networks, providing high-performance, high interface density, and low latency. CE8800 series switches provide high-density 100GE, 40GE, 25GE, and 10GE interfaces, and support abundant data center features and high-performance stacking technologies, helping enterprises and carriers build cloud-oriented data center networks.

Currently, CE8800 series switches include the CE8860 series and CE8850 series.

- CE8860 series: The CE8860-4C-EI is a 2U high switch that supports flexible cards and interface split. Each switch supports a maximum of 32 100GE, 64 40GE, 128 25GE, or 128 10GE interfaces.
- CE8850 series: The CE8850-32CQ-EI is a 1U high 100GE fixed switch. Each switch supports a maximum of 32 100GE, 32 40GE, 128 25GE, or 130 10GE interfaces.

On data center networks, CE8800 series switches can function as high-density TOR/EOR access switches. When a small number of servers are deployed on a network, CE8800 switches can function as core and aggregation switches, or as spine and leaf switches in the spine-leaf architecture. The switches can be deployed as standalone switches or set up a stack or an M-LAG. They also support mainstream overlay technologies such as BGP EVPN VXLAN.

For more information about CE8800 series switches, visit:

http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce8800

1.4 CE7800 Series

CE7800 series switches are next-generation 40GE fixed switches designed for data centers and high-end campus networks, providing high-performance, high interface density, and low latency. CE7800 series switches provide high-density 40GE QSFP+ interfaces, and support abundant data center features and high-performance stacking technologies, helping enterprises and carriers build cloud-oriented data center networks.

Currently, CE7800 series switches include the CE7850 series and CE7855 series. Each switch provides 32 40GE QSFP+ optical interfaces, and each of these interfaces can be split into four 10GE interfaces.

On data center networks, CE7800 series switches can function as core and aggregation switches, or as spine and leaf switches in the spine-leaf architecture. The switches can be deployed as standalone switches or set up a stack, an M-LAG, or an SVF system. They also support mainstream overlay technologies such as BGP EVPN VXLAN.

For more information about CE7800 series switches, visit:

http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce7800

1.5 CE6800 Series

CE6800 series switches are next-generation 10GE fixed switches designed for data centers and high-end campus networks, providing high-performance, high interface density, and low latency. CE6800 series switches have an advanced hardware structure design with the industry's highest density of 10GE/25GE access interfaces. They provide 40GE/100GE uplink interfaces. The switches support a variety of data center features and high-performance stacking technologies.

Currently, CE6800 series switches include the CE6880 series, CE6870 series, CE6860 series, CE6850 series, and CE6810 series.

- CE6880 series: The switches use the Huawei-developed Ethernet network processor to meet customers' requirements for flexible and fast service customization and implement refined O&M. Each switch provides a maximum of 48 10GE SFP+ downlink optical interfaces or 48 10GBASE-T downlink electrical interfaces, two 40GE/100GE QSFP28 uplink optical interfaces, and four 40GE QSFP+ uplink optical interfaces. Each QSFP28 optical interface can be split into four 10GE interfaces.
- CE6870 series: Each switch provides a large buffer of 4 GB to easily handle traffic surges on data center networks caused by video, search, and other applications. Each switch provides a maximum of 48 10GE SFP+ downlink optical interfaces or 48 10GBASE-T downlink electrical interfaces, and a maximum of six 40GE/100GE QSFP28 uplink optical interfaces. Each QSFP28 optical interface can be split into four 10GE or 25GE interfaces.
- CE6860 series: Each switch provides a maximum of 48 10GE/25GE SFP28 downlink optical interfaces and a maximum of eight 40GE/100GE QSFP28 uplink optical interfaces. Each QSFP28 optical interface can be split into four 10GE or 25GE interfaces.
- CE6850 series: includes the CE6856HI, CE6855HI, CE6850U-HI, CE6851HI, CE6850HI, and CE6850EI models. Each switch provides a maximum of 48 10GE SFP+ downlink optical interfaces or 48 10GBASE-T downlink electrical interfaces, and a maximum of six 40GE QSFP+ uplink optical interfaces. Each QSFP+ optical interface can be split into four 10GE interfaces.
- CE6810 series: includes the CE6810EI series and CE6810LI series. Each switch provides a maximum of 48 10GE SFP+ downlink optical interfaces and a maximum of four 40GE QSFP+ uplink optical interfaces. Each QSFP+ optical interface can be split into four 10GE interfaces. CE6810LI switches are typically used as Layer 2 switches.

On data center networks, CE6800 series switches primarily function as high-density 10GE access switches, or as leaf switches in the spine-leaf architecture. The switches can be deployed as standalone switches or set up a stack, an M-LAG, or an SVF system. All models

except the CE6810 and CE6850EI models support mainstream overlay technologies such as BGP EVPN VXLAN.

For more information about CE6800 series switches, visit:

http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce6800

1.6 CE5800 Series

CE5800 series switches are next-generation fixed gigabit switches that provide 40GE uplink interfaces.

Currently, CE5800 series switches include the CE5855 series, CE5850 series and CE5810 series. A CE5800 series switch provides 48 or 24 10/100/1000BASE-T downlink electrical interfaces and a maximum of four 10GE SFP+ uplink optical interfaces and two 40GE QSFP+ uplink optical interfaces. Each QSFP+ interface can be split into four 10GE interfaces.

On data center networks, CE5800 series switches primarily function as high-density gigabit access switches, or as leaf switches in the spine-leaf architecture. The switches can be deployed as standalone switches or set up a stack, an M-LAG, or an SVF system (only as leaf switches).

For more information about CE5800 series switches, visit:

http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/ce5800

1.7 Summary

This chapter briefly describes CE series switches. Huawei provides both data center switches that use commercial chips and data center switches that use Huawei-developed chips, such as CE12800E and CE6880EI series switches.

These two types of switches are different as follows:

- The switches that use commercial chips provide basically the same features and performance specifications as non-Huawei switches using the same commercial chips. Commercial chips provide high throughput and software capabilities required by most networks, and support high interface density. However, the functions and features of switches that use these chips are limited by capabilities of the chips. Therefore, the flexibility of these switches is lower than that of switches that use Huawei-developed chips.
- The switches that use Huawei-developed chips can more precisely meet customers' customization requirements and be used more flexibly. Switches of this type are more powerful in terms of openness, automation, and independent O&M. One disadvantage of the switches is that their forwarding performance may decrease when increasing numbers of services are configured.

Both commercial chips and Huawei-developed chips have advantages and disadvantages. During network planning and design, select the appropriate data center switches according to service requirements and future network planning.

Chapter 2: Architecture

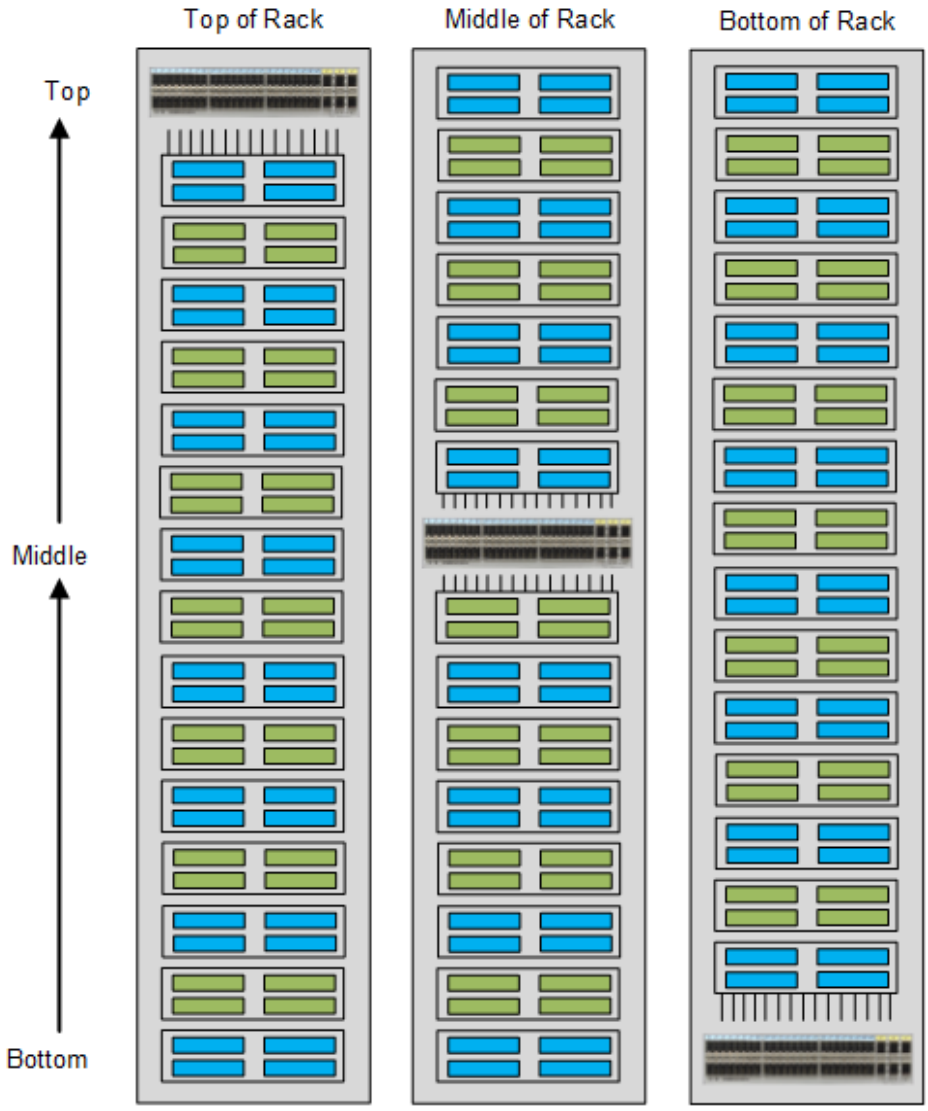
1.1 Overview

Chapter 1 provided an overview of the CloudEngine series data center switches. Chapter 2 will elaborate on the deployment of switches in data center equipment rooms, that is, the physical architecture. According to the positions where the switches are deployed in the cabinet or the access modes of the servers, the physical architecture of a switch can either be Top of Rack (TOR), Middle of Rack (MOR), or End of Row (EOR). The following sections describe these architectures.

1.2 TOR

In the TOR architecture, one or two switches are deployed in each server cabinet, and the servers in the cabinet are directly connected to the switches. While the name TOR implies that the switch is deployed at the top of the server cabinet, in reality, as shown in Figure 1-1, the switch can also be deployed in the middle or at the bottom of the cabinet. Generally, the deployment of switches on the top of the cabinet facilitates cabling. For this reason, deployment at the top of rack is most widely implemented.

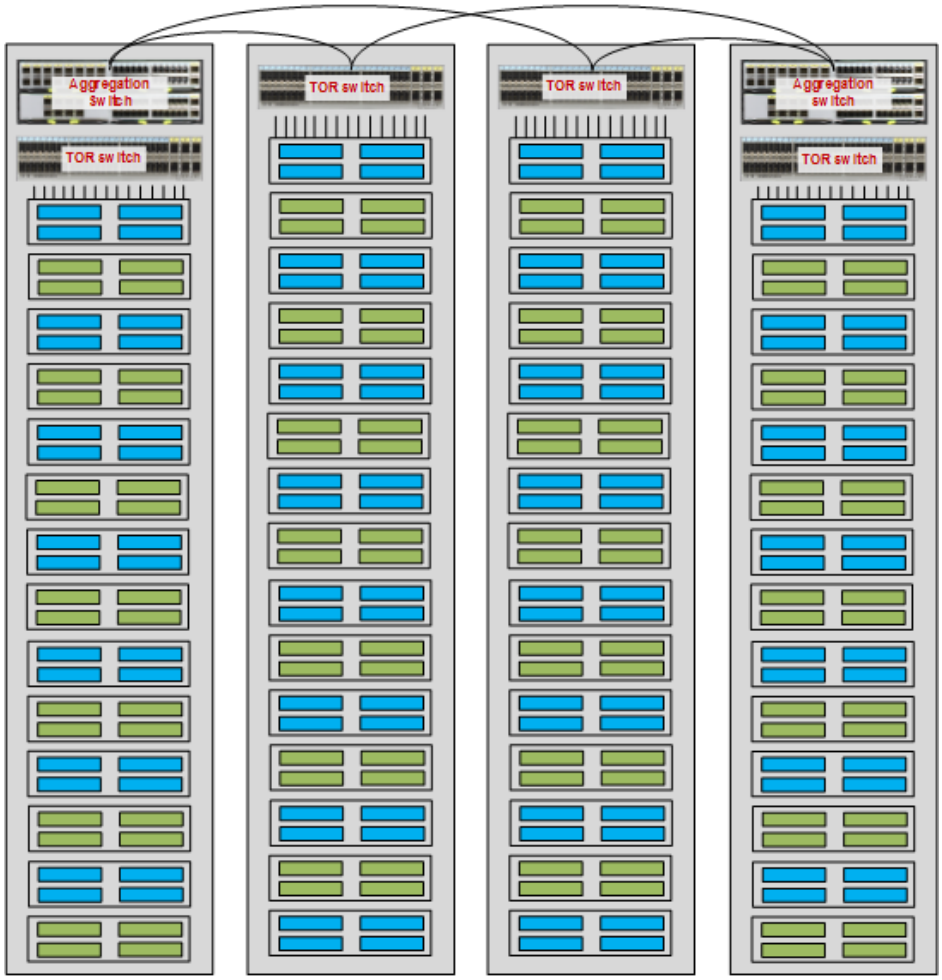
Figure 1-1 TOR architecture



Switches deployed in a server cabinet in TOR mode are called TOR switches. TOR switches are usually 1 U to 2 U fixed switches, such as Huawei CE5800 and CE6800 series switches.

The main benefit of the TOR architecture is that it simplifies the server-switch connection. The GE, 10GE, and 25GE interfaces on the server can be directly connected to the TOR switches through short cables, and then the 10GE, 40GE, or 100GE interfaces on the TOR switches can be dual homed to the aggregation switches through uplink optical fibers, as shown in Figure 1-2. This design shortens the cables used, simplifies cable management and network structure, and helps drive the development of environmentally friendly and energy-saving data centers. Additionally, this facilitates cable changes, if they are required during subsequent service expansion.

Figure 1-2 Connecting TOR switches to aggregation switches



In the TOR architecture, each cabinet can be considered as an independent management entity. Servers and switches can be upgraded by cabinet. During upgrades, the traffic forwarding of other cabinets is not affected, and the impact on services is minimized.

Optical fibers are often used to connect TOR switches to upstream devices because optical fibers are a better investment for long-term deployment than copper cables. Additionally, optical fibers can support higher bandwidth when higher speeds are required. Different optical fibers with different bandwidth capabilities can be flexibly selected.

Therefore, when choosing TOR switches, consider the number and rate of downlink interfaces that connect the switches to the servers and the flexibility of uplink interfaces. When choosing TOR switches, pay attention to the following points:

- If the servers provide GE interfaces, select CE5855EI series switches including the CE5855-48T4S2Q-EI and CE5855-24T4S2Q-EI. The CE5855-48T4S2Q-EI and CE5855-24T4S2Q-EI provide 48 and 24 downlink 10/100/1000BASE-T Ethernet electrical interfaces respectively. Both of them provide two 40GE QSFP+ Ethernet optical interfaces and four 10GE SFP+ Ethernet optical interfaces. Each 40GE interface can be split into four 10GE interfaces.
- If the servers provide 10GE interfaces, select CE6856HI series switches including the CE6856-48S6Q-HI and CE6856-48T6Q-HI. The CE6856-48S6Q-HI and CE6856-48T6Q-HI provide 48 downlink 10GE SFP+ Ethernet optical interfaces and

10GBASE-T Ethernet electrical interfaces respectively. Both of them provide six 40GE QSFP+ Ethernet optical interfaces. Each 40GE interface can be split into four 10GE interfaces.

- If TOR switches that provide large buffers are needed, use CE6870EI series switches. CE6870 series switches are classified into the CE6870-48S6CQ-EI, CE6870-24S6CQ-EI, and CE6870-48T6CQ-EI according to the type and number of their downlink interfaces. Figure 1-3 shows the CE6870-48S6CQ-EI as an example. This switch provides six uplink 40GE or 100GE QSFP28 Ethernet optical interfaces. Each 40GE and 100GE QSFP28 Ethernet optical interface can be split into four 10GE interfaces and four 25GE interfaces respectively. The diversity of uplink interfaces enables the CE6870EI to meet various service requirements and to maximize the return on investment (ROI) over the long run. Each CE6870 switch provides a large buffer of 4 GB to easily handle traffic surges on data center networks caused by video, search, and other applications.

Figure 1-3 CE6870-48S6CQ-EI

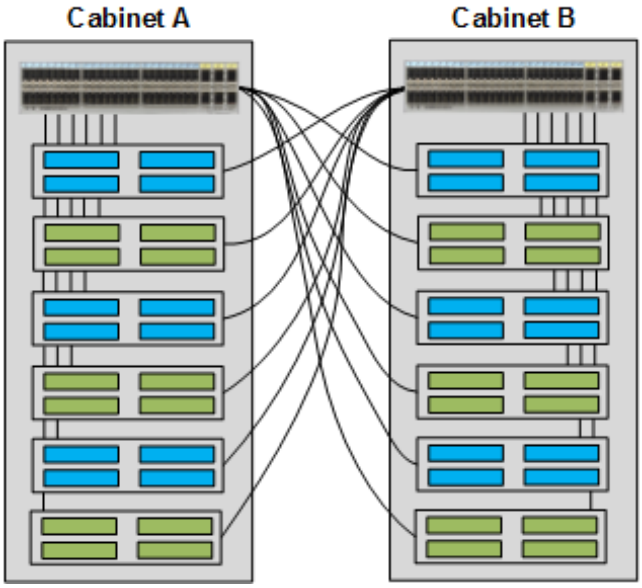
The TOR architecture also has disadvantages, with the primary disadvantage being that it expands the management domain of the data center equipment room. Each cabinet is deployed with several switches, which means there are a lot of switches in each equipment room. Every switch must be configured, managed, and maintained. If there are 10 rows of cabinets in an equipment room, 10 cabinets are placed in each row, and two TOR switches are deployed in each cabinet, there are 200 TOR switches to be managed and maintained. Although the configurations of these switches are basically the same, the labor costs are still high and the possibility of incorrect configurations increases.

To solve these problems, CE series switches provide the following solutions. For newly delivered or unconfigured switches, the Zero Touch Provisioning (ZTP) function can be used to automatically configure switches in a batch. By default, the ZTP function is enabled on CE series switches. Switches running ZTP can automatically obtain and load version files (including the system software, configuration file, license file, patch file, and user-defined file) from a USB flash drive or file server, freeing network engineers from onsite configuration. CE series switches also support diverse device virtualization technologies, for example, stacking. These technologies simplify the device management plane, reduces labor costs, and improves deployment efficiency. Chapter 5 will explore these technologies in detail.

Another disadvantage of the TOR architecture is interface waste. Currently, most TOR switches can provide 48 GE, 10GE, or 25GE downlink interfaces. When two TOR switches are deployed in each cabinet, there are 96 downlink interfaces, and a lot of servers need to be deployed in the cabinet to fully use these interfaces.

Cross connections between adjacent cabinets can reduce the waste of interfaces. As shown in Figure 1-4, one 48-interface TOR switch is deployed in each cabinet. Half of the interfaces are used to provide access for servers in the cabinet, and the other 24 interfaces are used to provide access for servers in the adjacent cabinet. A disadvantage of this solution is that extra cables are needed. However, this solution is more cost-effective than the deployment of two TOR switches in each cabinet.

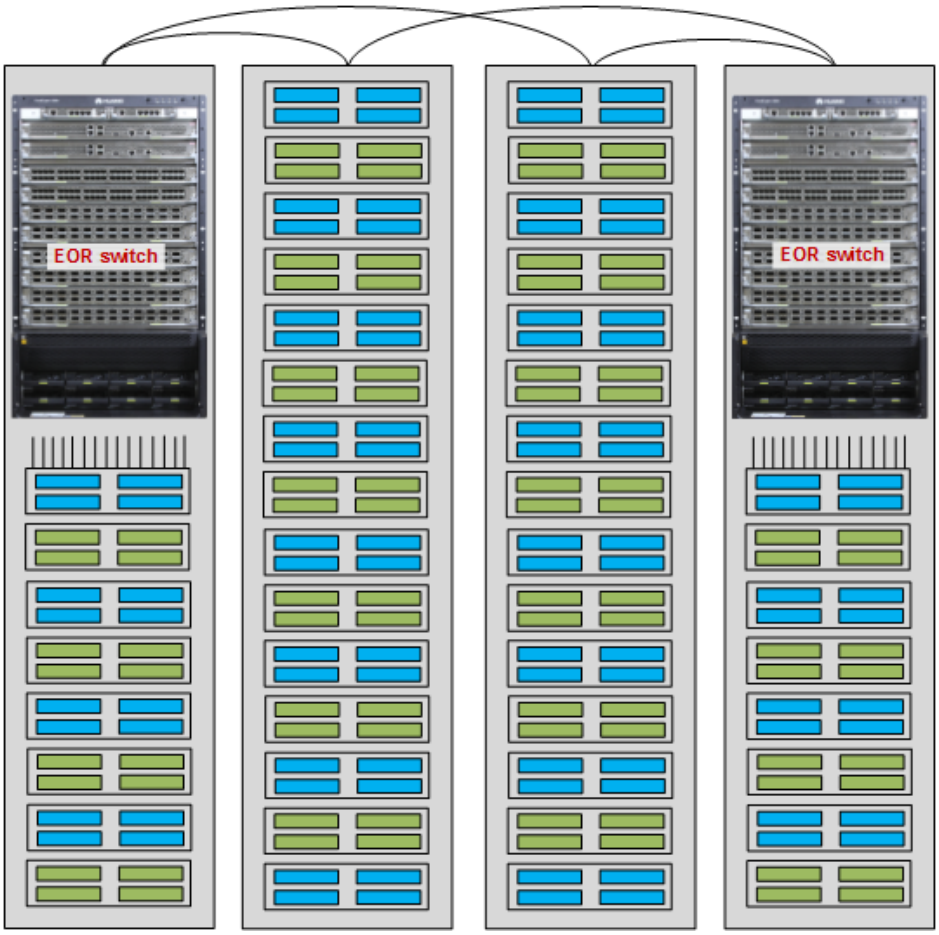
Figure 1-4 Cross connection between adjacent cabinets



1.3 EOR/MOR

In contrast to the TOR architecture, EOR architecture provides unified network access points at the end of each row of cabinets. As shown in Figure 1-5, EOR switches are deployed at the end of each row of cabinets to support server access.

Figure 1-5 EOR architecture



To ensure reliability, each row of cabinets is configured with two network cabinets, which are located at the head and end of the row. The server network interface cards (NICs) are connected to the distribution frame in the cabinet using a relatively short RJ45, direct attach cable (DAC), or optical fiber. The network cables, optical fibers, and copper cables on the distribution frame are bundled and then channeled through either overhead cable troughs or under floor bundles to the end-of-row network cabinets.

In the EOR architecture, access switches are deployed in 1 or 2 cabinets for centralized management and maintenance. This requires more cables to connect the server cabinets to network cabinets. If the server cabinet is far away from a network cabinet, longer cables are needed. As a result, the cable maintenance workload is heavy and flexibility is poor.

Middle of Row (MOR) architecture is an improvement over the EOR architecture because it provides a unified network access cabinet for the servers. The network cabinet is placed in the middle-of-row cabinets to shorten the distance between it and server cabinets, simplifying cable management and maintenance. However, compared with the TOR architecture, the EOR and MOR architectures are disadvantageous in cabling and cable management and maintenance. The following describes the EOR architecture, which is also applicable to the MOR architecture unless otherwise specified.

EOR switches are usually modular switches, for example, Huawei CE12800 series switches. If there is a small number of servers in an equipment room, use CE8800 and CE7800 series switches as EOR switches.

Compared with fixed switches, modular switches have the following advantages:

- A lot of access interfaces of different types. Line processing units (LPUs) of different quantity and rates can be configured on modular switches to flexibly control the number and rate of access interfaces. For example, CE12800 series switches support LPUs with a maximum of 36*100GE, 36*40GE, 48*10GE, and 48*GE interfaces. The interfaces can be split in different ways for flexible data center server access.
- High reliability. Modular switches provide redundant hardware, such as multiple switch fabric units (SFUs), power modules, and fan modules, to improve reliability.
- Maximized ROI. If a higher access rate is required in a data center, only replacement of the current LPU with one that has a higher rate is needed, instead of replacing the entire device. From the perspective of the entire life cycle, modular switches are more cost-effective.

In the EOR architecture, the data center's management domain is considerably reduced because it is managed by row instead of by rack. However, this also means that once an EOR switch fails or fails to be upgraded, the entire row of servers is affected. Therefore, higher requirements are posed for EOR switches.

1.4 Summary

Table 1-1 summarizes the advantages and disadvantages of the TOR and EOR/MOR architectures.

Table 1-1 Advantages and disadvantages of the TOR and EOR/MOR architectures

Physical Architecture	TOR Architecture	EOR/MOR Architecture
Advantages	• Simple cabling, convenient cable maintenance, and high scalability All servers are connected to the TOR switches in the same cabinet. Only the uplinks of the switches are connected to the aggregation switches outside the cabinet to simplify cabling. When a server needs to be upgraded (for example, from 10GE to 25GE), only small-scale changes in cable connections are required. This makes the TOR architecture highly scalable. • Cabinet-based modular management and small-scale fault impact The TOR architecture supports modular management based on cabinets. When a device is	• Simple management and high reliability The number of access switches to be managed and maintained is relatively small. Most EOR switches are modular switches, and their key components feature a redundant design. Therefore, the high reliability of the entire system can be guaranteed. When service expansion or upgrades are required, you can simply add or replace LPUs instead of the entire device. • High interface use The servers are centrally connected to EOR switches so that the interfaces on the switches can be fully used.

Physical Architecture	TOR Architecture	EOR/MOR Architecture
	faulty or upgraded, the impact on services is minimized.	
Disadvantages	• Port waste Every server cabinet is restricted by the output power. The number of servers that can be deployed is limited. As a result, the access interface usage of switches in the cabinet is insufficient. • Complex management and maintenance of TOR switches The workload of managing and maintaining network devices is heavy because a large number of access switches exist in the TOR architecture.	• Complex cabling and difficult maintenance Many cables are required between server cabinets and network cabinets, and the cables must be longer if the network cabinet is far away from a server cabinet. As a result, cable management and maintenance create heavy workloads and have limited flexibility. • Large-scale fault impact When an EOR switch becomes faulty, all servers in the row where the switch resides will be affected.

In conclusion, both the TOR and EOR/MOR architectures have their own advantages and disadvantages, and they can each be applied to different scenarios. When selecting an architecture, consider the service type of the data center, service characteristics, and conditions of different data centers including the investment and management costs. For traditional user data centers where the service data volume is not large and the scalability requirement is not high, the EOR/MOR architecture is still favored. For user data centers that focus on distributed architectures and require high scalability, the TOR architecture will be widely used.

Chapter 3: Cabling

1.1 Overview

Chapter 2, which discussed the TOR and EOR architectures for switches in data center equipment rooms, presented the importance of physical cabling design in data centers.

The use of the optimal cable for the appropriate scenarios is a prerequisite for constructing high-speed data center networks. When cabling infrastructure is designed, the signal attenuation, transmission distance, cable routing and termination, cost, installation and removal, and future network upgrade require thorough consideration.

Most suppliers provide three types of cables: network cables, referred to as twisted pairs in this document, optical fibers, and DAC.

This chapter discusses each of these types of cables and how to choose the optimal cables for each data center network design.

1.2 Network Cable

According to the frequency and signal-to-noise ratio (SNR), common network cables include Category 5 cable (Cat 5), Category 5 enhanced (Cat 5e), and Category 6 cable (Cat 6). These are twisted pair cables that use RJ45 connectors, with a maximum transmission distance of 100 m. Network cables also include Category 1 cable (Cat 1), Category 2 cable (Cat 2), Category 3 cable (Cat 3), Category 4 cable (Cat 4), Category 6a (Cat 6a), and Category 7 cable (Cat 7). Generally, a higher number indicates a later version, more advanced technology, and higher bandwidth and cost.

Depending on whether the shield layer is available, network cables can be classified into shielded twisted pair (STP) and unshielded twisted pair (UTP). STP cables can reduce radiation and prevent information from being intercepted and external electromagnetic interference from entering. Compared with the same type of UTP cables, STP cables boast higher transmission rate, but they are more expensive and more difficult to install. UTP cables feature low cost, light weight, and are easy to bend. They rarely cause great impact on common networks. Therefore, UTP cables are more widely used. However, to implement a full-duplex transmission rate of up to 10 Gbit/s, only Category 7 STP-7 can be used.

The different categories of network cables are described as follows:

- Cat 1

  Cat 1 cables, used as telephone cables before the early 1980s, are mainly used for voice transmission, but not for data transmission.

- Cat 2

Cat 2 cables have a transmission frequency of 1 MHz and are used for voice transmission as well as data transmission at a maximum transmission rate of 4 Mbit/s. This type of cable is commonly used on old token ring networks that use the 4 Mbps standard token passing protocol.

- Cat 3

  Cat 3 cables have a transmission frequency of 16 MHz and are used for voice transmission as well as data transmission at a maximum transmission rate of 10 Mbit/s on 10Base-T networks. This type of cable is specified in the ANSI/TIA-568.C.2 standard as the lowest-class cable for 10Base-T networks.

- Cat 4

  Cat 4 cables have a transmission frequency of 20 MHz and are used for voice transmission as well as data transmission at a maximum transmission rate of 16 Mbit/s on token-based local area networks (LANs) and 10Base-T/100Base-T networks.

- Cat 5

  Cat 5 cables have a transmission frequency of 100 MHz and are used for voice transmission as well as data transmission at a maximum transmission rate of 100 Mbit/s on 100Base-T and 10Base-T networks. Offering high winding density and a high-quality insulation shield, this type of cable is most commonly used as Ethernet cable.

- Cat 5e

  Cat 5e cables have a transmission frequency of 100 MHz and are mainly used on gigabit Ethernet (GE) networks. With low attenuation and crosstalk, a higher attenuation-to-crosstalk ratio (ACR) and SNR, and small latency error, performance of Cat 5e cables is greatly improved.

- Cat 6

  Cat 6 cables have a transmission frequency of 250 MHz and are applicable to networks that require a transmission rate higher than 1000 Mbit/s. The appearance and structure of Cat 6 twisted pair cables are different from those of Cat 5 or Cat 5e twisted pairs. A Cat 6 twisted pair uses an additional insulation cross framework, and the four wire pairs are placed in four respective grooves of the cross framework. Additionally, the diameter of Cat 6 cables is larger.

- Cat 6a

  Cat 6a cables have a transmission frequency of 500 MHz, which is twice that of Cat 6 cables. The maximum transmission rate can reach 10 Gbit/s. This type of cable is mainly used on 10GE networks. As an improvement of Cat 6 cables, Cat 6a cables are UTP cables specified in ANSI/EIA/TIA-568B.2 and ISO category 6/class E standards. Cat 6a cables greatly reduce the crosstalk and attenuation and improve the SNR.

- Cat 7

  Cat 7 cables have a transmission frequency of at least 600 MHz and a transmission rate of up to 10 Gbit/s. This type of cables is mainly used to adapt to the application and development of 10GE technologies. In the ISO category 7/class F standard, Cat 7 cables are the latest STP cables. Unlike other types of network cables, Cat 7 cables use GigaGate45 (CG45) as the connector.

Table 1-2 describes the basic parameters of several network cables.

Table 1-2 Basic parameters of network cables

Network Cable Type	Usage Scenario	Transmission Frequency	Maximum Transmission Rate	Transmission Distance

Network Cable Type	Usage Scenario	Transmission Frequency	Maximum Transmission Rate	Transmission Distance
Cat 5	100Base-T and 10Base-T networks	1 MHz to 100 MHz	100 Mbit/s	100 m
Cat 5e	1000Base-T networks	1 MHz to 100 MHz	1000 Mbit/s	100 m
Cat 6	1000Base-T networks	1 MHz to 250 MHz	1000 Mbit/s or 10 Gbit/s	100 m or 37 m to 55 m
Cat 6a	10GBase-T networks	1 MHz to 500 MHz	10 Gbit/s	100 m
Cat 7	10GBase-T networks	1 MHz to 600 MHz	10 Gbit/s	100 m

In the TOR architecture, network cables are required in each cabinet. In the EOR architecture, network cables are used to connect servers and switches. The distance between the servers and switches and the cabling scale must be considered for future network upgrades.

If network cables are required in network deployment in TOR scenarios, CE5855-48T4S2Q-EI, CE6810-32T16S4Q-LI, CE6850-48T4Q-EI, or other Huawei CE series switches that provide electrical interfaces can be used. In EOR scenarios, the use of high-performance switches is recommended, such as the CE6870-48T6CQ-EI and CE6880-48T4Q2CQ-EI, or CE12800 series switches with CE-L48GT or CE-L48XT series cards.

1.3 Optical Fiber

Optical fibers can be classified based on their optical transmission modes into multimode fibers (MMFs) and single-mode fibers (SMFs).

This section describes the differences between MMFs and SMFs.

MMF

An MMF has a relatively thick fiber core and can transmit optical signals of multiple modes. However, the MMF mode dispersion is large and accentuates as the transmission distance increases. Therefore, MMFs are often used with multi-mode optical modules in short-distance and low-cost transmission scenarios.

On the basis of fiber diameters and modal bandwidths, MMFs are classified into OM1, OM2, OM3, and OM4. Table 1-3 lists the fiber diameter and modal bandwidth of the different classes of optical fibers. The G.651 standard optical fibers are commonly used MMFs that can transmit between 800 nm to 900 nm or 1200 nm to 1350 nm optical signals.

MMFs are marked with "MM". Generally, the optical fibers of the OM1 and OM2 classes are orange, and those of the OM3 and OM4 classes are light green.

Figure 1-6 Appearance of an OM1/OM2 MMF

Figure 1-7 Appearance of an OM3/OM4 MMF

Table 1-3 MMF classification

Fiber Class	Fiber Diameter (μm)	Modal Bandwidth (MHz x km)
OM1	62.5	200
OM2	50	500
OM3	50	2000
OM4	50	4700

The transmission distances of MMFs depend on the interface type, center wavelength, and fiber class. Table 1-4 describes the specifications of common MMFs.

Table 1-4 Specifications of common MMFs

Application Type	Center Wavelength (nm)	Fiber Class	Maximum Transmission Distance (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

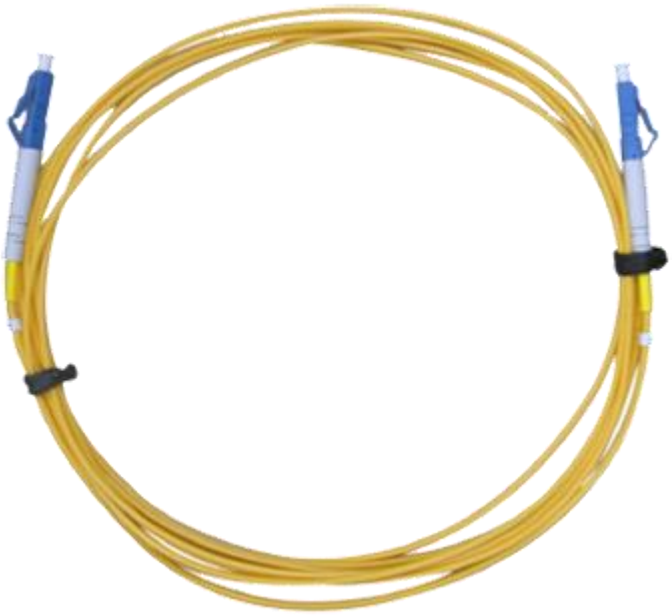
SMF

SMFs have a thin fiber core and can transmit only one mode of optical signals. Therefore, the mode dispersion is small, making SMFs suitable for long-distance transmission.

G.652 standard optical fibers are commonly used SMFs that can transmit between 1260 nm to 1360 nm or 1530 nm to 1565 nm optical signals.

SMFs are marked with "SM" and generally yellow, as shown in Figure 1-8.

Figure 1-8 Appearance of an SMF

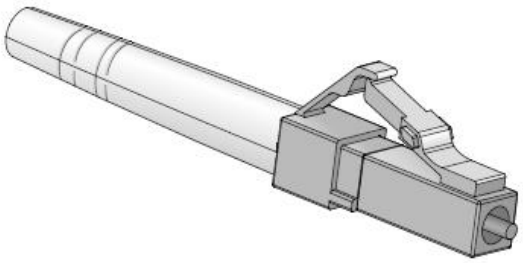


Connector Type

The Huawei-developed optical modules of CE series switches support two types of optical connectors: lucent connectors or local connectors (LCs) and Multi-fiber Push On (MPO) connectors.

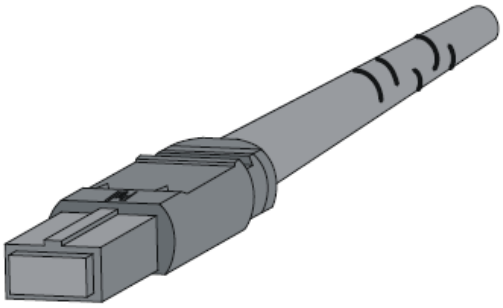
- Figure 1-9 shows the appearance of an LC.

  Figure 1-9 Appearance of an LC



- Figure 1-10 shows the appearance of an MPO connector.

Figure 1-10 Appearance of an MPO connector



The MPO connectors used by Huawei optical modules are female connectors with guide pins.

In TOR scenarios, MMFs are used to connect devices inside a cabinet. When the transmission distance is less than or equal to 400 m, MMFs can also be used at the aggregation layer of the network. In EOR scenarios, MMFs are used to connect switches and servers. If the distance between devices to be connected is within the transmission distance of the MMFs, MMFs can also be used at the core or aggregation layer of the network.

SMFs are often used for long-distance transmission.

Therefore, when optical fibers are required for network deployment in TOR scenarios, CE6851-48S6Q-HI, CE6855-48S6Q-HI, or other Huawei CE series switches that provide optical interfaces can be used. In EOR scenarios, the use of high-performance switches is recommended, such as the CE6870-48S6CQ-EI and CE7855-32Q-EI, or CE12800 series switches with cards providing optical interfaces.

Active optical cables (AOCs) can also be used. AOCs integrate optical modules and optical fibers. Specifically, an AOC encapsulates two optical modules with optical fibers. The optical modules at both ends of the AOC must have built-in laser components because the transmission media are optical fibers.

The optical modules on the AOCs are fixed and some optical components are not needed in the manufacturing process. Therefore, they are cost-saving for users. In addition, AOCs have low requirements on the environment and their fiber connectors do not present cleaning problems, ensuring the high reliability of the cables.

However, because the cable length is fixed, AOCs offer low flexibility of configuration. This limits the applicability of this type of cable to short-distance transmission scenarios.

Figure 1-11 Appearance of an AOC

1.4 DAC

DACs are also referred to as Twinax copper cables or high-speed cables. DACs have a fixed length and a fixed connector at each end, as shown in Figure 1-12.

Figure 1-12 DAC

DACs are classified into active cables and passive cables. Active DACs have built-in amplifiers and equalizers to improve signal quality, but are expensive. In most cases, if the transmission distance is less than 5 m, passive DACs can be used. If the transmission distance is longer than 5 m, active DACs can be used.

The connectors of the DACs are of the same type as the interfaces on the optical modules. However, in contrast to the optical modules, the connectors of DACs do not have expensive optical lasers or other electronic components. This greatly reduces their cost and power consumption. Because of this, DACs are widely used in short-distance connection in data center networks.

In TOR scenarios, DACs are the best choice for short-distance cabling in a cabinet. In EOR scenarios, if the transmission distance is less than 10 m, DACs can also be used.

Table 1-5 provides basic information about the different types of DACs that may be used in a data center, including the bending radius of the DACs. Like many other cables, DACs are sensitive to bending, and it affects the cable transmission rate. Like many other cables, DACs are sensitive to poor bending that affects the cable transmission rate.

Table 1-5 Basic information about DACs

DAC Cable Name	Maximum Transmission Distance (m)	Electrical Characteristics	Connector Type	Minimum Clearance for Cable Routing (mm)	Minimum Bending Radius (mm)
SFP-10G-CU1M	1	Passive	SFP+<->SFP+	60	35
SFP-10G-CU3M	3	Passive	SFP+<->SFP+	60	35
SFP-10G-CU5M	5	Passive	SFP+<->SFP+	60	35
SFP-10G-AC7M	7	Active	SFP+<->SFP+	60	35
SFP-10G-AC10M	10	Active	SFP+<->SFP+	60	35
QSFP-40G-CU1M	1	Passive	QSFP+<->QSFP+	75	50
QSFP-40G-CU3M	3	Passive	QSFP+<->QSFP+	75	50
QSFP-40G-CU5M	5	Passive	QSFP+<->QSFP+	75	50
QSFP-4SFP10G-CU1M	1	Passive	QSFP+<->4*SFP+	QSFP+ interface: 100 SFP+ interface: 60	QSFP+ interface: 50 SFP+ interface: 35
QSFP-4SFP10G-CU3M	3	Passive	QSFP+<->4*SFP+	QSFP+ interface: 100 SFP+ interface: 60	QSFP+ interface: 50 SFP+ interface: 35
QSFP-4SFP10G-CU5M	5	Passive	QSFP+<->4*SFP+	QSFP+ interface: 100 SFP+ interface: 60	QSFP+ interface: 50 SFP+ interface: 35
QSFP28-100G-CU1M	1	Passive	QSFP28<->QSFP28	90	70

DAC Cable Name	Maximum Transmission Distance (m)	Electrical Characteristics	Connector Type	Minimum Clearance for Cable Routing (mm)	Minimum Bending Radius (mm)
QSFP28-100G-CU3M	3	Passive	QSFP28<->QSFP28	90	70
QSFP28-100G-CU5M	4	Passive	QSFP28<->QSFP28	90	70
SFP-25G-CU1M	1	Passive	SFP28<->SFP28	70	40
SFP-25G-CU3M	3	Passive	SFP28<->SFP28	70	40
SFP-25G-CU3M-N	3	Passive	SFP28<->SFP28	70	40
SFP-25G-CU5M	5	Passive	SFP28<->SFP28	70	40
QSFP-4SFP25G-CU1M	1	Passive	QSFP28<->4*SFP28	QSFP28 interface: 100 SFP28 interface: 70	QSFP28 interface: 50 SFP28 interface: 40
QSFP-4SFP25G-CU3M	3	Passive	QSFP28<->4*SFP28	QSFP28 interface: 100 SFP28 interface: 70	QSFP28 interface: 50 SFP28 interface: 40
QSFP-4SFP25G-CU3M-N	3	Passive	QSFP28<->4*SFP28	QSFP28 interface: 100 SFP28 interface: 70	QSFP28 interface: 50 SFP28 interface: 40
QSFP-4SFP25G-CU5M	5	Passive	QSFP28<->4*SFP28	QSFP28 interface: 100 SFP28 interface: 70	QSFP28 interface: 50 SFP28 interface: 40

NOTE

The differences between SFP-25G-CU3M-N copper cables and SFP-25G-CU3M copper cables are as follows:

The SFP-25G-CU3M-N copper cable (26AWG) is thicker than the SFP-25G-CU3M copper cable (30AWG). Therefore, the transmission loss is smaller. When the SFP-25G-CU3M-N copper cable (26AWG) is used, the forward error correction (FEC) function does not need to be enabled on the interfaces.

The same difference exists between QSFP-4SFP25G-CU3M-N copper cables and QSFP-4SFP25G-CU3M copper cables.

1.5 Summary

Table 1-6 summarizes the advantages and disadvantages of network cables, optical fibers, and DACs that have been discussed in this chapter.

Table 1-6 Advantages and disadvantages of network cables, optical fibers, and DACs

Cable Type	Advantage	Disadvantage	Usage Scenario
Network cable	• Low price • Easy installation	• Low transmission rate • Short transmission distance	To reduce costs, network cables can be used for interconnections between electrical interfaces. Currently, Cat 6 and Cat 6a cables can support a higher transmission rate.
Optical fiber	• Large transmission capacity • High transmission rate • Long transmission distance • High anti-interference performance	• High deployment cost • Complex installation • High requirements on the environment	Optical fibers offer large-capacity, highly reliable data transmission over long distances. All-optical networks are an inevitable trend for the future evolution of data centers.
DAC	• Low deployment cost • High transmission rate • High anti-interference performance	• Short transmission distance • Fixed cable length and poor flexibility	DACs can be used for the reliable, short-distance interconnection of optical interfaces on data center networks. The deployment cost of DACs is lower than that of optical fibers.

For more information about cables, see CloudEngine 12800 Hardware Description or CloudEngine 8800&7800&6800&5800 Hardware Description.

Chapter 4: Oversubscription Ratio

1.1 Overview

Traffic oversubscription occurs when, due to architecture or device faults, the data packets on a network cannot be forwarded at the line rate without packet loss. During traffic oversubscription, some interfaces on the network device are congested and forced to discard some packets. The degree of oversubscription is assessed using the oversubscription ratio. Specifically, this is the ratio of the total bandwidth of all southbound (downlink) interfaces of a system to that of all northbound (uplink) interfaces of the system.

For example, if 10 servers need to be deployed, with each connecting to an access switch through a 10GE interface, there is a total of 100 Gbit/s (10 x 10 Gbit/s = 100 Gbit/s) southbound bandwidth. Assuming that this switch has two 40GE interfaces that can be used to connect to the aggregation switches at the upper layer, there is a total of 80 Gbit/s (2 x 40 Gbit/s = 80 Gbit/s) northbound bandwidth. In this case, the oversubscription ratio is 1.25:1 (100 Gbit/s ÷ 80 Gbit/s = 1.25).

The causes of traffic oversubscription can be classified into two types:

- The switch does not support line-rate forwarding, and traffic oversubscription may occur within the switch.
- Due to the network architecture design, traffic oversubscription occurs when the packets are forwarded regardless of whether they are forwarded at the line rate.

This section explains both these causes of oversubscription.

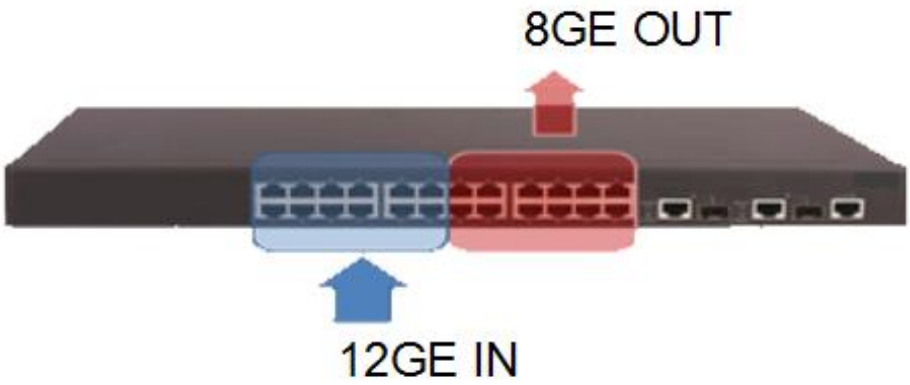
NOTE

In actual scenarios, the network-layer protocol overhead may affect but is not crucial to the oversubscription ratio. To simplify description, the network-layer protocol overhead is not considered in the calculation of the packet transmission rate and bandwidth oversubscription in this chapter.

Oversubscription caused by non-line-rate forwarding switches

Assume that a switch can forward packets at a maximum of 8 Gbit/s line rate. If the switch simultaneously forwards traffic from its first 12 interfaces to its last 12 interfaces at a certain time and each interface works at a 1 Gbit/s speed, congestion occurs within the switch. In this case, forwarding oversubscription occurs, as shown in Figure 1-13. The switch receives 12 GB traffic but forwards only 8 GB traffic per second. In this case, the oversubscription ratio is the input bandwidth (12 Gbit/s) divided by the output bandwidth (8 Gbit/s), that is, 1.5:1.

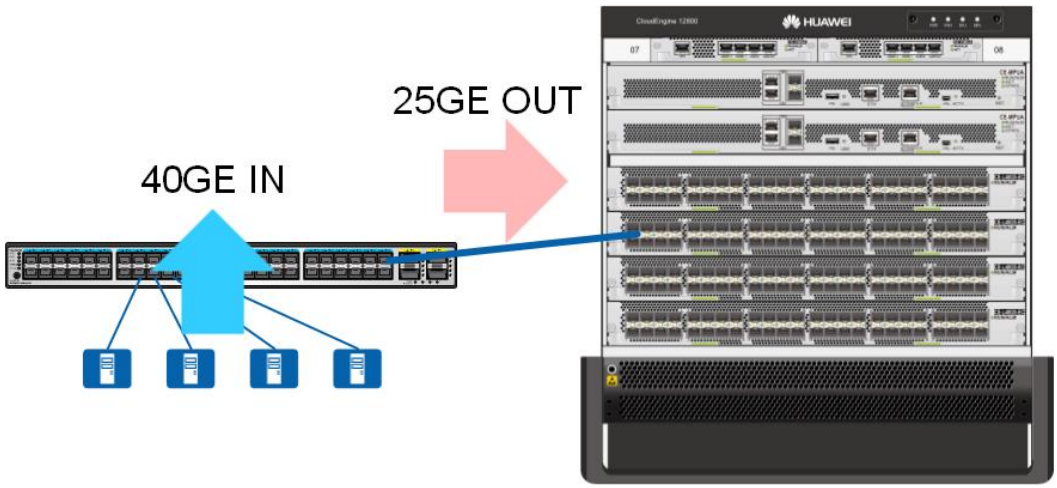
Figure 1-13 Oversubscription caused by non-line-rate forwarding switches



Oversubscription caused by the network design

In Figure 1-14, four servers are connected to an access switch through 10GE links, and the access switch is connected to the core switch through a 25GE link. Therefore, the downlink bandwidth of the access switch is 40 Gbit/s, and the uplink bandwidth is 25 Gbit/s. The oversubscription ratio is the downlink bandwidth (40 Gbit/s) divided by the uplink bandwidth (25 Gbit/s), which is 1.6/1.

Figure 1-14 Oversubscription caused by the network design



The ideal oversubscription ratio is 1:1. However, a low-oversubscription-ratio design indicates that switches with a higher uplink interface bandwidth must be used and such switches cost more than those without higher uplink interface bandwidth. When cost is not an issue, an oversubscription ration of 1:1 is desirable. On the other hand, servers do not work at capacity or occupy 100% bandwidth at all times. Even if the oversubscription ratio is not 1:1, packet loss may not definitely occur due to traffic congestion and services still can run properly. This makes it critical to find the most suitable oversubscription ratio.

The oversubscription ratio reflects the capability of a network to forward traffic at the line rate and is often used to measure network performance. On campus networks, the traffic oversubscription ratio is relatively large because the overall traffic volume is small. However,

data center networks have high performance requirements, making the traffic oversubscription design essential.

1.2 Traffic Oversubscription Design

Before designing a network traffic oversubscription ratio, we need to understand the service applications and features to be deployed on the network and to determine the network services and traffic models. Additionally, the volume and ratio of the east-west and north-south traffic must be thoroughly considered for the design of a suitable oversubscription ratio and selection of suitable devices. Overall, the following aspects are crucial to the design of the oversubscription ratio:

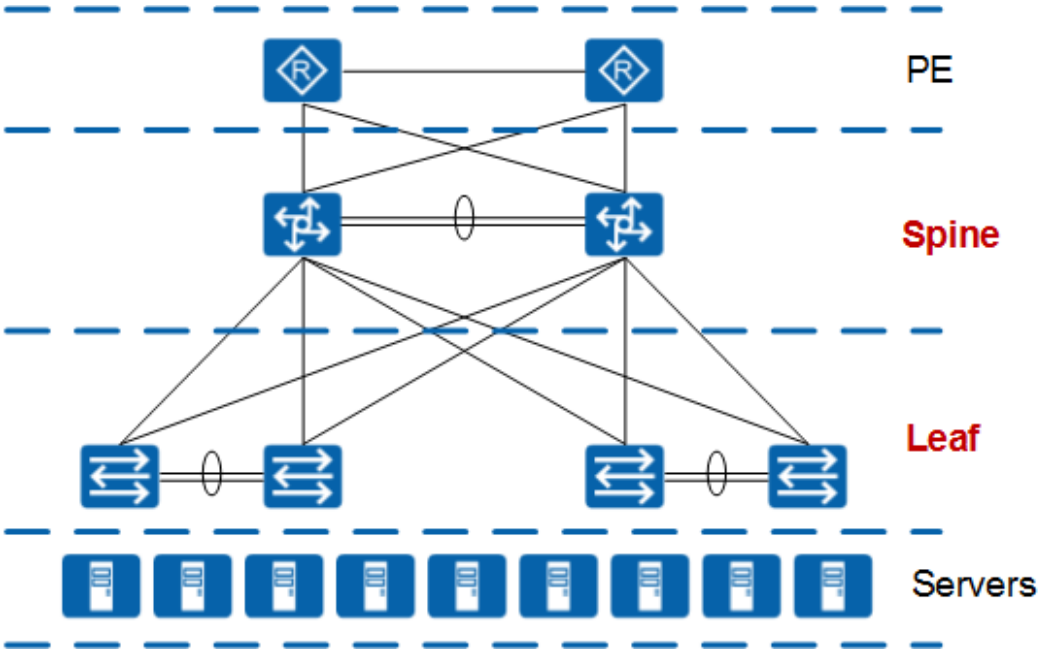
- Network architecture
- Available link design
- Device model selection

In practice, unless there is a high requirement on the traffic oversubscription ratio, a simple method can be used. That is, the oversubscription ratio can be designed based on the bandwidth o the available uplink interfaces on the switch.

Network Architecture

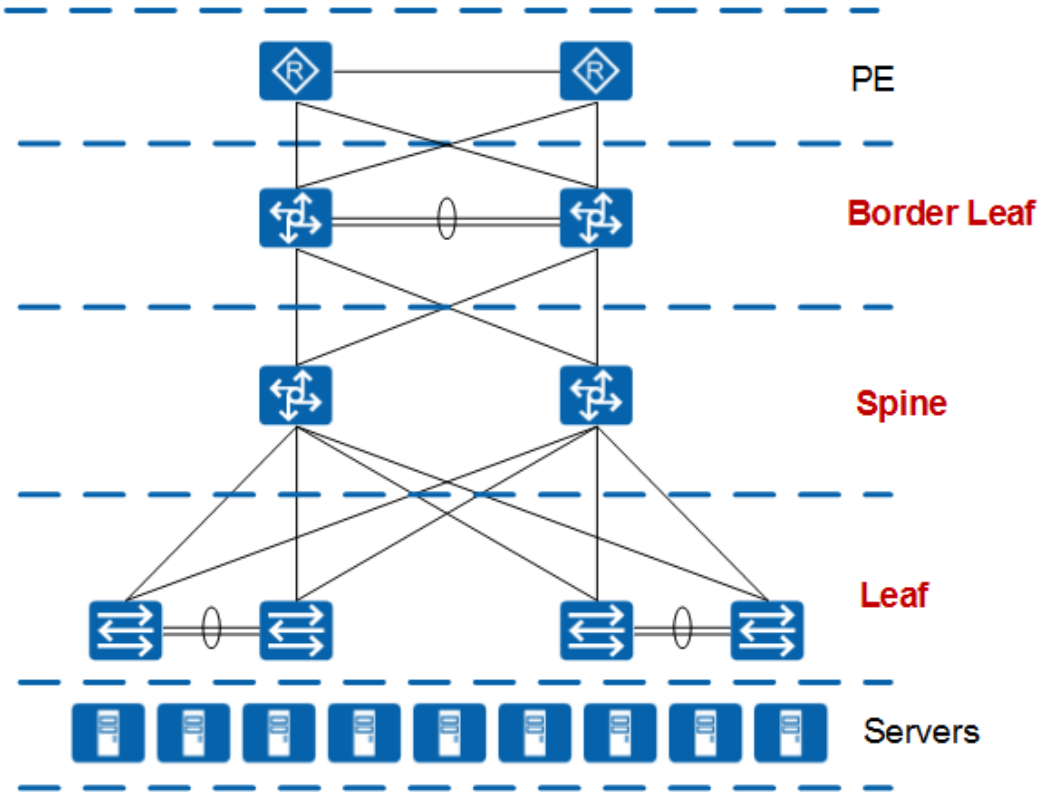
Currently, the two-layer network architecture is used widely. The two-layer architecture refers to the flattened spine+leaf architecture, as shown in Figure 1-15. This architecture has an overall lower oversubscription ratio compared with the multi-layer architecture and is applicable to environments such as data centers that have high requirements for network performance.

Figure 1-15 Two-layer network architecture in a data center



Currently, many data center networks adopt a networking design with three layers: border leaf, spine, and leaf, as shown in Figure 1-16. This design is actually also a two-layer architecture because the border leaf and leaf belong to the same layer. However, in government, finance, or some special areas, switches at the spine layer must be separated from the gateways at the same layer and security isolation must be deployed to meet the requirements on service architecture or network security.

Figure 1-16 Data center network architecture with the border leaf, spine, and leaf layers



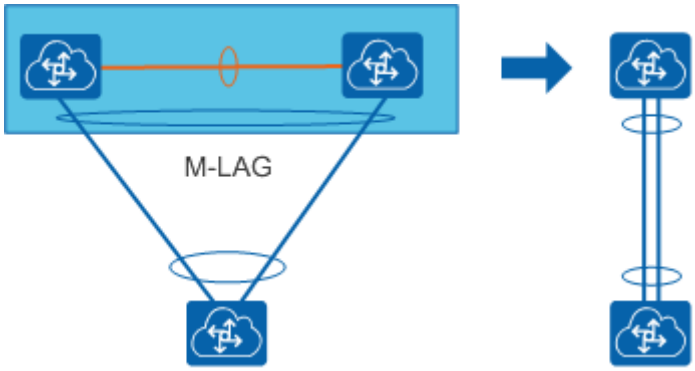
The network architecture should be selected based on site requirements. For example, the two-layer design is more suitable for application systems with high performance requirements. The three-layer or multi-layer design is more applicable to application systems with high security requirements.

Available Link Design

On traditional data center networks, the spanning tree protocol (STP) and virtual router redundancy protocol (VRRP) are typically used together to provide server access reliability. In addition, the servers connect to the networks through multiple NICs to enhance redundancy. These redundant links, however, can be used only when the active link is faulty. The usage of links and devices is low and the oversubscription ratio of the network is affected.

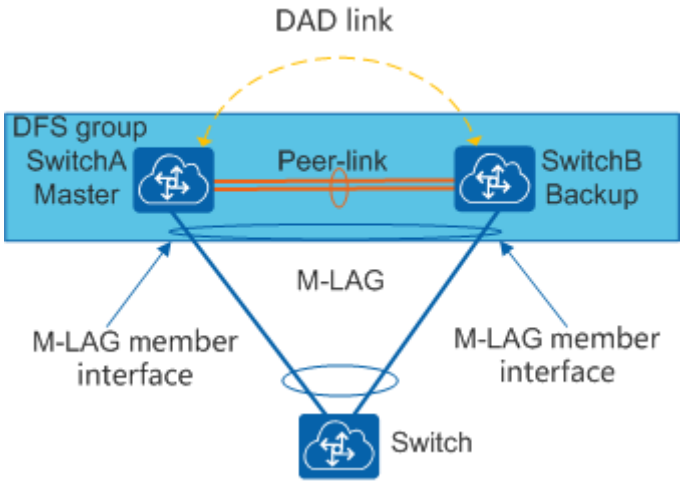
To address these problems, the use of M-LAG technology to construct data center networks is recommended. M-LAG is a horizontal virtualization technology that virtualizes two dual-homed devices into one, as shown in Figure 1-17. M-LAG prevents loops on a Layer 2 network. All the links where M-LAG member interfaces reside participate in traffic forwarding, which avoids link waste.

Figure 1-17 M-LAG network



When M-LAG is used, interface planning needs to be done. This is because certain interfaces must be reserved for the peer-link and dual-active detection (DAD) link in the M-LAG, as shown in Figure 1-18.

Figure 1-18 Interface usage in an M-LAG



For details about interfaces on M-LAG links, see Table 1-7.

Table 1-7 Description of interfaces on M-LAG links

Link Interface	Description
Peer-link	A peer-link is between two directly connected devices and has link aggregation configured. It is used to exchange negotiation packets and transmit part of traffic. After an interface is configured as a peer-link interface, other services cannot be configured on the interface. To improve the peer-link reliability, you are advised to use multiple links for aggregation.

Link Interface	Description
DAD link	A DAD link is used for M-LAG master and backup devices to exchange DAD packets at Layer 3.
M-LAG member interface	M-LAG member interfaces are the Eth-Trunks on M-LAG master and backup devices that are connected to the user-side host or switch. To improve reliability, you are advised to configure link aggregation in LACP mode.

The link design on the server access side is similar. When the two server NICs work in active/standby mode, the active link can be designed to work independently and the standby link is enabled only when the active link is faulty. When the two server NICs work in load balancing mode, all uplinks can be used. The design of link availability, combined with the design of the oversubscription ratio, can improve the actual available bandwidth on the network and network forwarding performance.

Device Model Selection

In the preceding examples, it is assumed that all the interfaces on all switches can forward packets at the line rate. If a switch cannot forward packets at the line rate, the oversubscription ratio on the switch needs to be considered. To ensure the high performance of the data center network, the use of switches with full line-rate forwarding capability is recommended.

Huawei CloudEngine series switches support line-rate forwarding. Table 1-8 uses the CE6870-48T6CQ-EI as an example to describe how to determine whether a switch supports line-rate forwarding.

Table 1-8 CE6870-48T6CQ-EI performance parameters

Item	CE6870-48T6CQ-EI
10GE BASE-T interface	48
10GE SFP+ interface	N/A
100GE QSFP28 interface	6
Switching capacity	2.16 Tbit/s or 19.44 Tbit/s

Table 1-8 lists that the CE6870-48T6CQ-EI has forty-eight 10GE interfaces and six 100GE interfaces. The following formula can be used to calculate the total bandwidth provided by all interfaces on the CE6870-48T6CQ-EI:

Total bandwidth = Number of interfaces x Interface speed x 2 (in full duplex mode)

If the total bandwidth is less than or equal to the switching capacity, the switch can be considered to have the line-rate forwarding capability from the perspective of switching capacity. As for the CE6870-48T6CQ-EI in the example, the following result can be obtained:

(48 x 10 Gbit/s + 6 x 100 Gbit/s) x 2 = 2160 Gbit/s (equals the 2.16 Tbit/s switching capacity)

Therefore, the CE6870-48T6CQ-EI supports line-rate forwarding. The CE6870-48T6CQ-EI also has a switching capacity of 19.44 Tbit/s in stacking scenarios. A maximum of nine CE6870-48T6CQ-EI switches can be stacked. In this case, the calculation formula is:

2.16 Tbit/s x 9 = 19.44 Tbit/s (equals the 19.44 Tbit/s switching capacity)

Therefore, when the CE6870-48T6CQ-EI switches are stacked, line-rate forwarding can also be implemented.

NOTE

This section uses fixed switches as an example. Modular switches are more complex when it comes to the calculation of line-rate forwarding. The LPUs of modular switches must work with the switch fabric units (SFUs) to implement line-rate forwarding. You can use the forwarding performance evaluation tool to select the required LPUs and SFUs.

The following sections describe the network oversubscription ratio design for the border leaf-spine-leaf architecture on data center networks.

1.2.1 Connecting Servers to Leaf Switches

Server access involves the selection of cabling modes and access switches. Common cabling modes include the TOR and EOR/MOR, as shown in Figure 1-19 and Figure 1-20. These two cabling modes can be used to form multiple deployment modes. In specific projects, different deployment modes or their combinations are determined based on the service deployment requirements.

Figure 1-19 TOR mode

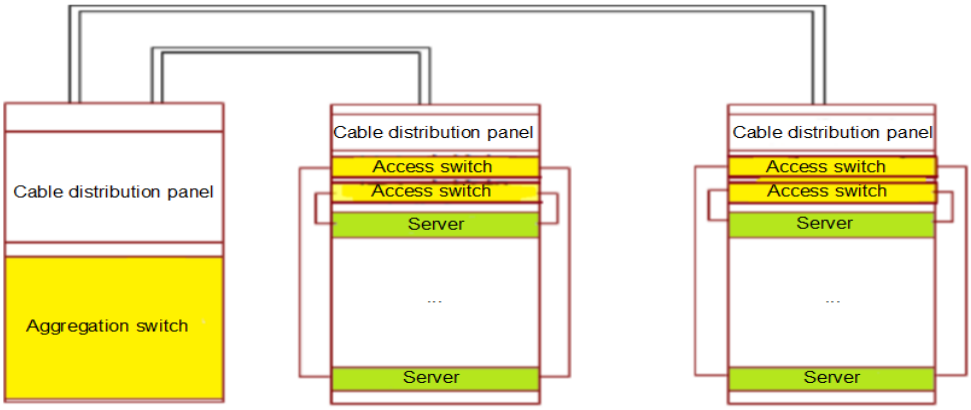


Figure 1-20 EOR/MOR mode

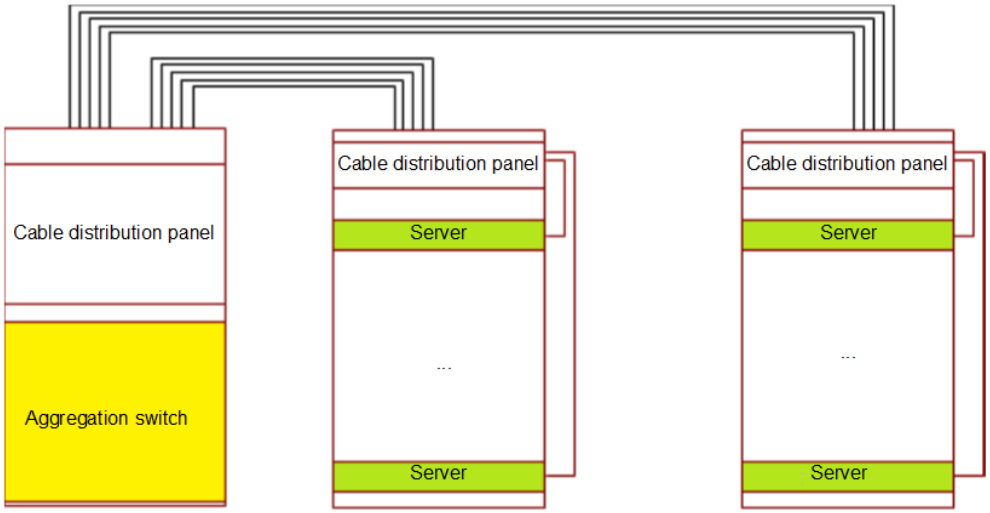


Table 1-9 compares the TOR access mode with the EOR/MOR access mode.

Table 1-9 Comparison of server access modes

Deployment Mode	TOR	EOR/MOR
Server type	1U, 2U, or 4U rack servers or blade servers (straight-through)	1U, 2U, or 4U rack servers or blade servers (switch module)
Usage scenario	High-density server cabinets	Low-density server cabinets or high-density blade server cabinets
Cabling	Simplified cabling between server cabinets and network cabinets	Complex cabling
Maintenance	A large number of access devices and complex network management and maintenance Simple cable maintenance and high scalability	Few access devices and simple maintenance Complex cable maintenance

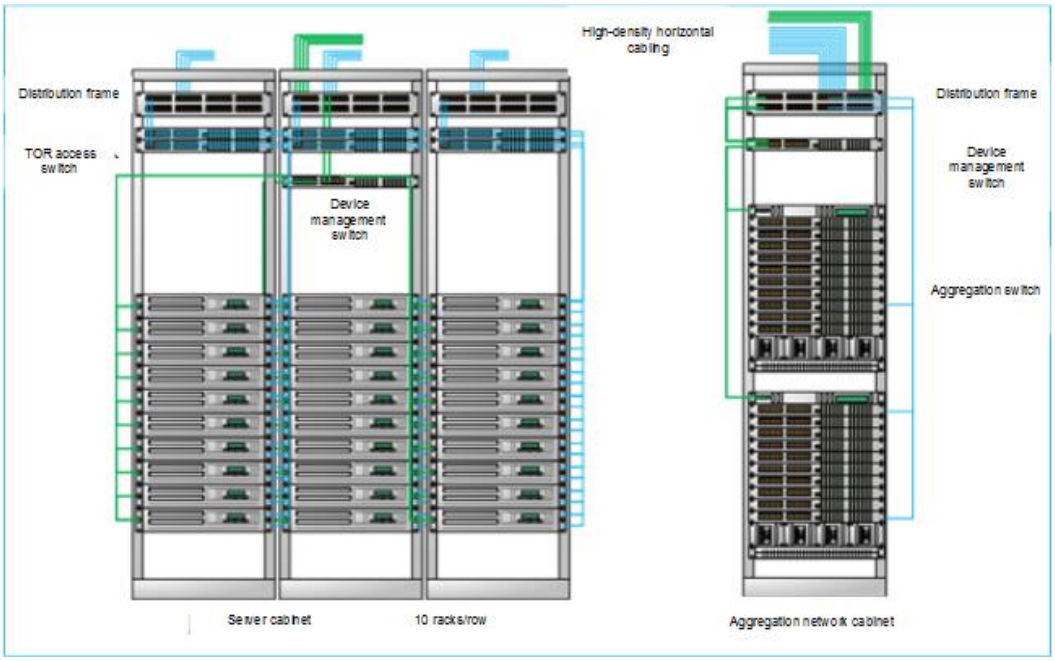
In most cases, servers are deployed in high density on data center networks. In these scenarios, the TOR access mode is commonly adopted. Table 1-10 describes the deployment roadmap for different servers in different scenarios.

Table 1-10 Typical server deployment scenarios

Server Density	Server Type	I/O Module Configuration	Service Interface	Storage Interface	Service Management Interface	Recommended Access Switch Deployment
Low-density partition	Rack server 2 U servers are used as an example.	Fixed GE/10GE modules with FC and GE/10GE interfaces	2 GE/10GE interfaces	0-2 FC or GE/10GE interfaces	0-2 GE/10GE interfaces	EOR/MOR/TOR
High-density partition	Rack server 2 U servers are used as an example.	Fixed GE/10GE modules with FC and GE/10GE interfaces	2 GE/10GE interfaces	0-2 FC or GE/10GE interfaces	0-2 GE/10GE interfaces	TOR
	Blade server 10 blades/chassis are used as an example. 8 U	Pass-through modules (Each interface is directly connected to the blade through the backplane.)	20 GE straight-through interfaces	0, 10, or 20 FC straight-through interfaces or GE straight-through interfaces	0, 10, or 20 GE interfaces	TOR The number of server interfaces must be less than that of the switch interfaces.
		Switching modules (A switch module is equivalent to an access switch.)	2-4 10GE interfaces	0-6 FC or GE/10GE interfaces	0-2 GE/10GE interfaces	EOR/MOR

The following section uses high-density rack servers as an example to describe the access solution. In Figure 1-21, 10 rack servers (usually 8 to 12 servers) are deployed in each rack, and 10 racks (usually 8 to 12 racks) are deployed in each row. Assume that the data center has four such rows of racks.

Figure 1-21 High-density rack server access solution



Each server has two 10GE service interfaces and one FE interface connecting to the BMC management interface. To ensure reliability, each server is connected to the network in M-LAG mode. Two adjacent racks form an M-LAG system. The TOR switches on the rack must support 200 Gbit/s bandwidth (10 Gbit/s x 10 x 2 = 200 Gbit/s).

According to experience values, the oversubscription ratio at the access layer is controlled at about 3:1. This ratio depends on the uplink bandwidth designed for the access switches. Currently, a single uplink interface of the access switch can reach 40 Gbit/s bandwidth. Theoretically, the 1:1 oversubscription ratio can be implemented using four such interfaces. Additionally, at least four aggregation switches need to be deployed at the spine layer, and an extra aggregation switch needs to be deployed each time an uplink interface is added. Therefore, the deployment costs for this design are high. In actual deployment, two aggregation switches are often configured. The access switch connects to the aggregation switch through two 40GE interfaces and provides 80 Gbit/s (40 Gbit/s x 2 = 80 Gbit/s) uplink bandwidth. In this way, a 2.5:1 (200 Gbit/s ÷ 80 Gbit/s = 2.5) oversubscription ratio is obtained, which is a desirable ratio as it is smaller than 3:1.

According to the preceding analysis, it is recommended that CE6870-24S6CQ-EI switches, shown in Figure 1-22, be used as TOR access switches. This switch has 24 10GE SFP+ Ethernet optical interfaces and six 40GE/100GE QSFP28 Ethernet optical interfaces. Its uplink interfaces also support 100GE optical modules. If no aggregation switch is added, the 1:1 oversubscription ratio can be supported by using 100GE interfaces. In addition, CE6870 series switches provide a large buffer of 4 GB to easily handle traffic surges on data center networks caused by video, search, and other applications. CE6870-24S6CQ-EI

Figure 1-22 CE6870-24S6CQ-EI

CE6870 series switches also provide 48 10GE interfaces to support high-density server access requirements. If large buffers are not required, CE6851-48S6Q-HI switches can also be used as TOR access switches.

NOTE

Here, the access of server and switch management interfaces is not considered. The management interfaces do not require large bandwidths. Generally, the access switches with low prices are used as long as they meet the interface access requirements, for example, Huawei S5700 series switches in MOR/EOR deployment mode.

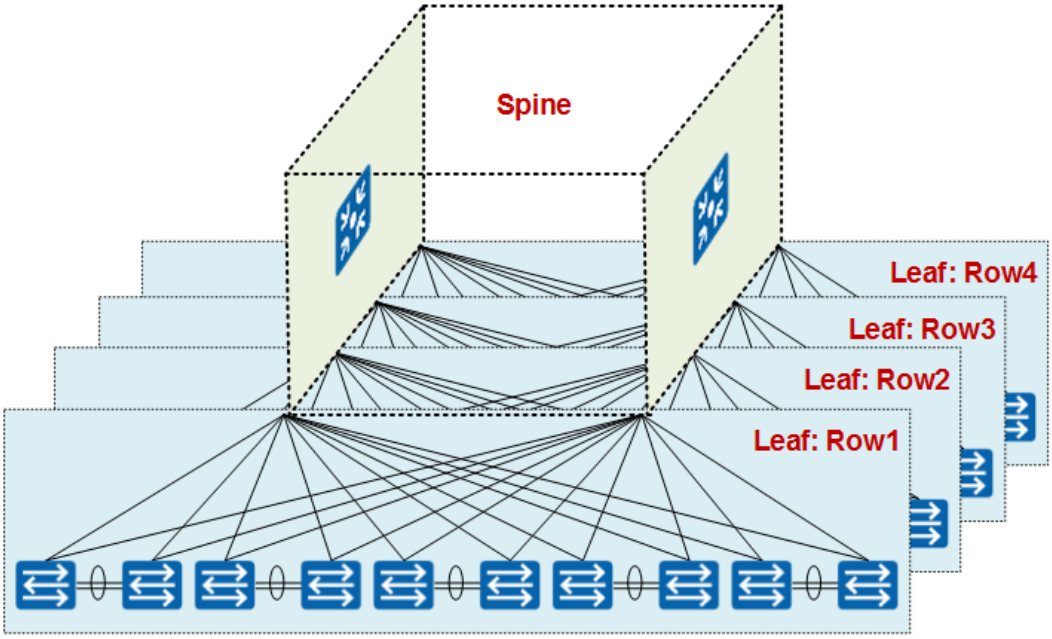
1.2.2 Leaf-Spine Access

This section focuses on the selection of switches at the spine layer. If ten TOR switches on each row of racks need to be connected to the aggregation switch and each TOR switch connects to the devices at the spine layer through two 40GE interfaces, there is a total of 80 interfaces (2 × 10 x 4 = 80) and 3200 Gbit/s bandwidth (80 x 40 Gbit/s = 3200 Gbit/s).

In this case, these interfaces can be divided into several spine nodes to access to the devices at the spine layer. To avoid the network problems caused by single point of failures (SPOFs), each spine node has at least two spine switches.

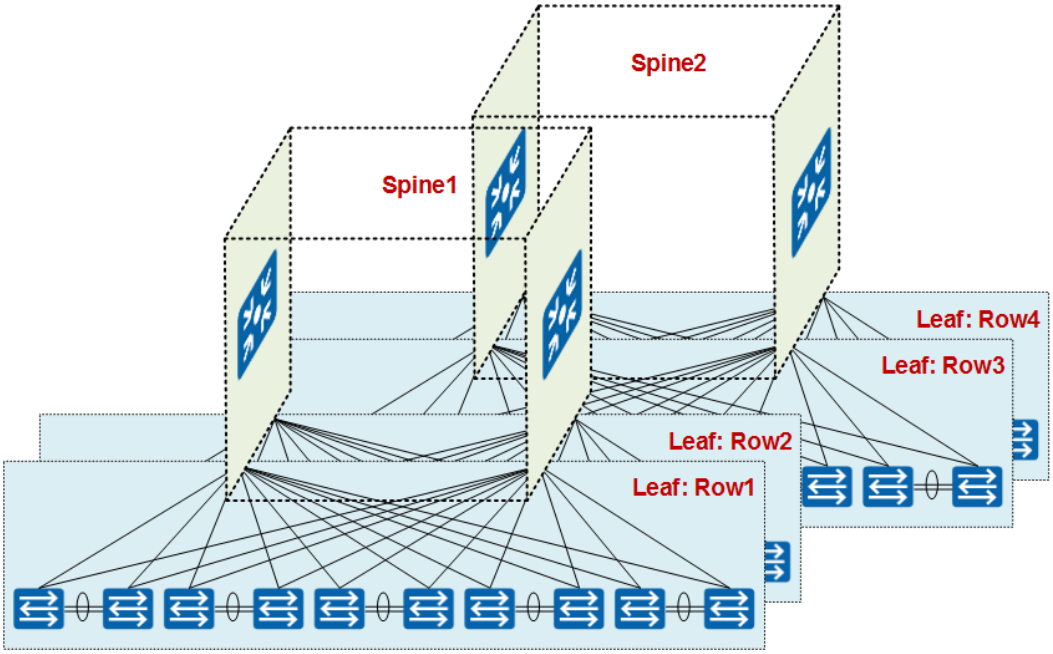
If these TOR switches are connected to one spine node, this node has 80 interfaces accessed and a total of 3200 Gbit/s bandwidth, as shown in Figure 1-23. In this case, if 100GE uplinks are used to connect to the devices at the border leaf layer (usually two devices with northbound interfaces connected to the egress routers), 400 Gbit/s bandwidth (4 x 100 Gbit/s = 400 Gbit/s) can be provided and the oversubscription ratio is 8:1 (3200 Gbit/s ÷ 400 Gbit/s = 8).

Figure 1-23 Access to one spine node



If the TOR switches are connected to two spine nodes, each node has 40 interfaces accessed and a total of 1600 Gbit/s bandwidth, as shown in Figure 1-24. In this case, if 100GE uplinks are used to connect to the devices at the border leaf layer, 400 Gbit/s bandwidth (4 x 100 Gbit/s = 400 Gbit/s) can be provided and the oversubscription ratio is 4:1 (1600 Gbit/s ÷ 400 Gbit/s = 4).

Figure 1-24 Access to two spine nodes



It should be noted that, on a data center network, the oversubscription ratio is not the only basis for dividing the spine nodes. Instead, they are mainly divided based on service and function requirements. In addition, the routing and address resolution protocol (ARP) specifications of the switches are limited, and virtual machines are applied at a large scale (a virtualization ratio of 1:30 or higher poses higher requirements on switch specifications). Therefore, the number of servers on the spine node is relatively small.

In this scenario, the two-spine-node design is adopted, that is, devices in Row1 and Row2 share a spine node, and devices in Row3 and Row4 share the other spine node. In this case, modular switches or high-performance fixed switches supporting flexible cards can be used as the spine switches.

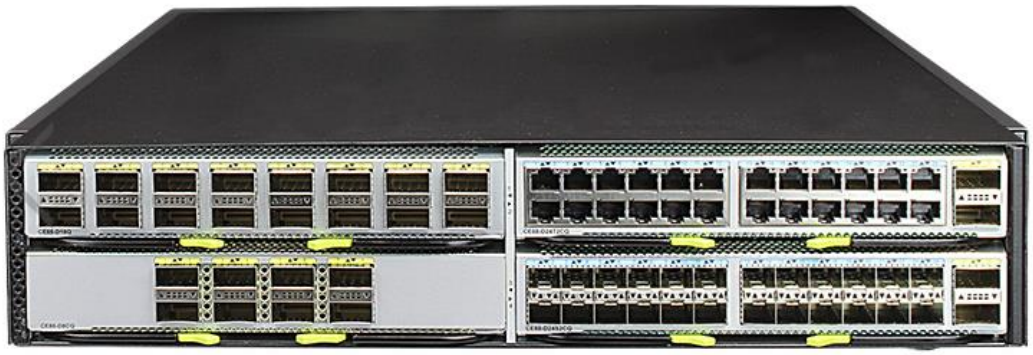
Here, the CE12804, shown in Figure 1-25, is recommended. The CE12804 provides four line processing unit (LPU) slots. CE12800 series switches can provide 4, 8, 12, or 16 LPU slots as required. Different cards can be selected flexibly, such as high-density 40GE cards and high-density 100GE cards. Both cards can provide up to 144 interfaces (36 x 4 = 144) of the corresponding bandwidth. The cards can be flexibly selected and combined based on the site requirements to facilitate subsequent adjustment or capacity expansion.

Figure 1-25 CE12804

Therefore, two 24-interface 40GE cards (CE-L24LQ-EC1) can be selected for each CE12804 to connect to the downstream devices and one 4-interface 100GE card (CE-L04CF-EF) can be selected for upstream connections.

In fact, there is a low-cost solution, that is, using the CE8860-4C-EI switches as spine switches. The CE8860-4C-EI, shown in Figure 1-26, is a 2 U switch with flexible cards and can house a maximum of four cards. Three 16-interface 40GE cards (CE88-D16Q) and one 8-interface 40GE/100GE card (CE88-D8CQ) can be selected. The CE88-D16Q cards are used for downstream connections and the CE88-D8CQ card is used for upstream connections.

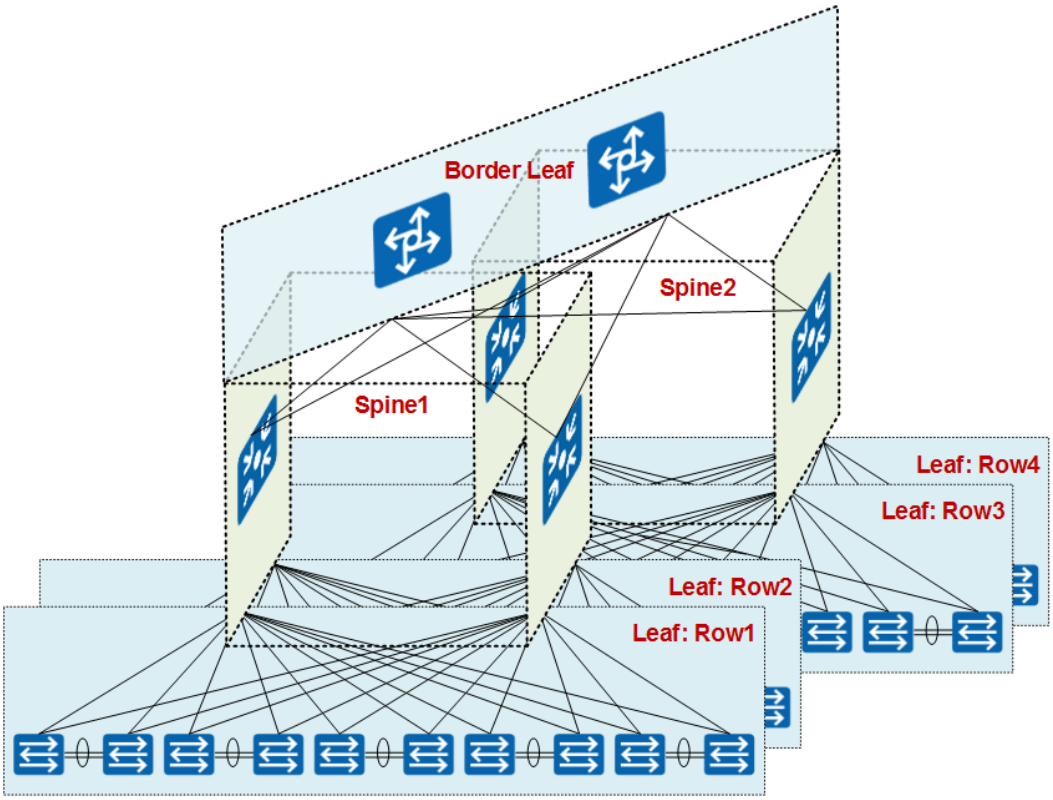
Figure 1-26 CE8860-4C-EI



1.2.3 Spine-Border Leaf Access

This section describes the design of the border leaf layer, shown in Figure 1-27. The northbound interfaces at the border leaf layer are mainly connected to the egress routers. The southbound interfaces are connected to the switches at the spine layer and are responsible for the east-west traffic forwarding at the spine layer. The interfaces of the egress routers purchased by customers are critical to the border leaf-layer design. These interfaces are more expensive than those on lower-layer network devices and are often 10GE or 40GE interfaces. This means that the oversubscription ratio of the border leaf layer is larger.

Figure 1-27 Spine-border leaf access

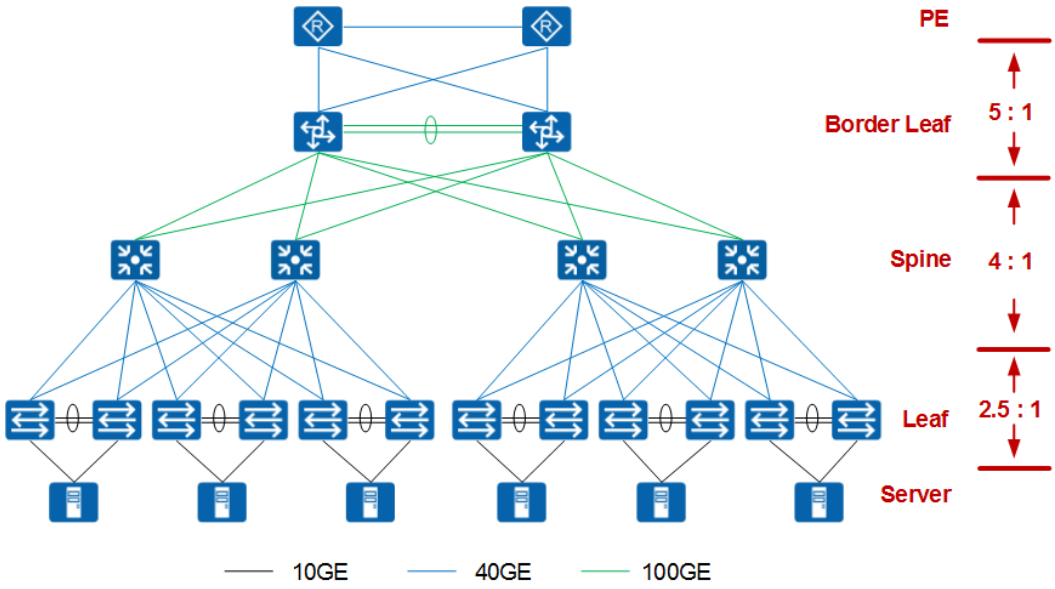


If four 40GE outbound interfaces are used, the bandwidth is 160 Gbit/s and the oversubscription ratio is 5:1 (800 Gbit/s ÷ 160 Gbit/s = 5). However, according to the current statistics, about 75% of traffic is generated within the data center, that is, the east-west traffic. The remaining 25% of traffic, or north-south traffic, is about 200 GB (800 GB × 25% = 200 GB). In this case, the oversubscription ratio is 1.25:1 (200 Gbit/s ÷ 160 Gbit/s = 1.25), which is acceptable.

Because the switches at the border leaf layer do not need to provide many interfaces, the CE6870-24S6CQ-EI can be used.

Figure 1-28 shows the logical diagram of the overall network oversubscription ratio design.

Figure 1-28 Logical diagram of the overall network oversubscription ratio design



1.3 Summary

The oversubscription ratio of a data center network must be designed based on network services and traffic models. Additionally, the volume and ratio of the east-west and north-south traffic must be thoroughly considered for the design of a suitable oversubscription ratio and selection of suitable devices. Based on experience values, you can refer to the following design:

- At the leaf layer to which the servers connect, the oversubscription ratio should be less than 3:1.
- At the spine layer, the oversubscription ratio is close to or smaller than that of the leaf layer.
- At the border leaf layer, the oversubscription ratio is relatively large and can be designed based on the egress router bandwidth on the customer side.

When the two-layer architecture is adopted, the east-west and north-south traffic volume at the spine layer is larger. Therefore, high-performance switches need to be used and the bandwidth for the interconnection between spine nodes must be increased.

Huawei CloudEngine series switches feature line-rate forwarding, high interface density, and large buffer capacity, making them the best choice for building high-performance networks with a low oversubscription ratio. For more information about the forwarding performance and interfaces of Huawei CloudEngine series switches, visit http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches.

Chapter 5: Fabric Network

1.1 Overview

The preceding chapters described how to deploy switches in a data center network, how to connect switches through cables, and how to calculate the traffic oversubscription ratio. These chapters helped you understand the physical architecture of Huawei CE series data center switches. The following chapter describes the logical architecture of CE series switches. You might think that the logical architecture just involves well-known mature technologies, such as the loop prevention protocols, IP addressing, as well as Layer 2 and Layer 3 forwarding mechanisms such as Layer 2 VLAN+xSTP and Layer 3 routing. This chapter provides more details about the data center network.

With the development of cloud computing, server virtualization technologies have been widely applied in data center networks. During server migration, IP addresses and running statuses of VMs must remain unchanged to ensure uninterrupted services. Therefore, dynamic migration of VMs can be performed only within the same Layer 2 domain, and not across the Layer 2 domains. This requires a large Layer 2 network. Traditional Layer 2 technologies reduce the scale of Layer 2 domains (for example, VLAN) to control broadcast storms or block redundant devices and links to eliminate loops (for example, xSTP). As a result, the number of supported hosts, convergence performance, and bandwidth efficiency of network resources on these networks cannot meet requirements of data center networks.

Huawei CE series switches can solve this problem. These switches provide a variety of Layer 2 expansion capabilities, including device virtualization (stack) and M-LAG. These capabilities are key to constructing logical data center networks. The following sections describe the advantages and disadvantages of these capabilities in different scenarios.

NOTE

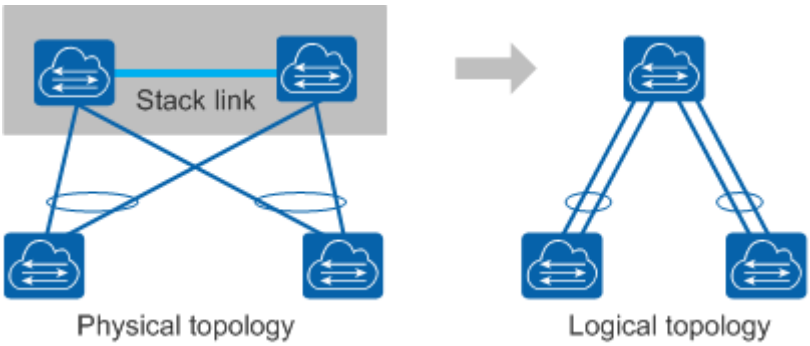
For more information about large Layer 2 network technologies in data centers, visit:

1.3 BGP.

1.2 Stack Scheme

A stack virtualizes multiple physical devices on the same layer into a single logical device through cables, simplifying network configuration and management. Together with the inter-device link aggregation technology, stack technology implements highly reliable backup for devices and links and prevents Layer 2 loops. Compared with traditional Layer 2 loop prevention technologies, stack technology provides a clearer logical topology and higher link efficiency.

Figure 1-29 Stack



Huawei CE series switches support cluster switch system (CSS) technology (stacking of modular switches) and intelligent stack (iStack) technology (stacking of fixed switches). CSS and iStack have the following characteristics:

1. Many-to-one virtualization: CSS or iStack virtualizes multiple physical devices into one logical device that has a unified control plane and provides unified management.

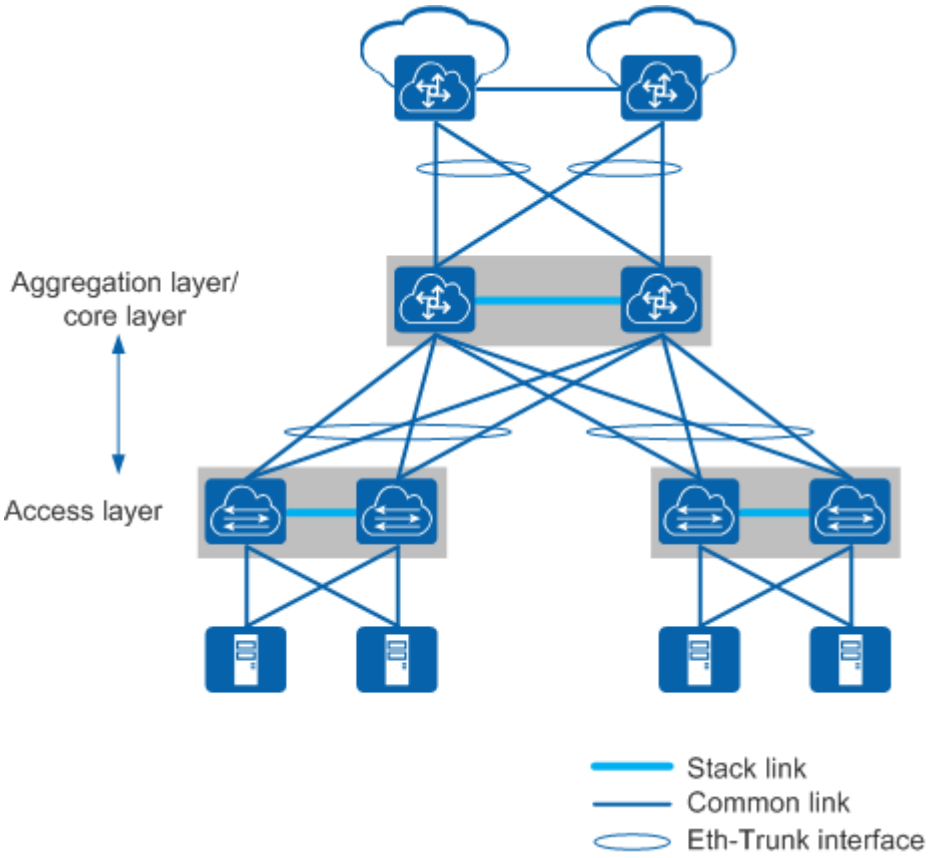
2. Unified forwarding plane: Physical devices in a CSS or an iStack use a unified forwarding plane that shares and synchronizes forwarding information in real time.

3. Inter-device link aggregation: Links between physical devices in a CSS or an iStack are aggregated into a single Eth-Trunk interface for interconnection with downstream devices.

In the data center network shown in Figure 1-30, stacks are deployed on the access, aggregation, and core layers to form a simple and loop-free network.

- On the access layer, low-cost CE8800&7800&6800&5800 series switches are deployed to set up a stack. Terminals, such as servers or other network devices, are dual-homed to the stack to ensure high reliability of access links.
- On the aggregation/core layer, high-performance CE12800 series switches are deployed to set up a stack, which connects to access devices through inter-chassis links to form a loop-free network.

Figure 1-30 Stacks in a data center



The following describes the deployment scheme:

- CSS or iStack technology ensures device reliability. When a device fails, the other one automatically takes over all services.
- CSS, LAG, and iStack technologies help build a reliable end-to-end architecture, ensuring continuous service availability for data centers.
- Multiple access switches form an iStack and two aggregation switches form a CSS.
- Access and aggregation switches are fully meshed through multiple 10GE or 40GE links, ensuring link reliability.
- Aggregation and core switches are fully meshed through high-speed 40GE links, ensuring non-blocking traffic forwarding between aggregation and core switches.

The scheme has the following characteristics:

- Simplified management and configuration

  After a stack is set up, multiple physical devices are virtualized into one logical device. You can log in to any member device to configure and manage all member devices, reducing the number of devices to be managed by the network by more than one half.

  Additionally, protocols such as xSTP and Virtual Router Redundancy Protocol (VRPP) do not need to be deployed, simplifying network configuration.
- High bandwidth efficiency

  Link aggregation ensures that bandwidth efficiency can reach 100%, while STP blocks congested links. Per-flow load balancing supports multiple load balancing modes.

- Fast convergence

  Link aggregation implements link failure convergence within milliseconds, eliminating the impact of link or node faults on services. STP, however, implements link failure convergence only within seconds.

- Flexible scalability

  When services increase, customers only need to add devices to networks without modifying network configurations. Network upgrades are performed smoothly and customer investments are protected.

Stack technology has been widely used in low-end fixed switches and high-end modular switches. It is very mature and stable. However, it still has its shortcomings:

- It just virtualizes devices on the same layer, implementing horizontal virtualization.

  Although horizontal virtualization improves the network structure and reduces the number of managed nodes, the number of network layers remains unchanged, and there are still many managed nodes. A large-scale data center has high-density access and a large number of access switches. To improve reliability, typically, multiple access switches are virtualized. In most cases, two access switches are virtualized into one logical switch. Therefore, after horizontal virtualization is implemented, there are still many managed nodes. If there are 40 access switches and every two access switches are virtualized, there are still 20 managed nodes.

- Unified control plane limits the number of physical nodes supported by the system.

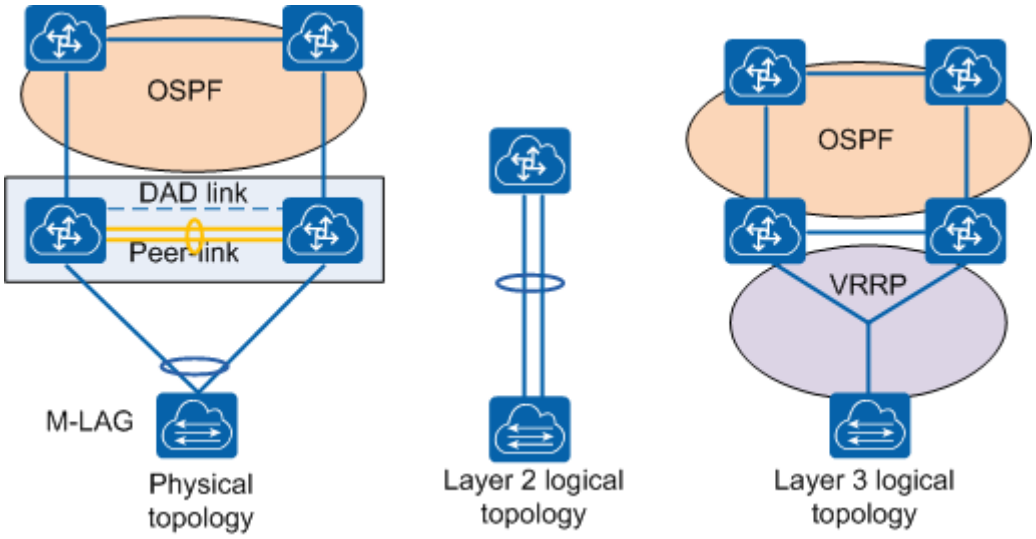
  After network devices have been virtualized into a stack system, control planes of these devices are integrated into one. As a result, MPUs of network devices work in active/standby mode. That is, only MPUs of the master device are working properly, while other MPUs are in standby state. Therefore, the number of physical nodes supported by the entire system is limited by the processing capability of the master node. For example, typically, 2 modular devices are virtualized into a stack system, and 16 fixed devices are virtualized into a stack system. Currently, a large-scale stack system supports a maximum of 10,000 to 20,000 hosts. This can meet access requirements of only small- and medium-sized data centers but not super-large data centers.

1.3 M-LAG Scheme

M-LAG implements link aggregation among multiple devices. One device is connected to two devices through M-LAG to provide device-level link reliability.

On Layer 2, M-LAG can be considered as a horizontal virtualization technology, which virtualizes two physical devices into a single Layer 2 logical device. M-LAG prevents loops on a Layer 2 network and implements redundancy, without performing laborious spanning tree protocol configuration. M-LAG greatly simplifies the network and configuration. Compared with the traditional xSTP loop prevention mechanism, M-LAG provides a clearer logical topology and higher link efficiency.

Figure 1-31 M-LAG-based physical topology and Layer 2 as well as Layer 3 logical topologies



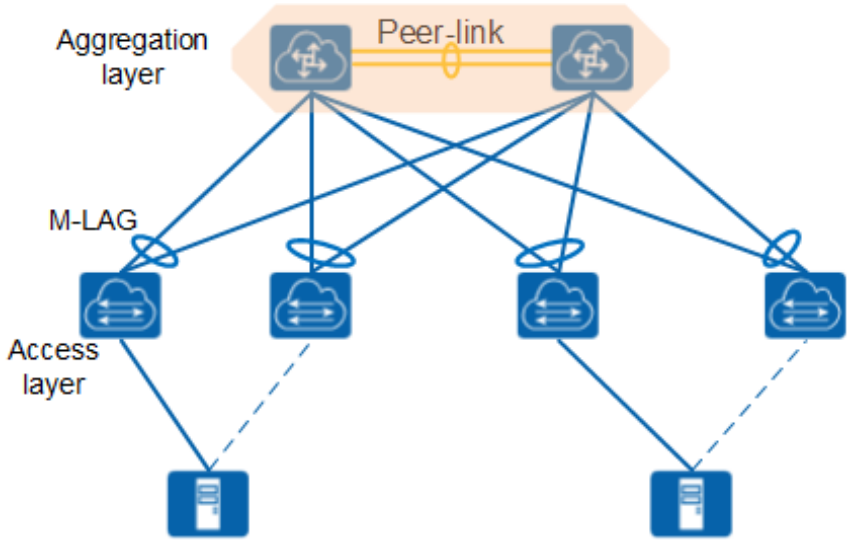
In Figure 1-31, the two M-LAG switches provide M-LAG interfaces for Layer 2 service access. A peer-link is configured between the two switches to exchange M-LAG packets and forward horizontal service traffic between the switches. In the Layer 3 logical topology, the two switches are two independent devices, can be managed by independent NMSs, and function as independent OSPF nodes. In addition, M-LAG supports preferential forwarding of local traffic, minimizing east-to-west traffic between the two switches. M-LAG supports dual-active detection (DAD). The two M-LAG devices are independent, so the in-band or out-of-band IP reachability detection can be used for DAD, without requiring additional cabling.

Deployment Scheme

- **Scheme 1: Aggregation switches set up an M-LAG system.**

  M-LAG enables a loop-free logical network between aggregation and access switches, making STP deployment unnecessary. Two aggregation switches are configured to set up an M-LAG system and the link between the switches is configured as a peer-link. The ports on the two aggregation switches connected to the same access switch set up an inter-chassis Eth-Trunk.

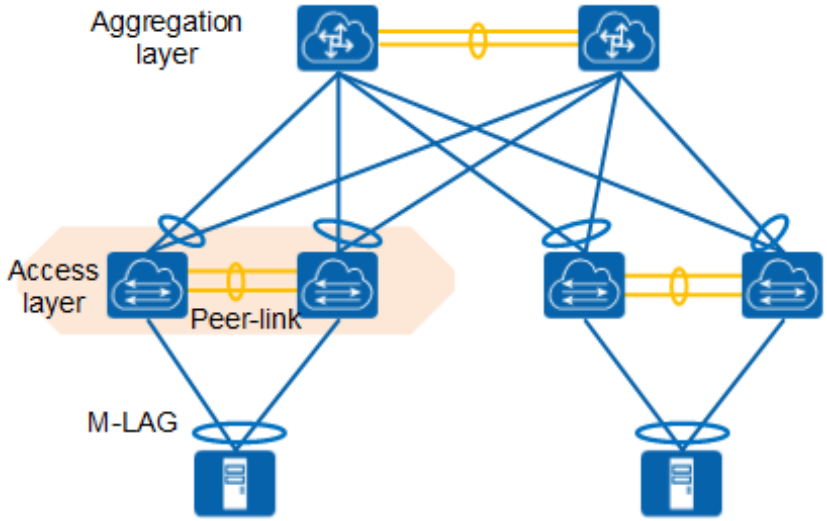
Figure 1-32 Deploying M-LAG on aggregation switches



M-LAG provides a clearer logical topology and higher link efficiency than the traditional STP loop prevention mechanism. M-LAG devices have independent control and management planes and use the same protocol for information synchronization. M-LAG offers higher reliability than stack. Additionally, M-LAG devices can be upgraded independently, facilitating device maintenance.

- **Scheme 2: Access switches set up an M-LAG system.**

  M-LAG also applies to scenarios where a server is dual-homed to two access switches with two NICs working in active-active mode. When a server is dual-homed to two access switches, the two NICs on the server use the same MAC address and implement flow-based load balancing. Therefore, in an M-LAG system, the ports on the two access switches connected to the server are configured as an Eth-Trunk, and MAC addresses and ARP entries of the two ports are synchronized between the two ports. Deploying M-LAG on access switches



Scheme Characteristics

M-LAG technology is essentially a control plane virtualization technology. Unlike stack technology, M-LAG only needs to synchronize information related to interfaces and entries,

not all device information. This makes the control plane coupling of M-LAG much looser than that of stack. Some defects in stack technology are eased in M-LAG, including the three major issues facing stack, as follows:

- Reliability issues: M-LAG does not need to synchronize all device information but only some protocol plane information, ensuring higher reliability than stack.
- Maintenance issues: Two M-LAG devices can be upgraded independently. Only the protocol planes are coupled, shortening the service interruption time.
- Limited scalability: M-LAG aims to solve the access-side multipath problem and typically works with routing or some large Layer 2 technologies to implement network-side multipath forwarding.

1.4 Chapter Summary

Table 1-11 summarizes the virtualization technologies discussed in this chapter.

Table 1-11 Advantages and disadvantages of stack and M-LAG

Virtualization Technology	Stack	M-LAG
Advantages	• Simplified management and configuration • High bandwidth efficiency • Fast convergence • Flexible scalability, maximizing the return on investment (RoI)	• Inter-device link aggregation, ensuring high reliability • Traffic load balancing, providing high link efficiency • Independent control and management planes, ensuring simple device upgrade
Disadvantages	• Only devices on the same layer can be virtualized. • A single control plane creates reliability, maintenance, and scalability problems.	The problem of network scalability cannot be resolved, and routing or other large Layer 2 technologies are required to implement network-side multipath forwarding.

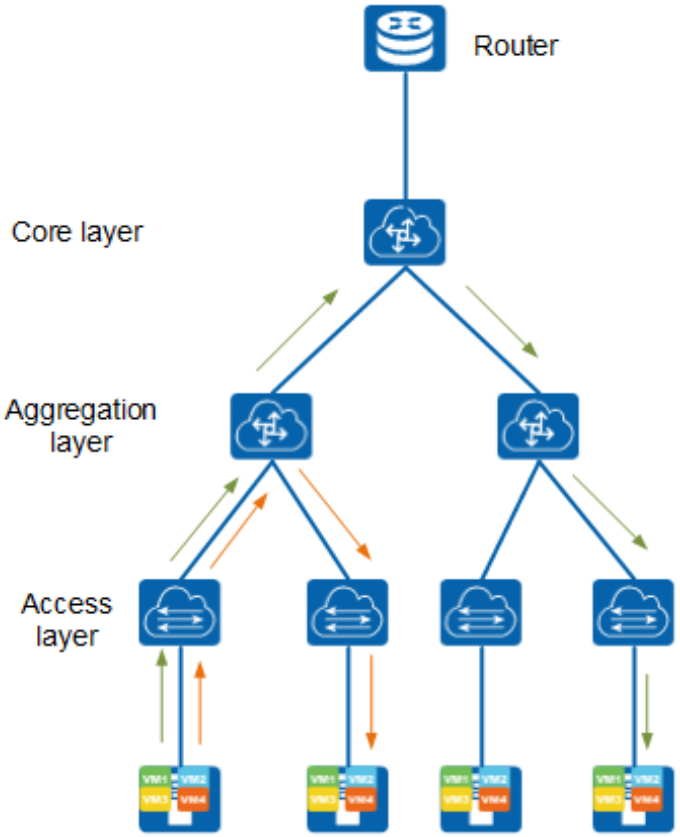
Now, you should understand requirements of large-scale Layer 2 data center networks. Huawei CE series switches have used a variety of technologies and solutions to meet these requirements and support future development of data center networks. Stack virtualizes multiple devices into a single logical device on the control plane. If a physical device needs to be virtualized into multiple logical devices, CE series switches use virtual system (VS) technology, which virtualizes a physical system (PS) into multiple isolated logical systems. Each VS functions as an individual physical device to process services. For more information about VS technology, visit: http://forum.huawei.com/enterprise/en/thread-406591.html.

Chapter 6: IP Fabric & Layer 3 Routing

1.1 IP Fabric

Until several years ago, most data center networks were still implemented based on the traditional three-layer architecture. This architecture came from the campus network design. Figure 1-33 shows a standard traditional three-layer network structure.

Figure 1-33 Traditional three-layer network traffic model



This traditional three-layer architecture is practical for most traffic models, including campus networks with southbound and northbound configurations. In addition, it is widely used, mature, and stable. However, with the development of technologies, this architecture no longer satisfies the needs of data center networks.

The data center network transmission mode keeps changing, with most traditional networks using the vertical (north-south) transmission mode. In this mode, a host communicates with

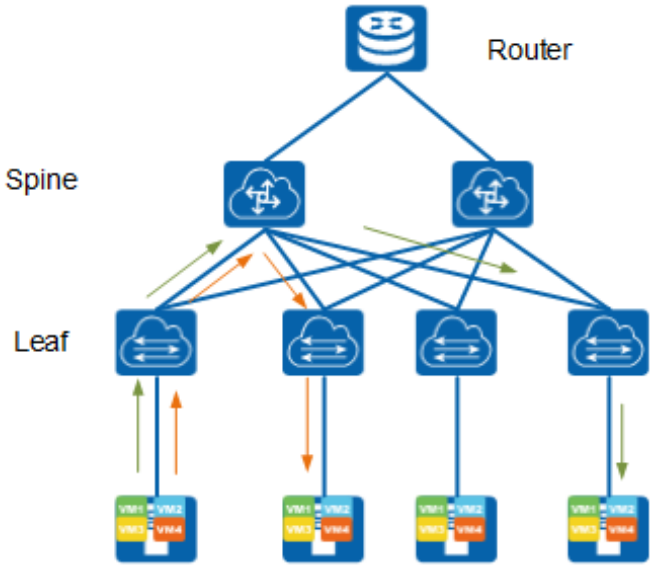
the hosts on different network segments in the network through switches and routers. Hosts on the same network segment are typically connected to the same switch and can communicate with each other directly.

As cloud computing develops, horizontal (east-west) traffic dominates data centers, involving almost all cloud computing, virtualization, and Big Data services. The horizontal traffic model has become a bottleneck for data transmission on vertically designed networks. This is because data must pass through many unnecessary nodes (including routers and switches). Inter-host access traffic needs to pass through many uplink ports, greatly degrading the transmission performance. The original three-layer network design worsens the degradation of performance. This means that the current mainstream three-layer network architecture can no longer meet requirements of data center networks.

In Chapter 5, the flat fabric network was described. The whole fabric network is a two-layer network, in which network loops and multi-path forwarding can be resolved using device virtualization technologies including stack, and the inter-device link aggregation technology M-LAG. However, this network supports a limited number of devices and has low scalability. For this reason, the IP fabric network has been introduced.

What is an IP fabric network? IP fabric is an overlay tunneling technology that is based on IP networks. Figure 1-34 shows IP fabric networking that is based on the fat-tree spine-leaf topology.

Figure 1-34 Two-layer IP fabric



In this type of networking, communication between any two servers involves a maximum of three devices, and each spine node and leaf node are fully meshed. This networking facilitates network scaling simply by increasing the number of spine nodes. Traffic can be transmitted between server nodes in almost all data center structures by traversing a certain number of switches. This architecture consists of multiple high-bandwidth direct links, avoiding network transmission slowdowns caused by network bottlenecks and supporting high forwarding efficiency and low latency.

Technologies, IP routing, VXLAN, or Transparent Interconnection of Lots of Links (TRILL), can be deployed between spine and leaf nodes based on service requirements.

- IP routing between spine and leaf nodes

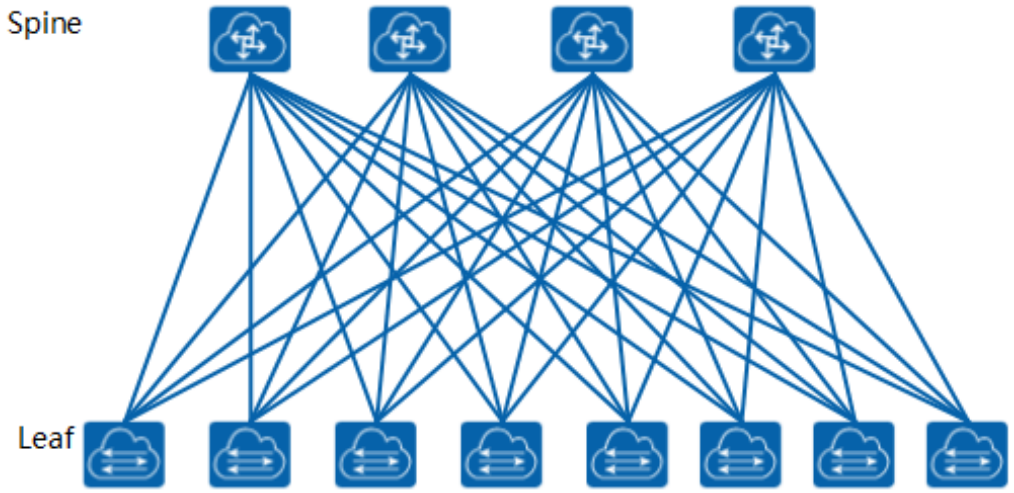
Layer 3 to edge typically applies to collaborative computing services, including the search service. These services have a small traffic oversubscription ratio (1:1 to 2:1) and require a highly efficient, non-blocking network.

- Deploy VXLAN or TRILL between spine and leaf nodes.

  Large Layer 2 network applies to data center networks that require large-scale resource sharing or VM migration.

An IP fabric network facilitates simplified network expansion, with expansion limited only by the number of supported devices and their ports, as shown in Figure 1-35.

Figure 1-35 Spine-leaf network architecture



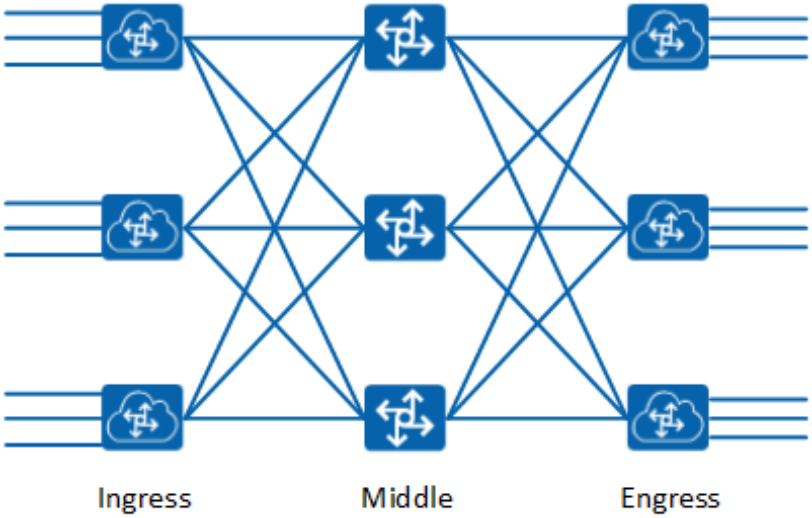
In Figure 1-35, there are four spine devices. Each leaf device has four uplinks, each of which is connected to a different spine device. In this topology, the maximum number of supported leaf devices is determined by the maximum number of ports supported by each spine device. If a spine switch supports 40*40GE connections, the maximum number of supported leaf devices will be 40. Because some ports of the spine switch will be used for uplink connections, it is preferable to set the maximum number of supported leaf devices to 36.

1.2 Spine-Leaf Network Architecture

The flat network architecture of spine and leaf devices comes from the Clos network. Clos networks are named after Charles Clos, a researcher in Bell Labs. He proposed the model in 1952, as a method to overcome the performance and cost challenges of the electromechanical switches then used in the telephone network. Clos used mathematical theory to prove that, if the switches were organized in a hierarchy, it was feasible to achieve non-blocking performance in a switching array (now known as a fabric). Before this, the non-blocking architecture could only be implemented in NxN crossbar mode.

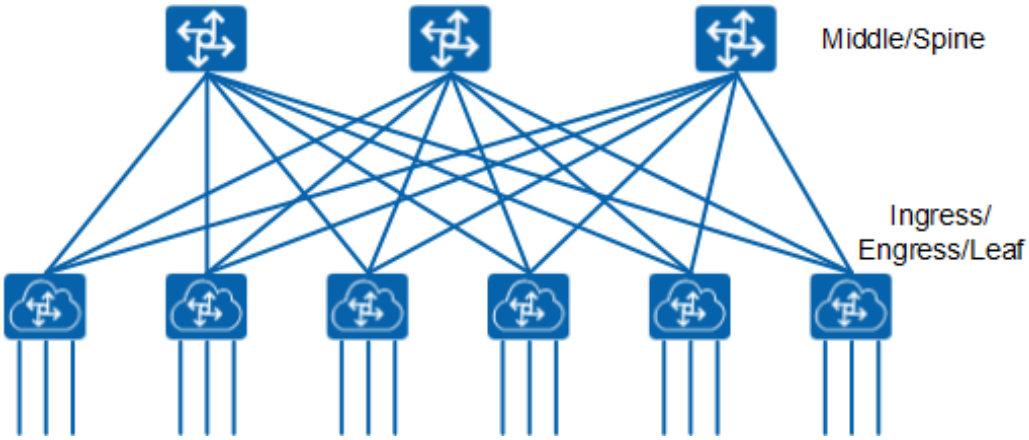
The Clos model proposed by Charles Clos was three-layer network architecture. In Figure 1-36, the three-layer Clos network architecture consists of one ingress node, one middle node, and one egress node.

Figure 1-36 Three-layer Clos network architecture model (with ingress, middle, and egress nodes)



If this architecture is folded in half and turned on its side, it becomes the spine-leaf architecture discussed previously, as shown in Figure 1-37.

Figure 1-37 Folded three-layer Clos network architecture model



The number of access connections is still equal to the number of connections between spine and leaf nodes in the folded three-layer Clos network architecture. As mentioned in the following sections of this chapter, traffic can be distributed on all available links without leading to traffic overload. As more connections are connected to leaf switches, the link bandwidth oversubscription ratio increases and can be reduced by increasing the link bandwidth between the spine and leaf switches.

In addition to supporting overlay technologies, the spine-leaf network architecture provides a more reliable networking because the spine and leaf nodes are fully cross-connected and the failure of a single switch at any layer does not affect the entire network structure. Therefore, the failure of one switch at any layer will not cause the entire structure to become invalid.

1.3 BGP

In section 1.1 IP Fabric, some technologies, including basic Layer 3 routing protocols and large Layer 2 technologies TRILL and VXLAN, can be used in IP fabric networks.

TRILL technology prepends extra frame headers to Layer 2 packets and uses route calculation to control data forwarding on the entire network. TRILL prevents broadcast storms on redundant links and implements Equal Cost Multipath (ECMP) routing. In this way, the Layer 2 network scale can be expanded to the entire network, without being limited by the number of core switches. However, TRILL uses Intermediate System-to-Intermediate System (IS-IS) to calculate and synchronize the network topology on the control plane, which adds the network complexity. In addition, the encapsulation and decapsulation of original packets also reduce the overall forwarding efficiency, and the TRILL protocol can only be handled by new chips. Therefore, to implement this technology, existing devices must all be replaced, so the investment cost is high. This is the primary reason why TRILL has not been widely implemented. To learn more about TRILL technology, visit 1.3 BGP.

VXLAN, as a typical overlay network technology, will be described in the next chapter. Now, we will focus on traditional Layer 3 routing protocols. Layer 3 is the control plane of the network architecture and distributes routing information to all switches in the network. Many Layer 3 routing protocols are available, and the best practice is to support any of the three mainstream open standard protocols: OSPF, IS-IS, or BGP. All routing protocols can advertise route prefixes on the network; however, these protocols support different networking scales, have different implementation methods, and provide different functions.

Both OSPF and IS-IS use the flooding technology to send update packets and other routing information. An area can be created to limit the number of flooded packets. However, this method does not take advantage of the Shortest Path First (SPF) routing protocol. BGP is created by group and supports a lot of prefixes and peers. BGP ensures network security, flexibility, stability, reliability, and efficiency. The Internet and most carriers use the BGP protocol as the routing protocol at the control layer.

1.3.1 BGP Basics

A network is divided into different autonomous systems (ASs) to facilitate network management. The Exterior Gateway Protocol (EGP) is used to dynamically exchange routing information between ASs. However, EGP advertises only reachable routing information without selecting optimal routes or considering loop prevention. Therefore, EGP cannot meet network management requirements.

BGP was designed to replace EGP and performs the following functions: BGP selects optimal routes, prevents routing loops, transmits routing information efficiently, and maintains a lot of routes. Although BGP is used to transmit routing information between ASs, it is not the best choice in some scenarios. For example, on the egress connecting a data center to the Internet, static routing instead of BGP is used for communication with external networks to prevent a huge number of Internet routes from affecting the data center internal network.

As an exterior routing protocol on the Internet, BGP has been widely used among Internet service providers (ISPs).

BGP has the following characteristics:

- Unlike an Interior Gateway Protocol (IGP), such as OSPF and Routing Information Protocol (RIP), BGP functioning as an EGP controls route advertisement and selects optimal routes between ASs rather than discovering or calculating routes.
- BGP uses TCP as the transport layer protocol, which improves BGP reliability.

  - BGP selects inter-AS routes, which requires high stability. TCP has high reliability and is used to enhance BGP stability.
  - BGP peers must be logically connected and establish TCP connections. The destination port number is 179 and the local port number is a random value.
- During route updates, BGP transmits only updated routes, greatly reducing bandwidth consumption. Therefore, BGP applies to the Internet where many routes need to be transmitted.
- BGP is a distance-vector routing protocol and prevents routing loops.
  - Between ASs: BGP routes carry information about the ASs along the path. The routes that carry the local AS number are discarded, thereby preventing inter-AS loops.
  - Within an AS: BGP does not advertise the routes learned in the AS to BGP peers in the AS, therefore preventing intra-AS loops.
- BGP uses routing policies to filter and select routes flexibly.
- BGP provides a mechanism for preventing route flapping, improving Internet network stability.
- BGP is easy to expand and adapts to the network development.

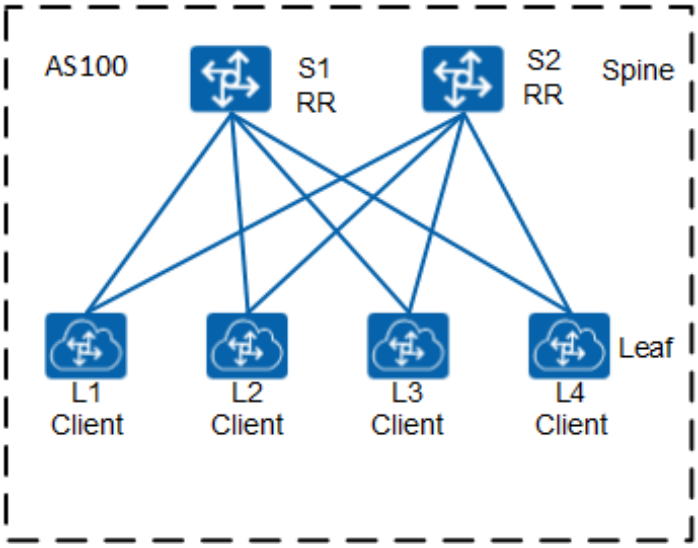
1.3.2 BGP Network Design

When deploying BGP, select between Internal BGP (IBGP) and External BGP (EBGP). Although IBGP and EBGP may have only slight differences, these slight differences may lead to significant changes in data center deployment. The biggest difference between IBGP and EBGP lies in the way they use ASs. This section compares IBGP and EBGP to explain how each switch allocates route prefixes and advertises routes.

IBGP

In IBGP, all spine and leaf switches reside in a single AS, as shown in Figure 1-38. In BGP, IBGP peers must be fully meshed to ensure connectivity between IBGP peers. Full-mesh is required in IBGP because full-mesh can prevent routing loops in IBGP. IBGP requires that IBGP speakers do not transmit the route prefixes learned from IBGP peers to other IBGP peers. Therefore, IBGP peers must be fully meshed. If there are n routers in an AS, n (n-1)/2 IBGP connections need to be established. If there are a lot of IBGP peers, many network resources and CPU resources will be consumed. BGP confederation or route reflector (RR) can be used between IBGP peers to solve this problem.

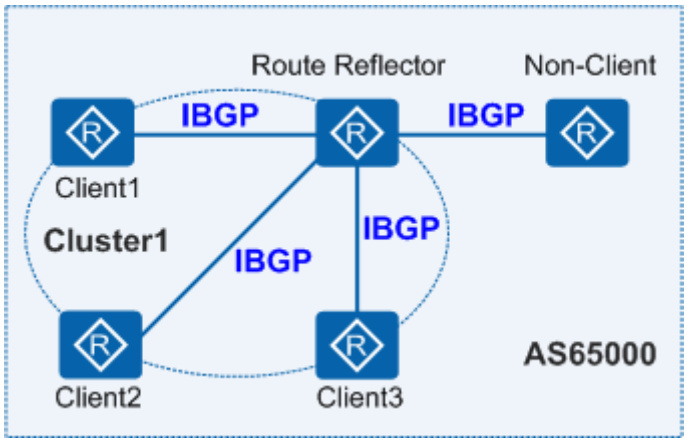
Figure 1-38 IBGP



A confederation divides an AS into sub-ASs. Full-mesh IBGP connections are established in each sub-AS. EBGP connections are established between sub-ASs. ASs outside a confederation still consider the confederation to be an AS. After the confederation divides an AS into sub-ASs, it assigns a confederation ID (the AS number) to each router within the AS. This brings two benefits: First, the original IBGP attributes are retained, including the Local_Pref attribute, MED attribute, and Next_Hop attribute. Second, the confederation-related attributes are automatically deleted when being advertised outside a confederation. Therefore, the administrator does not need to configure the rules for filtering information such as sub-AS numbers at the egress of a confederation.

An RR is an IBGP router which re-advertises routes to other IBGP routers. In Figure 1-39, the following roles are involved in RR scenarios in an AS.

Figure 1-39 RR



- RR: A BGP device that can reflect the routes learned from an IBGP peer to other IBGP peers. An RR is similar to a designated router (DR) on an OSPF network.
- Client: An IBGP device whose routes are reflected by the RR to other IBGP devices. In an AS, clients need to be directly connected to the RR only.

- Non-client: An IBGP device that is neither an RR nor a client. In an AS, a non-client must establish full-mesh connections with the RR and all the other non-clients.
- Originator: A device that originates routes in an AS. The Originator_ID attribute helps eliminate routing loops in a cluster.
- Cluster: A set of the RR and clients. The Cluster_List attribute helps eliminate routing loops between clusters.

An IBGP router cluster can be created and connected to an RR. Clients in a cluster need to exchange routing information only with the RR in the same cluster. Therefore, clients need to establish IBGP connections only with the RR. This reduces the number of IBGP connections in the cluster. However, the RR does not send each route. Instead, the RR sends only the optimal path to its peers. If there are multiple spine switches and multiple links between the spine and leaf switches, the link efficiency may be low. To solve this problem, BGP load balancing can be enabled on the BGP RR so that four equal-cost routes can be advertised to the leaf switches and traffic can be distributed based on ECMP.

Table 1-12 compares BGP confederation and RR in terms of configuration, connection, and application.

Table 1-12 Comparisons between an RR and a BGP confederation

RR	BGP Confederation
Retains the existing network topology and ensures compatibility.	Requires the logical topology to be changed.
Requires only an RR to be configured because clients do not need to know that they are clients of an RR.	Requires all devices to be reconfigured.
Requires full-mesh connections between clusters.	Does not require full-mesh connections between sub-ASs of a confederation because the sub-ASs are special EBGP peers.

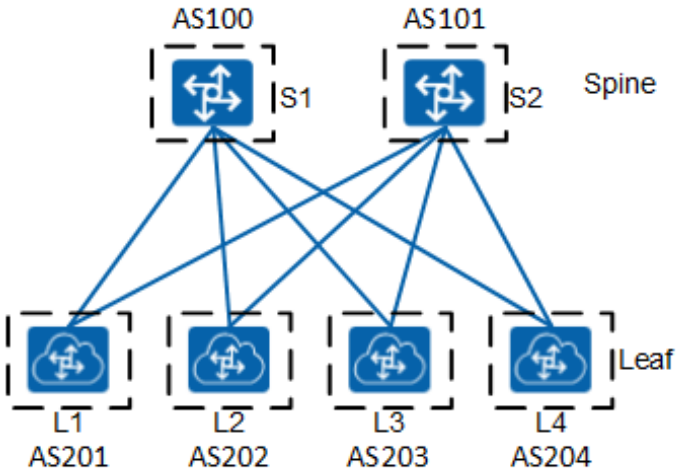
EBGP

In EBGP, each spine or leaf switch resides in a different AS, as shown in Figure 1-40. In contrast to IBGP, which has the routing loop prevention mechanism, EBGP does not have split horizon. EBGP prevents routing loops by using the AS_Path attribute. The AS_Path attribute records all of the ASs that a route passes through from the source to the destination in the vector order. To prevent inter-AS routing loops, a BGP device does not accept the routes whose AS_Path list contains the local AS number.

As discussed in the preceding sections, IBGP does not require physical connections between IBGP peers and requires only logical connections. EBGP peers generally require physical connections between EBGP peers.

The only challenge is the quantity of ASs used in the IP fabric network. Each switch has its own BGP AS number. The private AS number ranges from 64512 to 65535, where 1023 private AS numbers are available. If the IP fabric network has more than 1023 switches, the public AS numbers (not recommended in data centers) or 4-byte private AS numbers must be used. CE series switches support 4-byte private AS numbers in a range from 4200000000 to 4294967295 (or from 64086.59904 to 65535.65535).

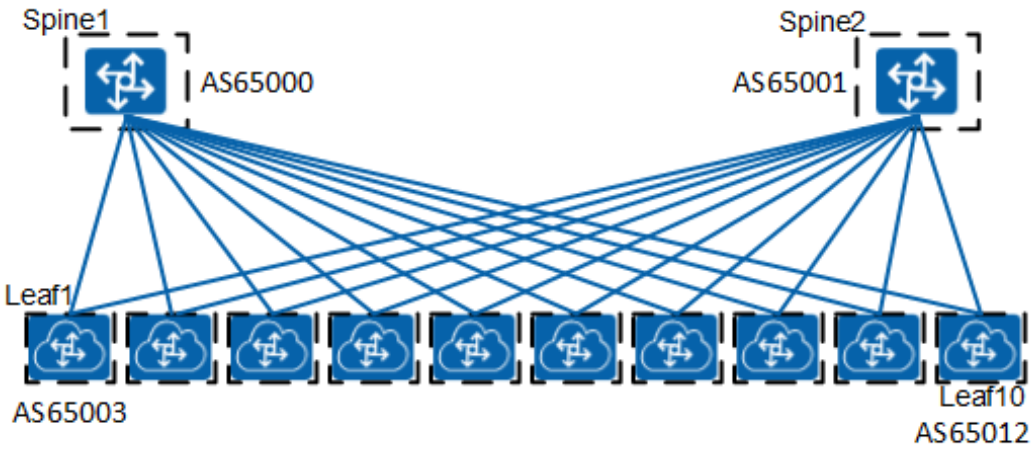
Figure 1-40 EBGP



BGP Application in the Data Center Network

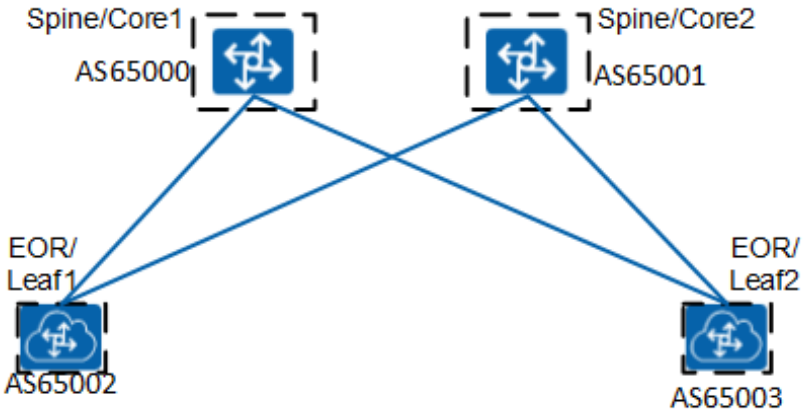
The following describes BGP in specific network architecture scenarios. In the DC1 scenario, there are five rows of racks, which are arranged into a spine-leaf network architecture. It is recommended that EBGP be used to establish a VXLAN underlay network. The network design when EBGP is used in each row of racks is shown in Figure 1-41. Each spine or leaf switch has its own AS number.

Figure 1-41 Running EBGP in the DC1 scenario



The EBGP design for the DC2 scenario, where EOR switches are deployed, is shown in Figure 1-42.

Figure 1-42 Running EBGP in the DC2 scenario



AS numbers AS65000 and AS65001 can be assigned to the first row of racks, and then the AS number of each device increases by 1 sequentially. This EBGP design is the same as that in Figure 1-41. If IBGP is selected for VXLAN overlay route exchange in a DC, the design will be simple because it requires only the allocation of the same AS number to all the devices in DC1. The same method can be used in DC2, as shown in Figure 1-43 and Figure 1-44.

Figure 1-43 Running IBGP in the DC1 scenario

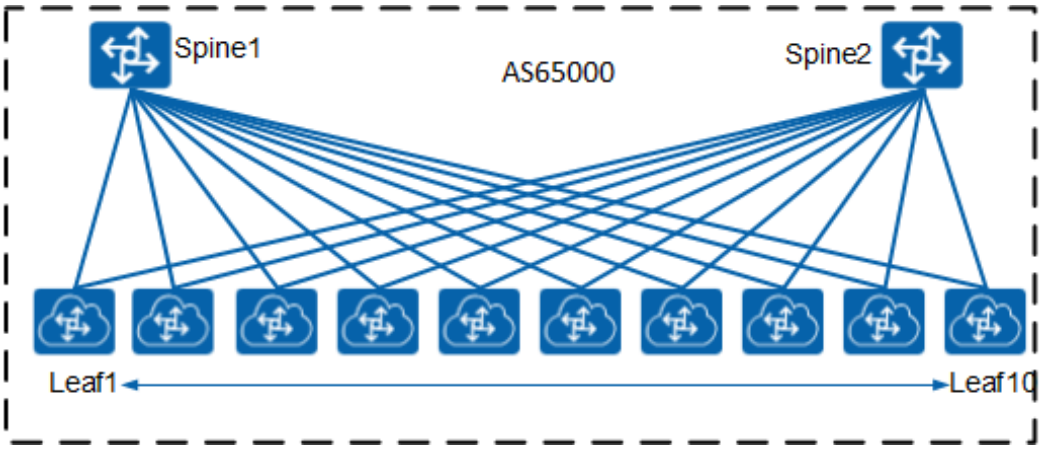
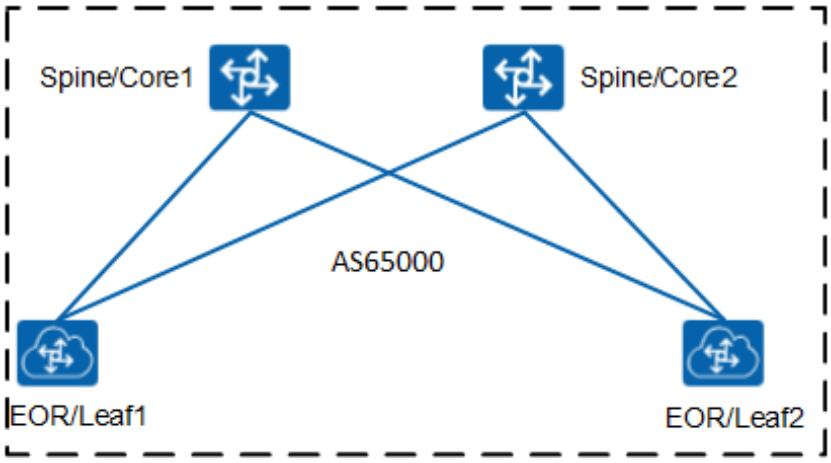


Figure 1-44 Running IBGP in the DC2 scenario



1.4 Chapter Summary

This chapter described the origins of the IP fabric network and the selection of Layer 3 forwarding routing protocols. In the flat spine-leaf network architecture, BGP can be used to construct a Layer 3 underlay network, which runs stably and has high scalability.

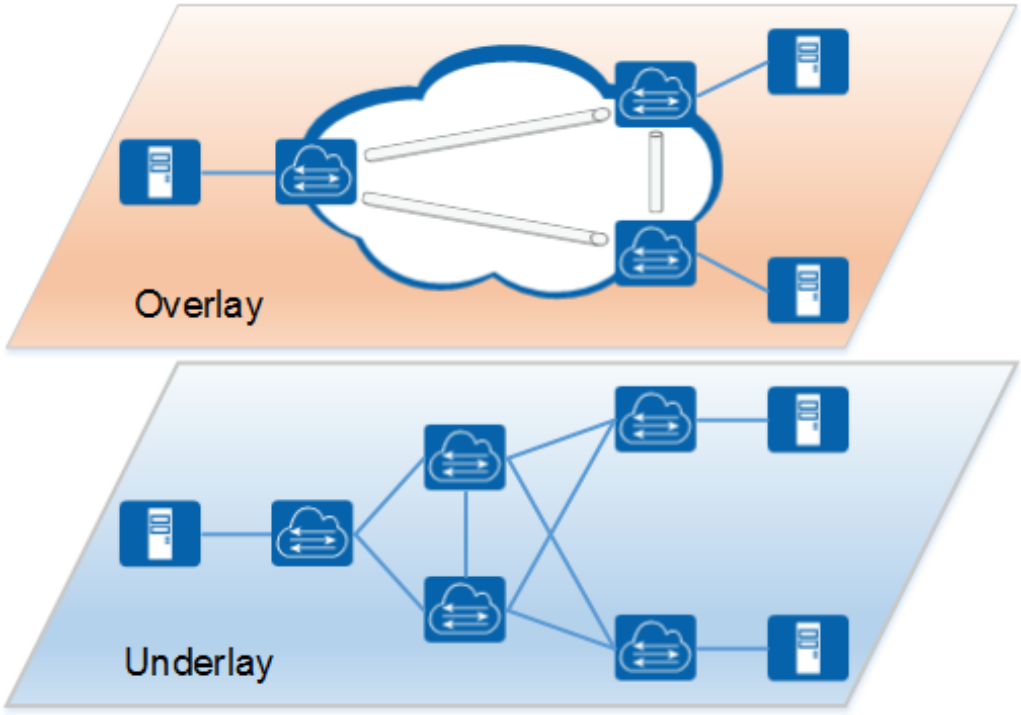
How can a large Layer 2 overlay network be constructed based on the underlay IP network? The following chapter will describe what overlay networks are, why they are needed, and how CloudEngine series switches use VXLAN technology to implement them.

Chapter 7: Overlay Networking

1.1 Overview

An overlay network is a logical network built over an existing physical network (underlay network) to virtualize network resources, as shown in Figure 1-45.

Figure 1-45 Overlay networking



The control and forwarding planes of an overlay network are independent from those on the underlay network. For a terminal (for example, a server) connected to an overlay network, the underlay network is transparent, thereby separating the bearer network from the service network.

Why is the overlay network needed?

As one of the core cloud computing technologies, server virtualization has been widely used in data centers. As enterprise services develop, the number of VMs grows rapidly and VM migration becomes a common service. However, this introduces the following problems to traditional networks:

- The VM scale is limited by network specifications.

  On a legacy Layer 2 network, data packets are forwarded at Layer 2 based on MAC address entries. However, the number of VMs is limited by the MAC address table capacity.

- Network isolation capabilities are limited.

  The current mainstream network isolation technology is VLAN. The VLAN ID defined in IEEE 802.1Q has only 12 bits and can represent only 4096 VLANs, which is insufficient for a large Layer 2 network that needs to isolate a large number of tenants or tenant groups.

- The VM migration scope is limited by the network architecture.

  To ensure service continuity during VM migration, the IP and MAC addresses of the VM must remain unchanged. Therefore, the service network must be a Layer 2 network and provide multipathing redundancy backup and reliability. Traditional technologies such as STP and device virtualization apply only to small-and medium-scale networks.

To address the preceding problems and meet the network capability requirements of cloud computing virtualization, overlay network technology is evolved. This technology solves the problems in the following ways:

- Addresses VM scale limitations imposed by network specifications

  The data packets sent by a VM are encapsulated in IP data packets, which are represented as encapsulated parameters on the network. This greatly reduces the requirements on MAC address specifications of large Layer 2 networks.

- Provides greater network isolation capabilities

  The overlay technology extends the number of isolation identifier bits to 24 bits, which greatly increases the number of tenants that can be isolated.

- Addresses VM migration scope limitations imposed by the network architecture

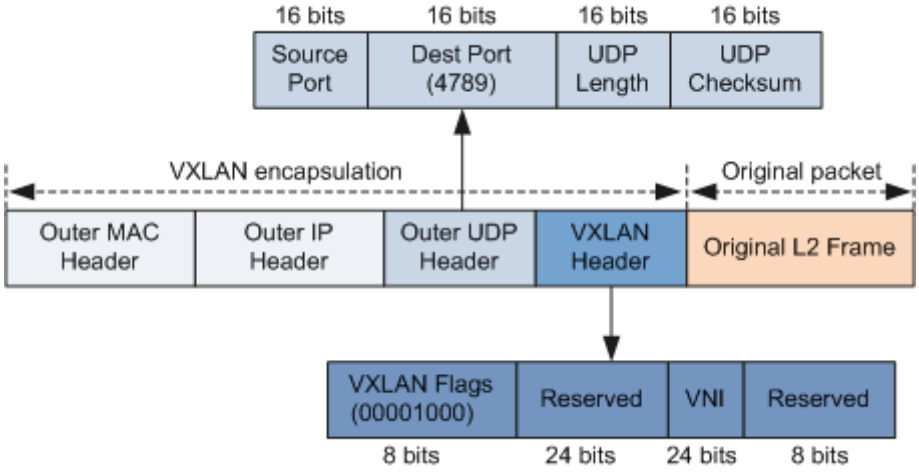
  An overlay network encapsulates Ethernet packets into IP packets and transmits them over routes. The routing network frees VM migration from the limitations imposed by the network architecture. In addition, the routing network provides high scalability, self-healing capability, and load balancing capability.

Overlay technologies include Virtual Extensible LAN (VXLAN), Network Virtualization using Generic Routing Encapsulation (NVGRE), and Stateless Transport Tunneling Protocol (STT), of which VXLAN is the most widely recognized.

1.2 VXLAN

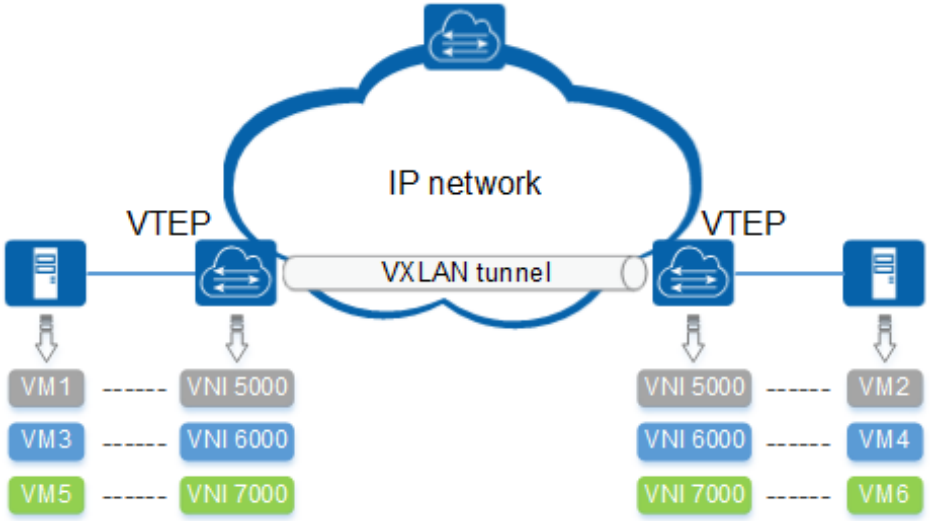
VXLAN is a Network Virtualization over Layer 3 (NVO3) technology defined by IETF and adopts the MAC-in-UDP packet encapsulation mode. In Figure 1-46, a virtual tunnel end point (VTEP) adds a VXLAN header to the original packet, which is then encapsulated into a UDP header. Finally, an outer IP header and an outer MAC header are added to the packet. The packet is then forwarded in accordance with standard Layer 2 and Layer 3 forwarding processes on the bearer network.

Figure 1-46 VXLAN packet format



The following figure shows the overlay network architecture based on VXLAN technology.

Figure 1-47 VXLAN network model



- VTEP

  A VTEP is an edge device on a VXLAN network. It is the start or end point of a VXLAN tunnel and is responsible for encapsulating and decapsulating VXLAN packets. VTEPs can be deployed on access switches or vSwitches (virtual switches on servers).

- VNI

  A VXLAN Network Identifier (VNI) is a network identifier similar to a VLAN ID and is used to identify a VXLAN Layer 2 network. A VNI represents a VXLAN segment. VMs in different VXLAN segments cannot communicate with each other at Layer 2.

- VXLAN tunnel

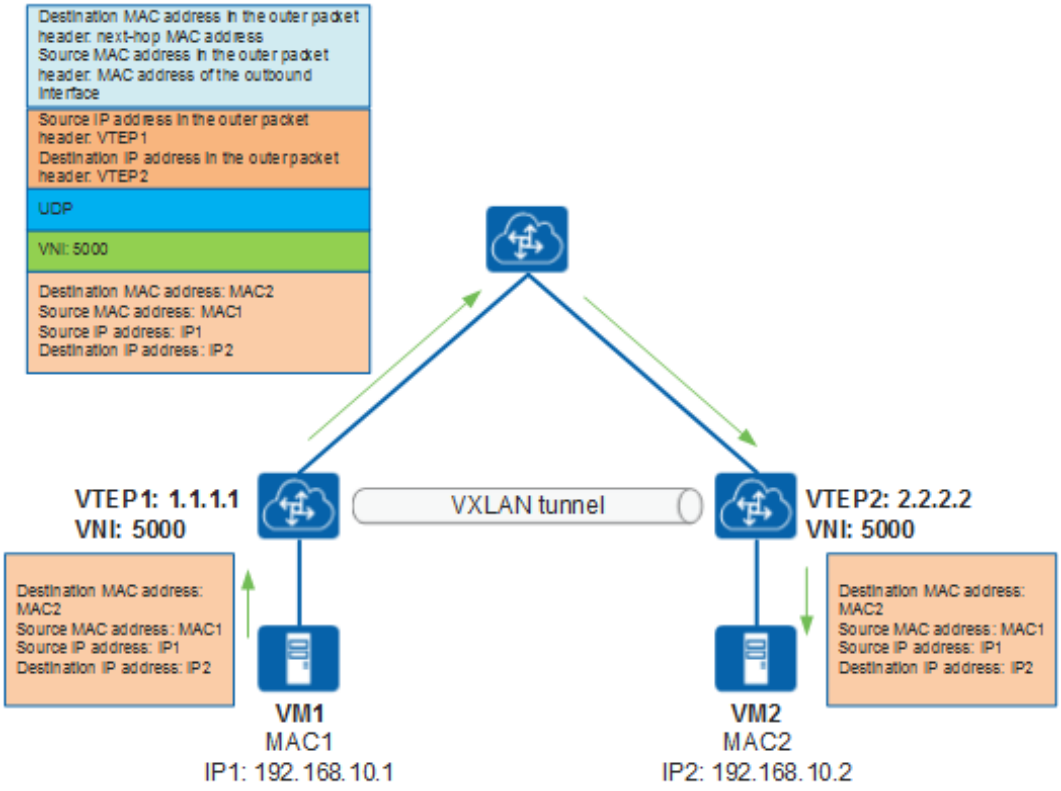
  A VXLAN tunnel is a logical tunnel established between two VTEPs for transmitting VXLAN packets. Service packets are encapsulated with VXLAN, UDP, and IP headers

(in that order) in the VXLAN tunnel. They are then transparently forwarded to the remote VTEP at Layer 3. The remote VTEP decapsulates the packets received.

VXLAN Packet Forwarding Process

The following describes the packet forwarding process on a VXLAN network for VMs on the same subnet.

Figure 1-48 VXLAN packet forwarding process



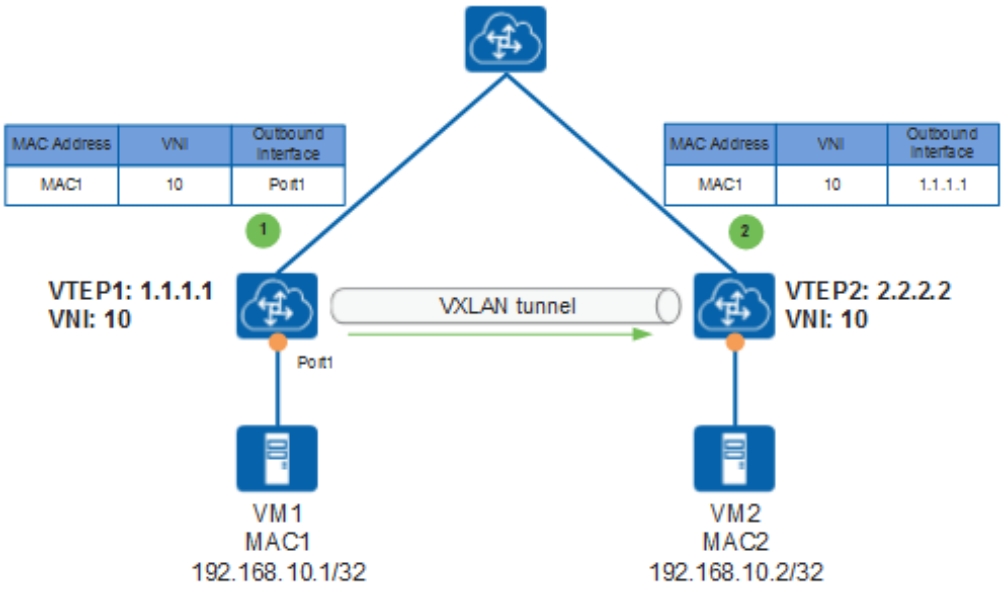
1. VM1 sends a packet destined for VM2.
2. After receiving the packet, VTEP1 performs VXLAN encapsulation. The IP address of VTEP2 is the destination IP address in the outer IP header added to the packet. VTEP1 transmits the encapsulated packet to VTEP2 through the IP network based on the outer MAC address and IP address of the packet.
3. VTEP2 decapsulates the received packet, obtains the original packet sent by VM1, and forwards the packet to VM2.

1.3 Layer 2 MAC Address Learning and BUM Packet Forwarding

On a VXLAN network, VMs on the same subnet can communicate with each other by querying their MAC address tables. In the following figure, VM1 sends a packet to VM2 through VTEP1. VTEP1 needs to learn the MAC address of VM2 from the packet.

Because no control plane is defined in the original VXLAN standard, the learned host MAC addresses cannot be transmitted between VTEPs. However, VXLAN has a MAC address learning mechanism similar to that of traditional Ethernets. After receiving a VXLAN packet, a VTEP records the source VTEP IP address, VM MAC address, and VNI to the local MAC address table. If the VTEP receives a packet in which the destination MAC address is the MAC address of the VM, the VTEP can implement VXLAN encapsulation and forwards the packet.

Figure 1-49 MAC address learning



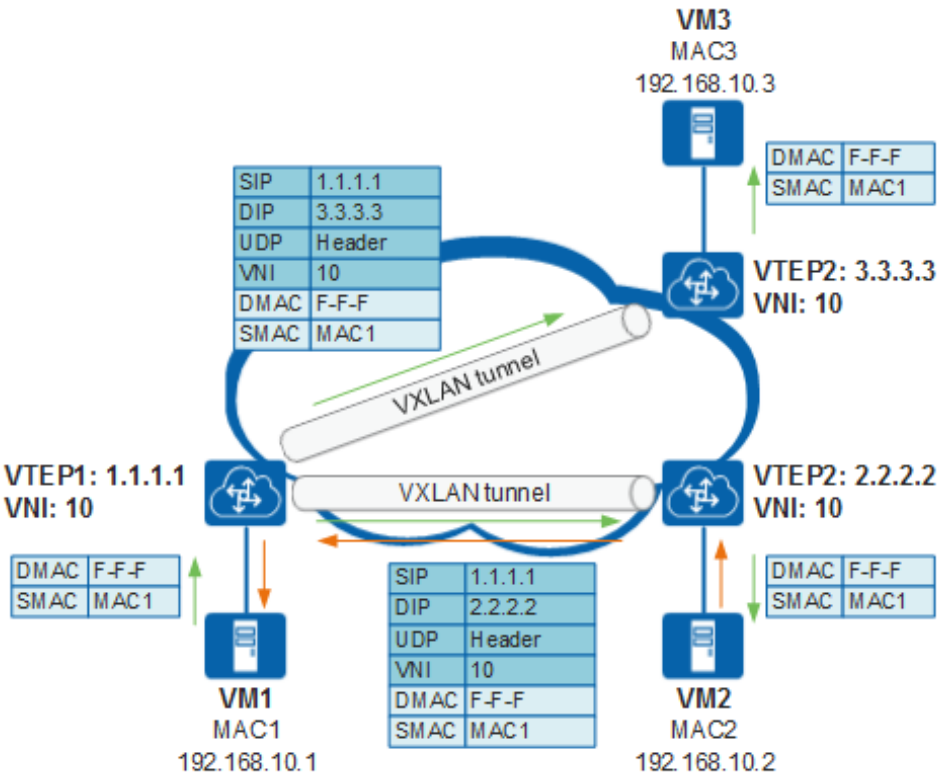
The following describes the process in which VTEP1 and VTEP2 learn the MAC address of VM1 (the process of learning the MAC address of VM2 is similar):

1. VM1 sends a packet destined for VM2.
2. VTEP1 encapsulates the received packet and forwards it to VTEP2. Meanwhile, VTEP1 learns the MAC address, VNI, and inbound interface of VM1.
3. VTEP2 decapsulates the received packet and learns the MAC address, VNI, and inbound interface (VTEP1) of VM1.

BUM Packet Forwarding

The preceding process is about the forwarding of known unicast packets, and therefore does not apply to scenarios where the VTEP receives a broadcast, unknown unicast, or multicast (BUM) packet with an unknown address. VTEPs forward traffic through flooding, which is also how devices on traditional Ethernets forward BUM packets.

Figure 1-50 BUM packet forwarding



In the preceding figure, VM1 wants to send packets to VM2. Because VM1 does not have the MAC address of VM2, VM1 sends an ARP broadcast packet to request VM2's MAC address. The following describes how VM1 obtains the MAC address of VM2:

1. VM1 broadcasts an ARP packet to request the MAC address of VM2.
2. Because the received request packet is a broadcast packet, VTEP1 searches for information about all tunnels in the broadcast domain, encapsulates the packet based on the obtained tunnel information, and sends the packet to all tunnels. In this way, VTEP1 forwards the packet to VTEP2 and VTEP3, which are on the same subnet.
3. VTEP2 decapsulates the received packet, obtains the original ARP packet sent by VM1, and forwards the packet to VM2. VTEP3 processes the packet in the same way and forwards it to VM3.
4. VM2 and VM3 compare the destination IP address in the ARP request with the local IP address. VM3 finds that the destination IP address is not the local IP address and discards the packet. VM2 finds that the destination IP address is the local IP address and responds to the ARP request.

   Because VM2 has learned the MAC address of VM1 at this stage, the ARP reply packet is a known unicast packet. The subsequent forwarding process is the same as that of known unicast packets.
5. VM1 receives the ARP reply from VM2 and learns the MAC address of VM2. The subsequent forwarding process is the same as that of known unicast packets.

1.4 VXLAN Gateway Deployment

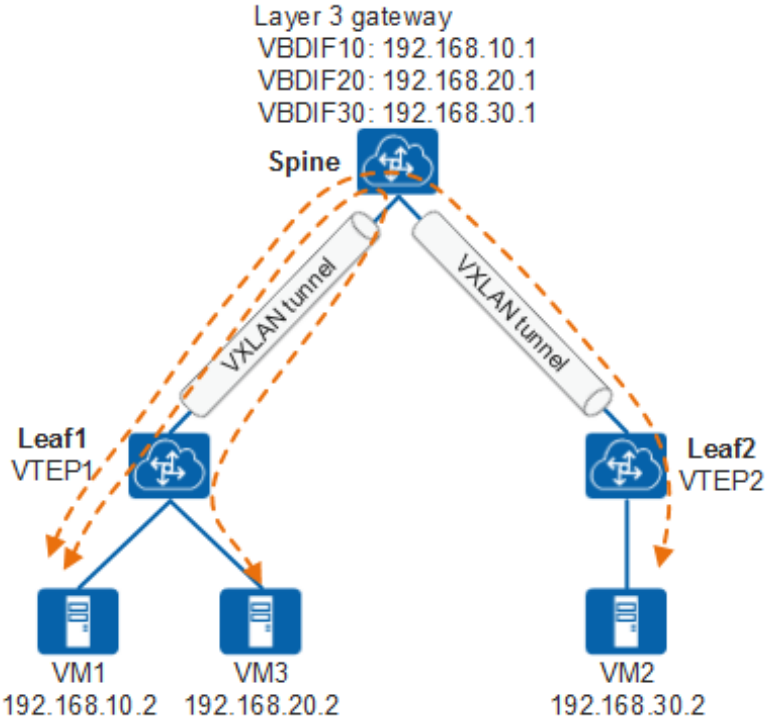
Different VLANs need to communicate with each other through Layer 3 gateways. VXLANs with different VNIs use a similar method to communicate with each other.

In the typical spine-leaf VXLAN networking, Layer 3 gateways can be deployed in centralized or distributed mode on the VXLAN, depending on their locations.

Centralized Gateway Deployment

In centralized gateway networking, Layer 3 gateways are centrally deployed on one spine node. In the following figure, traffic across subnets is forwarded through Layer 3 gateways to implement centralized traffic management.

Figure 1-51 Centralized gateway networking

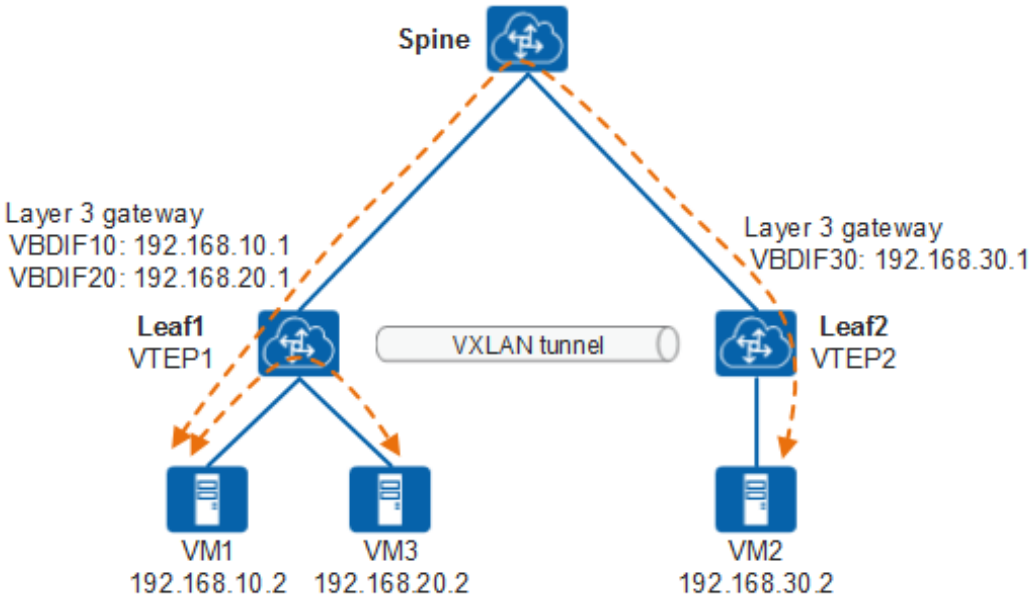


In centralized gateway networking, traffic on different subnets is centrally managed, and gateways are easy to deploy and manage. However, inter-subnet traffic on the same leaf node needs to be forwarded through the spine node, which means the traffic forwarding path is not optimal. Additionally, all terminal tenant entries forwarded through Layer 3 are generated on the spine node, which supports only a limited number of entries and may become a bottleneck as the number of tenants increase.

Distributed Gateway Deployment

In distributed VXLAN gateway networking, leaf nodes function as VTEPs of VXLAN tunnels as well as Layer 3 VXLAN gateways. Spine nodes are unaware of the VXLAN tunnels and only forward VXLAN packets.

Figure 1-52 Distributed gateway networking



VXLAN Layer 3 gateways can be deployed on a leaf node to implement inter-subnet communication on the same node. This allows traffic to be directly forwarded by the leaf node without passing through the spine node, conserving a large amount of bandwidth. A centralized Layer 3 gateway has to learn the ARP entries of all VMs on the network. In distributed gateway deployment, in contrast, the leaf node functioning as the gateway only needs to learn the ARP entries of VMs attached to it. This eliminates the bottleneck caused by limited ARP entry specifications in centralized Layer 3 gateway scenarios, improving network expansion capabilities.

In a distributed gateway scenario, a control plane is needed to transmit host routes between Layer 3 gateways, which is required for communication between VMs. The next chapter will explore how Ethernet VPN (EVPN) is applied as a control plane technology on VXLAN networks.

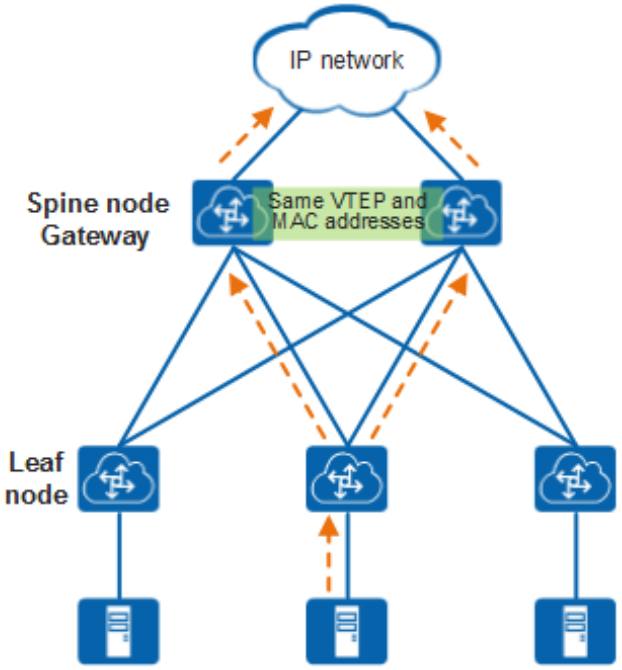
1.5 Active-Active Gateway

To ensure reliability, multiple gateways are deployed as backups on traditional networks. Similar to traditional networks, VXLAN networks also support active-active gateways on the overlay network.

Active-Active Gateway in the Centralized Gateway Deployment Scenario

In the typical spine-leaf networking, leaf nodes function as Layer 2 gateways and spine nodes function as Layer 3 gateways. Multiple spine nodes are configured with the same VTEP address and MAC address so that the nodes are virtualized into one VTEP. In this way, any spine node to which traffic is sent can provide the gateway service and forward packets to the next-hop device.

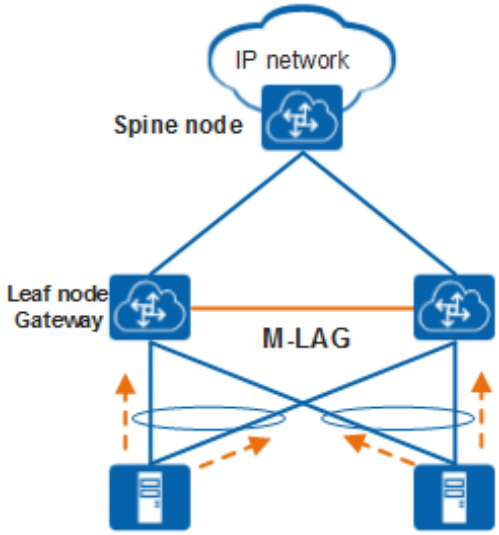
Figure 1-53 Active-active gateway in the centralized gateway deployment scenario



Active-Active Gateway in the Distributed Gateway Deployment Scenario

In the distributed gateway deployment scenario, the spine node functions as a transparent transmission device, and the leaf nodes function as Layer 3 gateways. If M-LAG is deployed on the leaf nodes, active-active gateways are implemented. That is, servers can be connected to multiple leaf nodes in active-active mode.

Figure 1-54 Active-active gateway in the distributed gateway deployment scenario



1.6 Summary

This chapter describes the VXLAN concepts and forwarding processes of VXLAN packets. VXLAN can construct a large Layer 2 network on top of an existing Layer 3 IP network. VXLAN can be deployed on physical switches or vSwitches residing on servers. Physical switches that support VXLAN can function as VTEPs, providing high processing performance and supporting the communication between non-virtualized physical servers. vSwitches functioning as VTEPs have low network requirements and do not require the switches to support VXLAN. However, the processing performance of a vSwitch is lower than that of a physical switch.

The CE12800, CE8800, CE7800, and CE6800 series switches (except the CE6810LI, CE6810EI, and CE6850EI) support VXLAN. In the typical spine-leaf architecture, you are advised to use CE12800, CE8800, or CE7800 switches as spine switches and use CE6800 switches as leaf switches.

As discussed in the preceding sections, because early VXLAN networks do not have a control plane, VXLAN learns MAC addresses through flooding. The next chapter will explore EVPN as a VXLAN control plane technology and how it implements automatic VXLAN tunnel establishment and MAC route learning.

Chapter 8: BGP EVPN

1.1 EVPN Overview

In the initial VXLAN solution (RFC7348), the control plane is not defined. The VXLAN tunnel is manually configured, and host addresses are learned through traffic flooding. This method is easy to implement, but it creates a lot of flooding traffic and makes network expansion difficult.

To solve these problems, EVPN is introduced as the VXLAN control plane. By referring to the BGP/MPLS IP VPN mechanism, EVPN defines several types of BGP EVPN routes by extending BGP. It advertises routes on the network to implement automatic VTEP discovery and host address learning.

Using EVPN as the control plane offers the following advantages:

- VTEPs can be automatically discovered and VXLAN tunnels can be automatically established, simplifying network deployment and expansion.
- EVPN can advertise Layer 2 MAC addresses and Layer 3 routing information at the same time.
- Flooding traffic is reduced on the network.

1.2 BGP EVPN Route Types

Traditional BGP-4 uses Update packets to exchange routing information between peers. An Update packet can advertise a type of accessible routes with the same path attributes, which are placed in Network Layer Reachability Information (NLRI) fields.

BGP-4 can manage only IPv4 unicast routing information, so Multiprotocol Extensions for BGP (MP-BGP) was developed to support multiple network layer protocols, such as IPv6 and multicast. MP-BGP extends NLRI fields on the basis of BGP-4. After extension, the description of the address family is added to the NLRI field to differentiate network layer protocols, such as the IPv6 unicast address family and VPN instance address family.

Similarly, EVPN defines a new sub-address family, that is, the EVPN address family in the L2VPN address family, and also introduces EVPN NLRI. EVPN NLRI defines the following types of BGP EVPN routes. After the routes have been advertised between EVPN peers, VXLAN tunnels can be automatically established and host addresses can be learned. The route types are as follows:

- Type 2 route — MAC/IP route: This is used to advertise the MAC address, ARP entry, and routing information of hosts.

- Type 3 route — inclusive multicast route: This is used for automatic discovery of VTEPs and dynamic establishment of VXLAN tunnels.
- Type 5 route — IP prefix route: This is used to advertise imported external routes or advertise routing information of hosts.

Advertised EVPN routes contain Route Distinguisher (RD) and VPN target (also known as route target) information. RD is used to differentiate different VXLAN EVPN routes. A VPN target is a BGP extended community attribute used to control the advertisement and receiving of EVPN routes. That is, a VPN target defines the peers that can receive EVPN routes from the local end as well as whether the local end can receive EVPN routes from peers.

There are two types of VPN targets:

- Export target: The VPN target attribute is set to export target when the local end sends EVPN routes.
- Import target: When receiving an EVPN route from a peer, the local end compares the export target in the received packet with the import target of its own. If they are the same, the local end accepts the route. Otherwise, the local end discards the route.

1.3 Type 2 Route

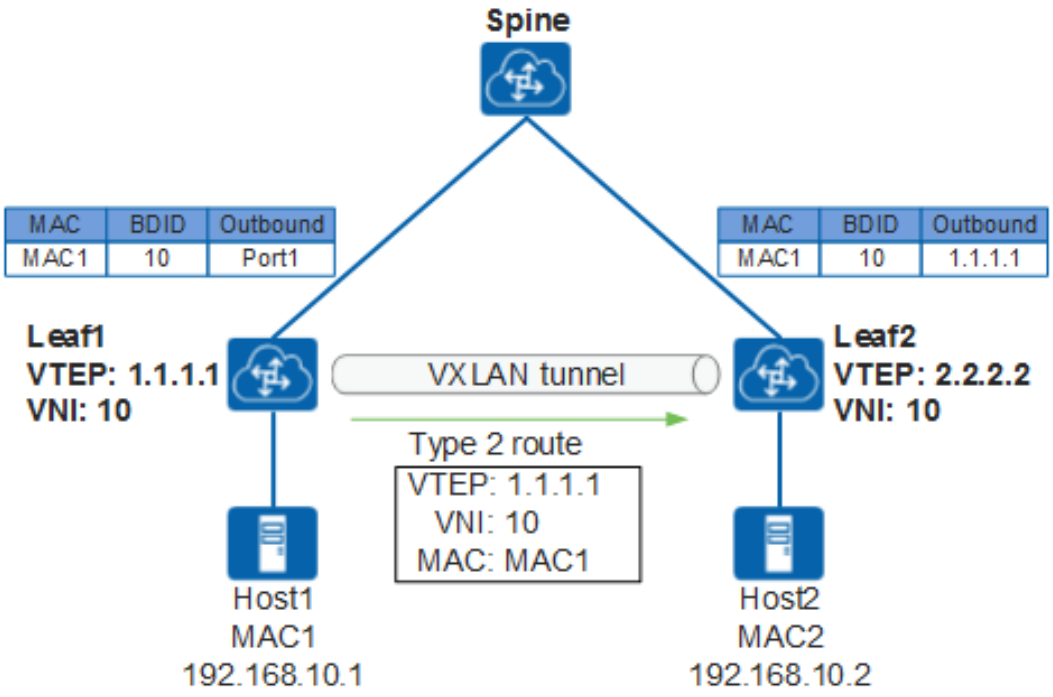
NLRI fields in type 2 routes are as follows:

Figure 1-55 NLRI fields in type 2 routes

Route Distinguisher	RD of an EVPN instance
Ethernet Segment Identifier	ESI of a connection to a peer
Ethernet Tag ID	VLAN ID
MAC Address Length	Length of host MAC address
MAC Address	Host MAC address
IP Address Length	Mask length of host IP address
IP Address	Host IP address
MPLS Label1	Layer 2 VNI
MPLS Label2	Layer 3 VNI

In Figure 1-55, the type 2 route carries the host MAC address and host IP address. Type 2 routes can be used to advertise host MAC addresses and IP addresses.

Figure 1-56 MAC address advertisement using a type 2 route



In Figure 1-56, Leaf1 learns the MAC address of Host1 after receiving a packet from Host1. It then generates a type 2 route and sends it to Leaf2. This route carries the export route target (ERT) of the EVPN instance, Host1 MAC address, and VTEP IP address of Leaf1.

After receiving the route from Leaf1, Leaf2 compares the ERT in the route with the import route target (IRT) of the local EVPN instance and determines whether to accept the route. If they are the same, Leaf2 accepts the route and learns Host1 MAC address. If they are different, Leaf 2 discards the route.

1.4 Type 3 Route

NLRI fields in type 3 routes are as follows:

Figure 1-57 NLRI fields in type 3 routes

Prefixes

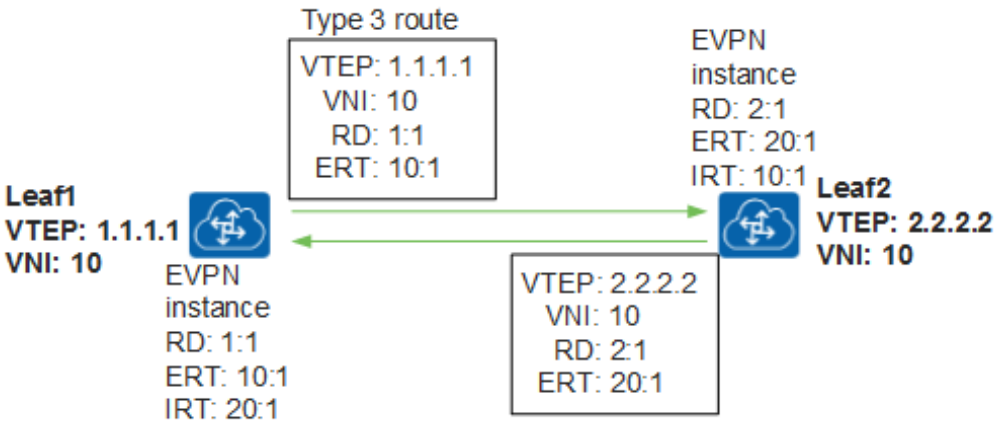
Route Distinguisher	RD of an EVPN instance
Ethernet Tag ID	VLAN ID (The value is all 0s.)
IP Address Length	Mask length of local VTEP IP address
Originating Router's IP Address	Local VTEP IP address

PMSI attributes

Flags	Flag bit (This field is meaningless in VXLAN.)
Tunnel Type	Tunnel type (The value 6 indicates VXLAN.)
MPLS Label	Layer 2 VNI
Tunnel Identifier	Tunnel information

Type 3 routes carry VTEP IP information, which is used for automatic VTEP discovery and dynamic VXLAN tunnel establishment.

Figure 1-58 VXLAN tunnel establishment using type 3 routes



In Figure 1-58, after a BGP EVPN peer relationship is established between Leaf1 and Leaf2, Leaf1 generates a type 3 route and sends it to Leaf2. This route carries the local VTEP IP address, VNI, and ERT of the EVPN instance.

After receiving the route from Leaf1, Leaf2 compares the ERT in the route with the IRT of the local EVPN instance and determines whether to accept the route. If they are the same, Leaf2 accepts the route and establishes a VXLAN tunnel to the peer. If the remote VNI is the same as the local VNI, Leaf2 creates an ingress replication list to forward broadcast, multicast, and unknown unicast packets.

1.5 Type 5 Route

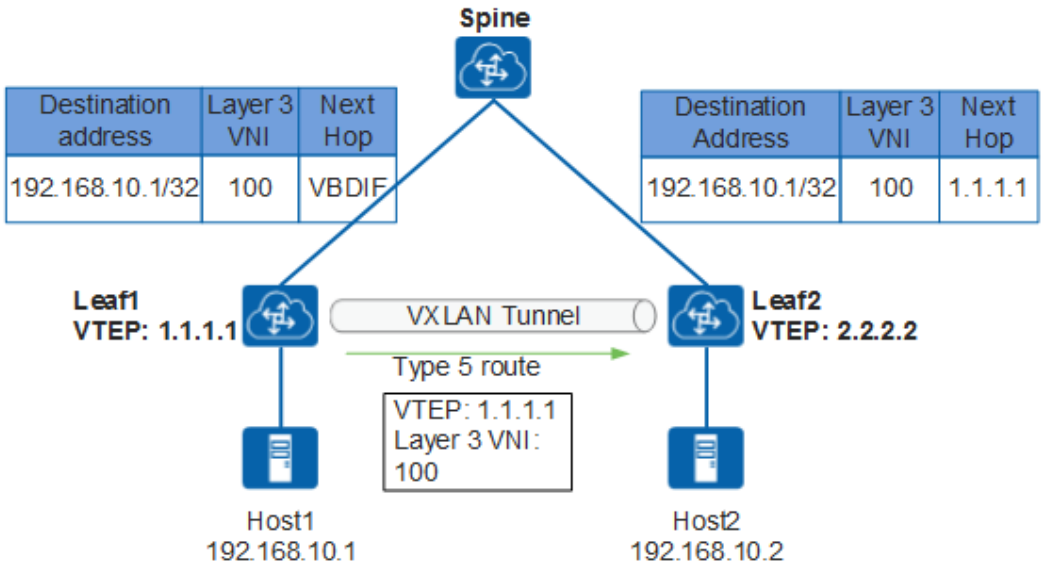
NLRI fields in type 5 routes are as follows:

Figure 1-59 NLRI fields in type 5 routes

Route Distinguisher	RD of an EVPN instance
Ethernet Segment Identifier	ESI of a connection to a peer
Ethernet Tag ID	VLAN ID
IP Prefix Length	Mask length of IP prefix
IP Prefix	IP prefix
GW IP Address	Default gateway address
MPLS Label	Layer 3 VNI

Type 5 routes carry routing information, which is used for route advertisement. Different from type 2 routes, type 5 routes can advertise both 32-bit host routes and Ethernet segment routes.

Figure 1-60 Type 5 route advertisement



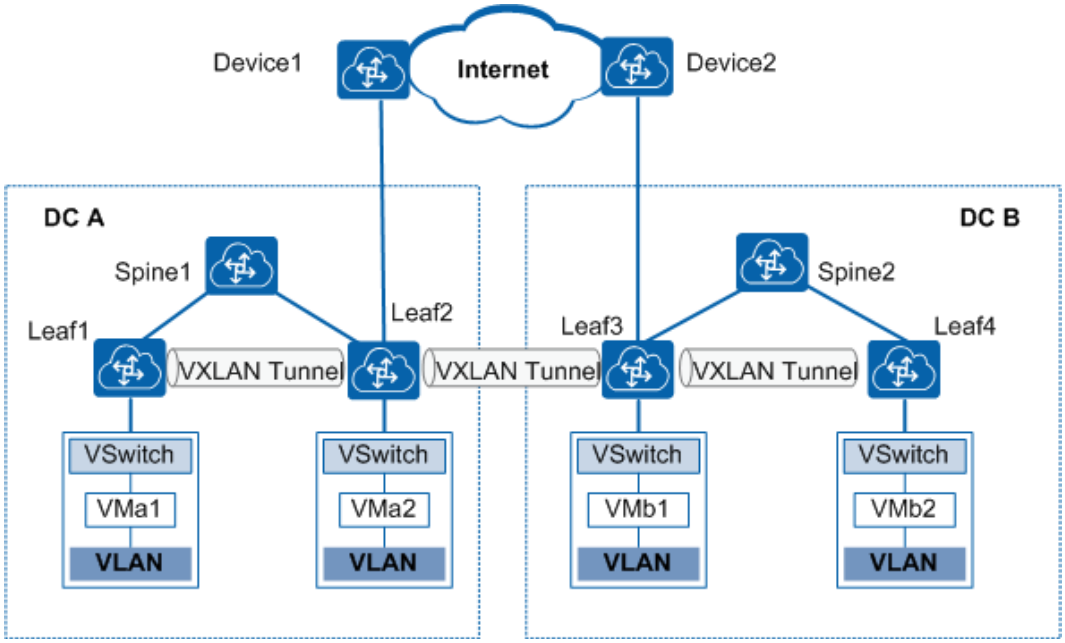
Type 5 routes can advertise local private routes (such as static routes, direct routes, and routes of other routing protocols) to other EVPN networks and generate host routes or Ethernet segment routes on the peer. Therefore, type 5 routes enable communication between hosts on a VXLAN network and a non-VXLAN network.

1.6 DCI Using BGP EVPN

BGP EVPN creates one VXLAN tunnel in each DC and one VXLAN tunnel between the two DCs to implement data center interconnect (DCI). As shown in Figure 1-61, BGP EVPN is used to create VXLAN tunnels in distributed gateway mode within both DC A and DC B so that the VMs deployed in each DC can communicate with each other. Leaf2 and Leaf3 are the edge devices within the DCs that connect to the backbone network. BGP EVPN is used to configure VXLAN tunnels on Leaf2 and Leaf3 so that the VXLAN packets received by one DC can be decapsulated, re-encapsulated, and sent to the peer DC. This process provides

end-to-end bearing for inter-DC VXLAN packets and ensures that VMs in different DCs can communicate with each other.

Figure 1-61 Using BGP EVPN to implement DCI



1.7 Summary

EVPN is a BGP-based technology and must be deployed on switches. Switches serving as VTEPs must encapsulate packets by using VXLAN. Servers connect to switches through interfaces or VLANs. These interfaces or VLANs are mapped to BDs, and each BD is bound to an EVPN instance. These EVPN instances exchange routes to implement VXLAN tunnel establishment and MAC address learning.

Summary

This book defines the key concepts in the data center network design and expands on the CloudEngine series switches, cabling, oversubscription ratio design, fabric network technology application, origin of the overlay network, and the VXLAN technology. After reading this book, you will be able to understand the advantages of Huawei CloudEngine series switches, problems they can help you overcome, and how to build your own data centers.

Data center network technologies are constantly changing, as are the design ideas and technical implementations discussed in this document. If you are interested in the development of these technologies, you are advised to visit the Enterprise technical support website to learn about the switch functions that are supported in the latest versions.

Currently, the software-defined networking (SDN) architecture is becoming more advanced and commercially available. More focus is being placed on the openness, low latency, and refined O&M capabilities of switches carrying fabric network traffic. To help customers adapt to the rapid changes of cloud services, Huawei innovatively launches the cloud data center network SDN solution, aiming to build simple, open, and flexible cloud data center networks for customers and accelerate enterprise digital transformation. For more information about the cloud data center network SDN solution, visit http://support.huawei.com/onlinetoolsweb/NetSolution/DataCenterNetwork/en/index.html.

Huawei will focus on the latest network architecture designs and technology applications and continuously update information. We hope you will pay close attention to these updates.

Further Reading

Table 1-13 lists the documents related to this book.

Table 1-13 Documents

Document	URL
Pre-sales documentation for CE series switches	http://e.huawei.com/en/products/enterprise-networking/switches/data-center-switches/
After-sales documentation for CE series switches	http://support.huawei.com/enterprise/en/index.html
CE series switches portfolio	http://e.huawei.com/en/material/onLineView?MaterialID=fe45e4fdd09e4dbd920b40cf35c757c1
Data Center Network Technology Red Treasure Book	http://forum.huawei.com/enterprise/en/thread-406591.html
Forwarding performance evaluation tool for CE series switches	http://support.huawei.com/onlinetoolsweb/proforward_tool/en/index.html

Acronyms and Abbreviations

Table 1-14 Acronyms and abbreviations

Term	Description
AS	An autonomous system (AS) is a part of a network. Generally, an AS is controlled by an organization and runs a routing protocol. Routing between different ASs is implemented through inter-area protocols.
AOC	An active optical cable (AOC) contains optical modules and optical fibers.
BD	A bridge domain (BD) is a Layer 2 broadcast domain in which data packets are forwarded on a VXLAN network.
BGP	The Border Gateway Protocol (BGP) is a distance vector routing protocol that allows devices in different ASs to communicate and select optimal routes. Internal/Interior BGP (IBGP) runs inside an AS and External/Exterior BGP (EBGP) runs between ASs.
DAC	A direct attach cable (DAC) has a fixed length and fixed connectors at both ends.
ECMP	Equal-cost multi-path routing (ECMP) implements equal-cost multi-path load balancing and link backup.
EGP	The Exterior Gateway Protocol (EGP) is used to dynamically exchange routing information between ASs.
EOR	End of Row (EOR) switches are deployed at the end of each row of cabinets to provide unified network access for servers.
EVPN	Ethernet Virtual Private Network (EVPN) is a VPN technology used for Layer 2 internetworking. EVPN uses BGP extensions to implement MAC address learning and advertisement between different sites on a Layer 2 network.
IGP	The Interior Gateway Protocol (IGP) is a routing protocol used inside an AS.
IS-IS	Intermediate System to Intermediate System (IS-IS) is an IGP protocol used inside an AS. IS-IS is also a link state routing protocol and uses the shortest path first (SPF) algorithm to calculate routes.
M-LAG	Multichassis Link Aggregation Group (M-LAG) is a mechanism that implements inter-device link aggregation. M-LAG connects one device to two devices to establish a dual-active system, improving link reliability from the card level to the device level.

Term	Description
MMF	A multimode fiber (MMF) can transmit optical signals of multiple modes.
MOR	The Middle of Row (MOR) architecture is an improvement over the EOR architecture because it provides a unified network access cabinet for servers. The network cabinet is placed in the middle of a row of cabinets to shorten the distance between it and server cabinets.
OSPF	The Open Shortest Path First (OSPF) protocol is a link-state IGP protocol developed by IETF.
SMF	A single-mode fiber (SMF) can transmit optical signals of one mode.
STP	The Spanning Tree Protocol (STP) prevents loops on a local area network (LAN). Devices running STP exchange information with each other to discover loops on the network and then block some interfaces to eliminate loops.
SVF	Super virtual fabric (SVF) is a vertical virtualization technology that virtualizes low-cost fixed leaf switches into interface cards of a parent switch. This technology increases the interface density of the parent switch and implements centralized management of switches, allowing for high-density access and simple management in data centers.
TRILL	Transparent Interconnection of Lots of Links (TRILL) uses IS-IS extensions to implement Layer 2 routing and applies Layer 3 link state routing technologies to Layer 2 networks.
TOR	The name Top of Rack (TOR) indicates that a TOR switch is deployed at the top of a cabinet. However, a TOR switch can also be deployed in the middle or at the bottom of a cabinet.
VLAN	Virtual Local Area Network (VLAN) technology divides a physical LAN into multiple broadcast domains, each of which is called a VLAN. Hosts within a VLAN can communicate with each other but cannot communicate directly with hosts in other VLANs. Therefore, broadcast packets are confined within a single VLAN.
VM	A virtual machine (VM) is a virtual computer that runs an operating system and applications, which is similar to a physical machine.
VNI	A VXLAN Network Identifier (VNI) is similar to a VLAN ID and identifies a VXLAN segment. VMs on different VXLAN segments cannot communicate at Layer 2.
VRRP	The Virtual Router Redundancy Protocol (VRRP) groups multiple routing devices into a virtual router. The virtual IP address of the virtual router is used as the default gateway address for communication with an external network. When a gateway device fails, VRRP selects another gateway device to transmit service traffic, ensuring reliable communication.
vSwitch	A virtual switch (vSwitch) implements Layer 2 or some Layer 3 network functions of a physical switch through software.
VTEP	A VXLAN Tunnel Endpoint (VTEP) encapsulates and decapsulates

Term	Description
	VXLAN packets. A pair of VTEP addresses identifies a VXLAN tunnel.
VXLAN	Virtual eXtensible Local Area Network (VXLAN) is a network virtualization technology.