

Data Center and Data Center Network Technologies



Foreword

- In the cloud and big data era, data centers are facing massive construction requirements. With the development of technologies and the improvement of user requirements, simplicity, efficiency, and reliability have become a new idea for future data center development, and the design concept of data centers is changing quietly.
- This course introduces the basic concepts of the data center and data center network.

Objectives

- On completion of this course, you will be able to:
 - Describe the concepts of data centers and data center network.
 - Understanding Common data center network Architectures.
 - Clarify key network technologies in the data center.

Contents

- 1. Data Center Overview**
2. Data Center Network Overview
3. Overview of Key DC Technologies

Why Do We Need a Data Center?

- With the development of enterprises, the amount of data that enterprises need to process every day is increasing. The processing power of personal computers in offices is no longer enough to meet the needs of enterprises. To provide more efficient methods for processing information and data, enterprises build or rent data centers to process massive data in a centralized manner, meeting enterprise development requirements.



Small business using personal PCs
Processing data

As the enterprise grows, more and
more data needs to be maintained.

Data is centrally processed in the data center, and
large enterprises use data through the data center

What is a Data Center?

- A data center, as the name suggests, is a data center where enterprises process and store massive amounts of data.
- A data center is actually a large-scale equipment room. Enterprises use the existing Internet lines and bandwidth resources of communications carriers to establish a standardized data center equipment room environment to provide all-round computing, storage, and security services for enterprises, governments, and individuals. The data center has the characteristics of high running speed, large storage capacity, and high security.

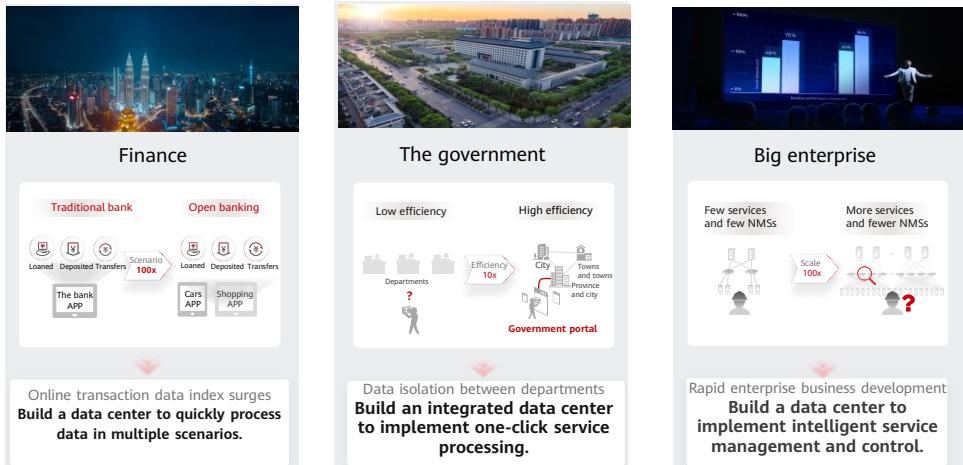


Data center equipment room



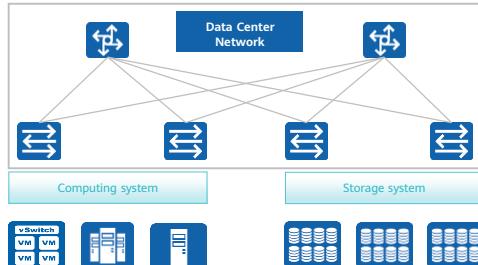
Data center cabinet

Typical Application Scenarios of Data Centers



Overall Data Center Architecture

- For enterprises, the data center is actually an extended version of the personal computer, which is responsible for computing, storing, and forwarding enterprise data. A modern data center consists of the following parts:
 - The computing system consists of a large number of servers and is the heart of the data center. It processes massive data in the data center.
 - A storage system consists of different types of storage devices. A storage device is a place where massive data is stored and is used for information storage.
 - The data center network consists of different types of network devices, such as switches and firewalls. It connects the computing and storage systems in the data center. All data interaction between the computing and storage systems is implemented through the data center network.



8 Huawei Confidential

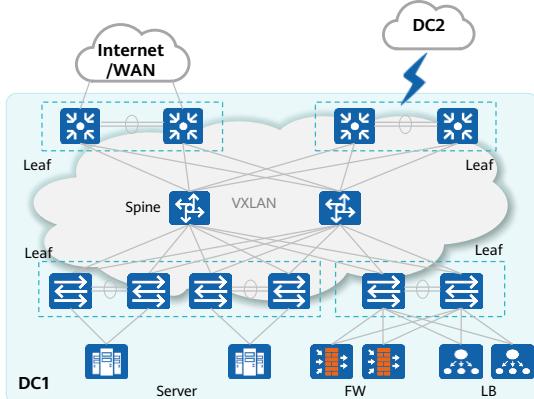
 HUAWEI

- Key devices in the data center equipment room include servers, network devices, and storage devices. Small- and medium-sized data centers are key devices, such as servers, which are characterized by small physical space, small requirements for network devices, and limited capacity expansion.

Contents

1. Data Center Overview
- 2. Data Center Network Overview**
 - Introduction to DCN
 - DCN Common Concepts
3. Overview of Key DC Technologies

Data Center Network

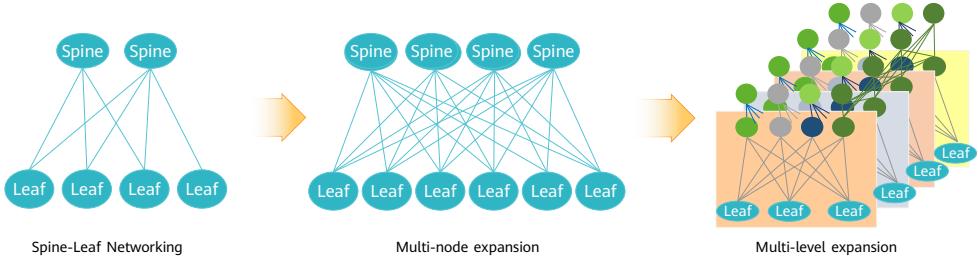


- The Data Center Network (data center network) is the infrastructure that carries services in a data center and is responsible for data forwarding.
- Multiple data center network can connect to branches of an enterprise or organization across regions. Data center networks can also connect to the Internet or WAN.

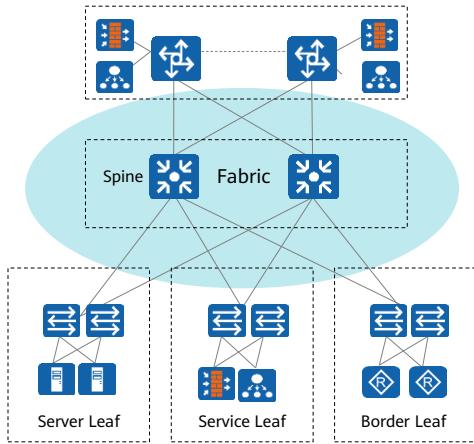
- The data center network uses the Spine-Leaf architecture and uses VXLAN (Virtual Extensible Local Area Network) Connectivity.
 - Spine: a backbone node and core node on a VXLAN network. It provides high-speed IP forwarding and connects to leaf nodes through high-speed interfaces.
 - Leaf: A leaf node, which provides VXLAN access for various network devices. Devices of different roles can be co-deployed based on the device type. (As shown in the figure, the border leaf node and service leaf node are co-deployed.) The specific types and functions will be described in detail later.
 - Value-added service (VAS): A device that provides L4-L7 services, such as a firewall or load balancer.

Advantages of the Spine-Leaf Architecture

- Spine-Leaf is a new network architecture of a data center, consisting of spine nodes and leaf nodes. Spine nodes are backbone nodes and provide high-speed IP forwarding. A leaf node provides the network access function.
- Spine nodes and leaf nodes are fully connected at Layer 3 and equal-cost multipathing is used to improve network availability.
- The Spine-Leaf architecture has high scalability.

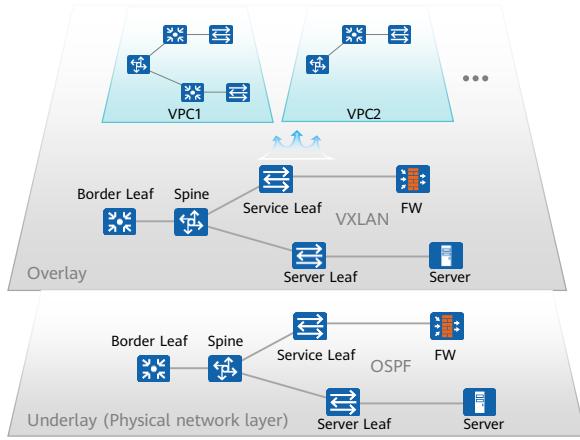


Basic Concepts of the Spine-Leaf Architecture



Terms	To explain
Spine	A core node on a VXLAN fabric network. It provides high-speed IP forwarding and connects to leaf nodes through high-speed interfaces.
Leaf	A leaf node, which is a VXLAN fabric function access node and provides various network devices to access the VXLAN network.
Fabric	A group of spine and leaf nodes are interconnected to form the basic physical network topology of the data center.
Service Leaf	A leaf node, which provides Layer 4 to Layer 7 value-added services, such as firewall and load balancer, to access the VXLAN fabric network.
Server Leaf	A leaf node, which provides computing resources, such as virtualized and non-virtualized servers, to access the VXLAN fabric network.
Border Leaf	A leaf node is a leaf node that connects external traffic of a data center to the VXLAN fabric network of the data center and connects to routers or transmission devices.

VXLAN-based Data Center Network Layer

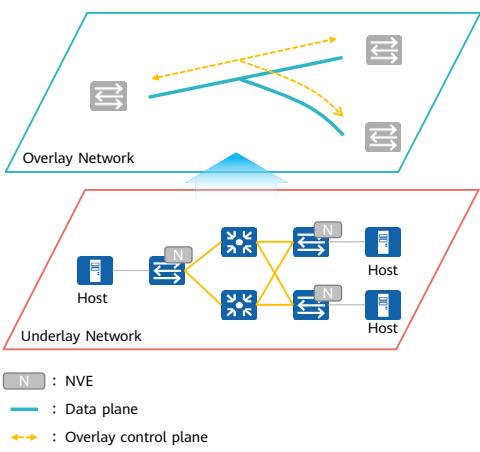


- A VPC is a logically isolated network created by tenants based on the VXLAN technology. It can also be called a security domain. A VPC usually represents a department or a service.

- Use virtualization technologies (such as VXLAN) to build a logical topology based on any physical network and enable logical tunnels to build a large Layer 2 network.

- A physical network established by a physical device.
- Provides interconnection capabilities for all services in the data center.
- Basic bearer network for service data forwarding.

Underlay and Overlay



Overlay

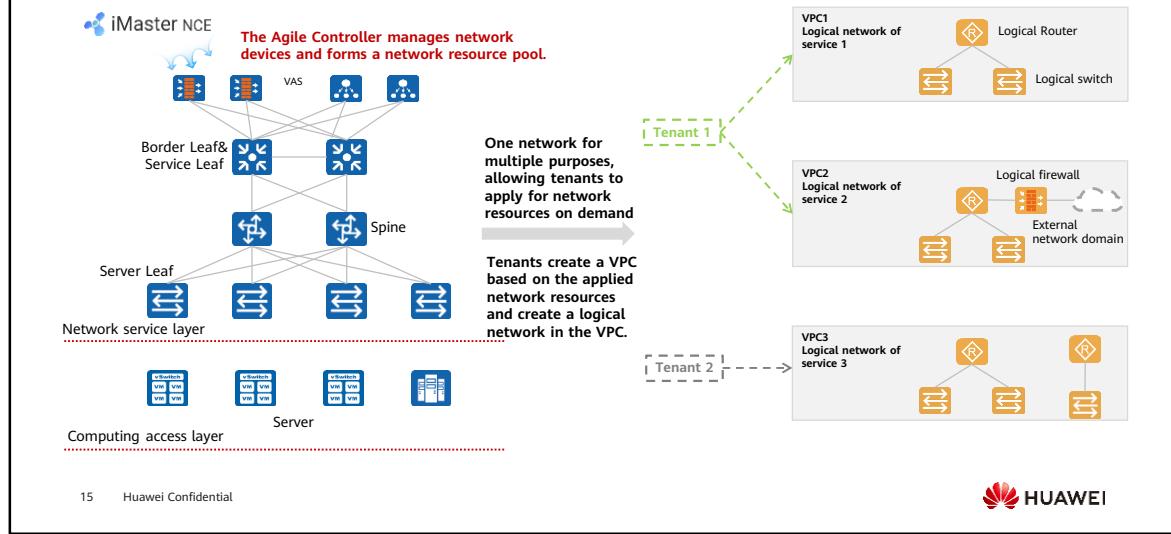
- VXLAN is a logical network established on the underlay network.
- It has an independent forwarding plane and control protocol.
- The underlay physical network is transparent to the devices that are not connected to the VXLAN tunnel endpoints.

Underlay

- The underlay network consists of various physical network devices and is a bearer network of the overlay network.
- After the overlay technology is implemented on the underlay network, a logical network is formed on the basis of the underlay network.
- The underlay network provides basic capabilities, such as reachability and reliability, for the upper-layer overlay network.
- The underlay network has independent control plane protocols and forwarding plane protocols. Generally, OSPF or EBGP is used as the control plane protocol, and IPv4 is used as the forwarding plane protocol.
- The underlay network is logically isolated from the overlay network and is unaware of overlay network routes.



Typical Data Center Network Scenarios



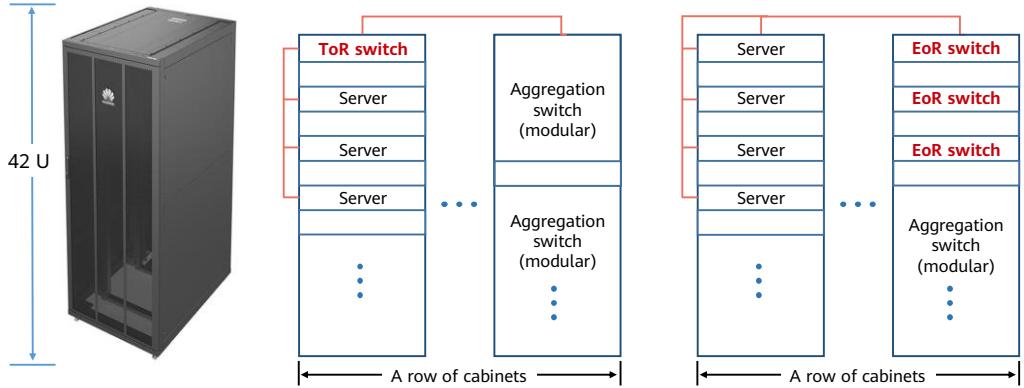
- iMaster NCE (Fabric) is an autonomous driving control system developed by Huawei for data center network scenarios. It integrates management, control, analysis, and AI functions. The following sections will describe the functions, features, and application scenarios.

Contents

1. Data Center Overview
2. **Data Center Network Overview**
 - Introduction to DCN
 - **DCN Common Concepts**
3. Overview of Key DC Technologies

Integrated Cabling

- The integrated cabling of a DC has three important concepts: Top of Rack (ToR), End of Row (EoR), and Middle of Row (MoR).

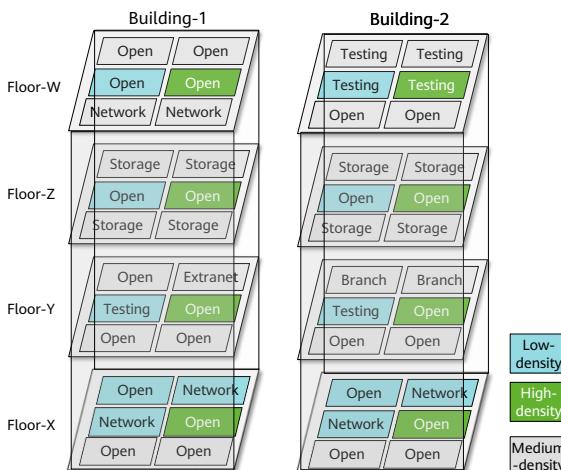


17 Huawei Confidential

HUAWEI

- The common height of standard cabinets within a DC is 42 U ($1 \text{ U} = 4.445 \text{ cm}$) and the height of each cabinet unit is 4.445 cm.
- Top of Rack (ToR):** ToR switches are deployed at the top of a cabinet and servers in the cabinet are connected to a switch through optical fibers or network cables. ToR switches are connected to the aggregation switches at the upper layer. This applies to the scenario with a large number of access devices or a high-density single cabinet. The distributed access of servers can reduce the network connections between server cabinets and network cabinets, simplifying the connection management. At the same time, access switches are distributed in multiple cabinets, causing difficulties in centralized maintenance and management. This is common in cloud data centers.
- End of Row (EoR):** Access switches are deployed on one or two cabinets of a cabinet group in a centralized mode. All servers of the row of cabinets are connected to the switches through horizontal cabling. This is common in traditional DCs. If cables are connected in the EoR mode, many cable connections will be aggregated from multiple server cabinets to network cabinets, causing difficulties in connection management while bringing conveniences to centralized maintenance and management of switches.
- Middle of Row (MoR):** The connection modes of MoR switches and EoR switches are similar. Access switches are deployed in one or two cabinets of a cabinet group in a centralized mode, but network cabinets are in the middle of the cabinet group. In this situation, connections from server cabinets to network cabinets are slightly simplified compared with the EoR mode, and switches are managed in a centralized way. It is a compromise solution between the ToR mode and the EoR mode.

Equipment Room Module



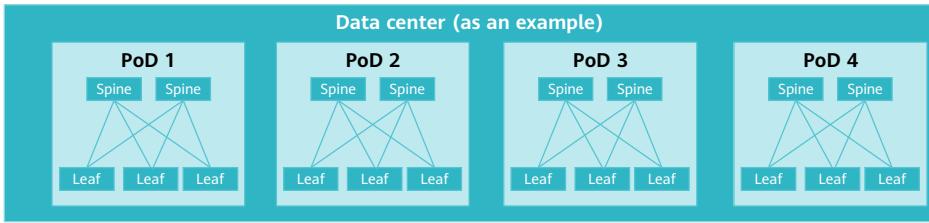
18 Huawei Confidential

- For example, each floor of each building in a financial DC is divided into multiple equipment room modules.
- From the perspective of functions, equipment room modules can be divided into different types, including network modules, storage modules, open server modules, and test modules, with different power densities.
- As shown in the left figure, based on power densities, each floor is divided into three areas: high-density area, medium-density area, and low-density area.
- Network module: responsible for network access, as the core of the WAN and LAN. The power consumption varies for different devices.
- Storage module: used for housing storage devices in a centralized mode
- Open server module: used for housing servers in a centralized mode
- Test module: used for housing test devices in a centralized mode



- Network module: responsible for network access. Network modules feature differences both in power consumption and device types, such as large-scale core network devices with high power consumption and devices with low power consumption, such as load balancing, firewalls, switches, and routers.
- Storage module: used for deploying storage devices in a centralized mode, including NAS storage and SAN storage, as well as tape libraries and virtual tape libraries.
- Open server module: used for deploying standard servers in a centralized mode, including PC servers, blade servers, and small-sized servers. Servers feature high standardization and high density.
- Testing module: used for deploying test devices in a centralized mode. Testing modules feature high flexibility, with lower security requirements compared with production modules. The modules can also be adjusted at any time based on testing requirements, with low management requirements.
- In the data center of a financial institution, based on special purposes, other special equipment room modules can be planned.
 - Internet module: responsible for accessing Internet applications. The module should be accessed to the Internet with high security requirements. It is vulnerable to various Internet attacks, such as online banking, websites and e-commerce applications.

PoD

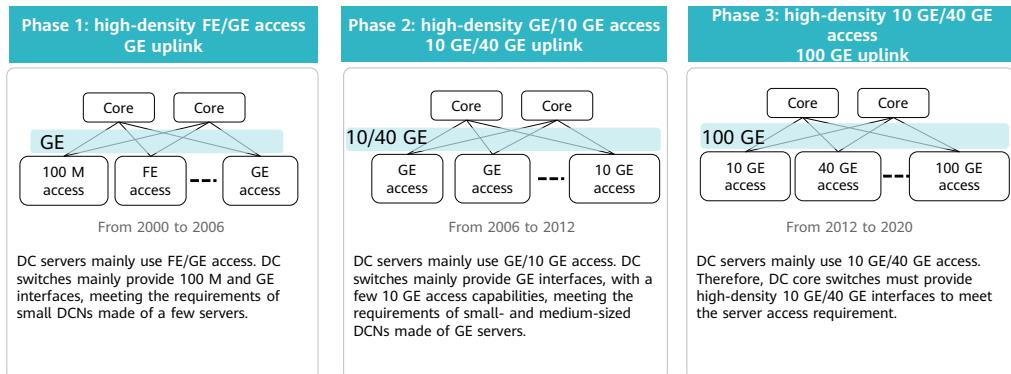


- To facilitate the resource pool-based operation and management of a DC, a DC is divided into one or more physical partitions and each partition is called a Point of Delivery (PoD). PoD is a common concept of DCs for physical design and a modular design entity integrating network, storage, and computing.
- PoDs can be defined based on actual service requirements:
 - In large DCs, equipment room modules can be defined as a PoD.
 - In midsize DCs, every two or multiple rows of cabinets can be defined as a PoD.
 - In small DCs, one or more cabinets can be defined as a PoD.

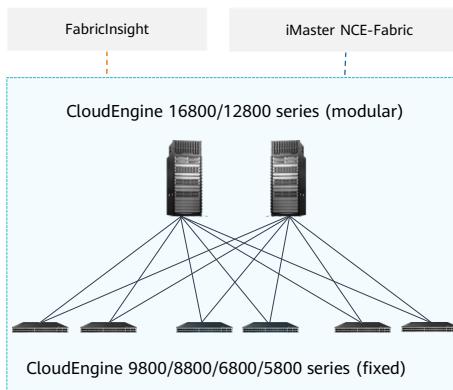
- Definitions of the PoD scope vary with different enterprises' user habits. For example, some large enterprises consider an equipment room module is wider than a PoD. In an enterprise, a PoD consists of 48 ToR devices and 4 spine switches.

Data Center Switch

- The data center switch usually refers to a hardware switch. It has gone through three phases: in phase 1, FE access and GE uplinks are applied; in phase 2, GE access and 10GE uplinks are applied; now, in phase 3, 10 GE/40 GE access and 100 GE uplinks are applied.



DC Switches for the AI Era



The CloudEngine (hereinafter referred to as CE) series refers to Huawei's high-performance switches designed for next-generation data centers, including:

- CE 16800 series and the CE 12800 series, mainly used for high-speed data forwarding in data centers
- CE 9800/8800/6800/5800 series, mainly used for high-density access in DCs

Contents

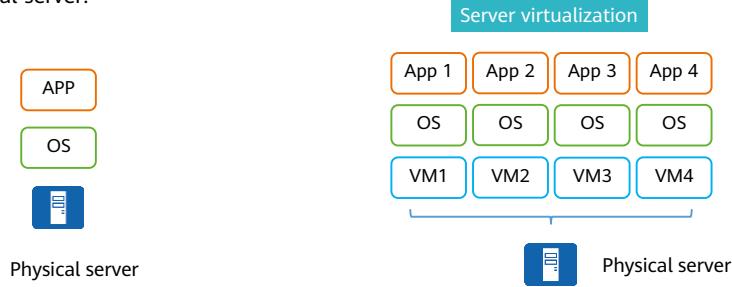
1. Data Center Overview
2. Data Center Network Overview

3. Overview of Key DC Technologies

- DC Key IT Technologies
 - DC Key Network Technologies

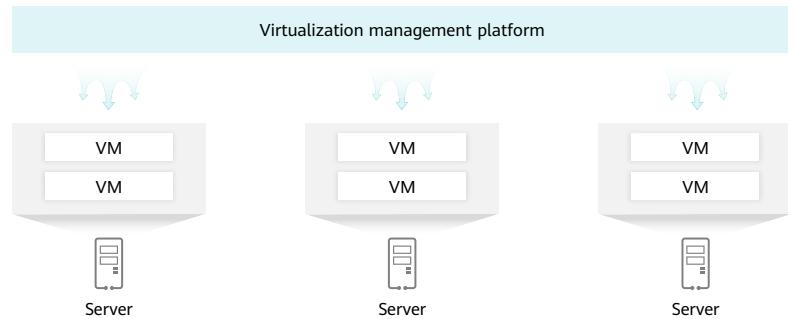
Introduction to Server Virtualization

- Server virtualization is a virtual technology with which you can run multiple virtual machines on a physical server to obtain advantages, including higher physical resource usage, rapid service deployment, and elasticity.
- Eventually, users can install and run multiple applications and services on these virtual machines like using a physical server.



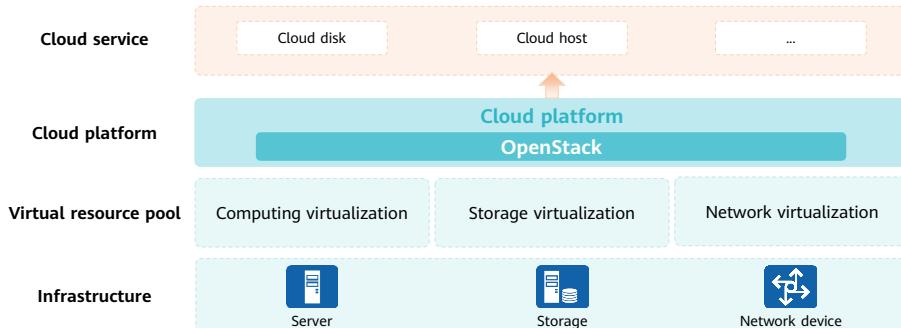
Server Virtualization: Virtualization Management Platform

- As services grow, the number of virtual server clusters reaches hundreds to thousands. Therefore, a virtualization management platform is required for centralized management.
- The virtualization management platform provides a simple user interface and various functions, such as monitoring and managing virtual resources, simplifying the creation process of VMs, configuring resource scheduling policies, and executing rules. Mainstream virtualization platforms in the industry include Huawei VRM, VMware vCenter, and Microsoft System Center.



Introduction to Cloud Computing and OpenStack

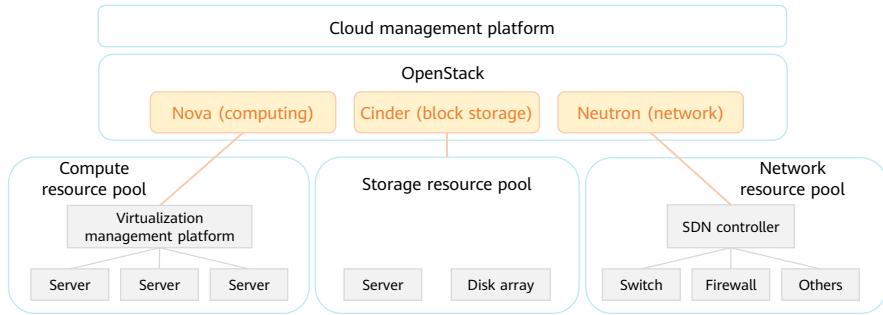
- OpenStack is an open-source cloud operating system that controls large-sized computing, storage, and network resource pools of the data center. After OpenStack is deployed, users can manage resources through web UIs, CLIs, or APIs.
- OpenStack does not simply mean cloud computing, but a cloud platform as a key component of cloud computing. OpenStack aims to offer resource management, including managing the computing, storage, and network resource pools of heterogeneous vendors.



- OpenStack and cloud computing:
 - OpenStack is a framework for building a cloud OS. The cloud OS integrates and manages various hardware devices and bears various upper-layer applications and services to form a complete cloud computing system. Therefore, OpenStack is the core software component of a cloud computing system and the basic framework for building a cloud computing system.
- OpenStack and virtualization:
 - OpenStack is a cloud OS framework. To build a complete cloud OS, especially to implement resource access and abstraction, OpenStack needs to be integrated with the virtualization software to implement compute resource pooling of servers. In the resource pooling process, physical resources are virtualized by the virtualization software.

Introduction to OpenStack Core Components

- OpenStack is decomposed into several service components, each of which supports the plug-and-play mode to meet diversified service requirements.
- There are many core projects of OpenStack resource management. Nova manages compute resources. Cinder manages block storage resources. Neutron manages network resources.
- The upper layer of OpenStack connects to the cloud management platform. The cloud management functions include but are not limited to: operations for tenants, cloud provisioning services, accounting, and multi-cloud monitoring.



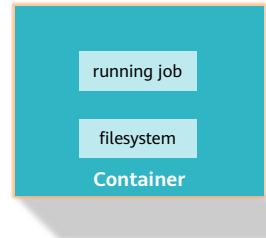
- To build a cloud OS, a large number of software components need to be integrated so that they can work together to provide functions and services required by system administrators and tenants. However, OpenStack cannot independently provide all capabilities required by a complete cloud OS.
- For example, OpenStack cannot independently access and abstract resources, and needs to work with underlying virtualization software, software-defined storage (SDS), and software-defined networking (SDN). OpenStack cannot independently provide comprehensive application lifecycle management capabilities, and needs to integrate various management software platforms at the upper layer. OpenStack does not have complete system management and maintenance capabilities. When OpenStack is put into production, it needs to integrate various management software and maintenance tools. The man-machine interface provided by OpenStack is not powerful enough.
- For details, see *Technical Principles and Applications of the OpenStack Cloud Platform*.

Introduction to Containers

- Container is an OS-level virtualization technology. Containers are more lightweight and efficient than VMs.
- For example, the Linux operating system can be divided into the kernel space and the user space. The kernel of an operating system supports multiple isolated user space instances. An advantage of the container technology is the integration of applications and their operating environment. This enables fast transportation of an application and greatly simplifies the process of development-test-deployment-O&M.



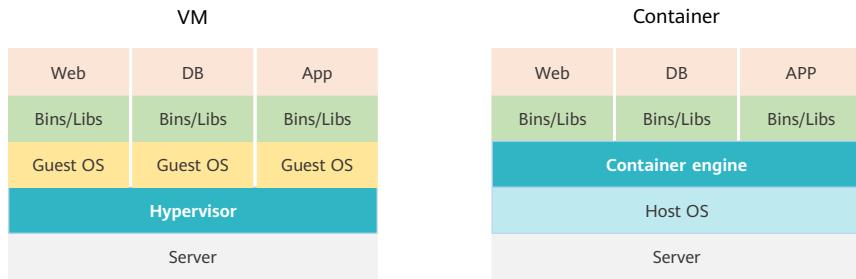
Container



- The Linux operating system and drivers run in the kernel space and applications run in the user space.
- Container can be more precisely defined as the entity for running a container image.
- Container image:
 - An application and its dependencies (including all files and directories of the OS) can be packaged into an image.
 - The image contains all dependencies required for application running. You only need to run the image in the isolated sandbox without any modification or configuration.
 - Images focus on packaging applications and their runtime environments in a unified format. This ensures high consistency between the local environment and the cloud environment.

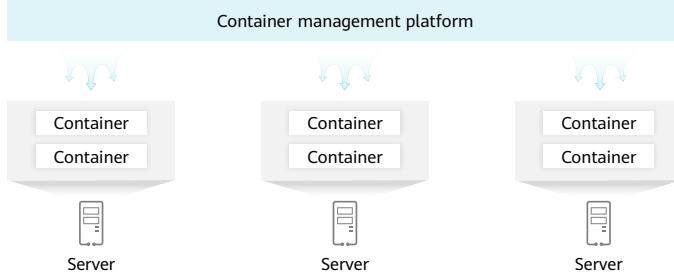
Comparison Between Containers and VMs

- Compared with VMs, containers share the same operating system kernel. They are of low independence, only providing the process-level isolation.
- Containers perform better than VMs, in terms of the startup speed, running performance, and server resource usage.



Container Management Platform

- The container engine is only a daemon for container management on a single node, while the scale of nodes managed by enterprise DCs or the public cloud is quite large. An independent container management platform is needed to implement large-scale container management.
- A mature container management platform should at least contain the following two major functions: application orchestration management and cluster resource scheduling.
- There are three platforms of cluster resource management scheduling and application orchestration in the industry: Kubernetes, Swarm, and Mesos.



30 Huawei Confidential



- Resource management and scheduling of container clusters: The resource status of managed nodes is collected to complete the resource management of tens of thousands of nodes; the container resource requests of users are handled based on specific scheduling policies and algorithms.
- Application orchestration and management: In terms of different types of applications in a data center, such as the Web service and task processing in batches, basic management capabilities commonly applied by different applications are abstracted and exposed to users through APIs. As such, users can achieve automatic application management by using the preceding API capabilities of the container management platform when developing and deploying their own applications. Through application orchestration and management, users can enable application model customization and one-click automatic deployment. Users enable a gray upgrade based on the one-click application deployment of application templates, which significantly simplifies the application management and deployment.
- The Kubernetes ecosystem is a community project launched by Google, covering the container cluster resource management and distribution, as well as application management components of different applications, such as copy reliability management, service discovery and load balancing, gray upgrade, and configuration management.
- The Mesos ecosystem is actively promoted by Mesosphere, Twitter, and other companies.
- The Docker ecosystem is proposed by Docker with an aim to develop towards the upper layer of the container ecosystem by introducing container schedule components of Swarm container resource management and Compose application orchestration components.

Introduction to Storage Types

- Block storage, file storage, and object storage are three common concepts in enterprise DC storage. Different data types and service scenarios in enterprises have different storage requirements, and the three different storage types providing different storage services are briefly described here:

Storage Type	Typical Application	Typical Device Model	Application Scenario
Block storage	High-performance applications	Disk arrays and hard disks	High I/O database
File storage	File-sharing applications based on LANs	FTP and NFS servers	<ul style="list-style-type: none"> • Common scenario: file sharing • High-performance scenario: video processing and animation rendering/high-performance computing
Object storage	Applications with a large amount of data and a rapid growth of storage capacity	Distributed servers with built-in large-capacity hard disks	Video storage of VOD/video surveillance, image storage, disk storage, static web page storage, and remote backup storage/archiving

Introduction to the Storage System

- From the storage product perspective, storage products in a data center are classified into two modes: centralized storage and distributed storage.

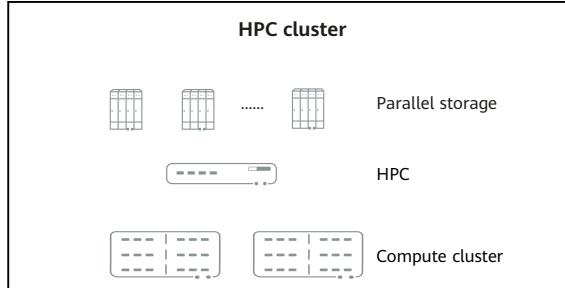


- Centralized storage: One or more primary computers form a central node where data is centrally stored and all service units and functions are deployed on a storage system.
- In a centralized system, each terminal or client is only responsible for the input and output of data, while the storage and processing of data are completely decided by a host.
- Distributed storage: The storage system stores data on multiple independent devices.
- A distributed storage system adopts a scalable system architecture and enables multiple storage servers to share the storage load. This improves the scalability, reliability, availability, and access efficiency.

- The most distinct feature of a centralized system is the simple deployment architecture. As centralized systems are often based on mainframes with outstanding performances at the bottom layer, there is no need to consider the multi-node deployment of services or the distributed collaboration among multiple nodes.
- Traditional network storage systems store data on centralized storage servers, which may become the bottleneck of system performance and a vulnerable point in terms of availability and security, failing to meet the requirements of large-scale storage applications.
- Features of the centralized storage:**
 - Devices, diversified in types, often establish external connections through an IP or FC network.
 - Extensibility must be ensured. Restricted by the controller's capabilities, the network scalability is limited and storage capabilities are in the PB level.
 - Devices need to be replaced after the lifecycle ends and all the data needs to be migrated.
- Features of the distributed storage:**
 - High scalability: Based on standard hardware, the distributed storage supports multiple types of storage protocols and models.
 - High elasticity: Based on the distributed architecture, the number of storage nodes reaches several thousand, with an amount of EB-level data.
 - Flexible capacity expansion: Capacity expansion is conducted based on standard hardware.

High-Performance Computing

- High Performance Computing (HPC) is a branch of the computer science. HPC improves the computing speed to a manner of tera operations per second (TOPS) through a cluster architecture, parallel algorithm, and the parallel/distributed computing of related software, which cannot be achieved by a single computer.
- The HPC system supports software and hardware collaboration. A typical architecture of the system includes infrastructure, compute nodes, storage and file systems, network switching, cluster management, and resource scheduling.



33 Huawei Confidential



- HPC refers to the aggregation of the computing power to perform computing tasks, such as simulation, modeling, and rendering, which are beyond the capacity of standard workstations. In recent years, HPC is often considered equivalent to a computer cluster system, which uses the high-speed interconnection technology to connect multiple computer systems and handles large-scale computation problems with comprehensive computing capabilities of all the connected systems. In this sense, HPC is usually called HPC cluster.
- A cluster refers to a group of computers which provide network resources for users as a whole. Each computer of a cluster is considered as a node, which can be added or deleted. A computer is virtualized based on these nodes for users. From users' viewpoint, they only care about the computer being used, without considering the nodes.
- HPC uses high-end hardware or aggregates the computing power of multiple units and provides the ultra-high floating point computing capability solution to satisfy the computing requirements of computing-intensive, network-intensive, and data-intensive services, covering scientific research, weather forecasting, computational simulation, military research, CAD/CAE, biopharmaceuticals, gene sequencing, and image processing. This greatly shortens computing time and improve computing accuracy.
- In general, the HPC solution consists of hardware, including servers, storage devices, and switches, and software, including cluster software and application software.
- The main purpose of building an HPC system is to improve the computing speed. To improve the computing speed to a manner of tera operations per second (TOPS), high requirements are imposed on the system processor, memory bandwidth, computing method, system I/O, and storage. Each of these directly affects the computing speed.

HPC Power Measurement

- HPC power is measured by floating-point operations per second (FLOPS).

Rpeak is determined by the specification and number of CPUs.

$Rpeak = \text{CPU frequency (standard frequency)} \times \text{number of floating-point operations in each CPU clock period} \times \text{number of cores in the CPU}$

Rmax refers to the maximum performance of HPL in the entire cluster.

HPL efficiency measures the computing efficiency of the entire cluster.

$$\text{HPL efficiency} = \frac{Rmax}{Rpeak}$$

1 K FLOPS = one thousand FLOPS = 10^3
1 M FLOPS = one million FLOPS = 10^6
1 G FLOPS = one billion FLOPS = 10^9
1 T FLOPS = one trillion FLOPS = 10^{12}
1 P FLOPS = one FLOPS = 10^{15}
1 E FLOPS = ten thousand FLOPS = 10^{18}
1 Z FLOPS = ten million trillion FLOPS = 10^{21}

- HPL: The High-Performance Linpack Benchmark.

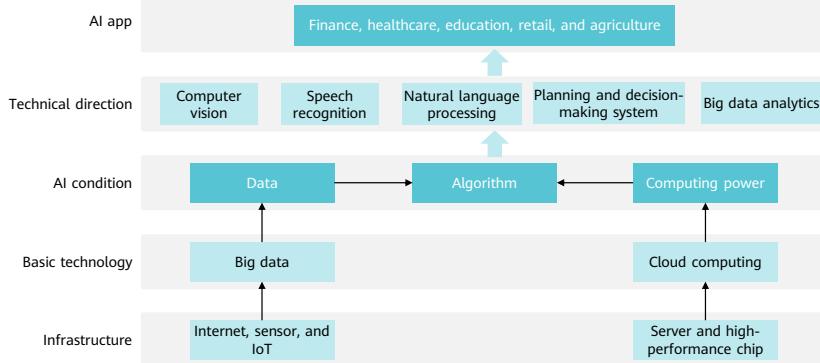
Introduction to AI

- AI is a technical science that studies and develops theories, methods, and applications for simulating and extending human intelligence.
- Machine learning simulates and implements human learning behaviors to obtain new knowledge. It is one of the core research areas of artificial intelligence.
- Deep learning originates from the research of artificial neural network. A multilayer sensor is a deep learning structure. Deep learning is a new research field in machine Learning. It simulates the mechanisms of the human brain to interpret data, such as the recognition of images, voice, and texts.



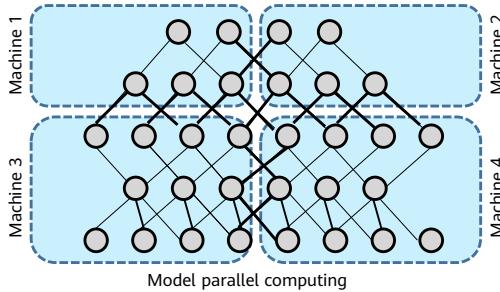
AI Industry Ecosystem

- The four elements of AI are the data, algorithm, computing power, and scenario. To meet the requirements of the four elements, AI is integrated into cloud computing, big data, IoT, and other industries.
- In the AI industry, networks are expected to provide the high-speed communication between computing nodes.



Network Requirements of AI Computing

- The development of machine learning and deep learning is accompanied by powerful computing requirements, which can hardly be met by only one computer. As such, distributed compute clusters are often established.
- Enterprises, such as Facebook, Baidu, and Alibaba, proactively build the machine learning and deep learning platform, which is usually built by 100 Gbps and faster network devices. The AI performance test result shows that networks can seriously affect the computing performance. In model parallel computing, each node computes one part of the algorithm. After the computing is complete, all data shards need to be transmitted to other nodes.



Network requirements of AI computing:

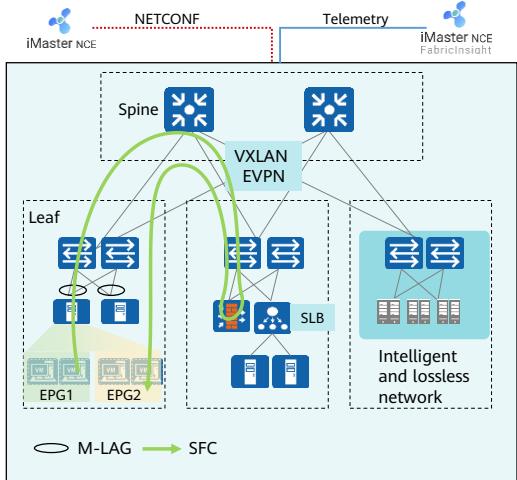
- High bandwidth, low delay, no packet loss
- Traffic control in the incast scenario
- Congestion control with quick responses
- Fast and efficient load balancing mechanism
- Differentiated scheduling of hybrid traffic

- For more information, see *Huawei AI certification*.

Contents

1. DC Overview
2. Data Center Network Overview
- 3. Overview of Key DC Technologies**
 - DC Key IT Technologies
 - DC Key Network Technologies

Overview of Key DCN Technologies



- There are multiple network technologies applied on DCNs. This course describes the following key DCN technologies:
 - SLB
 - M-LAG
 - VXLAN
 - EVPN
 - Telemetry
 - NETCONF
 - Microsegmentation
 - SFC
 - Intelligent and lossless network technologies

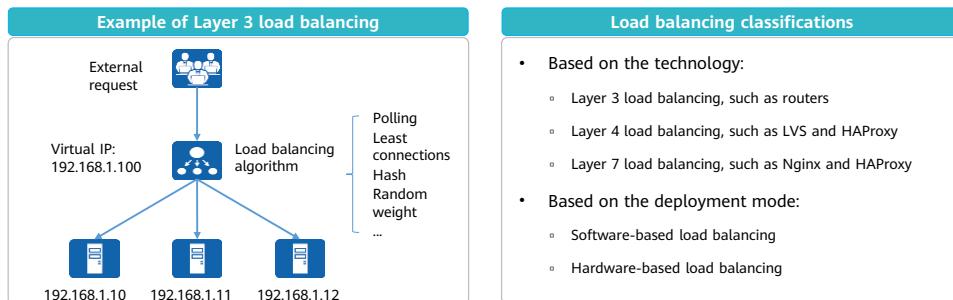
39 Huawei Confidential



- M-LAG: Multichassis Link Aggregation Group
- EVPN: Ethernet Virtual Private Network
- NETCONF: Network Configuration Protocol

Introduction to Load Balancing

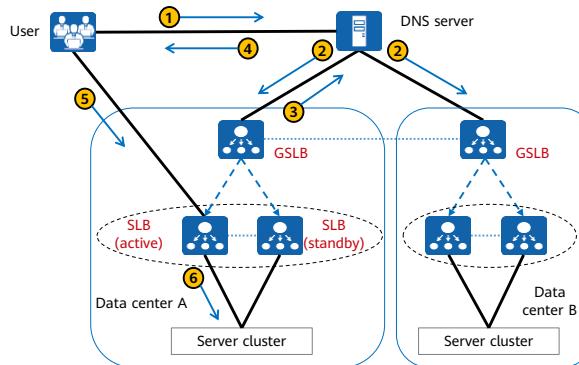
- Load balancing is a technology with which computer clusters can allocate loads. Proxy devices are used to receive external requests and share them to multiple internal servers. The proxy device is called load balancer.
- Layer 3 load balancing means IP-based load balancing. Similarly, Layer 4 load balancing means load balancing based on IP addresses and port numbers and Layer 7 load balancing means load balancing based on the application layer protocol (such as HTTP).



- The IP-based load balancing is called virtual IP in this example and is called floating IP on OpenStack.
- Load balancing algorithms determine healthy servers at the back end to be chosen. Common algorithms include:
 - Round robin: Select the first server in the first request list, and scroll the list downwards in order in a circular manner for conducting preceding requests.
 - Least connections: The server with the least connections is preferred.
 - Hash: A hash is created after the hash calculation of the requested source IP address and requested are sent to a certain server based on the hash.
 - Random LoadBalance: weight-based random allocation.

Load Balancing Applications in DCs

- Load balancing applications in DCs include server load balancing (SLB) and global server load balancing (GSLB). The former implements server load balancing within DCs and the latter implements load balancing between DCs.

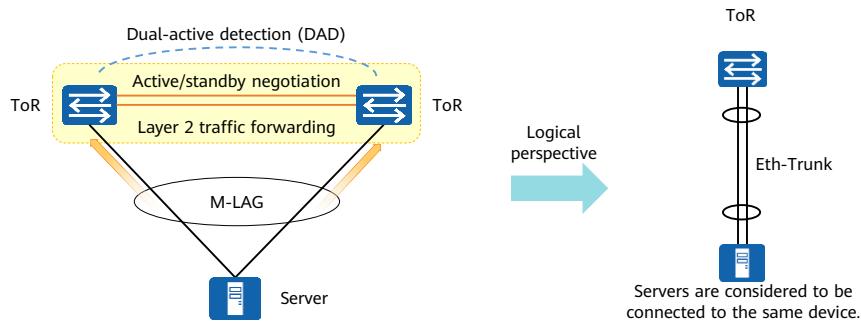


1. The user accesses <https://www.huawei.com/en/>, and applies for address resolutions from the DNS server.
2. The DNS server forwards the query request to the GSLB for resolution.
3. GSLB selects the optimal result, that is, the virtual IP address provided by the SLB, and sends the result to the DNS server.
4. The DNS server returns the user result.
5. The user accesses the virtual IP address provided by the SLB.
6. The SLB forwards the request to the specified server.

- The application scenario of GSLB is that enterprises establish multiple DCs in different areas. Users can access the nearest DC based on their locations. There are multiple GSLB solutions. This example describes the DNS-based GSLB solution that is used most widely within DCs.
- In the GSLB solution, domain name service providers forward the name server (NS) records of domain names to GSLB devices with smart DNS resolution functions and the records are resolved by GSLB devices. If GSLB devices are deployed in multiple places, they should all be added to the NS record to provide high availability. GSLB devices perform health checks to back-end servers and public IP addresses of other DCs. The results will be synchronized between GSLB devices of different DCs through proprietary protocols. Eventually, GSLB devices choose the optimal address resolution for DNS servers based on the GSLB policy and DNS servers send the optimal address to the user.
- Based on the differences of user requests, SLBs in a data center distribute the requests to multiple, hundreds, or even thousands of devices at the back end and ensures that the system selects the optimal server to process the requests according to the previously defined policy, which improves the availability and scalability of applications to some extent.

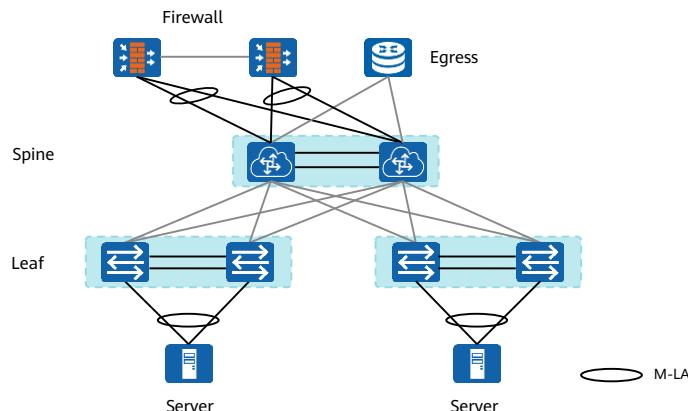
Introduction to M-LAG

- Multichassis link aggregation group (M-LAG) is an inter-device link aggregation technology. M-LAG improves link reliability from the board level to the device level. M-LAG provides traffic load balancing and backup protection.
- In a DC, M-LAG is established through the active/standby negotiation of two ToR switches, responsible for the access of other devices (such as servers and firewalls).



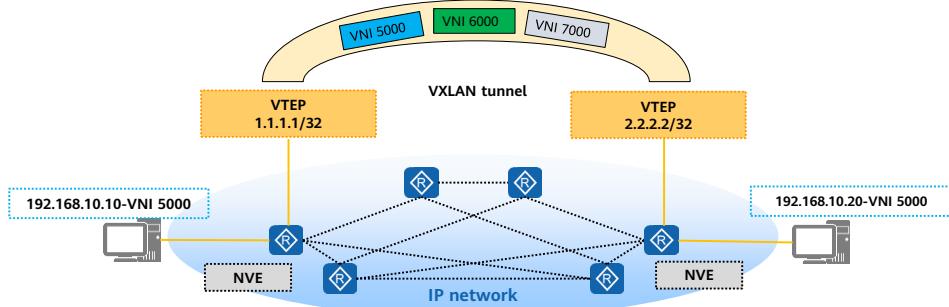
M-LAG Applications in DCs

- The current DCs usually adopt the spine-leaf architecture. In order to meet the requirement for high reliability, M-LAG is recommended for server or firewall access.



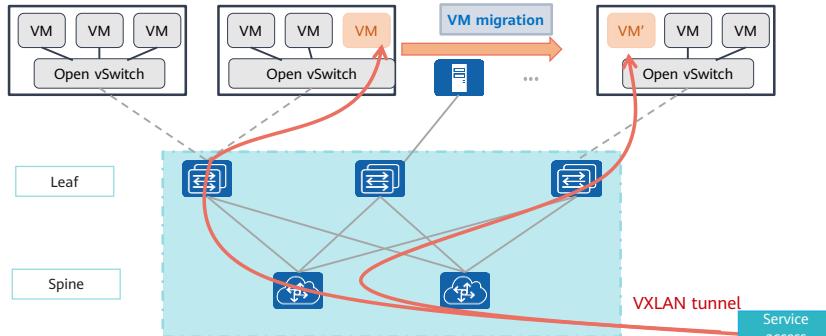
Introduction to VXLAN

- VXLAN is a VPN technology that can build a Layer 2 virtual network over a physical network with reachable routes. Routed networks relied on by the underlying VXLAN layer are not limited by the network architecture and support strong scalability.
- VXLAN packets contain some VXLAN network identifier (VNI) fields, which are similar to the VLAN ID and are used to identify different networks. Between two devices, there is only one VXLAN tunnel, which is similar to the Trunk link and is used to carry the permitted traffic of all the VNIs between devices.



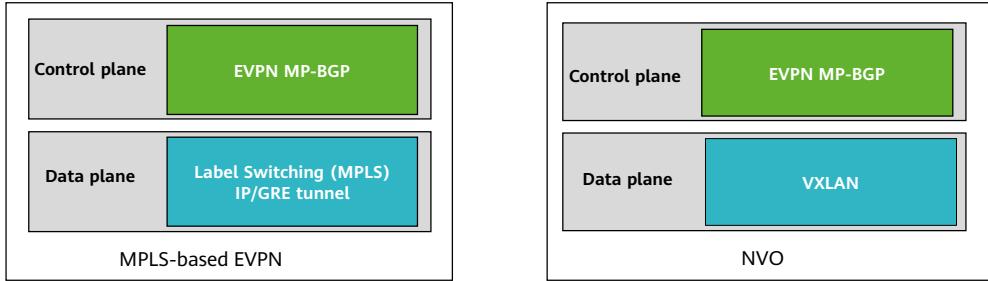
VXLAN Applications in Cloud DCs

- Many services are deployed on VMs in cloud DCs. VMs can be live migrated randomly in a server cluster. When VMs are migrated to a server under another leaf node, IP addresses and MAC addresses of VMs remain unchanged to prevent service interruptions.
- Spine-Leaf + VXLAN is the best practice in this scenario.



Introduction to EVPN

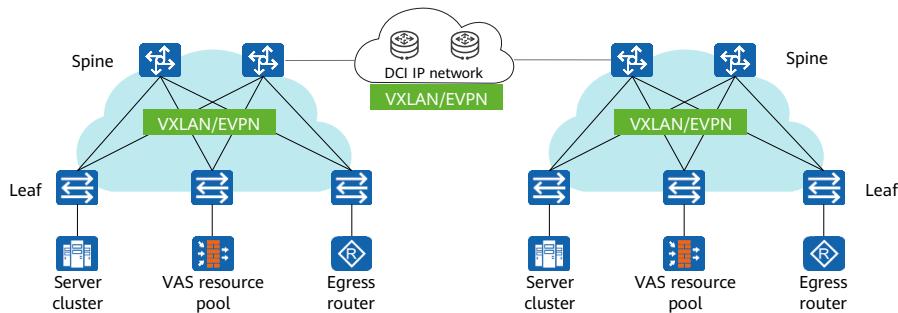
- Ethernet VPN (EVPN) is initially defined in RFC-7432. The MPLS-based VPN meets the requirements of high bandwidth and complicated QoS scheduling.
- Virtualization technologies are introduced into cloud DCs. As such, a host can carry multiple VMs which belong to different tenants. This raises new requirements for the network. As such, the network virtualization overlay (NVO) solution is adopted.



- **NVO:** Network Virtualization Overlay, A logical network is built on the existing IP network to shield differences between underlying physical networks and virtualize network resources. In this way, multiple logically isolated network partitions and multiple heterogeneous virtual networks can coexist on the same shared network infrastructure.

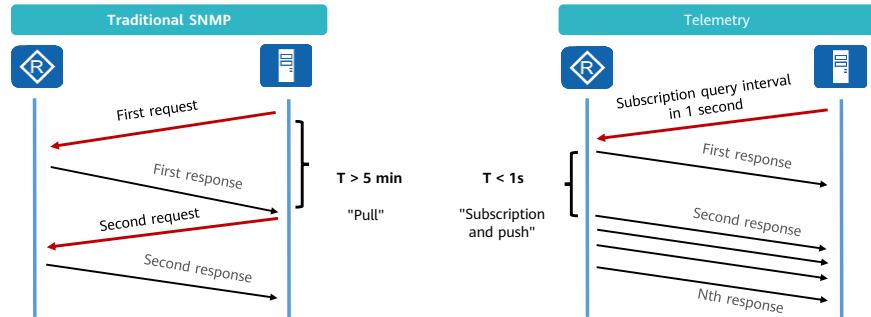
EVPN Applications in DCs

- The NVO solution is applied in DCs, that is, BGP EVPN works as the control plane to transmit routing information within a DC and between DCs and VXLAN works as the data forwarding plane to forward data packets.
- EVPN supports traffic transmission between Layer 2 and Layer 3 within a DC and between DCs.



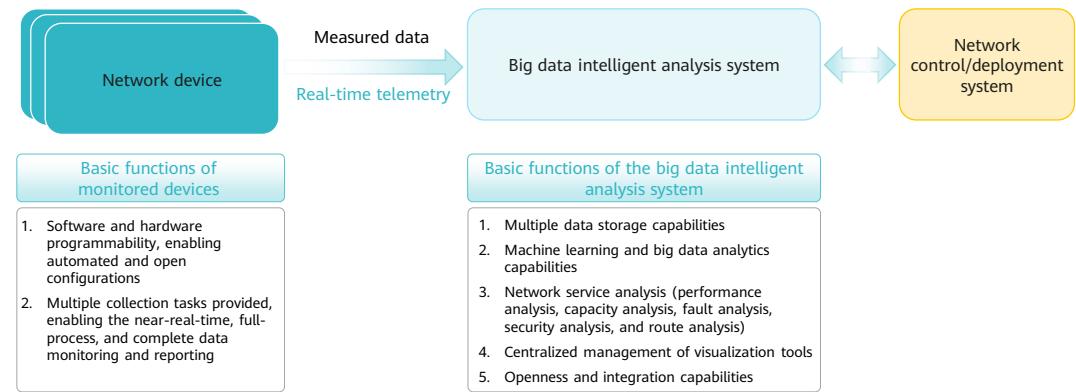
Introduction to Telemetry

- Telemetry, also known as network telemetry, is a technology that remotely collects data from physical or virtual devices at a high speed.
- Compared with SNMP, the telemetry is at the subsecond level in terms of the collection interval. A telemetry-enabled device proactively sends information in push mode, implementing real-time, high-speed, and precise data collection.



Telemetry Applications in DCs

- DCN collects high-precision device data based on the telemetry technology to build an intelligent O&M system.



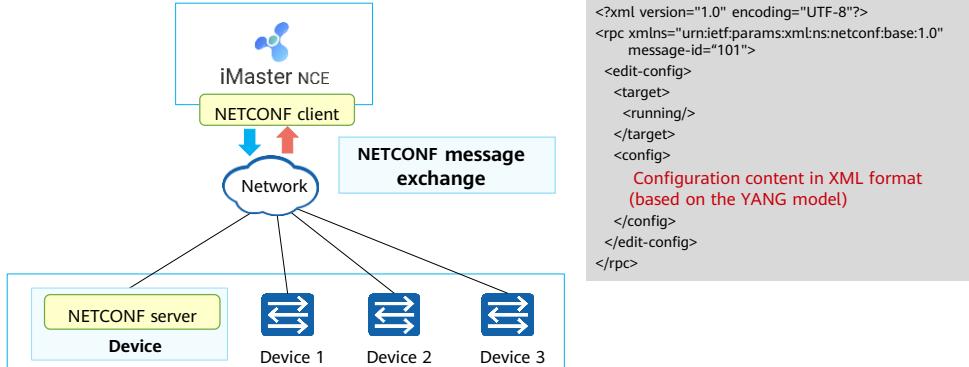
Introduction to NETCONF

- Network Configuration Protocol (NETCONF) provides a mechanism for managing network devices. To be specific, users can use NETCONF to add, modify, and delete configurations of network devices, as well as obtain configurations and status of network devices.
- Compared with CLI and SNMP, NETCONF has the following advantages in device configuration:

Function	NETCONF	SNMP	CLI
Interface type	Machine-machine interface: The interface definition is complete and standard, and the interface is easy to control and use.	Machine-machine interface	Man-machine interface
Operation efficiency	High: Data is modeled based on objects. Only one-time interaction is required for operations on an object. Operations such as filtering and batch processing are supported.	Medium	Low
Extended capability	Proprietary protocol capabilities can be extended.	Low	Medium
Transaction processing	Supported: transaction processing mechanisms such as trial running, rollback upon errors, and configuration rollback	Not supported	Partially supported
Secure transmission	Multiple security protocols: SSH, TLS, BEEP/TLS, and SOAP/HTTP/TLS	Available only in SNMPv3	SSH supported

NETCONF Applications in DCs

- In a DC, NETCONF is mainly used by a network controller to orchestrate services and deliver configurations to southbound devices.
- NETCONF messages are encoded in XML format, including the standard YANG model.



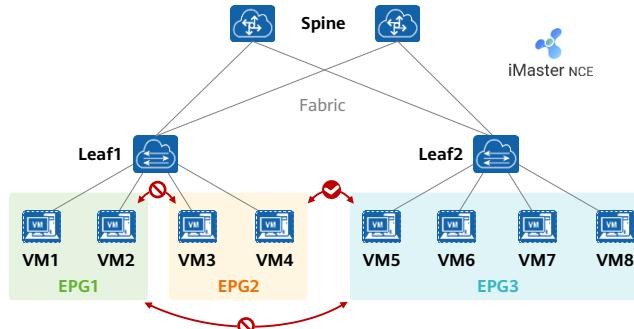
Introduction to Microsegmentation

- Microsegmentation is a security isolation technology that groups DC service units based on certain rules and deploys policies between groups to implement traffic control.
- Traditionally, subnets are created for DCs based on coarse-grained granularities such as VLAN IDs or VNIs. Microsegmentation supports more fine-grained and flexible grouping modes, for example, grouping based on IP addresses, MAC addresses, and VM names. This can further narrow down security zones to implement fine-grained service isolation and enhance network security.



Microsegmentation Applications in DCs

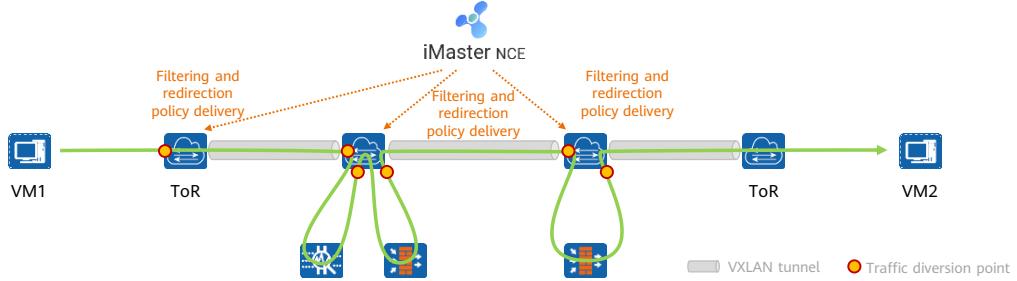
- In a DC, microsegmentation classifies servers or VMs into groups and defines access control policies between different groups to implement traffic control between service nodes.
- Microsegmentation can be implemented on a standalone CE switch or on a CE switch and an iMaster NCE-Fabric controller.



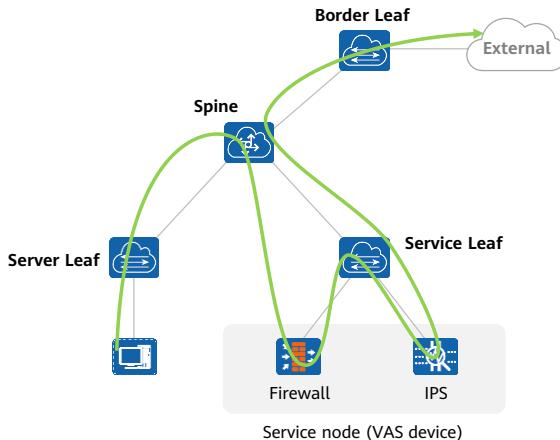
- Microsegmentation implements service isolation between different servers of a VXLAN network and ensures secure management and control for the VXLAN network. At the same time, the configuration and maintenance are simple, significantly reducing the costs.
- For more information, see *Technical Principles and Applications of Microsegmentation and SFC*.

Introduction to SFC

- Service Function Chaining (SFC) technology provides ordered services for the application layer.
- SFC creates a chain of service functions (SF), usually value-added service (VAS) devices, along which matched traffic passes through to obtain VASs. Typical VAS devices are firewalls, load balancers, deep packet inspection (DPI) devices, and intrusion prevention devices.
- iMaster NCE-Fabric can be used to directly orchestrate SFCs, which can be achieved through PBR or NSH.



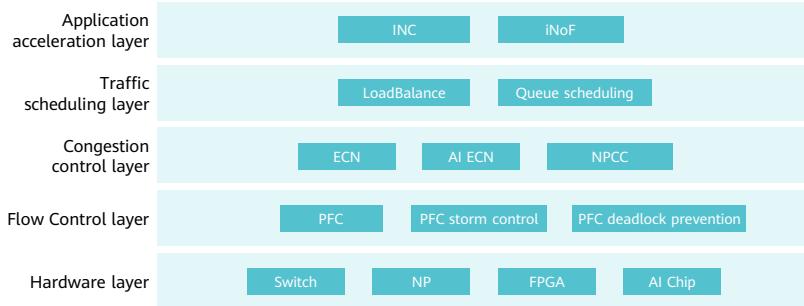
SFC Applications in DCs



- When data packets are transmitted on DCNs, they need to pass through various service nodes to ensure that DCNs flexibly divert traffic to the service nodes as planned, thus providing VASs for users. Typical service nodes are firewalls, intrusion prevention systems (IPS), and load balancers.
- With SFC, differentiated VASs can be provided on a network.

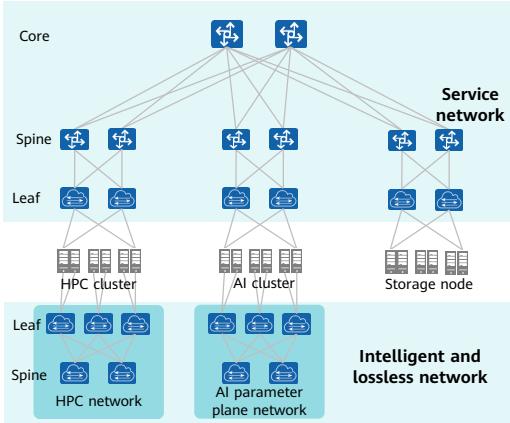
Introduction to the Intelligent and Lossless Network Technologies

- In DCs, lossy networks cannot satisfy the requirements of high-performance systems. An intelligent and lossless network uses the AI-ready hardware architecture and iLossless algorithm to maximize the throughput and minimize the latency without packet loss.
- The intelligent and lossless network technology architecture has five layers, which will be detailed later in the following courses.



- Flow control: matches traffic rates between the sender and the receiver to ensure zero packet loss.
- Congestion control: ensures the maximum throughput and minimum latency by controlling traffic rates in the case of network congestion.
- Traffic scheduling: implements load balancing for service traffic and network links to ensure the quality of different service traffic.

Intelligent and Lossless Network Applications in DCs



- Service networks are often deployed as TCP/IP lossy networks.
- Different industries or enterprises have different zone division classifications.

In DCs, the intelligent and lossless network solution is applicable different computing scenarios:

- For example, the HPC network and the distributed AI training network are generally deployed as closed networks.
- A two-layer or three-layer spine-leaf networking architecture is selected based on the access node scale of the cluster, and appropriate leaf switches and spine switches are selected based on the port bandwidth.



Quiz

1. (True or false) iMaster NCE-Fabric sends NETCONF messages to deliver configurations to network devices and NETCONF messages are encoded in XML format.
 - A. True
 - B. False
2. (Multiple-answer question) Which of the following devices are included in a DC? ()
 - A. Environmental control devices, such as air conditioners
 - B. Security devices, such as platform screen doors (PSDs)
 - C. IT devices, such as servers
 - D. Communication devices, such as switches and routers

1. A
2. ABCD

Summary

- As the closest area to the network industry and the computing industry, DCNs should quickly respond to IT requirements, featuring complicated and integrated structures and rapid technological development.
- This is the first course of the DCN series courses. You will understand what a DC and a DCN is, as well as their development histories.
- We will analyze more DC technical principles in detail to help you understand the hyper-converged DCN.

Thank you.

把数字世界带入每个人、每个家庭。
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



Technical Principles and Applications of Virtualization



Foreword

- Virtualization is defined in different ways by network engineers and IT engineers. With the convergence and application of cloud network technologies in data centers (DCs), it is important to explain basic concepts from different perspectives and clarify their differences.
- Server virtualization is a technology that allows multiple virtual servers to run on one physical server. From the perspective of IT engineers, server virtualization includes compute virtualization, storage virtualization, and network virtualization.
- From the perspective of network engineers, network virtualization refers to network device and network architecture virtualization technologies, such as stacking, Multichassis Link Aggregation Group (M-LAG), virtual system, and Virtual Extensible LAN (VXLAN), instead of server virtualization.
- This course introduces the background and related technologies of server virtualization, and further explains "network virtualization" in the eyes of IT and network engineers.

Objectives

- On completion of this course, you will be able to:
 - Describe the background of server virtualization.
 - Describe the applications of network virtualization in DCs.
 - Describe the technical fundamentals of network virtualization in server virtualization.

Contents

- 1. Server Virtualization**
 - Background
 - Technical Fundamentals
 - Deployment
- 2. Network Virtualization
- 3. Introduction to FusionCompute

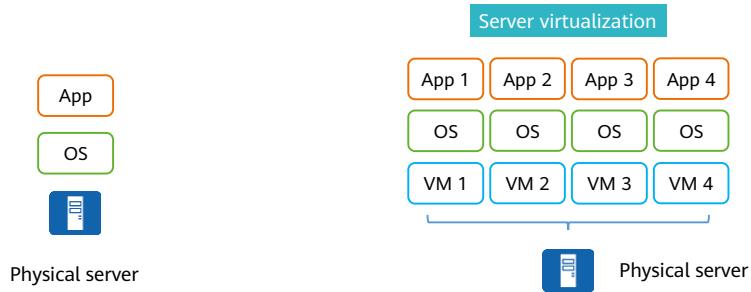
Overview and Objectives

- This section describes what is (server) virtualization from the perspective of IT engineers, what technologies are involved in (server) virtualization, and how to deploy server virtualization.
- You can learn the definition, development history, and key technologies of server virtualization.

- In this course, virtualization from the perspective of IT engineers is referred to as server virtualization.

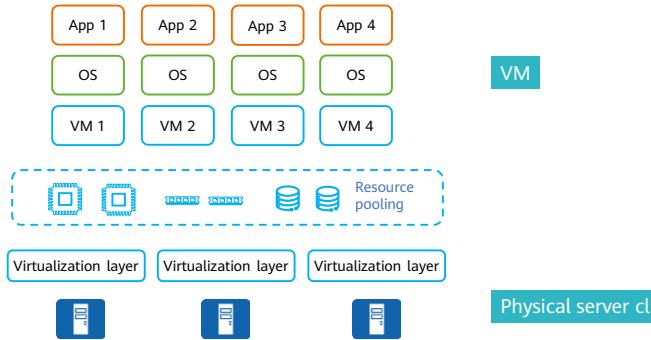
Server Virtualization Definition (1)

- Server virtualization is a technology that creates multiple virtual machines (VMs) on a physical server. It brings various benefits, including efficient physical resource utilization, rapid service provisioning, and elasticity.
- Each VM can run its own applications and services and act as a physical server.



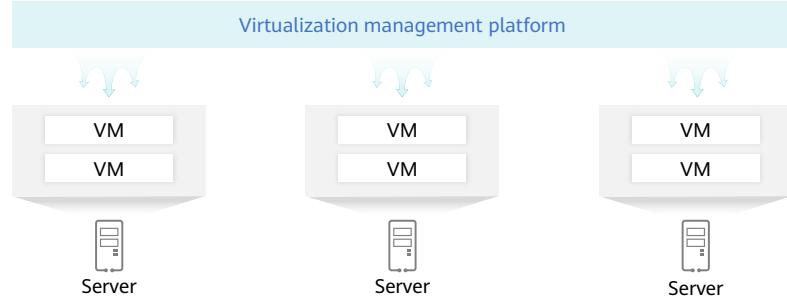
Server Virtualization Definition (2)

- As the clustering technology emerges, server virtualization provides the ability to have multiple physical servers operated in a cluster, which acts as a virtual resource pool.
- VMs can be migrated between physical servers in a cluster. This further unlocks flexibility, elasticity, and high availability of server virtualization.



Virtualized Server Cluster Management

- As services grow, the number of VMs in a cluster reaches hundreds to thousands. Therefore, a virtualization management platform is required for centralized management.
- The virtualization management platform provides a simple user interface and various functions, such as monitoring and managing virtualized resources. It simplifies VM creation and helps users configure and execute resource scheduling policies.



8 Huawei Confidential

 HUAWEI

- Different vendors have their own virtualization management platforms since they use different virtualization technologies, such as vCenter of VMware, FusionCompute VRM of Huawei, SystemCenter of Microsoft, and RHEV of Red Hat.

Server Virtualization Benefits

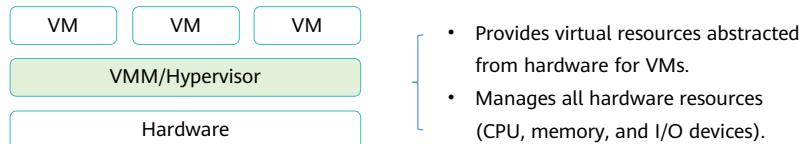
- Increased resource utilization: Without virtualization, servers in a DC use only 5% to 30% of their resources during normal operation. After virtualization, the utilization of virtualized server resources is dramatically improved to more than 60%.
- Reduced costs: Server virtualization provides the time-sharing feature for resources and allows dynamic adjustment of cluster resources. As such, DCs require fewer servers and less equipment room space and power.
- Improved flexibility: Clustering allows elastic VM provisioning and can flexibly cope with service requirements in peaks and off-peaks.
- Less system breakdown: High availability (HA) for VMs helps prevent VM services from being affected due to a faulty physical server.

Contents

- 1. Server Virtualization**
 - Background
 - Technical Fundamentals
 - Deployment
- 2. Network Virtualization
- 3. Introduction to FusionCompute

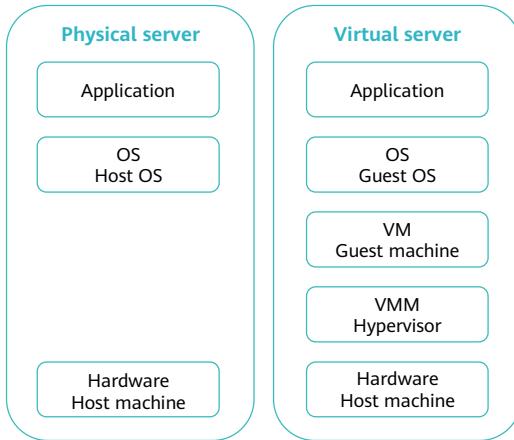
Server Virtualization Technologies

- There are three kinds of server virtualization: compute virtualization, storage virtualization, and network virtualization.
- A hypervisor, also known as a virtual machine monitor (VMM), is introduced to compute virtualization. It abstracts hardware into virtual resources to allow an OS to run directly on each VM. In this way, multiple OSs can run on a single physical server at the same time.
- A hypervisor virtualizes the following physical resources: CPU, memory, and input/output (I/O) resources.



- A CPU (Central Processing Unit) is one of the main devices of a computer, and a function of the CPU is to interpret computer instructions and process data in computer software.
- A hypervisor provides the following basic functions: Identify, capture, and respond to privileged CPU instructions or protection instructions sent by VMs (the privileged instructions and protection instructions will be described in the CPU virtualization section); schedule VM queues and return physical hardware processing results to related VMs.

Compute Virtualization - Basic Concepts



Guest OS: OS running on a VM

Guest machine: VM virtualized on a physical server

Hypervisor: virtualization software layer

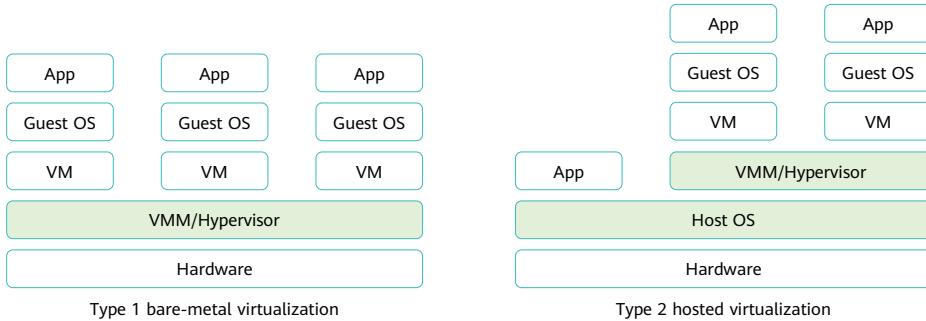
Host OS: OS running on a physical server

Host machine: physical server

- A host machine is a physical host that can run multiple VMs, and an OS installed and running on the host machine is a host OS. VMs running on a host machine are called guest machines. The OS installed and running on a VM is called a guest OS. The core of virtualization technologies is a hypervisor between the host OS and guest OS. It can also be called Virtual Machine Manager (VMM).
- In the physical architecture, a host uses a two-tier architecture from bottom to top: hardware (host machine) and host OS. Applications are installed on top of the host OS. In the virtualization architecture, a host uses a three-tier architecture from bottom to top: hardware (host machine), hypervisor, and guest machine installed with a guest OS. Applications are installed on the guest OS. Multiple guest machines can operate on a host machine.

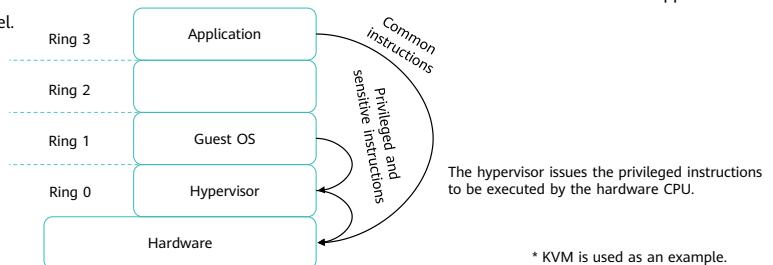
Compute Virtualization Technologies

- Virtualization can be implemented in two modes based on hypervisor deployment locations, referred to as Type 1 (or bare metal) and Type 2 (or hosted).
- Type 1 virtualization has a hypervisor run directly on the hardware, without the need of a host OS, while Type 2 virtualization lets a hypervisor run as a software program.



Compute Virtualization: CPU Virtualization

- A host OS sends three types of instructions: privileged instructions and common instructions in the physical scenario, and sensitive instructions specific to the virtualization scenario.
- Hierarchical protection domains, often called protection rings, are defined for CPU instructions. A CPU has four rings, numbered from 0 through to 3. Ring 0 is the most privileged level and interacts directly with the hardware. Ring 3, the least privileged ring, is where most applications reside.
- For example, when Kernel-based Virtual Machine (KVM) is used for CPU virtualization, guest OSs send all instructions to the hypervisor, and then the hypervisor schedules the instructions to the CPU for execution. Common instructions from applications are executed at the non-privilege level.

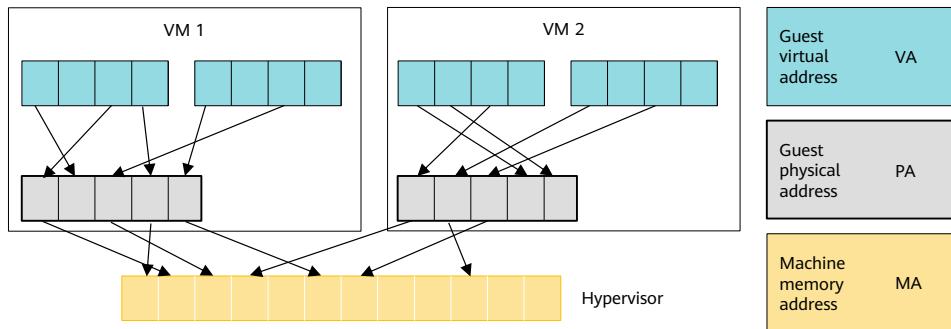


* KVM is used as an example.

- There are four CPU hierarchical protection domains, also called protection rings, numbered 0 (most privileged) to 3 (least privileged). Ring 0 has direct access to the hardware. Generally, only the OS and driver have this privilege. Ring 3 has the least privileges. All programs can run in Ring 3. To protect computers, some dangerous instructions can only be executed by the OS, preventing malicious software from randomly calling hardware resources. For example, if a program needs to enable a camera, the program must request the driver in ring 0 to enable the camera. Otherwise, the operation will be rejected.
- The instructions sent by a host OS are classified into two types: privileged instructions and common instructions.
 - Privileged instructions:** are instructions used to operate and manage key system resources. These instructions can be executed only at the highest privilege level, that is, Ring 0.
 - Common instructions:** are instructions that can be executed at the non-privilege level, that is, Ring 3.
- In a virtualization environment, another special instruction type is called sensitive instruction. A sensitive instruction is used for changing the operating mode of a VM or the state of a host machine. The instruction is handled by VMM after a privileged instruction that originally needs to be run in Ring 0 on the guest OS is deprived of the privilege.
- CPU virtualization can be further classified into full virtualization, para-virtualization, and hardware-assisted virtualization. For details, see *HCIA-Cloud Computing*.

Compute Virtualization: Memory Virtualization

- Memory virtualization centrally manages physical memory of a physical server and divides the physical memory into multiple virtual memories for VMs. As shown in the figure, the memory addresses of VMs are contiguous.



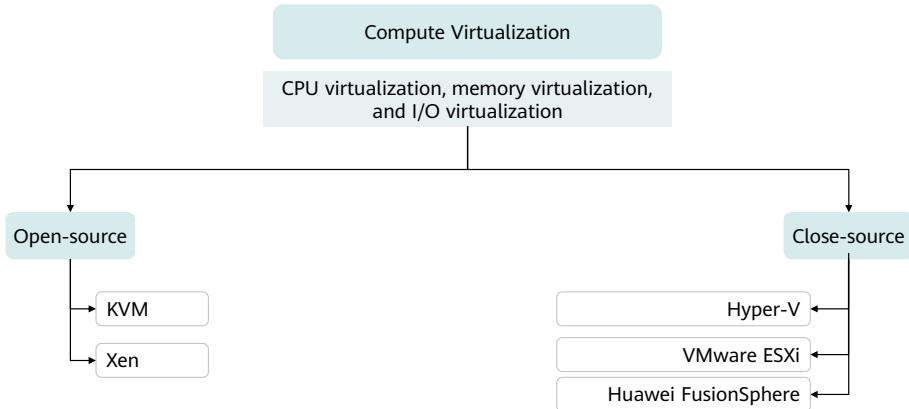
- Generally, a physical host uses the memory address space as follows:
 - The memory address space starts from the physical address 0.
 - Addresses in the memory address space are assigned contiguously.
- However, after virtualization is introduced, the following problems occur: There is only one memory address space that can start with the physical address 0. Therefore, it is impossible to ensure that the memory address space of all VMs on a physical host starts from the physical address 0. Although contiguous physical addresses can be assigned, this way of memory allocation leads to poor efficiency and flexibility.
- Memory virtualization involves the translation of three types of memory addresses, that is, VM memory address (VA), physical memory address (PA), and machine memory address (MA). To have multiple VMs run a physical host, addresses need to be translated in the following path: VA → PA → MA. The guest OS on a VM controls the mapping from the VA to the PA of the host memory. However, the guest OS cannot directly access the host memory. Therefore, the hypervisor needs to map the PA to the MA.
- For details about memory virtualization techniques, such as the shadow page table and huge page memory, see *Huawei Cloud Computing certification courses*.

Compute Virtualization: I/O Virtualization

- In a virtualization environment, a hypervisor implements I/O device sharing among VMs.
- The hypervisor intercepts the requests from VMs to I/O devices, simulates real I/O devices using software, and responds to the requests. In this way, multiple VMs have access to limited I/O resources.
- There are three ways to implement I/O virtualization: full virtualization, para-virtualization, and hardware-assisted virtualization.
 - Full virtualization: Software is used to simulate real hardware, such as the keyboard and mouse. Physical servers are responsible for device monitoring and simulation, resulting in poor performance.
 - Para-virtualization: Domain0, a privileged VM, is introduced to run hardware drivers. Guest OSs on other VMs access I/O devices through this privileged VM.
 - Hardware-assisted virtualization: An I/O device driver is directly installed on the guest OS, without any change to the OS. In this way, the time required for a VM to access the I/O hardware is the same as that in the traditional way. Hardware-assisted virtualization requires special hardware support, such as intelligent network interface cards (NICs).

- I/O virtualization creates a hardware middleware layer between the hypervisor and various available I/O processing units, allowing multiple guest OSs to reuse limited peripheral resources.
- I/O virtualization can be implemented in the following modes: full virtualization, para-virtualization, and hardware-assisted virtualization, among which hardware-assisted virtualization is the mainstream technology. For details, see Huawei Cloud Computing certification courses.

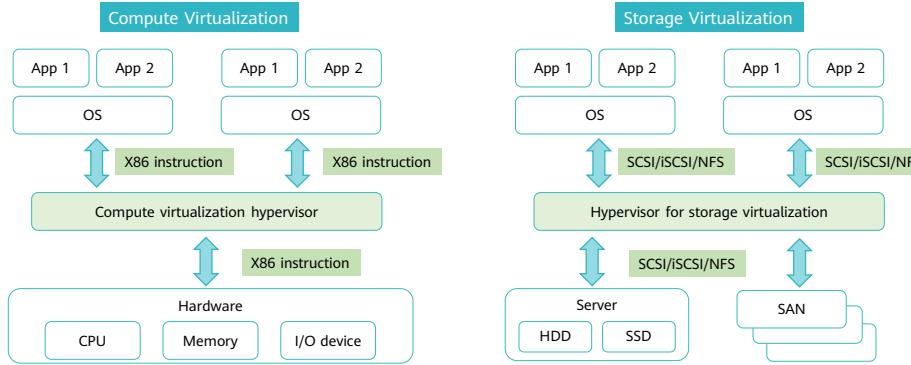
Mainstream Compute Virtualization Technologies



- There are many mainstream virtualization technologies, which can be classified into open-source (such as KVM and Xen) and closed-source virtualization technologies (such as Microsoft Hyper-V, VMware vSphere, and Huawei FusionSphere).
- Among open-source virtualization technologies, KVM is implemented in full virtualization mode and gains more popularity. Xen supports both para-virtualization and full virtualization modes, but is not widely used due to various causes. KVM, a module in the Linux kernel, is used to virtualize CPU and memory resources. It is a process running on the Linux OS. When KVM is used, QEMU is required to virtualize I/O devices (such as NICs and disks). Different from KVM, Xen directly runs on hardware, and VMs run on Xen. VMs in Xen are classified into two types: privileged VM (Domain 0) and common VM (Domain U). Domain 0 has the permission to directly access hardware and manage common VMs. Domain 0 must be started before other VMs. Domain U is a common VM and cannot directly access hardware. All operations on Domain U must be forwarded to Domain 0 through frontend and backend drivers. Domain 0 completes the operations and returns the results to Domain U.

Server Virtualization: Storage Virtualization

- In storage virtualization, the hypervisor intercepts I/O read and write instructions on the storage data plane, shields underlying storage differences, and provides storage resources for guest OSs in a unified manner.
- The significant difference between storage virtualization and compute virtualization is that storage virtualization aims to aggregate resources as a pool, instead of dividing resources as compute virtualization does.



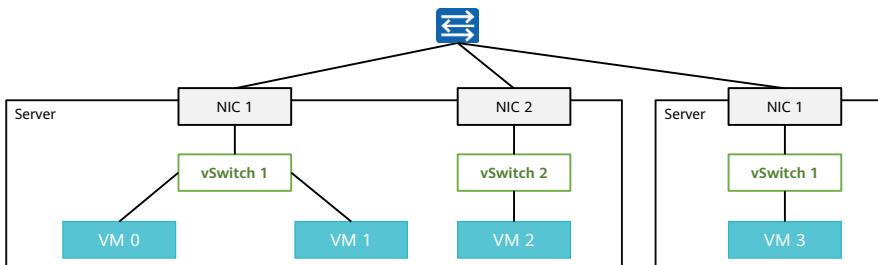
18 Huawei Confidential



- The concept of storage virtualization varies in different scenarios. From the perspective of server virtualization, storage virtualization provides storage technologies for disk mounting and file storage access of VMs. From the perspective of storage arrays, functions such as heterogeneous device management and storage gateways are also considered as storage virtualization.

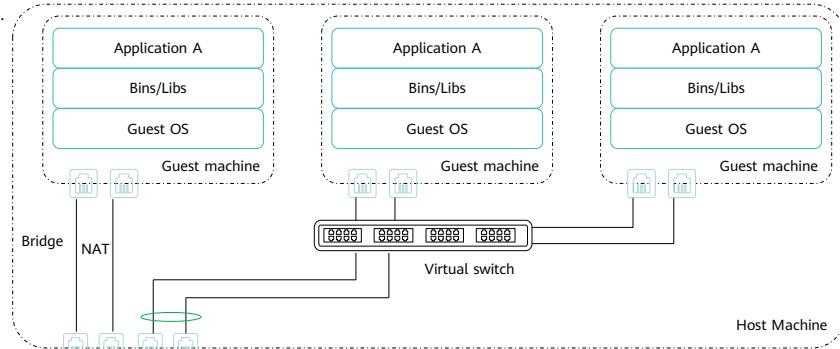
Server Virtualization: Network Virtualization

- Network virtualization creates a virtual network connecting compute and storage units. For example, this network allows communication between VMs on the same and different physical servers, and allows VM access to file systems.
- Network virtualization requires the participation of virtual network elements (NEs). In the following figure, virtual switches (vSwitches) set up a simple virtual network topology inside a server. In complex network virtualization scenarios, virtual networks may contain more virtual NEs and a network controller.



Server Virtualization: Virtual Network

- The virtual network topology varies according to network requirements of VMs. The following figure shows a simple virtual network topology. In a personal or small-scale virtualization system, VMs are bound to physical NICs using bridges or NAT. In a large-scale enterprise virtualization system, VMs are connected to physical networks through vSwitches.
- The virtual network provides VMs with various capabilities, such as Layer 2 communication, isolation, Quality of Service (QoS), and port mirroring.

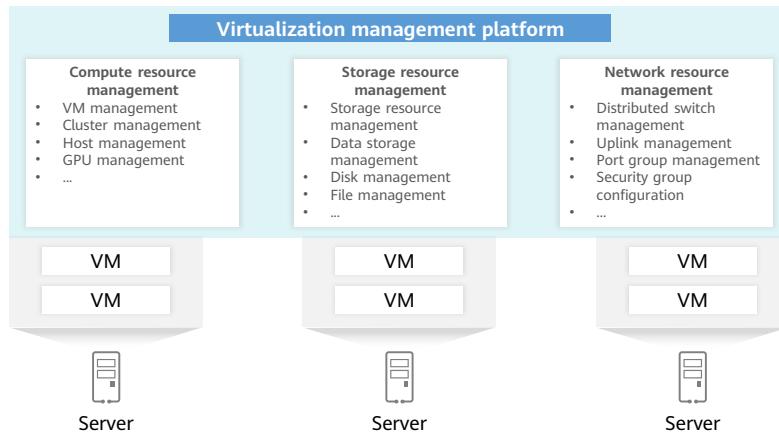


Contents

- 1. Server Virtualization**
 - Background
 - Technical Fundamentals
 - Deployment**
- 2. Network Virtualization
- 3. Introduction to FusionCompute

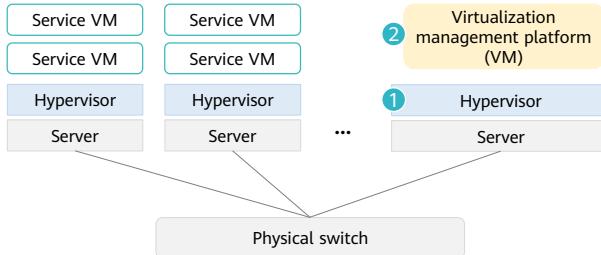
Virtualization Management Platform

Server cluster resource management and scheduling, VM operation and life cycle management.



Server Virtualization Deployment

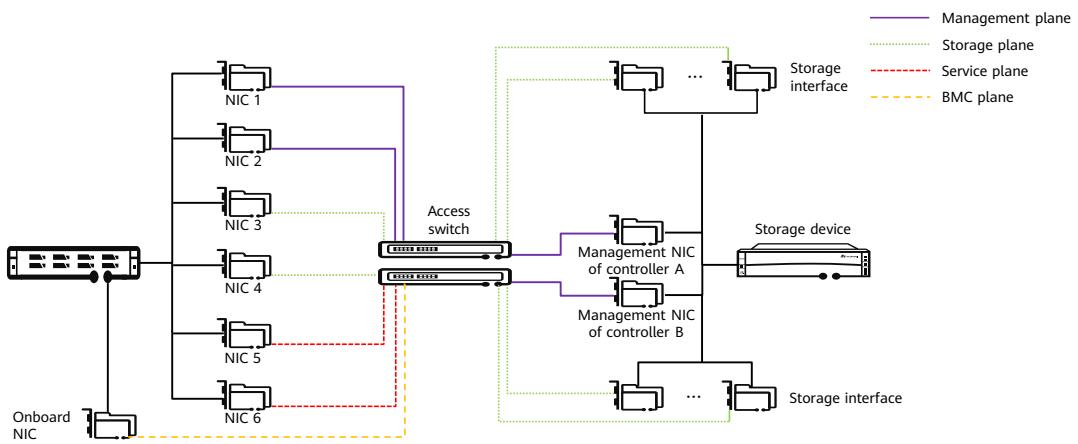
- Server virtualization requires a hypervisor and a virtualization management platform.
 - A hypervisor is deployed directly on a physical server to create VMs.
 - A virtualization management platform can be deployed as a VM on top of the hypervisor and manage all other VMs.



Virtualization management platforms and hypervisors of mainstream vendors include:

- VMware vCenter and ESXi
- Huawei CNA and VRM

Server Virtualization Topology



- Baseboard management controller (BMC) plane:
 - Plane used by the BMC network port on a host. This plane enables remote access to the BMC system of a server. It is similar to the management port of a switch.
- Management plane:
 - Plane used by the management system to manage all nodes in a unified manner and used by internal nodes for communication.
- Storage plane:
 - Network plane on which hosts communicate with storage units on storage devices.
- Service plane:
 - Plane used by service data of user VMs.

Section Summary

- A complete implementation of server virtualization requires multiple virtualization technologies working simultaneously:
 - Computing virtualization: includes CPU virtualization, memory virtualization, and I/O virtualization. A computing resource pool is used to integrate physical CPU and memory resources of a host into a computing resource pool, and then allocate virtual CPU and memory resources to provide computing capabilities for VMs.
 - Storage virtualization: The virtualization layer is compatible with various storage types. Virtual storage space provided by different storage types is integrated into storage resource pools and allocated to VMs as virtual volumes.
 - Network virtualization: provides VM NICs, virtual switches, and internal networks of servers to enable communication between VMs and between VMs and external networks.

Contents

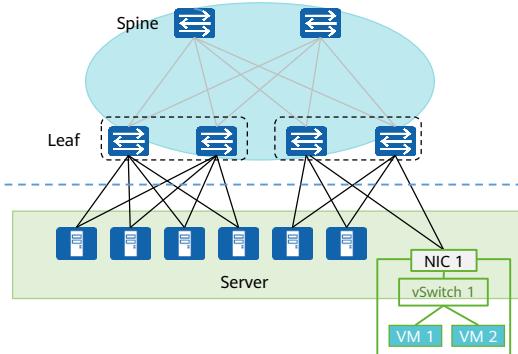
1. Server Virtualization
2. **Network Virtualization**
 - Overview
 - Fundamentals
3. Introduction to FusionCompute

Overview and Objectives

- Network virtualization focuses on virtual network configuration and connection inside servers. Traditional network engineers are unaware of this and cannot understand traffic forwarding paths from an overall perspective.
- In this section, you will learn the applications and fundamentals of network virtualization based on service traffic forwarding paths from the perspective of network engineers.

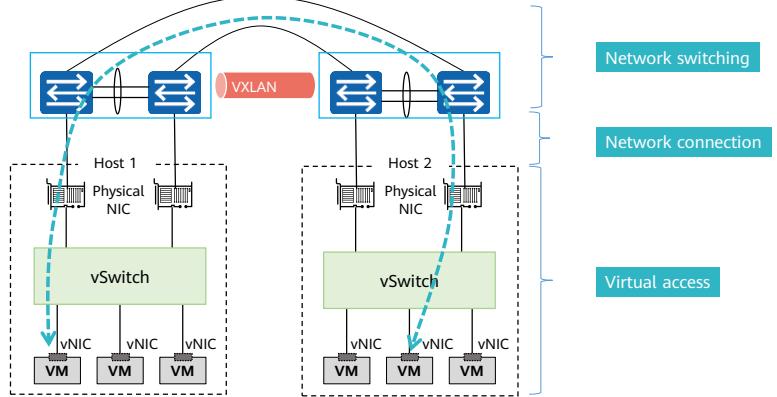
Network Virtualization in DCs

- In a DC, network virtualization mostly applies to the network layer and server layer.
 - Network virtualization at the network layer is classified into two types:
 - Device virtualization: such as stacking, M-LAG, and virtual system.
 - Network architecture virtualization: such as a large Layer 2 network in the spine-leaf architecture with VXLAN and BGP EVPN.
 - Network virtualization at the server layer: sets up virtual networks connecting virtual NEs inside servers to implement network connectivity after server virtualization.

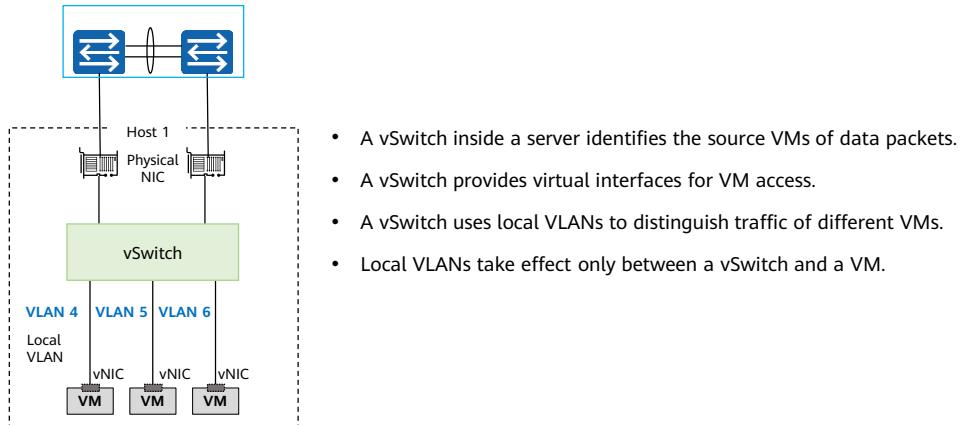


End-to-End Network Virtualization

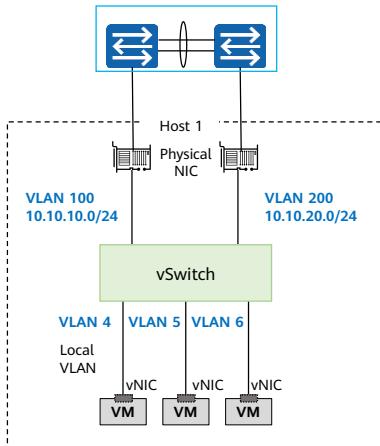
- Traffic is forwarded along the following path: VM -> vSwitch -> physical NIC (based on mappings) -> physical switch -> destination device. This process involves three phases: virtual access, network connection, and network switching.



From a VM to a vSwitch



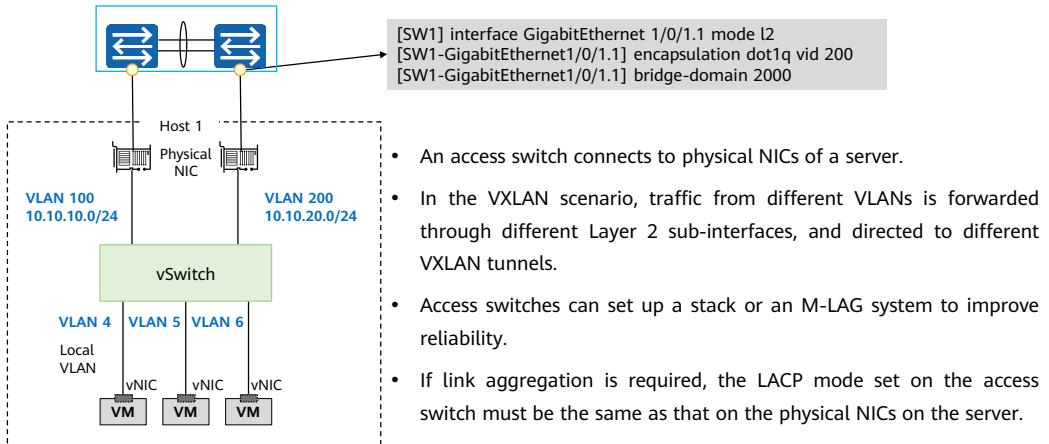
From a vSwitch to a Physical NIC



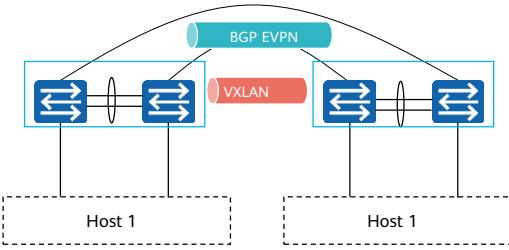
- When sending a packet received from a VM to a physical NIC, a vSwitch removes the local VLAN tag from the packet and adds a new VLAN tag that identifies the VLAN to which the physical interface corresponding to the physical NIC belongs.
- In most cases, packets from different network segments are encapsulated with different VLAN tags.
- Allowed VLANs, bonding mode, and vSwitch connection mode can be configured for physical NICs on servers.

- Bond: The Linux NIC bonding function is used to bond host network ports to improve network reliability.

From a Physical NIC to an Access Switch

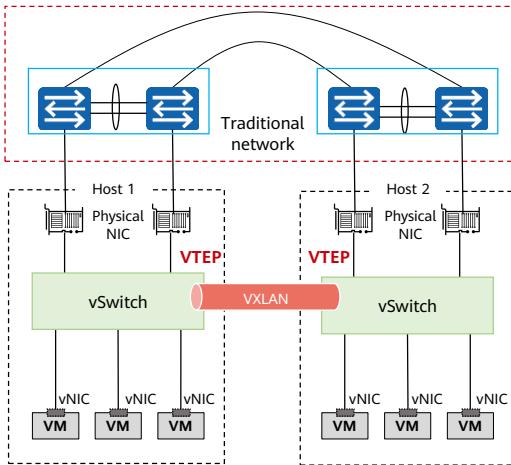


From the Local Access Switch to the Remote Switch



- VXLAN and BGP EVPN are used between switches to build a large Layer 2 network.
- On the control plane, BGP EVPN is used to transmit IP and MAC addresses of VMs, establish VXLAN tunnels, and import external routes.
- On the data plane, VXLAN tunnels are used to forward traffic.

New Traffic Model: Host Overlay



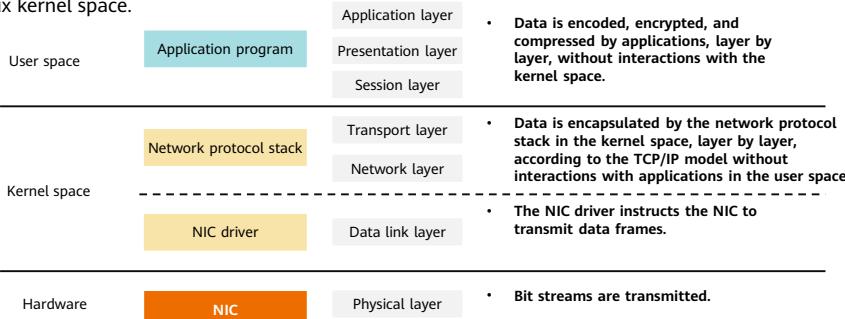
- VMs on a host use Open vSwitches (OVSS) to differentiate networks. VXLAN tunnels are established between OVSS on different hosts to set up the large Layer 2 network required by VM communication. Hardware switches only provide connectivity, and therefore require only the traditional network configuration.
- This scenario applies mostly in host overlay networking.

Contents

1. Server Virtualization
2. **Network Virtualization**
 - Overview
 - **Fundamentals**
3. Introduction to FusionCompute

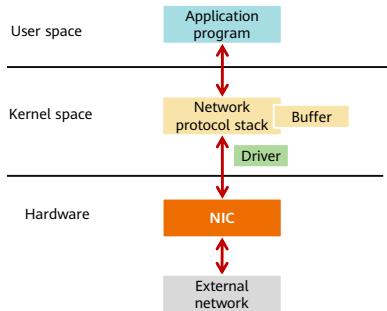
Server OS Basics

- Compared with network devices, the server OS network works in a different way but still follows the OSI model.
- Use Linux as an example. Linux consists of the user space and kernel space, also referred to as user and kernel modes, respectively. Simply speaking, the user space is where application programs run whereas the kernel space controls hardware resources to support program running in the user space. The network protocol stack runs in the Linux kernel space.



How Does a Server NIC Send and Receive Data?

- A physical NIC sends and receives data as follows (when the CPU executes data copying):
 - Sending data: The kernel reads data from the network protocol stack and writes it to the physical NIC. The NIC then sends the data to the destination external network.
 - Receiving data: Upon data receipt, the physical NIC triggers an interrupt to the CPU, which then instructs the kernel to read data and place it in the memory. The network protocol stack then parses the data.

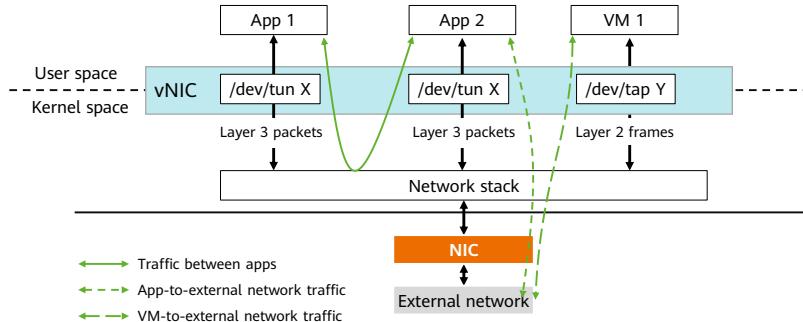


- The NIC driver needs to register the physical NIC in the kernel space so that the NIC can function properly. After registration, an NIC interface name is available.
- Interface properties, such as the IP address and mask, can be set for a physical NIC. These properties are configured in the network protocol stack of the kernel space.
- A physical NIC connects to the network protocol stack in the kernel space on one end and connects to an external network on the other end.

- Currently, an intelligent server NIC provides the Direct Memory Access (DMA) function, which allows data to be directly cached to the memory, bypassing the CPU. As such, the NIC is responsible for data transmission with the network protocol stack. After the DMA data transfer is complete, the DMA controller (DMAC) triggers an interrupt to the CPU, indicating that the transfer is completed. In this process, the CPU does not need to read or write data.

Linux Virtual Network Devices (TUN/TAP)

- The kernel can create virtual NICs (vNICs), which are similar to physical NICs, and provide NIC drivers for these vNICs to complete registration.
- TAP and TUN are vNICs defined in the Linux kernel. TUN reads and writes Layer 3 IP packets whereas TAP reads and writes Layer 2 Ethernet frames.
- A vNIC connects to the user space on one end and connects to the network protocol stack on the other end. Therefore, vNICs can neither directly send data packets to nor directly receive data packets from physical NICs.



38 Huawei Confidential

HUAWEI

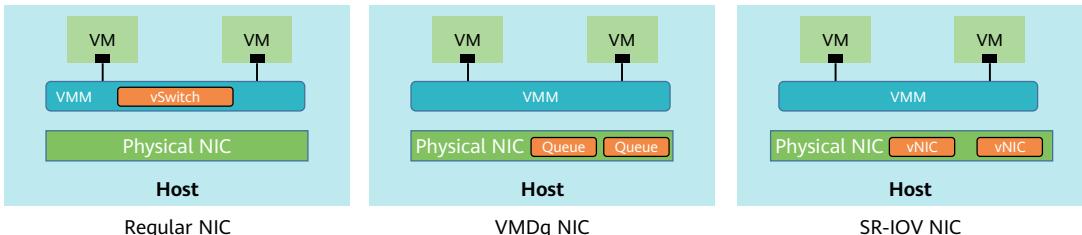
- TUN and TAP are two types of vNICs in a Linux system and provide packet reception and transmission functions. Compared with physical NICs, TUN and TAP provide almost the same functions, except that they do not provide the hardware functions of physical NICs. In addition, TUN and TAP are responsible for transferring data between the user space and the network protocol stack in the kernel space.
- In Linux, the character special files corresponding to TAP and TUN are `/dev/tapX` and `/dev/tunX`, respectively.
- TAP devices are usually used to connect to network devices, such as vSwitches. TUN devices are usually used to re-encapsulate data sourced from application programs in the user space, for example, encapsulating data using IPsec VPN.
- For more information, see
<https://www.kernel.org/doc/html/latest/networking/tuntap.html>.

vNIC vs Physical NIC

- The kernel allows same configurations on vNICs and physical NICs.
- For example, a vNIC can be configured with a MAC address, an IP address, and a subnet mask.
- Physical NICs and vNICs transfer data in different ways. Physical NICs transfer data as bit streams whereas vNICs copy data to and from the memory.

SR-IOV: Improves I/O Performance

- Single Root I/O Virtualization and Sharing Specification (SR-IOV) is a hardware-based virtualization solution that improves performance and scalability.
- SR-IOV enables efficient sharing of a physical Peripheral Component Interconnect Express (PCIe) device among VMs. This physical PCIe device can present itself as multiple virtual devices, of which each is directly attached to a VM and has an independent memory space, queues, interrupts, and command execution capability. As such, the physical PCIe device can perform direct I/O with attached VMs, achieving I/O performance that is comparable to native performance.



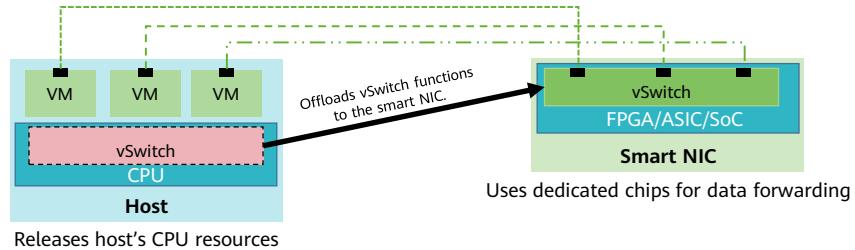
40 Huawei Confidential

HUAWEI

- SR-IOV requires special hardware.
- VMDq: virtual machine device queue.

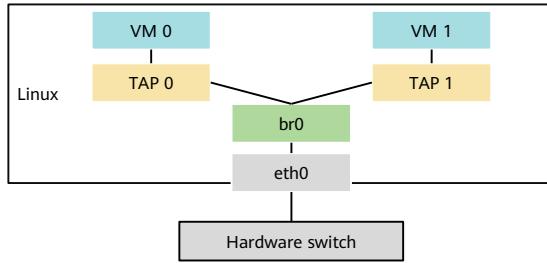
Smart NIC

- Smart NICs integrate wired networking and compute resources. They offload vSwitch functions from server CPUs, thus freeing up CPU compute resources to drive application performance. In this way, smart NICs expand NIC functions and provide higher performance.
- Smart NICs can also offload network virtualization protocols, such as Virtual Extensible LAN (VXLAN) and Network Virtualization using Generic Routing Encapsulation (NVGRE). They support the SR-IOV function as well.



Introduction to Linux Bridges

- A Linux Bridge is a virtual network device that works at Layer 2 in a Linux system. It is named after br in the OS.
- Network devices, such as TUN and TAP devices, can be added to a Linux Bridge as interfaces. Devices on a Linux Bridge can receive only Layer 2 data frames and forwards all received data frames to the Linux Bridge.
- Similar to a switch, a Linux Bridge supports various functions, such as MAC address learning, STP, and VLAN.



- Similar to a physical switch, a Linux Bridge looks up for the outbound port for forwarding a data frame in the MAC address table and updates the table. As such, a Linux Bridge can decide whether to forward the data frame to another interface, discard it, broadcast it, or send it to the upper-layer protocol stack.
- When Linux Bridges are used to set up virtual networks, bridge_nf of Linux Bridges works with iptables to implement the security group function in the cloud computing scenario.
- For more information, see <https://wiki.linuxfoundation.org/networking/bridge>.

Introduction to OVS

- An Open vSwitch (OVS) is a virtual switch running on an open-source virtualization platform. It supports OpenFlow and tunneling technologies such as GRE, VXLAN, and IPsec, and provides comprehensive functions in network security, monitoring, management, and QoS.
- OVS can be deployed across multiple physical servers.
- Compared with a Linux Bridge, an OVS gains popularity in virtualization and cloud computing scenarios due to its rich functions.



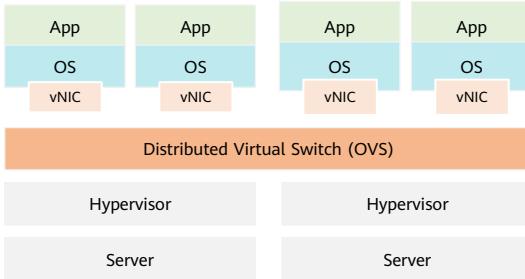
43 Huawei Confidential

 HUAWEI

- An Open vSwitch (OVS) is a software-based open-source virtual switch. It complies with the Apache 2.0 license. It supports multiple standard management interfaces and protocols, such as NetFlow, sFlow, SPAN, Remote Switched Port Analyzer (RSPAN), Command Line Interface (CLI), LACP, and 802.1ag. It can be deployed across multiple physical servers (similar to vSwitch from VMware and Nexus 1000V from Cisco). OVS supports the OpenFlow protocol and can be integrated with multiple open-source virtualization platforms.
- OVS supports but is not limited to the following features:
 - Supports traffic monitoring protocols, such as NetFlow, IPFIX, sFlow, and SPAN/RSPAN.
 - Supports fine-grained ACL and QoS policies.
 - Supports port bonding, LACP, and tunneling (VXLAN, GRE, and IPsec).
 - Supports the standard 802.1Q VLAN protocol.
 - Supports VM interface-based traffic management policies.
- For more information, see <https://docs.openvswitch.org/en/latest/intro/what-is-ovs/>.

Introduction to DVS

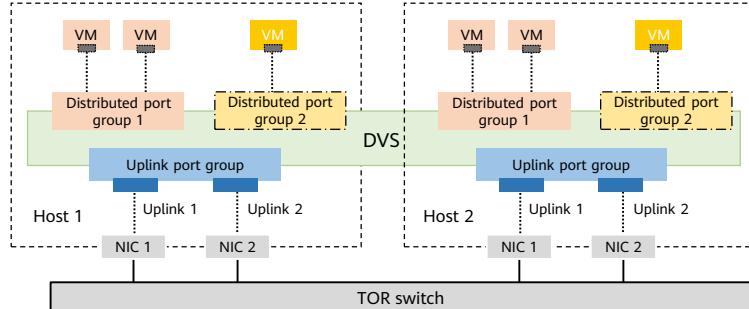
- A distributed virtual switch (DVS) is an abstract representation of multiple hosts defining the same name, network policy, and attribute. OVS is a kind of DVS.
- A DVS lets VMs maintain consistent network configuration and policy as they migrate across multiple hosts.



- A DVS acts as a virtual switch across hosts.
- A DVS provides VMs with consistent network experience regardless of physical host locations.
- A DVS is an OVS on a host (if OVS is used).

DVS Fundamentals

- Key concepts in DVS:
 - Distributed port group: provides VMs with network connections that span across hosts. A DVS can have multiple distributed port groups.
 - Uplink: At the host level, each uplink is connected to a physical NIC. Uplinks are used to configure physical connections of hosts.
 - Uplink port group: can have one or more uplinks. A DVS can have only one uplink port group.



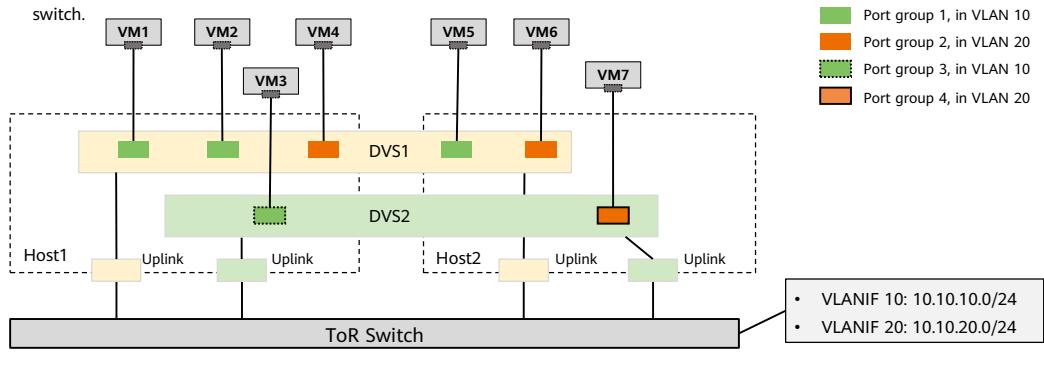
45 Huawei Confidential

 HUAWEI

- A DVS provides similar functions as a physical switch. Each host is connected to the DVS. A DVS connects to VMs through distributed port groups and connects to physical Ethernet adapters on hosts where the VMs reside. As such, a DVS implements communication between virtual and physical networks, since it connects hosts and VMs.
- A DVS functions as a single virtual switch across all associated hosts. It allows VMs to maintain consistent network configuration as they migrate across hosts.

How Do DVSs Allow VMs to Communicate

- DVS 1 and DVS 2 are created on the TOR switch, each of which has two port groups in VLANs 10 and 20, respectively. The gateways for VMs in the two VLANs are located on the TOR switch.
- VMs on the same DVS and in the same VLAN can communicate with each other directly inside the host where they reside. VMs on the same DVS but in different VLANs, as well as VMs on different DVSs, can communicate with each other only through a physical switch.

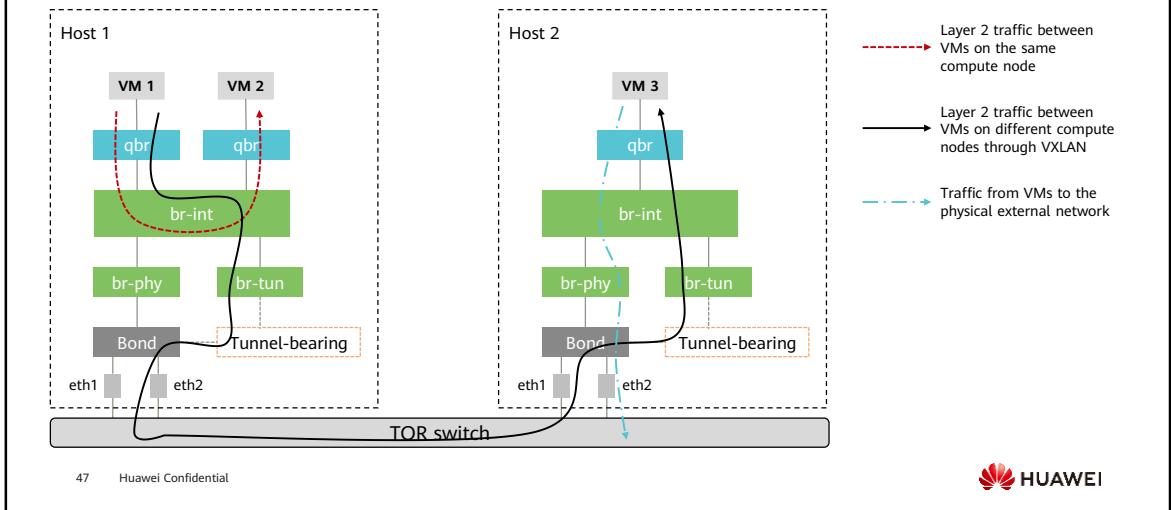


46 Huawei Confidential

HUAWEI

- As shown in the figure, DVS 1 has two port groups: port group 1 connecting VM 1, VM 2, and VM 5 and port group 2 connecting VM 4 and VM 6. DVS 2 also has two port groups: port group 3 connecting VM 3 and port group 4 connecting VM 7. The two DVSs have separate uplinks.
- VMs connected to the same port group can directly communicate with each other. VMs on the same host and DVS can communicate with each other directly through the DVS. For example, VM 1 and VM 2 can communicate through DVS 1. VMs on different hosts but on the same DVS, such as VM 1 and VM 5, can communicate with each other through uplinks of the DVS.
- VMs connected to different port groups (that is, in different VLANs), no matter whether they are located on the same DVS or host, can communicate with other only through the physical switch that allows inter-VLAN communication. For example, VM 1 in VLAN 10 can communicate VM 4, VM 6, and VM 7 in VLAN 20 only through the TOR switch.
- Traffic between VMs on different DVS but connected to port groups in the same VLAN (such as VM 1 and VM 3 connected to port groups 1 and 3, or VM 4 and VM 7 connected to port groups 2 and 4, respectively) is transmitted through DVS uplinks to the physical switch for forwarding.

OVS Application in Virtualization and Cloud Computing Scenarios



47 Huawei Confidential

HUAWEI

- qbr: Linux Bridge, which provides security group services for VMs and implements security isolation.
- br-int: one of the OVS core bridges. Layer 2 and Layer 3 traffic must pass through this bridge. Local VLANs help isolate different virtual networks on a host, and take effect only locally.
- br-phy: a physical bridge, one of the OVS core bridges. Physical NICs of a node are mounted on this bridge. It encapsulates traffic of different service VLANs based on the flow table, and then sends encapsulated packets to physical external networks through physical NICs.
- br-tun: a tunnel bridge, one of the OVS core bridges. This bridge is used to forward VXLAN traffic. Tunnel-bearing is a VTEP that encapsulates and decapsulates VXLAN packets.
- Bond: NIC bonding provided by Linux. It bonds the NIC ports on a host to improve network reliability.

Section Summary

- This section uses the Linux OS as an example to describe traffic forwarding principles and process on the underlying network deployed with server virtualization.
- This section focuses on the functions and principles of virtual network devices in a Linux system, such as TAP/TUN, Linux Bridge, and OVS.

Contents

1. Server Virtualization
2. Network Virtualization
- 3. Introduction to FusionCompute**

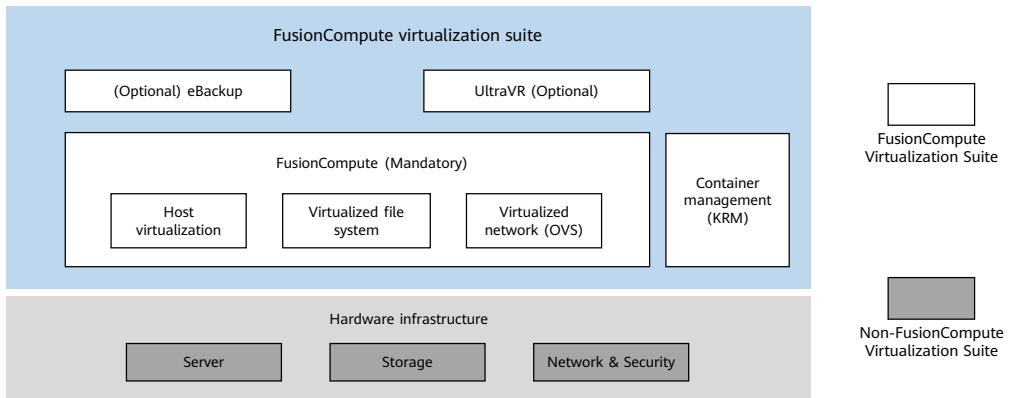
Introduction to FusionCompute Virtualization Suite

- Huawei FusionCompute virtualization suite is an industry-leading virtualization solution.
- The FusionCompute virtualization suite deploys virtualization software on servers so that one physical server can function as multiple servers. Achieve high consolidation ratios by consolidating existing workloads and utilizing remaining servers to deploy new applications and solutions, which greatly improves the efficiency of the data center infrastructure.
- The FusionCompute virtualization suite brings the following benefits to customers:
 - This feature helps customers improve resource utilization of data center infrastructure.
 - Help customers shorten the service rollout period by multiple times.
 - Help customers reduce data center energy consumption by multiple times.
 - With the high availability and strong recovery capability of the virtualized infrastructure, the solution quickly and automatically recovers services from faults, reducing data center costs and increasing system application uptime.

- Application scenario: This scenario applies to the scenario where enterprises use FusionCompute as the unified O&M management platform to operate and maintain the entire system. including resource monitoring, resource management, and system management.
- FusionCompute virtualizes hardware resources and centrally manages virtual resources, service resources, and user resources. It uses virtual computing, virtual storage, and virtual network technologies to virtualize computing, storage, and network resources. In addition, the unified interface is used to centrally schedule and manage these virtual resources, reducing service operating costs and ensuring system security and reliability.
- This section describes the features of the Fusioncompute virtual network. For details about other features and scenarios, see the FusionCompute Product Documentation.

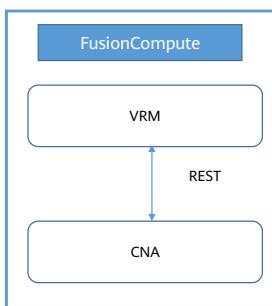
FusionCompute Architecture

- Shows the logical architecture of the FusionCompute virtualization suite.



- FusionCompute is a cloud operating system software that virtualizes hardware resources and centrally manages virtual resources, service resources, and user resources. It uses virtual computing, virtual storage, and virtual network technologies to virtualize computing, storage, and network resources.
- eBackup is a virtual backup software. It works with the snapshot function and CBT function of FusionCompute to implement the VM data backup solution of FusionCompute. (eBackup does not support virtualization deployment in Haiguang scenarios.)
- UltraVR is the DR service management software. It uses the asynchronous remote replication feature provided by the underlying SAN storage system to protect and restore key VM data.
- Container management: manages Kubernetes clusters and nodes, content libraries, projects, and container images.
- Note: FusionCompute is mandatory, and eBackup and UltraVR are optional. This section describes only mandatory components.

FusionCompute Logical Architecture

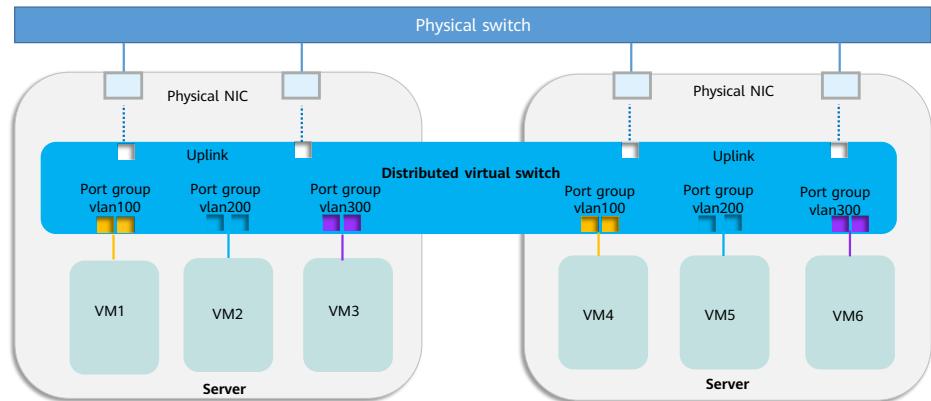


Module	Function
VRM	<ul style="list-style-type: none">Manages block storage resources in a cluster.Manages network resources (IP addresses and VLANs) in a cluster and allocates IP addresses to VMs.Manages the life cycle of VMs in a cluster and the distribution and migration of VMs on compute nodes.Manages dynamic resource adjustment in a cluster.Manages virtual resources and user data in a unified manner and provides services such as elastic computing, storage, and IP addresses.The provides a unified O&M management interface for O&M personnel to remotely access the through the WebUI. FusionCompute performs O&M on the entire system, including resource management, resource monitoring, and resource reporting.
CNA	<ul style="list-style-type: none">Provides the virtual computing function.Manages VMs on compute nodes.Manages computing, storage, and network resources on compute nodes.

- Virtual Resource Management (VRM): functions as a unified O M management platform and manages multiple CNA hosts.
- Computing Node Agent (CAN): deployed on a computing node, manages VMs on the computing node and mounts the VMs to the corresponding virtual volumes.

FusionCompute Virtual Network Management

- As shown in the figure, the virtual NIC of a VM is connected to the DVS through a port group, and then connected to the physical NIC of the host through the uplink of the DVS. In this way, the VM can communicate with the external network environment.



54 Huawei Confidential

 HUAWEI

- This section describes how to create network resources, such as distributed switches (DVSs) and port groups, and how to adjust and configure network resources on FusionCompute.
- A DVS is a virtual switch. It functions like a Layer 2 physical switch. It connects to VMs through port groups and connects to physical networks through uplinks.
- A port group is a virtual logical port. Similar to a network attribute template, a port group defines the mode in which VM NICs are connected to the network through a DVS.
 - VLAN mode:** No IP address is allocated to the VM NICs that use the port group. You need to manually allocate IP addresses to the VM NICs. However, the VMs are connected to the VLAN defined by the port group.
 - MUX VLAN mode:** The Layer 2 traffic isolation mechanism provided by the MUX VLAN enables some users to communicate with each other and isolate other users.
- Uplinks are used by the DVS to connect to the physical NICs of hosts and are used for VM data uplinks.

VM Provisioning (1)

Creation Mode	Description
Creating an empty VM	An empty virtual machine is like a blank physical computer without an operating system installed. When creating an empty VM, you can create it on a host or cluster and customize the CPU, memory, disk, and NIC specifications. After an empty VM is created, you need to install the OS on the VM. The procedure for installing the operating system is the same as that for installing the operating system on a physical machine.
Creating a VM Using a Template	Use a template to create VMs similar to the template. <ul style="list-style-type: none">Use an existing template to create VMs by converting the template to a VM and deploying VMs based on the template.Export the template used by other sites and import the template to create VMs at the site. When a template is converted to a VM, all attributes of the VM are the same as those of the template. After the conversion, the template does not exist. When a VM is deployed using a template or a VM is imported using a template, the following attributes are inherited from the template and other attributes can be customized. <ul style="list-style-type: none">VM OS type and versionNumber, capacity, and bus type of VM disksNumber of VM NICs If you have a virtual machine that you want to clone frequently, you can set the virtual machine as a template.
Creating a VM Using a VM	Clone a VM similar to an existing VM in the system. During VM cloning, the following attributes are inherited from the original VM. Other attributes can be customized. <ul style="list-style-type: none">VM OS type and versionNumber, capacity, and bus type of VM disksNumber of VM NICs If you have a virtual machine that you want to clone frequently, you can set the virtual machine as a template.

- A virtual machine, like a physical computer, is a virtual computer that runs an operating system and applications.
- A VM runs on a CNA and obtains required computing resources such as CPUs and memory, USB devices, network connections, and storage access from the CNA. Multiple VMs can run on one CNA at the same time. FusionCompute provides multiple methods for creating VMs.

VM Provisioning (2)

Creation Mode	Recommended Application Scenario
Creating an empty VM	<ul style="list-style-type: none">• Create a VM for the first time during the initial deployment of the system.• If no suitable template or VM is available in the system (the OS and hardware configuration are the same), you need to create an empty VM.• Create an empty virtual machine, install an operating system on it, and convert or clone the virtual machine to a template so that you can use the template to create a virtual machine.
Creating a VM Using a Template	<ul style="list-style-type: none">• A proper template is available in the system (the operating system and hardware configuration are the same). Using the template to create a VM can save time.• Export the template of another site and import the template to create VMs at the site.
Creating a VM Using a VM	When deploying multiple similar virtual machines, you can create, configure, and install different software on a single virtual machine, and then clone the virtual machine multiple times instead of creating and configuring each virtual machine separately.

Section Summary

- This section describes the features of FusionCompute, including the logical architecture of FusionCompute and the functions of each module, such as the functions of CAN and VRM.

Quiz

1. (Essay) What functions does a virtualization management platform provide?
2. (Multiple-answer question) Which of the following virtual network devices work at Layer 2 when server virtualization is deployed? ()
 - A. TUN
 - B. TAP
 - C. Linux Bridge
 - D. OVS

1. A virtualization management platform manages virtualized clusters in a unified manner, provides a simple management interface for users, monitors and manages virtual resources, simplifies the VM creation process, and configures and executes resource scheduling policies.
2. BCD

Summary

- With increasingly wide application of virtualization technologies in DCs, servers are being integrated into physical networks. Network engineers need to have basic knowledge of server virtualization.
- This course introduces the principles of server virtualization and server network virtualization from the perspective of IT engineers, and describes end-to-end traffic forwarding between servers and physical networks deployed with network virtualization.
- For more information about server and network virtualization principles, visit the websites on the More Information slide.

More Information

- <https://www.kernel.org/doc/html/latest/networking/tuntap.html>
- <https://wiki.linuxfoundation.org/networking/bridge>
- <https://docs.openvswitch.org/en/latest/intro/what-is-ovs/>

Thank you.

把数字世界带入每个人、每个家庭。

每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



Technical Principles and Applications of VXLAN



Foreword

- Cloud computing has become a new form of enterprise IT construction due to its advantages such as high system utilization, low labor and management costs, and high flexibility and scalability. In cloud computing, widely deployed virtualization is a basic technology mode. The wide deployment of server virtualization technology greatly increases the computing density of data centers (DCs). In addition, to implement flexible service changes, VMs need to be migrated without restrictions on the network.
- Virtual eXtensible Local Area Network (VXLAN) is an important overlay technology. It can solve the problems faced by traditional DCs, such as small VM migration scope, limited number of VMs, and limited network isolation capability. VXLAN is widely used in SDN network scenarios of DCs, such as cloud-network integration DC scenarios.
- This course describes the background, fundamentals, and application scenarios of VXLAN and EVPN.

Objectives

- On completion of this course, you will be able to:
 - Describe the network requirements of DCs and how VXLAN meets these requirements.
 - Describe basic concepts of VXLAN.
 - Describe fundamentals of VXLAN.
 - Understand concepts and fundamentals of EVPN.
 - Understand the combination of EVPN and VXLAN technologies.

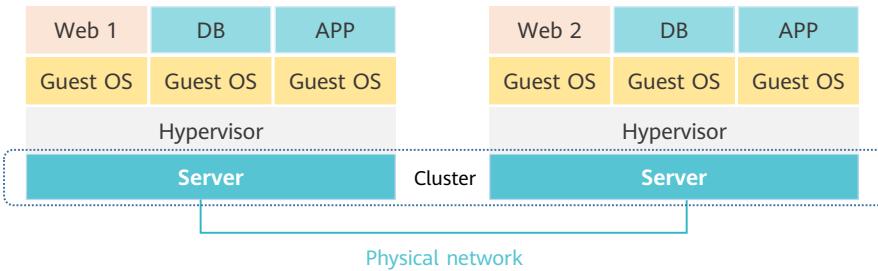
Contents

- 1. Background of VXLAN**
2. Basic Concepts and Fundamentals of VXLAN
3. EVPN VXLAN Fundamentals
4. VXLAN Deployment Cases in Typical Scenarios

Technical Background: Virtualization Is Widely Deployed by Enterprises

- Virtualization technologies reduce IT and O&M costs, and improve service deployment flexibility. More and more enterprises choose to use cloud computing or virtualization technologies in their DC IT facilities.
- After an enterprise chooses the virtualization architecture, services are deployed on VMs in server clusters.

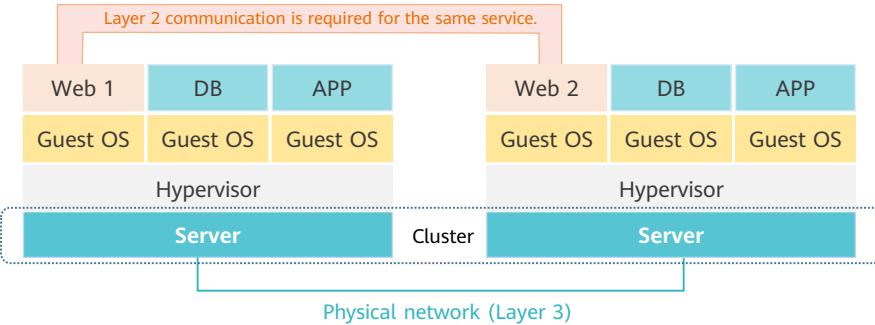
Services are deployed on VMs in server clusters.



New Network Requirement - Layer 2 Extension

- VMs in a virtualization or cloud computing cluster can be migrated flexibly. As a result, VMs running the same service (on the same network segment) may run on different servers, or the same VM (with the same IP address) may run on different servers (physical locations) at different times.
- Physical servers may be distributed in equipment rooms that are geographically distant from each other. Therefore, Layer 3 connectivity is required.

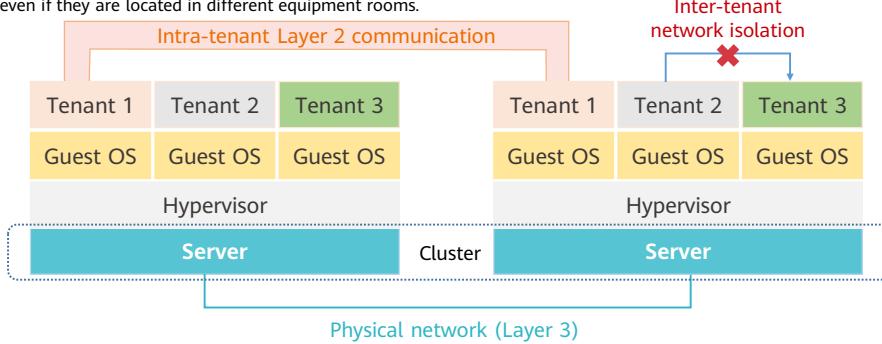
Layer 2 communication across a Layer 3 network is required.



- After servers are virtualized, services are encapsulated on VMs. VMs can be live migrated to any host in a cluster. One of the features of live migration is that the network status does not change. This requires that the IP addresses of VMs in different physical locations remain unchanged. Therefore, a large Layer 2 network is required to solve this problem.

New Network Requirement - Multi-Tenant Isolation

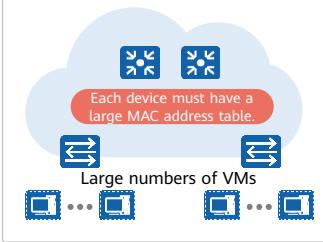
- In cloud-based scenarios, multi-tenancy is supported, that is, different tenants share physical resources. This poses two requirements on the network: inter-tenant isolation and intra-tenant communication.
 - Inter-tenant isolation: Tenants may be configured with the same MAC address and IP address. Physical network isolation needs to be considered, and a large number of users need to be isolated.
 - Intra-tenant communication: VMs on the same network segment of a tenant can directly communicate with each other at Layer 2, even if they are located in different equipment rooms.



Challenges Facing Traditional Networks

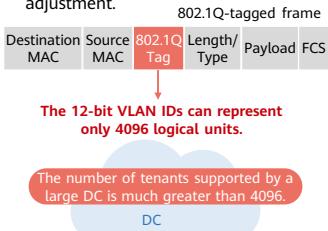
VM quantity limited by entry specifications of devices

- After servers are virtualized, the number of VMs increases greatly compared with the number of original physical machines. However, the MAC address table size of Layer 2 access devices is small, which cannot meet the requirements of the rapidly increasing number of VMs.



Limited network isolation capabilities

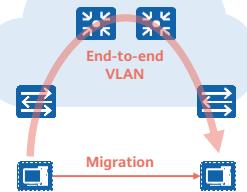
- The VLAN ID field has only 12 bits.
- In large virtualization and cloud computing service scenarios, the number of tenants is much greater than the number of available VLANs.
- VLANs on traditional Layer 2 networks cannot adapt to dynamic network adjustment.



Limited VM migration scope

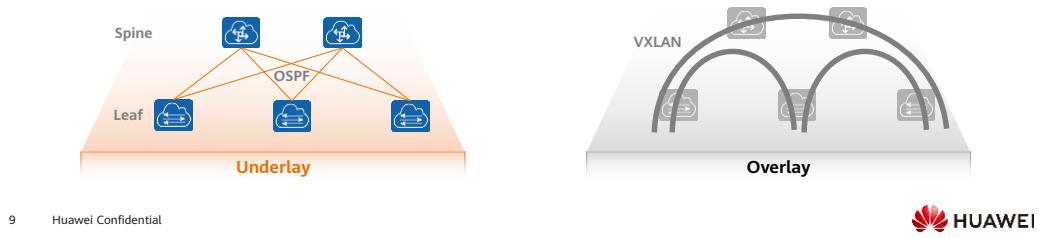
- VM migration must be performed on a Layer 2 network.
- VM migration on a traditional Layer 2 network is limited to a small scope.

VMs can be migrated only within a VLAN. The number of VLANs is limited.



Overview of VXLAN

- VXLAN is essentially a virtual private network (VPN) technology and can be used to build a Layer 2 virtual network (overlay network) on any physical network (underlay network) with reachable routes. VXLAN tunnels can be built between VXLAN gateways to implement communication within a VXLAN network as well as communication between a VXLAN network and a non-VXLAN network.
- VXLAN utilizes MAC-in-UDP encapsulation to extend Layer 2 networks. It encapsulates Ethernet packets into IP packets for these packets to be transmitted through routing, without considering the MAC addresses of VMs. In addition, Layer 3 networks are not limited by the network architecture and support large-scale scalability. VM migration through routed networks is also not limited by the physical network architecture.

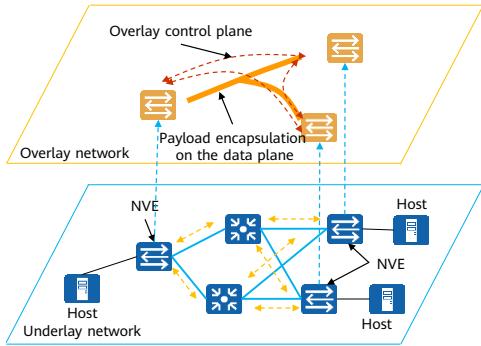


9 Huawei Confidential

 HUAWEI

- VXLAN resolves the following problems on traditional networks:
 - VM quantity limited by network specifications:
 - VXLAN encapsulates data packets sent from VMs into UDP packets, and encapsulates IP and MAC addresses used on the physical network into the outer headers. Devices on the network are aware of only the encapsulated parameters but not the inner data.
 - Only VXLAN network edge devices need to identify the MAC addresses of VMs, thereby reducing the number of MAC addresses that must be learned and enhancing device performance.
 - Limited network isolation capabilities:
 - VXLAN uses a VNI field similar to the VLAN ID field to identify users. The VNI field has 24 bits and can identify up to 16M VXLAN segments, effectively isolating and identifying a large number of tenants.
 - VM migration scope limited by the network architecture:
 - VMs with IP addresses on the same network segment are logically located in the same Layer 2 domain even if they are physically located on different Layer 2 networks. VXLAN builds a virtual large Layer 2 network over a Layer 3 network.
- Underlay network: It is a physical network that functions as the base layer of the upper-layer logical network.
- Overlay network: It is a logical network built on a physical network using a tunneling technology.

Underlay and Overlay



Overlay

- An overlay network is a logical network established on an underlay network through VXLAN.
- It has independent forwarding and control plane protocols.
- The underlay physical network is transparent to devices that are not connected to VXLAN tunnel endpoints (VTEPs).

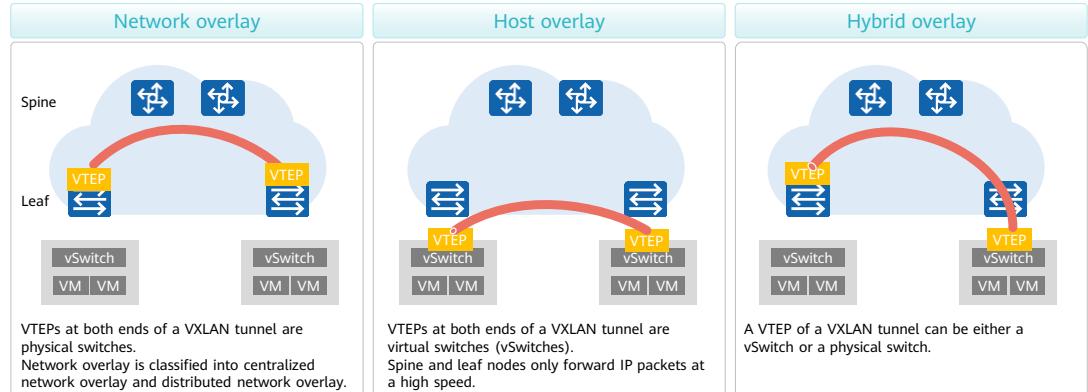
Underlay

- An underlay network consists of various physical network devices and is a bearer network of an overlay network.
- After an overlay technology is implemented on an underlay network, a logical network is formed based on the underlay network.
- The underlay network provides basic capabilities such as reachability and reliability for the upper-layer overlay network.
- The underlay network has independent control and forwarding plane protocols. Generally, OSPF or EBGP is used as the control plane protocol, and IPv4 is used as the forwarding plane protocol.
- The underlay network is logically isolated from the overlay network and is unaware of overlay network routes.



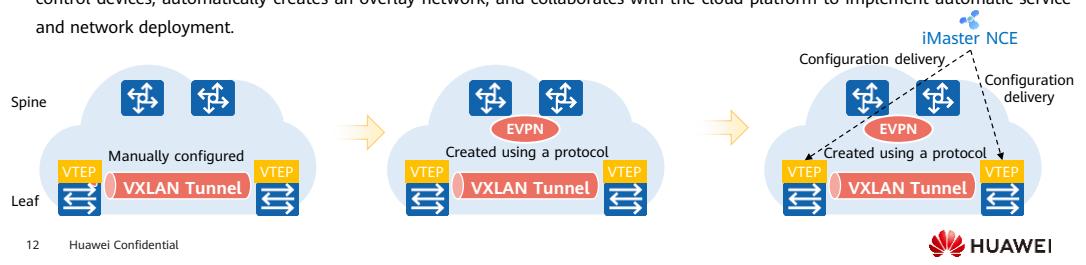
VXLAN Overlay Network Types

- VXLAN overlay networks are classified into network overlay, host overlay, and hybrid overlay networks based on the types of devices where VTEPs reside.



Overlay Protocol Development

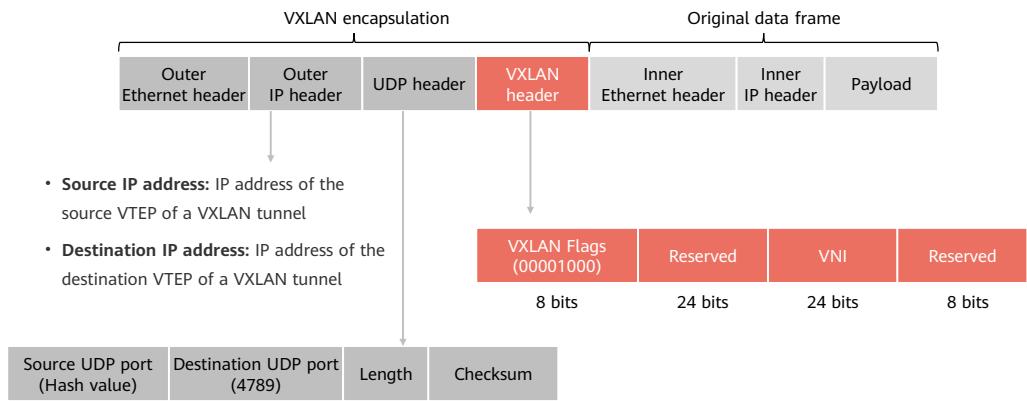
- To meet the requirements of multi-tenant and VM migration in cloud DCs, vendors are looking for an overlay protocol with optimal performance and the most flexible applications. VXLAN proposed in RFC 7348 meets the requirements.
- In the early stage, VXLAN is deployed in static mode, and VXLAN tunnels are manually created, which requires heavy configuration workload. In addition, VXLAN does not have a control plane. VTEP discovery and host information collection are implemented through traffic flooding on the data plane. As a result, a large amount of flooding traffic exists on the data center network (DCN). To address these problems, VXLAN works with Ethernet Virtual Private Network (EVPN) to implement automatic VXLAN tunnel establishment, automatic VTEP discovery, and host information advertisement.
- To facilitate control and deployment on a large Layer 2 network, an SDN controller is introduced. The controller uses NETCONF to control devices, automatically creates an overlay network, and collaborates with the cloud platform to implement automatic service and network deployment.



Contents

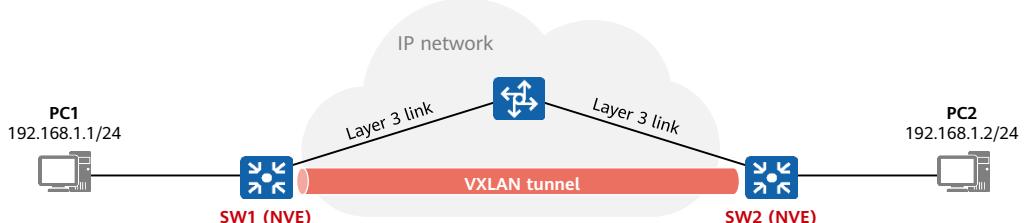
1. Background of VXLAN
2. **Basic Concepts and Fundamentals of VXLAN**
 - Basic Concepts
 - Fundamentals
3. EVPN VXLAN Fundamentals
4. VXLAN Deployment Cases in Typical Scenarios

VXLAN Packet Format



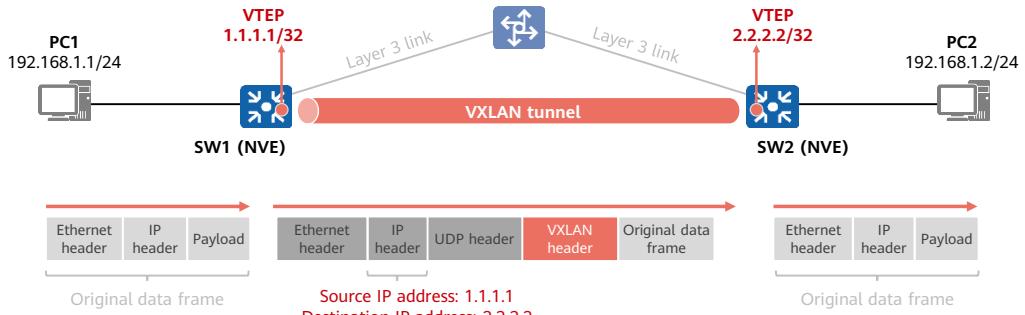
Basic Concepts of VXLAN: NVE

- Network Virtualization Edge (NVE):
 - A network entity that implements network virtualization functions. A hardware or software switch can work as an NVE.
 - NVEs run VXLAN and construct a Layer 2 virtual network over a Layer 3 network. SW1 and SW2 in the figure are NVEs.



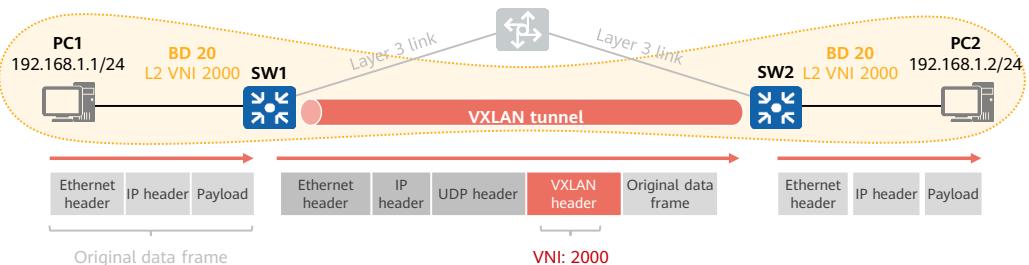
Basic Concepts of VXLAN: VTEP

- VXLAN tunnel endpoint (VTEP):
 - A VTEP is located on an NVE and performs VXLAN encapsulation and decapsulation.
 - In the outer IP header of VXLAN packets, the source IP address is the IP address of the source VTEP, and the destination IP address is the IP address of the destination VTEP.



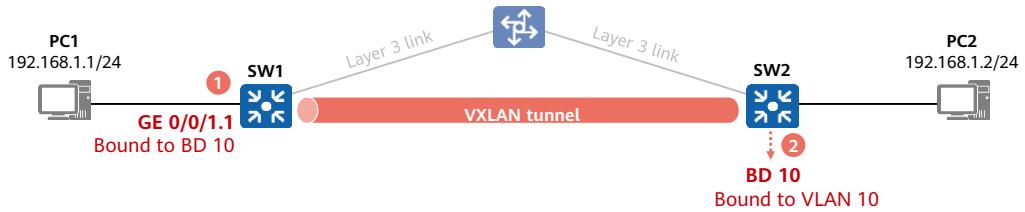
Basic Concepts of VXLAN: VNI and BD

- VXLAN Network Identifier (VNI):
 - An L2 VNI is similar to a VLAN ID and identifies a Layer 2 broadcast domain. VMs in different broadcast domains cannot communicate with each other at Layer 2.
 - An L3 VNI is used to identify a VPN instance. A Layer 3 VNI is associated with a VPN instance for inter-subnet forwarding of VXLAN packets.
 - A tenant can have one or more VNIs. The VNI field has 24 bits, and a maximum of 16M tenants are supported.
- Bridge domain (BD):
 - VLANs are used to divide broadcast domains on a traditional network. Similarly, BDs are used to divide broadcast domains on a VXLAN network. A BD identifies a large Layer 2 broadcast domain on a VXLAN network.
 - VNIs are mapped to BDs in 1:1 mode. Terminals in the same BD can communicate with each other at Layer 2.



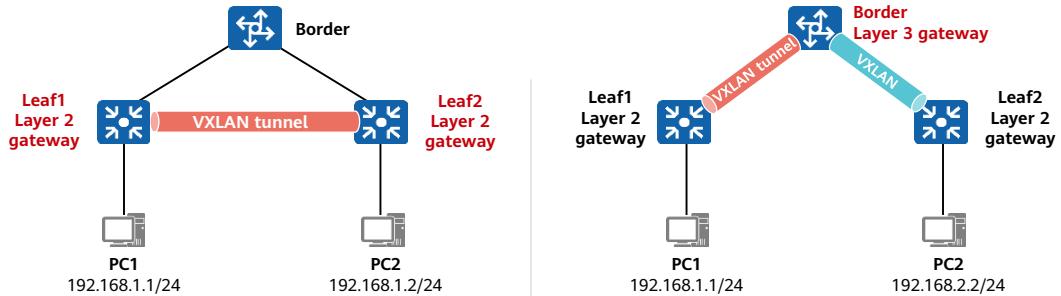
Basic Concepts of VXLAN: VXLAN Access Modes

- Service access points need to be configured on devices for service access to a VXLAN network. The following two access modes are available:
 - Access in Layer 2 sub-interface mode: For example, a Layer 2 sub-interface is created on SW1 and associated with BD 10. Specific traffic on the sub-interface is then forwarded to BD 10.
 - Access in VLAN binding mode: For example, VLAN 10 is configured on SW2 and associated with BD 10. All traffic from VLAN 10 is then forwarded to BD 10.



- After traffic from a traditional network enters a VXLAN network, the traffic is bound to a BD through Layer 2 sub-interface or VLAN binding mode. A VXLAN VNI is specified in the BD to implement mapping from the traditional VLAN network to the VXLAN network.
- When VLAN binding mode is used for VXLAN access, a BD cannot be configured with a VBDIF interface. Therefore, this mode applies only to Layer 2 service access.

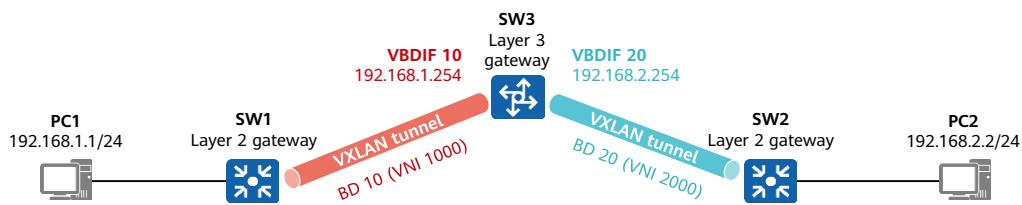
Basic Concepts of VXLAN: Layer 2 and Layer 3 VXLAN Gateways



Layer 2 gateway: forwards traffic to a VXLAN network and is used for intra-subnet communication between terminals on the same VXLAN network.

Layer 3 gateway: is used for inter-subnet communication between terminals on a VXLAN network and allows terminals to access external networks (non-VXLAN networks).

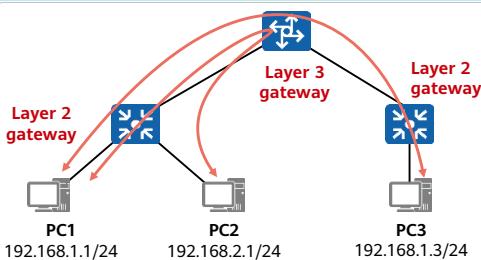
Basic Concepts of VXLAN: VBDIF Interface



- VLANIF interfaces are used for communication between broadcast domains on a traditional network. Similarly, VBDIF interfaces are used for communication between BDs on a VXLAN network.
- A VBDIF interface is a Layer 3 logical interface created for a BD on a Layer 3 VXLAN gateway.
- VBDIF interfaces allow users on different network segments to communicate through a VXLAN network, allow communication between VXLAN and non-VXLAN networks, and implement Layer 2 network access to a Layer 3 network.

Basic Concepts of VXLAN: Distributed and Centralized Gateways

Centralized gateway

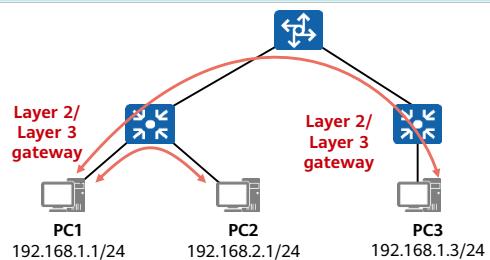


The Layer 3 gateway is deployed on one device. All inter-subnet traffic is forwarded by the gateway to implement centralized traffic management.

Advantage: Inter-subnet traffic is managed in a centralized manner, simplifying gateway deployment and management.

Disadvantage: The forwarding path is not optimal. The number of ARP entries supported is a bottleneck. Because a centralized Layer 3 gateway is deployed, the gateway needs to maintain a large number of ARP entries for terminals connected to the VXLAN network.

Distributed gateway



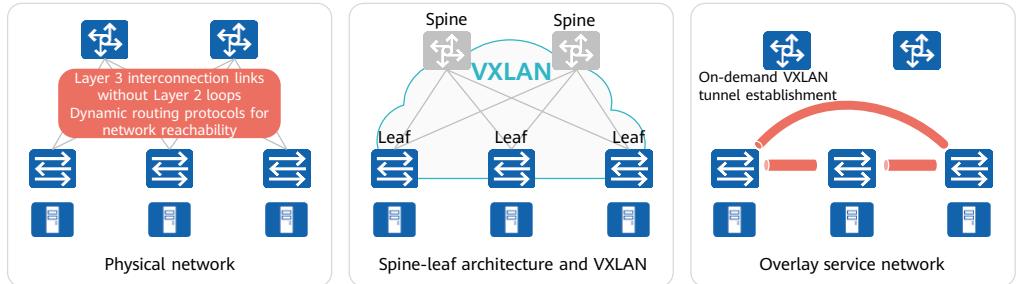
VTEPs function as both Layer 2 and Layer 3 gateways. Non-gateway nodes are unaware of VXLAN tunnels and only forward VXLAN packets.

Advantage: A VTEP only needs to learn ARP entries of terminals connected to it. Therefore, the number of ARP entries supported is no longer a bottleneck on distributed VXLAN gateways, and the network scalability is improved.

Disadvantage: Compared with centralized gateway deployment, this mode is complex to configure and implement.

Application of VXLAN in DCs

- VXLAN can be applied to a DCN that uses a two-layer spine-leaf physical architecture.
- It is recommended that a VXLAN network with distributed gateways be deployed in a DC. Spine nodes forward packets based on routes and are unaware of VXLAN during traffic forwarding. Leaf nodes provide network access for device resources such as servers, and perform VXLAN encapsulation and decapsulation.
- All services in the DC are carried by the VXLAN network.



Contents

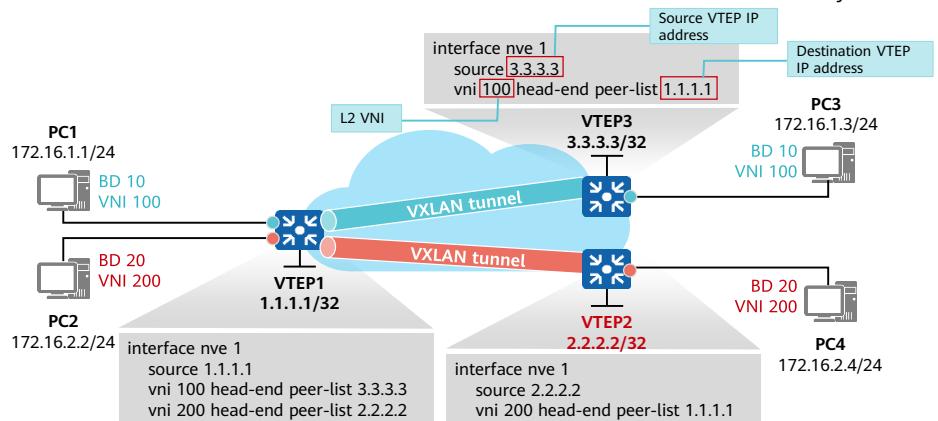
1. Background of VXLAN
2. **Basic Concepts and Fundamentals of VXLAN**
 - Basic Concepts
 - **Fundamentals**
3. EVPN VXLAN Fundamentals
4. VXLAN Deployment Cases in Typical Scenarios

VXLAN Tunnel Establishment

- A VXLAN tunnel is identified by a pair of VTEPs. Packets are encapsulated on VTEPs and then transmitted in the VXLAN tunnel through routing. A VXLAN tunnel can be successfully established as long as the VTEPs at both ends of the VXLAN tunnel have reachable routes to each other at Layer 3.
- VXLAN tunnels are classified into the following types based on the VXLAN tunnel creation mode:
 - Static VXLAN tunnels: created by manually configuring the local and remote VNIs, VTEP IP addresses, and ingress replication lists.
 - Dynamic VXLAN tunnels: dynamically established using BGP EVPN. After a BGP EVPN peer relationship is established between VTEPs, the VTEPs use BGP EVPN routes to transmit VNIs and VTEP IP addresses to dynamically establish a VXLAN tunnel.

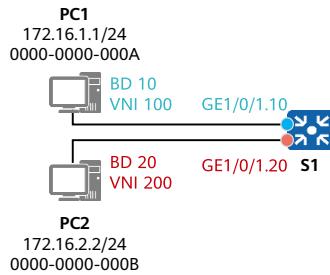
Static VXLAN Tunnel

- A static VXLAN tunnel is created through manual configuration. The VXLAN tunnel can be established successfully as long as the VTEPs at both ends of the tunnel have reachable routes to each other's IP address at Layer 3.



VXLAN MAC Address Entries

- VXLAN implements Layer 2 forwarding on the overlay network. Unicast data frames are still forwarded based on MAC address entries.
- When a VTEP receives a data frame from the local BD, the VTEP adds the source MAC address of the data frame to the MAC address table of the BD. The outbound interface in the MAC address entry is the interface that receives the data frame.
- This entry is used to guide the forwarding of data frames sent to terminals connected to the VTEP.

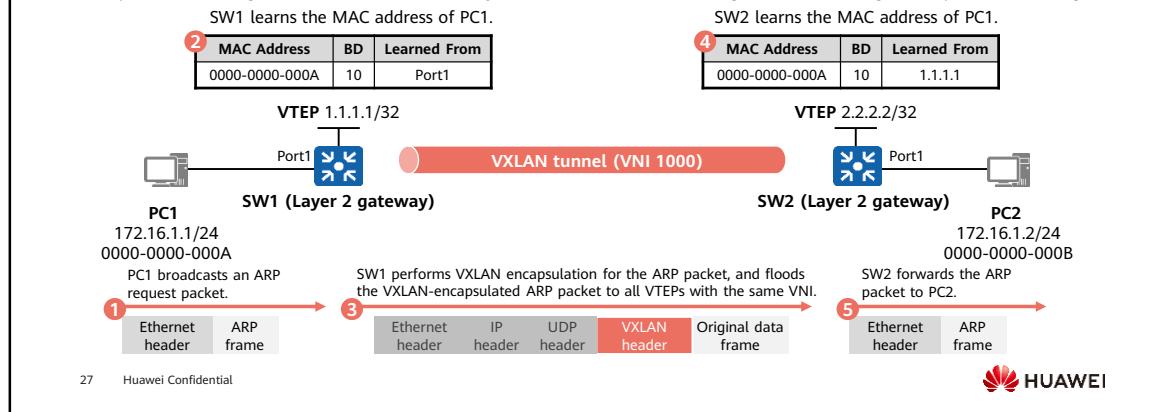


How can data frames be forwarded to the device connected to the remote VTEP?

<S1>display mac-address bridge-domain 10			
MAC Address	VLAN/VSI/BD	Learned-From	Type
0000-0000-000a	-/-/10	GE1/0/1.10	dynamic
<S1>display mac-address bridge-domain 20			
MAC Address	VLAN/VSI/BD	Learned-From	Type
0000-0000-000b	-/-/20	GE1/0/1.20	dynamic

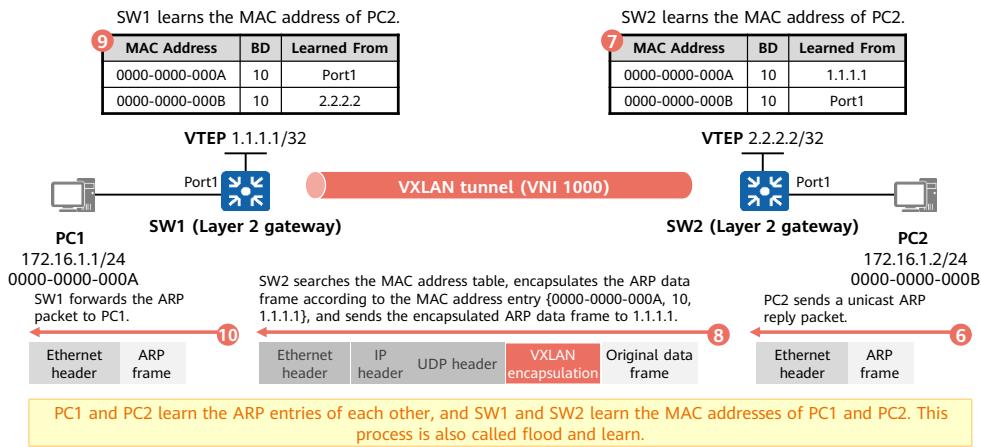
Dynamic MAC Address Learning (1)

- To forward data frames to a device connected to a remote VTEP, the local VTEP needs to learn the MAC address of the remote device first.
- Similar to the traditional MAC address entry generation process, the MAC address entry generation process depends on packet exchange between hosts. Generally, MAC address entries are generated through ARP packet exchange.



- The communication process between PC1 and PC2 is as follows:
 - To communicate with PC2, PC1 broadcasts an ARP request frame to obtain the MAC address of PC2.
 - After receiving the frame, SW1 determines the BD ID, destination VXLAN tunnel, and VNI of the traffic based on the service access point configuration. In addition, SW1 learns the MAC address of PC1 and records the BD ID and the interface that receives the frame in the corresponding MAC address entry.
 - SW1 performs VXLAN encapsulation for the ARP request packet and forwards the encapsulated packet based on the ingress replication list.
 - After receiving the VXLAN packet, SW2 decapsulates the packet to obtain the original data frame. In addition, SW2 learns the MAC address of PC1 and binds the MAC address to the VTEP address of SW1.
 - SW2 floods the ARP packet in the local BD. PC2 then receives the frame and learns the ARP information of PC1.

Dynamic MAC Address Learning (2)



- PC2 sends a unicast ARP reply packet.
 - SW2 has learned the MAC address of PC1 and forwards the packet in unicast mode. SW2 learns the source MAC address of PC2 and adds it to the MAC address table.
 - SW2 performs VXLAN encapsulation for the ARP reply packet and sends the encapsulated packet to the remote VTEP with the IP address 1.1.1.1.
 - After receiving the VXLAN packet, SW1 decapsulates the packet and records the source MAC address of PC2 in the MAC address table. The outbound interface of the corresponding MAC address entry is the remote VTEP.
 - SW1 forwards the data frame to PC1.
- PC1 and PC2 learn ARP entries of each other, and SW1 and SW2 learn corresponding MAC addresses.

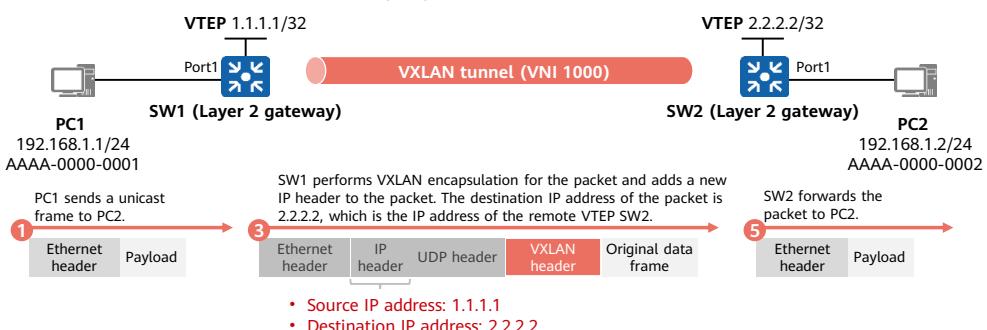
Intra-Subnet Forwarding of Unicast Packets with Known Destination Addresses

MAC Address	BD	Learned From
AAAA-0000-0001	10	Port1
AAAA-0000-0002	10	2.2.2.2

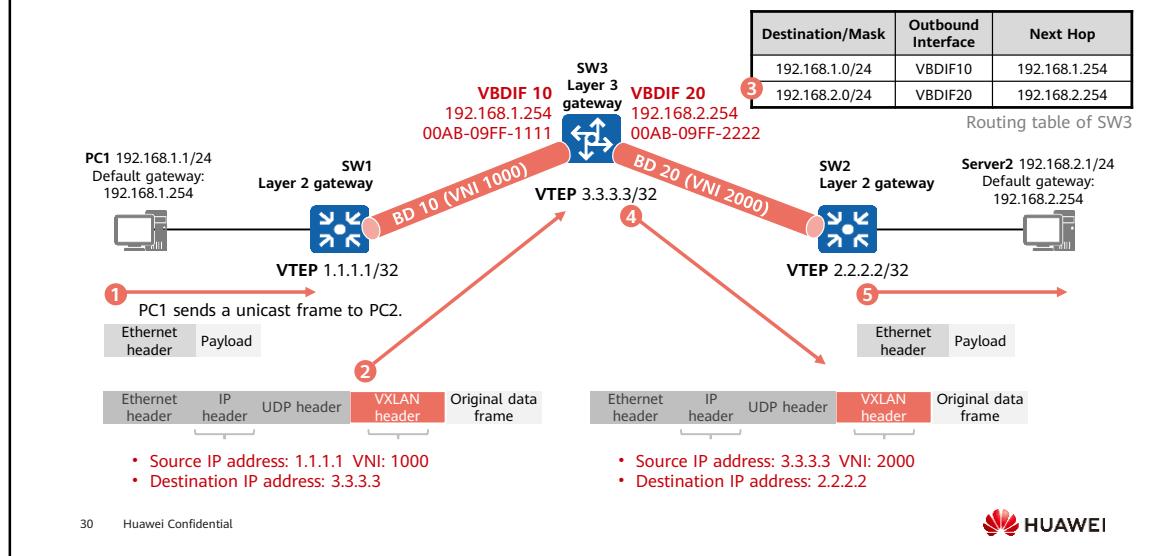
② SW1 searches its MAC address table for the MAC address of PC2 and finds the matching entry.

MAC Address	BD	Learned From
AAAA-0000-0001	10	1.1.1.1
AAAA-0000-0002	10	Port1

④ SW2 searches its MAC address table for the MAC address of PC2 and finds the matching entry.



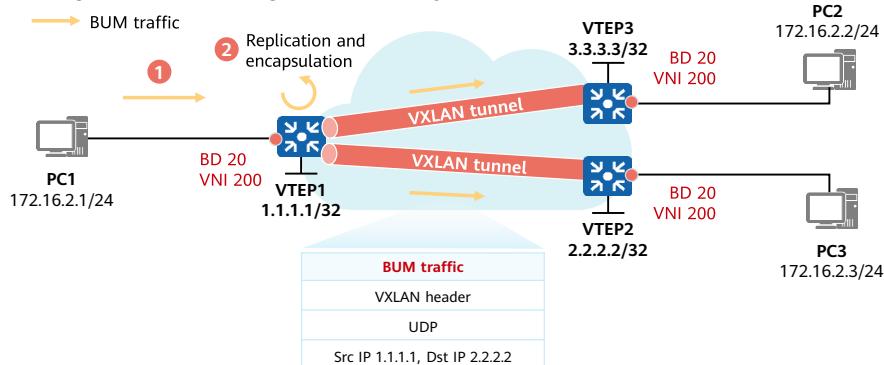
Inter-Subnet Forwarding of Unicast Packets



- PC1 wants to communicate with PC2. After local calculation, PC1 finds that it is on a different subnet from PC2. PC1 then sends the packet to the gateway.
- The destination MAC address of the data frame from PC1 to PC2 is 00AB-09FF-1111 (gateway MAC address). After receiving the data frame, SW1 searches the Layer 2 forwarding table and finds that the outbound interface is the remote VTEP (Layer 3 gateway). SW1 then adds a VXLAN header (VNI = 1000) to the data frame and sends the packet to SW3.
- After receiving the packet, SW3 performs VXLAN decapsulation for the packet and finds that the destination MAC address of the original data frame is 00AB-09FF-1111, which is the MAC address of VBDIF 10 on SW3. SW3 needs to search the Layer 3 forwarding table to forward the data frame.
- SW3 searches the routing table and finds that the destination IP address 192.168.2.1 matches the direct route generated by VBDIF 20 on SW3. SW3 then searches the ARP table for the destination MAC address of the packet and searches the MAC address table for the outbound interface of the packet. On SW3, the outbound interface in the MAC address entry corresponding to 192.168.2.1 is the remote VTEP with the IP address 2.2.2.2. SW3 performs VXLAN encapsulation for the packet and sends the encapsulated packet to SW2.
- After receiving the packet, SW2 performs VXLAN decapsulation for the packet and finds that the destination MAC address is not the MAC address of any interface on SW2. SW2 searches the Layer 2 forwarding table and forwards the packet from a local interface based on the MAC address table.

BUM Traffic Forwarding

- When transmitting broadcast, unknown unicast, and multicast traffic (BUM traffic), the local VTEP sends multiple copies of the traffic to remote VTEPs in the ingress replication list, implementing flood forwarding on the overlay network.



Contents

1. Background of VXLAN
2. Basic Concepts and Fundamentals of VXLAN
- 3. EVPN VXLAN Fundamentals**
 - Basic Concepts
 - BGP EVPN Routes
 - BGP EVPN Feature
4. VXLAN Deployment Cases in Typical Scenarios

Using BGP EVPN as the Control Plane Protocol

BGP EVPN not used	BGP EVPN used as the control plane protocol
<p>Problem 1: A total of $N \times (N-1)/2$ tunnels need to be created for N nodes, causing heavy configuration workload.</p>	<ul style="list-style-type: none"> Enable BGP EVPN on devices and establish BGP EVPN peer relationships between them. Devices advertise BGP EVPN routes to each other to complete related operations on the VXLAN control plane. VXLAN tunnels are automatically established through BGP EVPN, and forwarding entries are dynamically updated through BGP EVPN. <p>In actual deployment, a route reflector (RR) can be used to further reduce the number of established BGP EVPN peer relationships.</p>
<p>Problem 2: The flood and learn mechanism is used to learn MAC addresses, causing a large amount of flooding traffic.</p>	

33 Huawei Confidential

HUAWEI

- The static VXLAN solution does not have a control plane. VTEP discovery and learning of host information (including IP addresses, MAC addresses, VNIs, and gateway VTEP IP addresses) are performed through traffic flooding on the data plane. As a result, there is a lot of flooding traffic on VXLAN networks. To address this problem, BGP EVPN is introduced as the control plane of VXLAN. BGP EVPN allows VTEPs to exchange BGP EVPN routes to implement automatic VTEP discovery and host information advertisement, preventing unnecessary traffic flooding.
- Problems in configuring VXLAN in static mode:
 - If N devices need to establish VXLAN tunnels, you need to manually configure the ingress replication list a maximum of $N \times (N-1)/2$ times.
 - A static VXLAN tunnel only has the data forwarding plane.
 - Remote MAC addresses can be learned only through broadcast ARP packets.

Overview of BGP EVPN

- BGP EVPN extends BGP by defining several new types of BGP EVPN routes using Network Layer Reachability Information (NLRI) in the MP_REACH_NLRI attribute.
- These BGP EVPN routes can be used to transmit VTEP addresses and host information. Therefore, BGP EVPN is applied to VXLAN networks to transfer VTEP discovery and host information learning from the data plane to the control plane.



- **Type 2 routes (MAC/IP routes):** are used to advertise host MAC addresses, ARP entries, and IP routes.
- **Type 3 routes (inclusive multicast routes):** are used to transmit Layer 2 VNI and VTEP IP address information, implement automatic VTEP discovery, dynamic VXLAN tunnel establishment, and BUM packet forwarding.
- **Type 5 routes (IP prefix routes):** are used to advertise host IP routes and external network routes.

- In a network virtualization overlay (NVO) scenario, BGP EVPN is used together with VXLAN as the control plane protocol for VXLAN.

EVPN NLRI

- EVPN NLRI is carried in the path attribute MP_REACH_NLRI. The address family identifier (AFI) is 25, indicating L2VPN. The sub-address family identifier (SAFI) is 70.

Path Attribute - MP_REACH_NLRI	
Flags:	Optional, Non-transitive
Type Code:	MP_REACH_NLRI (14)
Length	
Address family identifier (AFI):	Layer-2 VPN (25)
Subsequent address family identifier (SAFI):	EVPN (70)
Next hop network address (4 bytes)	
Route Type (1 octet)	
Length (1 octet)	
Route Type specific (variable)	

—
EVPN NLRI
—

Extended Community

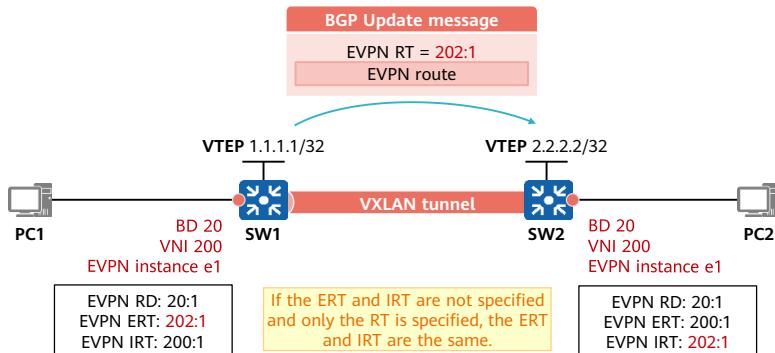
- Similar to MPLS VPN, BGP EVPN uses EVPN instances to control route sending and receiving. Similar to traditional IP VPN instances, EVPN instances also have RDs and RTs, and the extended community attribute is used to carry EVPN instance RTs during route transmission.
- In addition to the RT, BGP EVPN adds some new subtypes to the extended community attribute: MAC Mobility and EVPN Router's MAC Extended Community.

Path Attribute - EXTENDED_COMMUNITIES	
Flags: Optional, Transitive	
Type Code: EXTENDED_COMMUNITIES (16)	
Length	
Route Target (RT)	T
MAC Mobility	Extended Community
EVPN Router's MAC Extended Community	L

- For details about RDs and RTs, see HCIP – Datacom - Advanced Routing & Switching Technology - 08 MPLS VPN Basics.

EVPN VPN Instance

- After an EVPN instance is bound to a BD, MAC address entries in the BD are transmitted through BGP EVPN routes carrying the export VPN target (ERT) of the EVPN instance bound to the BD. After receiving the EVPN routes, the remote end compares the import VPN target (IRT) of the local EVPN instance with the ERT, adds the EVPN routes to the routing table of the corresponding EVPN instance, parses the EVPN routing table to obtain MAC address entries, and adds the MAC address entries to the MAC address table of the BD bound to the local EVPN instance.



Contents

1. Background of VXLAN
2. Basic Concepts and Fundamentals of VXLAN
- 3. EVPN VXLAN Fundamentals**
 - Basic Concepts
 - BGP EVPN Routes
 - BGP EVPN Feature
4. VXLAN Deployment Cases in Typical Scenarios

MAC/IP Route (1)

- Type 2 routes (MAC/IP routes): are used to advertise MAC addresses, ARP entries, and host IP routes.

Packet format	Field description
Route Distinguisher (8 bytes)	Route distinguisher (RD) configured for an EVPN instance.
Ethernet Segment Identifier (10 bytes)	Unique ID of the connection between local and remote devices.
Ethernet Tag ID (4 bytes)	VLAN ID configured on the local device.
MAC Address Length (1 byte)	Length of the host MAC address carried in the route.
MAC Address (6 bytes)	Host MAC address carried in the route.
IP Address Length (1 byte)	Mask length of the host IP address carried in the route.
IP Address (0, 4, or 16 bytes)	Host IP address carried in the route.
MPLS Label1 (3 bytes)	Layer 2 VNI carried in the route.
MPLS Label2 (0 or 3 bytes)	Layer 3 VNI carried in the route.

MAC/IP Route (2)

- Contents carried in BGP EVPN Type 2 routes vary in different scenarios.

Host MAC address advertisement

Route Distinguisher
Ethernet Segment Identifier
Ethernet Tag ID
MAC Address Length = MAC address length
MAC Address = MAC address
IP Address Length
IP Address
MPLS Label1 = VNI (Layer 2)
MPLS Label2

When hosts on the same subnet communicate with each other, host MAC addresses containing host MAC address information and Layer 2 VNIs are advertised.

Host ARP advertisement

Route Distinguisher
Ethernet Segment Identifier
Ethernet Tag ID
MAC Address Length = MAC address length
MAC Address = MAC address
IP Address Length = IP address length
IP Address = IP address
MPLS Label1 = VNI (Layer 2)
MPLS Label2

In a centralized VXLAN gateway scenario, ARP routes containing host IP address information, MAC address information, and Layer 2 VNIs are advertised.

Host IP route advertisement

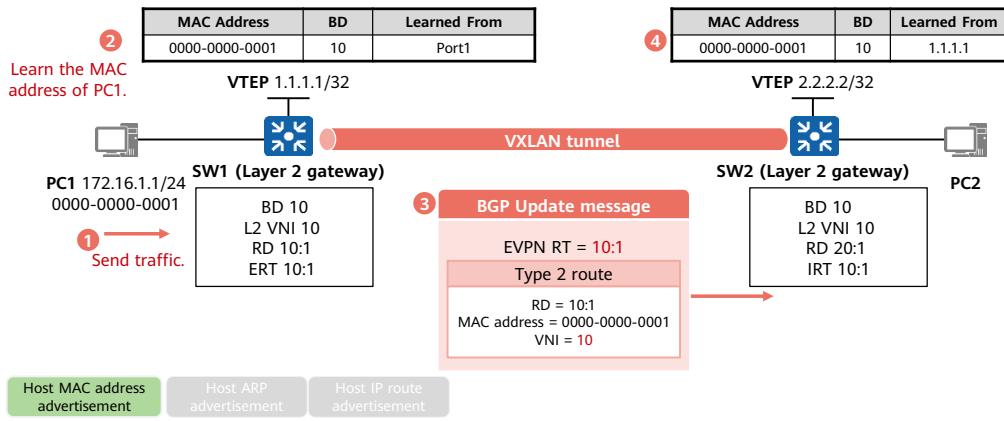
Route Distinguisher
Ethernet Segment Identifier
Ethernet Tag ID
MAC Address Length = MAC address length
MAC Address = MAC address
IP Address Length = IP address length
IP Address = IP address
MPLS Label1 = VNI (Layer 2)
MPLS Label2 = VNI (Layer 3)

When hosts on different subnets communicate with each other in a distributed gateway scenario, IRB routes containing host MAC address information, IP address information, Layer 2 VNIs, and Layer 3 VNIs are advertised.

- The contents of the first three fields (RD, Ethernet Segment Identifier, and Ethernet Tag ID) of BGP EVPN Type 2 routes are the same in different scenarios, and the contents of the last six fields vary in different scenarios.

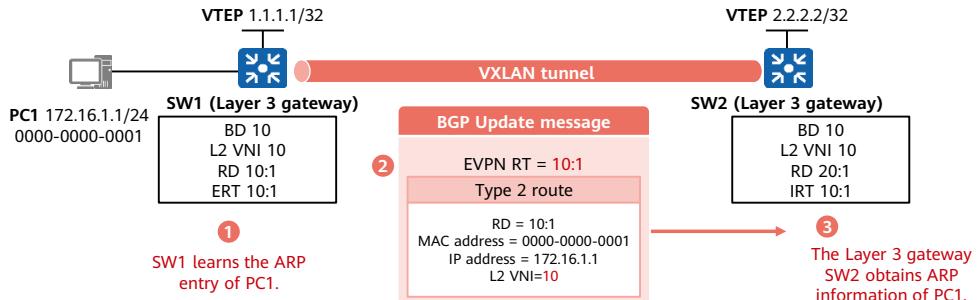
Host MAC Address Advertisement

- This slide shows how BGP EVPN uses Type 2 routes to implement dynamic MAC address learning. This function is used to implement intra-subnet communication through VXLAN.



Host ARP Advertisement

- This slide shows how Type 2 routes implement host ARP advertisement when BGP EVPN is used to construct a DCN in a distributed gateway scenario.



When BGP EVPN is used in a centralized gateway scenario, the inter-subnet packet forwarding process is similar to that in a static VXLAN scenario, and is not described here.

Host MAC address advertisement

Host ARP advertisement

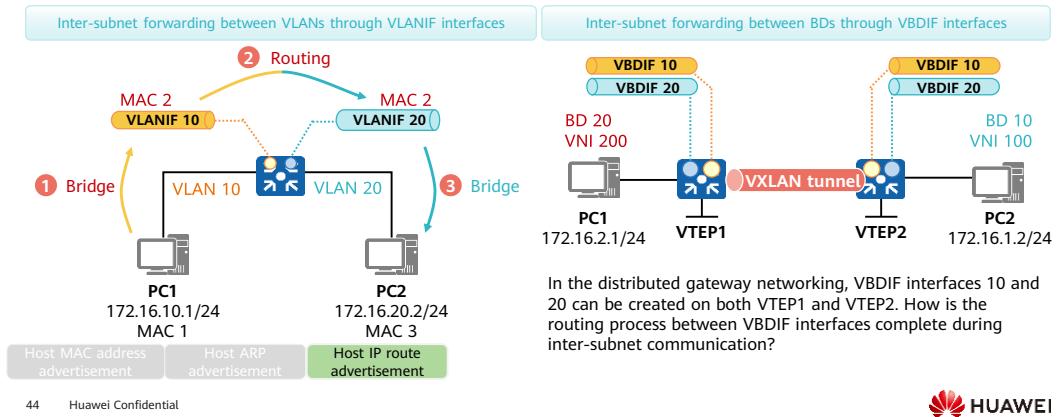
Host IP route advertisement

- A MAC/IP route can carry both the MAC address and IP address of a host. As such, this type of route can be used to transmit host ARP entries between VTEPs, thereby implementing host ARP advertisement. The MAC Address and MAC Address Length fields identify the MAC address of the host, whereas the IP Address and IP Address Length fields identify the IP address of the host. In this case, MAC/IP routes are also called ARP routes. Host ARP advertisement applies to the following scenarios:

- ARP broadcast suppression. After a Layer 3 gateway learns the ARP entry of a host on its subnet, it generates host information that contains the host IP and MAC addresses, L2VNI, and gateway's VTEP IP address. The Layer 3 gateway then advertises an ARP route carrying the host information to a Layer 2 gateway. When the Layer 2 gateway receives an ARP request, it searches for host information corresponding to the destination IP address in the request. If the host information exists, the gateway replaces the broadcast MAC address in the ARP request with the destination unicast MAC address, and unicasts the packet, thereby implementing ARP broadcast suppression.

Inter-Subnet Communication in a Distributed Gateway Scenario

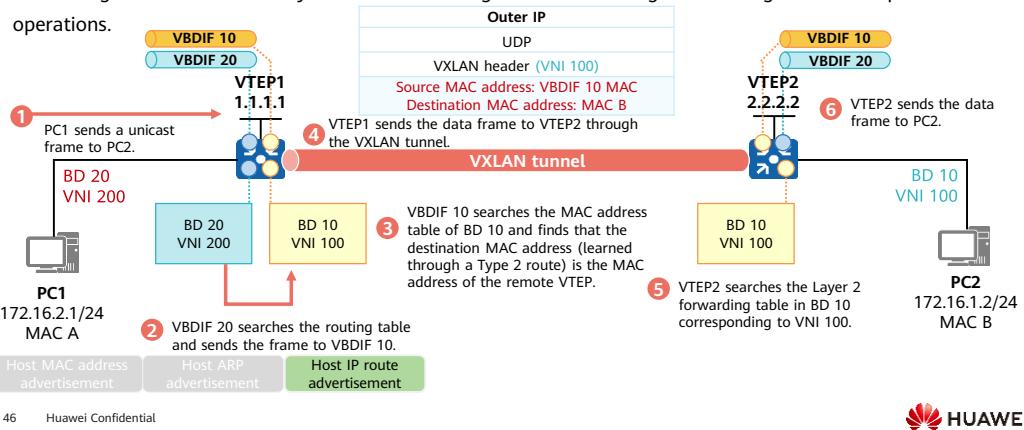
- In the distributed gateway networking, VTEPs function as both Layer 2 and Layer 3 gateways. In this networking, inter-subnet communication can be implemented in different modes. According to the processing mode of the ingress VTEP that receives packets, inter-subnet communication can be classified into asymmetric integrated routing and bridging (IRB) and symmetric IRB.



- Details about inter-subnet forwarding between VLANs through VLANIF interfaces:
 - Based on the local IP address, local mask, and peer IP address, PC1 finds that PC2 is not on the same network segment as itself. Therefore, PC1 determines that the communication is Layer 3 communication and sends the traffic destined for PC2 to the gateway. In the data frame sent by PC1, the source MAC address is MAC1 and the destination MAC is MAC2.
 - After receiving a packet destined for PC2 from PC1, the switch decapsulates the packet and finds that the destination MAC address is the MAC address of VLANIF 10. Therefore, the switch considers that the packet is destined for itself and sends the packet to the routing module for further processing.
 - The routing module parses the packet and finds that the destination IP address is 192.168.20.2, which is not an IP address of a local interface. Therefore, the packet needs to be forwarded at Layer 3. After the routing table is searched, a direct route generated by VLANIF 20 is matched.
 - Because the matched route is a direct route, the packet has reached the last hop. Therefore, the switch searches the ARP table for 192.168.20.2 to obtain the MAC address of the host with the IP address 192.168.20.2, and sends the MAC address to the switching module for re-encapsulation into a data frame.

Asymmetric IRB

- Asymmetric IRB: The ingress VTEP searches both the Layer 3 and Layer 2 forwarding tables for traffic forwarding at the same time, and the egress VTEP searches only the Layer 2 forwarding table for traffic forwarding. This forwarding mode is called asymmetric forwarding because the ingress and egress VTEPs perform different operations.



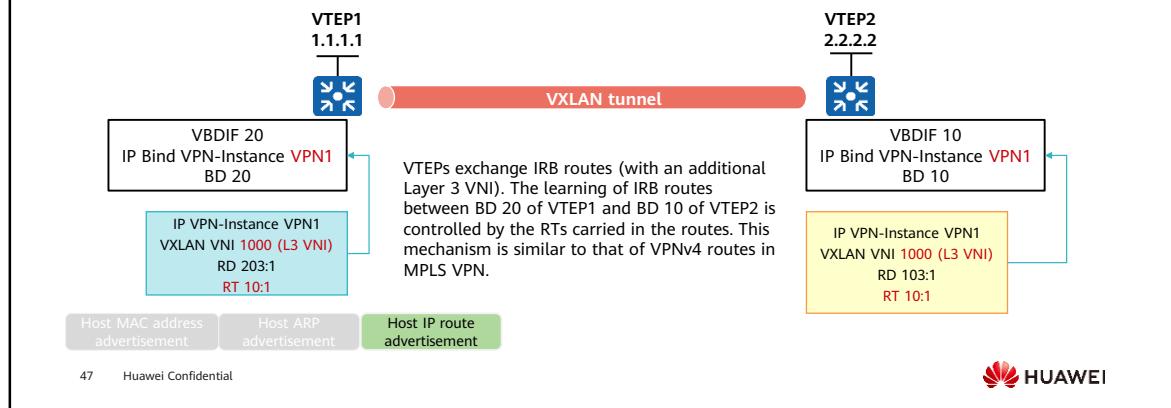
46 Huawei Confidential

HUAWEI

- During asymmetric IRB, host IP routes are not advertised between VTEPs. That is, VTEP1 and VTEP2 do not advertise 32-bit host routes (generated based on ARP information) generated by the local downstream PCs between them. Therefore, VTEP1 searches the routing table in step 2, and only the direct route generated by VBDIF 10 can be matched.
- In step 5, VTEP2 decapsulates the VXLAN packet and finds that the destination MAC address is not the MAC address of the local VBDIF interface corresponding to the BD. Therefore, VTEP2 searches the Layer 2 forwarding table for the MAC address entry of the BD based on the VNI carried in the packet, and then forwards the packet at Layer 2.

Symmetric IRB

- Symmetric IRB: Both the ingress and egress VTEPs search the Layer 3 forwarding table for traffic forwarding.
- Compared with asymmetric IRB, the concepts of an IP VPN instance and its bound Layer 3 VNI are added. (In asymmetric IRB, the VNI in the VXLAN header of packets transmitted between VTEPs is a Layer 2 VNI.) A VBDIF interface needs to be bound to an IP VPN instance. In this case, route learning and data forwarding of the VBDIF interface are restricted in the IP VPN instance, which is similar to the implementation in MPLS VPN.



- Huawei devices implement symmetric IRB.

EVPN RT and IP VPN RT (1)

- After an IP VPN instance is added, the RT carried in a Type 2 route advertised by BGP EVPN is still an EVPN RT. The only difference is that the remote end processes the received route differently.
 - If the RT carried in the route is the same as the import RT of the local EVPN instance, the route is accepted. After the EVPN instance obtains an IRB route, it can extract an ARP route from the IRB route to implement host ARP advertisement.
 - If the RT carried in the route is the same as the import RT (EVPN) of the local IP VPN instance, the route is accepted. The VPN instance then obtains the IRB route carried in the route, extracts the host IP address and Layer 3 VNI from the route, saves the host IP route in the routing table, and recurses the outbound interface based on the next hop of the route. The final recursion result is the VXLAN tunnel pointing to the VTEP.

Host MAC address advertisement

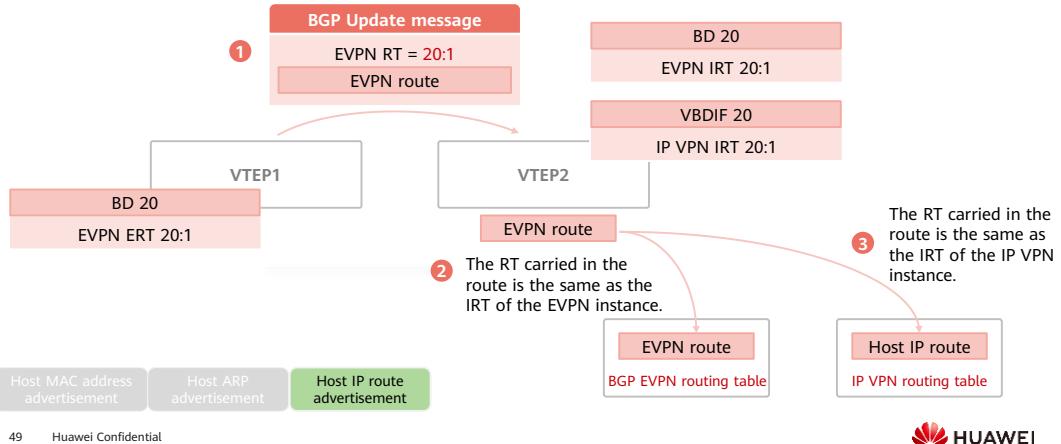
Host ARP advertisement

Host IP route advertisement

- In a BGP EVPN scenario, to use the RTs of an IP VPN instance to control the sending and receiving of EVPN routes, run the `vpn-target evpn` command to configure RTs for the IP VPN instance. Then, the export RT attribute is carried in the EVPN route to be sent to the remote BGP EVPN peer, the import RT attribute is used to determine which EVPN routes can be added to the routing table of the local IP VPN instance address family by matching the import RT attribute with the RT attribute carried in the EVPN route.
- Note: The RTs configured using the `vpn-target evpn` command are called RTs (EVPN).

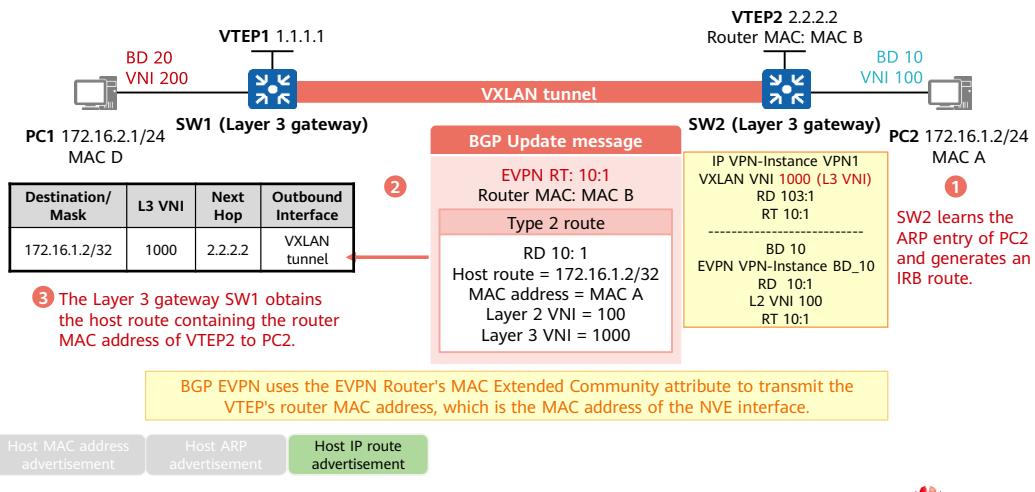
EVPN RT and IP VPN RT (2)

- A route is discarded only when the RT carried in the route is different from the EVPN IRT and IP VPN IRT (EVPN).

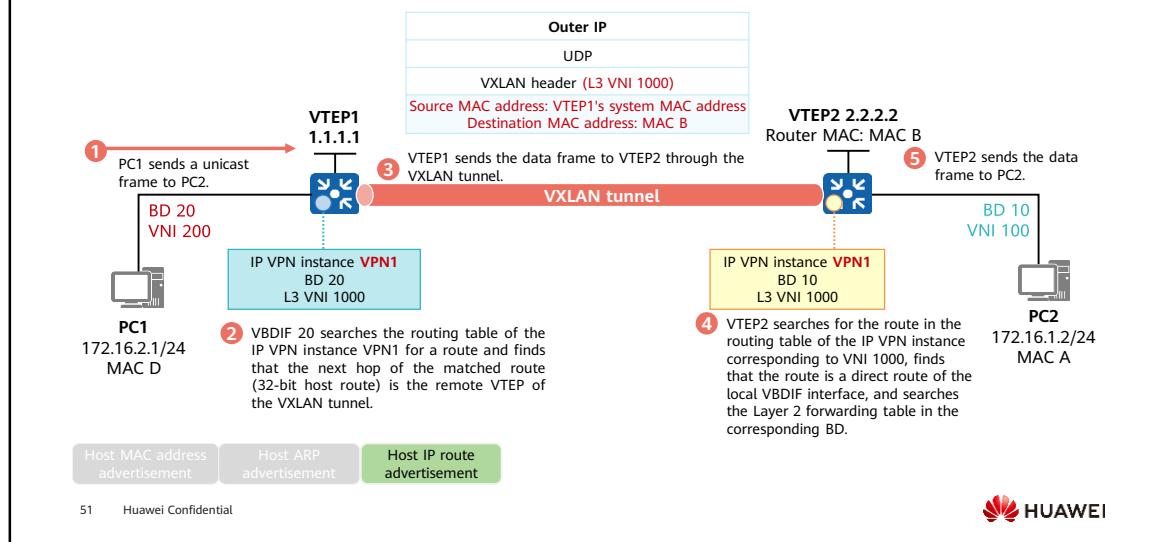


- VTEP1 sends a Type 2 BGP EVPN route (IRB type). The route carries the ERT (20:1) of the EVPN instance bound to the BD to which the route belongs.
- After receiving the BGP Update message, VTEP2 checks whether the RT (20:1) carried in the extended attribute of the BGP Update message is the same as the IRT of the local EVPN instance and the IRT (EVPN) of the IP VPN instance. If the IRT is the same as that of the EVPN instance bound to BD 20 and that of the IP VPN instance bound to VBDIF 20, the device adds the EVPN route to the EVPN routing table of BD 20 and the IP route contained in the EVPN route to the routing table of the IP VPN instance corresponding to VBDIF 20.

Symmetric IRB: Host IP Route Advertisement (IRB Route)



Symmetric IRB: Communication Process



- During symmetric IRB, VTEPs exchange 32-bit host routes generated based on ARP information. Therefore, VTEP1 searches the routing table for the 32-bit host route transmitted by VTEP2. Even if VBDIF 10 and the corresponding direct route exist on VTEP1, VTEP1 still forwards packets based on the 32-bit host route according to the longest match rule.
- In step 4, VTEP2 decapsulates the VXLAN packet and finds that the destination MAC address of the inner data frame is VTEP2's router MAC address (MAC B). VTEP2 then determines to search the Layer 3 table for traffic forwarding. VTEP2 finds the corresponding IP VPN instance based on VNI 1000 and searches for the corresponding route in the routing table of the IP VPN instance. It finds the direct route matching VBDIF 10, searches the local MAC address table, and sends the packet to a local host PC2.

Description of Type 3 Routes

- Type 3 route (inclusive multicast route)

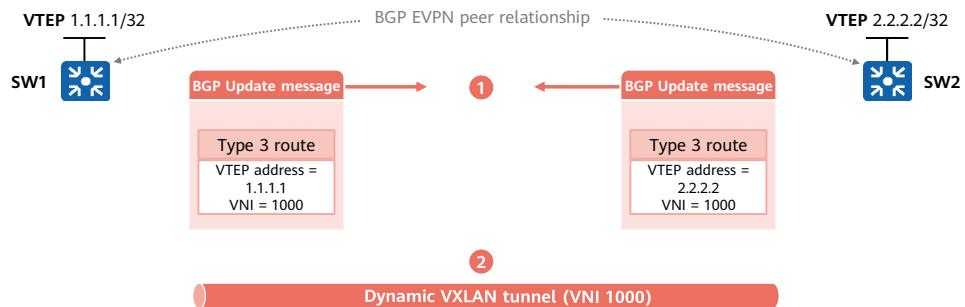
- Inclusive multicast routes are used for automatic VTEP discovery and dynamic VXLAN tunnel establishment on the VXLAN control plane.
- Through these routes, VTEPs that function as BGP EVPN peers transmit Layer 2 VNIs and VTEPs' IP addresses.
- The Originating Router's IP Address and MPLS Label fields carried in the routes indicate the local VTEP's IP address and Layer 2 VNI, respectively.

NLRI format	Route Distinguisher (8 bytes)	Route distinguisher (RD) configured for an EVPN instance.
	Ethernet Tag ID (4 bytes)	VLAN ID configured on the local device, which is all 0s in this type of route.
	IP Address Length (1 byte)	Mask length of the local VTEP's IP address carried in the route.
	Originating Router's IP Address (4 or 16 bytes)	Local VTEP's IP address carried in the route.
PMSI attribute	Flags (1 byte)	This field is not used in VXLAN scenarios.
	Tunnel Type (1 byte)	In VXLAN scenarios, the value can be 6: Ingress Replication.
	MPLS Label (3 bytes) = Layer 2 VNI	Layer 2 VNI carried in the route.
	Tunnel Identifier (variable length)	This field is the local VTEP's IP address in VXLAN scenarios.

- Provider Multicast Service Interface (PMSI): an optional transitive BGP attribute. In VXLAN scenarios, the Tunnel Type field has a fixed value of 6, carrying the VTEP's IP address and Layer 2 VNI of the sender.

VXLAN Tunnel Establishment

- VTEPs exchange Layer 2 VNI and VTEP IP address information through Type 3 routes. If there are reachable routes between the local and remote VTEPs' IP addresses at Layer 3, a VXLAN tunnel is established between the VTEPs. Additionally, if the local and remote VNIs are the same, an ingress replication list is created for BUM packet forwarding.



Description of Type 5 Routes

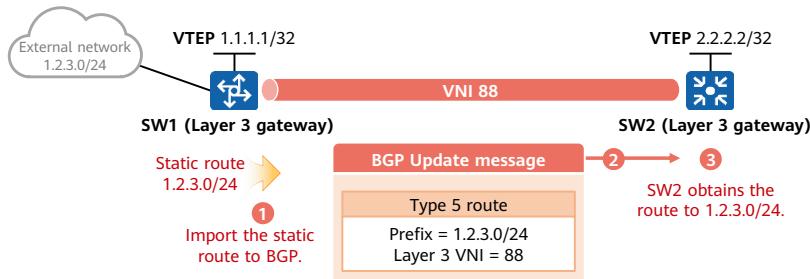
- Type 5 route (IP prefix route)

- The IP Prefix Length and IP Prefix fields in this type of route carry a host IP address or network segment address.
- If a host IP address is carried, the route is used for IP route advertisement in distributed VXLAN gateway scenarios. In this case, the route functions the same as an IRB route on the VXLAN control plane.
- If a network segment address is carried, the route can be advertised to allow hosts on the VXLAN network to access an external network.

Packet format	Field description
Route Distinguisher (8 bytes)	Route distinguisher (RD) configured for an EVPN instance.
Ethernet Segment Identifier (10 bytes)	Unique ID of the connection between local and remote devices.
Ethernet Tag ID (4 bytes)	VLAN ID configured on the local device.
IP Prefix Length (1 byte)	Mask length of the IP prefix carried in the route.
IP Prefix (4 or 16 bytes)	IP prefix carried in the route.
GW IP Address (4 or 16 bytes)	Default gateway address. It is used in specific scenarios.
MPLS Label (3 bytes)	Layer 3 VNI carried in the route.

Application Scenario of Advertising IP Prefix Routes

- For an external network of a VXLAN network, a VTEP can advertise external routes to the entire VXLAN network through Type 5 routes to allow hosts on the VXLAN network to access the external network.



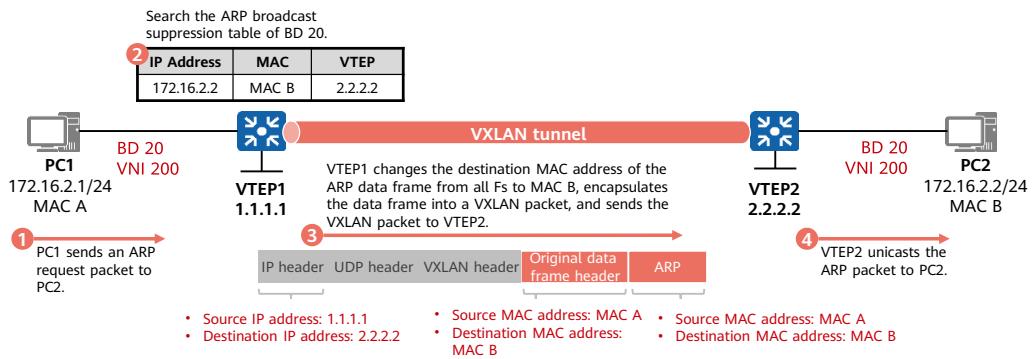
- Similar to Type 2 IRB routes, Type 5 routes carry the router MAC address of the VTEP through the EVPN Router's MAC Extended Community attribute during route transmission. In addition, Type 5 routes carry only the Layer 3 VNI. Therefore, the forwarding process is also IRB.

Contents

1. Background of VXLAN
2. Basic Concepts and Fundamentals of VXLAN
- 3. EVPN VXLAN Fundamentals**
 - Basic Concepts
 - BGP EVPN Routes
 - **BGP EVPN Feature**
4. VXLAN Deployment Cases in Typical Scenarios

ARP Broadcast Suppression

- BGP EVPN Type 2 routes enable VTEPs to learn MAC addresses without depending on communication between hosts. However, ARP requests between hosts still need to be flooded on the VXLAN overlay network, which consumes a large number of network resources.
- ARP broadcast suppression can be implemented based on BGP EVPN routes to reduce broadcast traffic.



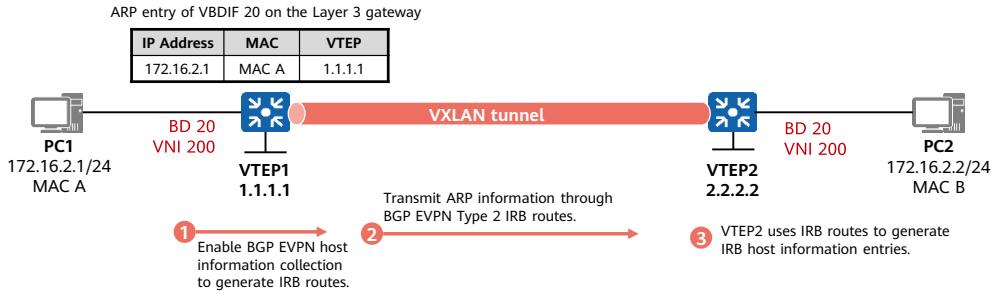
57 Huawei Confidential

HUAWEI

- ARP broadcast suppression effectively reduces the burden of the gateway in processing ARP packets. When the gateway receives an ARP request packet, it searches the ARP broadcast suppression table, which stores the mapping between the IP address and MAC address of the destination device. If a matching entry is found, the gateway replaces the broadcast MAC address in the ARP request packet with the MAC address of the destination device. The gateway then sends the ARP request packet through the interface corresponding to the destination MAC address.

Host Information Collection

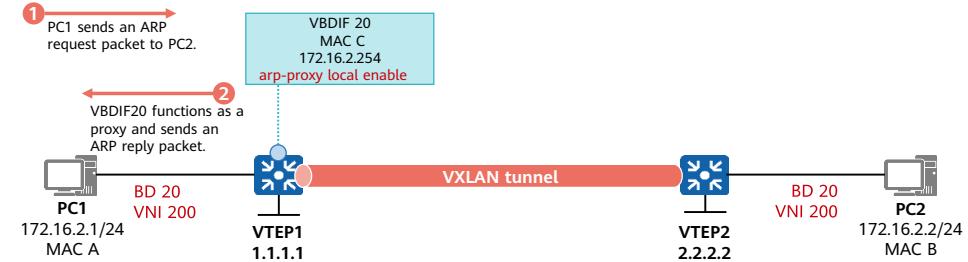
- The implementation of ARP broadcast suppression depends on the ARP broadcast suppression table. The generation of ARP broadcast suppression entries depends on Type 2 routes (IRB routes and host ARP advertisement) generated by BGP EVPN.
- By default, a Layer 3 gateway does not generate BGP EVPN routes based on local ARP information. You need to manually enable BGP EVPN host information collection. VTEPs then generate IRB routes based on ARP information.



- An ARP route carries the following valid information: host MAC address, host IP address, and Layer 2 VNI. An IRB route carries the following valid information: host MAC address, host IP address, Layer 2 VNI, and Layer 3 VNI. Therefore, IRB routes include ARP routes and can be used to advertise both host IP routes and host ARP entries.

Local Proxy ARP (1)

- After BGP EVPN host information collection is enabled on the entire network, the Layer 3 gateway learns 32-bit host routes of all hosts. In this way, the Layer 3 gateway can use host routes to perform Layer 3 symmetric IRB for traffic in the same BD.
- You can enable local proxy ARP on the VBDIF interface of the Layer 3 gateway. The VBDIF interface responds to ARP requests from downstream hosts for IP addresses on the same network segment. The Layer 3 gateway then performs Layer 3 forwarding for access to the IP addresses on the same network segment.

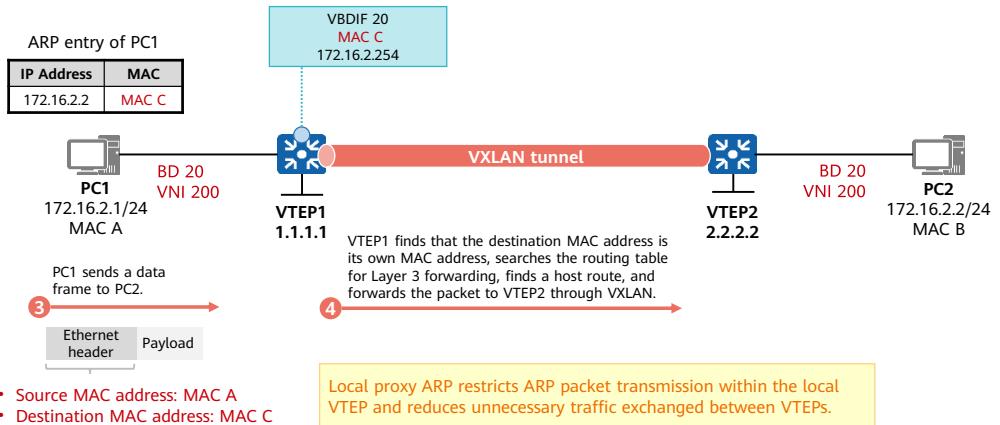


59 Huawei Confidential

HUAWEI

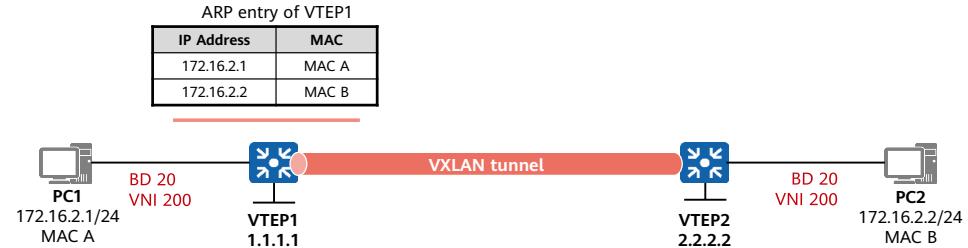
- On a VXLAN network, a BD is a broadcast domain. After receiving BUM packets, a VTEP broadcasts the packets in the BD. To reduce broadcast traffic, the network administrator usually configures access-side isolation or port isolation on the access side to isolate access users in a BD and prevent Layer 2 communication. However, with the increase of user services, users have higher requirements for communication. To meet the requirements, the network administrator can enable local proxy ARP on a VBDIF interface so that isolated access users in a BD can communicate with each other.

Local Proxy ARP (2)



Anycast Gateway

- When local proxy ARP is enabled, a VTEP only needs to maintain local ARP entries. ARP information transmitted by other VTEPs through BGP EVPN routes is not used during packet forwarding. In this case, the VTEP does not need to maintain ARP entries learned from other VTEPs.
- After the distributed gateway function is enabled, the VTEP processes only ARP packets received from user-side hosts and deletes learned network-side ARP entries.



61 Huawei Confidential



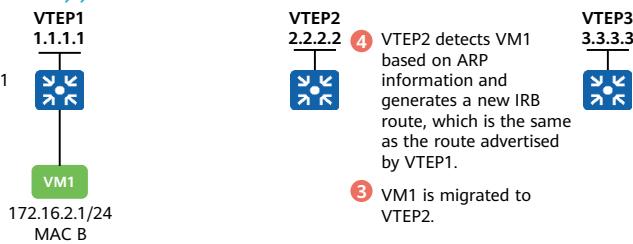
- Generally, the VBDIF interfaces with the same ID on different VTEPs are configured with the same MAC address. After the distributed gateway function is enabled, the VBDIF interfaces have the same IP address and MAC address, but no ARP conflict is reported. In addition, when hosts and VMs are migrated to different VTEPs, ARP resolution does not need to be performed on the gateway.

MAC Mobility (1)

- ② BGP EVPN route, with the sequence number of the extended community attribute MAC Mobility being 0.

MAC Mobility - Seq 0
Prefix = 172.16.2.1/24
MAC B
Next hop: VTEP1 (1.1.1.1)

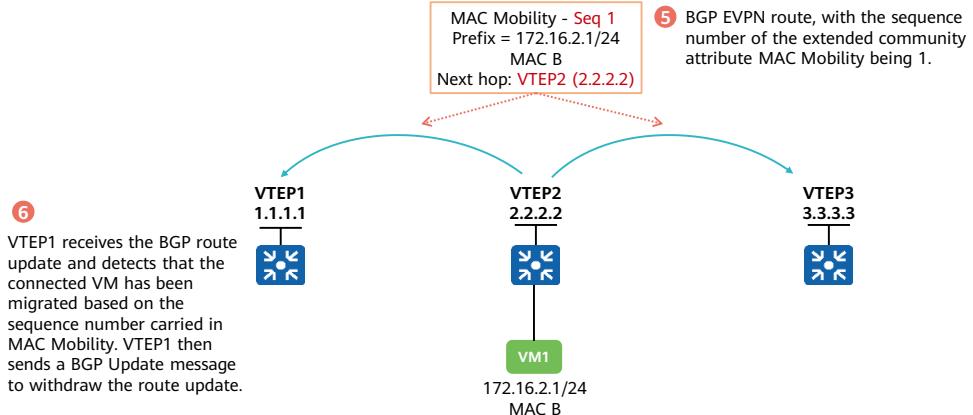
- ① VTEP1 learns ARP information of VM1 and generates and advertises an IRB route.



- ④ VTEP2 detects VM1 based on ARP information and generates a new IRB route, which is the same as the route advertised by VTEP1.

- ③ VM1 is migrated to VTEP2.

MAC Mobility (2)

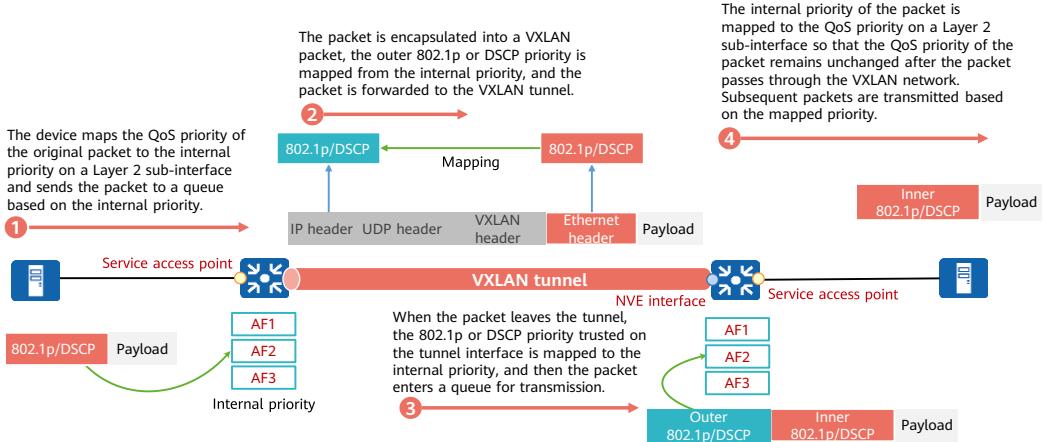


- The MAC Mobility extended attribute is used to announce the location change of a host or VM when the host or VM is migrated from one VTEP to another VTEP.

VXLAN QoS (1)

- Certain fields in the packet header record QoS information so that network devices can provide differentiated services.
- Packets carry different types of precedence field depending on the network type. For example, packets carry the 802.1p field on a VLAN network, the DSCP field on an IP network, and the EXP field on an MPLS network. If packets traverse different types of networks, the mapping between the precedence fields must be configured on the gateway. This configuration ensures that the packet priorities are retained regardless of the network type.
- VXLAN QoS provides differentiated quality assurance for VXLAN packets based on their internal priorities, which are assigned by devices to differentiate the service classes of packets. In VXLAN QoS implementation, devices map QoS priorities carried in original packets to internal priorities, and map internal priorities to the priorities of VXLAN packets.

VXLAN QoS (2)



65 Huawei Confidential

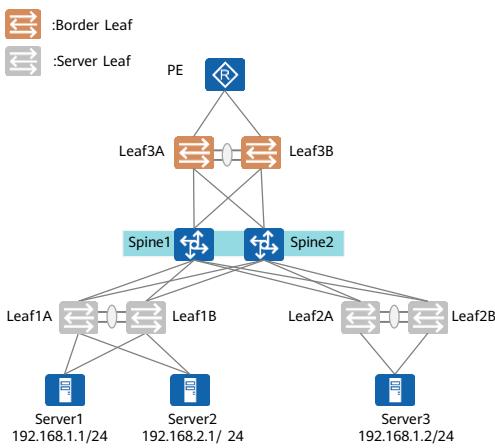
HUAWEI

- In step 2, after the device encapsulates the packet into a VXLAN packet, the QoS priority of the encapsulated packet is as follows:
 - By default, the outer 802.1p value of the encapsulated packet is mapped from the internal priority, and the inner 802.1p value of the encapsulated packet remains unchanged. After the **qos phb marking 8021p disable** command is configured in the Ethernet interface view, the outer 802.1p value is 0, and the inner 802.1p value remains unchanged.
 - By default, the outer DSCP value of the encapsulated packet is 0, and the inner DSCP value of the encapsulated packet remains unchanged. After the **qos phb marking dscp enable** command is configured in the Ethernet interface view, the outer DSCP value is mapped from the internal priority, and the inner DSCP value remains unchanged.
- After VXLAN encapsulation is complete, the local VTEP maps the internal priority based on the DSCP or 802.1p field in the outer packet before the packet arrives at the remote VTEP.

Contents

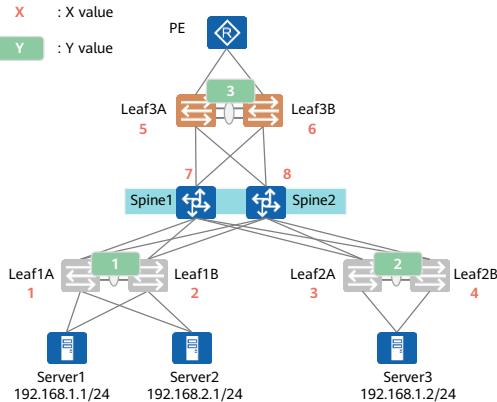
1. Background of VXLAN
2. Basic Concepts and Fundamentals of VXLAN
3. EVPN VXLAN Fundamentals
- 4. VXLAN Deployment Cases in Typical Scenarios**

Distributed Gateway (1)



- Networking requirements:
 - The entire network uses BGP EVPN to construct a VXLAN network with distributed gateways. Spines function as RRs to reflect EVPN routes to implement Layer 2 and Layer 3 communication between servers.
 - M-LAG is configured on all leaf nodes to ensure access link reliability.
 - Configure an egress route on Leaf 3 (Border Leaf) to allow Server 1 on the intranet to access the Internet.
- Configuration procedure:
 - Configure the M-LAG on the leaf node. (The configuration is not mentioned here.)
 - Configure the interface IP address and OSPF. (The configuration is not mentioned here.)
 - Configure BGP and enable BGP EVPN peers.
 - Configure a VXLAN tunnel.
 - Configure EVPN and VPN instances.
 - Configure a VXLAN Layer 3 gateway.
 - Configure service access points and egress routes.

Distributed Gateway (2)

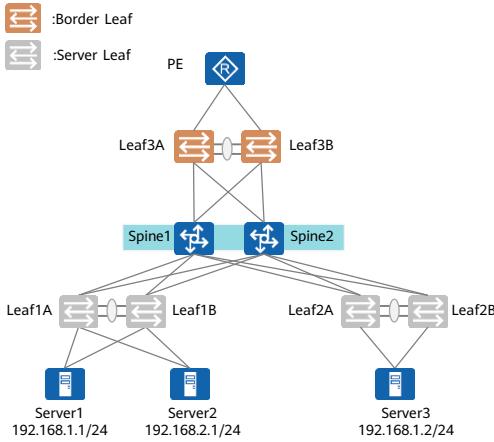


- Router ID planning: All devices use the IP address of the Loopback0 interface as the router ID. The IP address planning is 10.X.X.X, where X indicates the device ID, which is marked on the left.
- VTEP IP address planning: All devices use the IP address of the Loopback1 interface as the VTEP IP address. The IP address planning is 11.Y.Y.Y, where Y is marked on the left.

Question: Why do two leaf nodes in an M-LAG share the same VTEP IP address?

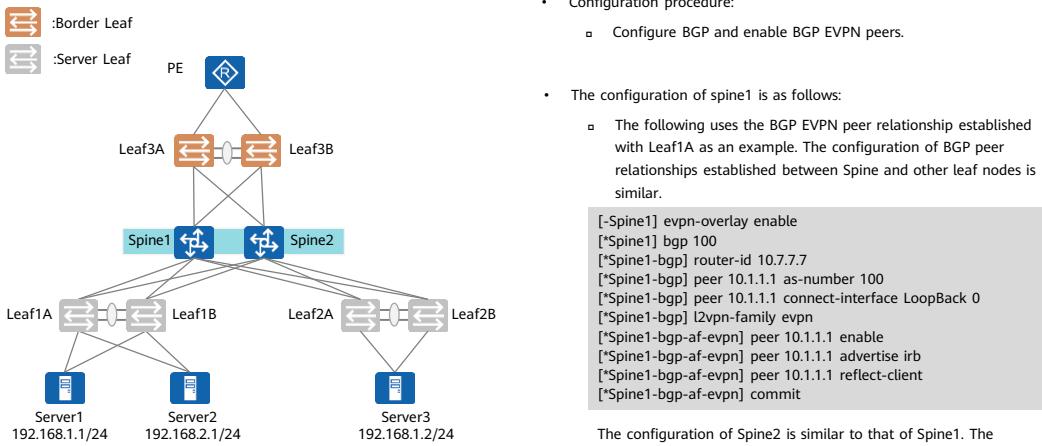
- Consideration: On a distributed VXLAN network, two leaf switches that form an M-LAG system function as dual-active access gateways. Ensure that the IP addresses and MAC addresses of the NVE interfaces on the two switches are the same to ensure normal traffic forwarding on the VXLAN network. On a typical data center network, all spine switches are fully connected, and at least two leaf switches are forwarded through IGP routes (to implement backup and load balancing).

Distributed Gateway (3)



- Configuration procedure:
 - Configure the M-LAG on the leaf node. (The configuration is not mentioned here.)
 - Configure the interface IP address and OSPF. (The configuration is not mentioned here.)
- Configuration notes:
 - After an M-LAG system is established for leaf nodes, monitor-links or best-effort routes need to be configured to forward traffic in the M-LAG system when the M-LAG system is faulty. You are advised to deploy best-effort routes on border nodes and monitor-links on server leaf nodes.
 - Each leaf device independently establishes OSPF and BGP neighbor relationships with the spine device. When an OSPF address is advertised, the VTEP IP address needs to be advertised to the OSPF process to ensure reachable routes during VXLAN tunnel establishment.

Distributed Gateway (4)

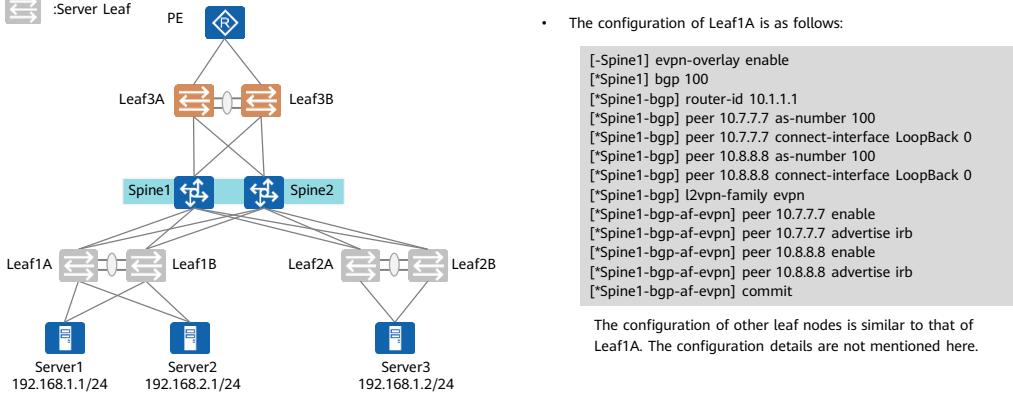


The configuration of Spine2 is similar to that of Spine1. The configuration details are not mentioned here.

Distributed Gateway (5)

:Border Leaf

:Server Leaf



- Configuration procedure:

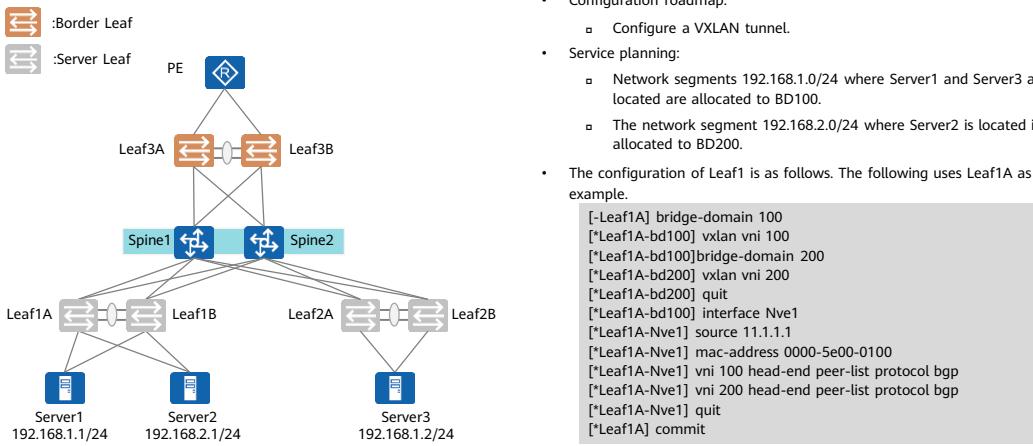
- Configure BGP and enable BGP EVPN peers.

- The configuration of Leaf1A is as follows:

```
[*Spine1] evpn-overlay enable
[*Spine1] bgp 100
[*Spine1-bgp] router-id 10.1.1.1
[*Spine1-bgp] peer 10.7.7.7 as-number 100
[*Spine1-bgp] peer 10.7.7.7 connect-interface LoopBack 0
[*Spine1-bgp] peer 10.8.8.8 as-number 100
[*Spine1-bgp] peer 10.8.8.8 connect-interface LoopBack 0
[*Spine1-bgp] 12vpn-family evpn
[*Spine1-bgp-af-evpn] peer 10.7.7.7 enable
[*Spine1-bgp-af-evpn] peer 10.7.7.7 advertise irb
[*Spine1-bgp-af-evpn] peer 10.8.8.8 enable
[*Spine1-bgp-af-evpn] peer 10.8.8.8 advertise irb
[*Spine1-bgp-af-evpn] commit
```

The configuration of other leaf nodes is similar to that of Leaf1A. The configuration details are not mentioned here.

Distributed Gateway (6)



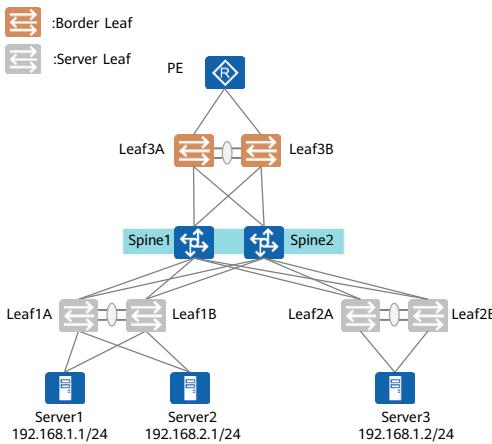
- Configuration roadmap:
 - Configure a VXLAN tunnel.
- Service planning:
 - Network segments 192.168.1.0/24 where Server1 and Server3 are located are allocated to BD100.
 - The network segment 192.168.2.0/24 where Server2 is located is allocated to BD200.
- The configuration of Leaf1 is as follows. The following uses Leaf1A as an example.

```
[*Leaf1A] bridge-domain 100
[*Leaf1A-bd100] vxlan vni 100
[*Leaf1A-bd100]bridge-domain 200
[*Leaf1A-bd200] vxlan vni 200
[*Leaf1A-bd200] quit
[*Leaf1A-bd100] interface Nve1
[*Leaf1A-Nve1] source 11.1.1.1
[*Leaf1A-Nve1] mac-address 0000-5e00-0100
[*Leaf1A-Nve1] vni 100 head-end peer-list protocol bgp
[*Leaf1A-Nve1] vni 200 head-end peer-list protocol bgp
[*Leaf1A-Nve1] quit
[*Leaf1A] commit
```

The configuration of Leaf2 and Leaf3 is similar to that of Leaf1. The configuration details are not mentioned here.



Distributed Gateway (7)



- Configuration roadmap:
 - Configure EVPN and VPN instances.
- The configuration of Leaf1 is as follows. The following uses Leaf1A as an example.

```
[*Leaf1A] bridge-domain 100
[*Leaf1A-bd100] evpn
[*Leaf1A-bd100-evpn] route-distinguisher 2:2
[*Leaf1A-bd100-evpn] vpn-target 100:1
[*Leaf1A-bd100-evpn] vpn-target 1000:1 export-extcommunity
[*Leaf1A-bd200-evpn] bridge-domain 200
[*Leaf1A-bd200-evpn] route-distinguisher 3:3
[*Leaf1A-bd200-evpn] vpn-target 200:1
[*Leaf1A-bd200-evpn] vpn-target 1000:1 export-extcommunity
[*Leaf1A-bd200-evpn] commit
```

EVPN Instance Configuration

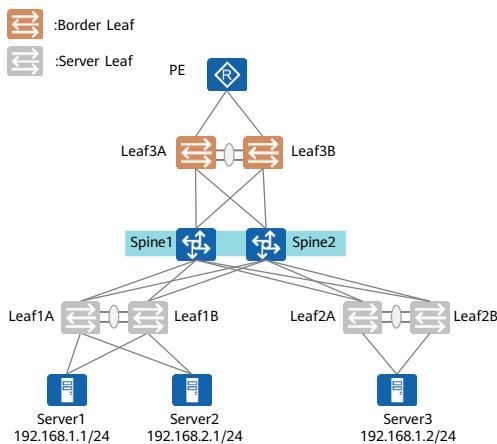
```
[*Leaf1A] ip vpn-instance vpn1
[*Leaf1A-vpn-instance-vpn1] vxlan vni 10000
[*Leaf1A-vpn-instance-vpn1] route-distinguisher 22:22
[*Leaf1A-vpn-instance-vpn1-af-ipv4] vpn-target 1000:1
[*Leaf1A-vpn-instance-vpn1-af-ipv4] vpn-target 1000:1 evpn
[*Leaf1A-vpn-instance-vpn1-af-ipv4] commit
```

IP VPN Instance Configuration

The configuration of Leaf2 and Leaf3 is similar to that of Leaf1. The configuration details are not mentioned here.

- Leaf2 needs to be configured with only the EVPN instance and IP VPN instance of BD100.

Distributed Gateway (8)

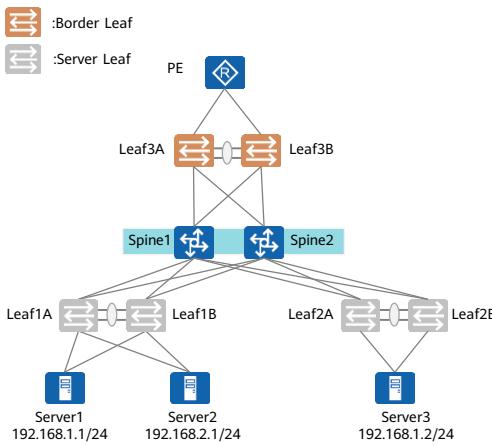


- Configuration roadmap:
 - Configure a VXLAN Layer 3 gateway.
 - The following figure shows the configuration of Leaf1. The following uses Leaf1A as an example.

```
[*Leaf1A] interface Vbdif100
[*Leaf1A-Vbdif100] ip binding vpn-instance vpn1
[*Leaf1A-Vbdif100] ip address 192.168.1.1 24
[*Leaf1A-Vbdif100] mac-address 0000-5e00-0102
[*Leaf1A-Vbdif100] arp collect host enable
[*Leaf1A-Vbdif100] arp distribute-gateway enable
[*Leaf1A-Vbdif100] quit
[*Leaf1A] interface Vbdif200
[*Leaf1A-Vbdif100] ip binding vpn-instance vpn1
[*Leaf1A-Vbdif100] ip address 192.168.2.1 24
[*Leaf1A-Vbdif100] mac-address 0000-5e00-0102
[*Leaf1A-Vbdif100] arp collect host enable
[*Leaf1A-Vbdif100] arp distribute-gateway enable
[*Leaf1A-Vbdif100] commit
```

The configuration of Leaf2 is similar to that of Leaf1. The configuration details are not mentioned here.

Distributed Gateway (9)



- Configuration roadmap:

- Configure service access points and egress routes.

Configure an egress route on Leaf3 and import BGP routes. The following uses Leaf3A as an example.

```
[*Leaf3A] ip route-static 0.0.0.0 0.0.0.0 100.1.1.2 vpn-instance vpn1
[*Leaf3A] bgp 100
[*Leaf3A-bgp] ipv4-family vpn-instance vpn1
[*Leaf3A-bgp-vpn1] default-route imported
[*Leaf3A-bgp-vpn1] import-route static
[*Leaf3A-bgp-vpn1] commit
```

IP address of the port connecting Leaf3 to PE

Configure all leaf nodes to advertise IP prefix routes to BGP peers.

```
[*Leaf1A] bgp 100
[*Leaf1A-bgp] ipv4-family vpn-instance vpn1
[*Leaf1A-bgp-vpn1] import-route direct
[*Leaf1A-bgp-vpn1] advertise l2vpn evpn
[*Leaf1A-bgp-vpn1] commit
```

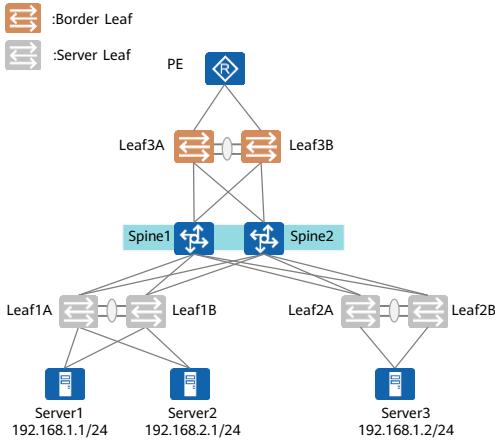
Configure a service access point. The following uses Server1 as an example:

```
[*Leaf1] interface Eth-Trunk1/1 mode l2
[*Leaf1-GE1/0/1.1] encapsulation dot1q vid 10
[*Leaf1-GE1/0/1.1] bridge-domain 100
[*Leaf1-GE1/0/1.1] commit
```

Server access port

Server1 data carries VLAN 10.

Distributed Gateway (10)



- Result verification:

- Run the display vxlan tunnel command on Leaf1A to check the VXLAN tunnel.

```
[<-Leaf1A] display vxlan tunnel
Number of vxlan tunnel: 1
Tunnel ID Source Destination State Type Uptime
-----
4026531841 11.1.1.1 11.2.2.2 up dynamic 0032h21m
4026531842 11.1.1.1 11.3.3.3 up dynamic 0032h25m
```

- After the configuration is complete, Layer 2 and Layer 3 communication can be implemented between different servers.
- Check the egress routes on Leaf1 and Leaf3, for example, Leaf1A.

```
[<-Leaf1A] display ip routing-table vpn-instance vpn1
Destination/Mask Proto Pre Cost Flags NextHop Interface
0.0.0.0/0 Static 60 0 D 100.1.1.2 vbdif1000
```

- After the configuration is complete, Server1 accesses the external network through Leaf1, Spine, and Leaf3 in sequence.



Quiz

1. (True or false) BGP EVPN Type 2 host IP routes can be used to transmit ARP information. ()
 - A. True
 - B. False
2. (Single-answer question) Which of the following statements about BGP EVPN is false? ()
 - A. Carrying routes through MP_REACH_NLRI.
 - B. Carrying the RT through the extended community attribute
 - C. Carrying L2 VNI and L3 VNI in MP_REACH_NLRI
 - D. Carrying the next hop address of a route through the Next_Hop attribute

1. A
2. D

Summary

- VXLAN uses a Layer 3 routed network as its underlay network and uses tunnels to build an overlay virtual network, supporting a large number of tenant networks.
- VXLAN does not define a control plane. To limit the flooding of BUM traffic, VXLAN needs to use other control plane protocols to optimize BUM traffic forwarding.
- BGP EVPN extends BGP by defining several types of BGP EVPN routes. These BGP EVPN routes can be used to transmit VTEP addresses, host information, and routing information, effectively helping VXLAN limit the flooding of BUM traffic.

Thank you.

把数字世界带入每个人、每个家庭。

每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



Technical Principles and Application of M-LAG



Foreword

- The data center carries core computing functions of enterprise production. The network has requirements for high-performance load balancing and high service reliability. Important service systems have requirements for uninterrupted services during device upgrade. This puts forward a high requirement on the availability of the network system.
- The CloudFabric solution uses Multichassis Link Aggregation Group (M-LAG) and Virtual eXtensible Local Area Network (VXLAN) to implement end-to-end reliability, ensuring that service systems can run properly in device failure and upgrade scenarios.
- This document describes the principles and applications of the M-LAG technology.

Objectives

Upon completion of this course, you will be able to:

- Describe the definition, usage, and features of M-LAG.
- Differences Between Stacking and M-LAG.
- Describe the technical principles of M-LAG.
- Describe the network deployment mode and typical application networking of M-LAG.

Contents

- 1. Overview of M-LAG**
2. M-LAG Fundamentals
3. M-LAG Failure Protection
4. M-LAG Deployment
5. M-LAG Best Practices

Overview of LAG

- SW1 and SW2 are connected by using multiple links, for example, four links. The four links can be bundled into an Eth-Trunk.
 - Increase the bandwidth (sum of the bandwidth of the four links)
 - Improve reliability (where some links are down, other links can take over the forwarding task)
 - Load balancing (Traffic is allocated to different links based on the 5-tuple hash algorithm to improve bandwidth utilization.)
- However, if SW1 or SW2 fails, the traffic transmitted through SW1 or SW2 is interrupted. In this case, board-level link aggregation cannot meet reliability requirements.



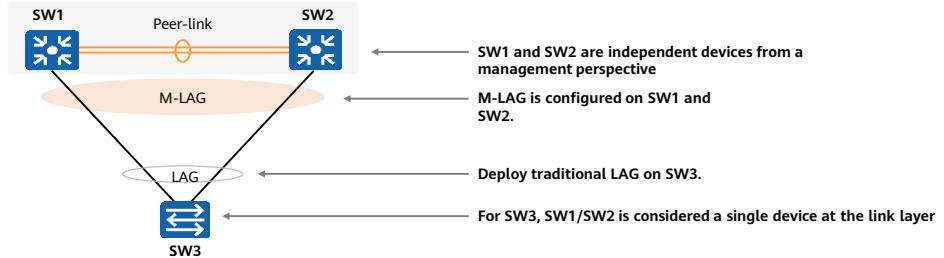
5 Huawei Confidential

 HUAWEI

- Huawei devices use Eth-Trunk as the link aggregation technology. You can configure an Eth-Trunk on a device and add multiple interfaces (for example, four interfaces) to the Eth-Trunk.

Overview of M-LAG

- M-LAG (Multichassis Link Aggregation Group, Inter-Device Link Aggregation Group): A mechanism that implements inter-device link aggregation. This mechanism improves the reliability of link aggregation from the link level to the device level. In addition, M-LAG member devices forward traffic through load balancing, forming a dual-active system.
- M-LAG is also a virtualization technology. From the perspective of the peer device connected to the M-LAG port, the M-LAG port is connected to a logical switch.

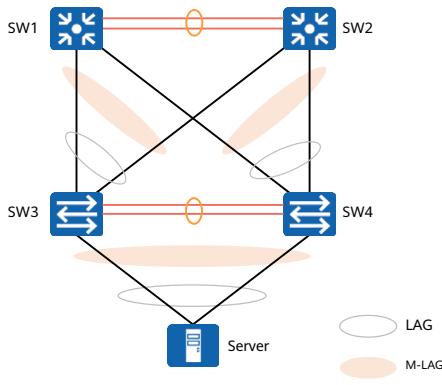


6 Huawei Confidential

 HUAWEI

- There are several options to improve network reliability, such as STP+VRRP and stacking. However, these options have obvious problems, such as:
 - STP+VRRP
 - The STP blocking mechanism leads to low Layer 2 link usage.
 - The Master/Backup backup function of VRRP leads to low resource utilization.
 - Only the Master/Backup mode is supported for server access.
 - Stacking technology (to be compared later)
 - Fast stack upgrade reduces the service interruption time, but increases the upgrade time and increases the upgrade risk. The control plane is centralized, and faults may spread on member devices.
 - The master control plane needs to control the forwarding planes of all stack members, increasing the CPU load.
- Therefore, M-LAG is often used in data center network to improve network reliability.

Advantages of M-LAG



- To meet the requirements for higher network reliability, M-LAG uses link aggregation between multiple devices to achieve higher reliability and improve link utilization.
- Advantage:
 - Implements inter-device link aggregation, improving Layer 2 link utilization.
 - The active-active gateway technology of M-LAG improves the utilization of device and link resources.
 - Servers can use link aggregation to implement active-active access devices and implement load balancing.

Comparison Between M-LAG and Stack



A stack implements virtualization on the management plane, protocol plane (control plane), and data plane, and member devices are highly coupled.

M-LAG implements virtualization on some data planes and some protocol planes (control planes) and has low coupling between member devices.

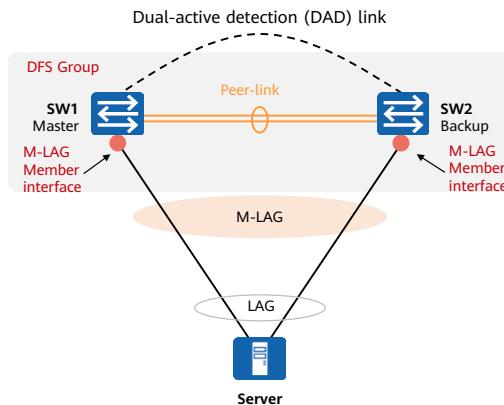
Dimension	Stacked	M-LAG
Reliability	Protocol planes are centralized, and faults may spread on member devices.	Excellent. Protocol planes are independent (partially centralized), and fault domains are isolated.
O&M	Excellent. The number of management nodes is reduced and the configuration is simple.	Two switches need to be managed.
Fault convergence performance	Excellent. The convergence performance is close to that of a single device.	Failover information needs to be passed through a protocol.
Upgrade complexity	High: Fast stack upgrade shortens the service interruption time, but increases the upgrade time and increases the upgrade risk.	Excellent. The two switches are upgraded independently without interrupting service access. The risk is low and applications are unaware of the upgrade.
Service interruption time during an upgrade	Longer upgrade: In the typical networking, the service interruption time is about 20s to 1 minute, which is closely related to the service volume.	Short: Traffic is interrupted in seconds.

- Application scenarios of stacking:
 - There is no requirement on the interruption duration during software upgrade.
 - Simple maintenance is required.
- Application scenarios of M-LAG:
 - The service interruption time during the software upgrade is high.
 - Higher reliability.
 - It is acceptable to add a certain degree of maintenance complexity.

Contents

1. Overview of M-LAG
2. **M-LAG Fundamentals**
 - Basic Concepts of M-LAG
 - Basic Features of M-LAG
 - M-LAG Traffic Forwarding Process
3. M-LAG Failure Protection
4. M-LAG Deployment
5. M-LAG Best Practices

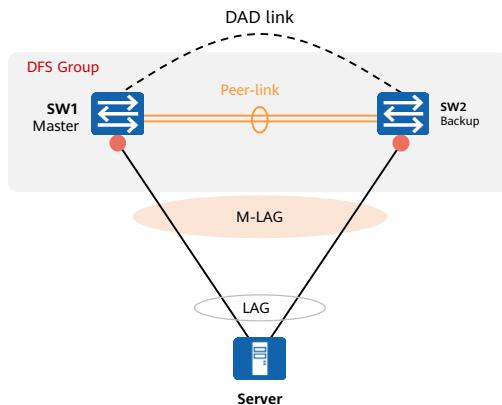
Basic Concepts of M-LAG (1)



- **Dynamic fabric service group (DFS) group:** It is used to pair M-LAG devices. The interface status and entries between M-LAG dual-homing devices must be synchronized using the DFS group protocol.
- **Peer-link:** a Layer 2 link used to exchange negotiation packets, synchronize device information, and transmit some traffic. After an interface is configured as a peer-link interface, other services cannot be configured on the interface.
- **DFS master device (Master):** indicates the master device with M-LAG deployed.
- **DFS backup device (Backup):** M-LAG is deployed and the device is in the standby state.
- **M-LAG member interface:** Eth-Trunk interface on the M-LAG Master/Backup connected to user-side hosts or switching devices.

- A DFS group consists of a master device and a backup device. Under normal circumstances, both the master and backup devices forward service traffic and their forwarding behaviors are the same. The master and backup devices have different forwarding behaviors only when a fault occurs.
 - When no fault occurs, both the master and backup devices forward traffic.
 - When two master devices are detected, service interfaces on the backup device enter the Error-Down state.
- The peer-link is used to achieve the following:
 - Transmit DFS group protocol packets.
 - Transmit synchronization packets used for synchronizing MAC address entries and ARP entries between M-LAG master and backup devices.
 - Forward inter-device traffic sent from non-M-LAG member interfaces or traffic received from an M-LAG member interface when downstream devices are single-homed to the M-LAG due to a fault.
- By default, the peer-link allows packets from all VLANs to pass through. If you do not want the peer-link to allow packets from some VLANs to pass through, you need to configure the VLANs separately.
- To improve the reliability of the peer-link, you are advised to add multiple links to a LAG and configure the aggregated link as the peer-link. However, even if there is only one link, you need to add it to the LAG.

Basic Concepts of M-LAG (2)

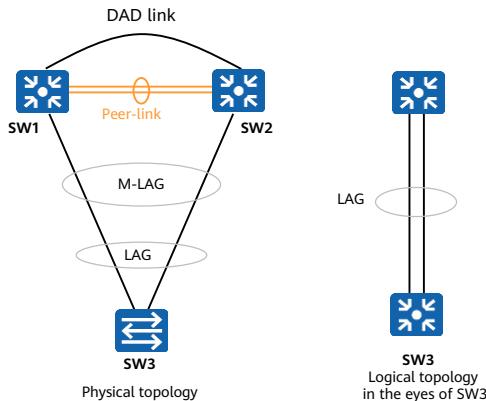


- **DAD link:** A DAD link, also called a heartbeat link, is a Layer 3 link used by Master/Backup in an M-LAG to send DAD packets.
- **HeartBeat (HB) DFS master device:** A device that negotiates the master state through a heartbeat link.
- **HB DFS standby device:** indicates the standby device negotiated through the heartbeat link.

- Under normal circumstances, the DAD link does not participate in any traffic forwarding behaviors in the M-LAG. It is only used to detect whether two master devices exist when a fault occurs. The DAD link can be an external link, for example, if the M-LAG is connected to an IP network and the two member devices can communicate through the IP network, the link that enables communication between the member devices can function as the DAD link. An independent link that provides Layer 3 reachability can also be configured as the DAD link, for example, a link between management interfaces of the member devices can function as the DAD link.
- Under normal circumstances, the HB DFS master/backup status does not affect traffic forwarding behaviors in the M-LAG. It is used only in secondary fault recovery scenarios.
 - If a fault on the original DFS master device is rectified and the peer-link is still faulty, the corresponding interfaces on the backup device are triggered to enter the Error-Down state based on the HB DFS master/backup status. This mechanism prevents abnormal traffic forwarding in the scenario where two master devices exist.

Base Protocol - LACP

- LACP needs to be configured on M-LAG member interfaces to detect faults such as link layer faults and incorrect link connections, improving link reliability.



According to the LACP principle, to enable SW3 in the scenario shown in the left figure to consider the two M-LAG member switches as one LACP node, the LACP packet received by SW3 must meet the following conditions:

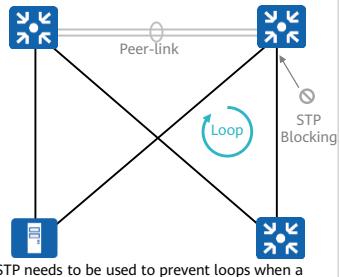
- Same LACP system ID.
- The LACP system priority is the same.
- The LACP port priority is the same.
- The LACP key (identifying one Eth-Trunk) is the same.
- LACP ports do not conflict.

- LACP can be deployed on the peer-link interface, M-LAG member interface, or Eth-Trunk interface of the access device.

Base Protocol - STP

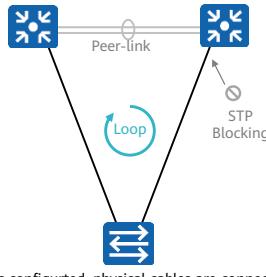
- M-LAG supports dual-homing and logical loop-free networks. This does not mean that STP is not required. For example, the following three scenarios are used.

Scenario 1: Preventing loops caused by incorrect cable connections or configurations



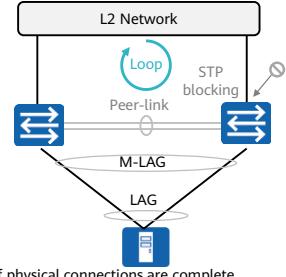
STP needs to be used to prevent loops when a port planned for connecting to a port on a server is incorrectly connected to a switch or uplink of the switch is connected to a non-M-LAG member interface.

Scenario 2: Connecting physical cables before configuring M-LAG interfaces



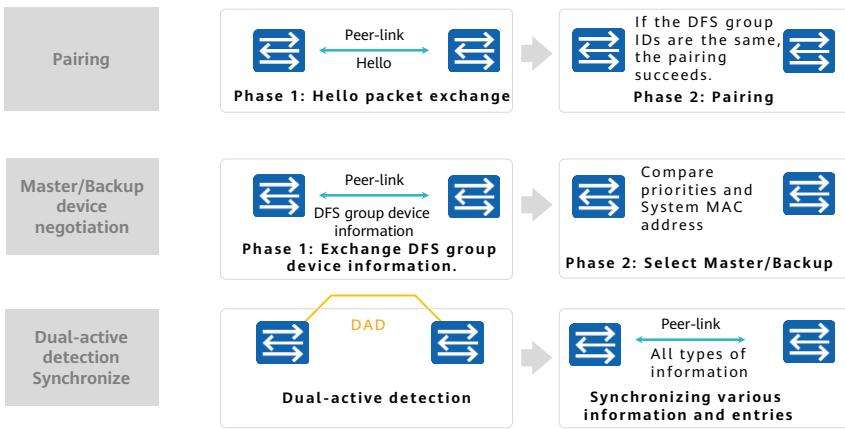
M-LAG is configured. physical cables are connected, and a loop occurs on the network. In this case, STP needs to be deployed to prevent loops.

Scenario 3: M-LAG Access to a Layer 2 Network



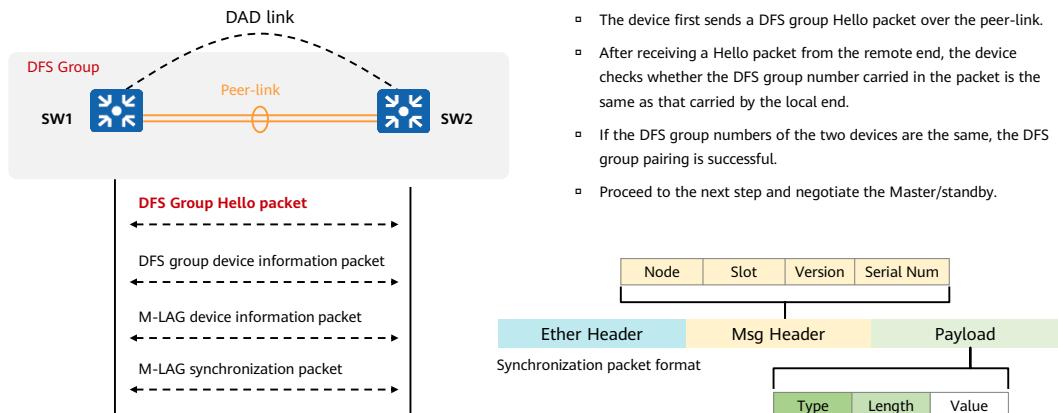
If physical connections are complete before M-LAG configuration, loops exist on the network. In this case, STP needs to be deployed to prevent loops.

M-LAG Setup Process



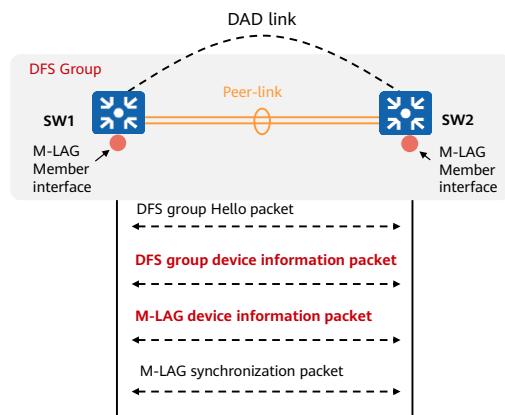
M-LAG Pairing

- After the M-LAG configuration is complete on two devices:
 - The device first sends a DFS group Hello packet over the peer-link.
 - After receiving a Hello packet from the remote end, the device checks whether the DFS group number carried in the packet is the same as that carried by the local end.
 - If the DFS group numbers of the two devices are the same, the DFS group pairing is successful.
 - Proceed to the next step and negotiate the Master/standby.



- A customized message header is encapsulated in the outer Ethernet header. The customized message header contains the following information:
 - Version: indicates the protocol version, which is used to identify the M-LAG version of M-LAG member devices.
 - Message Type: indicates the type of a packet, which can be Hello or Synchronization.
 - Node: indicates the device node ID.
 - Slot: indicates the slot ID of the card that needs to receive messages. For a fixed device, the value is the stack ID.
 - Serial Number: indicates the protocol serial number, which is used to improve reliability.
- The user-defined message header contains the normal packet data, including the information that needs to be exchanged or synchronized. For example, the DATA field of a Hello packet contains the DFS group ID, priority, and MAC address of the device. However, the synchronization packet DATA contains some entries and status information.

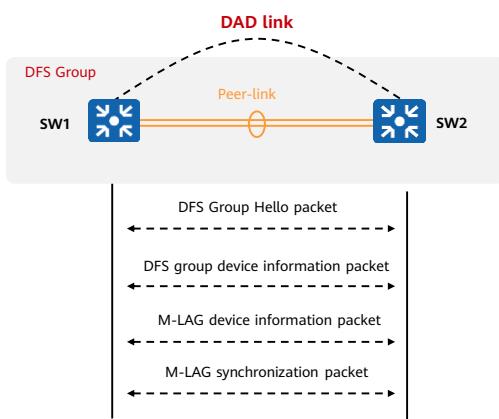
Negotiate Master/Backup



- DFS Group Negotiation Master/Backup
 - After the pairing succeeds, the two devices send a DFS group information packet to the peer through the peer-link. The device determines the Master/Backup status of the DFS group based on the DFS group priority and system MAC address carried in the packet. (If the priority is higher, the device functions as the master. If the priority is the same, the device compares the MAC address of the device. If the MAC address is smaller, the device functions as the master.)
 - In normal cases, the forwarding behavior of the master and backup devices is the same. The forwarding behavior of the Master/Backup device is different only in the case of a fault.
- M-LAG member interface negotiation Master/Backup
 - In addition to Master/Backup negotiation, member interfaces also use M-LAG packets to negotiate the master/backup status.

- M-LAG member interface negotiation Master/Backup:
 - After the DFS group negotiates the Master/Backup status, the two M-LAG devices send M-LAG device information packets over the peer-link. The packets carry the configurations of M-LAG member interfaces. After the information about the M-LAG member interfaces is synchronized, the Master/Backup status of the M-LAG member interfaces is determined.
 - When member interface information is synchronized from the peer end, the M-LAG member interface whose status changes from Down to Up first becomes the master M-LAG member interface, and the M-LAG member interface on the peer end becomes the backup.
- The forwarding behavior of the Master/Backup role of the M-LAG member interface is different only in the M-LAG multicast access scenario.
 - In versions earlier than V200R003C00, only the M-LAG member interface in the master state forwards multicast traffic to receivers. In V200R003C00 and later versions, the M-LAG member interface in the master/backup state can forward multicast traffic to receivers. Load balancing is implemented. When the versions of two devices in an M-LAG system are different, the multicast traffic forwarding rule of the earlier version prevails.

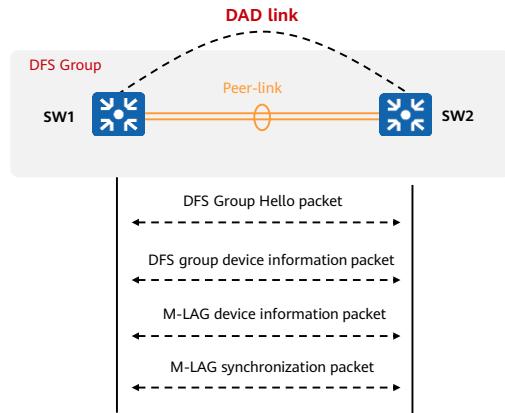
Dual-Active Detection



- After the M-LAG Master/Backup is negotiated, the two devices send M-LAG DAD packets at an interval of **1s** over the DAD link. Once the device detects a peer-link fault, it sends three DAD packets at an interval of **100 ms** to accelerate the detection. When the two devices can receive the packets from the peer device, the active-active system starts to work properly.
- After the peer-link fails, the DAD determines that the other device is running. The service port on the standby device is set to the Error-down state.
- Key deployment points:
 - Independent links are used to carry DAD traffic. Peer-links cannot be reused.
 - You are advised to deploy the DAD through the out-of-band management network port, which reduces costs.
 - DAD can also be deployed on an independent Layer 3 service interface.

- If the peer-link fails and the two member switches continue to run, network services will be affected.
 - Forwarding entries (ARP and MAC addresses) on SW1 and SW2 are not synchronized, which may cause forwarding exceptions.
 - In the V-STP scenario, messages cannot be synchronized through the peer-link. As a result, STP calculation may be abnormal.
- To improve the reliability of the M-LAG system, you need to configure DAD. Normally, DAD links do not participate in any forwarding behavior of the M-LAG. This command is used only when the DFS group pairing fails or the peer-link fails. Therefore, the M-LAG does not work properly even if DAD fails. The DAD link can be carried over an external network. (For example, if M-LAG is connected to an IP network, two dual-homing devices can communicate with each other through the IP network. In this case, the interworking link can be used as a dual-active detection link.) You can also configure a reachable Layer 3 link as the DAD link (for example, through the management interface).
 - (Recommended) DAD links communicate with each other through management network ports. The IP addresses of management network ports bound to DFS groups must be able to communicate with each other. The management network ports must be bound to VPN instances to isolate DAD packets from service traffic.

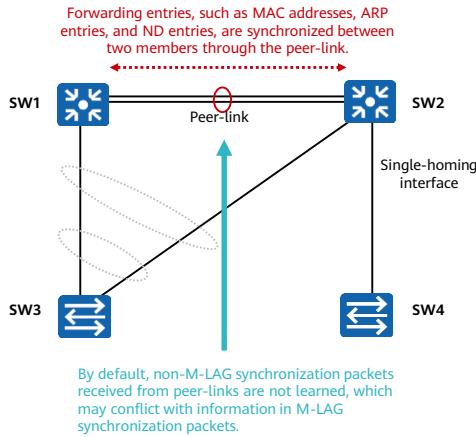
Synchronizing M-LAG Device Data - Synchronizing Device Information



- To ensure that the connected devices regard the M-LAG system as a logical device, the two switches in the M-LAG system must have the same device information (partial) and forwarding entries. This ensures that the fault of either device does not affect traffic forwarding and services are not interrupted.
- Therefore, after the M-LAG system works properly, the two devices send M-LAG synchronization packets over the peer-link to synchronize the information about the peer end in real time. The device information includes the device name, system MAC address, software version, M-LAG status, STP protocol packets, and VRRP packets.

- The synchronization information includes the device name, system MAC address, software version, M-LAG status, STP status, VRRP priority, DR priority, ACL, and LACP information.

M-LAG Device Data Synchronization - Forwarding Entry Synchronization



- Common forwarding entries that need to be synchronized include the MAC address table, ARP table, ND table, and IGMP multicast table.

- Synchronization principles:**

- The entries learned on the M-LAG interface must be synchronized to the peer device. After receiving the message, the peer device changes the interface corresponding to the entry to the M-LAG interface on the local device.
- The entries learned on the isolated port must be synchronized to the peer device. After receiving the message, the peer device changes the interface corresponding to the entry to the peer-link.

- Why does the peer-link disable the learning of related entries or protocols?**

- If the peer-link interface learning function is enabled, the peer-link interface may conflict with the forwarding entry synchronized by the M-LAG DFS protocol.
- Because the peer-link interface is disabled from learning related entries or protocols, the entries learned by the isolated interface need to be synchronized to the peer end.

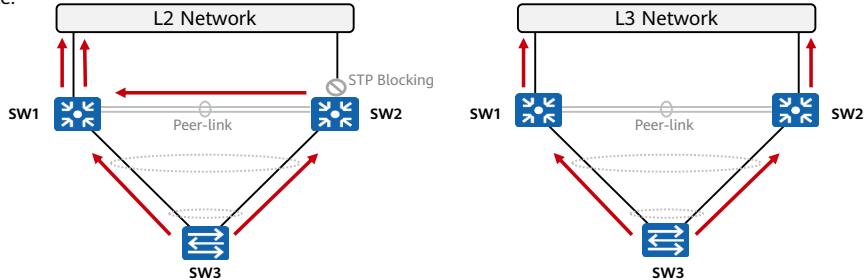
- The synchronization information includes MAC addresses, ARP entries, ND entries, IGMP entries, and DHCP snooping entries.

Contents

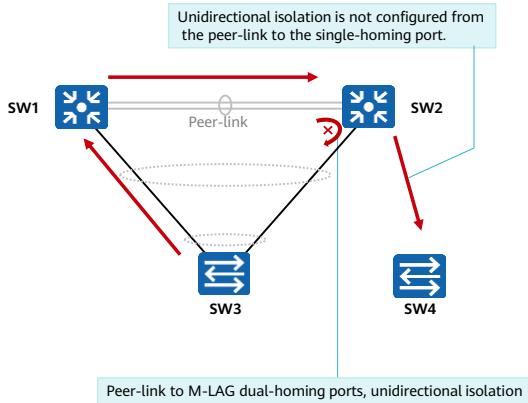
1. Overview of M-LAG
2. **M-LAG Fundamentals**
 - Basic Concepts of M-LAG
 - **Basic Features of M-LAG**
 - M-LAG Traffic Forwarding Process
3. M-LAG Failure Protection
4. M-LAG Deployment
5. M-LAG Best Practices

Local Preferential Forwarding

- Preferential local forwarding applies only to known unicast traffic (upstream and downstream).
 - Layer 2 unicast traffic: If the outbound interface in the MAC address table contains both the peer-link interface and the local interface, the local interface is preferentially sent. (If there is no local outbound interface, the peer-link can be used.)
 - Layer 3 unicast traffic: Check the routing table of each active-active gateway and forward traffic based on the local table.

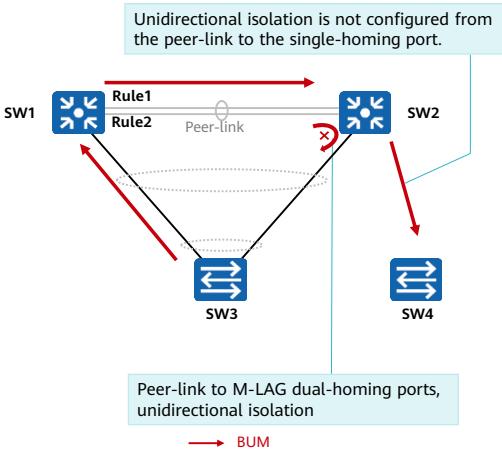


Unidirectional Isolation to Prevent Loops - BUM Packets (1)



- In M-LAG networking, a loop exists from the physical perspective. Loops greatly affect Layer 2 forwarding. How does M-LAG solve this problem?
 - For packets forwarded at Layer 2, the M-LAG uses a unidirectional isolation technology to prevent loops on the Layer 2 network.

Unidirectional Isolation to Prevent Loops - BUM Packets (2)



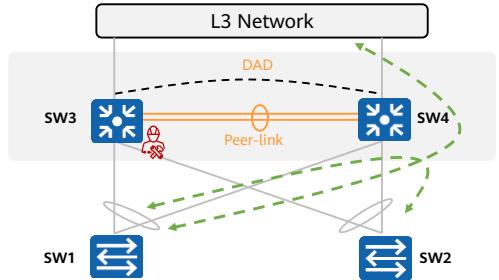
- When SW3 is connected to M-LAG in active-active mode, global ACL configurations are delivered in the following sequence by default:
 - Rule 1: Allows **Layer 3 unicast packets** with the source interface being the peer-link interface and the destination interface being the M-LAG member interface to pass through.
 - Rule 2: **All packets** with a peer-link interface as the source interface and an M-LAG member interface as the destination interface are rejected.
- Unidirectional isolation: M-LAG devices use the ACL rule group to implement unidirectional isolation between peer-link interfaces and M-LAG member interfaces. **Flooding traffic** such as broadcast traffic from a peer-link interface to an M-LAG member interface is isolated.

- Prerequisites for the unidirectional isolation mechanism to take effect
 - When M-LAG master and backup devices are negotiated, the system checks whether the access device is dual-homed to the M-LAG using M-LAG synchronization packets. If the access device is dual-homed to the M-LAG, the two M-LAG devices deliver the unidirectional isolation configuration of the corresponding M-LAG member interface to isolate traffic from peer-link interfaces to M-LAG member interfaces. Unidirectional isolation in the M-LAG loop prevention mechanism takes effect only for flooding traffic such as broadcast traffic.
 - If the access device is single-homed to the M-LAG, the M-LAG does not deliver the unidirectional isolation configuration of the corresponding M-LAG member interface.
- Canceling unidirectional isolation: When an M-LAG device detects that the local M-LAG member interface is in Down state, the device sends M-LAG synchronization packets through the peer-link to instruct the remote device to revoke the automatically delivered unidirectional isolation ACL rule group of the corresponding M-LAG member interface.

M-LAG Upgrade in Maintenance Mode

- If SW3 needs to be upgraded in the networking shown in the following figure, switch traffic to SW4 by shutting down the interface or modifying the link cost of the routing protocol, and then upgrade SW3. After SW3 is upgraded, restore the interface status or the cost value of the routing protocol link and switch traffic back to SW3. As a result, packet loss occurs in north-to-south traffic due to routing protocol convergence or ECMP path switching, and packet loss occurs in south-to-north and east-west traffic due to Eth-Trunk interface status changes.
 - M-LAG upgrade in maintenance mode allows you to run commands in the maintenance mode view to switch traffic from the device to be upgraded to the backup device and then restart the device. This reduces the packet loss rate during the upgrade and improves upgrade reliability.
- Started**
 - Preparing for the upgrade (including the device status, upgrade files, and upgrade tools)**
 - Traffic switchover**
 - Upgrading the Device (You can upgrade the Main Device first)**
 - Verifying the Upgrade**
 - After the upgrade succeeds, the traffic is switched back and the next phase starts.
 - If the upgrade fails, the traffic is switched back. Perform the upgrade again after the check.
 - After 10 minutes, the device status and services are normal, and the other device can be upgraded.**

25 Huawei Confidential

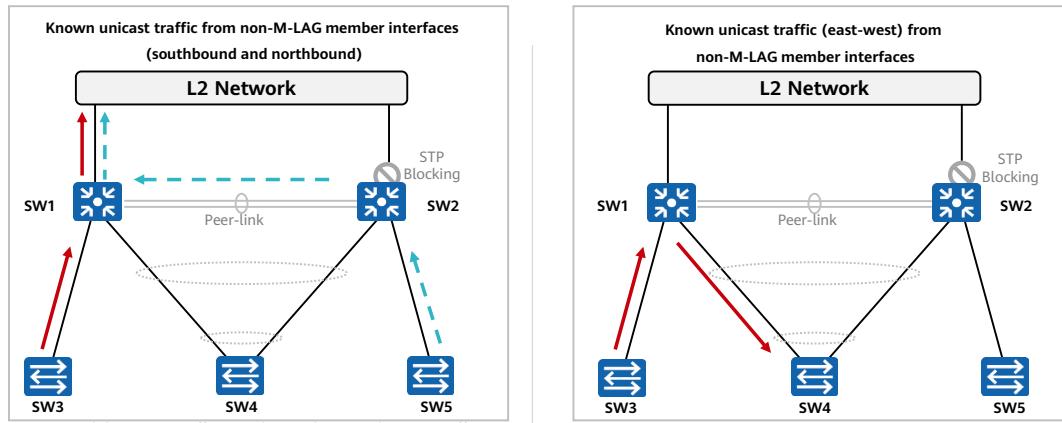


- Since V200R020C10, M-LAG can be upgraded in maintenance mode. M-LAG upgrade in maintenance mode allows you to run commands in the maintenance mode view to switch traffic from the device to be upgraded to the backup device and then restart the device. This reduces the packet loss rate during the upgrade and improves upgrade reliability.
- The upgrade in M-LAG maintenance mode is controlled by a license. By default, the upgrade in M-LAG maintenance mode is disabled on a newly purchased device. To use this function, apply for and purchase a license.
- Detailed operations:
 - SW1 and SW2 are dual-homed to the network through M-LAG, and SW3 and SW4 are dual-homed to the network through routing protocols.
 - In the M-LAG maintenance mode scenario, enter the maintenance mode of SW3 and perform the following configurations before the upgrade.
 - On SW3, change the OSPF and OSPFv3 cost values, or change the MED and Local_Pref values of BGP and BGP4+ to lower the route advertisement priority and switch the network-side traffic destined for SW3 to SW4.
 - On SW3, enable the Eth-Trunk member interface added to the M-LAG to be set to Down so that SW3's Eth-Trunk member interface added to the M-LAG sends dying packets to SW1 and SW2. SW1 and SW2 switch the traffic destined for SW3 to SW4 after receiving the dying packet.
- After the preceding operations are complete, upgrade SW4. After SW3 is upgraded, perform the following configurations to switch service traffic back to SW3.

Contents

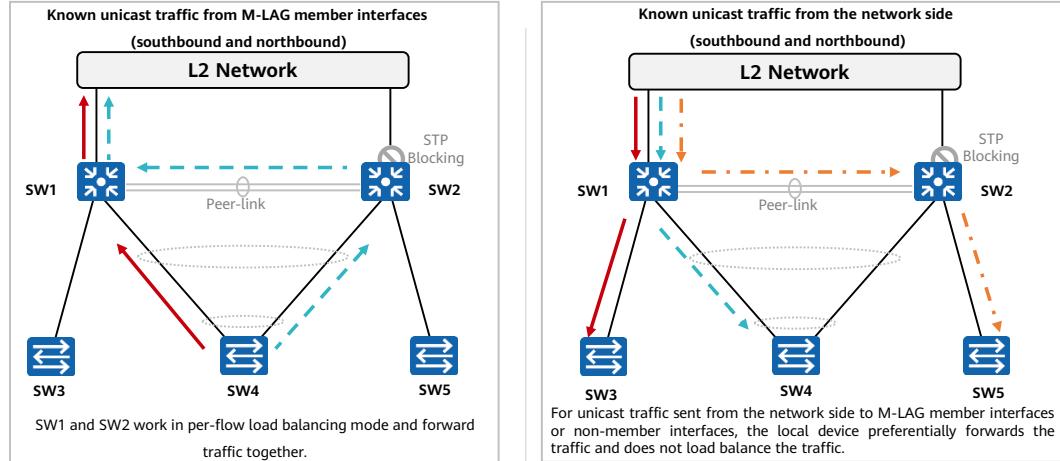
1. Overview of M-LAG
2. **M-LAG Fundamentals**
 - Basic Concepts of M-LAG
 - Basic Features of M-LAG
 - **M-LAG Traffic Forwarding Process**
3. M-LAG Failure Protection
4. M-LAG Deployment
5. M-LAG Best Practices

Known Unicast Traffic Forwarding: Connecting to Layer 2 Network (1)

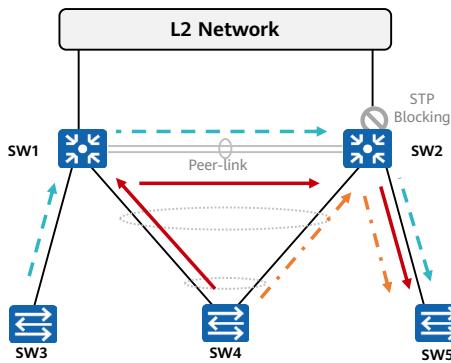


- For north-south Layer 2 traffic, M-LAG member devices forward network-side traffic based on the MAC address table. Due to the STP blocking interface, some traffic will be forwarded through the peer-link interface to the normal member devices.
- For east-west Layer 2 traffic, M-LAGs are configured for all devices and no isolated ports are available. Local Layer 2 traffic is preferentially forwarded through the M-LAG.

Known Unicast Traffic Forwarding: Connecting to Layer 2 Network (2)



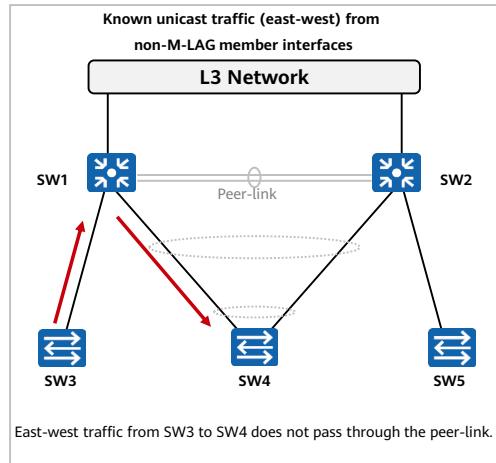
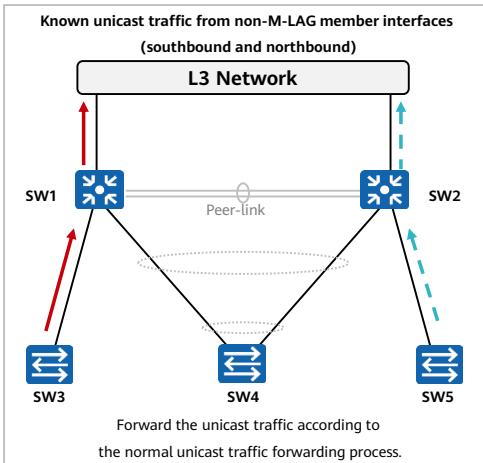
Known Unicast Traffic Forwarding: Connecting to Layer 2 Network (3)



Special Scenario

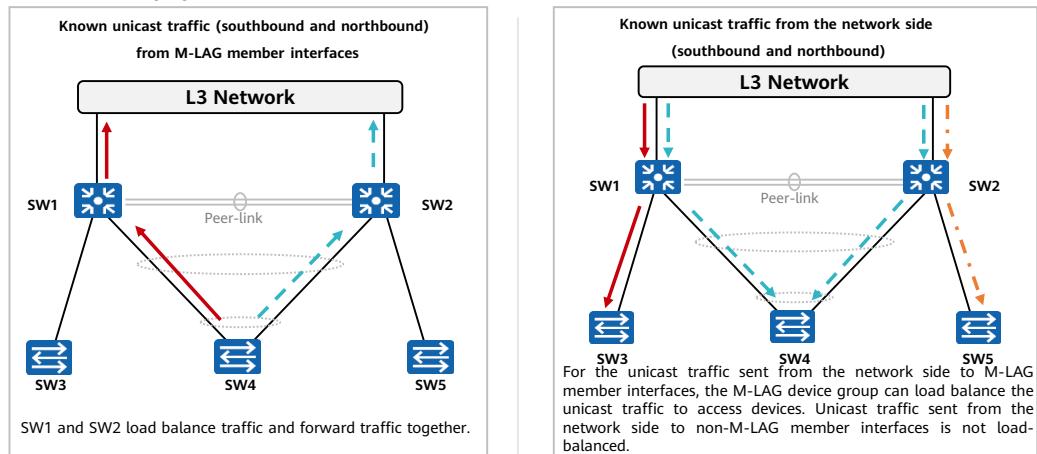
- Traffic from SW3 to SW5 will reach SW5 via SW1 and SW2.
- Traffic from SW4 to SW5:
 - The packets passing through SW1 are sent to SW2 through the peer-link.
 - The packets destined for SW2 will be forwarded directly to the destination through SW2.

Known Unicast Traffic Forwarding: Connecting to Layer 3 Network (1)



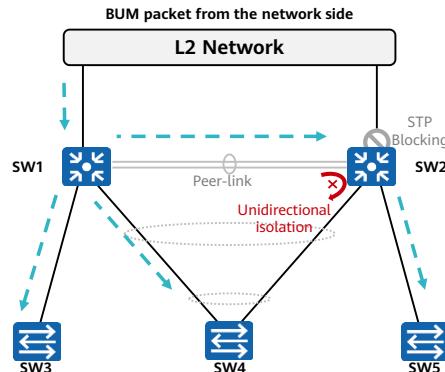
- For north-south Layer 3 traffic, M-LAG member devices preferentially forward the received network-side traffic locally based on the routing table to implement load balancing.
- For east-west Layer 3 traffic, M-LAG member devices preferentially forward local traffic.

Known Unicast Traffic Forwarding: Connecting to Layer 3 Network (2)



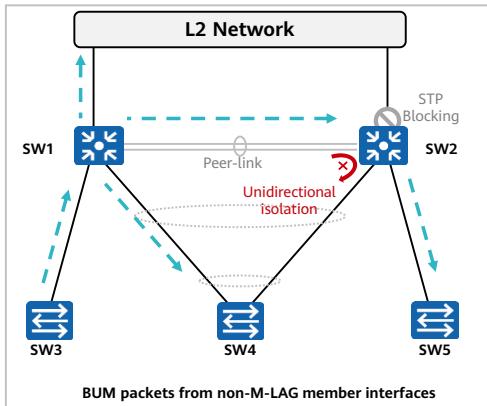
BUM Traffic Forwarding (1)

- SW1 floods the received traffic. When the traffic reaches SW2, SW2 does not forward the traffic to SW4 because the peer-link and M-LAG member interfaces are isolated unidirectionally.

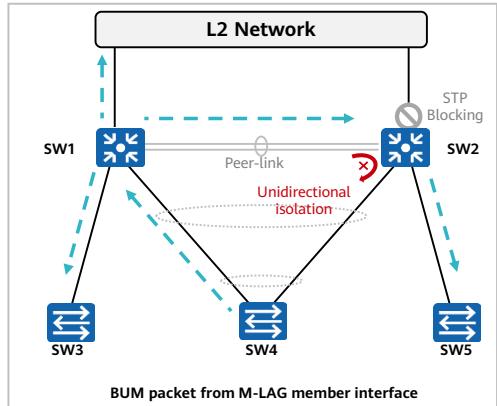


- BUM packets refer to broadcast, unknown unicast, and multicast packets. The Layer 2 forwarding process floods these packets.
- Packets received from a common port (dual-homing source port or single-homing source port) are flooded to the local port and to the peer-link.
- Packets received from the peer-link are flooded only to the single-homing interface. The unidirectional isolation technology is used to prevent the packets from being flooded to the dual-homing destination.

BUM Traffic Forwarding (2)



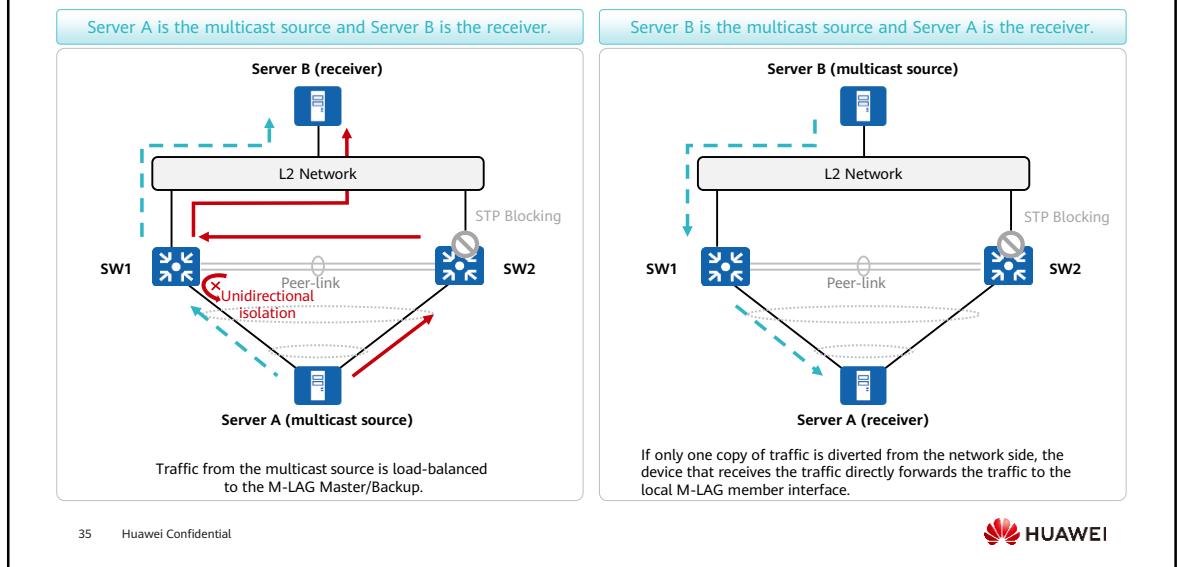
SW1 floods the received traffic. When the traffic reaches SW2, SW2 does not forward the traffic to SW4 because the peer-link and M-LAG member interfaces are isolated unidirectionally.



SW1 floods the received traffic. When the traffic reaches SW2, SW2 does not forward the traffic to SW4 because the peer-link and M-LAG member interfaces are isolated unidirectionally.

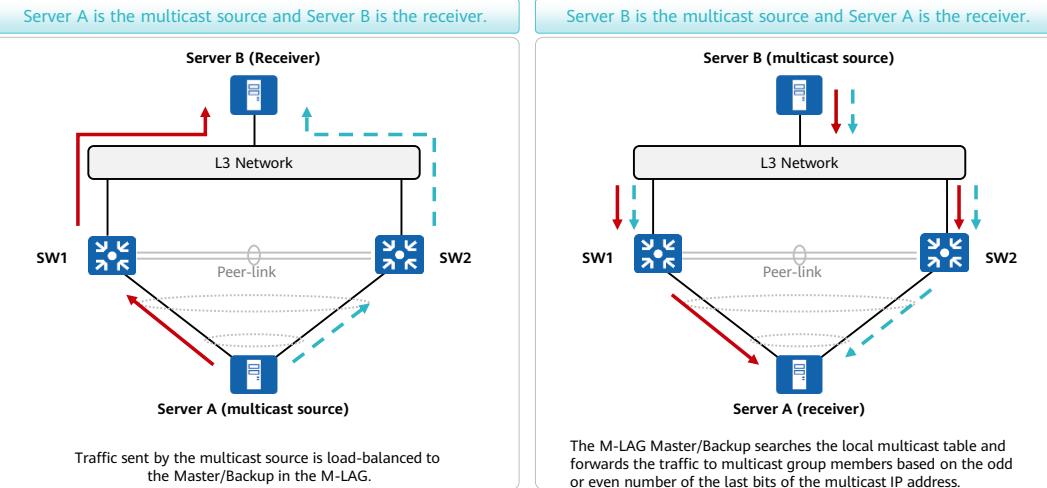
- The figure on the right shows only the packets sent from SW4 to SW1 and the packets sent from SW4 to SW2. The M-LAG member ports on SW1 are unidirectionally isolated.

Multicast: M-LAG Connecting to a Layer 2 network.



- If an M-LAG is connected to a Layer 2 network, the Layer 2 network must send only one copy of traffic to the M-LAG. Otherwise, loops may occur. As shown in the figure, assume that the M-LAG upstream interface on the right is blocked by STP.
- When Server A functions as the multicast source and Server B functions as the multicast group member, the traffic of the multicast source is sent to the M-LAG Master/Backup through load balancing. Because the upstream interface on the right M-LAG Master/Backup is blocked, the multicast outbound interface on the right device points to the peer-link.
- When Server B functions as the multicast source and Server A functions as the multicast group member, M-LAG Master/Backup can forward multicast traffic. When only one copy of traffic is diverted from the network side, the device that receives the traffic directly forwards the multicast traffic to the local M-LAG member interface.

Multicast: M-LAG Connecting to a Layer 3 network.



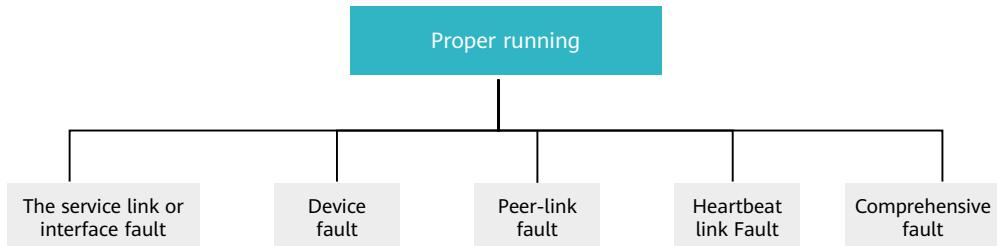
- When ServerAfunctions as a multicast source and ServerBfunctions as a multicast group member, traffic sent by the multicast source is load balanced to M-LAG master and backup devices. After receiving the traffic, M-LAG master and backup devices query the local multicast forwarding table and forward the traffic.
- When ServerBfunctions as a multicast source and ServerAfunctions as a multicast group member, both M-LAG master and backup devices divert traffic from the multicast source, query the local multicast forwarding table, and load balance the traffic to the multicast group member based on the following rules:
 - If the last digit of the multicast group address is an odd number (for example, 225.1.1.1, FF1E::1, or FF1E::B), the M-LAG device where the master M-LAG member interface resides forwards the traffic to the multicast group member.
 - If the last digit of the multicast group address is an even number (for example, 225.1.1.2, FF1E::2, or FF1E::A), the M-LAG device where the backup M-LAG member interface resides forwards the traffic to the multicast group member.

Contents

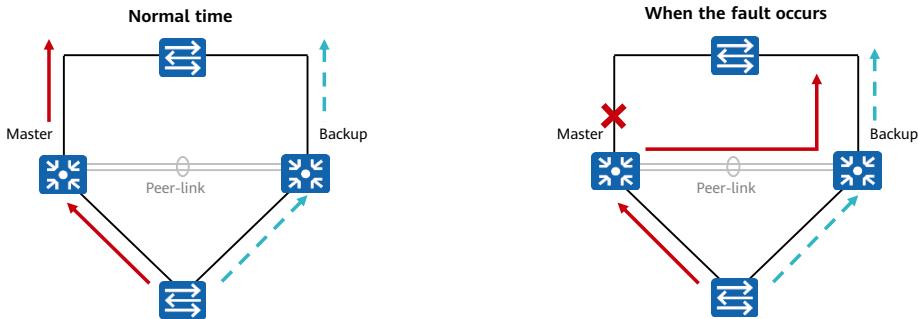
1. Overview of M-LAG
2. M-LAG Fundamentals
- 3. M-LAG Failure Protection**
4. M-LAG Deployment
5. M-LAG Best Practices

Introduction to M-LAG Failure Protection

- As an inter-device link aggregation technology, M-LAG improves link reliability from the card level to the device level. If a fault (link, device, or peer-link fault) occurs, M-LAG uses the fault handling mechanism to ensure that normal services are not affected.



M-LAG Service Link Fault - Uplink

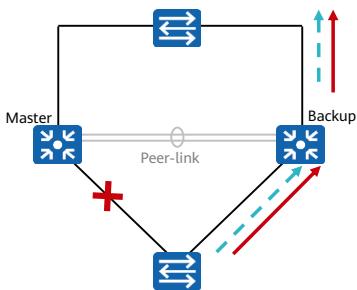


- Uplink faults do not affect DAD on the M-LAG Master/Backup or the active-active system.
- In an M-LAG-to-Ethernet scenario, if the uplink of the M-LAG master device fails, all traffic passing through the M-LAG master device is forwarded through the peer-link, as shown in the right figure.
- If the M-LAG connecting a Layer 3 network and the uplink is faulty, the route is unavailable. In this case, you need to configure best-effort path forwarding or configure Monitor-Link (which will be shown later) to disable the downlink interface when the uplink fails.

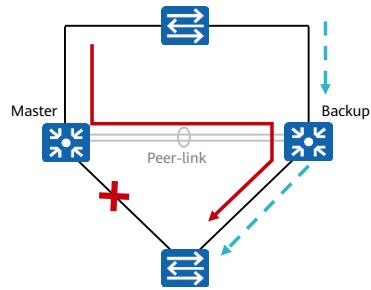
- When an M-LAG is connected to a common Ethernet network and the uplink of the M-LAG master device fails, traffic passing through the M-LAG master device is forwarded through the peer-link. (STP performs network convergence, and the blocked interface may be enabled.)
- If the DAD link is on a service network and the faulty uplink is the DAD link, the M-LAG works properly without being affected. If the peer-link also fails, DAD cannot be performed and packet loss occurs.

M-LAG Service Link Fault - Downlink

Uplink traffic when a fault occurs



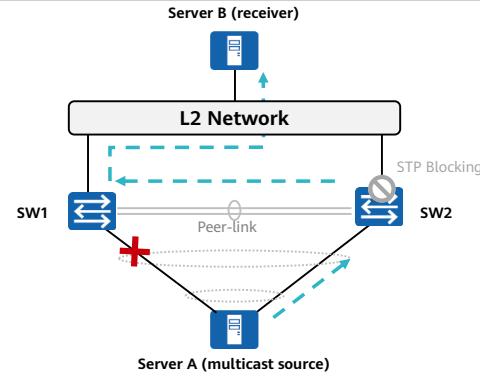
Downlink traffic when a fault occurs



- If a downstream M-LAG member interface fails, the DFS group Master/Backup status does not change. If the faulty M-LAG member interface is in the master state, the slave M-LAG member interface becomes the master. The MAC address of the faulty M-LAG member interface points to the peer-link interface.
- The unidirectional isolation mechanism between the peer-link and M-LAG member interfaces is enabled when the M-LAG master member interface fails to prevent traffic forwarding failure.
- After the faulty M-LAG member interface recovers, the status of the M-LAG member interface remains unchanged, and the M-LAG member interface that becomes the master remains the master.

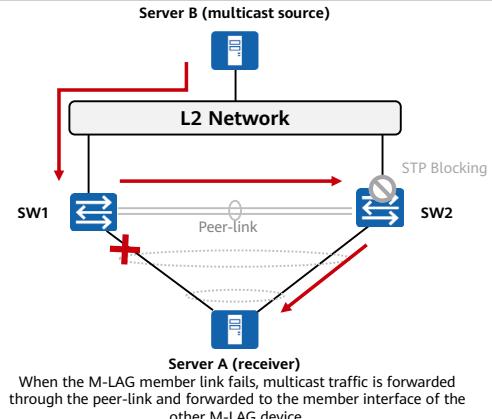
M-LAG Service Link Fault - Multicast: M-LAG Connecting to a Layer 2 Network

Server A is the multicast source and Server B is the receiver.



If an M-LAG member link fails, multicast services are not affected.

Server B is the multicast source and Server A is the receiver.

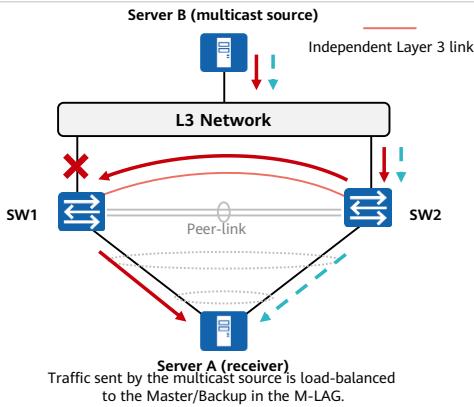


When the M-LAG member link fails, multicast traffic is forwarded through the peer-link and forwarded to the member interface of the other M-LAG device.

- A multicast traffic forwarding scenario is special because M-LAG master and backup devices load balance traffic depending on whether the last digit of the multicast group address is an odd or even number and an independent Layer 3 link is required between the M-LAG devices to forward Layer 3 packets (described on the next slide).
- As shown in the figure on the right, if the local M-LAG member interface fails, multicast traffic is forwarded to the member interface of the other M-LAG device through the peer-link.
- Assume that a multicast source is at the network side and a multicast group member is at the access side. If the M-LAG member interface on the M-LAG master device fails, the master device instructs the remote device to update multicast entries through M-LAG synchronization packets. M-LAG master and backup devices no longer load balance traffic depending on whether the last digit of the multicast group address is an odd or even number, and all multicast traffic is forwarded by the M-LAG backup device on which the M-LAG member interface is Up. If the M-LAG member interface on the M-LAG backup device fails, multicast traffic is forwarded in a similar manner.

M-LAG Service Link Fault - Multicast: M-LAG Connecting to a Layer 3 Network

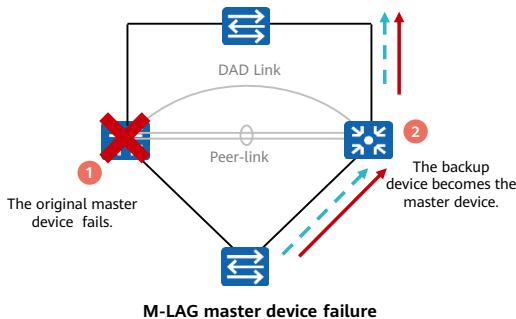
Server A is the receiver and Server B is the multicast source.



- To forward multicast traffic on a Layer 3 network, an independent Layer 3 link must be configured between two M-LAG devices.

- In the case of a fault, there may be only one uplink on the network side. In this case, an independent Layer 3 link is deployed between the M-LAG Master/Backup to transmit multicast packets.
- Multicast packets with the last bit of a multicast address being an odd number cannot be forwarded to the M-LAG master device (SW1 in this example) through the peer-link. Instead, the packets can be forwarded to the M-LAG master device only through an independent Layer 3 link.
- Similarly, if the backup device in the M-LAG system fails, the multicast packet whose last bit of the multicast address is an even number may also be forwarded to the master device by using the independent layer 3 link.

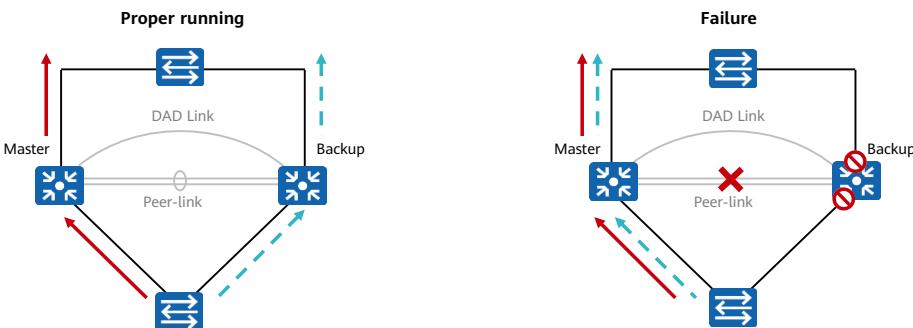
M-LAG Device Fault



- If the M-LAG Master Device fails:
 - The backup device becomes the master device and continues to forward traffic.
 - The Eth-Trunk on the M-LAG master device goes Down.
- If the M-LAG backup device is fails:
 - The Master/Backup status of the M-LAG does not change, and the Eth-Trunk on the M-LAG backup device goes Down.
 - The Eth-Trunk link on the M-LAG master device remains Up, and the traffic forwarding status remains unchanged.

M-LAG Peer-Link Fault

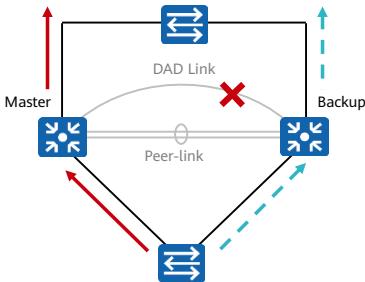
- If the peer-link fails but the DAD heartbeat status is normal, all interfaces on the backup M-LAG device, except the logical interface, management interface, and peer-link interface, enter the Error-Down state.



- You can run a command to configure logical interfaces on the M-LAG backup device to enter the Error-Down state if the peer-link fails but the DAD heartbeat status remains normal.
 - If the peer-link fails but the DAD heartbeat status is normal when M-LAG is used for dual-homing access on a VXLAN or IP network, the VLANIF interface, VBDIF interface, loopback interface, and M-LAG member interface on the M-LAG backup device enter the Error-Down state.
- After logical interfaces are configured to change to Error-Down state when the peer-link fails but the DAD heartbeat status is normal in an M-LAG, if a faulty peer-link interface in the M-LAG recovers, the devices restore VLANIF interfaces, VBDIF interfaces, and loopback interfaces to Up state 6 seconds after DFS group pairing succeeds to ensure that ARP entry synchronization on a large number of VLANIF interfaces is normal. If a delay after which the Layer 3 protocol status of the interface changes to Up is configured, the delay after which VLANIF interfaces, VBDIF interfaces, and loopback interfaces go Up is the configured delay plus 6 seconds.
- When the faulty peer-link recovers, the M-LAG member interface in the Error-Down state automatically restores to the Up state after 240s by default, and the other interfaces in the Error-Down state automatically restore to the Up state immediately.
- When the peer-link recovers, the M-LAG interface in the Error Down state automatically goes Up after 240 seconds by default, and the other M-LAG interfaces in the Error Down state immediately go Up.

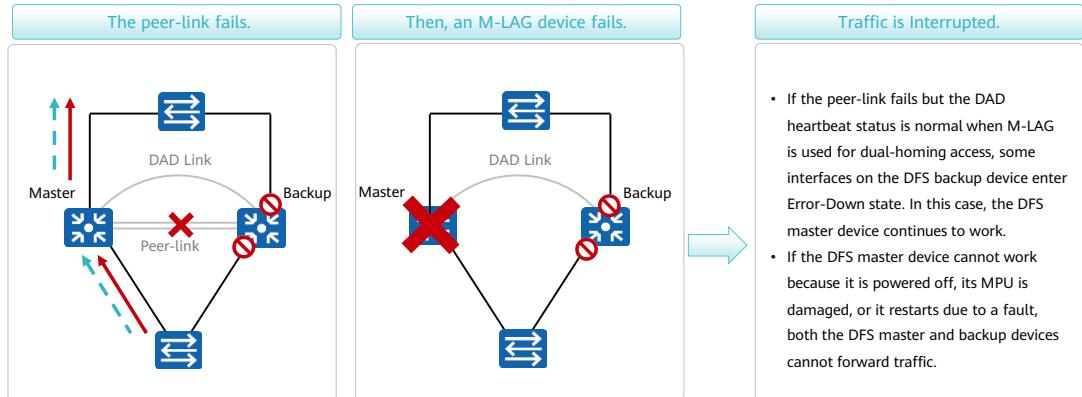
M-LAG Heartbeat Link Fault

- If the heartbeat link fails:
 - Services are not affected.
 - The dual-active peer-link fault cannot be identified during peer-link troubleshooting.
- Therefore, If the heartbeat link fails:
 - The failsafe mechanism is not triggered.
 - However, an alarm will be generated. You need to handle the alarm in a timely manner to prevent service abnormalities if the entire peer-link fault occurs.

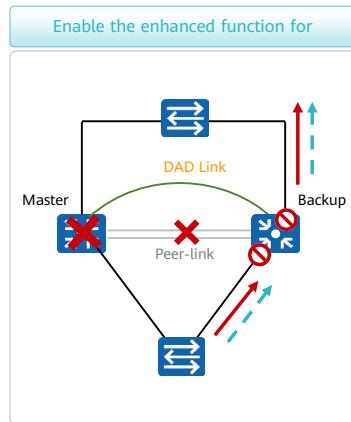


- After the heartbeat link fault is rectified, a heartbeat fault clear alarm is generated.

Peer-Link Fault + M-LAG Device Fault (Problem)



Peer-Link Fault + M-LAG Device Fault (Solution)



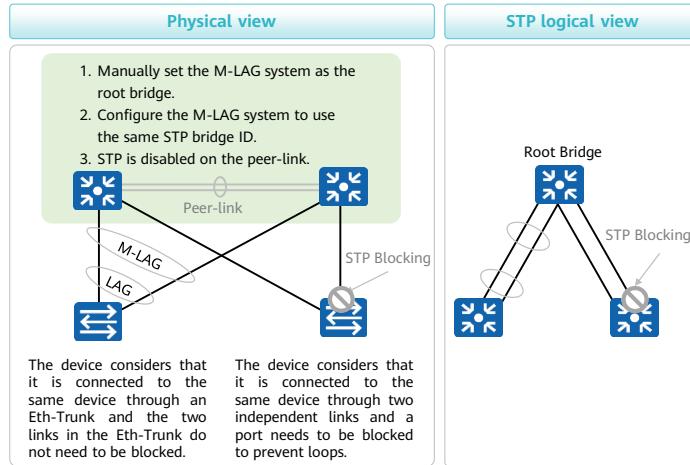
- Enable the enhanced secondary fault function:** If the enhanced secondary fault function has been enabled in the M-LAG, the backup device detects the fault of the DFS master device by using the dual-active detection (DAD) mechanism. (No M-LAG DAD heartbeat packet is received within a certain period.) After, the becomes the DFS master device and the interface that is in the ERROR DOWN state on the device goes Up and continues to forward traffic.

- Device fault rectification: If the fault on the original DFS master device is rectified but the peer-link fault persists, the following applies:
 - If the LACP M-LAG system ID is switched to the LACP system ID of the local device within a certain period, the access device selects only one of the uplinks as the active link during LACP negotiation. The actual traffic forwarding is normal.
 - If the default LACP M-LAG system ID is used, that is, it remains unchanged, two M-LAG devices use the same system ID to negotiate with the access device. Therefore, links to both devices can be selected as the active link. In this scenario, because the peer-link fault persists, M-LAG devices cannot synchronize information such as the priority and system MAC address of each other. As a result, two M-LAG master devices exist, and multicast traffic forwarding may be abnormal. In this case, the HB DFS master/backup status is negotiated through heartbeat packets carrying necessary information for DFS group master/backup negotiation (such as the DFS group priority and system MAC address). Some interfaces (for details, see Peer-LinkFault) on the HB DFS backup device are triggered to enter Error-Down state. The HB DFS master device continues to work.

Contents

1. Overview of M-LAG
2. M-LAG Fundamentals
3. M-LAG Failure Protection
- 4. M-LAG Deployment**
 - M-LAG Multi-Protocol Deployment
 - M-LAG Deployment Scenario
5. M-LAG Best Practices

Multi-Protocol Deployment - STP Solution 1: Root Bridge Solution



Advantages of the root bridge solution:

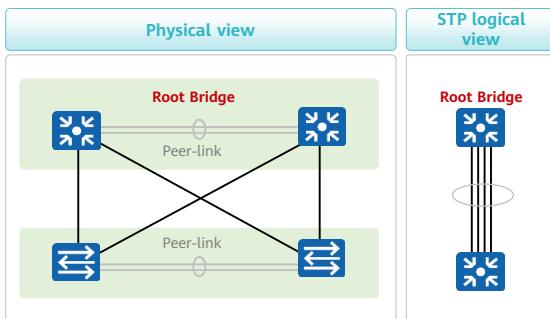
- Simple implementation: The STP protocol implementation does not need to be modified. Only the STP bridge ID parameter needs to be set.

Constraints on the root bridge solution:

- An M-LAG system can only serve as the STP root bridge, but cannot serve as a non-root bridge.

- Configuration suggestion: When configuring M-LAG based on the root bridge, set the bridge IDs of the two devices in the M-LAG to the same and set the root priority to the highest. This ensures that the two devices in the M-LAG are the STP root bridges.

Multi-Protocol Deployment - STP Solution 2: V-STP Solution



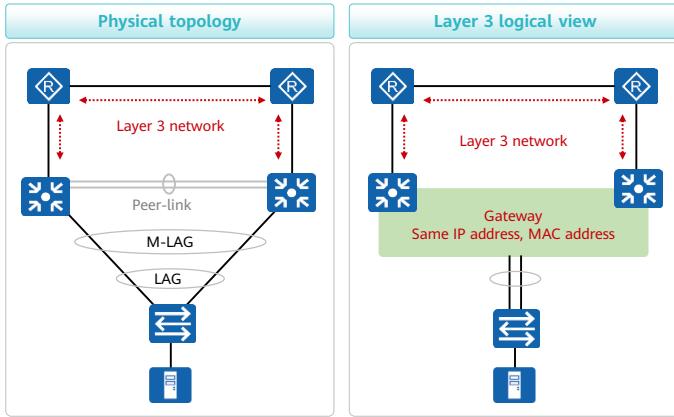
- After the V-STP mode is enabled on the M-LAG Master/Backup, the two devices are virtualized into one device using V-STP to calculate the port role and fast convergence once the M-LAG master/backup negotiation is successful.
- After the V-STP mode is enabled, the M-LAG backup device needs to synchronize the bridge MAC address and instance priority information of the M-LAG master device.
- The M-LAG backup device uses the bridge MAC address and instance priority information synchronized from the M-LAG master device to perform STP calculation and send and receive BPDU. This ensures that the STP calculation parameters are consistent after the M-LAG master device is virtualized into one device.

Multi-Protocol Deployment - STP Solution Comparison

- The root bridge mode and V-stp mode can be used to build a loop-free network. In root bridge mode, M-LAG devices must be manually specified as the same bridge. In V-stp mode, protocol information between M-LAG devices must be synchronized and displayed as one device for STP negotiation.

Mode	Configuration Method	Application Scenario	Application Limitations
Root Bridge Mode	Manually configure the two M-LAG devices as the root bridge and configure the same bridge ID to simulate the two devices as the same root bridge.	This mode applies to the deployment of a single-level M-LAG or the deployment of a multi-level M-LAG at the aggregation layer as the root bridge of a Layer 2 network.	<ul style="list-style-type: none">This mode supports only STP, RSTP, and MSTP.M-LAG cascading in all-root bridge mode is not supported.STP must be disabled on the peer-link interface.
V-stp mode	After the V-stp mode is enabled, the STP protocol status is synchronized between dual-homing devices so that the two devices use the same status for STP negotiation.	Applies to interconnection with traditional STP networks. This mode applies to multi-level M-LAG deployment.	<ul style="list-style-type: none">Only STP and RSTP are supported in this mode.The M-LAG member interface configurations on M-LAG master and backup devices must be the same.

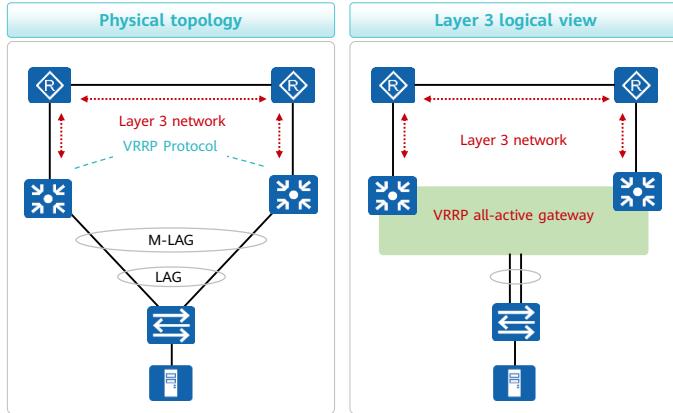
Multi-Protocol Deployment - Dual-Active Gateway (1)



- Scheme 1 (simulate the same gateway): Configure the same gateway IP address and MAC address for the two devices.

Solution 1 is preferred. This solution is easy to configure and reduces protocol costs. The Layer 3 gateways on the two M-LAG devices are **independent**. You can configure the same IP address and MAC address for the Layer 3 gateways on the two devices so that they function as dual-active gateways.

Multi-Protocol Deployment - Dual-Active Gateway (2)

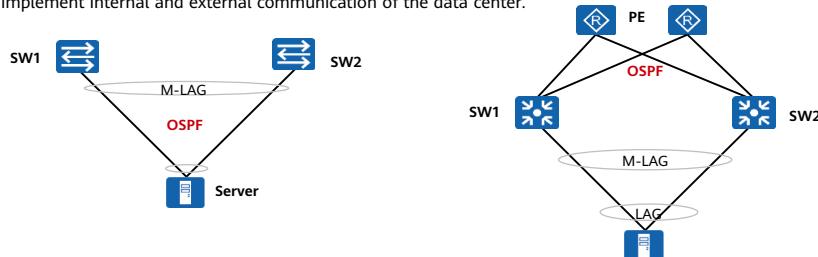


- **Solution 2 (virtual route redundancy protocol):** Configure VRRP on the two devices. An M-LAG supports VRRP dual-active mode.
 - Create a VRRP group on VLANIF or VBDIF interfaces and configure the same virtual IP and MAC addresses for them so that the M-LAG master and backup devices in the VRRP group function as dual-active gateways.

- If VRRP is deployed, VRRP information needs to be synchronized between the master and backup devices through the peer-link so that the virtual interfaces (VLANIF or VBDIF interfaces) of the master and backup devices have the same virtual IP address and virtual MAC address.
- M-LAG and VRRP are usually configured together in Data Center Interconnect (DCI) scenarios.

Multi-Protocol Deployment - Dynamic Routing Protocols such as OSPF

- M-LAG devices can function as access devices to connect to servers or as egress devices to connect to egress routers (PEs).
 - An M-LAG can be configured with a static route to the network segment where a server resides or use OSPF to dynamically exchange routing information with the server.
 - An M-LAG can function as a border leaf node and communicate with PE routers through OSPF to exchange routing information and implement internal and external communication of the data center.



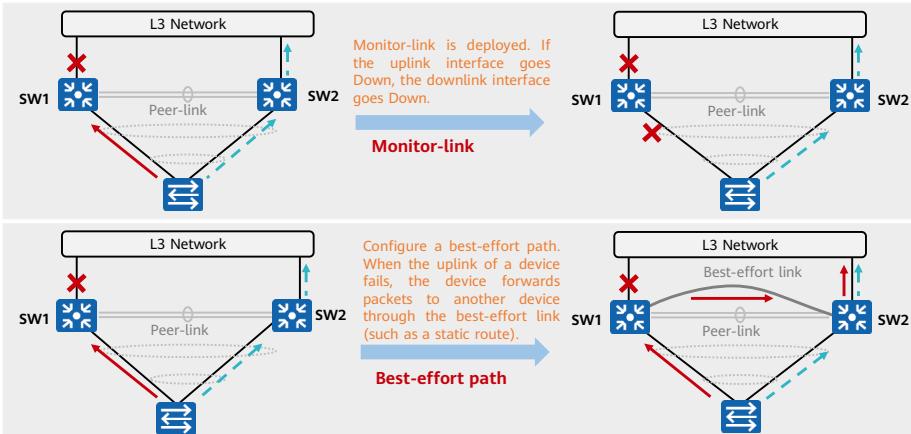
55 Huawei Confidential

 HUAWEI

- The server is dual-homed to the M-LAG and has static routes configured so that it can communicate with the M-LAG through Layer 3 routes. However, the network using static routes is difficult to configure and maintain and is lack of flexible and fast deployment capabilities, thereby cannot meet the requirements of rapidly growing services.
- To address this problem, M-LAG member devices need to establish neighbor relationships of dynamic routing protocols with the user-side device. Therefore, M-LAG member interfaces need to support dynamic routing protocols.
- Before configuring OSPF over M-LAG, you need to complete the following tasks:
 - Establish an M-LAG and an OSPF network.
 - Add M-LAG member interfaces to the corresponding VLAN.
 - Enable OSPF on the user-side device.

Multi-Protocol Deployment - Monitor Link or Best-effort Path

- When an M-LAG accesses a Layer 3 network, if the uplink of a device fails, packets cannot be forwarded through the peer-link. As a result, packets sent to the device are discarded.



56 Huawei Confidential

HUAWEI

- In a Layer 3 scenario, a bypass link must be configured between M-LAG master and backup devices. Otherwise, the uplink traffic that reaches the master device cannot reach the backup device through the peer-link.
- When Monitor Link is configured, if the downlink or M-LAG member interface of the other device fails, all traffic is discarded. Therefore, Monitor Link is not applicable to the scenario where the M-LAG functions as the egress gateway.

Contents

1. Overview of M-LAG
2. M-LAG Fundamentals
3. M-LAG Failure Protection
- 4. M-LAG Deployment**
 - M-LAG Multi-Protocol Deployment
 - **M-LAG Deployment Scenario**
5. M-LAG Best Practices

Overview of the M-LAG Deployment Solution

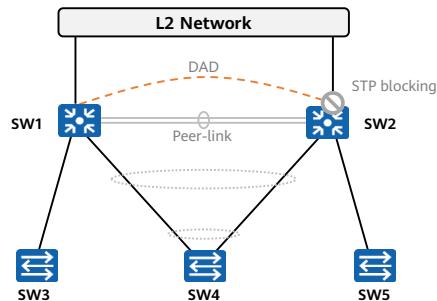
- M-LAG deployment modes are as follows:

M-LAG access network type	M-LAG access device type	M-LAG access mode	M-LAG deployment mode
Connecting to a Layer 2 network	Switch access	Single-homing access	Single-level M-LAG
Connecting to a Layer 3 network	Server access	Dual-homing access	Multi-level M-LAG
Connecting to a tunnel network	VAS device access		

- Single-level M-LAG deployment is the most common deployment. The preceding sections use the deployment as an example and are not described here.

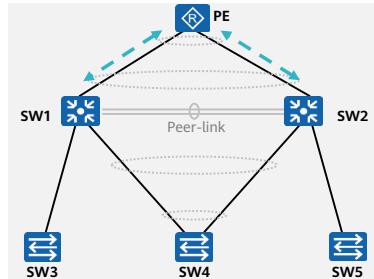
M-LAG Access Network Type - Connecting to a Layer 2 Network

- An M-LAG can connect to a Layer 2 network, such as an Ethernet network. Pay attention to the following point: To prevent loops, a link may be blocked by STP, and packets may need to be forwarded through the peer-link.



M-LAG Access Network Type - Connecting to a Layer 3 Network

- An M-LAG system can access a Layer 3 network. Note the following points:
 - The M-LAG system functions as the gateway of the access-side device. To function as a logical device, the M-LAG system must be deployed with active-active gateways.
 - If the ping test is performed between a device in the M-LAG system and a PE, packet loss may occur due to load balancing between the PE and the PE. (This is normal and does not affect services.)

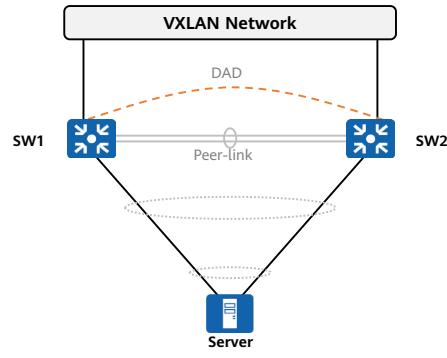


M-LAG Access Network Type - Connecting to a Tunnel Network

- A large Layer 2 network needs to be deployed in a data center. M-LAG can connect to the VXLAN network.

Note the following points:

- Configure VXLAN on the two devices in the M-LAG. The two devices function as tunnel endpoints and must be configured with the same tunnel endpoint IP address and function as dual-active gateways. When an underlay routing protocol (such as OSPF) is configured, the two devices have different router IDs.
- The M-LAG is considered as a logical device (one tunnel endpoint) to the remote device. Traffic sent to the M-LAG is load balanced to the M-LAG member devices.
- Configure Layer 2 sub-interfaces on the M-LAG interfaces of leaf nodes to transmit traffic.

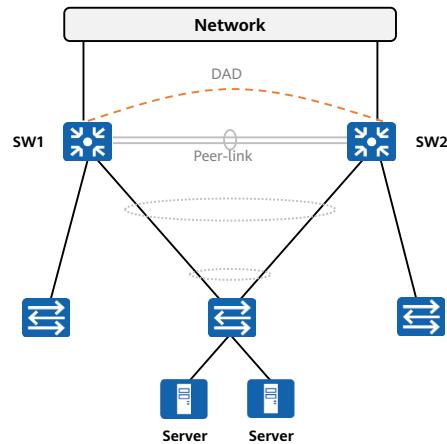


- Configuration suggestions:

- Configure M-LAG on leaf switches to support dual-active access of server NICs.
- Configuring EVPN: Configure the same VTEP on two leaf nodes.
- Configuring Layer 2 functions: Create a BD and specify a VNI.
- Create a Layer 2 sub-interface on the M-LAG interface of a leaf node and associate the Layer 2 sub-interface with the BD.
- Configuring Layer 3 functions: Create BDIF interfaces and configure IP and MAC addresses for the BDIF interfaces.
- Note: For distributed VXLAN gateways, BDIF interfaces on the same network segment must be configured with the same IP address and MAC address to support VM migration.

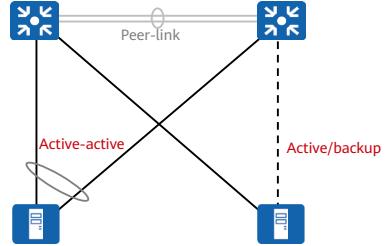
M-LAG Access Device Type - Switch Access

- A switch can function as the access device of the M-LAG system. Generally, the switch is not a data source but a Layer 2 transparent transmission device. In this case, note the following:
 - When a switch is dual-homed to a switch, only link aggregation can be configured to implement load balancing. The hash calculation result determines the device in the M-LAG system to which the switch sends packets.
 - If the access switch is not connected to other Layer 2 networks, STP does not need to be configured to prevent loops. The unidirectional isolation mechanism prevents loops between the access switch and the M-LAG system.



M-LAG Access Device Type - Server Access

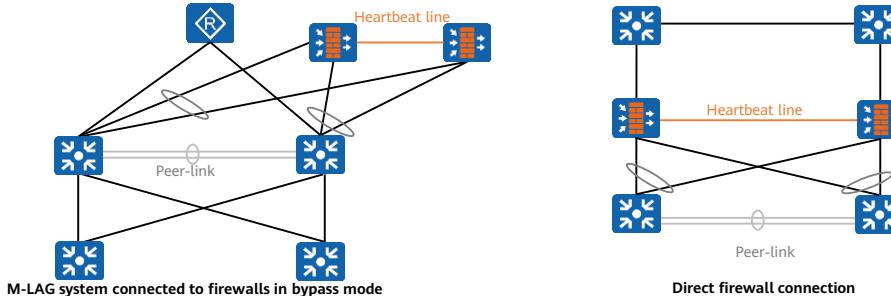
- The M-LAG system supports dual network ports of a server in active-active bond and active-backup bond mode.
- If two network ports on a server are connected in active-active mode:
 - LACP is recommended to provide higher reliability and failover performance.
 - When some servers are deployed in PXE mode, access switches must support dynamic LACP. When the server goes online, no configuration is configured. In this case, LACP negotiation fails and the Eth-Trunk goes Down. However, member interfaces can independently forward Layer 2 data so that the server can obtain the configuration file. After obtaining the configuration, the server negotiates aggregation parameters with the access switch through LACP.



- Note: The Linux operating system is used as an example. The operating system supports seven bonding modes.
 - 0. round robin and 4.lACP support load balancing between two network ports. They are two common dual-network-port active-active access solutions. Link binding must be configured on the peer switch.
 - 0. round robin: Data packets are sent to each interface in polling mode to implement load balancing.
 - 4.lACP: indicates that LACP is used to negotiate the working mode, load balancing, and redundancy of bound interfaces.
 - "1.active-backup": indicates the common active-standby mode. Link aggregation does not need to be configured on interfaces of the remote switch.

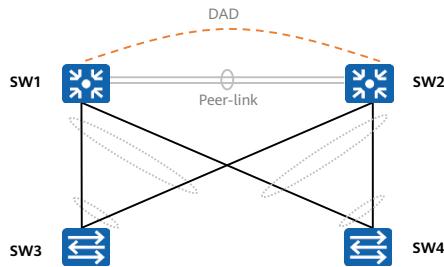
M-LAG Access Device Type - VAS Device Access (Firewall as an Example)

- VAS devices can be connected to the M-LAG system in bypass or direct connection mode.
 - In bypass mode, traffic can pass through the firewall based on the route, avoiding traffic bottlenecks on the firewall. In addition, the bypass mode is more conducive to network expansion.
 - In the direct connection mode, all traffic needs to pass through the firewall.



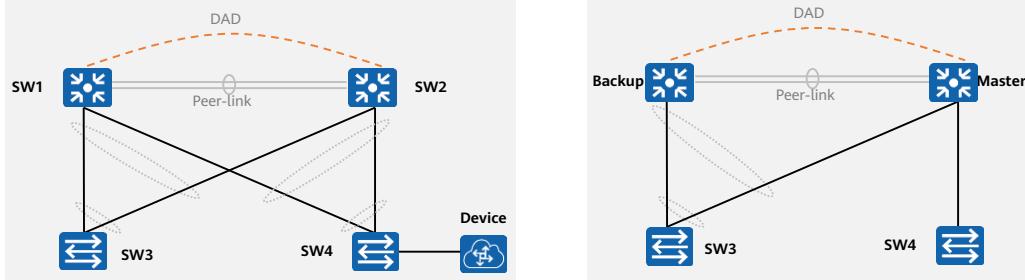
M-LAG Access Mode - Dual-Homing to an M-LAG

- Generally, access devices are dual-homed to an M-LAG, which is recommended and most commonly used.
- The networking where access devices are dual-homed to an M-LAG through link aggregation has the following advantages:
 - If the peer-link fails, fast convergence can be performed. In dual-active scenarios, traffic forwarding behaviors are consistent.
 - Dual-active redundant forwarding paths are provided, improving link reliability.



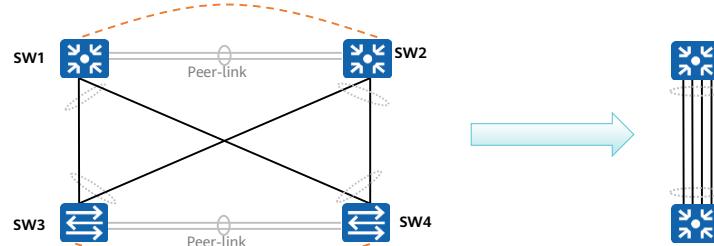
M-LAG Access Mode - Single-Homing to an M-LAG

- If a device cannot be dual-homed to an M-LAG, preferentially connect the device to another device that has been dual-homed to the M-LAG.
- If a device cannot be connected to another device that has been dual-homed to the M-LAG, you can connect the device to the M-LAG master device to prevent the device from being isolated upon failure of the peer-link. (If the peer-link fails, all interfaces except the stack interface, management interface, and peer-link interface on the backup device enter Error-Down state.) In addition, you are advised to use the VLAN that is not used by M-LAG member interfaces.



M-LAG Deployment Mode - Multi-Level M-LAG

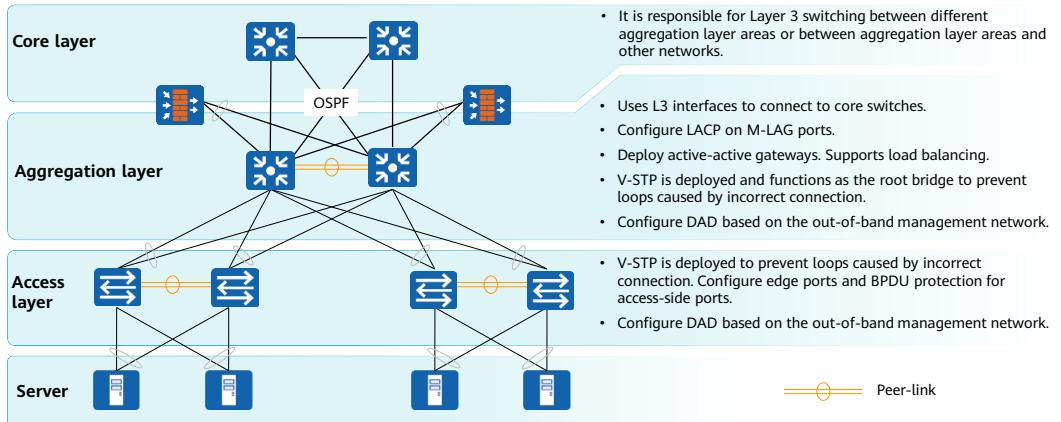
- Multi-level M-LAG interconnection is mainly used in large-scale data centers to build large Layer 2 networks. It not only simplifies networking, but also increases the number of dual-homing access servers while ensuring reliability.
- During the configuration of two-level M-LAG and in various fault scenarios, ensure that no loop occurs.



Contents

1. Overview of M-LAG
2. M-LAG Fundamentals
3. M-LAG Failure Protection
4. M-LAG Deployment
- 5. M-LAG Best Practices**

M-LAG Deployment Best Practice



- It is recommended that a dedicated L3 best-effort link be configured between M-LAG devices to meet the scenario where all upstream ports on a single member device fail.

Quiz

1. (Multiple-answer question) In an M-LAG, which of the following entries are synchronized between two devices? ()
 - A. MAC address entry
 - B. ARP entry
 - C. Routing entry
 - D. ACL entry
2. (Short-answer question) What are the functions of DAD links in an M-LAG? What are the deployment considerations?

1. AB
2. A dual-active detection (DAD) link, also called a heartbeat link, is a Layer 3 interconnection link used to exchange DAD packets between M-LAG master and backup devices. Under normal circumstances, the DAD link does not participate in any traffic forwarding behaviors in the M-LAG. It is only used to detect whether two master devices exist when a fault occurs. The DAD link can be an external link, for example, if the M-LAG is connected to an IP network and the two member devices can communicate through the IP network, the link that enables communication between the member devices can function as the DAD link. An independent link that provides Layer 3 reachability can also be configured as the DAD link, for example, a link between management interfaces of the member devices can function as the DAD link.

Summary

- M-LAG is a mechanism that implements inter device link aggregation. Two access switches in the same state in an M-LAG can perform link aggregation negotiation with a connected device. M-LAG allows two devices to establish a dual-active system, improving link reliability from the card level to the device level.
- This course describes the basic concepts, fundamentals, failure protection principles, and typical applications of M-LAG on data center networks.
- The CloudFabric solution uses M-LAG and VXLAN to implement end-to-end reliability, ensuring that service systems can run properly in device failure and upgrade scenarios.

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。
Bring digital to every person, home, and
organization for a fully connected,
intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product performance, market position, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



Huawei CloudFabric Data Center Network Solution



Foreword

- Nowadays, the cloud-based digital architecture has become the key to digital transformation. ICT infrastructure has undergone profound cloud transformation, and cloud computing has been widely used. As a key ICT infrastructure, data center networks (DCNs) also need to undergo technological transformation based on service requirements in the cloud computing scenario.
- To meet the service requirements and challenges for traditional DCNs in the cloud computing scenario, Huawei launches the CloudFabric Hyper-Converged DCN Solution, which is also called Huawei CloudFabric Solution.
- This course describes the overall architecture, functions, and features of the CloudFabric Solution based on service scenarios and requirements, and further introduces the core components and functions of the solution.

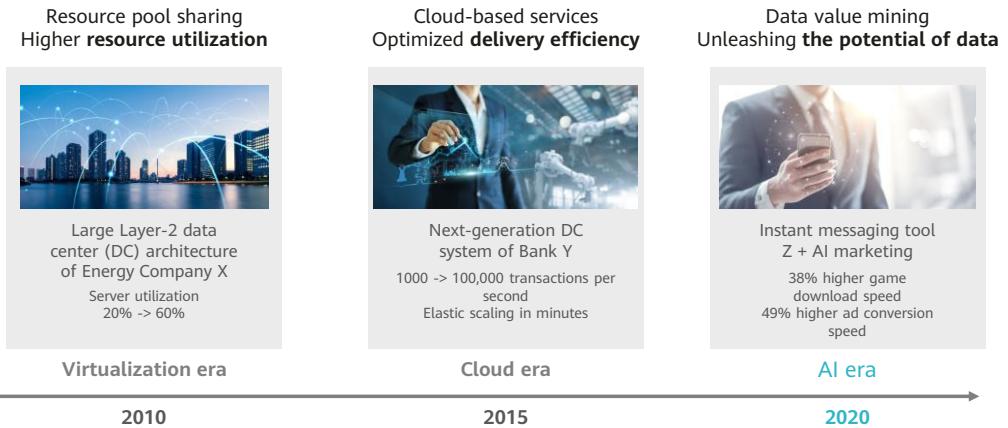
Objectives

- On completion of this course, you will be able to:
 - Describe the development trend and challenges of DCNs.
 - Describe the architecture and core components of the CloudFabric Solution.
 - Describe the application scenarios of the CloudFabric Solution.
 - Describe typical functions and features of the CloudFabric Solution.

Contents

- 1. Development Trends and Challenges of DCNs**
2. Huawei CloudFabric Solution
3. Solution Features

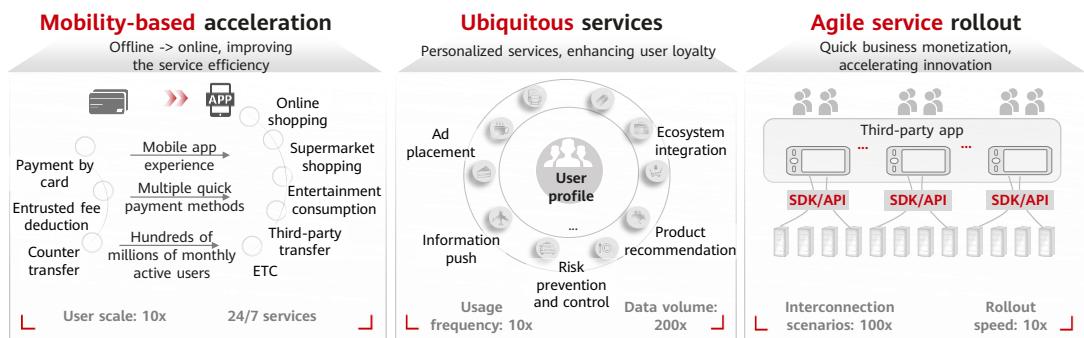
DC Mission: Shift from a Service Center to a Value Center



5 Huawei Confidential



DCNs Are Evolving to Multi-DC, Multi-cloud Networks



Centralized -> Distributed

More complex system architecture



Single-DC -> Multi-DC

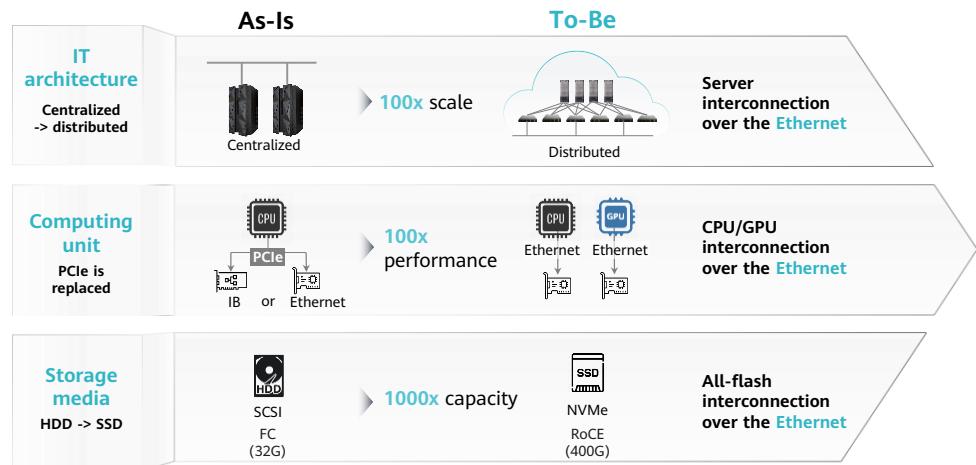
DC scale increased by 100 times



Private cloud -> Hybrid cloud

Virtualization scale increased by 100 times

DCNs Are Evolving Toward All-Ethernet



7 Huawei Confidential

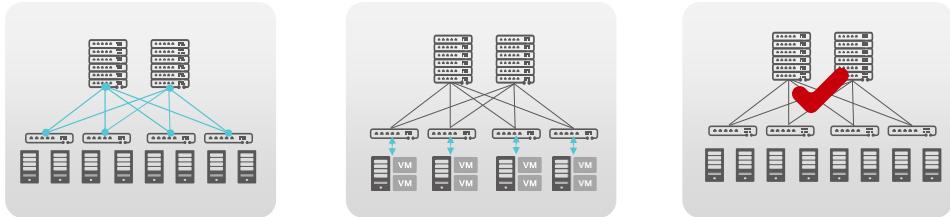
 HUAWEI

- The network connects computing and storage servers to support the IT architecture of the entire DC. This means that the network needs to be adjusted accordingly upon any change of the connected servers or the IT architecture. Such change, however, is commonplace in DCs. Specifically, three major changes of the IT architecture, computing, and storage are driving DCNs to evolve from the original multi-protocol mode to all-Ethernet.
 - The IT architecture has evolved from centralized to distributed, and large-scale node interconnection has become a new norm on the Ethernet.
 - PCIe buses are being removed from computing units, no matter whether they are CPUs or GPUs. This aims to break through the bus speed bottleneck. Instead, Ethernet ports are used to directly provide higher computing power.
 - From the perspective of storage media, HDDs are upgrading toward all-flash, improving storage performance 100-fold. Traditional FC, however, provides only 32G bandwidth, which cannot meet the high throughput requirements of all-flash. In this context, the Ethernet with up to 400G bandwidth becomes the de facto standard for the next-generation storage network.
- Note:
 - PCIe: PCI Express
 - IB: InfiniBand, an Input/Output (I/O) technology
 - HDD: Hard Disk Drive

- SSD: Solid State Disk

Challenge 1: AI-Powered DCs Pose Challenges to Networks

- Rapid construction of ultra-large DCNs
Requiring **fast network construction**
- Unified management of compute resource pooling platforms, such as VMs and containers
Requiring **network linkage and rapid login and logout**
- Frequent service changes bring a large number of network changes
Requiring **intelligent network evaluation and verification**



Full DCN lifecycle
Automatic requirement sorting

Planning and design

Installation and deployment

Service rollout

Service change

Monitoring and maintenance

Fault handling



Challenge 2: The Network Changes and O&M Have Exceeded Human Limits

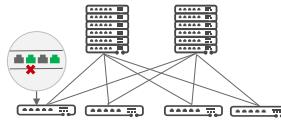
Bank A: The single DC construction volume in 2021 is **greater** than that in 2020.
Deployment and rollout of 30 switches: **3+ person-weeks**
> 3 days for rolling out a service, cross-DC, N work orders and conferences.



Manual network construction and change operations
Manual operations such as solution design, evaluation, and decision-making account for 80% of the entire process.

Long service rollout period

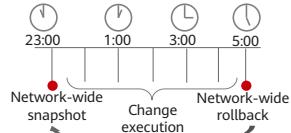
Bank B: **14,000+** changes in a year
The network is interrupted for 40 minutes because a legacy server port is deleted by mistake.



Experience-dependent change, no prevention or detection methods
Nearly 40% of faults are caused by human errors, causing multiple major accidents.

Error-prone configuration change

Abnormal alarms are generated due to unexpected situations caused by changes. Network-wide emergency recovery is the top priority. Ensure that the **network recovery time is less than 30 minutes**.



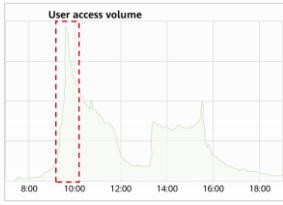
Urgent rollback: abnormal alarms, changes that are not completed within the specified time, customer-defined...
Network exceptions or faults cannot be quickly rectified
Average fault locating time: > 76 minutes
Average critical incident recovery time: > 40 minutes

Slow network recovery

Challenge 3: Difficulties Faced by Traditional O&M

Difficult health check

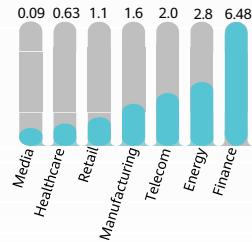
Fluctuating securities market, resulting in the daily needs to cope with service peaks.



It takes **three person-hours** to perform routine inspection before the market opens every day. This increases difficulties in confidently keeping up with the general market trends.

Difficult fault locating

Hundreds of millions of cross-bank transactions per day, requiring 24/7 uninterrupted services.

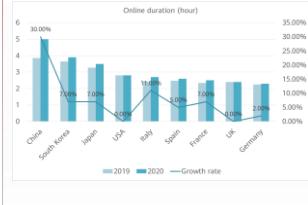


Survey on loss caused by fault-triggered interruptions ①

The complicated architecture results in difficult fault locating. It takes **76 minutes** on average to locate a fault.

Difficult network change

Enormous increase in Internet traffic, requiring network changes every week.



About **70%** of network faults are caused by human errors as changes are manually compared and verified.

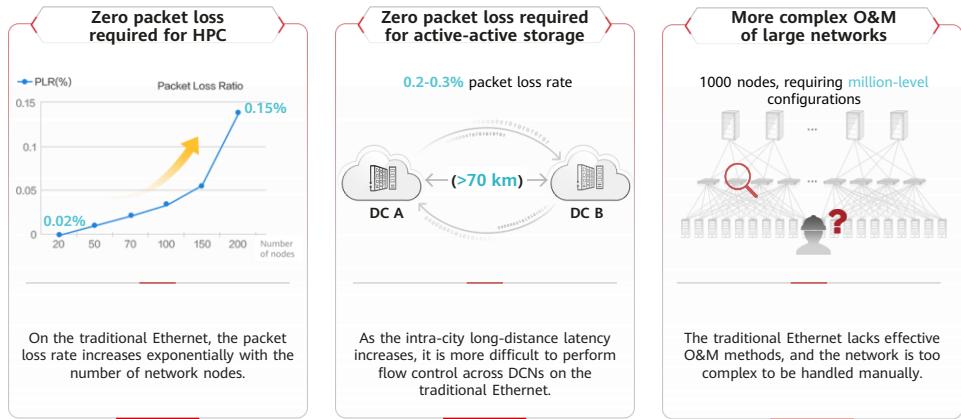
10 Huawei Confidential

 HUAWEI

- Note:

- Source ①: *Network Computing, the Meta Group and Contingency Planning Research*
 - Source ②: *App Annie*

Challenge 4: Three Challenges Faced by All-Ethernet Evolution

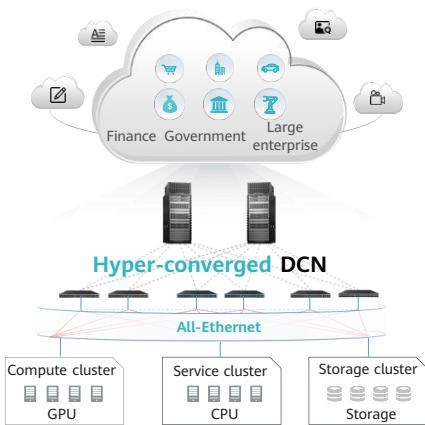


- Network evolution toward all-Ethernet faces three challenges. It is well known that the Ethernet is natively prone to packet loss, which remains unresolved for more than 40 years since the debut of the Ethernet. As the network scale increases, the packet loss rate increases exponentially. In intra-city active-active storage scenarios, long-distance transmission causes an extra latency of hundreds of microseconds, making it even harder for network flow control to implement zero packet loss. The Ethernet lacks effective O&M methods. As services are migrated to clouds, the network scale increases 100-fold, and the number of relationships between network objects such as ports and policies reaches millions. Manual network O&M no longer can meet requirements.

Contents

1. Development Trends and Challenges of DCNs
2. **Huawei CloudFabric Solution**
 - Solution Features
 - Overall Architecture
 - Application Scenarios
 - Core Components and Key Services
3. Huawei CloudFabric Typical Scenarios - Computing Scenario

Huawei CloudFabric Solution



Full-lifecycle automation

Automated network planning, construction, maintenance, and optimization
Intent-driven, network as a service (NaaS).

Lossless Ethernet

Local and long-distance lossless data transmission,
Converged computing and storage networks.

Network-wide intelligent O&M

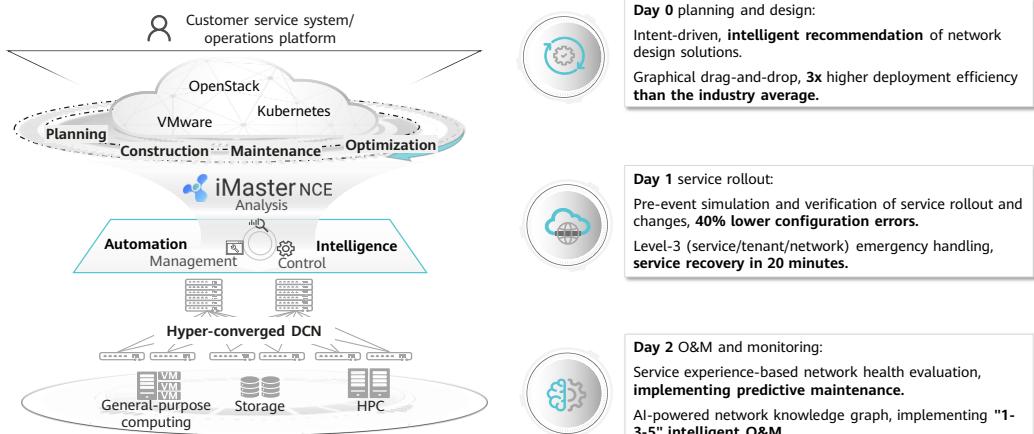
Predictive maintenance of devices, ports, optical modules, networks, and services, ensuring interruption-free services.

13 Huawei Confidential



- Based on the development trends and challenges of DCs, Huawei launches the CloudFabric Hyper-Converged DCN Solution, which can:
 - Implement full-lifecycle automation of services and improve the service TTM by 90%.
 - Build a lossless Ethernet network to implement lossless HPC and implement lossless long-distance transmission so as to build intra-city active-active storage networks over Ethernets.
 - Implement fast fault detection, intelligent analysis, and fast fault remediation, as well as proactive fault prediction in a large number of fault scenarios.
- Huawei CloudFabric Solution is built on Huawei DC flagship core switches — CloudEngine 16800/12800 series — and high-performance fixed switches — CloudEngine 8800/7800/6800/5800 series. It works with Huawei DC controller — iMaster NCE-Fabric, intelligent network analysis platform — iMaster NCE-FabricInsight, and security solution — HiSec, ideal for providing customers with simplified operation experience throughout the DCN lifecycle spanning network planning and construction, service rollout, O&M and monitoring, and change optimization. It also implements intelligent remediation of network faults, and can detect, analyze, and isolate network faults in real time. In addition, the CloudFabric Solution can meet the evolution requirements of DCs to an all-Ethernet architecture. It can integrate computing and storage networks, enable a lossless Ethernet, and improve computing and storage performance.

Feature 1: Full-Lifecycle Automation

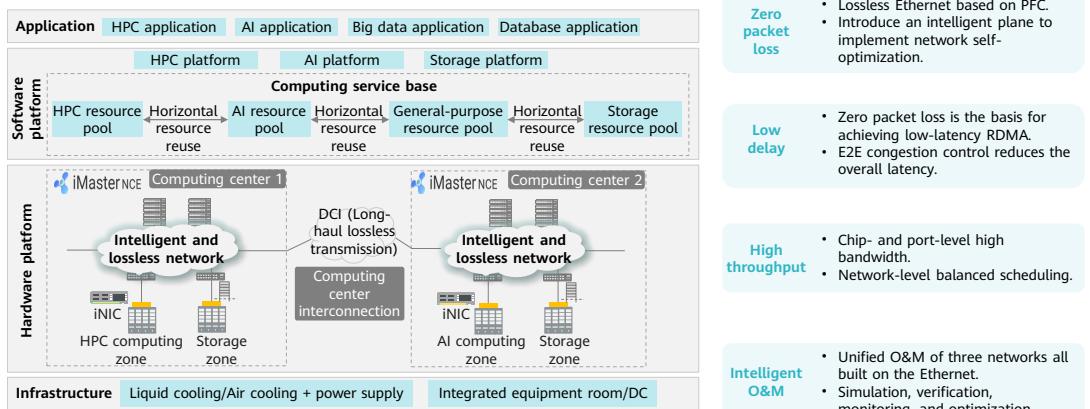


14 Huawei Confidential

HUAWEI

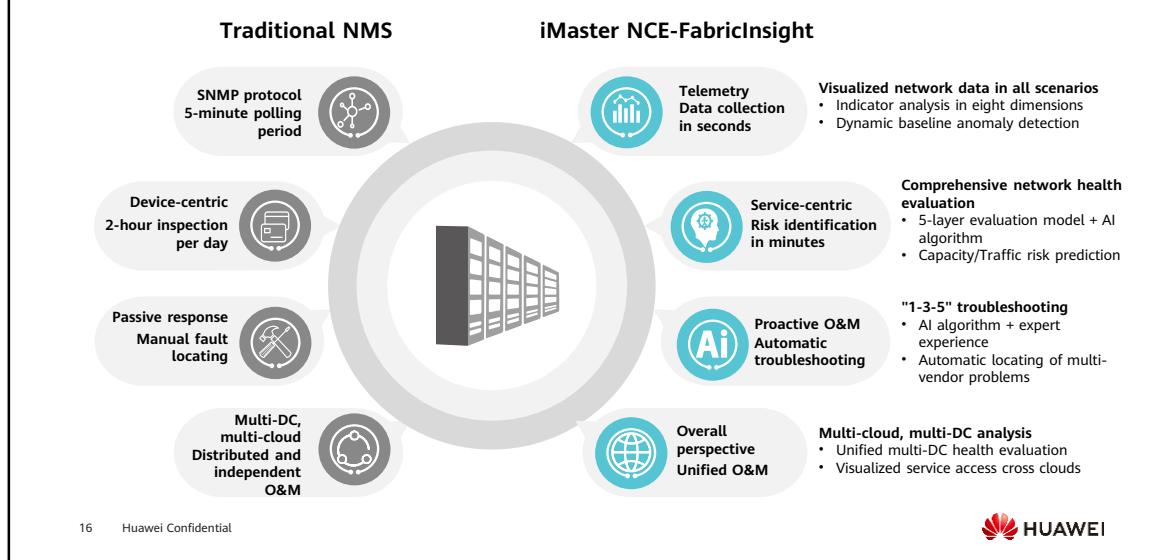
- Currently, network configuration automation has been implemented through SDN on many DCNs. However, service design and planning, technical review, and effect verification still need to be manually performed, involving multiple departments and roles. The entire process is time-consuming and inefficient, which has become the bottleneck of service rollout.
- The CloudFabric solution introduces intelligent algorithms in:
 - Design phase: The factors that affect network design are broken down into three evaluation dimensions: resource, quality, and reliability. In this way, the network solution can be generated and recommended in seconds.
 - Verification phase: The network topology, device configuration, and traffic information are calculated together to implement second-level verification of massive configurations on the entire network.
- The CloudFabric Solution can implement automated management and control throughout the network lifecycle spanning network planning and construction, service rollout, O&M and monitoring, and change optimization.

Feature 2: Lossless Ethernet



- The intelligent lossless algorithm overcomes the packet loss problem of Ethernet, which has remained unresolved for 40+ years. This helps to achieve zero packet loss under 100% throughput, meeting the ultimate network performance requirements of HPC and high-performance storage services and doubling the computing power and storage I/O performance at the same cluster scale.
 - The CloudFabric solution provides an all-Ethernet HPC network for HPC scenarios. Based on Huawei's unique iLossless™ algorithm, the solution solves the Ethernet packet loss problem that remains unresolved for many years and achieves zero packet loss under 100% throughput, providing the ultimate network performance required by HPC services with unchanged network scale and doubled computing power.
 - The CloudFabric solution provides an active-active all-Ethernet storage network for storage scenarios. Based on the iLossless™ algorithm for short-distance transmission, the iLossless-DCI algorithm is proposed to solve the packet loss problem in long-distance transmission scenarios. The solution increases network bandwidth by 10 times from 32GE to 400GE and significantly improves the storage input/output operations per second (IOPS) performance.
- Note:
 - PFC: priority-based flow control
 - RDMA: remote direct memory access
 - Three networks: front-end service network, diversified computing network, and storage network

Feature 3: Network-Wide Intelligent O&M



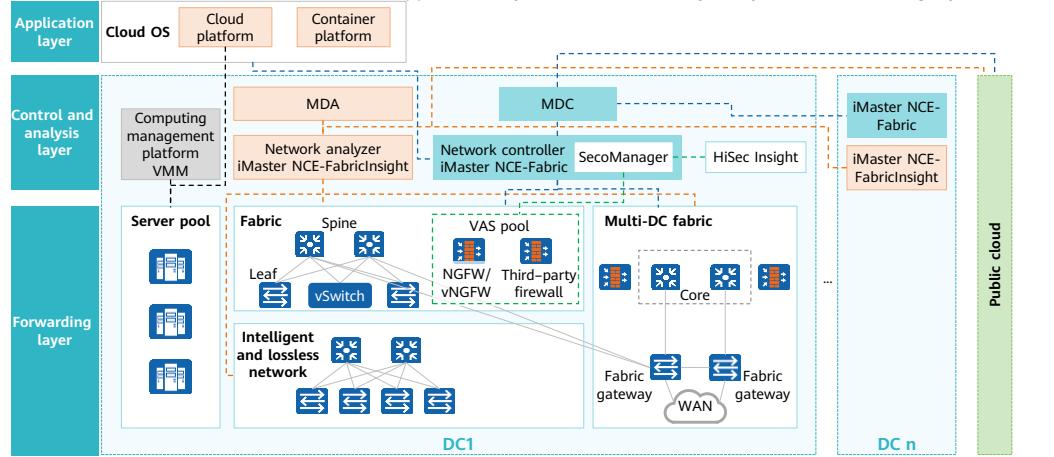
- The CloudFabric Solution uses telemetry technology to collect multi-dimensional data from the network, and uses the intelligent analysis platform to analyze network-wide O&M data. In addition to visualization of various O&M data, the CloudFabric Solution provides multiple key O&M capabilities.
 - Network health evaluation: A multi-dimensional evaluation system in terms of the device, network, protocol, overlay, and service is built to integrate configuration data, entry data, log data, and KPI performance data on the network with the help of telemetry. The intelligent analysis platform can detect issues and risks in each dimension of the network in real time. The detection scope covers the network working status, network capacity, component sub-health, and service traffic exchange. In this way, O&M personnel can view the overall experience quality of the entire network.
 - Rapid root cause locating: Based on knowledge graph, known DCN faults can be detected within 1 minute, located within 3 minutes, and rectified within 5 minutes. Unknown faults learning and fault inference are also supported to help O&M personnel deeply explore the root causes of unknown faults.
 - Automated assurance for service changes: Network data after configuration changes are collected to perform modeling to check whether the actual network forwarding behavior is consistent with users' service intents. O&M personnel can use the verification result to check whether the change meets the expectation and causes issues. If an intent fails verification, they can locate the failure cause, greatly improving the O&M efficiency in network change scenarios. In addition, important services can be periodically and automatically verified to ensure normal and reliable running of the services.

Contents

1. Development Trends and Challenges of DCNs
2. **Huawei CloudFabric Solution**
 - Solution Features
 - Overall Architecture
 - Application Scenarios
 - Core Components and Key Services
3. Huawei CloudFabric Typical Scenarios - Computing Scenario

CloudFabric Solution Architecture

- The CloudFabric Solution consists of the application layer, control and analysis layer, and forwarding layer.



18 Huawei Confidential

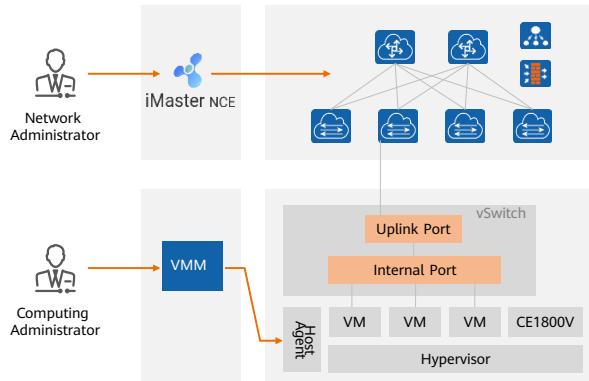
- Application layer:
 - Cloud OS:
 - Cloud platform: OpenStack-based cloud operating platform, including open-source OpenStack and Huawei FusionSphere, which collaboratively manage computing, storage, and network resources.
 - Container platform: creates and provisions containers.
- Control and analysis layer:
 - Computing management platform: The Virtual Machine Management (VMM) implements computing plane virtualization and resource management. vCenter and System Center are common computing management platforms.
 - Network controller: iMaster NCE-Fabric is used to centrally manage and control cloud DCNs. It provides automatic mapping from applications to physical networks, and implements resource pool deployment, and visualized O&M, helping customers dynamically schedule service-centric network services.
 - VAS controller: SecoManager implements centralized security policy management and control for firewalls, monitors events in real time, comprehensively analyzes security events such as attacks, and provides statistical reports in different formats. All of this helps customers master the cyber security status at any time.

Contents

1. Development Trends and Challenges of DCNs
2. **Huawei CloudFabric Solution**
 - Solution Features
 - Overall Architecture
 - **Application Scenarios**
 - Core Components and Key Services
3. Huawei CloudFabric Typical Scenarios - Computing Scenario

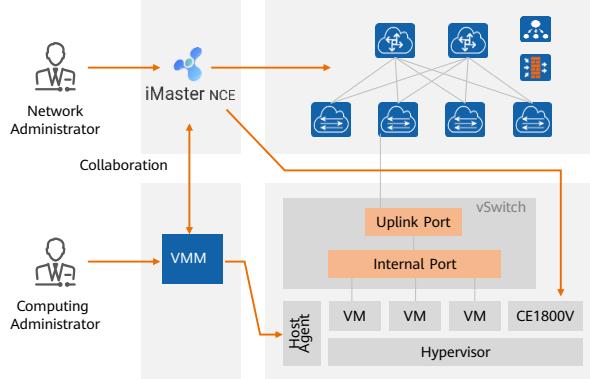
Hosting Scenarios Overview

- iMaster NCE and network are deployed without the cloud platform and VMM. The network administrator uniformly manages networks through the GUI provided by iMaster NCE.



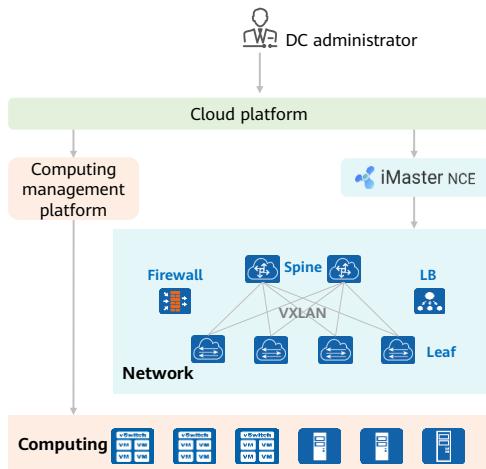
Computing Scenarios Overview

- The network administrator uniformly manages the physical and virtual networks through the GUI provided by iMaster NCE. The network system collaborates with compute resources, which are managed by the computing administrator.



- The computing scenario and hosting scenario are network virtualization scenarios, which are essentially only network devices for virtualization and management.

Cloud-Network Integration Scenarios Overview



23 Huawei Confidential

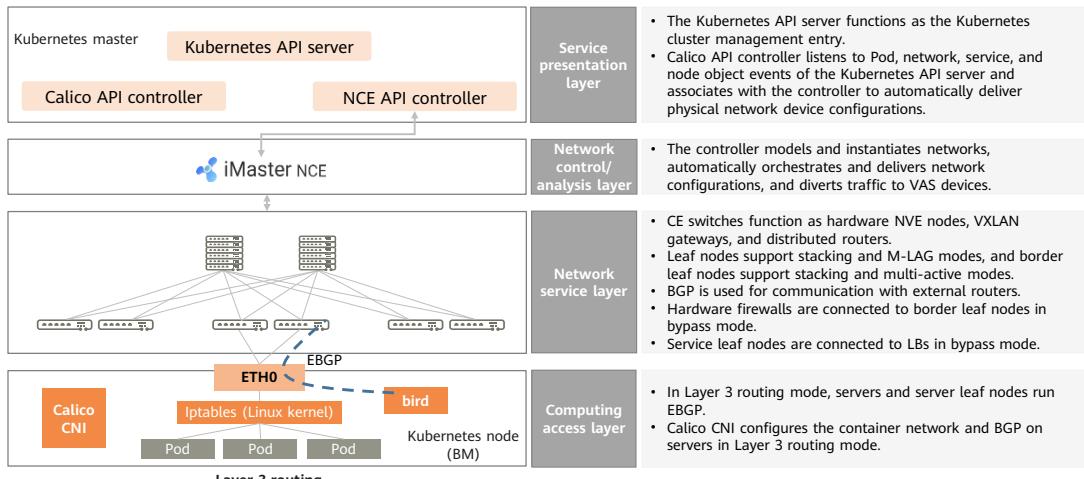
 HUAWEI

In this solution, user operations are performed on the cloud platform. The cloud platform interconnects with the controller (iMaster NCE-Fabric) and computing management platform (VMM) to implement association between computing and network services.

- The cloud platform delivers network-related service instructions to the controller for processing. The controller translates these service instructions into network configurations, and then delivers the configurations to corresponding network devices.
- The cloud platform delivers computing-related service instructions to the VMM for processing. The VMM then manages the lifecycle of compute resources.

- The cloud-network integration scenario applies to unified provisioning of computing and network resources based on the cloud platform. The cloud platform is the only portal for provisioning services and managing compute and network resources. It uses standard northbound APIs of the SDN controller to implement dynamic provisioning of tenant network resources as well as rapid network provisioning and resource adjustment, shortening service provisioning time. The benefits are as follows:
 - Unified computing and network service provisioning
 - This eliminates the isolation between the IT system and network system in the traditional service system. The cloud platform or orchestration platform collaborates with the SDN controller to provision and maintain services, providing a unified GUI for customers.
 - Information sharing between departments, improving efficiency
 - This breaks down the barriers between IT and network departments in traditional enterprises and implements unified collaboration, greatly improving their working efficiency.
 - Built on the quasi-standard platform
 - The system built based on OpenStack, the open-source cloud platform, can maximize customers' return on investment and be compatible with other systems and devices.

Container Network Scenarios Overview



24 Huawei Confidential

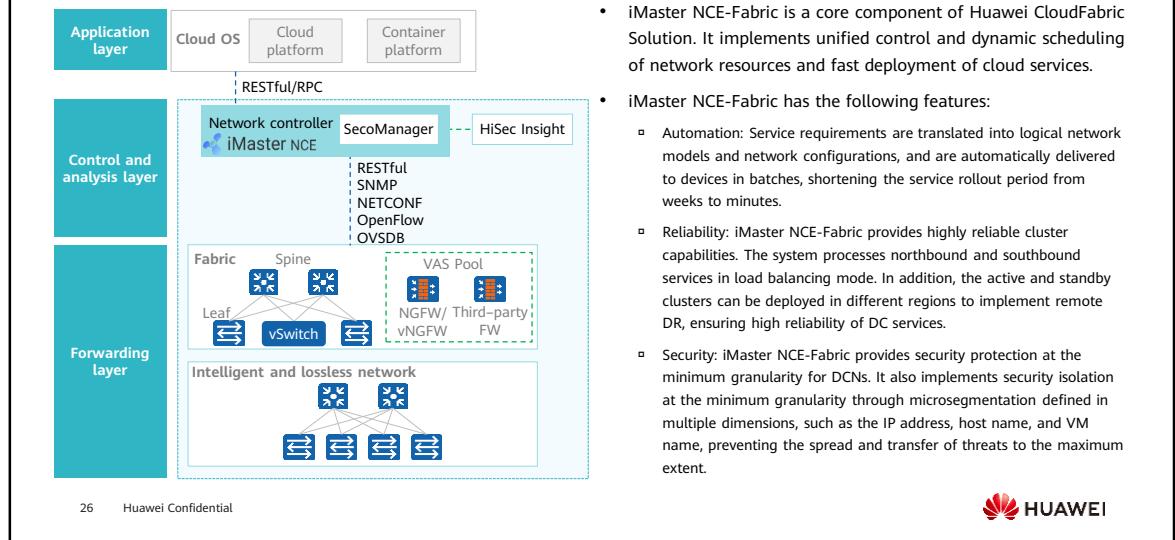
HUAWEI

- Container networks can be classified into independent deployment and interworking between container networks and physical networks. shows the interworking between container networks and physical networks. In this scenario, physical networks can be associated with container networks to implement automatic deployment, avoiding manual configuration errors, shortening service provisioning time, and providing stronger O&M features.

Contents

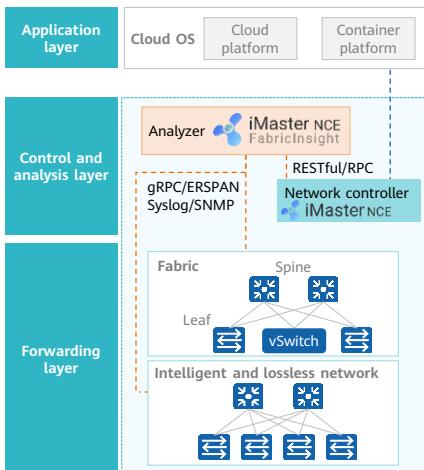
1. Development Trends and Challenges of DCNs
2. **Huawei CloudFabric Solution**
 - Solution Features
 - Overall Architecture
 - Application Scenarios
 - **Core Components and Key Services**
3. Huawei CloudFabric Typical Scenarios - Computing Scenario

Core Component: iMaster NCE-Fabric



- iMaster NCE-Fabric is designed based on open platforms, allowing it to connect to cloud platforms through northbound interfaces, to physical switches, vSwitches, and firewalls through southbound interfaces, and to the computing management platform through eastbound and westbound interfaces. These capabilities implement management and control of network resources and collaborative provisioning of compute and storage resources, resulting in an efficient, simple, and open DC. Based on the multi-engine capability of the data base, iMaster NCE-Fabric works with iMaster NCE-FabricInsight to provide L3 autonomous driving capabilities. The combination of system-based automated processing and manual assisted processing greatly reduces labor costs and error rates, as well as implementing conditional autonomy.
- iMaster NCE-Fabric interface description:
 - **Between the control layer and application layer:**
 - The two layers are interconnected via RESTful or RPC. The control layer receives service instructions from the application layer and returns status information to the application layer.
 - **Between the control layer and forwarding layer:**
 - The controller uses SNMP to discover and obtain physical device information, uses NETCONF to deliver configurations to physical devices, and uses OpenFlow to deliver flow tables to physical devices which are used for constructing detection packets during O&M.

Core Component: iMaster NCE-FabricInsight



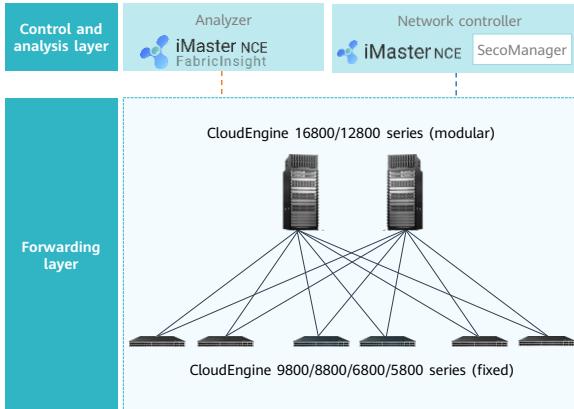
28 Huawei Confidential

- iMaster NCE-FabricInsight is an intelligent network analysis platform of CloudFabric. It detects fabric network status and application behavior status in real time, breaks network and application boundaries, and helps customers detect network and application problems in a timely manner from the application perspective, ensuring continuous and stable running of applications.
- Features of iMaster NCE-FabricInsight:
 - Display service flows and network-wide KPIs through Telemetry in seconds, implementing correlation analysis of services, network paths, and network devices and giving intuitive insights into network health.
 - Train the knowledge inference engine based on machine learning to analyze root causes of dozens of faults in minutes, implement edge intelligence based on software and hardware, and perform comprehensive analysis of TCP and UDP flows.
 - Construct dynamic baselines, identify device, queue, and port exceptions, and proactively predict traffic and optical module faults.



- iMaster NCE-FabricInsight interface description:
 - **Inside the control layer:** The control layer uses RESTful or RPC to synchronize configuration and status information between control units.
 - **Between the control layer and forwarding layer:** The analyzer at the control layer connects to network devices through Google Remote Procedure Call (gRPC) or ERSPAN to collect and send device data. Syslog or SNMP is used to collect device status, alarms, and logs.
- The overall architecture of iMaster NCE-FabricInsight consists of three parts: network device, collector, and analyzer.
 - **Network device:** includes Huawei CloudEngine (CE) switches, NetEngine (NE) routers, and some third-party devices. Devices report performance metrics such as interface traffic in Telemetry mode based on the gRPC protocol. Devices are connected to iMaster NCE-FabricInsight as gRPC clients. Users can run commands to configure the telemetry function on the devices. The devices then proactively establish a gRPC connection with the target collector and send data to the collector. The current version supports the following sampling metrics: CPU and memory usage at the device and card levels; number of sent and received bytes, number of discarded sent and received packets, and number of sent and received error packets at the interface level; number of congested bytes at the queue level; packet loss behavior data.
 - **Collector:** receives data reported by network devices via telemetry, including performance metric data reported through gRPC. The collector parses the metrics, combines and compresses the metric data, and reports the data to the analyzer.
 - **Analyzer:** receives performance metrics from switches. In addition, the analyzer establishes dynamic baselines for some performance metrics based on the AI algorithm, detects exceptions, and displays the analysis result on the GUI.

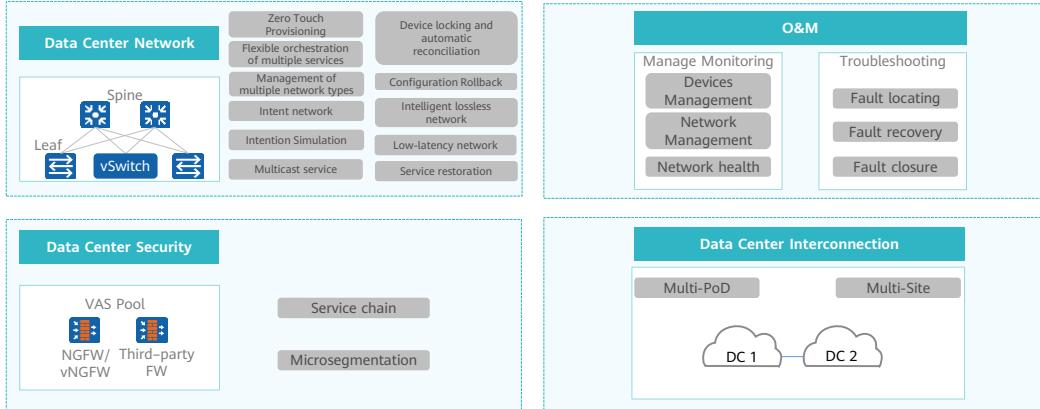
Core Component: CloudEngine Series Switches



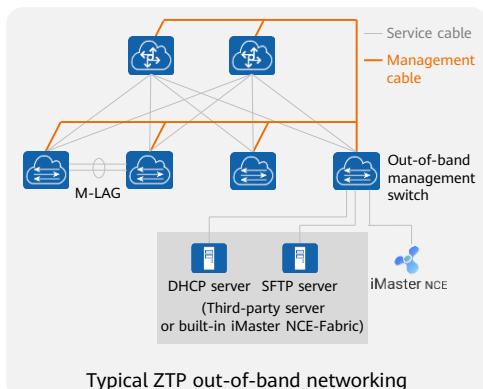
- Huawei CloudEngine (CE for short) series switches are high-performance cloud switches designed for next-generation DCs, including the industry's first DC switches designed for the intelligence era (CE16800 series) and next-generation high-performance core switches designed for DCs (CE12800 series), and high-performance aggregation/access switches (CE9800/8800/6800/5800 series).

Key Service Overview

- Huawei CloudFabric provides mission-critical services at multiple layers of data centers, efficiently building agile data centers.



Initial Network Construction: ZTP



Application scenario

- Zero Touch Provisioning (ZTP) allows newly delivered or unconfigured devices to automatically load version files, deploy the underlay network, and be managed by iMaster NCE-Fabric after they are powered on.

Deployment solution

- Out-of-band deployment:** iMaster NCE-Fabric connects to the management interfaces of all devices to be brought online through the out-of-band management switch.
- In-band deployment:** The management network and service network share service network ports, and no independent switch needs to be deployed.

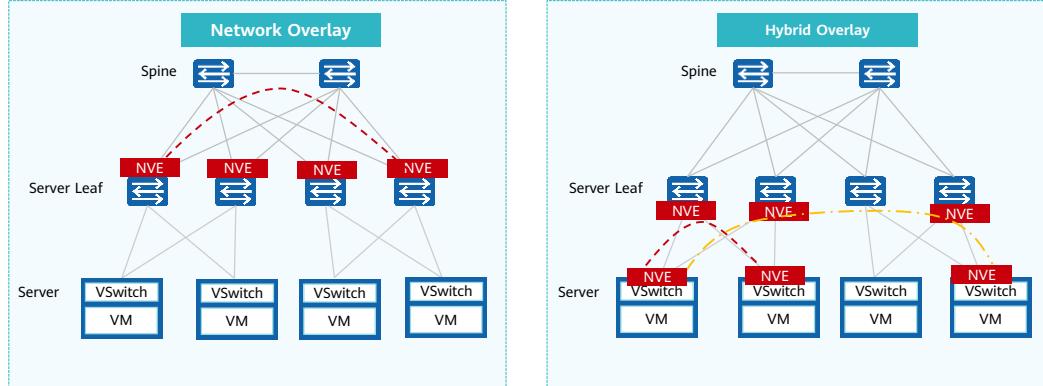
Deployment mode

- Typical configuration mode:** A plan file is automatically generated, reducing the workload of filling in the topology template.
- User-defined import mode:** Topology planning is required, which is highly refined.

- In the traditional deployment mode, administrators need to manually configure each newly delivered or unconfigured device after hardware installation, which lowers deployment efficiency and results in high labor costs. This is where the ZTP-based simplified deployment function of iMaster NCE comes in. The function enables users to complete network topology planning and fabric resource planning, automatically bring devices online, execute device configuration scripts, and deliver underlay network configurations to devices in batches on a visualized GUI. This reduces labor costs and improves deployment efficiency. ZTP-based simplified deployment enables rapid rollout and management of DCN devices.
- In the CloudFabric solution, the physical DC network uses the spine-leaf architecture and supports horizontal on-demand capacity expansion. The roles on the network include spine nodes, server leaf nodes, border leaf nodes, service leaf nodes, and DCI gateways. There are often a large number of server leaf nodes, which require automatic service rollout. Therefore, ZTP mainly focuses on server leaf nodes.
 - Server leaf nodes support M-LAG and standalone networking, which are applicable to different server access scenarios. M-LAG networking is recommended because high reliability is achieved when servers are dual-homed to switches in M-LAG mode. In addition, each M-LAG device has its own control plane, simplifying upgrade and maintenance.
- The CloudFabric solution supports two network architectures: three-layer networking and two-layer networking, which are both supported by ZTP.
 - Three-layer networking architecture: Spine nodes, border leaf nodes, and service leaf nodes are separately deployed, which applies to large network scenarios.
 - Two-layer networking architecture: Spine nodes, border leaf nodes, and service leaf nodes are combined, which applies to small and midsize network scenarios.

Management of Multiple Network Types

- iMaster NCE-Fabric supports multiple overlay fabric types to meet different user requirements in different application scenarios, such as network forwarding performance, server access type, and VXLAN tunnel encapsulation points.



32 Huawei Confidential

HUAWEI

- In a network overlay network, all overlay devices are physical devices, and VXLAN tunnels on the overlay network are encapsulated on physical switches. This networking has the advantages of high forwarding performance and reliability, and can connect to multiple servers. Servers do not need to support VXLAN tunnel encapsulation. Network overlay is applicable to new data centers that have high requirements on forwarding performance and security, and SDN networks and traditional networks need to communicate with each other.
- On a hybrid overlay network, overlay devices include physical and virtual network devices. Overlay VXLAN tunnel encapsulation can be implemented on either the physical switch or the virtual switch where the host server resides. The hybrid overlay can not only use the high-performance forwarding of physical network devices, but also improve performance by reusing existing physical network devices and overlaying physical servers. Therefore, hybrid overlay networking is more flexible and provides customers with more choices. Hybrid overlay networking is applicable to scenarios where network capacity expansion, hardware costs are sensitive, network reuse is emphasized, VXLAN and hardware decoupling is required, and SDN networks and traditional networks need to communicate with each other.

Three-Level Rollback

Network-wide rollback	Tenant snapshot	Service-level rollback
<ul style="list-style-type: none"> Network-wide rollback is used to resolve major faults on the entire network. For example, if network configurations are deleted due to changes, many services are interrupted. In this case, network-wide configurations can be rolled back to those before the changes or interruptions, enabling quick service recovery. Before changes, you can back up network-wide configurations on iMaster NCE-Fabric. When a problem occurs due to changes, the configurations can be quickly restored to the backup point, resolving major network faults. You can manually save data in real time or periodically on the GUI. You need to proactively back up data. 	<ul style="list-style-type: none"> The tenant snapshot function is used to back up and restore network service configurations by tenant, and apply to multi-tenant services. Backup and restoration operations performed by a tenant do not affect the provisioning of other tenants' services, including backup and restoration of network service configurations by other tenants. The tenant snapshot function allows a tenant to set a backup point and save all its service configurations at the backup point. If needed, service configurations can then be restored to a specific snapshot point. Additionally, iMaster NCE-Fabric can compare the current configurations with the configurations at the snapshot point, or compare the configurations from two given snapshot points, and perform configuration rollback to eliminate differences. The tenant snapshot function supports manual backup and restoration as well as automatic and periodic backup. 	<ul style="list-style-type: none"> Service-level rollback helps quickly restore original network configurations to recover services when a network exception occurs due to a fine-grained single-point service provisioning failure. You do not need to manually back up data for service-level rollback, but need to manually restore data. iMaster NCE-Fabric automatically backs up each service that is provisioned. When an exception occurs, iMaster NCE-Fabric can quickly restore the service to the status before the service is provisioned.

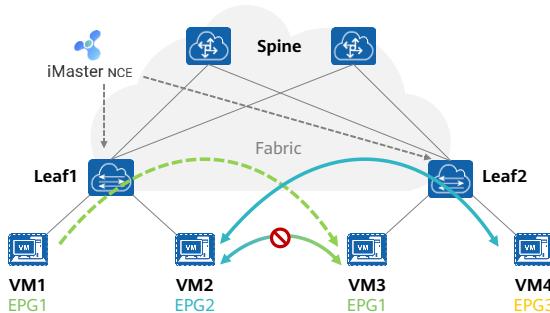
33 Huawei Confidential



- iMaster NCE-Fabric provides three-level rollback, meeting the reliability requirements of different scenarios and ensuring quick service recovery. This feature covers 70% to 80% of routine change scenarios. For example, the fast rollback feature is available for single-point service provisioning exceptions and independent tenant services.

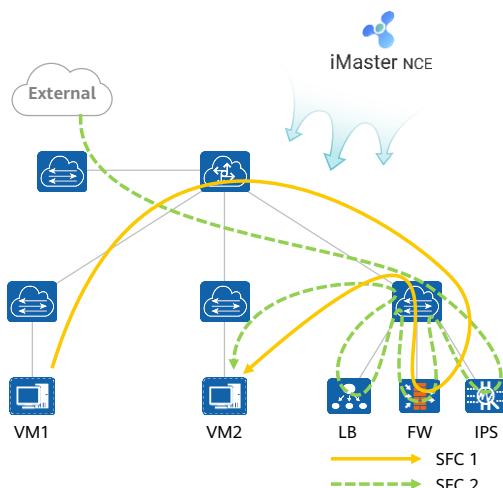
- Network-wide rollback features:
 - iMaster NCE-Fabric saves the snapshots of the entire network, including those of iMaster NCE-Fabric and its managed devices.
 - You can manually save the snapshots in real time or periodically.
 - During restoration, iMaster NCE-Fabric delivers commands to devices to restore data. The devices restore specific configurations based on specified snapshot point labels and do not need to be restarted.
- Tenant snapshot features:
 - You can manually save the snapshots in real time or periodically.
 - iMaster NCE-Fabric divides different tenant spaces for tenant backup so that operations between tenants do not affect each other.
 - Differences between rollback points can be previewed for further examinations.
- Service-level rollback features:
 - Service operations are automatically saved.
 - Snapshots are automatically stored in mirroring mode.
 - The linkage technology enables rollback of multiple operations to the previous state.

Microsegmentation



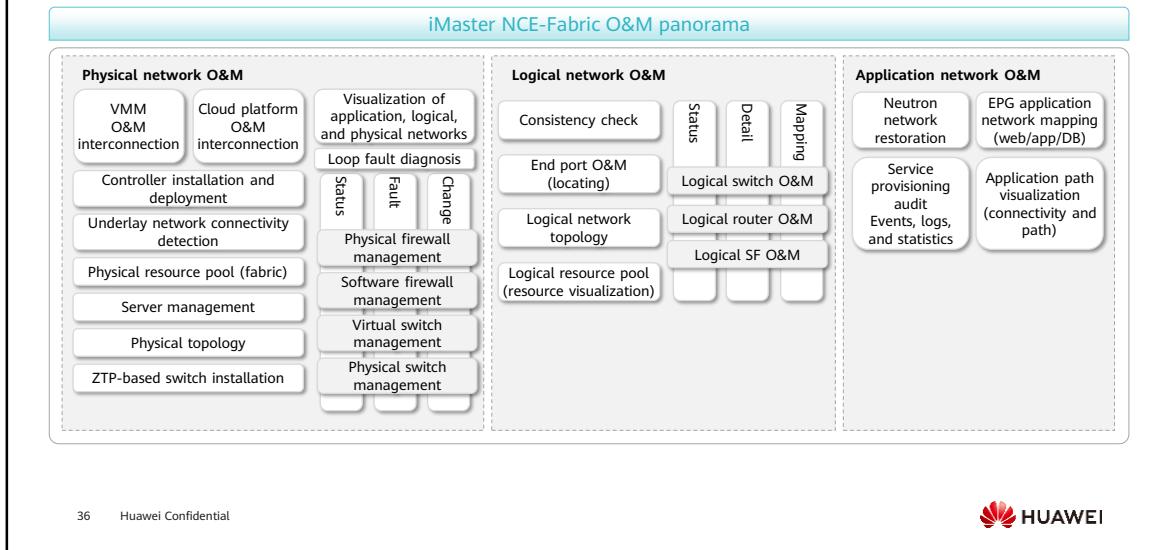
- Microsegmentation allocates servers to different EPGs and defines GBPs between EPGs to implement traffic control between servers.
- Microsegmentation can be implemented either on CE switches or on iMaster NCE-Fabric. iMaster NCE-Fabric configures EPGs and GBPs and delivers the configurations to CE switches through NETCONF interfaces.

Service Chain



- SFC is a technology that logically connects services on network devices to provide an ordered service set for the application layer. SFC adds service function path (SFP) information to original packets to enable packets to pass through SFs along **the specified path**.
- SFC can be implemented in Policy-based Routing (PBR) mode or Network Service Header (NSH) mode. When creating a fabric network on the controller, you must specify the PBR or NSH mode. Install logical switches, logical routers, or external gateways on the controller to divide EPGs and define SFCs between EPGs.

iMaster NCE-Fabric O&M

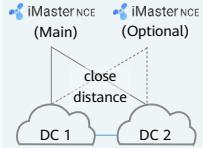


- iMaster NCE-Fabric centrally manages and controls cloud DCNs and provides automatic mapping from applications to physical networks, resource pool deployment, and visualized O&M, helping customers build service-centric dynamic network service scheduling capabilities.
- In addition to network planning and deployment, iMaster NCE-Fabric also provides DCN service O&M, including: topology visualization, loop detection, path detection, traffic statistics collection, three-level rollback, and data consistency verification.

Multi-DC Service

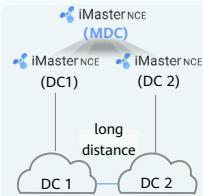
- With the development of services, more and more applications are deployed in data centers. The resources of a single data center cannot meet the increasing service requirements. Therefore, multiple data centers are required to deploy services.

Multi-PoD solution



The computing and network resources of multiple DCs are unified and managed by a cloud platform and a set of iMaster NCE (Fabric).

Multi-site solution



In the multi-DC scenario, the computing and network resources of each DC are independent resource pools and are managed by the cloud platform and iMaster NCE (Fabric) in their respective DCs.

Contents

1. Development Trends and Challenges of DCNs
2. Huawei CloudFabric Solution
- 3. Huawei CloudFabric Typical Scenarios - Computing Scenario**

- This course describes only the computing scenario in detail. For details about other scenarios, see the related sections of HCIE-DCN.

Introduction to Computing

**Challenge 1**

Service types are becoming more refined, and an increasing number of devices are deployed, resulting in increasingly high configuration and management costs.

Challenges for enterprise IT

Challenge 2

IT resources always seem to be insufficient while the resource utilization is low. The resource utilization is unbalanced, and resources cannot be flexibly scheduled.

Cloud computing provides various advantages such as resource pooling, elastic scaling, and on-demand self-service provisioning, helping enterprises cope with the preceding challenges.

However, some enterprises cannot fully implement cloud-based services at a time.

Technical factors

- Service systems are complex and have different requirements on the running environment.
- The application scale of each service system can be estimated within a certain period of time and will not scale greatly.

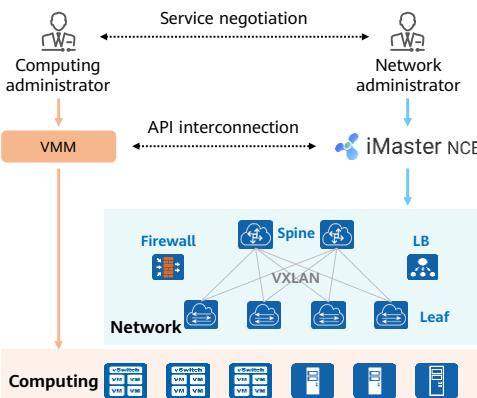
Non-technical factors

- Generally, enterprises have IT department and network department but no cloud platform department. The IT department and network department are responsible for computing and network requirements, respectively. These two departments cannot be integrated in a short period of time.

Some enterprises that cannot achieve cloud computing at a time start with automation reconstruction on networks. That is, they associate network resources with compute resources, and then gradually transform their networks toward the scenario where a unified cloud platform will be deployed, which is the Cloud-Network Integration scenario.

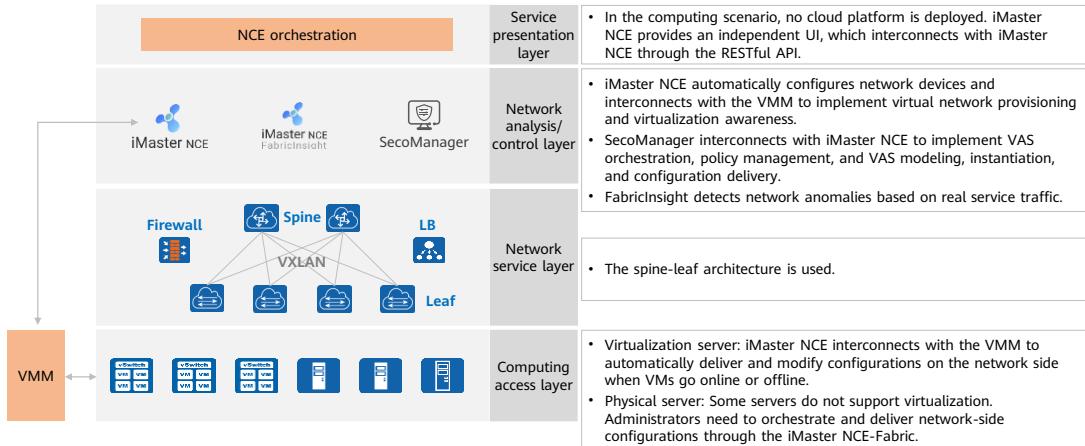
- Currently, enterprises face the following challenges in their IT systems:
 - Service types are becoming more refined, and an increasing number of devices are deployed, resulting in increasingly high configuration and management costs.
 - IT resources always seem to be insufficient while the resource utilization is low. The resource utilization is unbalanced, and resources cannot be flexibly scheduled.
- Cloud computing provides various advantages such as resource pooling, elastic scaling, and on-demand self-service provisioning, helping enterprises cope with the preceding challenges.

CloudFabric Solution Overview in the Computing Solution

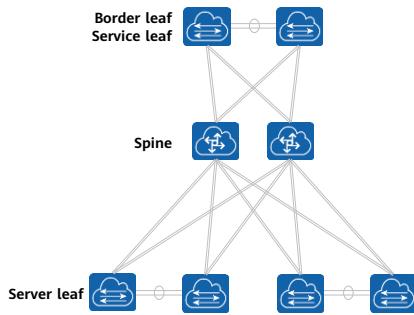


- In this scenario, no cloud platform is involved. The network administrator configures network services through the controller, and the computing administrator configures compute resources. The network administrator and computing administrator perform service negotiation through the enterprise's internal service process.
- The controller can interconnect with the VMM to implement service automation. The controller delivers network configurations to the computing platform through APIs. The computing platform notifies the controller of the VM online and offline information, and the controller delivers the configuration to the corresponding API to complete E2E service configuration.
- The computing solution implements automatic network configuration to the maximum extent, reducing the configuration workload of the network administrator.

CloudFabric Solution Architecture in the Computing Solution



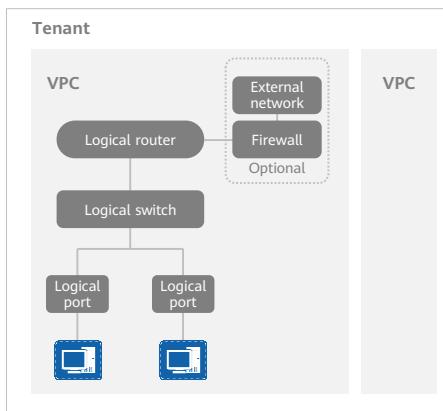
Networking Solution



- The spine-leaf architecture is used. The border leaf node and service leaf node are co-deployed, and the spine node is deployed separately and connects to VAS devices and external networks. The spine and leaf nodes are fully meshed to implement ECMP load balancing.
- The distributed network overlay solution is recommended. iMaster NCE centrally manages the gateways and automatically delivers service configurations.
- OSPF or BGP is used as the routing protocol on the forwarding plane of the underlay network, and BGP EVPN is used as the control plane protocol. A BGP EVPN peer relationship is established between VTEPs.

Service Model: Tenant Service Model (1)

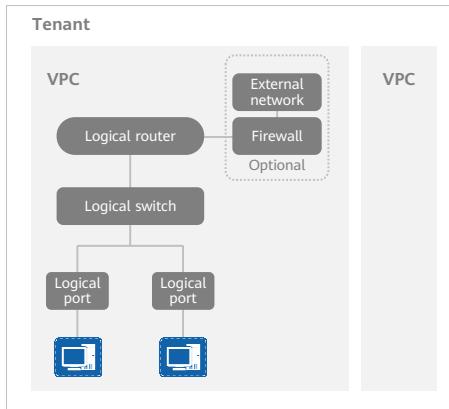
- Understanding the terminology, background, and implementation principles associated with service provisioning is helpful for quickly mastering tenant network interconnection skills in a computing scenario.



- Tenant:** is the minimum unit for enterprise service management.
- Virtual Private Cloud (VPC):** provides secure and reliable information processing, storage, and transmission services to tenants through the virtualization and encryption technologies based on network, storage, and compute resources. Multiple VPCs can be created for a tenant based on service requirements.
- Logical router:** is virtualized by a network device where virtualization software is running, and is connected to VMs on different networks, so that VMs can communicate with each other on a Layer 3 network. One network device can be virtualized into multiple logical routers for different tenants.
- Logical switch:** connects to different VMs to ensure that the VMs can communicate with each other at Layer 2. One network device can be virtualized into multiple logical switches for different tenants.

- One network device can be virtualized into multiple logical routers for different tenants. Multiple tenants can share a network device. For each tenant, a logical router functions as an independent and real router with independent hardware and software resources and running space. Services on different logical routers do not affect each other. In terms of experience, there is no difference between a logical router and a real router.
- One network device can be virtualized into multiple logical switches for different tenants. Multiple tenants can share a network device. For each tenant, a logical switch functions as an independent and real switch with independent software and hardware resources and running space. Services on different logical switches do not affect each other. In terms of experience, there is no difference between a logical switch and a real switch.

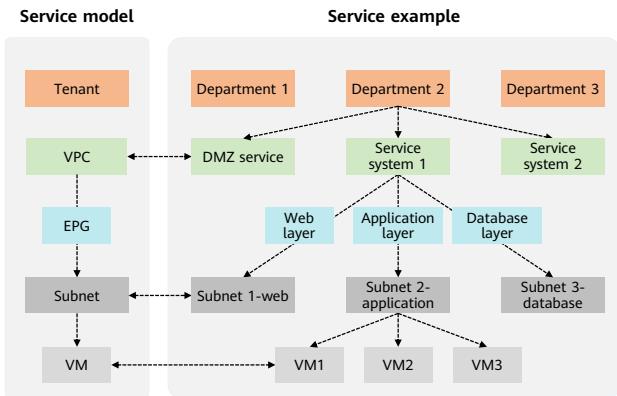
Service Model: Tenant Service Model (2)



- **Logical port:** functions as an access point for VMs to access the network. One physical port on a network device can be virtualized into multiple logical ports for different tenants. For each tenant, a logical port functions as an independent and real port.
- **External network:** networks outside the tenant's management, such as Internet or other tenant networks connected through VPNs.
- **Firewall:** The firewall function is provided by a physical firewall or virtual firewall.
- **VM:** virtual machine.

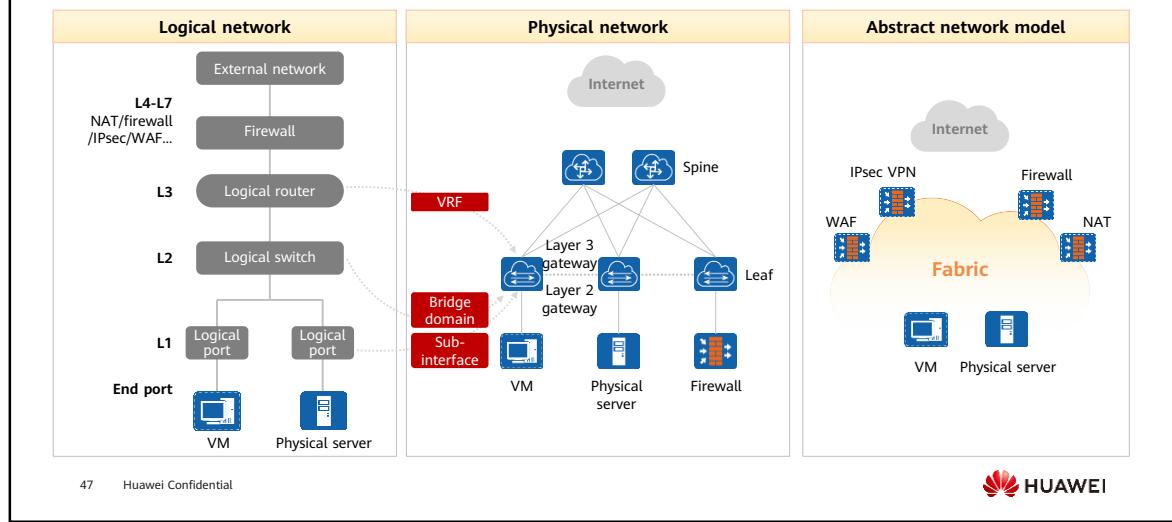
- Located at the border of a network, a firewall implements secure access control between the external network and internal network, which enhances the network protection capability. It protects service data flows between the Untrust and Trust zones based on 5-tuple information. It can also be used for access control between subnets. You can choose whether to deploy firewalls based on whether the tenant needs to access an external network. For security purposes, deploy a firewall when a tenant is connected to an external network.
- In the computing scenario, VMs are provisioned by the VMM connected to iMaster NCE-Fabric. The VMM manages compute resources, and iMaster NCE-Fabric manages network resources.

Example of a Tenant Service Model



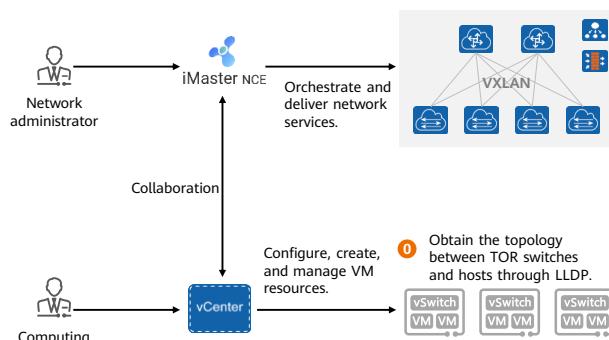
- **Tenant:** A tenant can apply for independent compute, storage, and network resources, and can be regarded as a service system or department.
- **VPC:** Each VPC is a security domain and can be regarded as a collection of services that have the same security policy. A VPC is mapped to a VRF.
- **EPG:** An EPG is a set of service ports. Service ports in an EPG have the same security policy. An EPG can have one or more subnets. Security policies can be easily configured using EPGs.
- **Subnet:** A subnet indicates a network segment. A VPC can have one or more subnets.
- **VM:** A VM is connected to only one subnet, and one subnet can have multiple VMs.

Relationship Between a Physical Network and a Logical Network



- The physical network uses the spine-leaf architecture. VMs, switches, and firewalls access the network through switches at the leaf layer. VMs and physical machines function as computing nodes. Firewalls function as network nodes and provide NAT, IPsec VPN, WAF, and firewall (packet filtering) network services in SFC.
- Common packets of VMs, physical machines, and firewalls are encapsulated into VXLAN packets and transmitted on a fabric network constructed by switches at leaf and spine layers. VXLAN Layer 2 gateways encapsulate the common packets into VXLAN packets at the access layer. VXLAN Layer 2 gateways provide data transmission services within a subnet and are called internal gateways.
- For communication between an intranet and the Internet, between an intranet and an external private network, and between subnets within an intranet, VXLAN Layer 3 gateways are required for route query and data forwarding. VXLAN Layer 3 gateways connect to the Internet using PE routers (not displayed in the figure). VXLAN Layer 3 gateways transform the VXLAN packets sent from an intranet to the Internet or an external private network to common packets and forward these packets. VXLAN Layer 3 gateways are also called external gateways. If VMs and physical machines need to communicate with other subnets or external networks, the IP addresses of gateways must be set to the IP addresses of VXLAN Layer 3 gateways by default.

Service Provisioning Process (1)

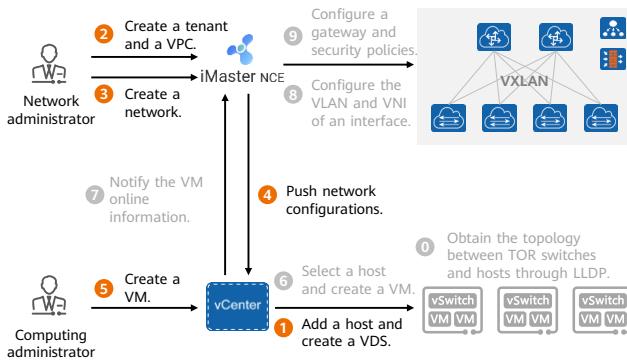


The network administrator and computing administrator need to cooperate with each other to deliver services in the computing scenario:

- The network administrator orchestrates and delivers network services on iMaster NCE.
- The computing administrator configures, creates, and manages VM resources on the VMM (vCenter in this example).

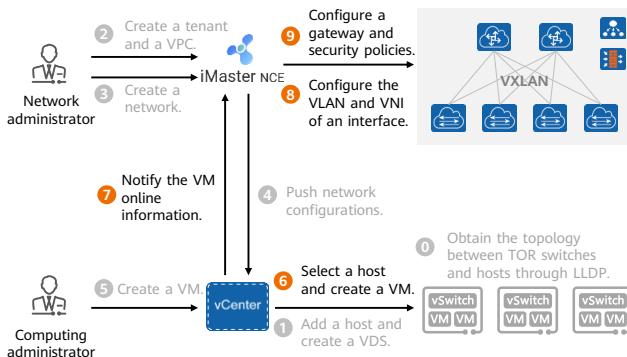
The topology, namely, connections between ports on TOR switches and physical servers, has been discovered between servers and TOR switches through LLDP before service provisioning in the computing scenario.

Service Provisioning Process (2)



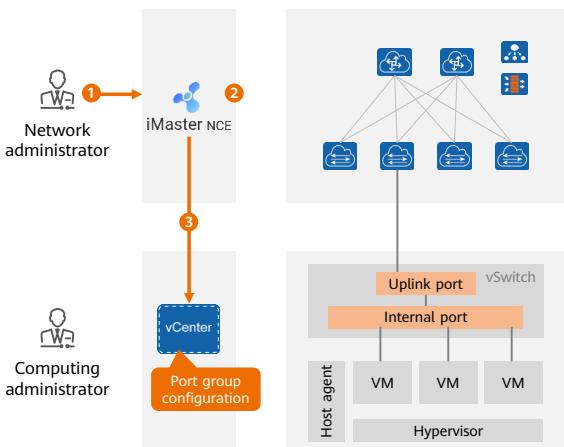
1. The computing administrator enables the VMM to manage servers and creates a virtual switch VDS.
2. The network administrator creates a tenant and a VPC on iMaster NCE.
3. The network administrator orchestrates the logical network in the VPC, including vRouters and subnets.
4. iMaster NCE synchronizes information about VLANs corresponding to the subnets to the VMM. The VMM then creates a port group for each subnet.
5. The computing administrator creates a VM and configures VM parameters, including the port group to which the VM belongs.

Service Provisioning Process (3)



- When receiving the instruction from the computing administrator, the VMM selects a host that it manages, creates a VM, and allocates resources to it.
- After confirming that the VM is online, the VMM notifies iMaster NCE of the information.
- iMaster NCE obtains the information about the host of the VM from the VMM, determines the TOR switch port for connecting to the host based on the LLDP information, and then delivers VLAN and VNI mappings to the port.
- iMaster NCE delivers the VM gateway and security policy configurations based on the configurations performed by the network administrator in the VPC. Then the VM can access the network properly.

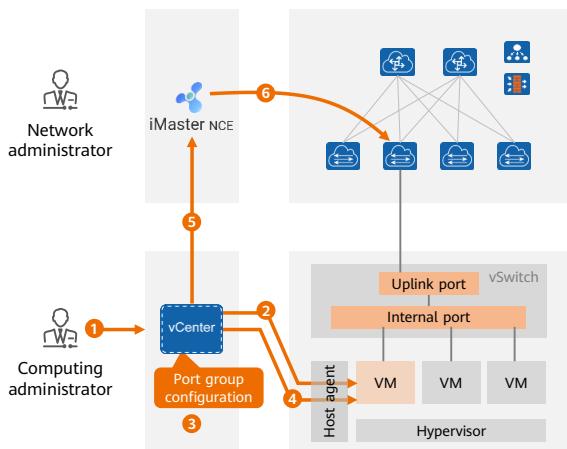
Network Resource Provisioning Process



When the network administrator configures and provisions network resources, the computing administrator is not aware of it.

1. The network administrator edits the logical networks required by services on the tenant network orchestration page of iMaster NCE.
2. iMaster NCE computes and saves the mapping between VLANs and VNIs based on the VNI range allocated by the administrator. These configurations and mappings are stored on iMaster NCE and have not been delivered to switches since VMs have not gone online.
3. iMaster NCE connects to the VMM through the WebService interface and transfers the preceding information. The VMM creates a port group required by the local network on the virtual switch and binds the local VLAN ID to the port group.

VM Online Process

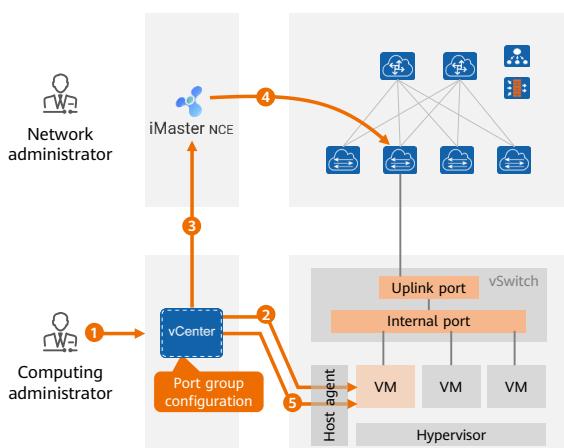


53 Huawei Confidential



- The computing administrator configures compute resource quota (vCPU, RAM, and operating system) on the VMM based on service requirements.
- The computing administrator creates a VM. The VMM dynamically allocates compute resources based on the existing configurations, selects a host, and loads the VM based on the configuration in step 1.
- The computing administrator checks the port group information synchronized from the network side on the VMM and manually binds the NIC of the VM to the corresponding port group.
- The VMM synchronizes the port group configurations to the corresponding host and binds the VM to the port group.
- iMaster NCE detects the VM online and port group binding information through the WebService interface, and obtains the location where the VM goes online (including the VM ID and the ID of the host where the VM is located).

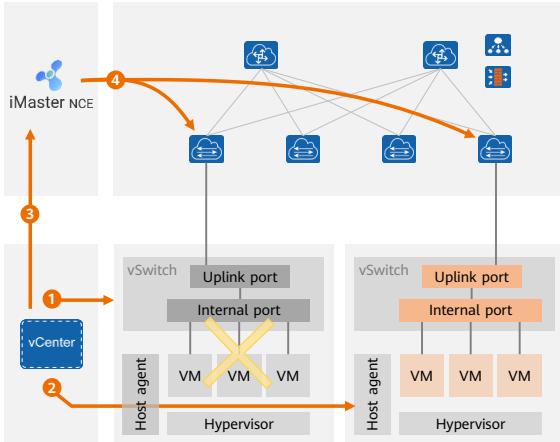
VM Offline Process



The VM offline process is also performed automatically, which cannot be detected by the network administrator.

1. The computing administrator brings a VM offline through the VMM.
2. The VMM queries the database, finds the host to which a specified VM belongs, brings the VM offline, removes the binding between the VM and the port group, and reclaims compute resources.
3. iMaster NCE detects the VM offline information and unbinding between the VM and port group through the WebService interface, and obtains the location where the VM goes offline.
4. iMaster NCE obtains the connection between the host and TOR switch port through LLDP, queries the database using the port group as the index to obtain the mapping between the local VLAN and VNI, and checks whether any VM still uses the local VLAN on the same port. If no VM uses the local VLAN, iMaster NCE removes the mapping between the local VLAN and VNI through NETCONF.
5. The VMM checks whether any other VM on the host is bound to the current port group. If no VM is bound to the port group, the VMM reclaims the port group configuration.

Automatic VM Migration Process

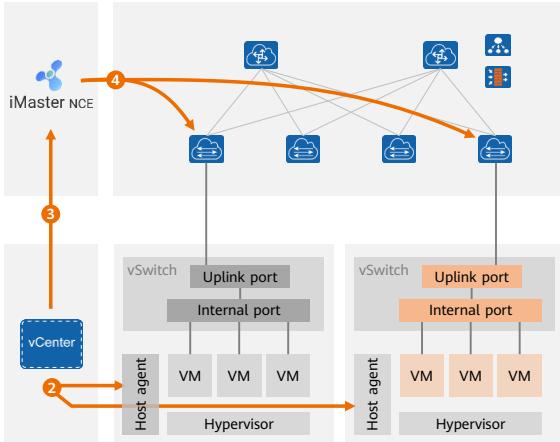


If a host or VM is faulty, the system restarts all VMs on the faulty host on other hosts.

1. The VMM detects a fault on the host.
2. The VMM schedules resources and restarts all VMs of the faulty host on other hosts.
3. iMaster NCE subscribes to VMM events, detects the VM migration, and obtains the locations of the hosts before and after VM migration.
4. iMaster NCE finds TOR switches and corresponding ports before and after the migration through LLDP. It deletes the mapping between VLANs and VNIs on the TOR switches before the migration through NETCONF, and delivers the mapping between VLANs and VNIs on the new TOR switches.

- Both the computing administrator and network administrator are unaware of the automatic migration process. The compute resources are automatically migrated and network configurations are automatically adjusted based on the collaboration between the VMM and iMaster NCE.

Manual VM Migration Process



The computing administrator can manually migrate VMs on the VMM GUI.

1. The computing administrator triggers VM migration through the VMM.
2. The VMM finds the host to which a VM belongs, performs re-scheduling in the VMM cluster, selects a new target host, and migrates the VM.
3. iMaster NCE subscribes to VMM events, detects the VM migration, and obtains the location of the host before and after VM migration.
4. iMaster NCE finds TOR switches and corresponding ports before and after the migration through LLDP. It deletes the mapping between VLANs and VNIs on the TOR switches before the migration through NETCONF (VMs with the same port group do not exist on hosts), and delivers the mapping between VLANs and VNIs on the new TOR switches.

Quiz

1. Which of the following components is used to deploy networks in Huawei CloudFabric solution?()
 - A. iMaster NCE-Fabric
 - B. SecoManager
 - C. iMaster NCE-FabricInsight
 - D. MDA

1. A

Summary

- Huawei CloudFabric Solution redefines the O&M, deployment, and interconnection of DCNs to build intelligent, simplified, ultra-broadband, open, and secure cloud DCNs. Leveraging iMaster NCE-Fabric and iMaster NCE-FabricInsight, the solution implements full-lifecycle automation, lossless Ethernet, and network-wide intelligent O&M.
- Due to limited space, this course only briefly introduces the key features of the solution. The following courses will further explain the technical implementation principles and application scenarios of the solution.

Thank you.

把数字世界带入每个人、每个家庭。

每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



CloudFabric Data Center Network Planning and Design



Foreword

- Huawei's CloudFabric hyper-converged data center network (DCN) solution (CloudFabric solution for short) provides customers with intelligent, lossless, and ultra-broadband infrastructure networks. The solution supports one-click automatic deployment, AI-powered intelligent O&M, and on-demand self-service customization to quickly complete the planning, design, and deployment of industry DCN solutions.
- This course describes the DCN planning and design process of the CloudFabric solution, including DCN architecture design, underlay and overlay network design, network security design, and network management and O&M design.

Objectives

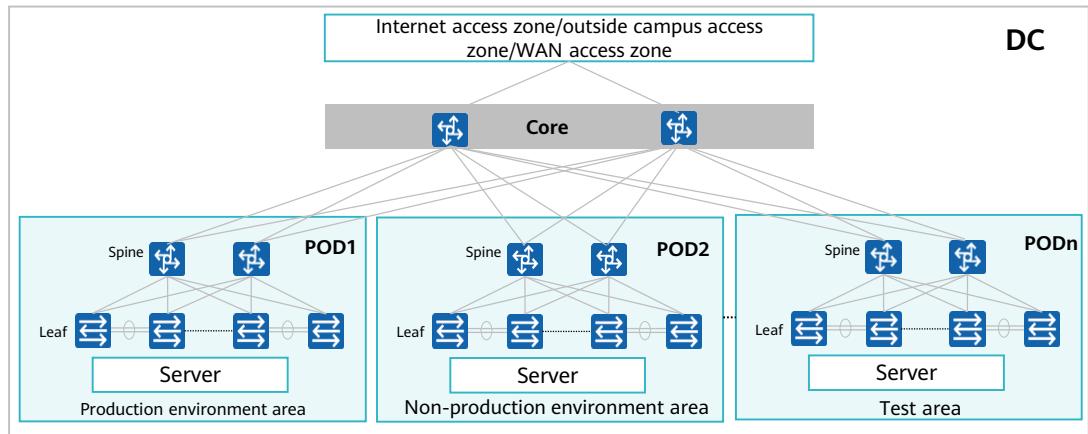
- On completion of this course, you will be able to:
 - Be familiar with the architecture of Huawei's CloudFabric solution.
 - Complete the underlay and overlay network design for a DCN based on actual requirements.
 - Complete the high reliability, network security, and network management and O&M design for a DCN based on actual requirements.

Contents

- 1. Data Center Network Overview**
2. Network Architecture Design and Data Planning
3. Underlay Network Design
4. Overlay Network Design
5. Network Security Design
6. Network Management and O&M Design

Typical Networking of The Data Center Network

- Shows the typical data center network networking.

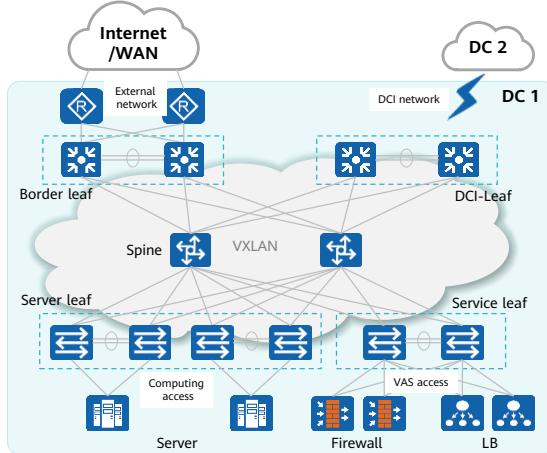


5 Huawei Confidential

HUAWEI

- Point of delivery (POD): A data center can be divided into one or more physical partitions to facilitate resource pooling and management. Each physical partition is called a POD. A POD is the basic deployment unit of a DC. Each DC can be deployed with multiple PODs, and a physical device can belong to only one POD. POD can be a standardized construction of equipment room modules based on POD or defined based on actual business requirements.
 - In a large data center, PODs can be defined based on the entire equipment room module.
 - A medium-sized data center can define a POD in the unit of two or more rows of cabinets.
 - In a small data center, multiple cabinets can be used to form a POD.

Data Center Network Architecture



- A DCN is an infrastructure for carrying DC services.
- Multiple DCNs can connect to branches of enterprises or organizations in different areas. In addition, DCNs can connect to the Internet or local area networks (LANs).
- The Spine-Leaf architecture is recommended for the underlay network.

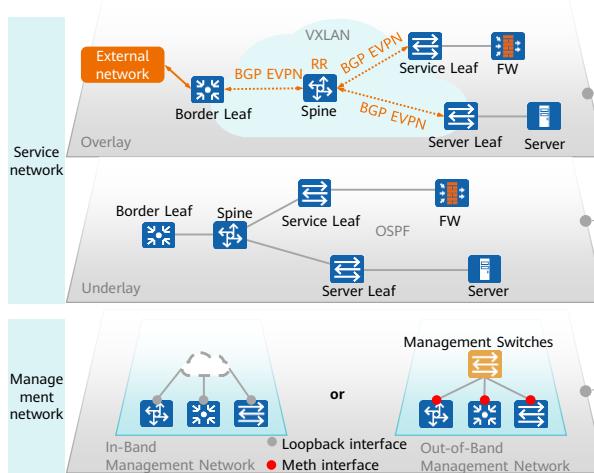
- The Spine-Leaf architecture is a new network architecture for a data center. It consists of spine nodes and leaf nodes. Spine nodes are backbone nodes and provide high-speed IP forwarding. A leaf node provides the network access function. In the standard Spine-Leaf architecture, leaf nodes are similar to line cards of modular switches and are responsible for receiving external traffic. Spine nodes are similar to the SFUs of modular switches and are responsible for traffic exchange between leaf nodes.
- Data Center Interconnect (DCI): Two data center network are interconnected to implement service interworking and service migration across data centers.

Physical Network Role

Roles	Function Description
Spine	A backbone node, which is the core node of the VXLAN fabric network and provides the high-speed IP forwarding function and connects to functional leaf nodes through high-speed interfaces.
Service Leaf	Leaf node, which provides Layer 4 to Layer 7 value-added services, such as firewall and load balance, to access the VXLAN fabric network.
Server Leaf	Leaf node, which provides computing resources, such as virtualized and non-virtualized servers, to access the VXLAN fabric network.
Border Leaf	Leaf node, which connects external traffic of the data center to the VXLAN fabric network of the data center and connects to external routers or transmission equipment.
DCI Leaf	Leaf node, which provides cross-DC service interworking and migration functions.

- DCI leaf nodes are also called DCI gateways or fabric gateways.

Network Layer



Bottom-up Design

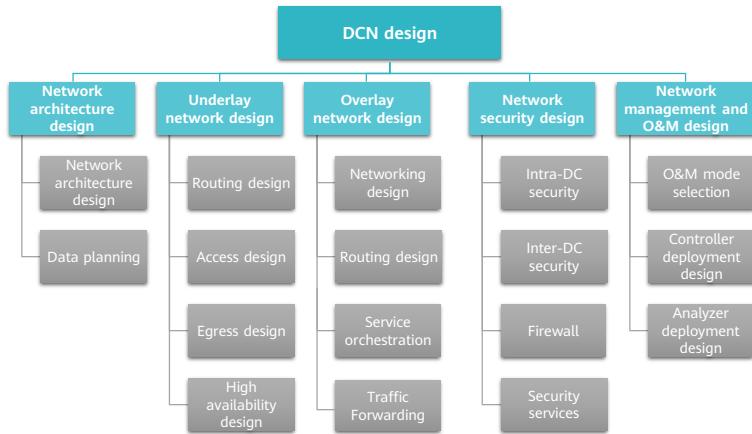
Overlay: A logical network established using the VXLAN protocol on the underlay network. Network resources are pooled through iMaster NCE. When creating a logical network in a VPC, you can invoke the network resources in the resource pool. A VPC usually represents a department or a service.

Underlay: A physical topology established by physical network devices, such as switches and routers, provides interconnection capabilities for all services in a data center and is the basic bearer network for service data forwarding in the data center.

Management network: Manages all physical devices on the service network. There are two types of management: in-band management and out-of-band management.

- Management network:
 - Inband management does not occupy service interfaces. Generally, loopback interfaces are used as management addresses and interwork with each other through the underlay network.
 - Out-of-band management: An independent management switch is configured to connect to the management interface (Meth interface) of the device to manage the network devices.

DCN Design Overview



9 Huawei Confidential

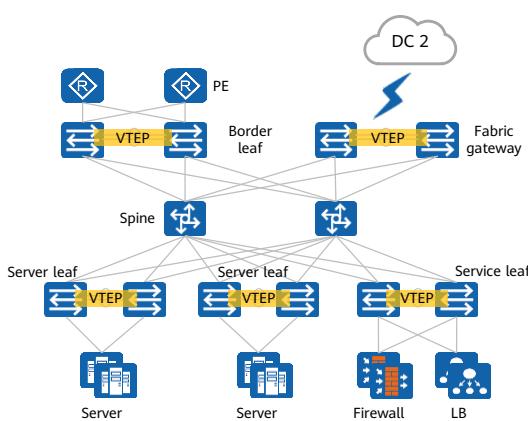


- Note: This course uses this solution design as an example to describe the DCN design process. The design and deployment parameters and device quantity involved in this course are examples. You can design a DCN based on actual service requirements.

Contents

1. Data Center Network Overview
- 2. Network Architecture Design and Data Planning**
3. Underlay Network Design
4. Overlay Network Design
5. Network Security Design
6. Network Management and O&M Design

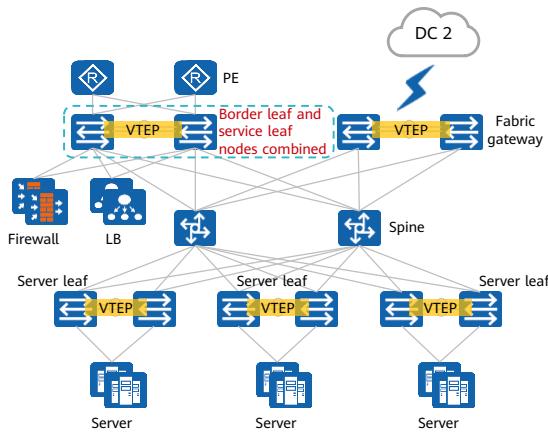
Standard Fabric Network Design



Standard fabric architecture and role separation solution

- The spine-leaf architecture is used. All roles are independently deployed and can be flexibly expanded. Spine and leaf nodes are fully meshed to form highly reliable redundant links.
- OSPF or EBGP is used to implement connectivity of the underlay network and VTEP address reachability, establish BGP EVPN peer relationships, and guide VXLAN packet forwarding.
- M-LAG is deployed on server leaf nodes and service leaf nodes to ensure access reliability, and active-active gateways are deployed on border leaf nodes to ensure reliability.
- Evaluate the DC scale and oversubscription ratio based on the number of access servers, interface bandwidth, and interface type, select proper switch models, and flexibly configure the numbers of spine and leaf nodes.

Converged Fabric Network Design



Combination of border Leaf and service leaf nodes

- **Combination design:**

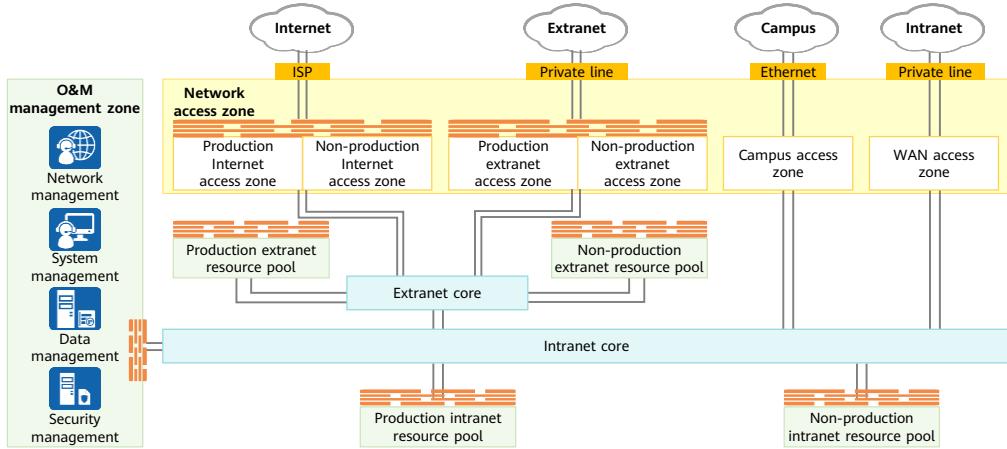
- Active-active gateways are deployed, which are configured as a DFS group to synchronize entries and are connected to VAS devices in an M-LAG.
- In the upstream direction, border leaf nodes are connected to PEs or core nodes of a DC in square looped or dual-homed Layer 3 networking. In the downstream direction, border leaf nodes are fully meshed to spine nodes.
- A dynamic routing protocol (OSPF or EBGP) or static routes run between border leaf nodes and core nodes (or PEs).

- In the converged network, only two roles can be deployed. In addition, three roles or even four roles can be deployed on the converged network. The investment depends on the deployment scenario scale and cost.
- Three-role integration: convergence of border leaf, service leaf, and spine nodes.
- Four roles are integrated: border leaf, service leaf, spine, and server leaf nodes.

Comparison Between Different Networking Solutions

Item	Standard Networking	Converged Networking
Fabric scale	Large	Large
Scalability	High Border leaf nodes, spine nodes, and VAS devices can be expanded independently.	Minor Border leaf nodes are scalable, but VAS resource scalability is poor.
Initial investment	Relatively high	Relatively low
Device selection requirements	Medium Resources such as the hardware ACL and routing table are distributed on devices of different roles. The requirement on spine nodes is lowered, and line-rate forwarding based on IP routes is required.	Relatively high There are high requirements for resources such as the hardware ACL and routing table. The requirement on spine nodes is lowered, and line-rate forwarding based on IP routes is required.
Application scenario	This architecture applies to large- and medium-sized fabrics, supporting about 100 server leaf nodes and 2000 physical servers. The north-south egresses have strong scalability and support four-active border leaf nodes. Scenarios that have strong VAS capacity expansion requirements are supported.	Border leaf and service leaf nodes are combined. The service configurations and forwarding plane resources for different device roles need to be deployed on one device, posing high requirements on device models, and the function scalability is low. This architecture does not support four-active border leaf nodes. Spine nodes are independently deployed. The scalability is not limited and four or more spine nodes can be deployed. The fabric supports large-scale server access. A single group of border leaf nodes or service leaf nodes supports a maximum of 6,000 VMs, and multiple groups of border leaf nodes or service leaf nodes are supported.

Case: Logical Zone Design for a DCN

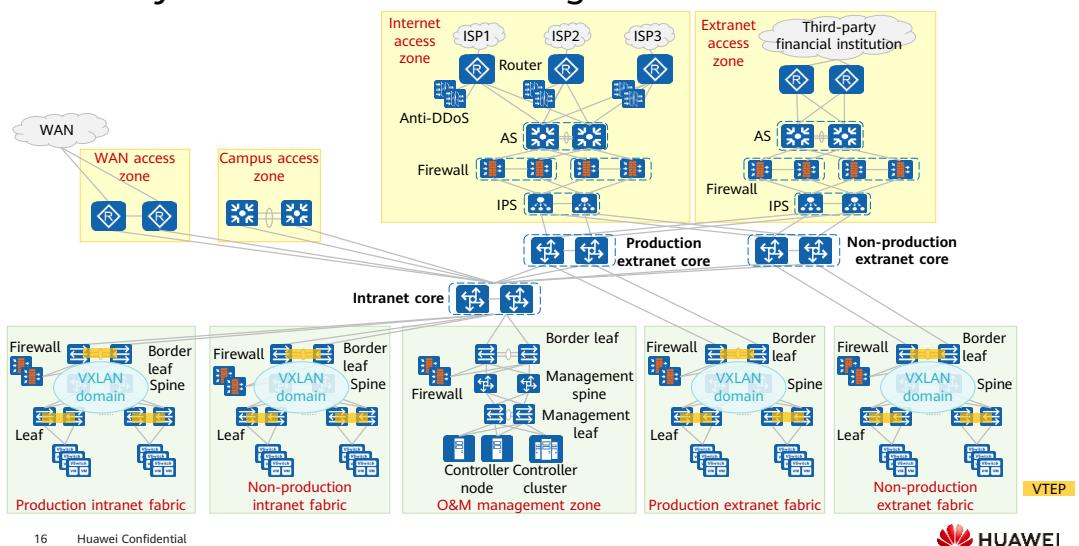


15 Huawei Confidential

HUAWEI

- Zone description:
 - The DCN consists of the production Internet access zone, non-production Internet access zone, production extranet access zone, non-production extranet access zone, campus access zone, WAN access zone, O&M management zone, production extranet zone, non-production extranet zone, production intranet zone, and non-production intranet zone, and core switching zone.
 - The out-of-band management network is deployed in each zone.
 - Firewalls are connected to the border of each zone in bypass mode for isolation.
- Note:
 - Out-of-band management: iMaster NCE-Fabric connects to the out-of-band management network ports on network devices through an independently deployed out-of-band management switch, and manages and controls the network devices through an independent out-of-band network.
 - In-band management: No independent management switch and network are configured. iMaster NCE-Fabric directly connects to the service network through a service switch, and manages and controls network devices through the underlay layer of the service network.

Case: Physical Architecture Design for a DCN

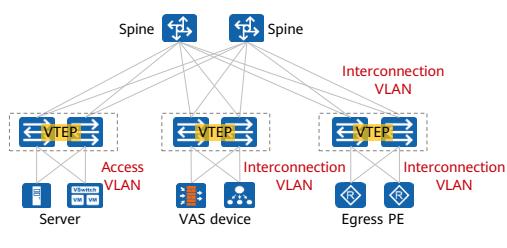


- In the DCN, VXLAN is deployed on the production intranet, non-production intranet, production extranet, and non-production extranet to build a fabric resource pool.
- Switch positioning:
 - The resource pool zone uses an architecture where border leaf and service leaf nodes are combined.
 - Extranet core switches are connected to the Internet access zone, extranet access zones, extranet resource pool zone, and DC core switches to control the advertisement of intranet routes.
- Firewall positioning:
 - Firewalls are deployed at the border of the resource pool zone and are connected to the border leaf nodes in bypass mode to perform access control on all traffic entering and leaving the zone. The cloud platform drives the SDN controller to automatically deliver the traffic diversion policy of the border firewalls of the resource pool zone.
 - Firewalls in the Internet access zone and extranet access zone meet the two-layer heterogeneous deployment requirements and perform access control on all traffic entering and leaving the access zones.

VLAN Planning

- The following VLANs need to be planned for the underlay network: interconnection VLANs between some devices, VLANs reserved for Layer 3 main interfaces, and default reserved VLANs of the system.
- The following VLANs need to be planned for the overlay network: access VLANs (VLANs for VMs and external networks to access the tenant network) of the tenant network and interconnection VLANs between gateways and VAS devices.

VLAN planning example for the underlay network:

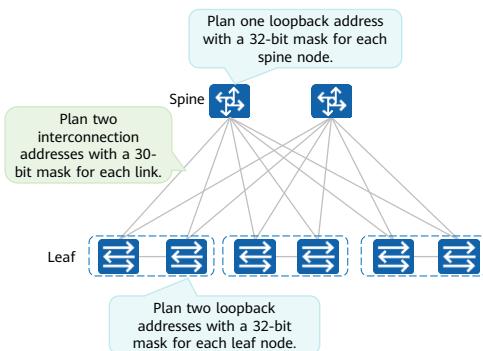


VLAN Type	VLAN Planning	Planning Suggestions
Device interconnection VLANs	2 to 30	Plan interconnection VLANs in advance based on the actual service design.
Reserved VLANs for Layer 3 main interfaces	4000 to 4062	<ul style="list-style-type: none"> Leaf node: Plan 16 VLANs, for example, VLANs 4047 to 4062. Spine node: Plan 63 VLANs, for example, VLANs 4000 to 4062. The reserved VLANs can be dynamically adjusted as required.
Default reserved VLANs	4064 to 4094	You are advised to retain the default value. The reserved VLAN range can be changed on a CE switch using CLI so that the default reserved VLAN range does not overlap with the planned or existing ones.

- VLAN planning for the underlay network:
 - Device interconnection VLANs: These VLANs provide VLANIF interfaces to establish links between some devices when an underlay network is manually constructed.
 - Reserved VLANs for Layer 3 main interfaces: For some CE series switches equipped with FD-X series cards, configure a reserved VLAN dedicated for Layer 3 main interfaces before switching the interface mode to Layer 3.
 - Default reserved VLANs: These VLANs are used as a channel of the internal control plane of a switch or a channel for transmitting user service data of some features.
 - Note: The number of VLANs required by the underlay network is relatively fixed.
- VLAN planning for the overlay network:
 - Note: The number of VLANs required by the overlay network is calculated based on the number of compute nodes and VAS devices.

IP Address Planning

- The IP addresses of the DCN are classified into service, management, and interconnection IP addresses.



- Service addresses: are the IP addresses of servers, hosts, and gateways.
 - It is recommended that gateway IP addresses use the same last digits, for example, gateways use IP addresses suffixed by .254.
 - The IP address range of each service must be clearly distinguished, and the IP addresses of each type of service terminals must be contiguous and can be summarized.
 - An IP address segment with a 24-bit mask is recommended.
- Management address: is the IP address configured for a loopback interface created on each Layer 3 network device.
 - A loopback address uses a 32-bit mask. A core device uses a smaller loopback address than other devices.
- Interconnection address: It is recommended that interconnection IP addresses use a 30-bit mask and core devices use a smaller host IP address.

- IP address planning principles:

- IP addresses must be managed and allocated uniformly on the entire network.
- IP address allocation should be simple and easy to manage, reflect network layers, simplify network management and network expansion, and be visualized.
- Extensibility of IP address planning must be ensured. That is, some IP addresses should be reserved at each layer so that IP addresses to be summarized can be contiguous during network expansion.
- IP addresses should be contiguous. The routes with contiguous addresses can be summarized easily on the hierarchical network. This reduces the routing table size and speeds up route calculation and route convergence.
- IP address allocation must be flexible to allow optimization of various traffic, security, and routing policies and make full use of the address space.

- IP address allocation principles:

- The network ID with the variable length and host address mask are used for IP address allocation. Some IP addresses need to be reserved based on the number of hosts on network segments. This ensures that IP addresses can be summarized and prevents the waste of IP addresses.
- To facilitate route summarization, assign IP addresses on the same network segment to devices in the same network area. If the preceding requirements cannot be met due to limitations, ensure that routes can be summarized at the aggregation layer.

VNI and VPC Planning

VXLAN Network Identifier (VNI) planning

VNI:

- Unique ID of a DC
 - Unique ID of an equipment room module in a DC
 - Unique ID of OpenStack in an equipment room module of a DC
- VNI range:
- Active DC: VNIs from 1000000 to 1999999, a total of 1 million VNIs
 - Intra-city DR DC: VNIs from 2000000 to 2999999, a total of 1 million VNIs
 - Remote DR cloud DC: VNIs from 3000000 to 3999999, a total of 1 million VNIs

VNI allocation example:

- DC ID (leftmost first digit): The value ranges from 1 to 9. It can be 1 (active DC), 2 (intra-city DR DC), 3 (remote DR DC), or 4 to 9 (reserved).
- Equipment room ID (leftmost second and third digits): The value is a decimal number of 100,000 or 10,000 digits. The value ranges from 01 to 99.
- OpenStack ID (leftmost fourth digit): The value ranges from 1 to 9.
- Subnet ID (leftmost fifth, sixth, and seventh digits): The value ranges from 001 to 999. That is, each OpenStack can use 999 subnets.

Virtual Private Cloud (VPC) planning

Fabric-VPC:

- VPC-Srv-DMZ: production extranet
- VPC-Mgt-DMZ: service assurance extranet
- VPC-Srv-Intranet: production intranet
- VPC-Mgt-Intranet: service assurance intranet
- VPC-Mgt: in-band management

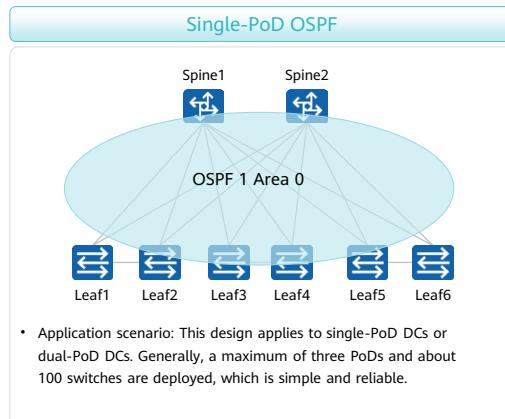
VNI allocation example: If the active DC equipment room is the first enabled cloud network equipment room module, the possible VNI in the fabric is **1011001**.

Contents

1. Data Center Network Overview
2. Network Architecture Design and Data Planning
- 3. Underlay Network Design**
4. Overlay Network Design
5. Network Security Design
6. Network Management and O&M Design

Underlay Routing Design: OSPF (1)

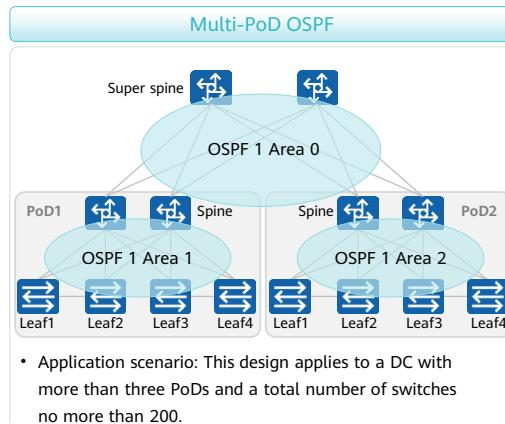
- OSPF is recommended on the underlay network if the total number of switches is less than 200.



- Solution design:**
 - Area division: All devices are planned in Area 0.
 - Configure the Loopback0 address as the VTEP IP address. Plan the same IP address for each group of active-active leaf nodes.
 - Configure a globally unique Loopback1 address for each device as the router ID.
 - Directly connect spine and leaf nodes through Layer 3 routed interfaces, and set the network type to P2P.
- Route optimization configuration:**
 - Configure BFD for OSPF to shorten the route convergence time.
 - Configure the OSPF route calculation interval and intervals for updating and receiving LSAs to optimize route convergence in case of faults.
 - Configure the period during which the maximum cost is retained after the interconnection interfaces between the spine and leaf nodes changes from Down to Up to optimize the switchback convergence performance.

Underlay Routing Design: OSPF (2)

- OSPF is recommended on the underlay network if the total number of switches is less than 200.



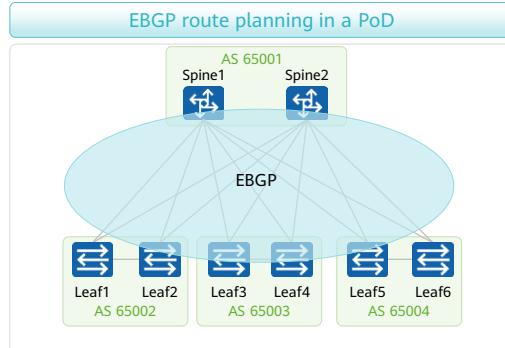
23 Huawei Confidential

HUAWEI

- Other planning description:
 - When planning the connections between spine nodes and super spine nodes, ensure that all nodes in OSPF Area 0 are reachable.

Underlay Routing Design: EBGP (1)

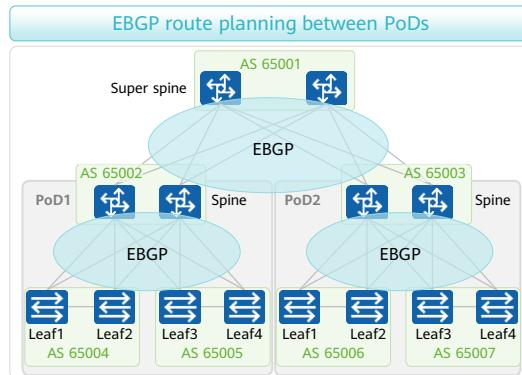
- EBGP is recommended on the underlay network if the total number of switches is greater than 200.
- Application scenario: This design applies to a DC with more than three PoDs and a total number of switches more than 200.



- Solution design:
 - AS partitioning: Deploy each group of active-active leaf nodes in an AS and spine nodes at the same layer in an AS.
 - Peer establishment: Use IP addresses of Layer 3 routed interfaces to establish EBGP peer relationships.
 - Route advertisement: Advertise loopback addresses.
 - Enable BGP load balancing to implement underlay load balancing.
- Route optimization configuration:
 - It is recommended that BFD for BGP be configured to shorten the route convergence time.
 - When the BGP peer relationship status changes from Down to Up, set the BGP route priority to the lowest to optimize the switchback convergence performance.

Underlay Routing Design: EBGP (2)

- EBGP is recommended on the underlay network if the total number of switches is greater than 200.
- Application scenario: This design applies to a DC with more than three PoDs and a total number of switches more than 200.



25 Huawei Confidential

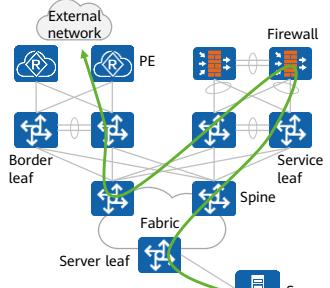
- Solution design:
 - AS partitioning: Deploy each group of active-active leaf nodes in an AS, spine nodes at the same layer in an AS, and the super spine node group in an AS.
 - Peer establishment: Fully mesh spine nodes with super spine nodes, and establish EBGP peer relationships through interconnection interface addresses.
 - In principle, a network segment is deployed in a PoD, and routes are summarized through spine nodes.
 - Use network segment routes for cross-PoD access to reduce the number of cross-PoD routes.

Comparison Between Routing Protocols on the Underlay Network

Item	OSPF	EBGP
Convergence speed	The convergence speed is fast.	The convergence speed is faster.
Protocol deployment	The protocol deployment is simple, but there are few control methods. The protocol depends on the cost and needs to be adjusted on the entire network.	The configuration is complex and various route control methods are available.
Network scale	Applicable to small- and medium-sized networks. OSPF has high calculation consumption and limited scalability.	Applicable to medium- and large-sized networks. BGP has low calculation consumption and good scalability.
Fault domain	The fault domain is large.	The routing domain is independent in each area, and the fault domain is controllable.
Application scenario	<ul style="list-style-type: none"> • Applicable to small- and medium-sized DCs and seldom used in large-sized DCs. • A single area is deployed for small- and medium-sized networks, multiple areas are deployed for large-sized networks with a three-layer architecture. • It is recommended that the number of router IDs be less than 200 and the number of OSPF neighbors be less than 100. • It is recommended that the number of neighbors in a single PoD be less than 100 to prevent a large routing domain from affecting network performance. 	<ul style="list-style-type: none"> • Applicable to large- and medium-sized DCNs. • Multiple PoDs and multi-layer spine nodes are deployed, and routes are transmitted between PoDs through EBGP. • It is recommended that the number of peers be less than 500. • It is recommended that the number of peers in a single PoD be less than 100 to prevent a large routing domain from affecting network performance.

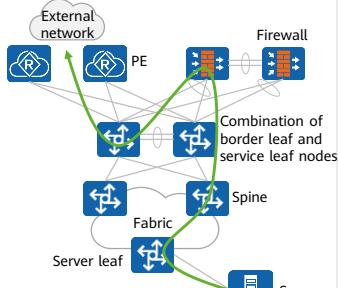
Firewall Access Design (1)

Firewalls connected to service leaf nodes in bypass mode



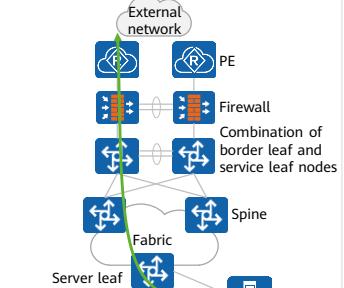
- This is a standard architecture and features high scalability. Multiple service leaf node groups are supported to connect to more VAS devices.
- This mode is recommended if load balancing needs to be performed among multiple border leaf nodes on the same egress network.

Firewalls connected to border leaf nodes in bypass mode



- This mode has low physical costs but poor scalability. It is a typical deployment mode for small- and medium-sized DCs.
- A physical device plays multiple roles, consuming more resources.

Firewalls connected to border leaf nodes in inline mode

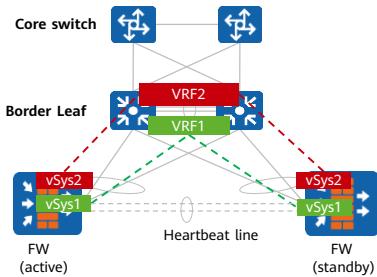


- External traffic must pass through the firewall, which applies to scenarios with high security requirements.
- The scalability of the firewall is poor. The DCN must keep stable for a certain period of time.

- The following factors must be considered for firewall access design:
 - Select the resource pool type (Huawei firewall, managed third-party firewall, or non-managed firewall) and network mode (inline or bypass) based on the firewall product model and customer service requirements.
 - Select the hardware device model and card type of the service leaf or border leaf node based on the firewall access bandwidth (10G or 40G).
 - Physically, both one-armed and two-armed connections are supported. The one-armed connection has advantages in terms of the cost (saving ports) and fault model (faults occur on both uplink and downlink logical links). As such, the one-armed mode is recommended.
 - Firewalls can be connected to service leaf nodes in bypass mode, to border leaf nodes in bypass mode, or between border leaf nodes and PEs in inline mode.

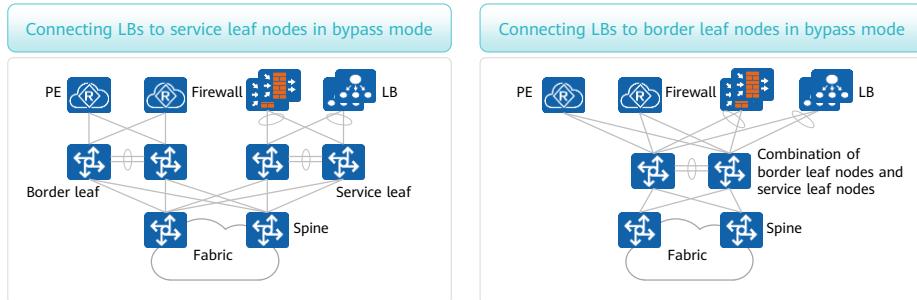
Firewall Access Design (2)

- Networking design:
 - Border Leaf (BL) and Service Leaf are co-deployed.
 - The firewall is connected to the BL node in bypass mode and logically connected between VRF1 and VRF2.
 - Firewalls are deployed in active/standby mirroring mode. Each firewall is connected to two blade servers through two 10GE ports. M-LAG is deployed on the BL to connect to the active and standby firewalls.
 - The firewall differentiates different service traffic through virtual systems.
 - Two 10GE links are deployed between active/standby firewalls as heartbeat synchronization links.
 - Traffic entering and leaving the fabric needs to pass through the firewall for security access control.
- Route design:
 - OSPF runs between the border leaf switch and core switch, and service VRF is used to isolate VPC routes.



SLB Access Design (1)

- Load balancing applications in DCs include server load balancing (SLB) and global server load balancing (GSLB). The former implements server load balancing within a DC, and the latter implements load balancing between DCs.

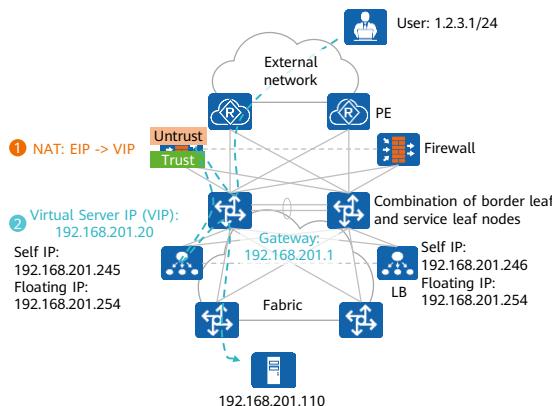


- LBs can be connected to border leaf nodes or service leaf nodes in bypass mode based on their deployment locations on the network.
- It is recommended that LBs be deployed in the same manner as firewalls.

- The LB is a key component of the high-availability network infrastructure. A cluster consisting of multiple servers replaces a single server to provide services externally. A large number of service requests are distributed to multiple servers, solving the high concurrency and high availability problems in the network architecture. In this way, resource usage is optimized, throughput is maximized, response time is minimized, and overload is avoided.
- The following factors must be considered for LB access design:
 - Select a proper LB model, working mode, and scheduling algorithm based on service characteristics and requirements.
 - Based on the interface bandwidth of the LB, select the hardware device model and card type of the service leaf or border leaf node.
 - One-armed and two-armed connection modes are supported. In actual deployment, LBs are generally connected in one-armed mode to reduce costs and improve reliability.

SLB Access Design (2)

- Solution design:



- LBs are deployed in active/standby mode. Heartbeat links are deployed between LBs.
- LBs are connected to border leaf nodes in bypass mode and connected to border leaf nodes through Eth-Trunks in M-LAG mode.
- LBs are connected in one-armed mode and connected to the VXLAN at Layer 2. The floating IP address, VIP, and server IP address are configured in the same network segment.
- The real server gateway is not the F5 floating IP address but the Layer 3 VXLAN gateway of the switch.
- The network needs to process gratuitous ARP packets and supports failover between floating IP addresses and VIPs.

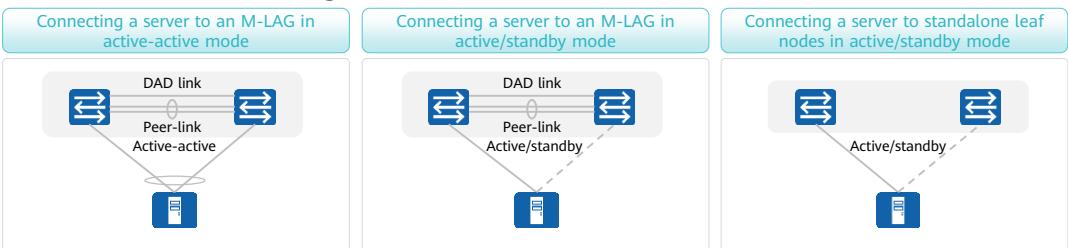
- Solution deployment description:

- In the cloud-network integration scenario, the LB management capability is provided by the cloud platform of each vendor. You need to query the product capability of each vendor.
- For details about the LB automation capability in the network virtualization scenario, see the controller product documentation.
- The LB can be configured with SNAT or configured as the server gateway. The specific solution is determined by the service orchestration of the LB/server and is not limited by the CloudFabric solution.
- The floating IP addresses and service VIPs of LBs and server IP address can be in the same subnet or different subnets. You are advised to deploy them in the same subnet. In this case, you do not need to configure a static route destined for a service VIP on a switch.

- NAT Server load balancing:

- The client sends a request to the load balancing device at the front end of the server cluster. The virtual service on the load balancing device receives the request, selects a real server based on the scheduling algorithm, translates the destination address of the request packet to the address of the selected real server, and sends the request to the real server.
- The real server sends a response packet to the load balancing device, which changes the source IP address in the response packet to the VIP, and then forwards the response packet to the user.

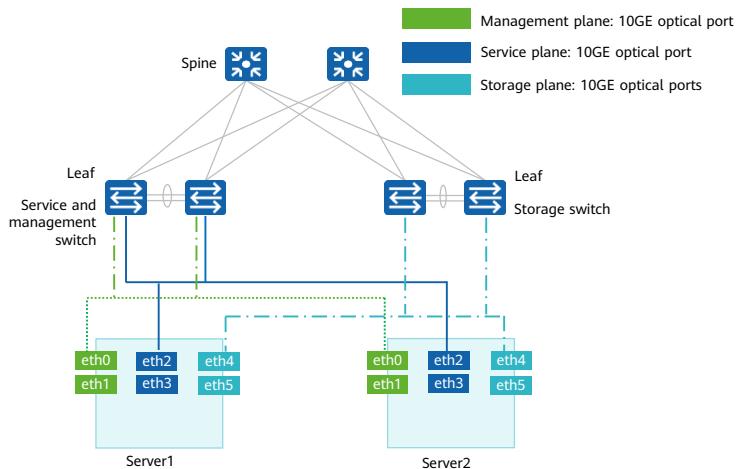
Server Access Design (1)



	Standalone Switches	Switches in an M-LAG
Server access mode	Active/standby	Active/standby or active-active
Availability	High. The two switches are deployed independently and faults are isolated.	High. Control planes are independent and fault domains are isolated.
Cost	The cost is low. No cable needs to be deployed between switches.	The cost is moderate. Peer-links and heartbeat links need to be deployed.
Version upgrade	The two switches are upgraded independently, without interrupting services. Upgrade risks are low.	The two switches are upgraded independently, without interrupting services. Upgrade risks are low.
Application scenario	Active and standby NICs of the server are connected.	Active/standby and load balancing access modes are deployed on the network, and the access solution on the entire network is unified. This mode applies to M-LAG.

- Consider the following factors when selecting models of and designing server leaf nodes:
 - Select an access mode. Servers often use M-LAG, stacking, and standalone modes. M-LAG active-active deployment is recommended because it can ensure service continuity during the upgrade of access switches.
 - Select server leaf nodes (hardware devices) based on the server access bandwidth (10GE/25GE access) and the ratio of server leaf nodes' uplink bandwidth to spine nodes' downlink bandwidth.
 - Determine the number of server leaf nodes based on the number of servers.
 - Select the model of server leaf nodes depending on whether microsegmentation or IPv6 deployment or evolution towards them is required.

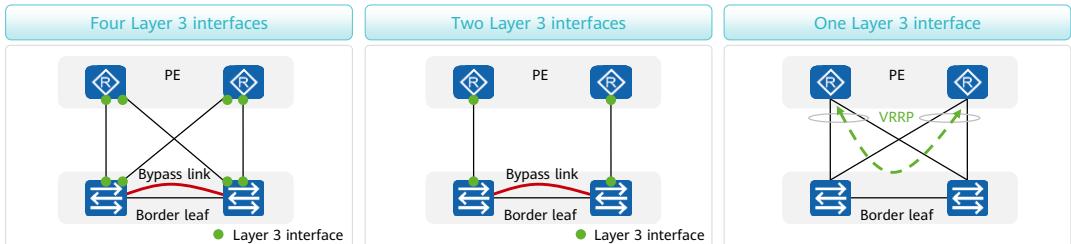
Server Access Design (2)



- Design scheme:
 - Servers in adjacent cabinets share two groups of leaf switches, which are connected to the service and management NICs and storage NICs of the servers.
 - The two leaf switches connect to the spine switches in the uplink.
 - M-LAG is deployed between leaf nodes to connect to servers. Two interfaces are used as peer-links. Each group of leaf nodes uses four uplink interfaces to connect to two spine switches.
 - Server SAN storage servers are deployed in storage node cabinets. The traffic between Server SAN nodes is heavy. It is recommended that storage nodes be connected to independent storage switches.
 - The management plane and storage plane need to be configured using iMaster NCE (Fabric) in advance. Therefore, the network administrator needs to plan access ports for the service plane and storage plane and connect storage NICs to storage switches.
- Server NIC planes are divided into the management plane (carrying management traffic), service plane (carrying service traffic), and storage plane (carrying storage traffic).
- In normal cases, a single node and different planes need to be dual-homed to access switches to ensure reliability.

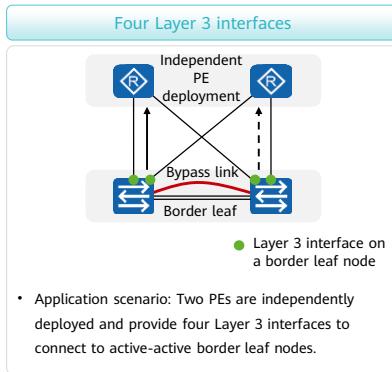
Egress Design Overview

- The egress network of a DC refers to the connection and configuration between border leaf nodes and egress PEs. Border leaf nodes and PEs can be connected in multiple modes. Select a connection mode based on the customer's existing border conditions.



- Physical networking design:**
 - It is recommended that PEs be deployed in a two-node cluster to ensure reliability.
 - Deploy border leaf nodes as active-active gateways in an M-LAG. In some scenarios, bypass links need to be deployed between border leaf nodes.
 - The interconnection topology between border leaf nodes and PEs can be square-shaped (two PEs have at least two physical ports) or dual-homed (two PEs have at least four physical ports), depending on the number of ports provided by the PEs. The dual-homed topology is recommended.
- Interconnection interface design:**
 - PEs can be connected to border leaf nodes through one, two, or four Layer 3 interfaces. It is recommended that PEs be connected to border leaf nodes through four Layer 3 interfaces.
 - Supported Layer 3 interfaces include VBDIF and VLANIF interfaces. (For details, see the device model and interconnection scenario.)
- Egress routing design:**
 - Border leaf nodes and PEs can interwork through dynamic or static routes. It is recommended that external routes be summarized and default routes be advertised within the DC.
- Note:** If four Layer 3 interfaces are used, a border leaf node group provides four Layer 3 interfaces (physical or logical interfaces) to connect to PEs.

Connecting Border Leaf Nodes to PEs Through Four Layer 3 Interfaces



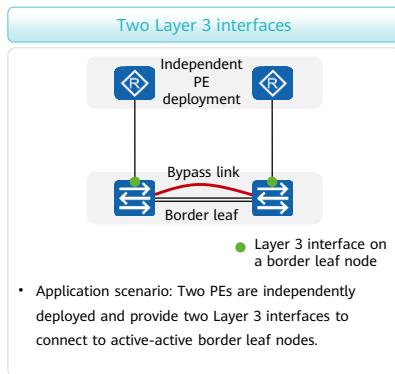
- Physical networking design:

- Four interconnection links form a dual-homed topology. Four independent Layer 3 interfaces need to be configured on the two PEs. If there are multiple cards, it is recommended that links be deployed across cards.
- A Layer 3 bypass link can be deployed.
- The M-LAG peer-link has at least two member links across cards to ensure reliability and bandwidth, and the member link cannot be configured as the bypass link.

- Solution description:

- A Layer 3 interface is configured on a border leaf node to connect to a Layer 3 interface on a PE.
- Dynamic routing protocols and static routes can be deployed between border leaf nodes and PEs. You are advised to deploy a dynamic routing protocol, for example, BGP.
- Fast switchover: Associate statics routes with NQA or dynamic routing protocols with BFD to detect the peer PE status, accelerating route convergence.
- Route convergence: Configure a delay for the interface connecting the border leaf node to the PE to go Up and a delay for advertising routes when the interface goes Up from Down to optimize the switchback performance.
- When border leaf nodes and spine nodes are deployed independently and there are only a few interconnection interfaces between them, a Monitor Link group needs to be deployed to associate the interfaces connecting border leaf nodes to spine nodes with the interfaces connecting the border leaf nodes to firewalls/LBs and PEs, preventing service interruption caused by multi-link faults.
- Deploy a dynamic routing protocol for the bypass link. The two border leaf nodes then can advertise egress routing information to each other for egress link protection.

Connecting Border Leaf Nodes to PEs Through Two Layer 3 Interfaces



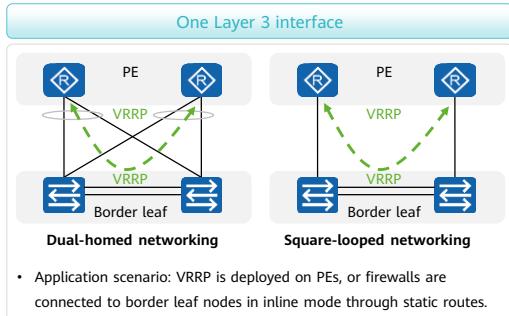
- Physical networking design:

- A square-looped topology is formed. If there are multiple cards, it is recommended that links be deployed across cards.
- A Layer 3 bypass link must be deployed.
- The M-LAG peer-link has at least two member links across cards to ensure reliability and bandwidth, and the member link cannot be configured as the bypass link.

- Solution description:

- A Layer 3 interface is configured on a border leaf node to connect to a Layer 3 interface on a PE.
- Dynamic routing protocols and static routes can be deployed between border leaf nodes and PEs. You are advised to deploy a dynamic routing protocol, for example, BGP.
- Fast switchover: Associate statics routes with NQA or dynamic routing protocols with BFD to detect the peer PE status, accelerating route convergence.
- Route convergence: Configure a delay for the interface connecting the border leaf node to the PE to go Up and a delay for advertising routes when the interface goes Up from Down to optimize the switchback performance.
- When border leaf nodes and spine nodes are deployed independently and there are only a few interconnection interfaces between them, a Monitor Link group needs to be deployed to associate the interfaces connecting border leaf nodes to spine nodes with the interfaces connecting the border leaf nodes to firewalls/LBs and PEs, preventing service interruption caused by multi-link faults.
- Deploy a dynamic routing protocol for the bypass link. The two border leaf nodes then can advertise egress routing information to each other for egress link protection. Fast Reroute (FRR) can be configured to improve the fault convergence performance.

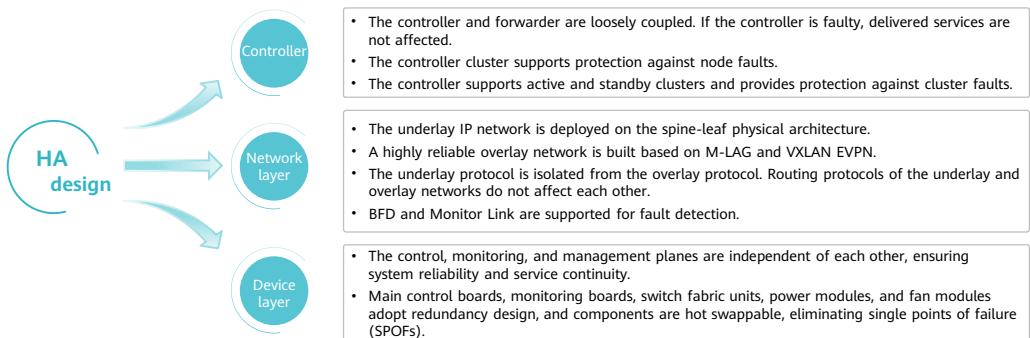
Connecting Border Leaf Nodes to PEs Through One Layer 3 Interface



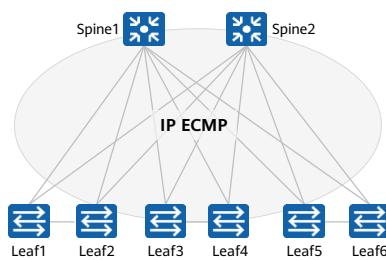
- Solution description:
 - Deploy Virtual Router Redundancy Protocol (VRRP) on the two PEs, and configure the same virtual IP address for them.
 - Deploy the border leaf nodes as an M-LAG to connect to PEs and configure the same IP address for the border leaf nodes.
 - Configure static routes between PEs and border leaf nodes to implement connectivity. The next hop of the static route configured on the border leaf node is the VRRP address of the PEs.
 - Fast switchover: Associate static routes with NQA to detect the peer PE status, accelerating route convergence.
 - Convergence optimization: Configure a delay for interconnection interfaces connecting border leaf nodes to PEs to go Up to optimize the switchback performance.
 - When border leaf nodes and spine nodes are deployed independently and there are only a few interconnection interfaces between them, a Monitor Link group needs to be deployed to associate the interfaces connecting border leaf nodes to spine nodes with the interfaces connecting the border leaf nodes to firewalls/LBs and PEs, preventing service interruption caused by multi-link faults.

HA Design Overview

- As the core department of the customer's IT infrastructure, the DC stores various data, runs a variety of services, and provides services to external networks. DC faults will cause great loss in every year. Therefore, stable and reliable running of DCs is critical.
- High availability (HA) design can be divided into three levels:



Spine HA Design

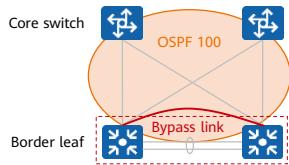


- Link redundancy: Spine nodes are fully meshed with all leaf nodes to form a full-mesh architecture.
- Device redundancy: Multiple high-density DC switches are deployed to implement device-level reliability. In a typical deployment scenario, two or four spine nodes are deployed. The number of deployed spine nodes depends on the live network scale.
- Network redundancy:
 - In the spine-leaf networking architecture, multiple spine nodes are deployed to construct an IP ECMP load balancing network, implementing network-level reliability.
 - Routing protocols (BGP/OSPF) are deployed to achieve a highly reliable architecture.

- Fault detection and tolerance design:
 - When the link of a spine node fails, leaf nodes quickly switch traffic to a normal link through ECMP routes on the underlay network.
 - If a spine node fails, leaf nodes quickly switch traffic to other spine nodes through ECMP routes on the underlay network.

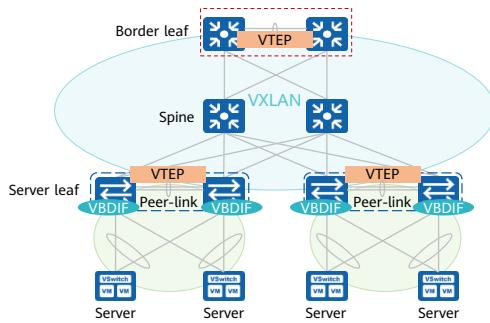
Border Leaf HA Design

- Link redundancy:
 - It is recommended that border leaf nodes be connected to different core switches and spine nodes, or to the same core switch and spine node through multiple links across cards.
 - At least two links are deployed between border leaf nodes as an Eth-Trunk (used as a peer-link). It is recommended that the links be deployed across cards.
- Device redundancy: Active-active device groups are deployed on border leaf nodes to implement device-level reliability.
- Network redundancy:
 - It is recommended that border leaf nodes be dual-homed to core switches and be fully meshed with spine nodes to build an IP ECMP network, eliminating Layer 3 loops.
 - Link aggregation protocols (LAG or M-LAG) and routing protocols (BGP or OSPF) are used to ensure a high-reliable architecture.



- Fault detection and tolerance design:
 - A bypass link is deployed between border leaf nodes to prevent traffic interruptions caused by faults of all uplink interfaces. The bandwidth must be at least equal to the total bandwidth of the uplinks of a single device. (Optional. The bypass link is required in square-looped topology.)

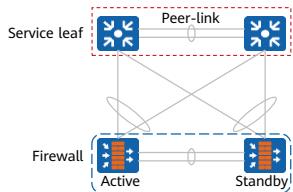
Server Leaf HA Design



- Link redundancy: At least two links are deployed between server leaf nodes as an Eth-Trunk (used as a peer-link). It is recommended that the links be deployed across cards.
- Device redundancy: Server leaf nodes are deployed as an active-active device group in an M-LAG to implement device-level reliability.
- Network redundancy:
 - Server leaf nodes are fully meshed with spine nodes to form an IP ECMP network, eliminating Layer 3 loops.
 - Link aggregation protocols (LAG or M-LAG) and routing protocols (BGP or OSPF) are used to ensure a high-reliable architecture.

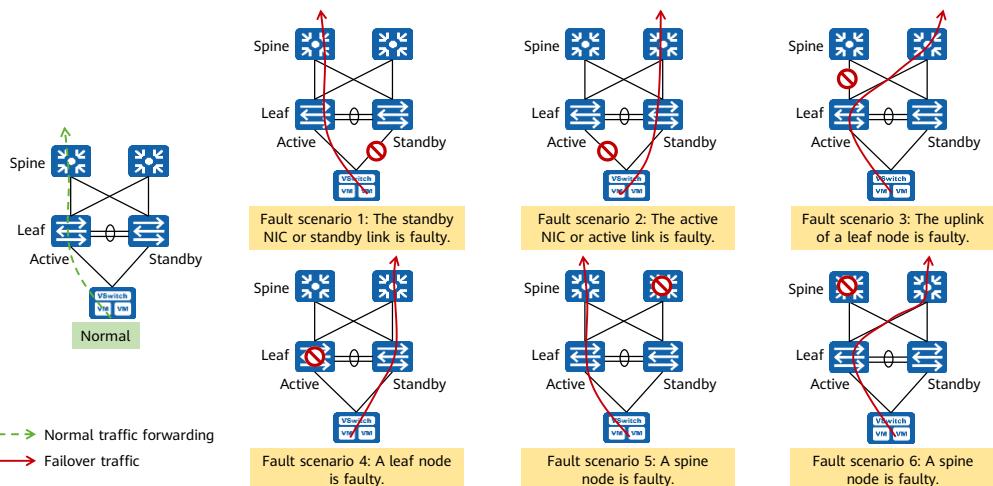
- Fault detection and tolerance design:
 - If possible, configure dual-active detection (DAD) based on out-of-band management network ports. Otherwise, configure DAD based on service network ports.
 - Deploy a Monitor Link group. If all uplinks fail, the associated downlink goes Down, preventing traffic interruptions.
 - Configure broadcast, unknown unicast, and multicast packets (BUM packets) storm suppression on the downlink ports of leaf nodes. When VLAN 1 is not used, packets from this VLAN are denied to prevent loops.

Firewall HA Design



- Link redundancy: Each firewall is dual-homed to the active-active service leaf device group in an M-LAG through link bundling.
- Device redundancy: Firewalls are deployed in active/standby mode and support hot standby.

Examples of HA Design Fault Scenarios

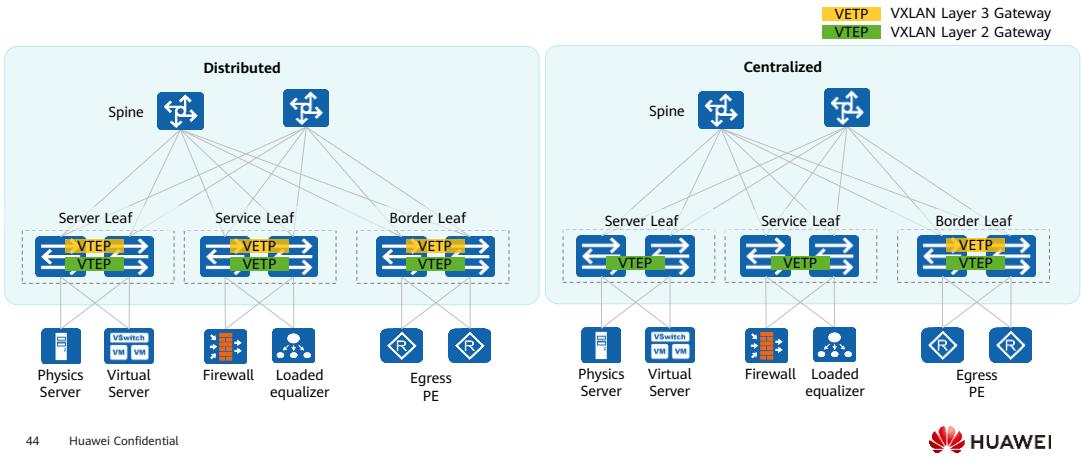


Contents

1. Data Center Network Overview
2. Network Architecture Design and Data Planning
3. Underlay Network Design
- 4. Overlay Network Design**
5. Network Security Design
6. Network Management and O&M Design

Overlay Networking Design (1)

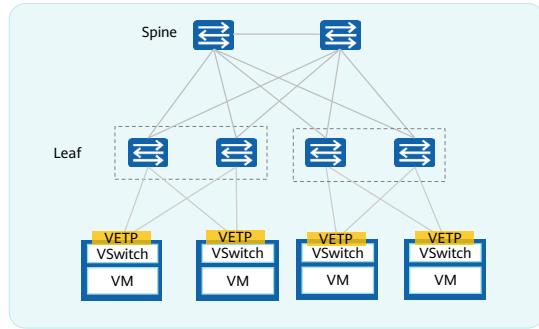
- There are three types of overlay networks. Network overlay indicates that the two endpoints of a VXLAN tunnel are physical switches. Network overlay networking modes are classified into centralized network overlay and distributed network overlay.



- Centralized network overlay: VXLAN Layer 3 gateways are deployed in a centralized manner, and leaf nodes function only as VXLAN Layer 2 bridges.
- Distributed network overlay: Leaf nodes function as VXLAN Layer 3 gateways, and spine nodes are used only for IP forwarding.

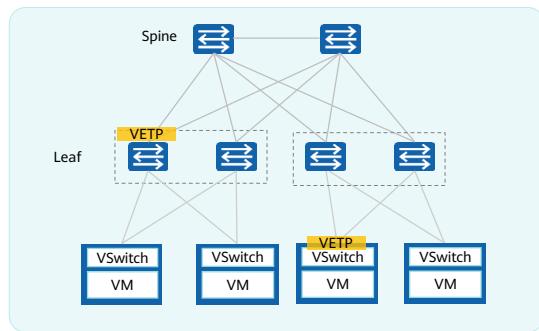
Overlay Networking Design (2)

- In host overlay mode, all VXLAN tunnel endpoints are deployed on software switches (installed on servers). That is, the start and end points of VXLAN tunnels are software switches.

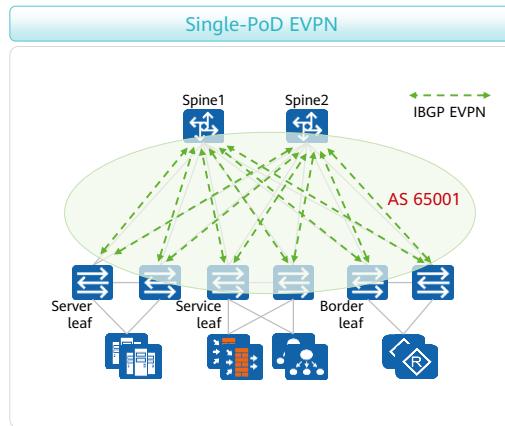


Overlay Networking Design (3)

- In hybrid overlay networking, VXLAN tunnel endpoints are deployed on hardware and software switches. That is, VXLAN tunnel start and end points are both hardware and software.



Overlay Routing Design: EVPN Deployment



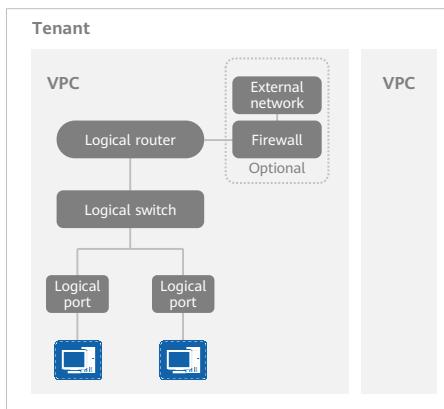
- Solution design:
 - Distributed VXLAN EVPN gateways are deployed. Each leaf node functions as the Layer 2 gateway and Layer 3 gateway, and traffic is forwarded along the shortest path.
 - IBGP EVPN is deployed, and loopback addresses are used to establish IBGP peer relationships.
 - Configuring IBGP RRs reduces the number of fully meshed connections between IBGP peers, simplifying configurations, and reducing device resource consumption. It is recommended that spine nodes be used as RRs.

- Other planning description:

- Run the **undo policy vpn-target** command on RRs to disable VPN target-based filtering for VPN routes or label blocks.
- Configure a delay for the interface connecting the border leaf node to the PE to go Up to optimize the traffic switchback performance.

Service Model: Tenant Service Model (1)

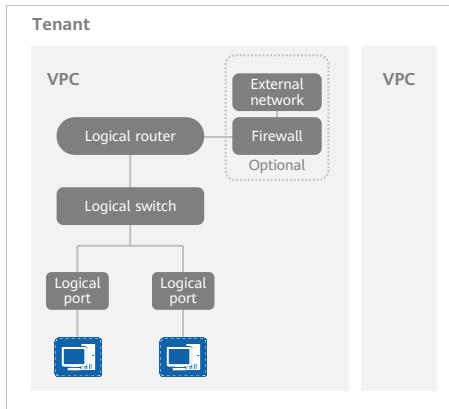
- Understanding the terminology, background, and implementation principles associated with service provisioning is helpful for quickly mastering tenant network interconnection skills in a computing scenario.



- Tenant:** is the minimum unit for enterprise service management.
- Virtual Private Cloud (VPC):** provides secure and reliable information processing, storage, and transmission services to tenants through the virtualization and encryption technologies based on network, storage, and compute resources. Multiple VPCs can be created for a tenant based on service requirements.
- Logical router:** is virtualized by a network device where virtualization software is running, and is connected to VMs on different networks, so that VMs can communicate with each other on a Layer 3 network. One network device can be virtualized into multiple logical routers for different tenants.
- Logical switch:** connects to different VMs to ensure that the VMs can communicate with each other at Layer 2. One network device can be virtualized into multiple logical switches for different tenants.

- One network device can be virtualized into multiple logical routers for different tenants. Multiple tenants can share a network device. For each tenant, a logical router functions as an independent and real router with independent hardware and software resources and running space. Services on different logical routers do not affect each other. In terms of experience, there is no difference between a logical router and a real router.
- One network device can be virtualized into multiple logical switches for different tenants. Multiple tenants can share a network device. For each tenant, a logical switch functions as an independent and real switch with independent software and hardware resources and running space. Services on different logical switches do not affect each other. In terms of experience, there is no difference between a logical switch and a real switch.

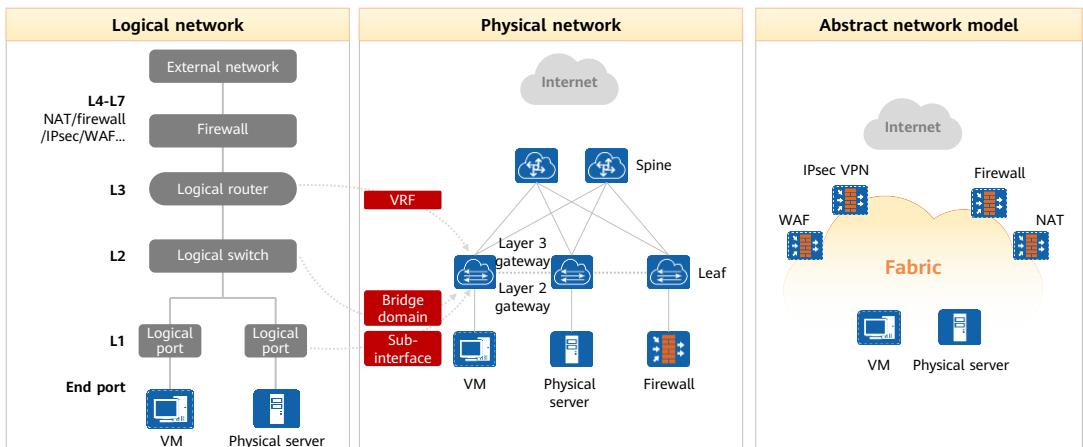
Service Model: Tenant Service Model (2)



- **Logical port:** functions as an access point for VMs to access the network. One physical port on a network device can be virtualized into multiple logical ports for different tenants. For each tenant, a logical port functions as an independent and real port.
- **External network:** networks outside the tenant's management, such as Internet or other tenant networks connected through VPNs.
- **Firewall:** The firewall function is provided by a physical firewall or virtual firewall.
- **VM:** virtual machine.

- Located at the border of a network, a firewall implements secure access control between the external network and internal network, which enhances the network protection capability. It protects service data flows between the Untrust and Trust zones based on 5-tuple information. It can also be used for access control between subnets. You can choose whether to deploy firewalls based on whether the tenant needs to access an external network. For security purposes, deploy a firewall when a tenant is connected to an external network.
- In the computing scenario, VMs are provisioned by the VMM connected to iMaster NCE-Fabric. The VMM manages compute resources, and iMaster NCE-Fabric manages network resources.

Relationship Between a Physical Network and a Logical Network

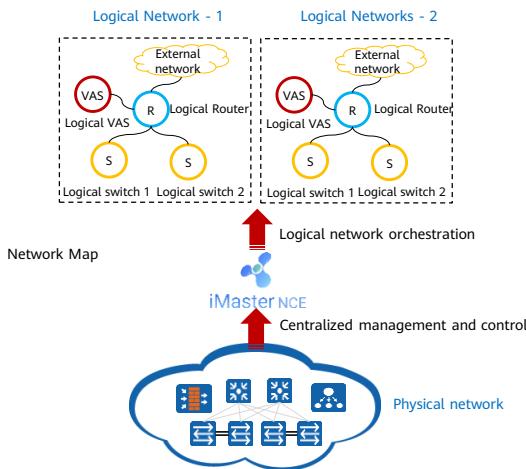


50 Huawei Confidential

HUAWEI

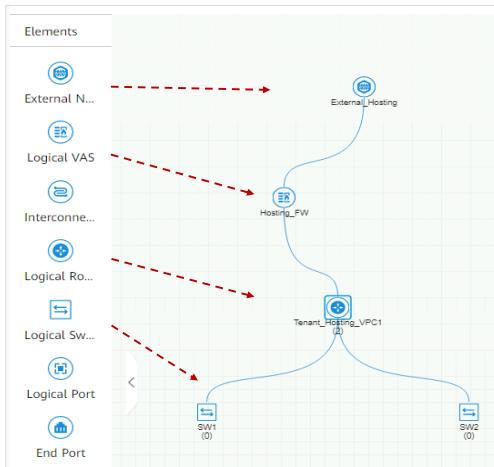
- A fabric network is a logical network constructed based on the VXLAN technology and provides services such as VM access, physical machine access, VPN access, Internet access, server load balancing, IP address translation, and ACL-based packet filtering. Tenants can focus on services and construct logical networks as required using services provided by a fabric network. You need to consider the following factors when constructing a tenant network:
 - Number of VPCs planned based on service types or security requirements.
 - Security policies planned for access authorization between VPCs.
 - Subnets planned for each VPC.
 - Routes planned for communication between subnets.
 - Resources such as VNIs and BDs allocated for each subnet.
- A logical network provides the following services based on the fabric network:
 - Logical port: Logical ports are located at the bottom of a logical network and provide access to the VXLAN network from VMs, physical machines, NAT devices, IPsec VPNs, firewalls, and WAFs.
 - Logical switch: Logical switches are located at the second layer of a logical network and provide the network switching service between logical ports.
 - Logical router: Logical routers are located at the third layer of a logical network and provide the network route service between logical ports.
 - NAT devices, IPsec VPNs, firewalls, and WAFs: They are located at the layer 4 to layer 7 of a logical network and provide advanced services.

Manually Orchestrating Logical Networks (1)



- Huawei iMaster NCE (Fabric) uses a logical model to define networks and divide multiple independent logical networks based on the physical network to virtualize network functions.
- Implementation principle:
 - A physical network is divided into multiple logical networks by configuring VRF/BD features.
 - On the Agile Controller-Campus, logical networks are created based on the logical network model that network engineers can understand and automatically map the logical networks to the network features such as VRF and BD on the physical network. In this way, the Agile Controller-Campus centrally manages and controls the switches on the physical network.

Manually Orchestrating Logical Networks (2)



53 Huawei Confidential

Solution Advantages

- The drag-and-drop configuration is intuitive and visible. Each operation step is guided by configuration, and the operation interface is user-friendly.

Application Scenario

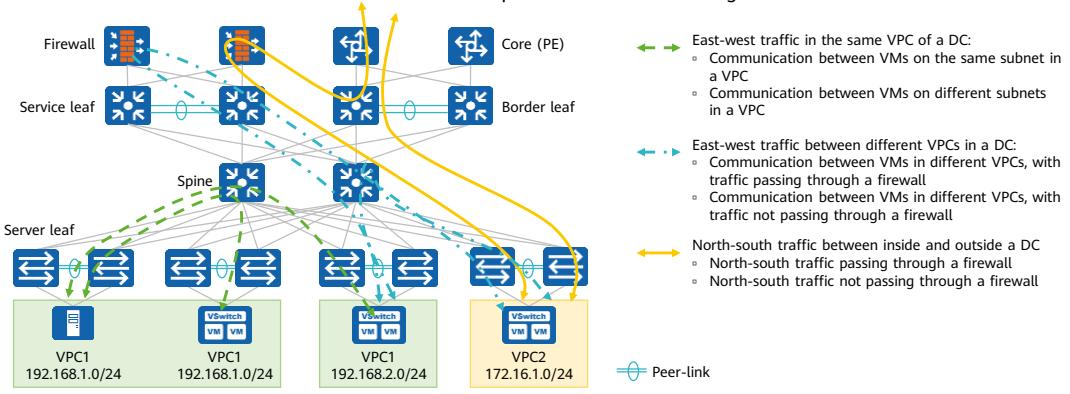
- This mode is applicable to users who are not familiar with configuration operations or manual configuration scenarios with small service scale.



- Create a single service VPC. In the VPC view, drag different logical units to complete network deployment.

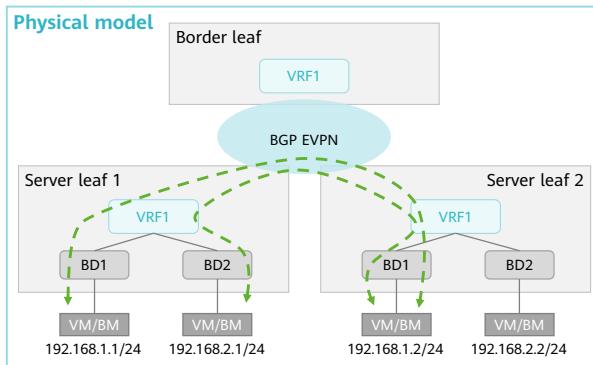
Overlay Forwarding Design Overview

- A DCN has various communication traffic, for example, north-south traffic between inside and outside the DC, and east-west traffic within the DC. Such traffic may or may not pass through the firewall. In these scenarios, the controller can be used for unified orchestration to implement traffic forwarding.



Intra-VPC Communication

- East-west traffic in a VPC is classified into intra-subnet Layer 2 traffic and inter-subnet Layer 3 traffic.

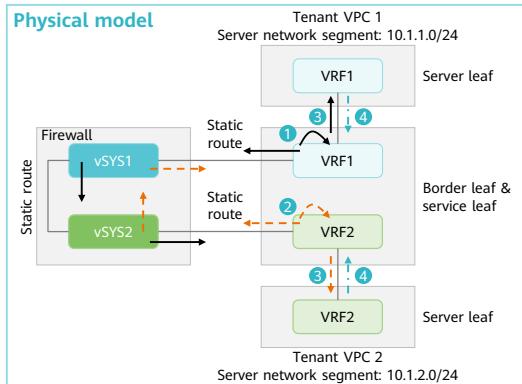


- Intra-subnet communication in a VPC:
 - Create a BD, associate it with an L2VNI, and create an EVPN instance. Then, create an access interface to associate the Layer 2 sub-interface with the BD.
 - The local leaf node learns the MAC address of the host and generates a MAC route. Then, the route is advertised to the remote leaf node through BGP EVPN, and the remote leaf node receives the route through route target (RT)-based route leaking.
- Inter-subnet communication in a VPC:
 - In addition to the preceding operations, you need to create a VBDIF interface for the gateway and associate the VBDIF interface with a VRF and an L3VNI.
 - After learning the host ARP entry and generating an IRB route, the local leaf node advertises the route to the remote leaf node through the BGP EVPN peer relationship. The remote leaf node receives the route through RT-based route leaking.

- From the perspective of the logical network, intra-VPC communication services are orchestrated as follows:
 - For communication within a subnet in a VPC, use the SDN controller to create a tenant and a VPC, and then create logical switches, logical ports, and end ports.
 - For communication across subnets in a VPC, associate different logical switches with the same logical router in addition to the preceding operations.

Inter-VPC Communication

- A VPC dynamically divides a physical network into logical network resource domains, including logical networks and logical VASs. Access between VPCs can be implemented through firewalls at two sides, a firewall at one side, or no firewall based on security access control requirements, which can be flexibly orchestrated on the SDN controller.



56 Huawei Confidential

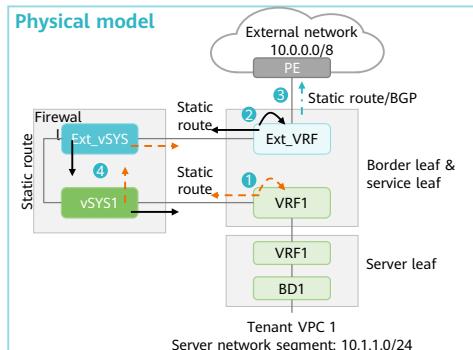


- If traffic does not pass through a firewall:

1. In VRF1 on the border leaf node, configure a static route to 10.1.2.0/24 with the next hop being the IP address of VRF2, import the route to BGP VPN-Instance VRF1, and advertise the route to EVPN.
2. In VRF2 on the border leaf node, configure a static route to 10.1.1.0/24 with the next hop being the IP address of VRF1, import the route to BGP VPN-Instance VRF2, and advertise the route to EVPN.
3. The border leaf node sends static routes destined for 10.1.2.0/24 and 10.1.1.0/24 to the server leaf nodes through BGP EVPN. Each server leaf node selects a route based on the VPN RT value. The RT value varies depending on the VPN.
4. The server leaf node sends the host routes destined for 10.1.1.1 and 10.1.2.1 to the border leaf node through BGP EVPN. The border leaf node selects a route based on the VPN RT value. The RT value varies depending on the VPN.

Communication Between a VPC and an External Network

- When an SDN network needs to communicate with an external network, you need to create an external network on the controller and associate it with an external gateway. If traffic needs to pass through a firewall, you need to create a logical VAS and associate the logical router and logical VAS with the external network to enable north-south service traffic.



57 Huawei Confidential

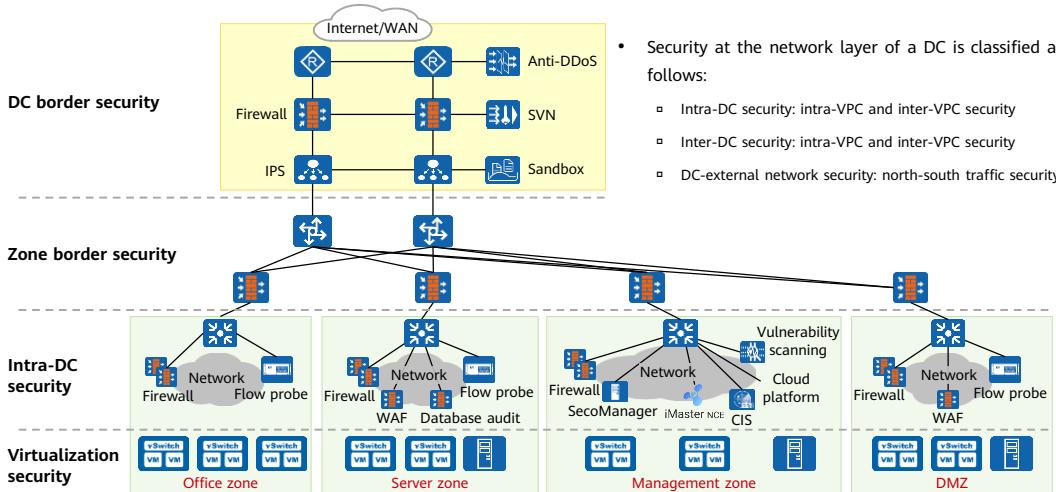
HUAWEI

- When traffic passes through firewalls at two sides and the firewalls are connected to border leaf nodes (combined with server leaf nodes) or service leaf nodes in bypass mode:
 - Deliver a static route destined for the external network with the next hop being the firewall interconnection IP address to the tenant VRF (VRF1) on the border leaf node. Import the static route to the tenant VRF and advertise the route to the server leaf node through BGP EVPN.
 - On the border leaf node, create an external gateway egress VRF (Ext_VRF), and configure a static route pointing to the network segment of a VM or server with the next hop being the firewall interconnection address. If the border leaf and service leaf nodes are deployed independently, the static route needs to be imported to the egress VRF and advertised to the border leaf node through BGP EVPN.
 - Static routes or BGP routes can be used between the external gateway egress VRF on the border leaf node and the PE.
 - In the egress vSYS on the firewall, configure a static route pointing to the tenant vSYS on the firewall and a static route pointing to the egress VRF on the border leaf node. In the tenant vSYS on the firewall, configure a static route pointing to the egress vSYS on the firewall and a static route pointing to the tenant VRF on the border leaf node.
- If traffic does not pass through a firewall:
 - Configure static routes between the tenant VRF and egress VRF on the border leaf node (service leaf node).

Contents

1. Data Center Network Overview
2. Network Architecture Design and Data Planning
3. Underlay Network Design
4. Overlay Network Design
- 5. Network Security Design**
6. Network Management and O&M Design

CloudFabric Security Architecture

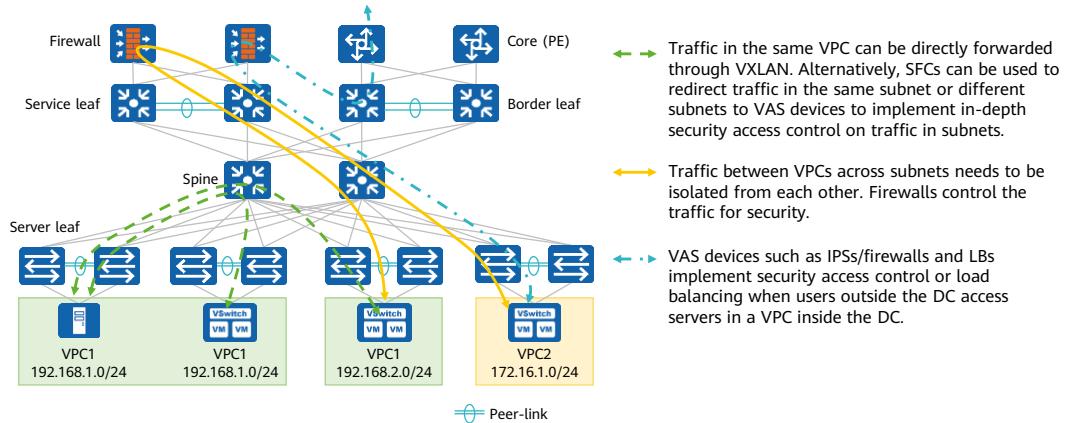


60 Huawei Confidential

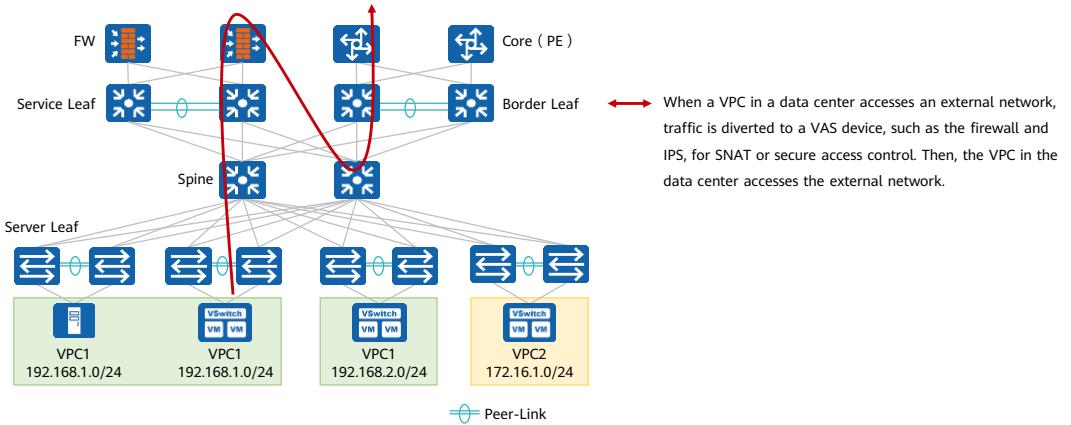


- Virtualization security:
 - Security groups of the cloud platform are used to protect VMs. The cloud platform adds VMs that require security control to different security groups using orchestration and defines security policies between security groups for access control.
- Intra-DC security:
 - Office zone:
 - Access control policies are configured on firewalls to protect security of access between intranet office users.
 - Cybersecurity Intelligence System (CIS) flow probes can be deployed to collect traffic in the office zone for in-depth threat detection.
 - Server zone:
 - Firewalls or web application firewalls (WAFs) are used to protect intranet servers.
 - Database audit is used to protect database servers.
 - CIS flow probes can be deployed to collect traffic in the server zone for in-depth threat detection, preventing intranet threats from being spread.

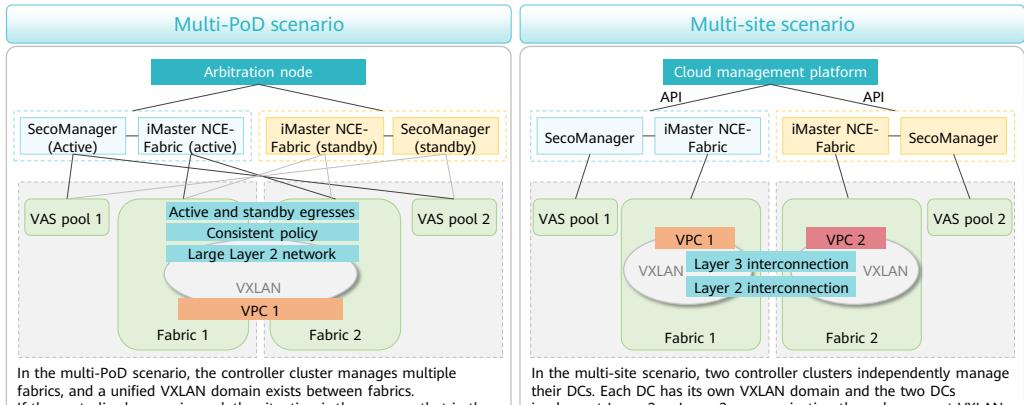
Intra-DC Security Deployment



Security Deployment Between DCs and External Networks



Inter-DC Security Deployment



In the multi-PoD scenario, the controller cluster manages multiple fabrics, and a unified VXLAN domain exists between fabrics.

If the centralized egress is used, the situation is the same as that in the single-DC scenario. If the active and standby egresses are used, a group of firewalls in active/standby mirroring mode must be deployed in each of the two fabrics. The controller cluster delivers the security policy to the two groups of firewalls. Firewalls in different DCs do not synchronize sessions.

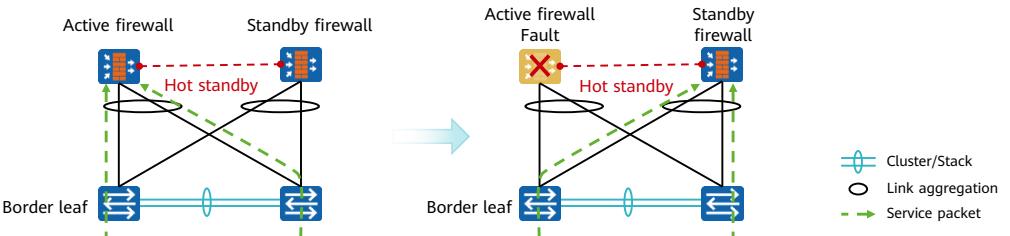
In the multi-site scenario, two controller clusters independently manage their DCs. Each DC has its own VXLAN domain and the two DCs implement Layer 2 or Layer 3 communication through segment VXLAN. During Layer 3 communication, traffic can be orchestrated to pass through firewalls in one DC or in both DCs, allowing security policies to be deployed flexibly.

- Note:

- Arbitration service: supports the site private network monitoring function. It periodically checks the network connectivity of the active, standby, and third-party sites and notifies these sites of the monitoring result through the communication links between the arbitration nodes. If the arbitration heartbeat is abnormal due to a network exception or site fault, the arbitration service uses an internal algorithm to provide the optimal site on the current network to implement automatic switchover between the active and standby sites.

Firewall Hot Standby Design

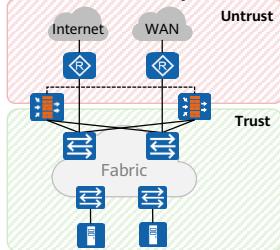
- You are advised to deploy firewalls in hot standby mode to improve reliability.



- As shown in the figure, firewalls are connected to border leaf nodes in bypass mode. Two firewalls are configured with the hot standby function and interconnected through heartbeat hot standby links.
- If the active firewall is faulty, the standby firewall takes over services from the active firewall and forwards service packets.

Security Zone Design

- A security zone, also known as a zone, is a collection of networks connected through one or more interfaces, where users have the same security attributes. There are three typical types of security zones: Trust, DMZ, and Untrust.
 - The Trust zone is a security zone with a high security level. It is typically used to define the zone where intranet users are located.
 - The DMZ is a security zone with a medium security level. It is typically used to define the zone where the servers that need to provide services for external networks are located.
 - The Untrust zone is a security zone with a low security level. It is typically used to define insecure networks such as the Internet.



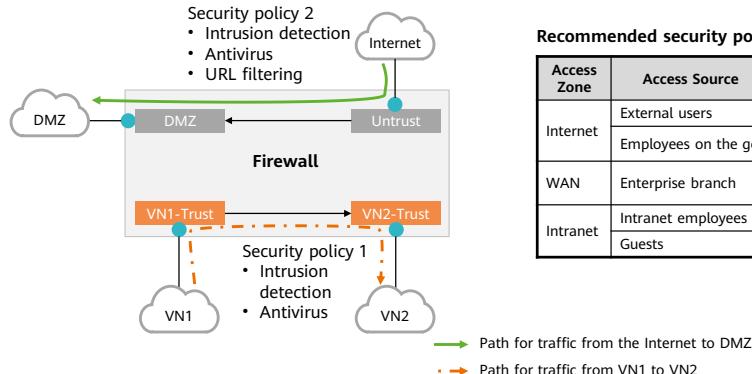
Security zone planning

- Generally, the intranet of a DC is considered secure, and security threats mainly come from the outside of the DC.
- Therefore, the Internet is divided into the Untrust zone, the DC intranet is divided into the Trust zone, and security devices are deployed at the egress to isolate the intranet from the external network and defend against external threats.

- Most security policies are implemented based on security zones. Each security zone identifies a network, and a firewall connects networks. Firewalls use security zones to divide networks and mark the routes of packets. When packets travel between security zones, security check is triggered and corresponding security policies are enforced. Security zones are isolated by default.

Security Policy Design

- After security zones are created on the firewall, these security zones are isolated from each other by default. To enable communication between security zones (for example, the campus intranet accesses the Internet), you need to configure Layer 3 connectivity and security policies on the firewall.



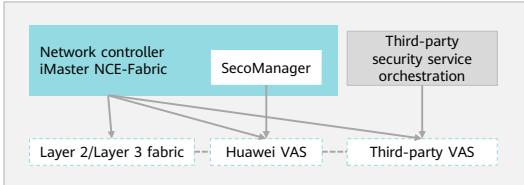
Recommended security policy design for common zones

Access Zone	Access Source	Trustworthiness	Recommended Security Policies
Internet	External users	Untrusted	Intrusion detection, URL filtering, antivirus
	Employees on the go	Medium	
WAN	Enterprise branch	Medium	URL filtering, antivirus
	Intranet employees	High	URL filtering, antivirus
Intranet	Guests	Low	

- As shown in the figure, after security policies are configured, virtual networks (VNs) on the intranet of the DC can communicate with each other, and the external networks can access servers in the DMZ. In addition, different security protection policies can be applied to traffic in different security zones.

Security Service Selection

- Huawei security service architecture:



- iMaster NCE-Fabric: provisions logical network services, orchestrates the bidirectional interconnection network between Huawei VAS devices and Layer 2 or Layer 3 fabric, and manages Huawei CE switches and delivers network configurations to them.
- SecoManager: orchestrates services for Huawei VASs, manages Huawei VAS devices, and delivers network configurations to them.
- Huawei VAS devices: Huawei firewalls provide service functions such as security policy, Elastic IP (EIP), SNAT, IPsec VPN, and content security detection.

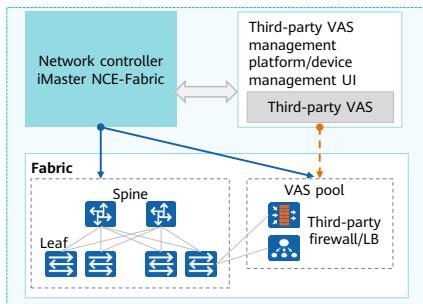
The CloudFabric solution provides the following security services:

- IPsec VPN
- Source Network Address Translation (SNAT)
- EIP
- Bandwidth management based on firewalls
- Anti-DDoS
- Security policy
- Content security detection
- Virtual system

- SecoManager description:

- SecoManager is a security controller that provides unified management for Huawei firewalls on a network.
- In the CloudFabric solution, SecoManager functions as a security controller to implement application-based independent security service provisioning. SecoManager provides security policies for applications and between applications to implement network visualization and improve network maintainability.
- By interworking with iMaster NCE-Fabric (SecoManager is installed on iMaster NCE-Fabric as a security service), SecoManager provides the following security capabilities for the solution: security policy service, SNAT, EIP, and IPsec.

Third-Party VAS Management Solution



- In the CloudFabric solution:

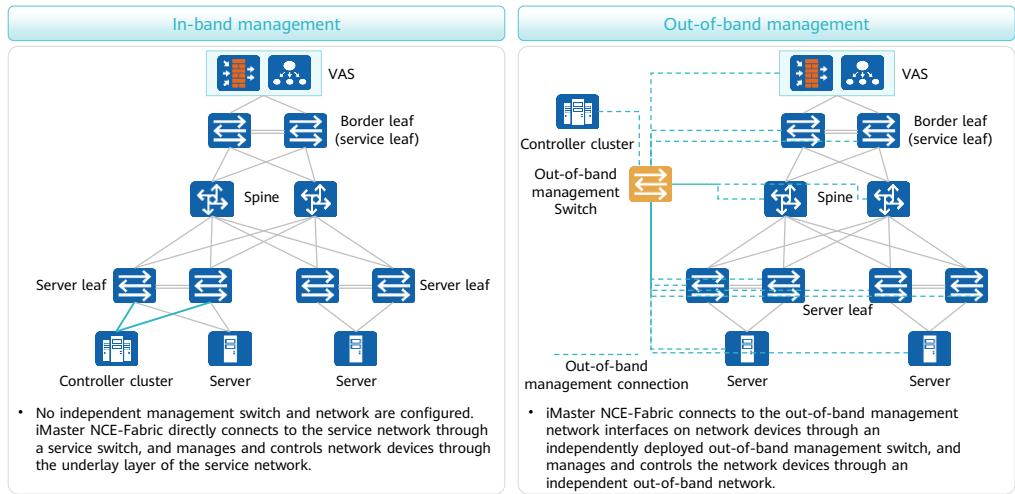
- For third-party VAS devices such as Check Point firewalls, the service manager mode is used. iMaster NCE-Fabric manages third-party VAS devices and is responsible for bidirectional network provisioning and bidirectional traffic diversion configuration. To provision L4-L7 services, administrators can redirect to the third-party VAS management UI from iMaster NCE-Fabric.
- For VAS devices of other vendors, the network policy mode is used. That is, iMaster NCE-Fabric does not manage third-party VAS devices. Only network provisioning and traffic diversion configuration from a fabric to third-party VAS devices are implemented on iMaster NCE-Fabric. Services provided by third-party VAS devices depend on the device capabilities.

- Service manager mode: iMaster NCE-Fabric manages fabrics and orchestrates the Layer 2 or Layer 3 interconnection network between Huawei VASs and a fabric. The third-party management platform orchestrates and delivers L4-L7 policies of third-party VASs.
- Service policy mode: iMaster NCE-Fabric manages fabrics and VASs, and orchestrates and delivers L2-L7 policies of third-party VASs.
- Network policy mode: iMaster NCE-Fabric does not manage third-party VAS devices. It is responsible for orchestrating the unidirectional interconnection network and traffic diversion from a fabric to third-party VAS devices.
- Note:
 - iMaster NCE-Fabric can manage third-party VAS devices such as Check Point firewalls.
 - iMaster NCE-Fabric uses SNMP to read device and link information and uses RESTful to deliver service commands to these devices.

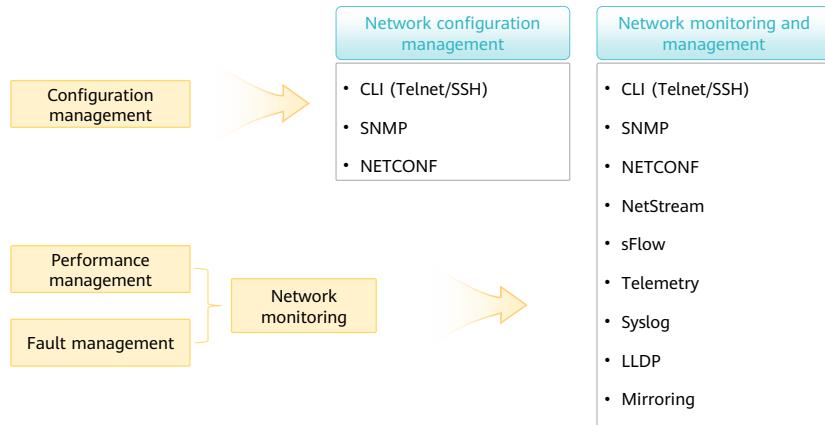
Contents

1. Data Center Network Overview
2. Network Architecture Design and Data Planning
3. Underlay Network Design
4. Overlay Network Design
5. Network Security Design
- 6. Network Management and O&M Design**

In-band Management/Out-of-band Management



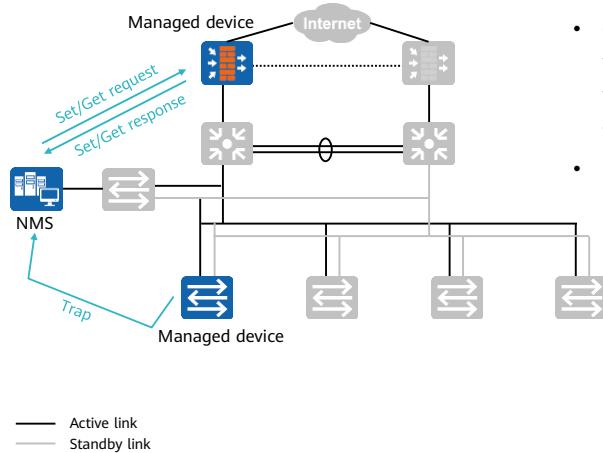
Network O&M Mode Selection



- You can select different network management modes based on DCN O&M requirements.

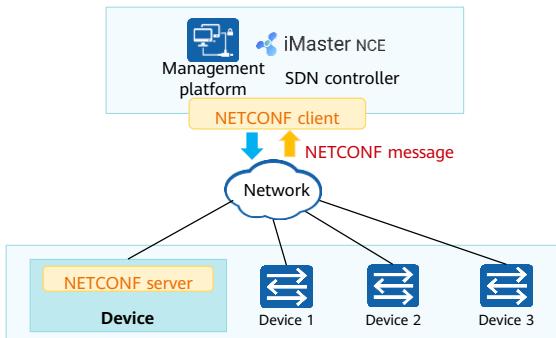
- The command-line interface (CLI) supports both network configuration management and network monitoring management.
- The Set function of SNMP supports network configuration management, and its Trap function supports network monitoring management.
- The Edit function of NETCONF supports network configuration management, and its Get function supports network monitoring management.

Network O&M Mode: SNMP



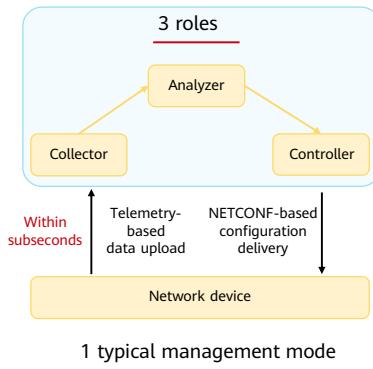
- Configure the SNMP management program in the network management station (NMS), enable the agent program on the managed device, and configure the SNMP protocol on the network.
- Using SNMP:
 - The NMS can obtain or change device information through the agent to implement remote monitoring and management.
 - The agent can report device status to the NMS in a timely manner.

Network O&M Mode: NETCONF



- NETCONF provides a set of mechanism for managing network devices. With this mechanism, users can add, modify, delete, back up, restore, lock, and unlock network device configurations. In addition, NETCONF provides transaction and session operation functions to obtain network device configuration and status information.
- NETCONF has three objects:
 - NETCONF client
 - NETCONF server
 - NETCONF message

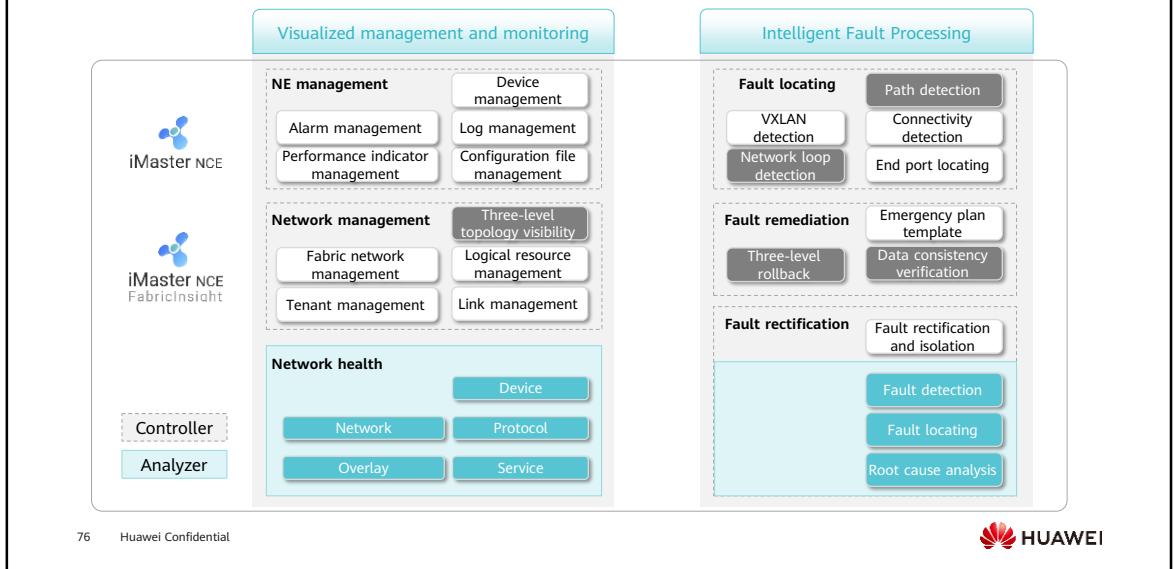
Network O&M Mode: Telemetry



- Telemetry, also known as network telemetry, is a technology for network monitoring, including packet check and analysis, intrusion and attack detection, intelligent data collection, and application performance management.
- Advantages of telemetry:
 - Supports multiple implementation modes, meeting diversified user requirements.
 - Collects a wide variety of data with high precision to fully reflect network status.
 - Continuously reports data with only one-time data subscription.
 - Locates faults rapidly and accurately.

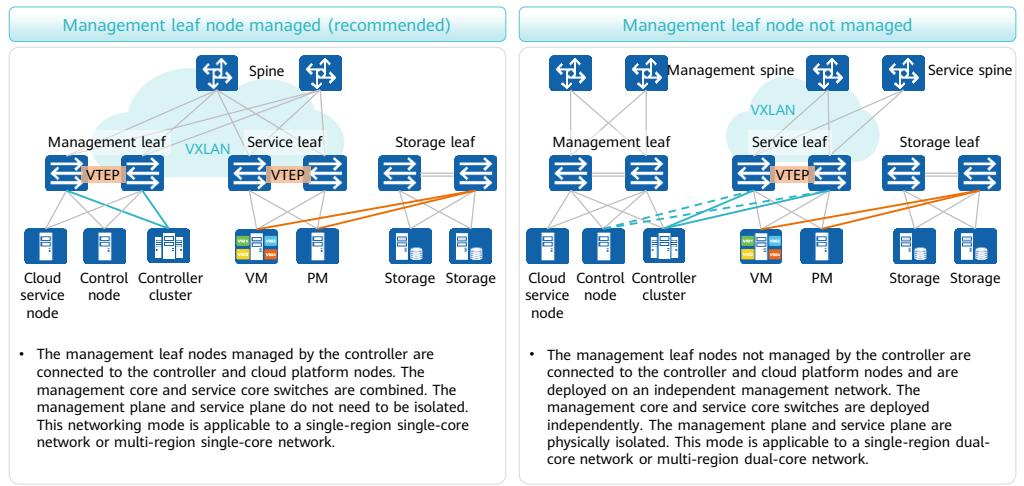
- The collector, analyzer, and controller are components of the network management system.
 - The collector receives and stores monitoring data reported by network devices.
 - The analyzer analyzes the monitoring data received by the collector and processes the data, for example, displays the data on the GUI.
 - The controller uses NETCONF to deliver configurations to devices, so as to manage network devices. The controller can deliver configurations to network devices based on the analysis data provided by the analyzer and adjust the forwarding behavior of network devices. It also controls the data that the network devices sample and report.

CloudFabric O&M Overview



- Intelligent O&M is implemented by the controller and analyzer. This course describes some O&M features.

iMaster NCE-Fabric Deployment Design



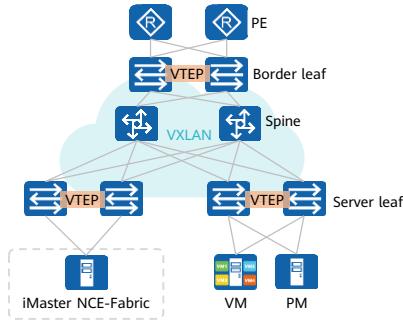
77 Huawei Confidential

HUAWEI

- iMaster NCE-Fabric is a next-generation SDN controller for DCNs and is a core component of the CloudFabric solution.
- The controller and SecoManager can be deployed in two modes: management leaf nodes managed by the controller and management leaf nodes not managed by the controller.
 - Networking where management leaf nodes are managed:
 - The management leaf node is managed by the DCN controller. The northbound and southbound gateways of the DCN controller are deployed on the management leaf node.
 - The southbound gateway of the DCN controller is deployed using a VLANIF interface. Direct routes are imported to the routing protocol on the underlay network to achieve communication with in-band management loopback interface addresses.
 - The DCN controller creates an overlay network for communication between network planes, such as the northbound gateway and cloud platform.

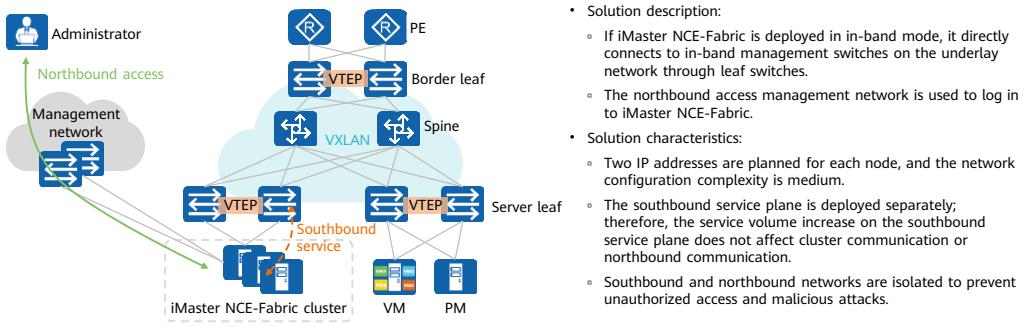
iMaster NCE-Fabric Deployment Design: Single-Node System

- To ensure high availability of network services, iMaster NCE-Fabric can be commercially deployed in a single-node system or cluster.
- In the single-node system deployment solution, iMaster NCE-Fabric is deployed on one node (PM or VM), saving hardware resources to the maximum extent.



iMaster NCE-Fabric Deployment Design: Cluster

- iMaster NCE-Fabric can be deployed in a cluster on PMs or VMs. PM cluster deployment is recommended to improve reliability. If a single node in the cluster is faulty, other nodes can still run properly.
- In the single-cluster deployment solution, iMaster NCE-Fabric is deployed in a three-server cluster in a DC to manage all switches in the DC. An iMaster NCE-Fabric cluster can also manage switches in multiple DCs.

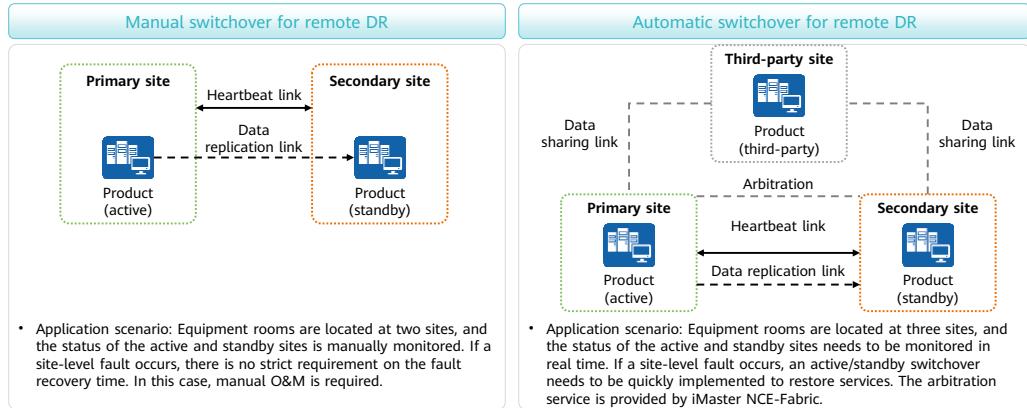


80 Huawei Confidential

HUAWEI

- The iMaster NCE-Fabric cluster consists of three network planes:
 - Internal communication network plane: Used for internal communication in a cluster, for example, communication between nodes and communication with the database.
 - Northbound management network plane: Used for northbound communication and Linux management, including cloud platform interconnection, web access, and Linux login.
 - Southbound service network plane: Used for communication with network devices in the southbound direction through NETCONF, SNMP, and OpenFlow.
- In actual deployment, the internal communication plane is integrated with the northbound management plane, and the southbound service plane is independently deployed. In addition, the southbound service plane manages DC switches in in-band mode. The figure shows the deployment solution.
- The cluster deployment solution has the following advantages:
 - Load balances services across multiple cluster nodes to ensure high reliability and performance.
 - Ensures the entire cluster runs normally even if a cluster node fails, improving reliability.
 - Supports flexible expansion to enhance the performance of the entire cluster, improving scalability.

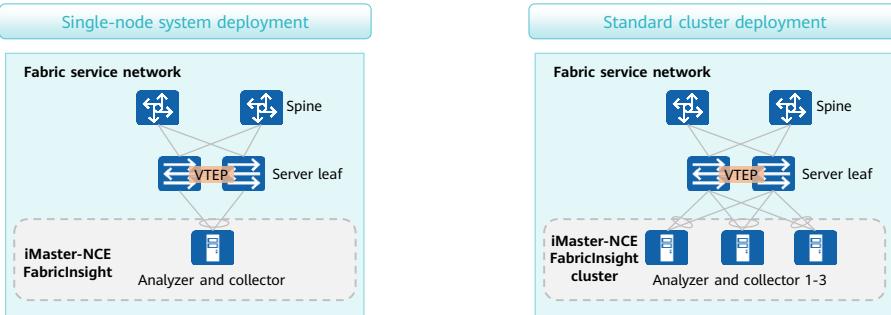
iMaster NCE-Fabric Deployment Design: Active/Standby Clusters



- In the multi-DC active/standby DR scenario, the active/standby cluster solution can be used.
- If the active DC is faulty, the standby DC and standby controller cluster become active and continue to provide services, improving DC DR reliability.

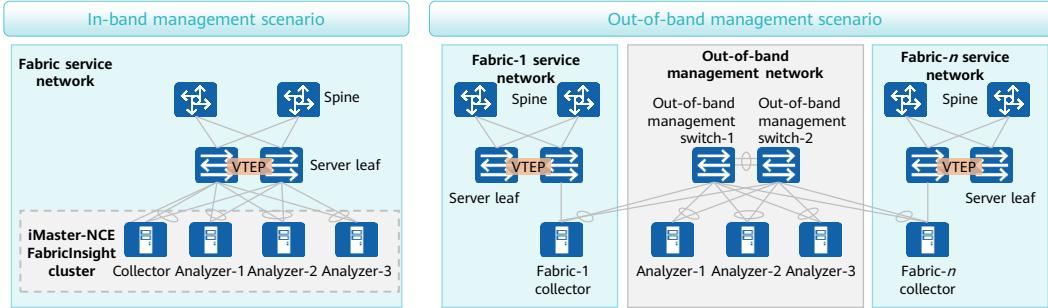
iMaster NCE-FabricInsight Deployment Design: Single-Node System and Standard Cluster

- In single-node system deployment (in-band management) of iMaster NCE-FabricInsight, the collector and analyzer are combined. Only one server needs to be connected to the leaf node.
- In standard cluster deployment (in-band management) of iMaster NCE-FabricInsight, the analyzer and collector are combined. That is, no independent collector server needs to be deployed.



iMaster NCE-FabricInsight Deployment Design: Advanced Cluster

- In advanced cluster deployment of iMaster NCE-FabricInsight, the collector and analyzer are deployed separately. It is recommended that iMaster NCE-FabricInsight be connected to an independent leaf node, preventing link congestion caused by increased traffic pressure on service links.



CloudFabric Software Deployment Mode Selection

Component	Mandatory	Deployment Mode	Description
iMaster NCE-Fabric	Yes	Single-node system deployment	The controller is deployed on one node.
		Cluster deployment	The controller cluster consists of N nodes. The controller can be installed on PMs or VMs.
iMaster NCE-FabricInsight	No	Single-node system deployment	In single-node system deployment, the collector and analyzer are combined. Only one server needs to be connected to the leaf node. A maximum of 100 CloudEngine devices can be managed.
		Standard cluster deployment	In standard cluster deployment, the analyzer and collector are combined. That is, no independent collector server needs to be deployed.
		Advanced cluster deployment	In advanced cluster deployment, the collector and analyzer are deployed separately. It is recommended that iMaster NCE-FabricInsight be connected to an independent leaf node, preventing link congestion caused by increased traffic pressure on service links.
SecoManager	No	Independent deployment	SecoManager is deployed on a server or VM as independent software.
		Combined with iMaster NCE-Fabric	SecoManager and iMaster NCE-Fabric are deployed on the same physical server or VM.

Quiz

1. (True or false) On a CloudFabric data center network with more than 200 switches, OSPF is recommended on the underlay network. ()
 - A. True
 - B. False
2. (Multiple-answer question) Which of the following deployment modes can be used to ensure high reliability of border leaf nodes? ()
 - A. Deploy border leaf nodes in active-active mode.
 - B. Cross-connect border leaf nodes to uplink core devices.
 - C. Fully mesh border leaf nodes with downlink spine nodes.
 - D. Deploy a bypass policy between border leaf nodes.

1. B
2. ABCD

Summary

- This course describes the planning and design of the CloudFabric DCN, including the network architecture design, underlay and overlay network design, network security design, network management and O&M design.
- On completion of this course, you will understand the typical methods of designing a DCN and be able to plan and design a DCN.

Thank you.

把数字世界带入每个人、每个家庭。

每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



CloudFabric Data Center Network Deployment - Computing Scenario



Foreword

- When the computing service management system is complex or the computing management and network management are not integrated enough and a unified cloud platform cannot be constructed, the computing and network can be associated to manage and provision services together.
- This course introduces how to build data center network. This solution is oriented to the computing linkage scenario. The controller connects to the computing virtualization platform instead of the cloud platform. The controller and computing virtualization platform deliver services together to implement collaborative provisioning of computing and network services.

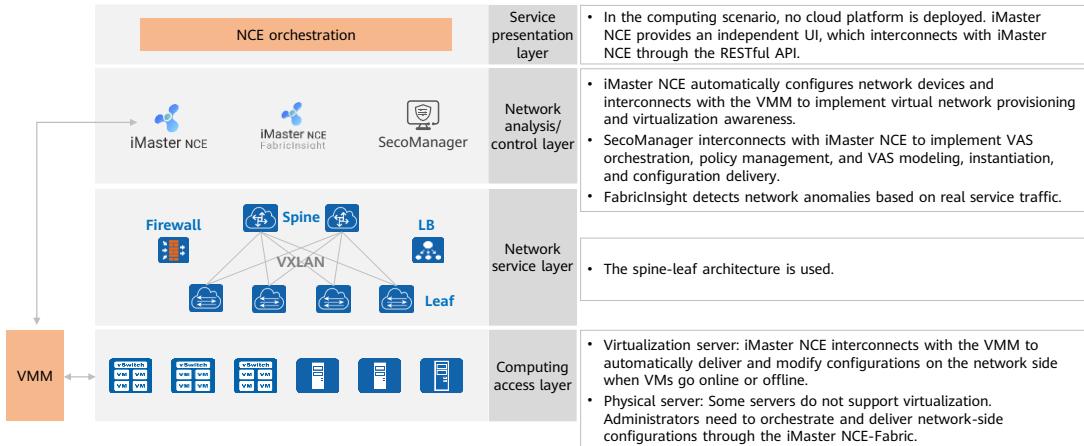
Objectives

- On completion of this course, you will be able to:
 - Describe the deployment process in the computing scenario.
 - Service provisioning in the computing scenario.
 - Understand the deployment in easy mode.

Contents

- 1. Deployment Process Overview**
2. Pre-configuration
3. Service Provisioning
4. Easy Deployment

Architecture of the Computing Scenario

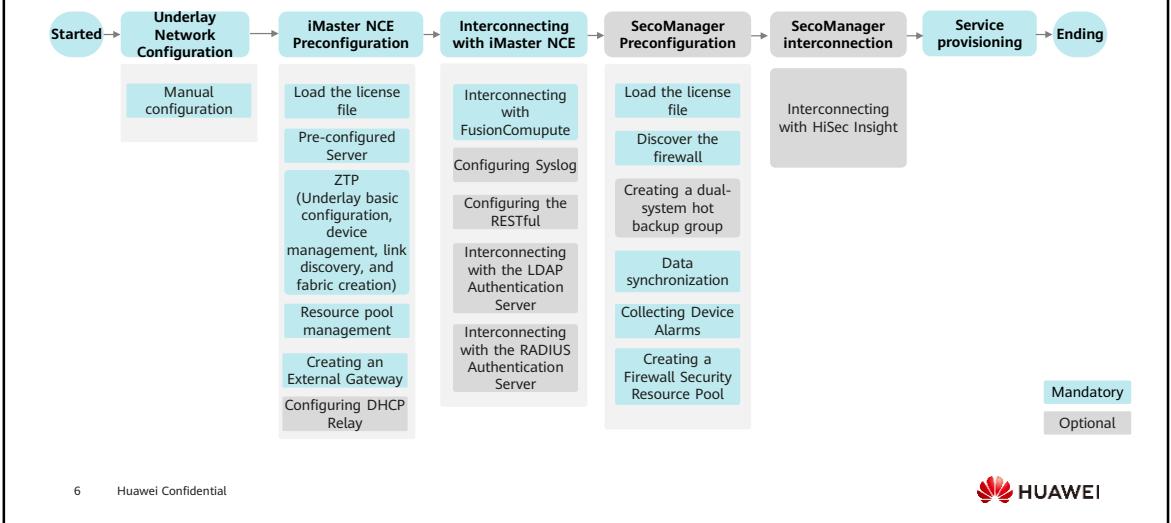


5 Huawei Confidential



- The architecture and deployment process of the rack leasing scenario are similar to those of the computing linkage scenario. This chapter uses the computing scenario as an example to describe the architecture and deployment process of the solution.

Deployment Process in the Computing Scenario



6 Huawei Confidential

 HUAWEI

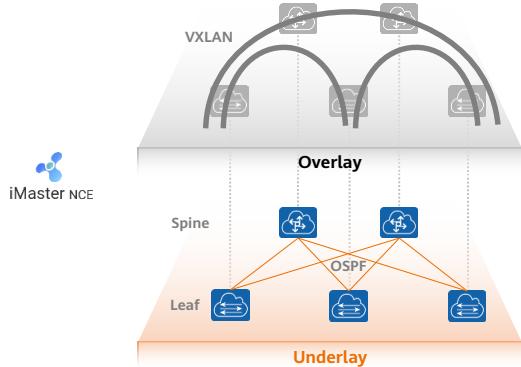
- In the deployment phase, the network administrator needs to perform operations such as basic underlay configuration and device management. These operations can be manually performed or completed using zero touch provisioning (ZTP). This slide uses ZTP as an example.

Contents

1. Deployment Process Overview
- 2. Pre-configuration**
 - Underlay Network Pre-configuration
 - iMaster NCE-Fabric Pre-configuration
3. Service Provisioning
4. Easy Deployment

Underlay Network Pre-configuration

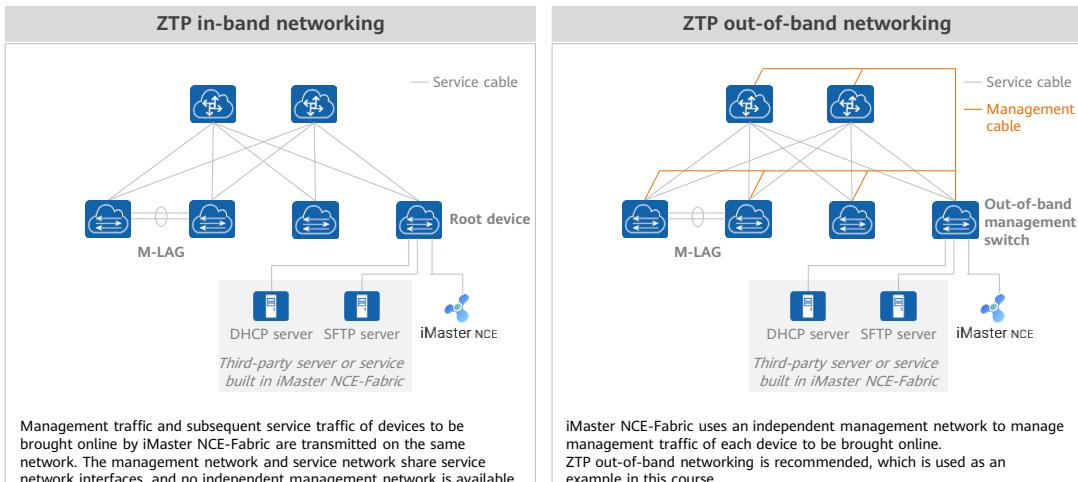
- An underlay network is the basic network for constructing a virtual extensible local area network (VXLAN) service network, which is an overlay network.
- The underlay network can be configured in ZTP or manual mode.



- ZTP allows newly delivered or unconfigured devices to automatically load version files, deploy the underlay network, and register with iMaster NCE for being managed after they start.
- This course will introduce both the ZTP and manual modes.

- In traditional deployment mode, the administrator needs to manually configure each newly delivered or unconfigured device after hardware installation, which lowers deployment efficiency and results in high labor costs. iMaster NCE-Fabric provides the ZTP-based simplified deployment function, which enables you to plan the networking topology and fabric resources, automatically bring devices online, execute device installation scripts, and deliver underlay configurations to devices in batches on the GUI. This reduces labor costs and improves deployment efficiency. ZTP-based simplified deployment enables quick rollout and management of DCN devices.
- The DC physical network of the CloudFabric solution uses the spine-leaf architecture and supports horizontal on-demand capacity expansion. The roles in the network include spine nodes, server leaf nodes, border leaf nodes, service leaf nodes, and DCI gateways. There are often a large number of server leaf nodes, which require automatic service provisioning. Therefore, ZTP currently focuses on server leaf nodes.
- Server leaf nodes support Multichassis Link Aggregation Group (M-LAG) and standalone networking, which are applicable to different server access scenarios. M-LAG networking is recommended because high reliability is achieved when servers are dual-homed to the M-LAG. In addition, each M-LAG device has its own control plane, simplifying upgrade and maintenance.

Basic Networking of ZTP



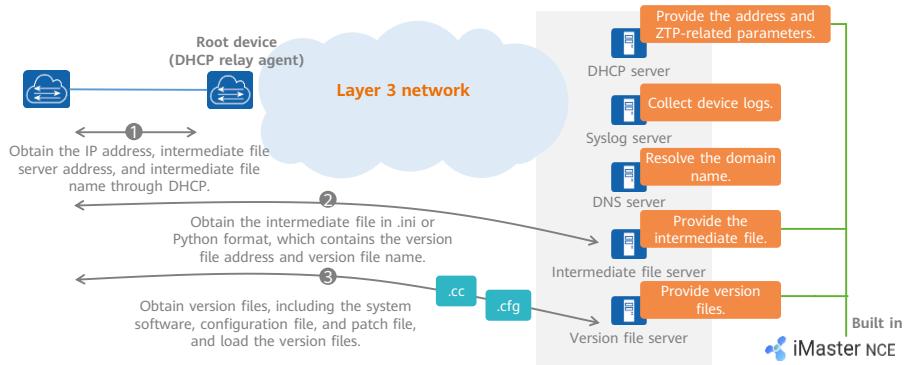
10 Huawei Confidential



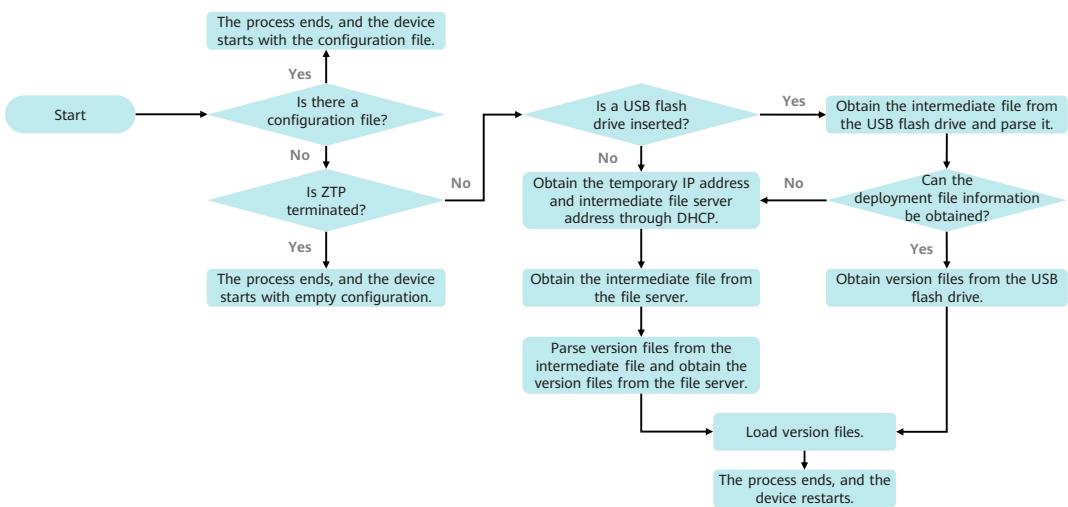
- For in-band networking:

- iMaster NCE-Fabric: used to execute ZTP tasks and manage the devices to be brought online.
- Root device: a device that has been managed by iMaster NCE-Fabric and connects to the devices to be brought online. The root device functions as the DHCP relay agent of the devices to be brought online and applies for a temporary IP address from the DHCP server for these devices. The root device is involved in in-band networking and needs to be manually managed.
- Spine and leaf nodes: CE switches that need to be brought online through ZTP. Currently, spine and server leaf nodes can be brought online through ZTP. To bring a border leaf node online through ZTP, bring the border leaf node online as a server leaf node and configure an external gateway for the border leaf node on iMaster NCE-Fabric.
- Device to go online: CE device that is to be brought online through ZTP.
- Online device: upper-level device of the devices to be brought online. In in-band networking scenarios, if spine nodes are brought online through ZTP, the online device is the root device; if server leaf nodes are brought online through ZTP, the online device is a spine node.

Standard ZTP Fundamentals



Standard ZTP Process

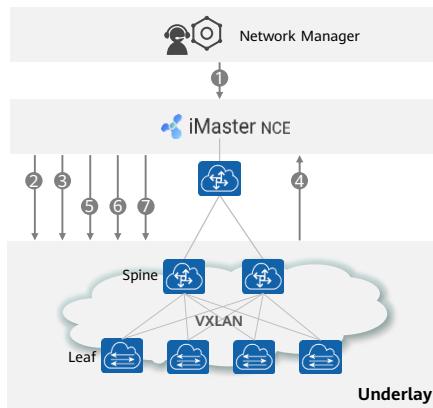


13 Huawei Confidential

HUAWEI

- The process of using the controller to bring devices online through ZTP is slightly different from this process and will be described later.
- Powering on and starting the device:
 - After the device is powered on, if the device has a configuration file, the device properly starts with the configuration file; if the device has no configuration file, the ZTP process starts.
 - If you have logged in to the device without a configuration file through the console port, you can choose whether to terminate the ZTP process as prompted. If you choose to terminate the ZTP process, the device starts with empty configuration.
- Obtaining the intermediate file and version files from the USB flash drive:
 - After the ZTP process starts, the unconfigured device first tries to obtain the intermediate file from the USB flash drive. If the device obtains the intermediate file, it parses the file and obtains information about the version files to be downloaded. After downloading the version files, the device restarts to complete automatic deployment. The device enters stage 3 if any of the following conditions occur: no USB flash drive is installed; the USB flash drive does not contain a required intermediate file; the device fails to obtain the version files.

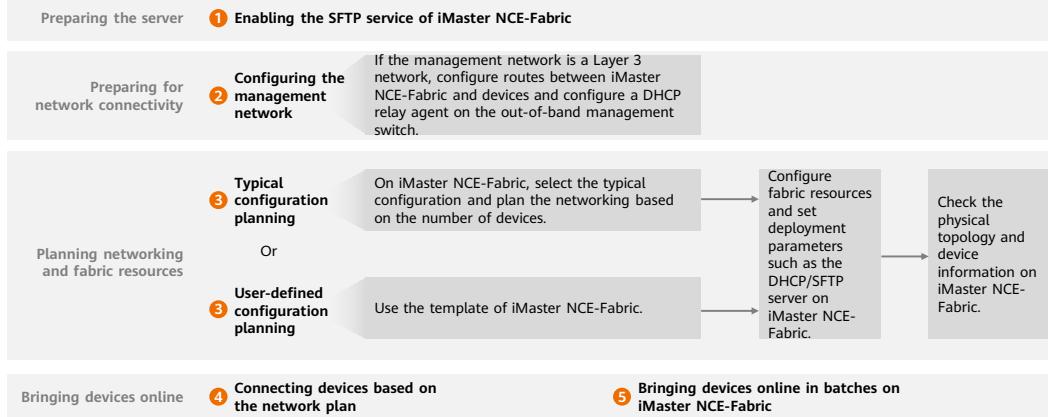
ZTP Deployment Using the Controller



ZTP Configuration Provisioning Process

- Application scenario: The Controller has been physically connected to the device (Huawei CE switch), the device license of the device has been imported to the Controller, and the root device has been manually managed.
- Process:
 1. The network administrator clicks to start the ZTP task.
 2. The iMaster NCE-Fabric advertises DHCP packets.
 3. The device to go online obtains the temporary IP address and southbound IP address of the iMaster NCE-Fabric from the DHCP packet sent by the iMaster NCE-Fabric.
 4. The device to go online uses the built-in certificate to initiate authentication to the iMaster NCE-Fabric.
 5. After the authentication succeeds, the iMaster NCE-Fabric determines the device role (Spine/Leaf) based on the device model.
 6. The iMaster NCE-Fabric delivers configurations such as the management IP address, SNMP, and NETCONF to devices to go online. After the devices to go online are restarted, the iMaster NCE-Fabric implements formal management using the management IP address.
 7. The controller delivers interconnection configurations, OSPF configurations, and BGP configurations to newly online devices through LLDP links.
 8. Devices on the entire network go online successfully and all links are established. The network topology is displayed on the iMaster NCE-Fabric.

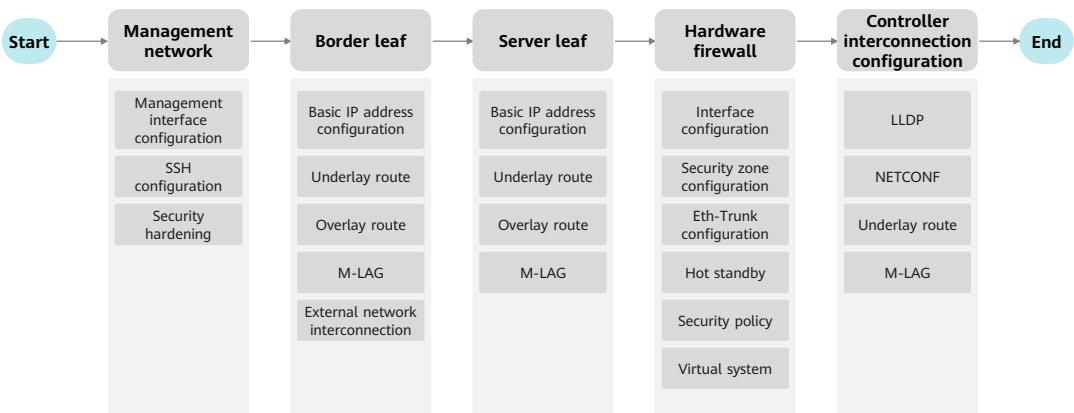
ZTP Process in Out-of-Band Networking (Using the Service Built In iMaster NCE-Fabric)



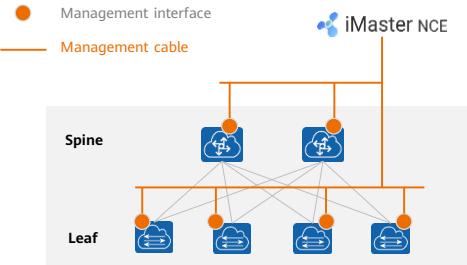
Typical Configuration Planning and User-Defined Configuration Planning

Item	Typical Configuration Planning	User-Defined Configuration Planning
Scenario	Applies to the scenario where the customer has no network plan.	Applies to the scenario where the customer has a network plan.
Configuration complexity	The configuration is simple. You only need to enter the numbers of spine and leaf nodes and the port range of the nodes. iMaster NCE-Fabric automatically generates a topology based on the number of devices and networking mode.	Compared with typical configuration planning, user-defined configuration planning is more complex. Using the iMaster NCE-Fabric template is more complex than using CloudFabric Designer.
Impact on follow-up operations	You can connect devices only after networking and fabric resource planning are complete on iMaster NCE-Fabric and a topology is generated.	After completing network planning, you can directly connect devices. Fabric resource planning can be performed during ZTP.

Process of Manually Configuring an Underlay Network



Management Network Configuration



Management interface configuration

- Configure the IP address and route of the management interface (MEth interface on the CE switch).

SSH configuration

- Enable the SSH function, create an SSH user, and specify the source interface of the SSH server.

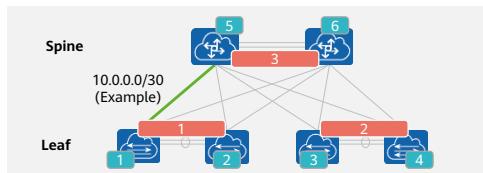
Security hardening

- Configure an ACL to allow only the controller to log in to the device using SSH.

Border Leaf - Basic IP Address Configuration

L0 11.0.XX

L1 11.0.0.X



Interconnection interface

- Plan a 30-bit network segment for interconnection interfaces.

Loopback0

- Configure the same VTEP address for the two spine nodes and the same VTEP address for the server leaf nodes configured in an M-LAG.

Loopback1

- Configure an independent router ID for each device.

- In this example, the spine, border leaf, and service leaf nodes are combined.

Border Leaf - Underlay Route

- Configure OSPF in a single area and enable OSPF on interconnection interfaces to optimize the OSPF convergence speed.

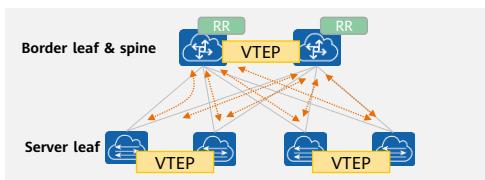
- Configuration example:

```
#  
ospf 1 router-id 11.3.3.3  
stub-router on-startup 600 include-stub  
area 0  
network 10.1.1.1 0.0.0.0  
...  
#  
interface 10GE1/0/20  
undo portswitch  
ip address 10.1.1.54 255.255.255.252  
ospf network-type p2p
```



Border Leaf - Overlay Route

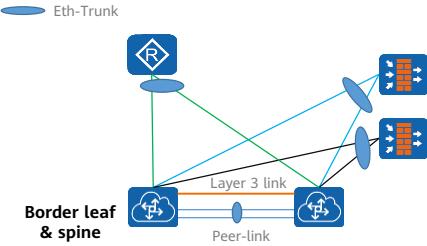
↔ BGP EVPN



BGP EVPN

- Use the IP address of Loopback1 as the source address to establish a BGP EVPN peer relationship. Spine nodes function as RRs, and leaf nodes function as RR clients.
- Configure the IP address of Loopback0 as the VTEP address, and configure the same source MAC address for NVE interfaces on the active-active gateways.

Border Leaf - M-LAG

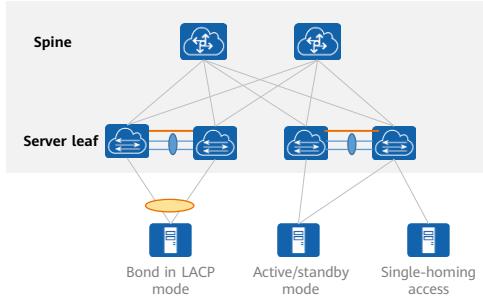


M-LAG:

- Configure Eth-Trunk interfaces as the peer-link interfaces and set the spanning tree mode to V-STP.
- Configure an M-LAG and establish links between the border leaf nodes and firewalls and between the border leaf nodes and external routers.

- Traffic from the corresponding interconnection management VLAN must be allowed to pass through the interconnection link between the border leaf nodes and firewalls.
- After creating Eth-Trunks with external routers, you need to configure interconnection between the border leaf nodes and external routers based on the external network connection mode. This configuration is performed on iMaster NCE-Fabric after the border leaf nodes are managed by iMaster NCE-Fabric.

Server Leaf - Active-active Gateways and M-LAG



Active-active gateways

- Deploy active-active gateways between server leaf nodes, configure a DFS group, and enable the active-active function.

M-LAG

- Configure an M-LAG based on the server access mode, such as access in LACP mode, active/standby mode, or single-homing mode.

Firewall

- Perform the following pre-configurations on firewalls.

No.	Task Name	Description
1	Interface IP address	Configure an IP address for hot standby heartbeat interfaces and the management interface.
2	Eth-Trunk	Configure Eth-Trunks for connecting to border leaf nodes and allow packets from the corresponding service VLANs to pass through.
3	Security domain and security policy	Add the hot standby heartbeat interfaces to the DMZ, Eth-Trunk and Virtual-if0 to the Untrust zone, and management interface to the Trust zone. Set the action in the security policy to permit by default.
4	Hot standby	Configure hot standby, enable the quick session backup function, configure a firewall to restart with the basic hot standby configuration and synchronize normal service configurations from the other normal firewall, and enable certain functions of the standby firewall.
5	Virtual system and Virtual-if name conversion	Enable the Virtual-if name conversion function and enable the virtual system function.

Interconnection with the Controller

- To enable the controller to manage devices, perform configurations, and discover links, configure the following functions on switches and firewalls.

No.	Task Name	Description
1	SNMP	Configure SNMPv3 and set the corresponding MIB view. The controller obtains LLDP link information from the MIB view specified in SNMP. The MIB view defined in SNMP is iso-view , and the OID MIB sub-tree of the MIB objects is iso .
2	NETCONF	Switch: Configure SSH, enable the NETCONF function, and enable the SNETCONF service. Firewall: Configure an API user, enable the NETCONF interface service, and configure the NETCONF port number.
3	LLDP	You need to enable LLDP globally on CE switches and firewalls so that the controller can discover links using LLDP.

Contents

1. Deployment Process Overview
2. **Pre-configuration**
 - Underlay Network Pre-configuration
 - iMaster NCE-Fabric Pre-configuration
3. Service Provisioning
4. Easy Deployment

Importing a License to iMaster NCE-Fabric

- Obtain the license file of iMaster NCE-Fabric and import it to iMaster NCE-Fabric.

Update License

License file: LICMasterNCE-Fabric_V100R021_00 (5 KB) Upload successful.

Replace all license files for a product.

Product name: iMaster NCE-Fabric Product version: V100R021

Resource Name	ID	Consumption %	Current Capacity %	New Capacity %	Update Type %	Risk Level %
iMaster NCE-Fabric -	LNCEFN0D01	0	--	10	Added	Suggestion
iMaster NCE-Fabric -	LNCEFN0E01	0	--	10	Added	Suggestion
iMaster NCE-Fabric -	LNCEFPV9W01	0	--	1000	Added	Suggestion
iMaster NCE-Fabric -	LNCEFFD01	0	--	1000	Added	Suggestion
iMaster NCE-Fabric -	LNCEFMCS01	0	--	10	Added	Suggestion

Total records: 5

10 page | Cancel | Apply

- This slide uses the traditional license management mode as an example.

iMaster NCE-Fabric Pre-configuration (1)

Task Name	Task Description
Pre-configured Server	To add the server and the links between the server and devices to iMaster NCE (Fabric), you need to configure the link discovery protocol on the server and check whether the host name is duplicate.
Device management	Configure a global policy for device management on iMaster NCE (Fabric). (for example, synchronizing device online data, saving device configurations periodically, and verifying device SSH fingerprints.) Then, iMaster NCE-Fabric discovers and manages network devices based on SNMP and NETCONF. After the management is complete, you need to create a device group to manage all-active devices and collect device alarms so that iMaster NCE-Fabric can successfully manage the target devices and device groups.
Link management	iMaster NCE-Fabric can learn the topology structure between devices and obtain the network connectivity status based on the link status and link details. iMaster NCE-Fabric supports automatic discovery, manual creation, and batch import. Automatic discovery is based on the fact that the devices at both ends of a link support the link automatic discovery protocol, such as LLDP. To manually create or import a file in batches, you need to enter the port information of the devices at both ends.
Managing Third-Party VAS Devices	(Optional) When CheckPoint, Palo Alto, Fortinet, and F5 load balancing are used, you can use iMaster NCE-Fabric to manage the devices in the computing interworking scenario. In this way, when VAS services are provisioned, iMaster NCE-Fabric can automatically deliver configurations such as service routes to the interconnection ports between the NCE and switches.

- After the license file is loaded, you need to perform the following pre- configurations to prepare for service provisioning.

iMaster NCE-Fabric Pre-configuration (2)

Task Name	Task Description
Resource pool management	<p>Creating a fabric: After device management and link discovery are successful, you need to create a fabric resource pool on iMaster NCE-Fabric and specify the egress gateway and DCI gateway (multi-DC scenario) to prepare for service provisioning. This scenario supports the creation of distributed network overlay (recommended) and centralized network overlay fabrics.</p> <p>Configure the best-effort link for the active-active CE device group: If the border leaf node functions as the active-active CE switch, configure the best-effort link to improve the reliability of the egress network.</p> <p>Setting the role of the firewall link: For the service interconnection link between the firewall and switch, you can set the role to internal, external, or internal and external links (recommended).</p> <p>Creating third-party L4-L7 resource pools: If third-party VAS devices are used, create third-party L4-L7 resource pools on iMaster NCE (Fabric).</p> <p>Associating fabrics with L4-L7 resource pools: Associate the created fabric with L4-L7 resource pools so that L4-L7 services can be associated with L2-L3 services when services are provisioned. If you need to associate Huawei L4-L7 resource pools, perform this operation only after the resource pools are created on the SecoManager.</p> <p>Configuring interconnection resources: When cross-VPC interconnection services pass through the firewall, you need to specify the value range of the interconnection VLAN and IP address between the switch and the firewall. iMaster NCE-Fabric automatically selects the VLAN and IP address from the range to deliver configurations. When a best-effort link is configured, iMaster NCE-Fabric automatically selects the interconnection IP addresses of the two ends of the best-effort link and delivers the IP addresses to devices.</p> <p>Configuring global resources: During service provisioning, iMaster NCE-Fabric uses a series of variable parameters. (such as the BD, global VNI, global VLAN, public IP address, and interworking IP address). Therefore, you need to set these parameters globally in advance so that iMaster NCE-Fabric can invoke the parameters.</p>

iMaster NCE-Fabric Pre-configuration (3)

Task Name	Task Description
Creating an External Gateway	To orchestrate services that access external data centers, you need to create an external gateway on iMaster NCE-Fabric and define the destination, outbound interface, and route delivery mode for accessing external networks.
Configuring DHCP Relay	(Optional) Create a DHCP relay so that VMs on the service network can automatically obtain IP addresses from the DHCP server.

iMaster NCE-Fabric Interconnection Commissioning

- After iMaster NCE-Fabric is preconfigured, you need to interconnect with the VMM on the computing virtualization platform to deliver computing services and network services.

Task Name	Task Description
Interconnecting with FusionCompute	To enable iMaster NCE-Fabric to detect VM login, logout, and migration through FusionCompute and implement automatic network service deployment, you need to configure interconnection between iMaster NCE-Fabric and FusionCompute.
Configuring FabricInsight	(Optional) If iMaster NCE-Fabric and iMaster NCE (FabricInsight) need to implement data association for intelligent fault handling, you need to interconnect them with each other.
Interconnecting with Northbound Services	(Optional) iMaster NCE-Fabric can interconnect with multiple systems through northbound interfaces to implement the following functions: <ul style="list-style-type: none">iMaster NCE-Fabric interconnects with eSight through the northbound SNMP protocol. After the interconnection is complete, eSight can synchronize internal alarms of iMaster NCE-Fabric.iMaster NCE-Fabric interconnects with the Syslog server to transfer iMaster NCE-Fabric logs to the Syslog server for centralized storage.
Interconnecting with Southbound Services	(Optional) Interconnect with the LDAP server and RADIUS server. <ul style="list-style-type: none">To enable users on the LDAP server or AD server to log in to iMaster NCE-Fabric, you need to configure interconnection between iMaster NCE-Fabric and these servers.To enable users in the user group on the RADIUS server to authenticate logins to iMaster NCE-Fabric, you need to configure interconnection between iMaster NCE-Fabric and the RADIUS server.

- iMaster NCE-Fabric can also interconnect with the vCenter of VMware.

SecoManager Pre-configuration and Interconnection Commissioning

- In the CloudFabric solution, the SecoManager and iMaster NCE-Fabric are combined. After the SecoManager is installed, the SecoManager and iMaster NCE-Fabric are automatically connected. When the SecoManager is installed and the license is activated, perform the following operations to prepare for service provisioning.

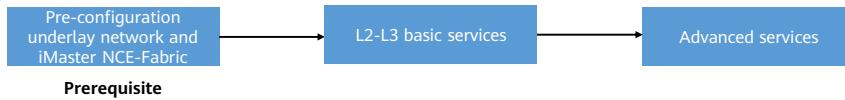
No.	Task Name	Description
1	Loading the license of the SecoManager	Log in to iMaster NCE-Fabric and load the license of the SecoManager.
2	Discovering firewalls	Discover Huawei firewalls on the SecoManager.
3	Creating a hot standby group for firewalls	(Optional) By default, the SecoManager can automatically identify hot standby groups. You do not need to perform this operation. If the SecoManager cannot automatically identify firewalls in a hot standby group, you need to manually create a hot standby group after the firewalls are discovered.
4	Synchronizing data	Before service provisioning, you are advised to perform difference discovery between the active and standby firewalls to ensure that the configurations on the active and standby firewalls are the same.
5	Collecting device alarms	Enable the SNMP trap function on iMaster NCE-Fabric to collect firewall traps.
6	Creating a firewall resource pool	Add the firewalls discovered by the SecoManager to a security resource pool to implement virtualization and provide virtual security resources for tenant services.
7	Associating a fabric with an L4-L7 resource pool	The resource pool created in the previous step must be associated with the fabric resource pool created on iMaster NCE-Fabric.
8	Setting the link role for firewalls	Set roles of links between firewalls and switches. Different link roles carry different traffic. There are three types of link roles: internal link, external link, internal and external link.

Contents

1. Deployment Process Overview
2. Pre-configuration
- 3. Service Provisioning**
 - Deploying Layer 2 and Layer 3 Basic Services
 - Deploying the VPC Interconnection Service
 - Deploying Value-added Services
 - Deploying Microsegmentation and Service Chain
4. Easy Deployment

Service Provisioning Overview

- Service provisioning refers to allocating appropriate network and computing resources to carry service applications. A data center administrator needs to allocate certain resources to tenants based on the service plan. Then the tenant administrator can configure and deploy network and computing services based on the resources.
- Service provisioning involves two steps: L2-L3 basic service invoking and other advanced services.
 - Layer 2 and Layer 3 basic services: Constructs the VPC basic network and associates with the VMM to connect VMs to the network.
 - Other advanced services: Invokes various advanced services, such as VPC interworking, value-added services, and service chain, based on service requirements.

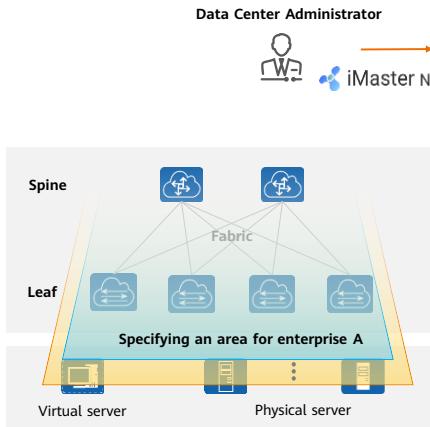


Overview of L2-L3 Basic Services

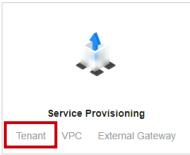
- Basic L2-L3 services refer to the basic networks created by the tenant administrator in the VPC, including logical routers and switches. Logical ports and user ports can be created in the following scenarios:
 - Computing association: VMs are connected to the network through VMM mapping, and logical ports and user ports are automatically generated.
 - Rack leasing: Manually create logical ports and user ports and specify the actual parameters for connecting servers to server leaf nodes.

- Configuration roadmap:
 - Create a tenant and allocate resources to the tenant.
 - Create a VPC in the tenant.
 - Create a logical router in the VPC of the tenant and create a subnet list.
 - Create a logical switch in the tenant's VPC and associate it with a logical router and subnet.
 - Orchestration server access:
 - Computing association scenario: Create VMM mappings and associate them with different logical switches. Create a VM on the VMM and select the corresponding port group to access the network.
 - In the rack leasing scenario, create a logical port and a user port and set parameters respectively.

Creating a Tenant and VPC



- Creating a Tenant and Allocating Resources



- Creating a VPC and Authorizing Related Resource Pools

Create VPC

If the VPC is not deployed, snapshots are not supported. To use snapshots, deploy the VPC as soon as possible.

Name: VPC1

Description:

Fabric: Fabric1

Multicast capability:

Cancel OK

- A tenant can create multiple VPCs based on service requirements.

Creating Logical Routers and Switches

- Orchestrate logical networks in a VPC and create logical routers and switches based on service requirements.
 - When creating a logical router, you can set related parameters, such as the subnet and gateway.
 - When creating a logical switch, you can associate it with a logical router to automatically complete service orchestration.

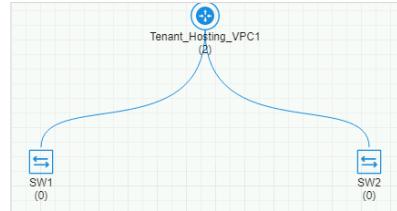
• Setting parameters related to the logical router

Subnet List	Route List	Loopback Interface	BGP Peer
Enter subnet <input type="text"/>			
<input type="checkbox"/> CIDR ↑ 192.168.10.0/24 192.168.20.0/24	Gateway IP Address 192.168.10.254	Protocol IP 192.168.10.254	DHCP Closed
			RA Mode --

• Associating a Logical Switch with a Logical Router

Select Logical Router

Enter a router name, VRF, CNI, or CIDR.			
<input type="radio"/> Tenant_Hosting_VPC1	Logical Router Name Tenant_Hosting_VPC1	VRF Name Tenant_Hosting_VPC1_5002	VNI 5002



Orchestration Server Access (1)

- Create VMM mappings for accessing VMs.

- Creating VMM Mappings

Create

VMM: FusionCompute

DVS: Computing_vSwitch

Port group: Select port group

IP-MAC binding: Off

VLAN: Automatic

VMM port group: [Empty]

QoS: Select Bandwidth Template

DHCP Isolation: Off

TCP checksum: Off

Logical Switch: SW1

VLAN: Automatic

Operation: Add

- View VMM Port Name

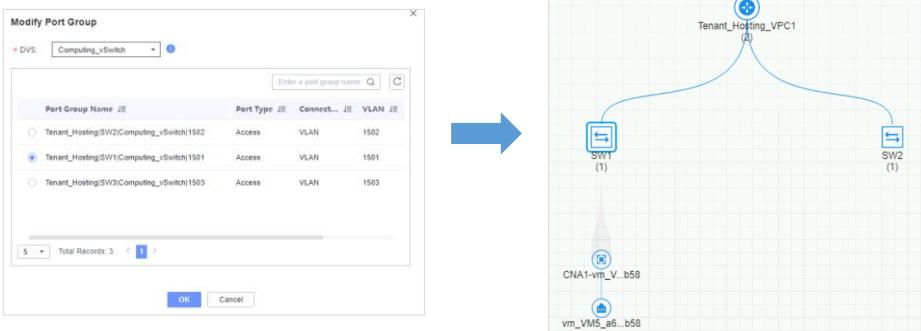
VMM	VDS/DVS	VLAN	DataCenter	VMM Port Group Name
FusionCompute	Computing_vSwitch	1501		Tenant_Hosting SW1 Computing_vSwitch 1501

Total records: 1

- The name of a port group is in the following format: tenant name|logical switch name|VDS name|VLAN ID.

Orchestration Server Access (2)

- The computing administrator creates a VM on the VMM and selects a port group to access the network. After the VM is started, iMaster NCE-Fabric automatically detects the logical port and user port connected to the VM.



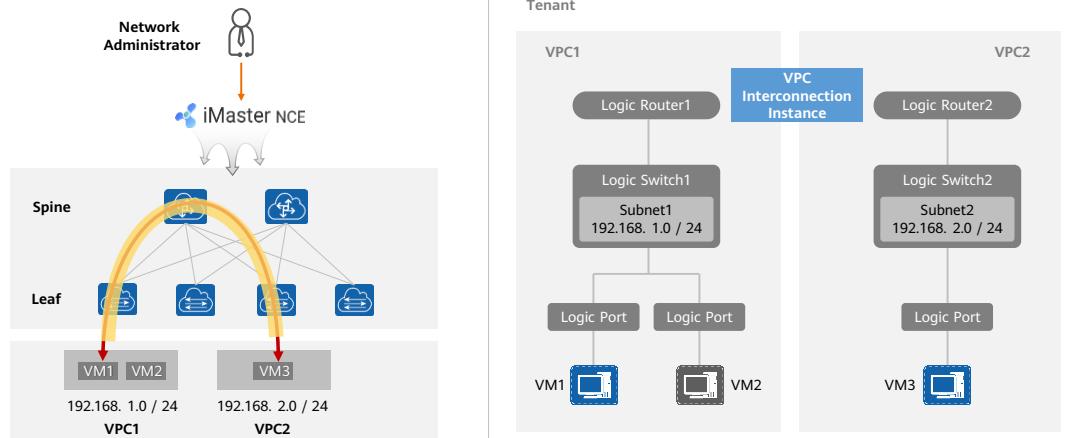
Contents

1. Deployment Process Overview
2. Pre-configuration
- 3. Service Provisioning**
 - Deploying Layer 2 and Layer 3 Basic Services
 - Deploying the VPC Interconnection Service**
 - Deploying Value-added Services
 - Deploying Microsegmentation and Service Chain
4. Easy Deployment

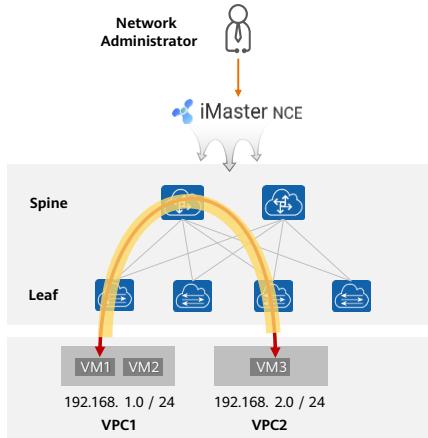
VPC Interconnection Service Overview

- By default, subnets under a VPC's logical routers communicate with each other at Layer 2 and Layer 3. However, networks between different logical routers, VPCs, and tenants cannot communicate with each other. To achieve such communication, you need to configure VPC interconnection.
- Based on the mutual access requirements, the VPC interworking scenarios are as follows:
 - Traffic not passing through the firewall.
 - Traffic passing through the firewall in only one direction.
 - Traffic passing through the firewall in both directions.

Traffic Model for Communication Between VMs Across VPCs (When Traffic Does Not Pass Through a Firewall)



Key Configurations for Communication Between VMs Across VPCs (When Traffic Does Not Pass Through a Firewall)



Key configuration: Configure VPC communication on iMaster NCE.

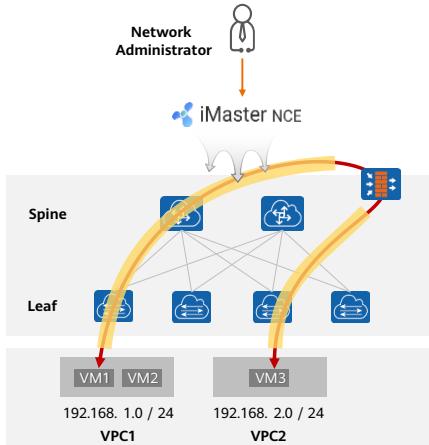
The screenshot shows the 'Create VPC Connection' dialog box. Key fields include:

- Name: vpc
- Source tenant: test
- Source VPC: router1
- Source router: router1
- Source subnet: 192.168.1.0/24
- Destination tenant: test
- Destination VPC: router2
- Destination router: router2
- Destination subnet: 192.168.2.0/24
- Source firewall: -Select-
- Source Egress gateway: GW220_221(Fabric01)
- Destination firewall: -Select-
- Destination Egress gateway: GW229(Fabric01)

44 Huawei Confidential

HUAWEI

Configuring Communication Between VMs Across VPCs (When Traffic Passes Through a Firewall)



Scenario Description

- A tenant deploys two different service systems that belong to different VPC logical networks. In terms of services, the two VPCs need to communicate with each other, and the inter-VPC traffic needs to pass through the firewall in one VPC. Therefore, the cross-VPC access service needs to be deployed.

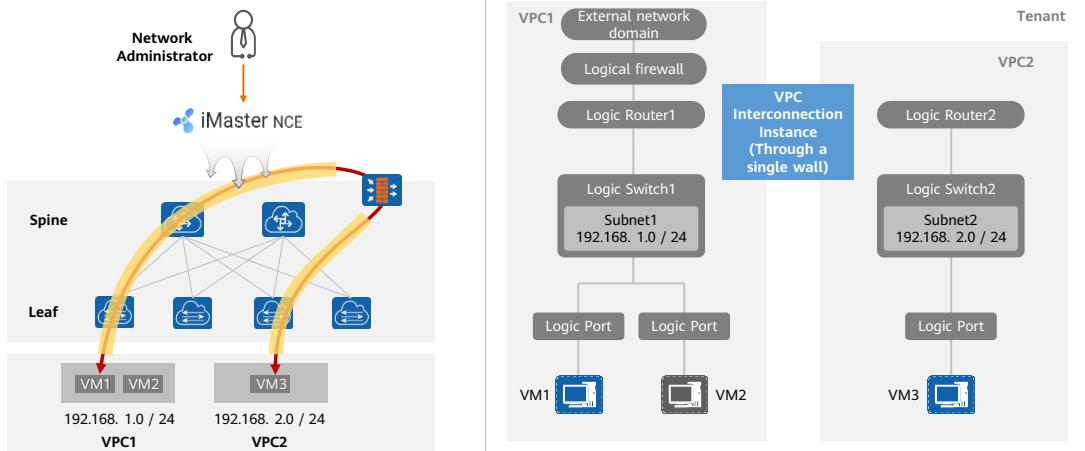
Configuration roadmap

- Create two VPCs and orchestrate the basic L2 and L3 networks in the VPCs. (For example, logical routers, logical switches, VMM mapping, and VM online).
- Create an external network domain and a logical firewall in VPC1, and configure internal and external links for the logical firewall.
- Create a security policy on the firewall to allow the subnets to communicate with each other.
- Create a VPC interworking instance and specify the logical firewall in VPC1 to implement interworking.

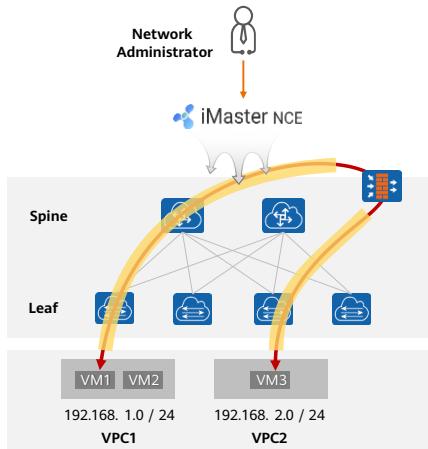


- If both firewalls need to pass through the firewall, configure the logical firewall in VPC2. The operations are the same as those in VPC1.

Logical Model for Configuring Communication Between VMs Across VPCs (When Traffic Passes Through a Firewall)



Key Configurations for Configuring Communication Between VMs Across VPCs (When Traffic Passes Through a Firewall)

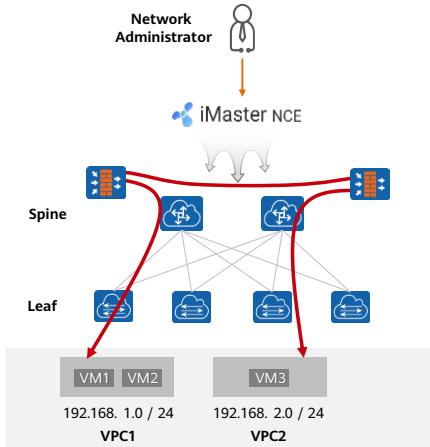


Key configuration: When configuring VPC communication on iMaster NCE, you need to specify the firewall.

Screenshot of the iMaster NCE interface showing the configuration for VPC1_VPC2:

- 名称: VPC1_VPC2
- 源组(2): DCN_Project5abbd6e137b4cfdf9b...
- 目的组(2): DCN_Project5abbd6e137b4cfdf9b...
- 源VPC: RouterA
- 目的VPC: RouterB
- 源逻辑路由器: RouterA
- 目的逻辑路由器: RouterB
- 源子网: 192.168.2.0/24
- 目的子网: 192.168.3.0/24
- 源防火墙: auto_van_bc4546393ccfaedf81ab
- 目的防火墙: SpineFabric01

Configuring Communication Between VMs Across VPCs (When Traffic Passes Through Firewalls)



Scenario Description

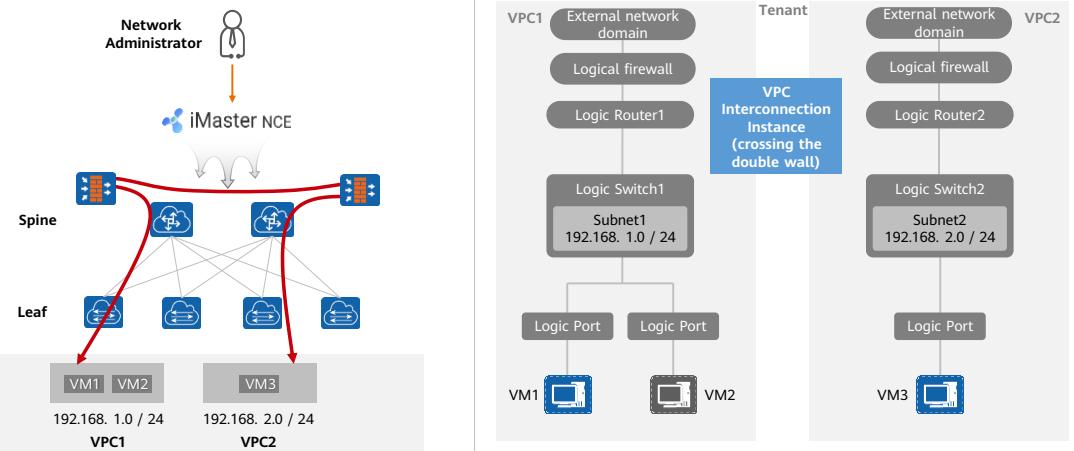
- A tenant deploys two different service systems that belong to different VPC logical networks. In terms of services, the two VPCs need to communicate with each other, and the inter-VPC traffic needs to pass through the firewall in one VPC. Therefore, the cross-VPC access service needs to be deployed.

Configuration roadmap:

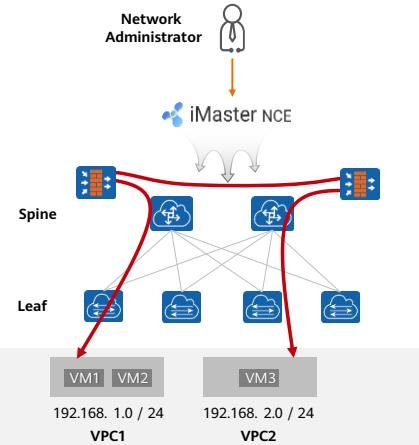
- Create two VPCs and orchestrate the basic L2 and L3 networks in the VPCs. (For example, logical routers, logical switches, VMM mapping, and VM online).
- Create an external network domain and a logical firewall in VPC1, and configure internal and external links for the logical firewall.
- Create a security policy on the firewall to allow the subnets to communicate with each other.
- Create a VPC interworking instance and specify the logical firewall in the VPC to implement interworking.



Logical Model for Configuring Communication Between VMs Across VPCs (When Traffic Passes Through Firewalls)



Key Configurations for Configuring Communication Between VMs Across VPCs (When Traffic Passes Through Firewalls)



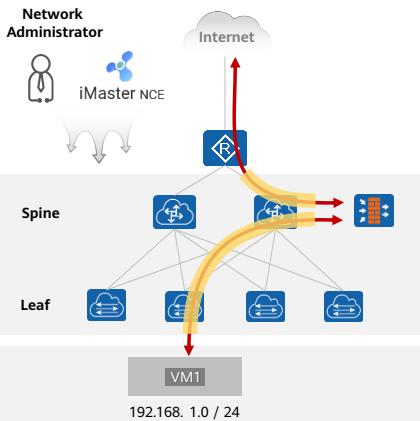
Key configuration: When configuring VPC communication on iMaster NCE, you need to specify the firewall.

The dialog box shows the configuration for creating a VPC connection named 'vpc'. It specifies source and destination tenants as 'test', source VPCs as 'router1' and 'router2', and destination VPCs as 'router2' and 'router1'. Source and destination subnets are set to '192.168.1.0/24' and '192.168.2.0/24' respectively. The 'Source firewall' field is highlighted with a red border and contains the value 'auto_vse_a4c1a5a92e2ea0b3046'. The 'Destination firewall' field contains 'auto_vse_cbfb0a9afac25fbabb5c'. A priority field is also present. At the bottom, there are 'Cancel' and 'OK' buttons.

Contents

1. Deployment Process Overview
2. Pre-configuration
- 3. Service Provisioning**
 - Deploying Layer 2 and Layer 3 Basic Services
 - Deploying the VPC Interconnection Service
 - Deploying Value-added Services**
 - Deploying Microsegmentation and Service Chain
4. Easy Deployment

Deploying the SNAT Service: Intranet VMs Access the Internet



Scenario Description

- The network segment of the internal hosts in the data center of company A is 192.168.1.0/24. The hosts need to access the Internet through SNAT.

52 Huawei Confidential

 HUAWEI

Configuration roadmap:

1. Create a tenant and a tenant VPC on iMaster NCE (Fabric).
 2. Create a logical router in the tenant VPC and add an IPv4 subnet.
 3. Create a logical switch in the tenant VPC and associate it with the logical router and subnet.
 4. Create VMM mappings in the tenant VPC and associate them with different logical switches.
 5. Create a VM on the VMM and connect the VM to the corresponding network.
 6. Create a logical VAS (firewall) in the tenant VPC and configure internal links.
 7. Create an external network domain in the tenant VPC, associate the domain with the created external gateway, and configure external links.
 8. **Create an SNAT policy in the tenant VPC and specify the SNAT type, source IP address, destination IP address, and public IP address.**
 9. **Create a security policy in the tenant VPC to allow the subnets or addresses for which SNAT needs to be performed.**
- Prerequisite:
 - Device discovery, global resource configuration, and interconnection resource configuration have been completed.
 - Fabric and L4-L7 resource pools have been created, associated with the resource pools, and roles of inter-device links have been configured.
 - An external gateway (of the L3 shared egress type) has been created and a public IP address has been configured for VM address translation.

- iMaster NCE-Fabric has been interconnected with VMM.

Deploying the SNAT Service: Configuring SNAT Policies and Security Policies

- 1 On the VPC1 orchestration page, click the FW Service tab and create an SNAT policy for the logical firewall.

Create NAT Policy

Policy Information

- Name: SNAT
- NAT type: NAT NAT64
- NAT model: Source address translation
- External network: External_Hosting
- Enable: ON
- Traffic profile: + Select

Pre-NAT Traffic Settings

- Source address: 192.168.1.0/24
- Destination address: any
- Service: any

Post-NAT Traffic Settings

- Source public IP address: 10.10.10.10

Cancel OK

2 For SNAT access to the Internet, you need to configure security policies to permit traffic. In the Tenant View, click the Security tab to go to the Security page.

Create policy

Policy Information

- Name: Security_Policy_SNAT
- Action: Permit
- Policy group: + Select

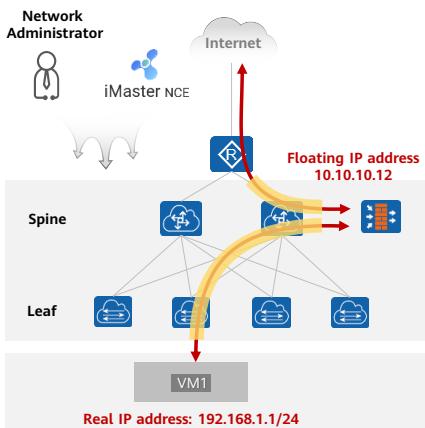
Policy Matching Condition

- Source VPC/External network: Tenant_Hosting_VPC1
- Destination VPC/External network: any
- Source addresses: 192.168.1.0/24
- Destination addresses: any
- Services: any

Cancel OK Confirm and Deploy

- For details, see the lab manual.

Deploying an EIP: Accessing Intranet Server from an External Network



Scenario Description

- The internal host (IP address: 192.168.1.1) in the data center of company A provides services externally. The external network accesses the floating IP address 10.10.10.12 to access the services provided by the internal host.

54 Huawei Confidential

 HUAWEI

• Configuration roadmap:

1. Create a tenant and tenant VPC on iMaster NCE (Fabric).
2. Create a logical router in the tenant VPC and add a subnet.
3. Create a logical switch in the tenant VPC and associate the logical router with the corresponding subnet.
4. Create VMM mappings in the tenant VPC and associate them with different logical switches.
5. Create a VM on the VMM and connect the VM to the corresponding network.
6. Create a logical VAS (firewall) in the tenant VPC and configure internal links.
7. Create an external network domain in the tenant VPC, associate the domain with the created external gateway, and configure external links.
8. **Create an EIP policy in the tenant VPC and specify the working mode, floating IP address, and fixed IP address of the EIP.**
9. **Create a security policy in the tenant VPC to allow the subnet or IP address for which the EIP needs to be executed.**
- EIP is also called floating IP address.
- Prerequisite:
 - Device discovery, global resource configuration, and interconnection resource configuration have been completed.
 - Fabric and L4-L7 resource pools have been created, associated with the resource pools, and roles of links between devices have been configured.
 - An external gateway (of the L3 shared egress type) has been created, and a public IP address has been configured for VM address translation.
 - iMaster NCE-Fabric has been interconnected with VMM.

Deploy EIP: Configure EIP Policies and Security Policies

- 1 On the VPC1 orchestration page, click the FW Service tab and configure an EIP policy for the logical firewall.

Create EIP

* Name: EIP

* Mapping mode: IP Address Mapping Protocol/Port Mapping

* Service mode: Unidirectional Bidirectional Bidirectional enhanced

* External network: External_Hosting [+ Select](#)

* Floating IP address: 10.10.10.12

* Fixed IP address: 192.168.1.1

Traffic profile: [+ Select](#)

Cancel OK

- 2 In the EIP scenario, you need to configure security policies to permit traffic. In the tenant view, click the Security tab to go to the Security page.

Create policy

Policy Information

* Name: Security_Policy_EIP

Action: Permit Deny

Policy group: [+ Select](#)

Policy Matching Condition

Source VPC/External network: any [+ Select](#)

Destination VPC/External network: any [+ Select](#)

Source addresses: any [+ Select](#)

Destination addresses: 192.168.1.2/32 [+ Select](#)

Services: any [+ Select](#)

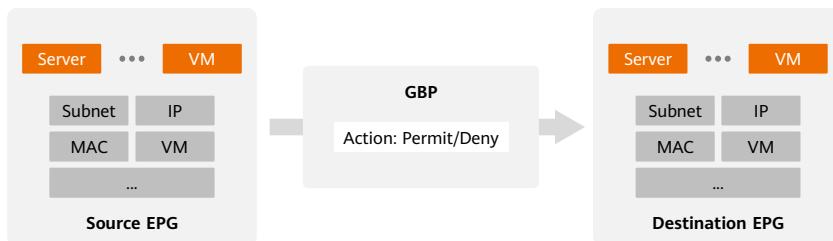
- For details, see the lab manual.

Contents

1. Deployment Process Overview
2. Pre-configuration
- 3. Service Provisioning**
 - Deploying Layer 2 and Layer 3 Basic Services
 - Deploying the VPC Interconnection Service
 - Deploying Value-added Services
 - Deploying Microsegmentation and Service Chain**
4. Easy Deployment

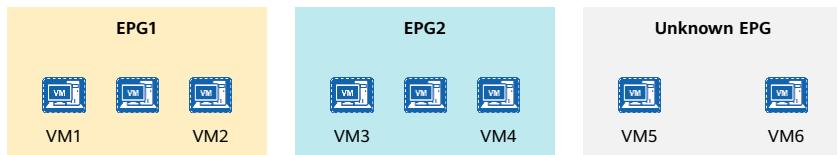
Microsegmentation Overview

- Microsegmentation is a security isolation technology that groups DC services based on certain rules and deploys policies between groups to implement traffic control.
- Traditionally, subnets are created for DCs based on coarse-grained granularities such as VLAN IDs or VNIs. Microsegmentation supports more fine-grained and flexible grouping modes, for example, grouping based on IP addresses, MAC addresses, and VM names. This can further narrow down security zones to implement more fine-grained service isolation and enhance network security.
- Microsegmentation implements service isolation between different servers of a VXLAN network and ensures secure management and control for the VXLAN network. In addition, the configuration and maintenance of microsegmentation are simple, significantly reducing the configuration and maintenance costs.



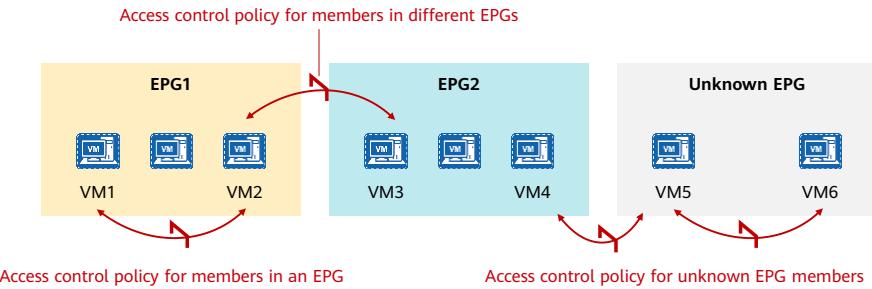
Basic Concepts of Microsegmentation - EPG

- End point group (EPG): A group of entities that carry services, such as servers and VMs. EPGs can be defined based on IP addresses, MAC addresses, VM names, and applications.
- After service entities on a network are allocated to EPGs, the VMs are classified based on the EPG:
 - Unknown EPG member: VMs that do not belong to any EPG (for example, VM5 and VM6).
 - EPG member: VMs that belong to any EPG (for example, VM1, VM2, VM3, and VM4).
 - Members in the same EPG: VMs that belong to the same EPG (for example, VM1 and VM2, or VM3 and VM4).
 - Members in different EPGs: VMs that belong to different EPGs (for example, VM1 and VM3).

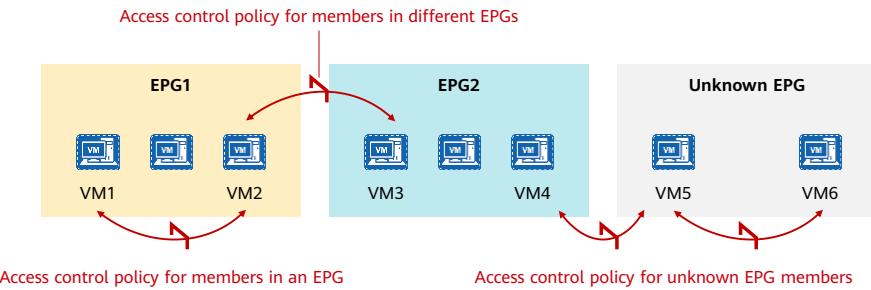


Basic Concepts of Microsegmentation - GBP

- Group-based policy (GBP): policy for traffic control within an EPG and between EPGs. A GBP can be configured based on EPGs, protocol numbers, and port numbers, which specifies the policies within an EPG, between EPGs, and between a known EPG and an unknown EPG.



Basic Concepts of Microsegmentation - Default GBP Policies

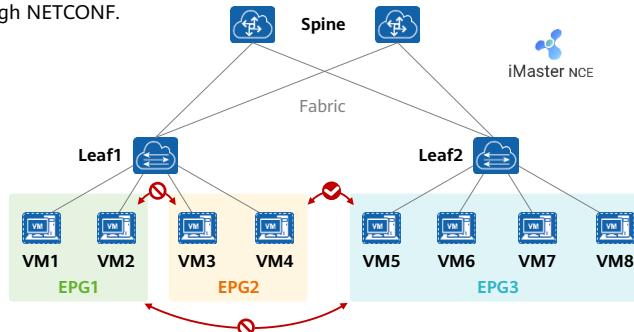


- ① By default, the access control policy for an unknown EPG member is permit. That is, unknown EPG members can communicate with each other, and an unknown EPG member and a known EPG member can also communicate with each other.
- ② By default, the access control policy for an EPG member is deny. That is, members in different EPGs cannot communicate with each other.
- ③ The default access control policy for members in an EPG varies according to CE switch models.

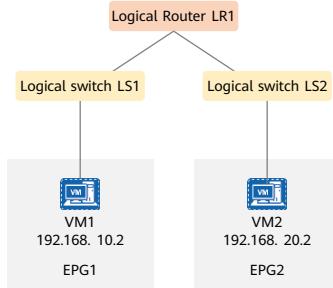
- Default access control policy for members in an EPG:
 - For the CE6857EI, CE6857E, CE6857F, CE6865EI, CE6865E, CE8861EI, and CE8868EI, the default access control policy is always permit for members in an EPG, which cannot be modified. That is, members in the same EPG can communicate with each other.
 - For the CE6881, CE6881K, CE6881E, CE6863, CE6863E, CE6863K, CE6820, and CE5881, the default access control policy is always none for members in an EPG, which can be modified. That is, access control is not performed for members in the same EPG. In this case, the devices perform access control for members in the EPG according to the default access control policy (policy 2 on this slide).

Microsegmentation Application Scenarios

- You can use microsegmentation to allocate servers or VMs to different EPGs and specify the GBP for members in different EPGs to implement traffic control between service functions (SFs).
- You can use a CE switch alone or use a CE switch and iMaster NCE-Fabric together to implement microsegmentation. If iMaster NCE-Fabric is used, it configures EPGs and GBPs and delivers the configurations to the CE switch through NETCONF.



Configuring Microsegmentation



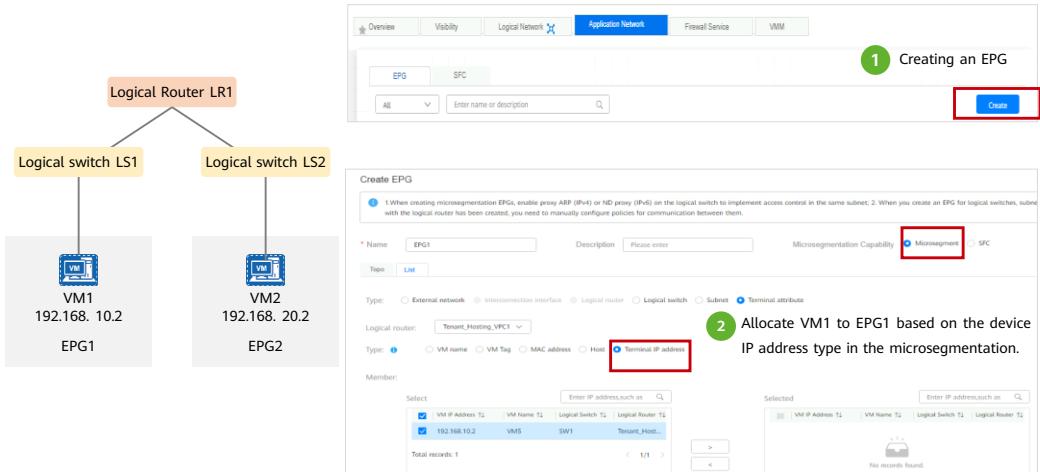
Scenario description:

- To isolate two VMs (such as VM1 and VM2) that are not dependent on firewalls within a VPC network, you can use iMaster NCE-Fabric to deploy microsegmentation. Associate VM1 with EPG1 and VM2 with EPG2.
- If necessary, you can create a service chain policy to allow mutual access between specified protocols and ports.

• Configuration roadmap:

1. Create a tenant and a tenant VPC on iMaster NCE (Fabric).
2. Create a logical router in the tenant VPC and add a subnet.
3. Create a logical switch in the tenant VPC and associate the logical router with the corresponding subnet.
4. Create VMM mappings in the tenant VPC and associate them with different logical switches.
5. Create a VM on the VMM and connect the VM to the corresponding network.
6. **Create two EPG servers of the host type.**
7. **Create a service chain template for microsegmentation.**
8. (Optional) Configure a service chain policy to allow the required communication protocols and ports.

Creating an EPG



63 Huawei Confidential

- Repeat the same steps to create another EPG and add VM2 as a member.

Creating a Service Chain

- To allow certain protocols and ports to pass between two VMs, create a service chain and related policies. When a microsegmentation-based SFC is created, SF nodes cannot be used between the source EPG and the destination EPG. That is, no value-added service is supported.

Create Service Chain

* SFC name

Description

* microsegmentatio... MicroSegment SFC

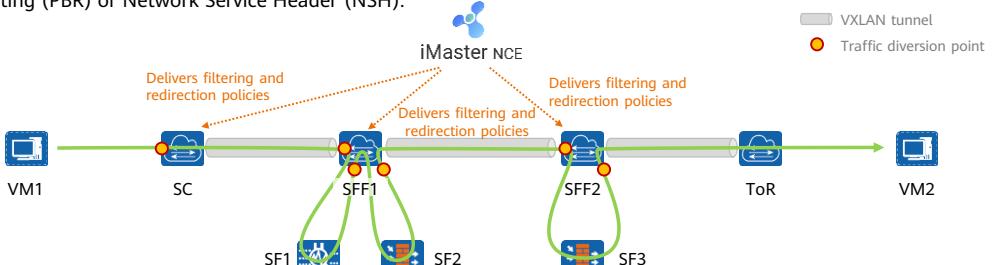
SFC Profile

! Need to complete the configuration of all SFs before submitting.

```
graph LR; SourceEPG[Source EPG] --- DestinationEPG[Destination EPG]
```

Service Chain Overview

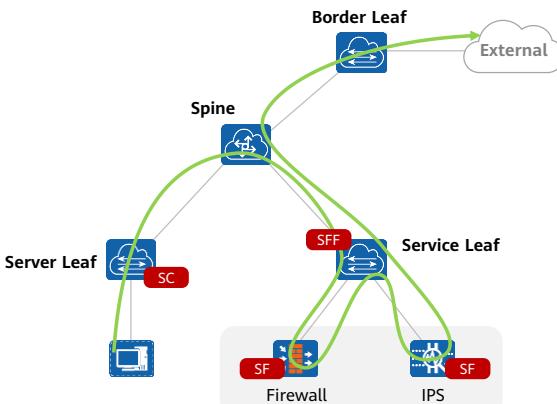
- Service function chain (SFC) is a technology that provides ordered services for the application layer.
 - After the SFC path is defined, the matching traffic can pass through the specified VAS device in sequence. (e.g., firewall, load balancing, in-depth detection, intrusion prevention, etc.) so as to obtain corresponding value-added services in turn.
 - Service chains are orchestrated on the Agile Controller-DCN, and may be implemented by using Policy-Based Routing (PBR) or Network Service Header (NSH).



65 Huawei Confidential

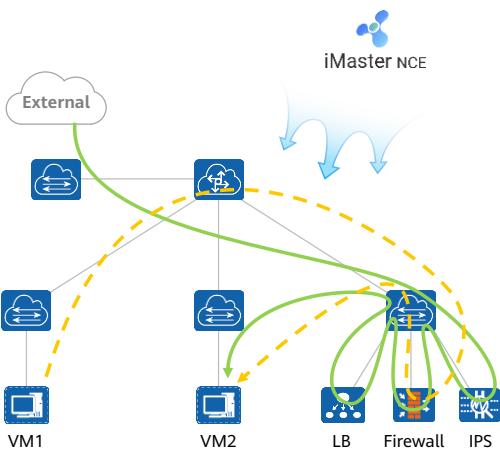
HUAWEI

Basic Concepts of SFC



- **SFC domain:** A domain where SFs are deployed.
- **EPG:** A set of service units with the same features. SFC defines the service function chain between a pair of EPGs, including an SFP and the SFC policy.
- **Service classifier (SC):** An SC is located at the ingress of an SFC domain. After packets enter the SFC domain, the SC classifies the packets.
- **SF:** SFs are devices that provide VASs, such as firewalls and load balancers.
- **Service function forwarder (SFF):** An SFF forwards the packets received from a network to its associated SFs.
- **SFP:** An SFP is a packet path calculated based on configurations.

SFC Application Scenarios



Security protection between the data center network and external networks is the core of network security. The external north-south access traffic can be flexibly diverted to different SFs (marked by the green line) based on the defined SFC to implement functions such as address translation and security filtering for internal and external networks.

When service units of different security levels need to communicate with each other, east-west traffic can be flexibly defined to pass through SFs (marked by the yellow line) in the resource pool in sequence based on user requirements for security protection.

- You can use a switch alone or use a switch and iMaster NCE-Fabric together to implement SFC. The controller orchestrates SFs, configures an SFP, and delivers the SFP configurations to the SC and SFFs (Huawei CE series switches) through NETCONF interfaces.

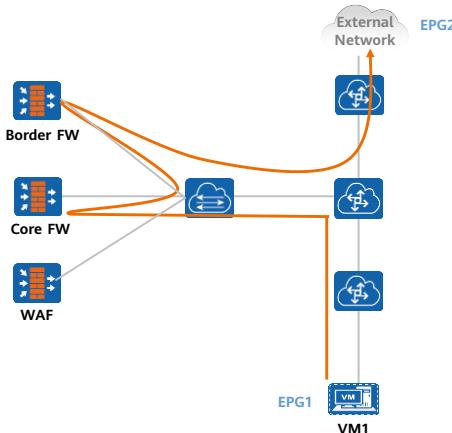
Configuring a Service Chain

Scenario description:

When intranet users need to access the Internet, the traffic orchestrated by intranet users passes through the core firewall for value-added service operations such as service isolation and security control, and then enters the border firewall for address translation.

Configuration roadmap:

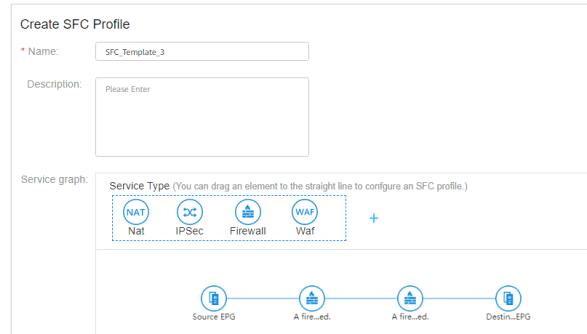
1. On iMaster NCE (Fabric), create tenants and tenant VPCs and orchestrate logical networks.
2. Create a service chain template for intranet users to access the Internet. The template passes through Firewall 1 (the core firewall provides security policy filtering) and then through Firewall 2 (the border firewall provides SNAT).
3. Create an EPG in the tenant VPC, set logical switches 1 and 2 as EPG1 (source EPG), and set the external network domain as EPG2 (destination EPG).
4. Create an SFC, associate it with an SFC template, and redirect traffic to the logical firewall.
5. Configure security filtering policies on logical firewall 1 and SNAT on logical firewall 2.



- When orchestrating logical networks, you need to create logical firewall 1 and configure internal links between the and logical routers. Create a domain between logical firewall 2 and the external network, and configure internal links between logical firewall 2 and the logical router and external links between logical firewall 2 and the external network domain.

Creating a Service Chain Template

- Creating an SFC template is to set the SFC path, that is, the service nodes that the source EPG server passes through and the sequence of the service nodes that the source EPG server passes through the destination EPG server.
- Create a service chain template and drag the required VAS node icons between two EPG servers based on service requirements. In this example, drag the firewall node icons between two EPG servers.

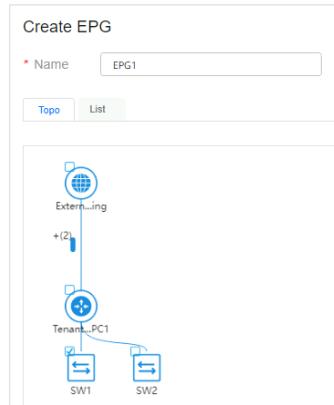


69 Huawei Confidential



Creating an EPG

- Take EPG1 as an example. Enter EPG1 in Name and select SW1 as the source EPG of the service chain. Create EPG2 in the same way. In the topology, select the external network domain Ext1 as the source EPG of the service chain.



Creating a Service Chain

- When creating a service chain, select the created template and set related parameters.

1 Setting the source and destination EPG servers

SFC Profile
Need to complete the configuration of all SFs before submitting.

Source EPG
* Source Router: Tenant_Hosting_VPC1
* Source EPG(Provider EPG): EPG1

Destination EPG
* Destination Router: Tenant_Hosting_VPC1
* Destination EPG(Provider EPG): EPG2

2 In the Service Function Node Configuration area, click and select a logical VAS created in the VPC as the SF role in the service chain.

SFC Profile
Need to complete the configuration of all SFs before submitting.

SF configuration
* Logical VAS: Hosts
Hosts Hosts

3 Configure policy actions.

Policy Configuration
Policy style: Existing Policy
Action: Forward

Similarities and Differences Between the Service Models & Basic Concepts of Microsegmentation and SFC

EPG	Service Model
<p>1. Granularity:</p> <ul style="list-style-type: none">In the SFC scenario, EPGs can be configured only based on logical routers, logical switches, and external network domains.In the microsegmentation scenario, EPGs can be configured based on logical switches, external network domains, subnets, and terminals with a finer granularity. In this scenario, terminals support four matching modes —matching based on the prefix and suffix of the host name, matching based on the prefix and suffix of the VM name, matching based on the MAC address, and matching based on the IP address. <p>2. Default isolation:</p> <ul style="list-style-type: none">In the microsegmentation scenario, automatic isolation is implemented after EPGs are configured.In the SFC scenario, after EPGs are created, you still need to configure SFC to implement traffic isolation or diversion.	<p>1. Significant differences related to the support for VASs:</p> <ul style="list-style-type: none">In the microsegmentation scenario, SF nodes cannot be deployed between the source EPG and destination EPG. That is, no VAS is supported. In the SFC scenario, multiple VASs are supported. <p>2. Both SFC and microsegmentation support east-west traffic and north-south traffic:</p> <ul style="list-style-type: none">In an SFC model for north-south traffic, the source EPG is a logical switch, subnet, or terminal (discrete IP address), and the destination EPG is an external network domain.In an SFC model for east-west traffic, the source and destination EPGs must be connected to the same logical router.

72 Huawei Confidential



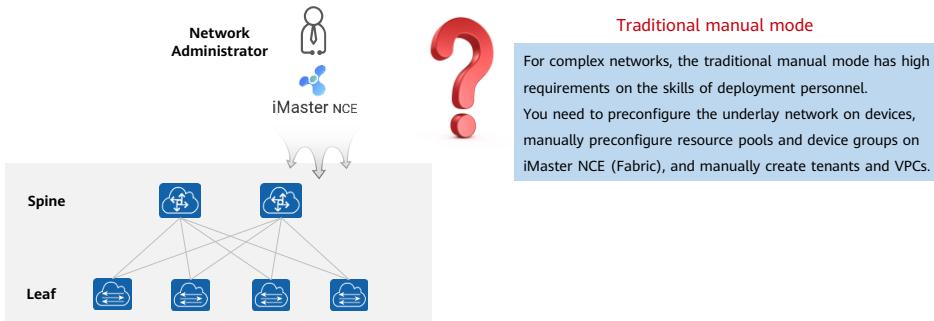
- In common service scenarios, a microsegmentation configuration model saves more ACL resources than an SFC configuration model.
- In the microsegmentation scenario, traffic between the source EPG and destination EPG cannot be redirected to SF nodes (except for north-south traffic, which usually passes through firewalls based on route forwarding).
- The configuration models of microsegmentation and SFC vary in that the SFC model adopts 5-tuple-based policies, while the microsegmentation model uses EPG-based policies.

Contents

1. Deployment Process Overview
2. Pre-configuration
3. Service Provisioning
- 4. Easy Deployment**

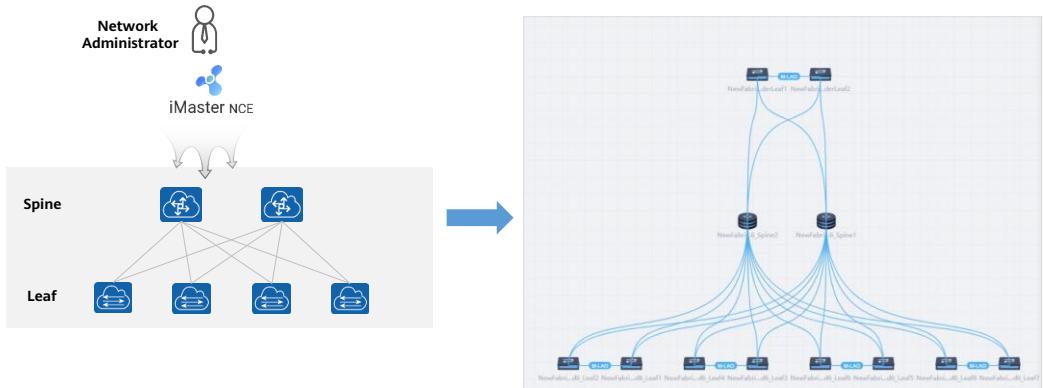
Overview of Easy Mode (1)

- The preceding figure shows the traditional deployment process. To facilitate quick network deployment, Huawei CloudFabric solution provides the Easy deployment function. That is, you can go to the dedicated page for Easy in iMaster NCE-Fabric. Based on the navigation tree, you can complete zero-touch deployment (ZTP) for switches, create tenants and VPCs, and provision basic network services in the VPC.



Overview of Easy Mode (2)

- When the Easy deployment mode is used, the network planning scheme can be automatically generated based on the number of devices and cable connections.

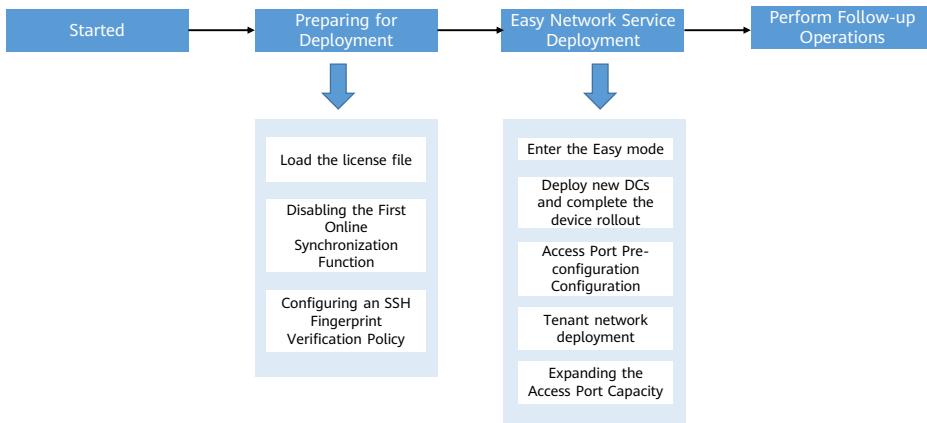


Easy Mode vs Manual Mode

Comparison Item	Easy Mode	Manual Mode
Networking Requirements	Low	Low
Service network connection planning	Manually plan the connection, or automatically generate the connection plan by iMaster NCE (Fabric).	Manually plan cable connections in the LLD.
Pre-Configure the Underlay Network on Switches	iMaster NCE-Fabric automatically generates planning parameters and configuration script files.	Parameters need to be manually planned and configured one by one.
Firewall and SecoManager pre-configuration	<ul style="list-style-type: none"> Underlay network pre-configuration on firewalls needs to be manually performed. SecoManager pre-configuration (Security menu) requires manual configuration. 	
Creating and configuring fabric resource pools, managed devices, device groups, device roles, tenants, and VPCs on iMaster NCE (Fabric)	Created by iMaster NCE-Fabric	Manually configure items one by one in the Configuration Wizard menu.
Layer 2 and Layer 3 Basic Service Orchestration in a VPC	Directly orchestrate on the Easy page.	Orchestrate on the Service Provisioning page.
Cross-VPC communication (through firewalls and not through firewalls)	Orchestrate on the Service Provisioning page.	
North-south access to external networks through firewalls (including SNAT and DNAT)	Orchestrate on the Service Provisioning page.	

Deployment Process

- The following figure shows the deployment process in Easy mode.



77 Huawei Confidential



- Preparations before deployment:

- Entering the Easy mode: Enter the dedicated Easy mode on iMaster NCE-Fabric to configure the network.
 - Create a DC to complete the switch online. On the page for creating a DC, set related parameters and use ZTP to quickly bring the switch online.
 - Access port pre-configuration: When a server is connected to a server leaf node, you can use the access port pre-configuration function to configure Eth-Trunk IDs, M-LAG IDs, and LACP mode for the ports in batches.
 - Tenant network deployment:
 - Configure the tenant, VPC, associated external gateway, subnet, and access port based on the actual service plan.
 - Enable an external gateway of the L3 shared egress type and configure interconnection interfaces to prepare for southbound and northbound service provisioning.
 - During tenant network deployment, you can associate VLANs and port groups to logical ports on logical switches in batches, improving configuration efficiency. Therefore, you need to create a port group before configuring the overlay network.
 - Access port capacity expansion: Create logical ports in batches as required to expand server ports.
- Follow-up operations: After the Easy network is deployed, the basic VPC network has been provisioned. If you need to deploy other advanced services, such as the firewall service, inter-VPC access service, and microsegmentation service, you can provision the services by yourself.

Quiz

1. (Multiple-answer question) Which elements are involved in the out-of-band networking of the ZTP function in the CloudFabric solution? ()
 - A. Independent management switch
 - B. Root device
 - C. iMaster NCE-Fabric
 - D. Spine and leaf nodes

1. ACD

Summary

- In the CloudFabric computing scenario, the network administrator is only responsible for network setup and service orchestration, and the VMM platform is managed and maintained by the computing management personnel.

Thank you.

把数字世界带入每个人、每个家庭。

每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.



CloudFabric Intelligent Data Center Network O&M Solution



Foreword

- As technologies such as cloud computing, big data, and artificial intelligence develop continuously and gain popularity in commercial use, the digital transformation of enterprises is deepening. However, traditional data centers (DCs) lag behind in this development trend, making cloud-based transformation become a must-have. However, the rapid increase in the scale and traffic of DCs has brought difficulties and challenges in network management and service operations. The traditional manual O&M model gradually becomes ineffective in terms of complicated application migration strategies, unstable service experience, difficult fault locating, and large-scale management of massive security policies.
- This course describes the CloudFabric intelligent data center network (DCN) O&M solution. With iMaster NCE-Fabric and iMaster NCE-FabricInsight, the solution can overcome challenges in the traditional passive O&M model and eliminate difficulties in fault locating, providing users with ubiquitous network application and network assurance.

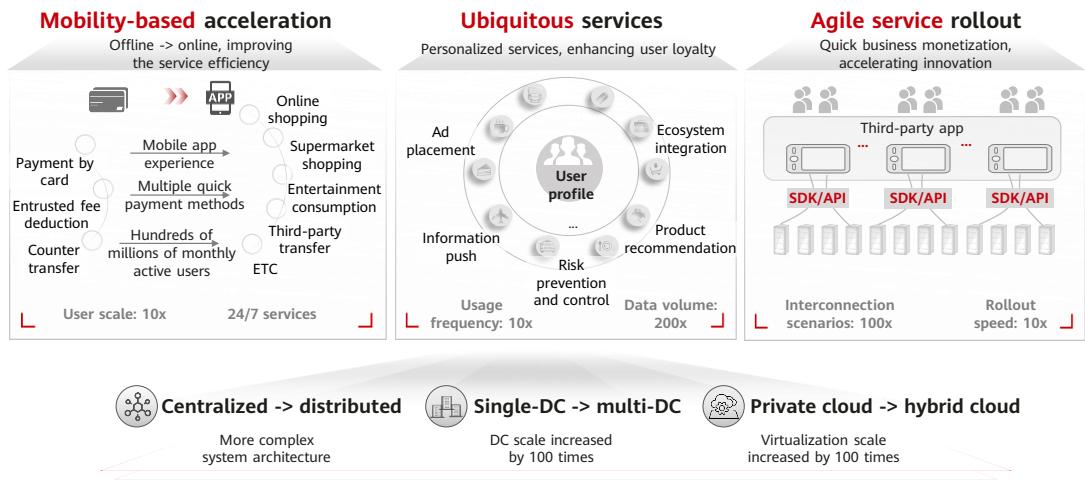
Objectives

- On completion of this course, you will be able to:
 - Describe pain points of the traditional DCN O&M and the Reason Why Intelligent O&M Is Required.
 - Describe O&M functions and features of iMaster NCE-Fabric.
 - Describe application scenarios of iMaster NCE-FabricInsight.
 - Describe main functions and features of iMaster NCE-FabricInsight.

Contents

- 1. DCN O&M Challenges and CloudFabric Intelligent DCN O&M Solution**
2. iMaster NCE-Fabric
3. iMaster NCE-FabricInsight

DCN Evolution to a Multi-Cloud and Multi-DC Mode



5 Huawei Confidential

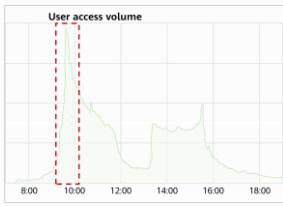
HUAWEI

- Note:
 - ETC: Electronic Toll Collection
 - SDK: Software Development Kit
 - API: Application Programming Interface

Challenges Faced by Traditional Manual O&M

Difficult health check

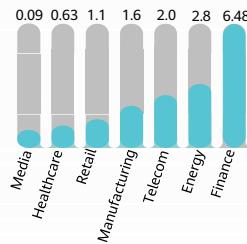
Fluctuating securities market, resulting in the daily needs to cope with service peaks.



It takes **three person-hours** to perform routine inspection before the market opens every day. This increases difficulties in confidently keeping up with the general market trends.

Difficult fault locating

Hundreds of millions of cross-bank transactions per day, requiring 24/7 uninterrupted services.

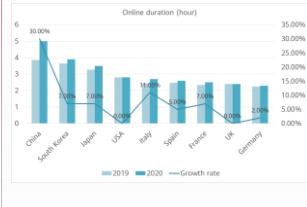


Survey on loss caused by fault-triggered interruptions ①

The complicated architecture results in difficult fault locating. It takes **76 minutes** on average to locate a fault.

Difficult network change

Enormous increase in Internet traffic, requiring network changes every week.



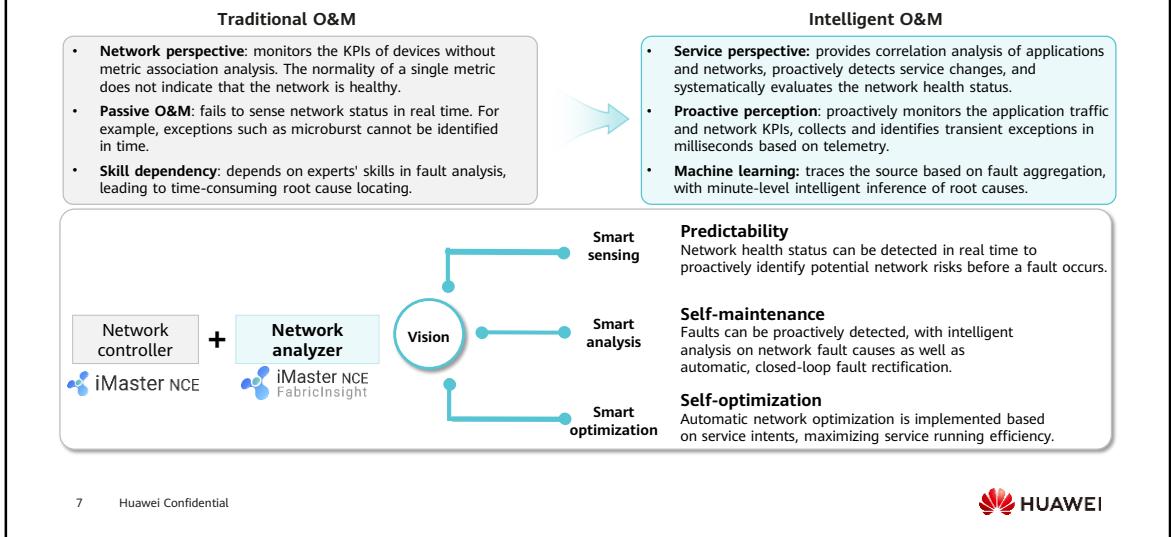
About **70%** of network faults are caused by human errors as changes are manually compared and verified.

6 Huawei Confidential

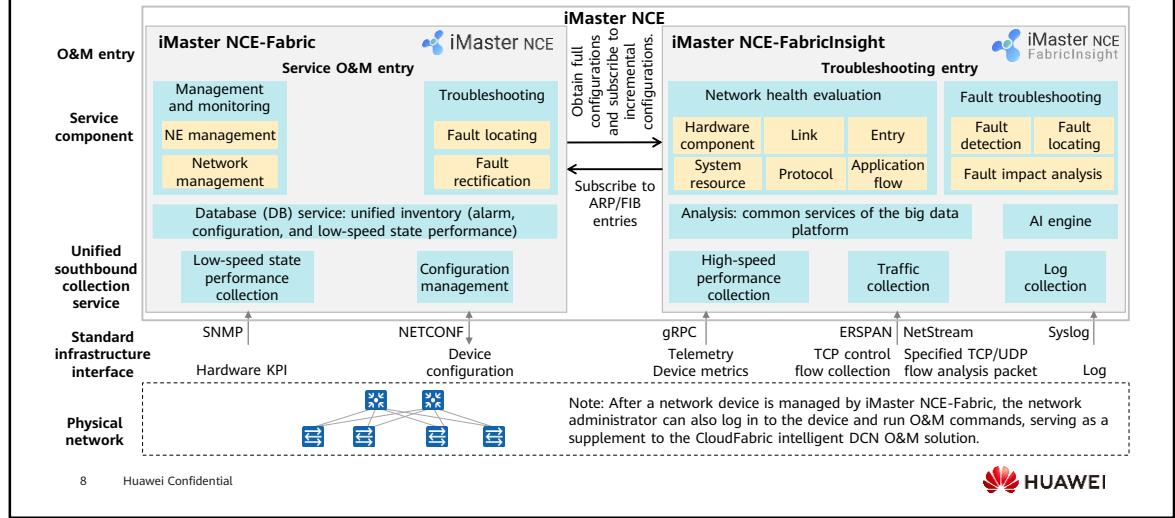
HUAWEI

- As applications are migrated dynamically and traffic increases sharply, DC O&M needs to be intelligent.
 - The fault domain expands with the increase in DC scale.
 - The boundaries of virtual networks extend to servers (such as vSwitches), blurring the O&M boundary between networks and IT systems.
 - It is more and more difficult to implement traditional O&M methods as the network needs to dynamically detect virtual machine (VM) migration and elastic scaling of applications, network configurations change frequently, there is a surge in traffic, and the application policies and mutual access relationships in DCs are increasingly complex.
 - To improve user experience and ensure high reliability of key services, faults need to be located and rectified in real time.
- Note:
 - Source ①: *Network Computing, the Meta Group and Contingency Planning Research*
 - Source ②: *App Annie*

Vision of the CloudFabric Intelligent DCN O&M Solution

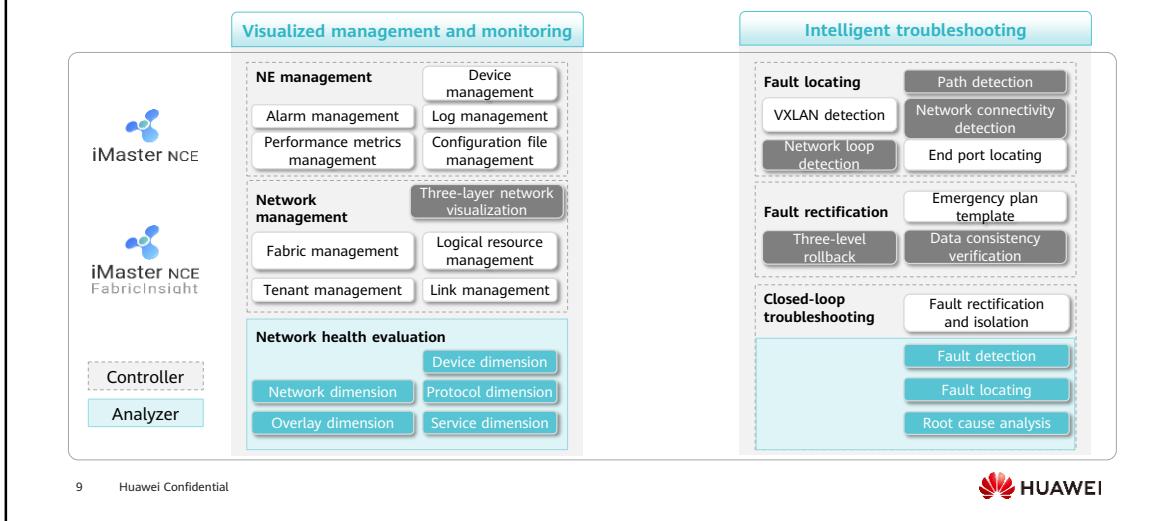


Overall Architecture of CloudFabric Intelligent DCN O&M Solution



- In the CloudFabric solution, iMaster NCE-Fabric provides some O&M capabilities, including network management, path detection, network reachability verification, and "1-3-5" fault rectification.
- Based on Huawei big data platform, iMaster NCE-FabricInsight receives data reported by multiple types of network devices, analyzes network data using intelligent algorithms, quickly detects network faults and O&M risks, quickly locates network faults, and displays key network events, providing a decision-making basis for network O&M.

O&M Function Panorama of the CloudFabric Intelligent DCN O&M Solution



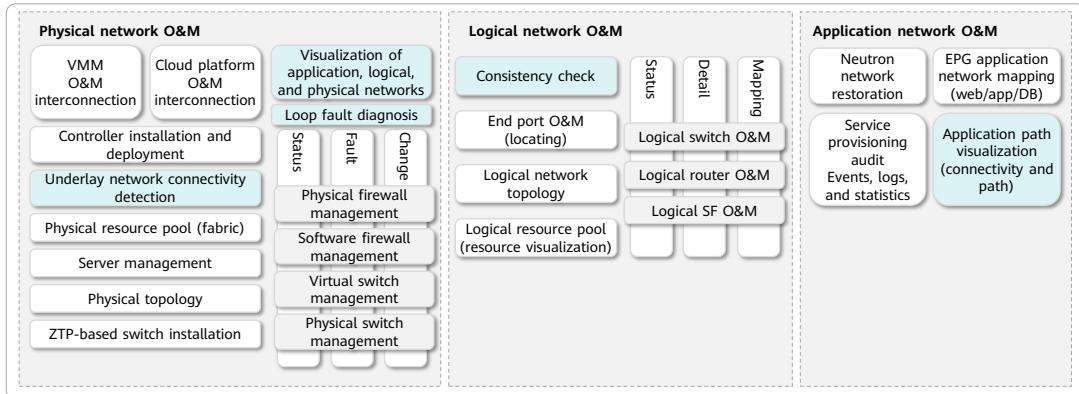
- An intelligent O&M system consists of the controller and analyzer. This course describes some O&M features.

Contents

1. DCN O&M Challenges and CloudFabric Intelligent DCN O&M Solution
2. **iMaster NCE-Fabric**
3. iMaster NCE-FabricInsight

iMaster NCE-Fabric

iMaster NCE-Fabric O&M panorama



11 Huawei Confidential



- iMaster NCE-Fabric centrally manages and controls cloud DCNs and provides automatic mapping from applications to physical networks, resource pool deployment, and visualized O&M, helping customers build service-centric dynamic network service scheduling capabilities.
- In addition to network planning and deployment, iMaster NCE-Fabric also provides DCN service O&M, including: topology visualization, loop detection, path detection, traffic statistics collection, three-level rollback, and data consistency verification.

iMaster NCE-Fabric O&M Panorama

Physical network O&M

- Service provisioning phase:
 - Physical resource pool management: supports the management of firewalls, load balancers, and DHCP servers.
 - **Physical network topology**
- O&M phase: monitoring and troubleshooting
 - **Loop detection**
 - Log management: During system running, iMaster NCE-Fabric can generate logs about system management operations (management logs) and system running (run logs), facilitating auditing and fault locating.
 - Device management: supports the management of hardware switches and firewalls, including device replacement and deletion.

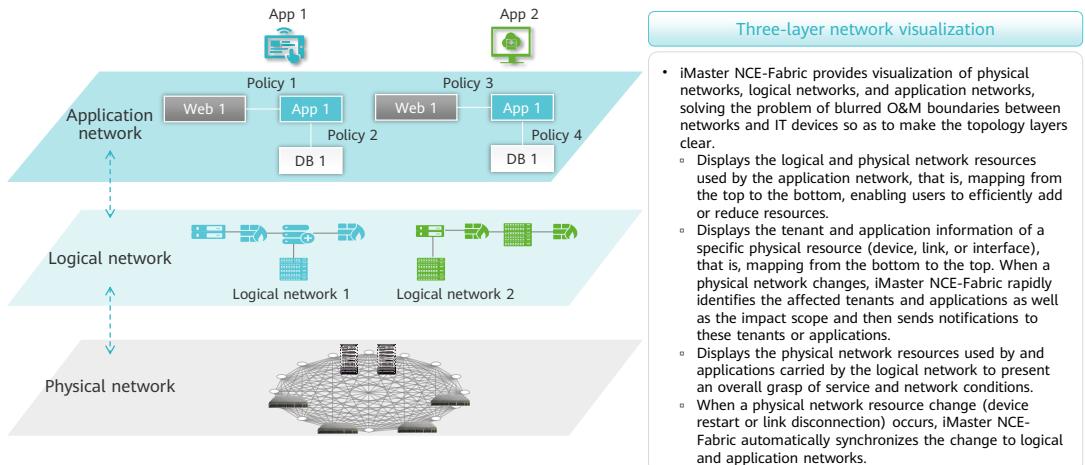
Logical network O&M

- Service provisioning phase:
 - **Logical network topology**
- O&M phase: monitoring and troubleshooting
 - **Consistency check**
- O&M phase: network quality
 - Logical interface performance monitoring: displays the performance data of logical switches, provides current performance data for users to query, and displays the tenant name, logical switch information, logical interface, physical interface, device IP address, number of sent/received packets, and number of sent/received bytes.

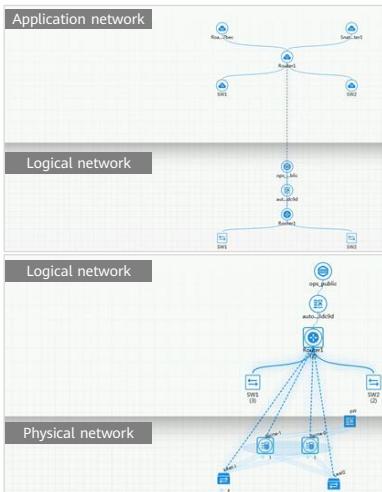
Application network O&M

- Service provisioning phase:
 - **Application network topology**
 - **Service connectivity detection**
- O&M phase: monitoring and troubleshooting
 - **Network path detection**

Three-Layer Network Visualization (1)



Three-Layer Network Visualization (2)



**100% service visualization:
mapping of the physical
topology to the logical
topology, and of the logical
topology to the application
network topology**



O&M tips

Application network -> logical network -> physical network
Query the logical and physical network resources used by tenants.
Add or reduce resources in a timely manner.

Application network -> logical network -> physical network
Query the applications running on physical networks.
Reversely identify the impact scope of physical network faults.

Application network -> logical network -> physical network
Query the physical resources used by logical networks.
Query the applications running on logical networks.
Present intuitive insights into service and network conditions.

Service Connectivity Detection

- Connectivity detection is used to detect the network connectivity between two VMs. With the simulation of Address Resolution Protocol (ARP) requests and ping processes, ARP request or Internet Control Message Protocol (ICMP) request packets are constructed and sent from the source VM to the destination VM. Network connectivity between two VMs is determined by checking whether the source VM receives the response packets from the destination VM.

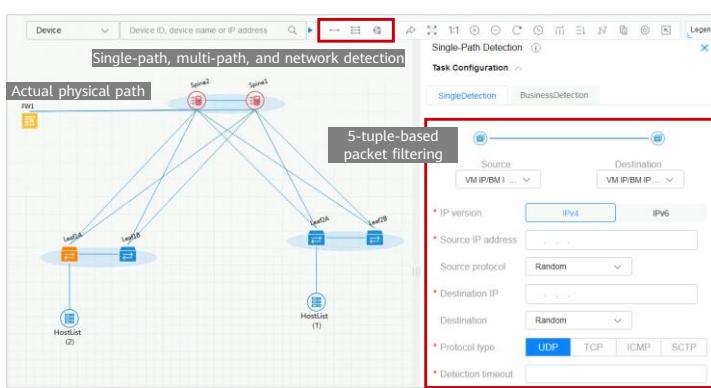
When the **Execution Result** is **Packet Loss Ratio 0%**, the source node and destination node are reachable to each other and network connectivity between the source and destination nodes is normal.

Execution Result											Execution Result
Defecto...	Source EndPort	Destination IP Address	Execution T...	Number of...	packets Sent	packets Rec...	Start Ti...	End Time	Interval (s)	Execution Time	Execution Result
IP	[REDACTED]	[REDACTED]	One-Time	10	10	10		3-19 11:42:43			Connected, Packet Loss Rat...

When the **Execution Result** is **Fail Go to Single-path Detection**, the source node and destination node are unreachable to each other and the path between the source node and destination node may be disconnected. Click **Single-path Detection**. The system then automatically switches to the single-path detection page to perform path detection.

Execution Result											Execution Result
Defecto...	Source EndPort	Destination IP Address	Execution T...	Number of...	packets Sent	packets Rec...	Start Ti...	End Time	Interval (s)	Execution Time	Execution Result
IP	[REDACTED]	[REDACTED]	One-Time	10	10	0		3-19 11:41:37			Fail Go to Single-path Detection

Service Path Visualization



Service path visualization

- iMaster NCE-Fabric can display the real physical paths of services based on application and logical networks. When the physical network is decoupled from the logical network, iMaster NCE-Fabric can quickly locate network faults, and detect and rectify unexpected service interruptions.
- Service path visualization provides the following functions:
 - Single-path detection
 - Multi-path detection
 - Network loop detection

Network Path Detection

- Path detection simulates the actual forwarding path of packets.

Single-path detection

- Single-path detection traces the actual physical paths between VMs, BMSs, containers, or network devices, and checks whether service flows are interrupted.
- Detection principle: iMaster NCE-Fabric sends a Packet-Out message to the source CE switch through an OpenFlow channel. This message simulates a service flow. 5-tuple information (including source IP address, destination IP address, source port, destination port, and protocol) and MAC address of this service flow are encapsulated into the message. The source CE switch forwards the Packet-Out message according to the service forwarding path. All devices that receive the Packet-Out message in the path report a Packet-In message to iMaster NCE-Fabric. iMaster NCE-Fabric then parses the Packet-In message and calculates the detection path based on the actual links.

Multi-path detection

- Multi-path detection traces multiple physical paths between NVE devices to check whether the service flows are interrupted.
- Detection principle: The implementation of multi-path detection is similar to that of single-path detection. The only difference is that a single detection packet is sent during single-path detection while the number of packets sent during multi-path detection is configurable. iMaster NCE-Fabric also can filter out duplicate paths.

Network Path Detection: Single-Path Detection

- Task configuration: Select the types of source and destination end ports or network devices for path detection from the drop-down list box and set task parameters. Filter packets based on 5-tuple information.

- Task result:

- If **Status** is displayed as **Finished**, the detection is successful and the paths are normal.
 - If **Status** is displayed as **Failed**, the detection task cannot be executed due to the failure to find the source node or other causes.
 - If **Status** is displayed as **Timeout**, the detection task is executed, but packet forwarding fails because the path is incomplete or interrupted.

Task Configuration

[SingleDetection](#) [BusinessDetection](#)

VM IP(BM ... ▾) VM IP(BM ... ▾)

* IP version: IPv4 IPv6

* Source IP:

Source protocol: Random

* Destination IP:

Destination: Random

* Protocol type: UDP TCP ICMP SCTP

* Detection:

Advanced

Task List				
Start	Stop	Delete	Ta...	☰
Task	Created At	Status	Operation	☰
<input type="checkbox"/> [REDACTED]	81.1.2...	15-27-...	Finished	

MQC Statistics Port		MQC Statistical Results	
Source --> Destination	Source --> Destination	Task Configurati...	
Hop	IP Address	Ingress-Band...	Egress-Band...
1	[REDACTED] 49.14	25GE1/0/47	25GE1/0/3
2	[REDACTED] 49.10	25GE1/0/4	25GE1/0/1
3	[REDACTED] 49.6	10GE1/0/7	10GE1/0/14
4	[REDACTED] 49.51	10GE0/0/2	
5	[REDACTED] 49.6	10GE1/0/14	10GE1/0/7
6	[REDACTED] 49.10	25GE1/0/1	25GE1/0/5
7	[REDACTED] 49.15	25GE1/0/2	25GE1/0/47

HUAWEI

- Customer requirements:
 - Actual physical paths of service flows can be displayed, or service flow interruptions can be detected.
 - IPv4 and IPv6 overlay network paths can be detected.
 - Single-path detection can be performed across fabric networks.
 - Single-path detection of container networks is supported.
 - IPv4 NSH-based SFC can be detected.
 - VM access paths can be detected.
 - Traffic statistics on a specified interface can be collected.
 - Single-path detection is supported in scenarios where IPv4 or IPv6 VMs are connected to the CE1800V.
 - Single-path detection is available to traffic that traverses a firewall.
 - Prerequisites:
 - iMaster NCE-Fabric is running properly.
 - Management IP addresses of switches have been configured and links between them have been set up.
 - SecoManager has been deployed on iMaster NCE-Fabric to detect the path passing through a firewall.

Network Path Detection: Multi-Path Detection

- Task configuration: Set multi-path detection task parameters.
- Task result:
 - If **Status** is displayed as **Finished**, the detection is successful and the paths are normal.
 - If **Status** is displayed as **Failed**, the detection task cannot be executed due to the failure to find the source node or other causes.
 - If **Status** is displayed as **Timeout**, the detection task is executed, but packet forwarding fails because paths between the source device and destination device are incomplete or interrupted.

The screenshot shows two main windows. On the left is the 'Multi-Path Detection' configuration window, which includes fields for 'Source Device' and 'Destination Device', 'VTEP IP' (set to IPv4), and 'Number of' paths (set to 10). On the right is the 'Task List' window, which displays a table of tasks. One task is listed with details: 'Status' is 'Finished', 'Created At' is '05-28 0...', and 'Operation' is 'Delete'. Below these windows is a table titled 'Detection Result' showing three network paths (Path1, Path2, Path3) with columns for Hop, IP Address, Ingress-Bandw., and Egress-Bandw. The data for Path1 is: Hop 1 (CPU, 25GE1/0/6), Hop 2 (25GE1/0/2, 25GE1/0/6), and Hop 3 (10GE1/0/10).

Hop	IP Address	Ingress-Bandw.	Egress-Bandw.
1	[redacted]	CPU	25GE1/0/6
2	[redacted]	25GE1/0/2	25GE1/0/6
3	[redacted]	10GE1/0/10	

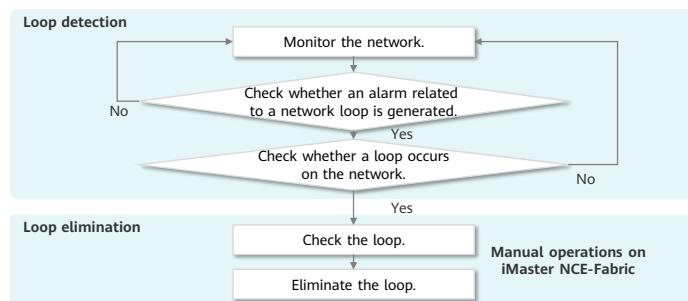
19 Huawei Confidential



- Customer requirements:
 - Actual physical paths of service flows can be displayed, or service flow interruptions can be detected.
 - IPv4 and IPv6 overlay network paths can be detected.
 - Multi-path detection can be performed across fabric networks.
- Prerequisites:
 - iMaster NCE-Fabric is running properly.
 - Management IP addresses of switches have been configured and links between them have been set up.

Network Loop Detection and Elimination (1)

- iMaster NCE-Fabric can automatically detect whether virtual extensible local area network (VXLAN) and virtual local area network (VLAN) loops occur on a fabric network through mechanisms such as traffic collection and event association. It then can locate and eliminate the loops, avoiding any impacts on traffic services caused by improper networking or network attacks.
- The following figure shows the network loop detection and elimination process of iMaster NCE-Fabric.



- When detecting loops, CE switches generate alarms. The alarms can be classified into different types, including the traffic threshold-crossing alarm, VLAN MAC address flapping alarm, and VXLAN MAC address flapping alarm. iMaster NCE-Fabric samples ARP packets based on the alarms reported by interfaces or sub-interfaces of CE switches and displays all suspected loops in a list.
 - When collecting multiple same packets within a specific period of time, iMaster NCE-Fabric determines a loop occurs and displays the loop information on the loop detection page and provides elimination suggestions. Only the local interface where the loop occurs is displayed in the loop detection result.
 - If a device interface or sub-interface sends a large number of normal packets, iMaster NCE-Fabric may fail to collect multiple same packets, and therefore cannot determine whether a loop exists. In this case, you can log in to the device and manually confirm whether a loop exists based on the suspected loop information.
- Customer requirements:
 - On a fabric network, traffic service exceptions may occur due to improper networking or network attacks. Customers require a traffic monitoring technology that samples packets on device interfaces to monitor the traffic status in real time and promptly find abnormal traffic as well as the source of attack traffic.
- Prerequisites:
 - iMaster NCE-Fabric is running properly.
 - The device to be monitored has been added to the fabric network and has available ACL resources.
 - Loop alarm reporting has been enabled on the device.

Network Loop Detection and Elimination (2)

- iMaster NCE-Fabric displays suspected loop information in the current record based on the received loop alarm. Users can delete or manually confirm the current record.

Suspected Loop						Current Record	Historical Record
Fabric Name	Management IP Address	Port	Time	Status	Operation		
Fabric	192.168.1.1	Eth-Trunk101/602	1-05-20 17:03:08	New			
Fabric	192.168.1.1	Eth-Trunk101/601	1-05-20 17:03:08	New			
Fabric	192.168.1.1	Eth-Trunk100/601	1-05-20 17:03:08	New			
Fabric	192.168.1.1	Eth-Trunk101/602	1-05-20 17:03:08	New			

- Information about the confirmed loops is listed on the **Loop Device** tab page. Users can view the details and perform port isolation.

Loop Device				Refresh
Enter fabric name	Latest refresh time:	Topology Preview		
Fabric Name	Detection Time	Number of Devices	Number of Ports	
Fabric_NeW_Hard	15-20 17:07:08	1	2	Port Isolation

• Loop elimination:

- Current records:

- After confirming that a suspected loop on a CE switch does not exist, you can delete the record of the suspected loop.
- If you need to confirm whether a suspected loop on a CE switch exists or not, you can delete the record of the suspected loop, and click the related button in the record of the suspected loop to delete the loop information. Click **Refresh** to view information about the confirmed loop and perform port isolation.

- Historical records: display the status of suspected loops that have been processed so that loop interfaces that have gone offline can be reconnected. On the **Historical Record** page, the status of a suspected loop can be one of the following:

- **Timeout:** After the suspected loop is confirmed and eliminated, the interface on the loop still reports a MAC address flapping alarm within the MAC address aging time of the CE switch. Information of the loop is displayed again in the current record of the suspected loop. The **Status** is **New** and changes to **Timeout** 2 minutes later.
- **Manually deleted:** indicates that users manually delete the record of a suspected loop. After the record is deleted, the status of this record is **Manually deleted** on the **Historical Record** page.

Consistency Check

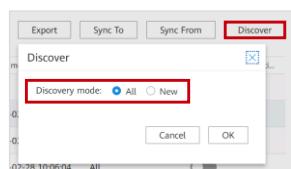
- iMaster NCE-Fabric sends a configuration query request to a forwarder and detects configuration inconsistencies between iMaster NCE-Fabric and the forwarder to facilitate subsequent inconsistency elimination. As such, you can perform either data reconciliation (synchronizing configurations from iMaster NCE-Fabric to the forwarder) or data synchronization (synchronizing configurations from the forwarder to iMaster NCE-Fabric).

Scenario	Inconsistency Discovery Mode	Inconsistency Elimination Mode
Data restoration based on iMaster NCE-Fabric	Inconsistency discovery is performed on services delivered by iMaster NCE-Fabric to the forwarder. Automatic discovery: 1. Full inconsistency discovery 2. Incremental inconsistency discovery	Manually overwrite inconsistent data on the forwarder with data on iMaster NCE-Fabric after inconsistency discovery is manually or automatically performed.
Data synchronization based on the forwarder	Manual discovery: 1. Full inconsistency discovery 2. Incremental inconsistency discovery	<ul style="list-style-type: none"> Automatic recovery: This mode enables the configuration of policies as required and allows the forwarder to automatically synchronize all service data from itself to iMaster NCE-Fabric when the forwarder goes online for the first time. Manual recovery: This mode manually synchronizes inconsistent data from the forwarder to iMaster NCE-Fabric after inconsistency discovery is manually or automatically performed.

- Synchronization policies include data reconciliation based on iMaster NCE-Fabric and data synchronization based on the forwarder.
 - Controller-based reconciliation is to overwrite inconsistent data on the forwarder with data on iMaster NCE-Fabric. If inconsistencies are caused by service data delivered by iMaster NCE-Fabric, synchronize the data from iMaster NCE-Fabric to the forwarder.
 - Forwarder-based synchronization is to synchronize forwarder data to iMaster NCE-Fabric. If inconsistencies are caused by manually configured data, synchronize the data from the forwarder to iMaster NCE-Fabric.
- Inconsistency discovery modes include full inconsistency discovery and incremental inconsistency discovery.
 - During full inconsistency discovery, iMaster NCE-Fabric collects all the data on the forwarder. During incremental inconsistency discovery, iMaster NCE-Fabric collects only the forwarder data that differs from the data collected last time.
 - Incremental inconsistency discovery consumes fewer performance resources than full inconsistency discovery, but requires at least one full inconsistency discovery to be performed in advance.

Data Reconciliation: Data Inconsistency Discovery

- Offline configurations on a forwarder may cause data inconsistencies between the forwarder and iMaster NCE-Fabric. iMaster NCE-Fabric supports manual and automatic data inconsistency discovery. After data inconsistency discovery, you can proactively initiate inconsistency elimination (reconciliation) based on iMaster NCE-Fabric.
- Manually perform data inconsistency discovery.
 - Select All or New.



- View the data inconsistency discovery result.

>	<input type="checkbox"/> cloud-spine2	188.0.4.12	Discovery_Error	Southbound device return process...	02-28 10:06:01	-02-28 1...
>	<input type="checkbox"/> cloud_leaf1	188.0.4.3	Discovered		-04-07 14:41:12	-04-07 1...

24 Huawei Confidential

HUAWEI

- Data inconsistency discovery results are as follows:
 - If **Status** is displayed as **Discovered**, data inconsistency discovery is completed.
 - If **Status** is displayed as **Invalid inconsistent data** and **Failure Reason** is displayed as **Device re-attachment is performed**, the link of the active node is switched and the inconsistent service data is invalid. In this case, perform inconsistency discovery again to resolve the issue.
 - If **Status** is displayed as **Discovery_Error** and **Failure Reason** is displayed as **The device is isolated**, you can cancel isolation of the device in the **Advanced Settings** area of the **Device Management** page.
 - If **Status** is displayed as **Discovery_Error** and **Failure Reason** is displayed as **SFTP users are not configured**, configure SFTP first.
 - If **Status** is displayed as **Discovery_Error** and **Failure Reason** is displayed as **Failed to send netconf package**, perform the following operations on the forwarder:
 - If the forwarder is running properly, run the ssh client first-time enable command on the forwarder to enable first authentication for the SSH client.
 - If the forwarder fails to connect to the southbound service IP address of iMaster NCE-Fabric through the SFTP client, run the sftp client-source -a X.X.X.X command on the forwarder. In this command, X.X.X.X indicates the source IP address used when the forwarder acts

as an SFTP client.

Data Reconciliation: Data Inconsistency Elimination (1)

- Eliminate data inconsistencies for devices.
 - Select the desired device and click **Sync To**. Data that exists on iMaster NCE-Fabric but not on the forwarder is synchronized to the forwarder.
- Eliminate data inconsistencies for instances. (Data that exists on iMaster NCE-Fabric but not on the forwarder is synchronized to the forwarder.)
 - Click the arrow in front of the desired device to check inconsistent features and data types.
 - Click . On the page that is displayed, click **Expand All** to view inconsistent data.
 - In the **Data from the controller** area, select the data to be overwritten and click **Sync To**. Data that exists on iMaster NCE-Fabric but not on the forwarder is delivered to the forwarder.

25 Huawei Confidential

HUAWEI

- Eliminate data inconsistencies for instances.
 - If multiple features are inconsistent, eliminate feature inconsistencies in the following sequence: system, gre, evc, ifm, ethernet, nvo3, dfs, syslog, vxlan, rtp, l3vpn, ifmtrunk, mlag, mstp, trafficanalysis, vrrp, acl, bfd, directrt, bgp, evpn, sflow, staticrt, smartlink, vlan, dhcp, dhcpcv6, nd, arp, qos, ospfv2, ospfv3, mac, sshs, feiarpstatus, sfc, l2mc, dgmp, mcastbase, mvpn, pim, and msdp.

Data Reconciliation: Data Inconsistency Elimination (2)

- Eliminate data inconsistencies for instances by deleting the data that exists on the forwarder but not on iMaster NCE-Fabric.
 - Enable **Delete Via Reconciliation**.
 - Click the arrow in front of the desired device to check inconsistent features and data types.
 - Click . On the page that is displayed, click **Expand All** to view inconsistent data.
 - In the **Data from the forwarder** area, select the data to be overwritten and click **Sync To**. Data that exists on the forwarder but not on iMaster NCE-Fabric is deleted.

The screenshot shows the 'Data Reconciliation' interface. At the top, there are buttons for 'Expand All', 'Sync To' (which is highlighted with a red box), and 'Sync From'. A legend indicates 'Discovered' (green dot). Below these are two main sections:

- Data from the controller (Device name = tor-14.54, Device IP = [REDACTED])**: Shows a list of configuration items under 'huawei-ac-ne-ce-sdnagent:sdnagent'.
- Data from the forwarder (Device name = tor-14.54, Device IP = [REDACTED])**: Shows a detailed tree structure under 'huawei-ac-ne-ce-sdnagent:sdnagent'. The 'sdnControllerInfos' node has a child 'sdnControllerInfo' which contains a 'controllerAddress' entry. The 'sdnAgentOpenflowInfo' node contains 'enable', 'connectAddress', and 'echoInterval' entries.

26 Huawei Confidential



- Eliminate data inconsistencies for instances.
 - If **Delete Via Reconciliation** is enabled, click **Sync To** to delete configuration that exists on the forwarder but not on the controller.
 - If multiple features are inconsistent, eliminate feature inconsistencies in the following sequence: msdp, pim, mvpn, mcastbase, dgmp, l2mc, sfc, feiarpstatus, sshs, mac, ospfv3, ospfv2, qos, arp, nd, dhcpcv6, dhcp, vlan, smartlink, staticrt, sflow, evpn, bgp, directrt, bfd, acl, vrrp, trafficanalysis, mstp, mlag, ifmtrunk, l3vpn, rtp, vxlan, syslog, dfs, nvo3, ethernet, ifm, evc, gre, and system.

Data Reconciliation: Data Inconsistency Elimination (3)

- Initiate data restoration.
 - If **Status** is displayed as **Reconciled**, data restoration is completed.

>	<input type="checkbox"/>	cloud_leaf2	188.0.4.14	Reconciled
>	<input type="checkbox"/>	cloud_leaf3	188.0.4.13	Reconciled

- If **Status** is displayed as **Invalid inconsistent data** and **Failure Reason** is displayed as **Device re-attachment is performed**, the active node link of the device has been switched and the inconsistent service data is invalid. In this case, perform inconsistency discovery and reconciliation again to resolve the issue.

- If the above issue cannot be resolved, contact Huawei technical support.

Three-Level Rollback

Network-wide rollback

- Network-wide rollback is used to resolve major faults on the entire network. For example, if network configurations are deleted due to changes, many services are interrupted. In this case, network-wide configurations can be rolled back to those before the changes or interruptions, enabling quick service recovery.
- Before changes, you can back up network-wide configurations on iMaster NCE-Fabric. When a problem occurs due to changes, the configurations can be quickly restored to the backup point, resolving major network faults.
- You can manually save data in real time or periodically on the GUI. You need to proactively back up data.

Tenant snapshot

- The tenant snapshot function is used to back up and restore network service configurations by tenant, and apply to multi-tenant services. Backup and restoration operations performed by a tenant do not affect the provisioning of other tenants' services, including backup and restoration of network service configurations by other tenants.
- The tenant snapshot function allows a tenant to set a backup point and save all its service configurations at the backup point. If needed, service configurations can then be restored to a specific snapshot point. Additionally, iMaster NCE-Fabric can compare the current configurations with the configurations at the snapshot point, or compare the configurations from two given snapshot points, and perform configuration rollback to eliminate differences.
- The tenant snapshot function supports manual backup and restoration as well as automatic and periodic backup.

Service-level rollback

- Service-level rollback helps quickly restore original network configurations to recover services when a network exception occurs due to a fine-grained single-point service provisioning failure.
- You do not need to manually back up data for service-level rollback, but need to manually restore data.
- iMaster NCE-Fabric automatically backs up each service that is provisioned. When an exception occurs, iMaster NCE-Fabric can quickly restore the service to the status before the service is provisioned.

- iMaster NCE-Fabric provides three-level rollback, meeting the reliability requirements of different scenarios and ensuring quick service recovery. This feature covers 70% to 80% of routine change scenarios. For example, the fast rollback feature is available for single-point service provisioning exceptions and independent tenant services.

- Network-wide rollback features:

- iMaster NCE-Fabric saves the snapshots of the entire network, including those of iMaster NCE-Fabric and its managed devices.
- You can manually save the snapshots in real time or periodically.
- During restoration, iMaster NCE-Fabric delivers commands to devices to restore data. The devices restore specific configurations based on specified snapshot point labels and do not need to be restarted.

- Tenant snapshot features:

- You can manually save the snapshots in real time or periodically.
- iMaster NCE-Fabric divides different tenant spaces for tenant backup so that operations between tenants do not affect each other.
- Differences between rollback points can be previewed for further examinations.

- Service-level rollback features:

- Service operations are automatically saved.
- Snapshots are automatically stored in mirroring mode.
- The linkage technology enables rollback of multiple operations to the previous state.

Network-Wide Rollback (1)

- When used by a single user, if a network configuration error occurs, iMaster NCE-Fabric can quickly restore the network configuration to a certain time point during which the configuration has been backed up. In this way, network services can be quickly restored in full mode, avoiding great loss caused by time-consuming network configuration restoration. iMaster NCE-Fabric can back up and restore its network service configurations and configurations of all managed CE devices.
- Network-wide data backup
 - Network-wide data backup includes the backup of iMaster NCE-Fabric database and configuration backup of CE devices.

The screenshot shows the 'Database Backup Details' section of the interface. A table lists a single task:

Task Name	Task Progress	Task Status	Start Time	End Time	Backup Reason	Operation
0531022353629 Back Up	100%	Backup Succeeded	05-31 02:23:53	05-31 02:29:26	III	Delete Edit

Below the table, a note says: "When Task Progress is 100%, the backup succeeds."

Further down, under 'Database Backup Details', it shows success statistics: Success: 1 Failure: 0. It also lists a single device entry:

Device Name	Management IP Address	Device Type	Device Model	Device Version	Location	Backup Time	Status	Failure Reason
CE6880_85.3	██████████	SWITCH	CE6863-48S6CQ	V200R020C00	Beijing China	05-31 02:23...	Backup Succeeded	-

- If the task fails, click > to view details about the backup task and failure cause. Then locate and rectify the fault based on the failure cause to back up the data again.

Network-Wide Rollback (2)

- Network-wide data restoration

- Network-wide data can be restored based on the existing backup points.
- When **Task Progress** is displayed as **100%**, the restoration succeeds.
- If **Status** is displayed as **Restore Failed**, you can view details about the restoration task and failure cause.
 - Database restoration failure: Locate and rectify the fault and perform the restoration task again.
 - Device restoration failure: Select a device in the **Device Name** column and click **Retry** to restore the device failing to be restored again.

The screenshot shows a user interface for network-wide data restoration. At the top, there is a dropdown menu labeled "Select backup point" with the value "0531022353629". Below this, a main panel displays "Device Restoration Details" and "Device Backup Details".

Device Restoration Details:

Task Name	Task Progress	Task Status	Start Time	End Time	Failure Reason
I530202121797	100%	Restore Failed	05-30 20:27:16	05-30 22:12:49	-

Device Backup Details:

Device Name	Management IP Ad...	Device Type	Device Model	Device Version	Location	Backup Time	Status	Failure Reason
CE6880_85.2	[REDACTED]	SWITCH	CE6863-48S6CQ	V200R020C00	local	05-30 20:27:16	Restore Failed	The checkpoint...
CE6880_85.3	[REDACTED]	SWITCH	CE6863-48S6CQ	V200R020C00	Beijing China	05-30 20:27:16	Restore Failed	The checkpoint...

Tenant Snapshot Management (1)

- If there are a large number of tenants on iMaster NCE-Fabric, tenant snapshot management can be implemented to restore the network service configurations of a single tenant to a certain time point. Tenant snapshot management supports the backup and restoration of tenant network configurations without affecting the delivery of network service configurations of other tenants on iMaster NCE-Fabric. For a single tenant, tenant snapshot management offers the following capabilities:
 - Manual or automatic creation of tenant snapshots.
 - Saving of snapshots to a remote server and importing of snapshots from a remote server.
 - Display of differences between current tenant configurations and snapshots
 - Rollback based on snapshot files.
 - Display of the rollback task status and historical records.
- Automatic backup or snapshot creation:



Recent task: Succeeded in creating the snapshot.

Snapshot file: 0ae4283d-e009-4503-b0b3-e3dda78b837020200609144629.zip Operation type: Manually Created Operation type: -06-09 14:46:29 Status: Success

Tenant Snapshot Management (2)

- Restoring a snapshot means to roll back the current tenant configuration to the backup configuration.

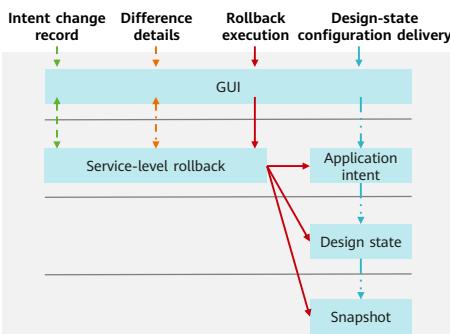
Recent task: The snapshot is being restored.

Restoration point snapshot: S29100922_Manual Status: Creating Progress: Automatic snapshot...Automatic compa... Automatic rollback [View Rollback Process](#)

Resource Snapshot	Resource Name	Execution Status	Operation	Resource Details	Error Message
I291112914_Manual	Comparison snapshot 1	S29100922_Manual	Creating	Operation time: 5-29 11:28:15	
Implementation details:					
Resource Type	Resource Name	Execution Status	Operation	Resource Details	Error Message
logicSwitch	sw5	Success	Delete	({"logicSwitch": {"appId": "2a180266-562d-45eb-b372-0f04fb51f5c", "id": "25e36418-0649-4e3f-8b36-b6392c286898"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}), {"logicSwitch": {"appId": "2a180266-562d-45eb-b372-0f04fb51f5c", "id": "25e36418-0649-4e3f-8b36-b6392c286898"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"logicSwitch": {"appId": "2a180266-562d-45eb-b372-0f04fb51f5c", "id": "25e36418-0649-4e3f-8b36-b6392c286898"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"logicSwitch": {"appId": "2a180266-562d-45eb-b372-0f04fb51f5c", "id": "25e36418-0649-4e3f-8b36-b6392c286898"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"logicSwitch": {"appId": "2a180266-562d-45eb-b372-0f04fb51f5c", "id": "25e36418-0649-4e3f-8b36-b6392c286898"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"logicSwitch": {"appId": "2a180266-562d-45eb-b372-0f04fb51f5c", "id": "25e36418-0649-4e3f-8b36-b6392c286898"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"routerSubnet": {"Router Subnet": "Router Subnet"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"logicPort": {"Ip01": "Ip01"}, "op": "Delete", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"routerSubnet": {"Router Subnet": "Router Subnet"}, "op": "Create", "status": "Success", "time": "2023-05-29T11:28:15Z"}}, {"switchSubnet": {"Switch Subnet": "Switch Subnet"}, "op": "Create", "status": "Waiting for ACK", "time": "2023-05-29T11:28:15Z"}}	

Service-Level Rollback (1)

- When the VPC design state data is submitted, iMaster NCE-Fabric automatically generates a configuration change record. Based on the selected configuration change history, the service-level rollback function can quickly roll back the service configurations that have been successfully delivered to network devices to the configurations before the data in the design state is submitted. This function is useful for scenarios where an exception occurs after configurations are delivered to network devices and emergency rollback is required. The following figure shows the implementation of service-level rollback.

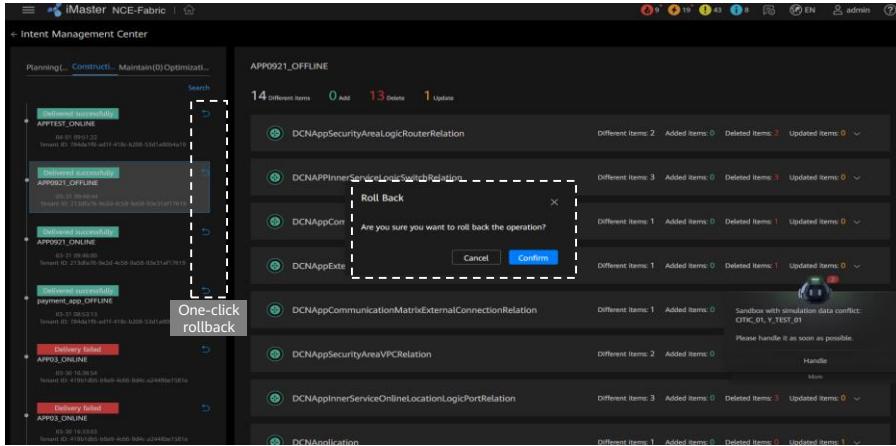


1. Design-state configuration delivery: When design-state configurations are submitted and delivered to network devices, the service-level rollback function records the configuration change data for subsequent rollback.
2. Intent change record: The service-level rollback function provides the change records and displays them on the UI of the intent management center.
3. Difference details: The service-level rollback function provides configuration differences at the logical layer and application layer and displays the data on the UI of the intent management center.
4. Rollback execution: The service-level rollback function performs atomic rollback of configurations at the logical layer and application layer. Different configurations at the application layer can be rolled back only after different configurations at the logical layer are successfully rolled back. If different configurations at the logical layer fail to be rolled back, service-level rollback fails.

- Currently, the following service-level rollback functions are provided:
 - Check the intent change records.
 - Check difference details in the intent change records.
 - Select the intent change history and perform rollback.
 - Check the rollback details after a successful rollback.
- Design state:
 - Indicates the process of VPC service orchestration, simulation, and verification in the Service Simulation app. Services orchestrated in design state will be delivered to the design-state database but not to real devices. The data will be submitted to the production-state database and delivered to the devices after you click **Submit**.

Service-Level Rollback (2)

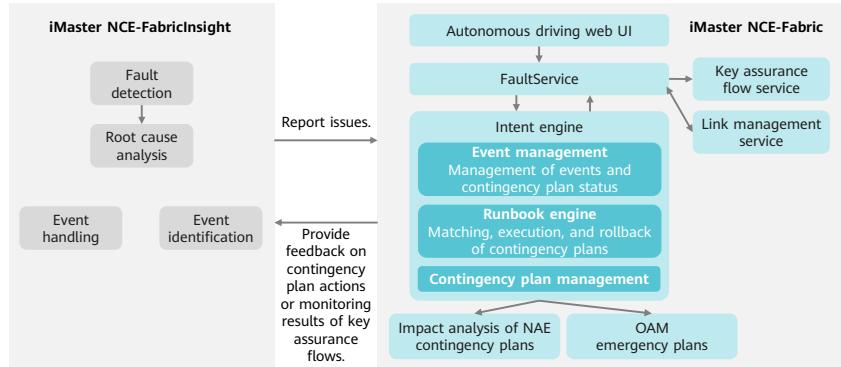
- You can search for the corresponding change records, click  in the **Search** column, and click **Confirm** to perform the rollback task.



Troubleshooting

- iMaster NCE-Fabric provides the intent-driven intelligent event function, which supports device monitoring, application and service fault monitoring, as well as display of fault details, rectification suggestions or plans. You can perform closed-loop management of faults based on user intents. The intelligent event function helps users quickly locate and rectify faults, shortening the time for fault locating and troubleshooting as well as enhancing service continuity.

- Functional architecture:



35 Huawei Confidential

HUAWEI

- Currently, the intelligent event function supports the following types of fault events: device fault, application fault, and service fault. iMaster NCE-Fabric needs to collaborate with iMaster NCE-FabricInsight to solve application and device faults, while iMaster NCE-Fabric can solve service faults independently.
 - Device fault: When a CE1800V, physical CE switch, or Huawei firewall is faulty, iMaster NCE-FabricInsight reports a fault event to iMaster NCE-Fabric, for example, a fan module of the switch is faulty.
 - Application fault: When a key assurance flow monitoring exception is detected, iMaster NCE-FabricInsight reports a fault event to iMaster NCE-Fabric.
 - Service fault: A new host access link is set up due to incorrect interface connection or server migration. In this case, the status of the existing host access link becomes unknown. When detecting the unknown host access link, iMaster NCE-Fabric generates a fault event.
- Note:
 - Web UI: Web user interface
 - FaultService: fault service
 - Runbook engine: text workflow engine
 - NAE: network automation engine
 - OAM: operation, administration, and maintenance

Handling Process of Various Fault Events

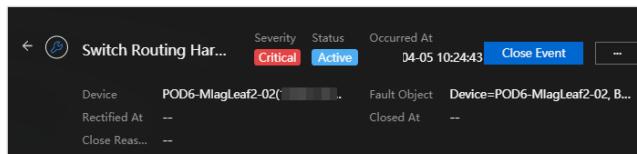
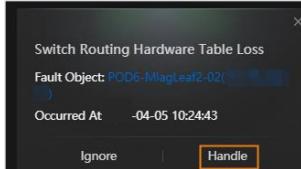
Device fault	Application fault	Service fault
<ol style="list-style-type: none">iMaster NCE-FabricInsight collects syslogs, device configurations, and device flow information to automatically detect faults, and analyze the faults and their impacts.iMaster NCE-FabricInsight sends fault details, root causes, and fault impacts to iMaster NCE-Fabric.iMaster NCE-Fabric analyzes fault information and provides suggestions as well as a fault rectification plan and its impacts.iMaster NCE-Fabric delivers a rectification plan. After detecting that the fault is rectified, iMaster NCE-FabricInsight updates the event status.	<ol style="list-style-type: none">iMaster NCE-Fabric sends the created key assurance flow task information to iMaster NCE-FabricInsight.iMaster NCE-FabricInsight monitors the traffic status of a specified task. When a flow exception is detected, it sends the exception information to iMaster NCE-Fabric.	<ol style="list-style-type: none">When detecting an unknown host access link, iMaster NCE-Fabric sends the fault information to the fault remediation module using closed-loop troubleshooting.The fault remediation module analyzes the network configuration associated with an unknown link.Fix the logical switch configuration associated with the unknown link. After the rectification is successful, the network configuration associated with the existing link will be migrated to a new port.After the unknown link is cleared, the status of the fault event is updated to Solved.
<ul style="list-style-type: none">Fault rectification method:<ul style="list-style-type: none">Notification: Notify the user of the fault.Suggestion: Provide rectification suggestions.Rectification plan delivery: Provide a rectification plan, which can be delivered and rolled back in one-click mode.	<ul style="list-style-type: none">Fault rectification method:<ul style="list-style-type: none">Display monitoring details of key assurance flow exceptions.	<ul style="list-style-type: none">Fault rectification method:<ul style="list-style-type: none">Fix the network configuration associated with an unknown link.Clear the unknown link.

Case: Switch Routing Hardware Table Loss (Rectification Plan Delivery) (1)

- **Case description:** When the routing hardware table of a switch is lost, iMaster NCE-Fabric detects and rectifies the fault.
- Click the number on the upper right of the intelligent twins on the iMaster NCE-Fabric GUI.



- In the dialog box that is displayed, click **Handle** on the **Switch Routing Hardware Table Loss** tab page to view the fault details.



- Prerequisites: iMaster NCE-Fabric has been connected to iMaster NCE-FabricInsight.

Case: Switch Routing Hardware Table Loss (Rectification Plan Delivery) (2)

- **Case description:** When the routing hardware table of a switch is lost, iMaster NCE-Fabric detects and rectifies the fault.
- Check the solution and impact analysis of the solution on the iMaster NCE-Fabric GUI shown below.

Hello, the "Switch Routing Hardware Table Loss" has been detected. The following solutions are recommended. Select one for the current event.

Recommend

Solution1: Smooth-Fault... Solution2: Reset-Slot Solution3: Reboot-Device

Deliver the inconsistent table entries between the software and hardware tables to the hardware table based on the software table entries.

Fault object: vpn instance=aaa, ip= [REDACTED]

Impact

No analysis data available.

Configuration to be Issued

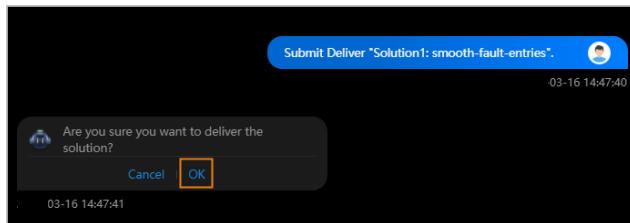
POD6-MlagLeaf2-02

```
<fibstatus xmlns="http://www.huawei.com/netconf/vrp" content-version="4.0" format-version="1.0"><ipv4FibDistribute><ipAddr></ipAddr><vpnName></vpnName><slotId></slotId></ipv4FibDistribute></fibstatus>
```

Analyze Again Deliver Roll Back

Case: Switch Routing Hardware Table Loss (Rectification Plan Delivery) (3)

- **Case description:** When the routing hardware table of a switch is lost, iMaster NCE-Fabric detects and rectifies the fault.
- Click **OK** in the **Are you sure you want to deliver the solution?** dialog box.



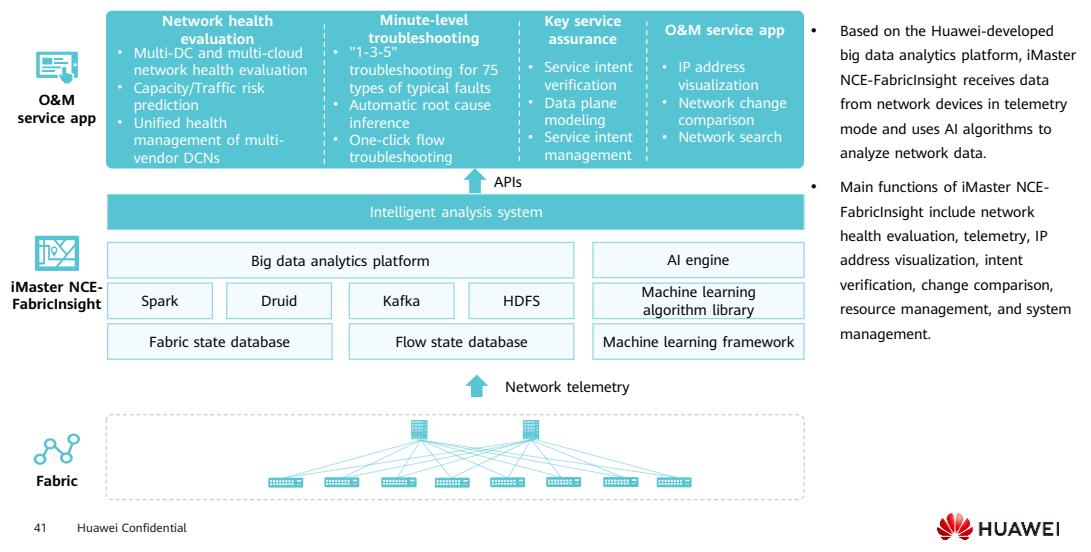
- After the fault is rectified, the event status changes to **Solved**. Click **Close Event** to close the event.

- If you need to roll back the rectification plan after it is successfully delivered, click **Roll Back** on the **Solution 1** tab page.

Contents

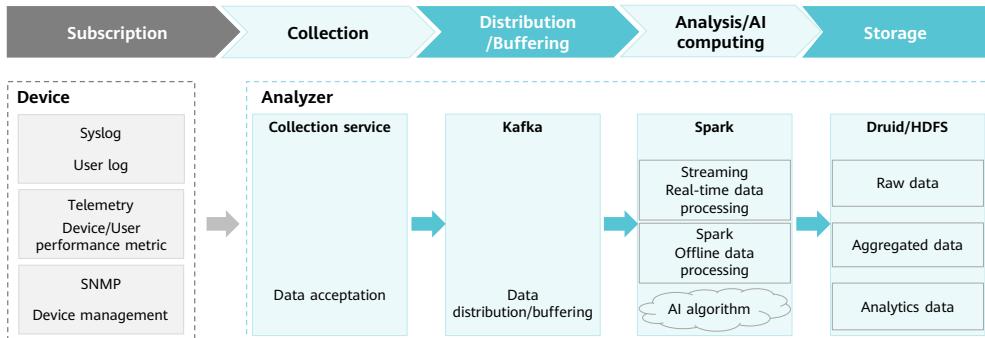
1. DCN O&M Challenges and CloudFabric Intelligent DCN O&M Solution
2. iMaster NCE-Fabric
- 3. iMaster NCE-FabricInsight**
 - Overview
 - Network Visualization and Health Evaluation
 - Fault Locating
 - Change Assurance

iMaster NCE-FabricInsight



- The overall architecture of iMaster NCE-FabricInsight consists of three parts: network devices, iMaster NCE-FabricInsight collector, and iMaster NCE-FabricInsight analyzer.
 - Network devices:
 - Huawei CE switches (For details about the models and supported specifications, see the specification list of the corresponding version.)
 - Devices report performance metrics such as interface traffic in telemetry mode based on Google Remote Procedure Call (gRPC). Devices are connected to iMaster NCE-FabricInsight as gRPC clients. Users can run commands to configure the telemetry function on the devices. The devices then proactively establish gRPC connections with the desired collector and send data to the collector. The current version supports the following sampling metrics: CPU and memory usage at the device and card levels; number of sent and received bytes, number of discarded sent and received packets, and number of sent and received error packets at the interface level; number of congested bytes at the queue level; packet loss behavior data. For details about metrics and device models, see the product specification list.

Data Processing



After data subscription, the collection service module collects data in seconds. The high-throughput distributed message system is used to buffer and distribute the collected data. Service modules perform data analysis and calculation based on the AI algorithm and expert experience, and save the processed data to the fast and column-oriented distributed data storage system. You can access the page to view the data and functions.

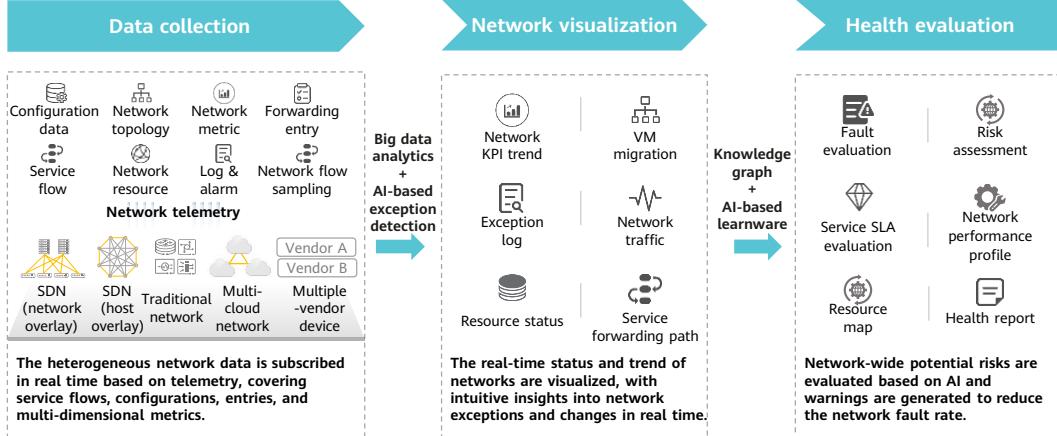
- Note:

- Kafka: the messaging middleware for storing and distributing data reported by devices.
- Spark: a universal parallel framework
- Streaming: flow processing
- Druid: a database for real-time analysis, used as a high-concurrency backend API requiring fast aggregation.
- HDFS: Hadoop Distributed File System

Contents

1. DCN O&M Challenges and CloudFabric Intelligent DCN O&M Solution
2. iMaster NCE-Fabric
- 3. iMaster NCE-FabricInsight**
 - Overview
 - Network Visualization and Health Evaluation
 - Fault Locating
 - Change Assurance

Visualized Analysis on Multi-Dimensional Data and Systematic Evaluation of Network-wide Health Risks



45 Huawei Confidential

HUAWEI

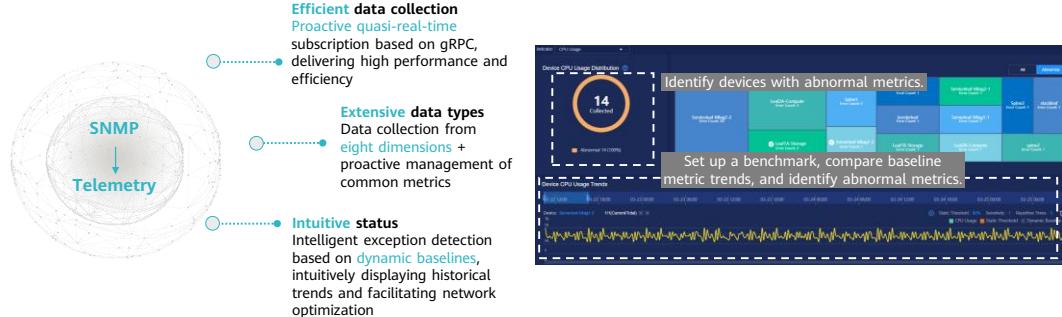
- iMaster NCE-FabricInsight provides network O&M services featuring visualization, automation, and intelligence:
 - Visualization: visible and clear
 - The concept of "visible" consists of two aspects: observed objects and real-time observation. Observed objects include physical objects such as devices, interfaces, and links. Real-time observation supports perception of millisecond-level symptoms, for example, identifying microburst traffic congestion on the network.
 - The concept of "clear" refers to the observation accuracy. On the one hand, a myriad of data needs to be collected. On the other hand, the data must be analyzed in real time.
 - Automation: proactive analysis
 - To proactively and intelligently detect issues on the network in a timely manner, the O&M system must be able to analyze massive data and identify abnormal events on the network. In addition, the O&M system needs to determine whether to generate issue models and recommend them to users based on machine learning algorithms.

Compliance of Real-time Analysis Requirements Based on Telemetry Technology

SNMP	Telemetry
 Simple statistics collection with manual decision-making	 Intelligent data analysis with automatic troubleshooting
 Unstructured data with low encoding and decoding efficiency	 GPB binary encoding and decoding with high transmission efficiency
<Pull> Request-response mode with a large sampling interval	<Push+gRPC> Continuous data push with only one-time data subscription
5/15 min Minute-level polling cycle, failing to meet the service requirements of real-time management	Near Realtime Quasi-real-time data acquisition
 Data analysis	 Transport format
 Data collection	 Data generation

The quasi-real-time data acquisition capability is the key to data analysis of the intelligent network O&M.

Telemetry-Powered Proactive Monitoring and Real-Time Network Visualization (1)



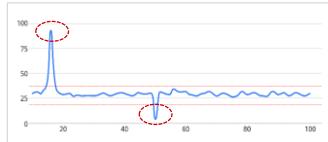
Telemetry-Powered Proactive Monitoring and Real-Time Network Visualization (2)

- Real-time monitoring of key metrics from eight dimensions, gaining deep insights into network status.

Measurement Objects	Measurement Metrics	Default Interval
Device	CPU usage/Memory usage	1 min
Board	CPU usage/Memory usage, and FIB/MAC address entry usage	1 min
Chip	Rules usage, Meters usage, Counters usage, and Slice / Banks usage	1 min
Interface	Number of received/sent packets, number of bytes, number of lost packets, number of error packets, number of broadcast packets, number of multicast packets, number of unicast packets, bandwidth usage, and number of ECN packets	1 min
	ARP & ND attack source tracing	
Queue	Buffer size	100 ms
Optical link	Transmitted/received optical power, current, voltage, and temperature	1 min
Packet loss behavior	Forwarding packet discarding and congestion-triggered packet loss	1 min
Entry	Details of FIB/ARP/ND entries	Dynamic subscription

Dynamic Baseline and Exception Detection (1)

Dataset & preprocessing



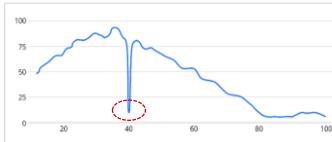
Value stability metric scenario
If values at sampling points are out of the valid range, they are called outliers.

Input: Time series data of metrics (value, time)

- Functions:**
- Automatic identification of collection frequencies
 - Automatic filling of missing data
 - Noise reduction data:** noise reduction of abnormal data
 - Special adaptation:** extra data processing during holidays

- Output:**
- Data features (value stability or period stability)
 - Metric collection interval

Dynamic baseline construction

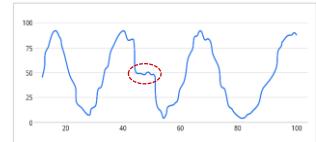


Differentiated stability metric scenario
If salient differences exist in time points before and after the sampling, it is called link comparison exception.

- Functions:**
- Period stability metric algorithm: time series decomposition
 - Value stability metric algorithm: Gaussian regression
 - Baseline boundary construction based on algorithms
 - Baseline sensitivity adjustment

- Output:**
- Prediction for the top and bottom baseline boundaries of the next collection interval

Exception detection



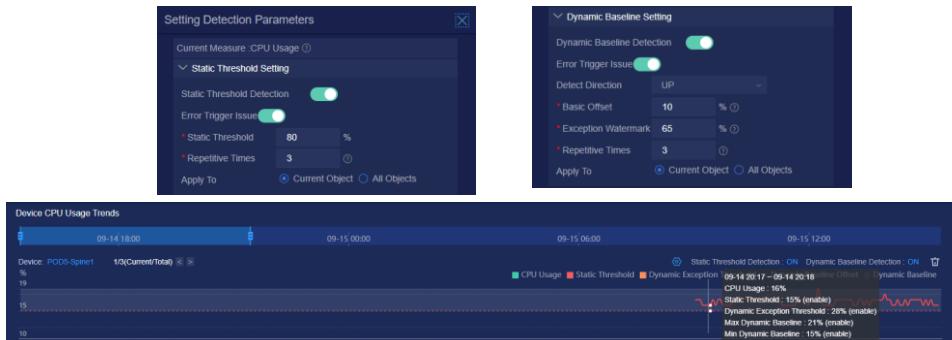
Period stability metric scenario
If salient differences exist between the sampling interval series and the overall trend, it is called parallel comparison exception.

- Functions:**
- Number of exceptions
 - Suppression and combination of problems
 - Problem notification

- Output:**
- Exception

Dynamic Baseline and Exception Detection (2)

- When a baseline exception occurs on a device, you can view associated flow information.
 - For static threshold detection, you can adjust the static threshold and number of repetitions.
 - For dynamic baseline exception detection, you can adjust the exception dynamic baseline, baseline offset, number of repetitions, and detection direction.



51 Huawei Confidential

HUAWEI

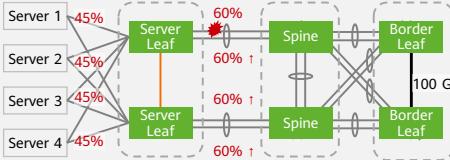
- Dynamic baseline exception detection:
 - Exception Watermark:** applies to ratio-type metrics (such as CPU usage and memory usage). Dynamic baseline exception detection is performed only when the metric value exceeds the exception dynamic baseline.
 - Baseline Offset:** applies to the dynamic baseline and is used to adjust the dynamic baseline detection range.
 - Repetitive Times:** indicates the number of times that the dynamic baseline or static threshold is exceeded consecutively.
 - Detection Direction:** indicates the direction in which dynamic detection is performed, including the scenarios where the metric value is detected only against the upper threshold, the metric value is detected only against the lower threshold, and the metric value is detected both against the upper threshold and lower threshold. The trend chart is displayed in green when the threshold is not exceeded and in red when the static threshold or the dynamic baseline is exceeded.
 - Issue Generation upon Threshold Exceeding:** If this toggle is switched on and the static threshold or dynamic baseline is exceeded, a pending issue is generated on the **Health** page.
- When a baseline exception occurs on an interface, you can view associated flow information.
 - You can adjust the static threshold and number of repetitions. You can also adjust the sensitivity of dynamic baseline detection based on detected

exceptions.

Interface Traffic Prediction with Precautions of Health Risks

Scenario: How to determine whether link traffic exceeds the threshold to provide a basis for capacity expansion

- How to evaluate the traffic during peak hours this year and make plans in advance?
- If the DCI link bandwidth usage is increased from 20% to 65% within two weeks, when is capacity expansion needed?

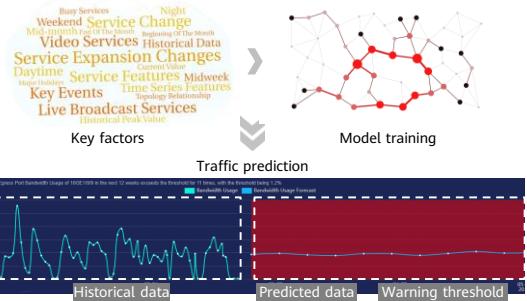


Challenges to traditional O&M:

- The rules of service traffic growth cannot be manually identified, making it hard to determine the proper time of capacity expansion to minimize the cost.
- Bottlenecks exist in capacity. Faults are passively alarmed and cannot be predicted.

Solution: The interface traffic trend in the next three months is predicted, with an algorithm accuracy of 90%.

- iMaster NCE-FabricInsight analyzes over 20 key factors from three dimensions, namely, historical time, space topology, and service attributes.
- iMaster NCE-FabricInsight predicts whether the interface traffic exceeds the threshold in the next three months through the learning and inference of deep neural network algorithm.



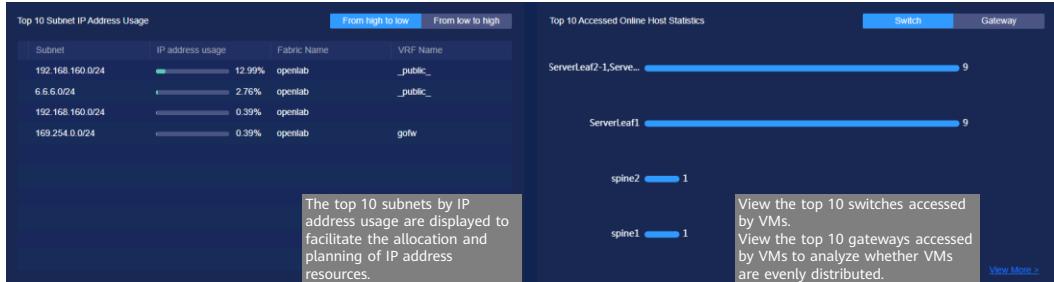
Interface Traffic Prediction

- You can create prediction tasks to predict the inbound and outbound bandwidth usage of interfaces. You can also view the prediction trend, threshold-crossing statistics (based on the static threshold configured on the Telemetry page), and deviation rate statistics.
- Every Sunday, AI predicts the inbound and outbound bandwidth usage trends of interfaces in the next 12 weeks based on the historical data of the last 66 days. If the data of the last 66 days is incomplete, traffic prediction reliability decreases, or even no prediction result is generated (the number of days in which historical data is stored is less than the threshold).

Task Name	Description	Scenario	Number of Thres...	Resource Type	Update Time	Task Status	Running Duration	Operation
> test	--	Prediction	--	Interface	2023-06-21 16:33:22	Running	46Seconds	
> Default Device Task	--	Detection	--	Device	2023-06-21 11:51:23	Running	4Hours42Minutes45Seconds	
Used model: dcn_astrain_baseline								
Created on: 2023-06-21 11:20:38								
Period: 1minute								
Included indicator: cpu usage; memory usage								
Indicator deduction number: 0								
> Default Board Task	--	Detection	--	Board	2023-06-21 11:51:22	Running	4Hours42Minutes46Seconds	
> Default Interface Task	--	Detection	--	Interface	2023-06-21 11:50:52	Running	4Hours43Minutes16Seconds	
> Default Chip Task	--	Detection	--	Chip	2023-06-21 11:50:51	Running	4Hours43Minutes17Seconds	

IP Address Visualization: Overview

- The **Overview** tab page collects the IP address statistics and ranking of the IP address statistics on a network from multiple dimensions, and clearly displays the statistics.
- The **Overview** tab page displays statistics on devices and hosts in the current system, including the host access mode, top 10 switches to which hosts are connected, top 10 gateways to which hosts are connected, top 10 fabrics by IP address usage, top 10 subnet usage, online IP address statistics and change trend, as well as invalid IP address statistics.
- Top 10 subnets by IP address usage
 - Top 10 devices connected to online hosts



54 Huawei Confidential

HUAWEI

- Use the analyzer of V100R021 as an example. Choose **Toolbox > IP 360** to access the IP address visualization page.

IP Address Visualization: IPv4 Distribution

- The **IPv4 Distribution** tab page displays IP address usage in the network-wide or subnet view. The IP address status can be: **Online**, **Transient**, **Offline**, **Exclude**, **Unknown**, **Selected**, and **Invalid IP Address ID**. The page also displays top 10 and bottom 10 subnets ranked by IP address usage in the current view.



55 Huawei Confidential

HUAWEI

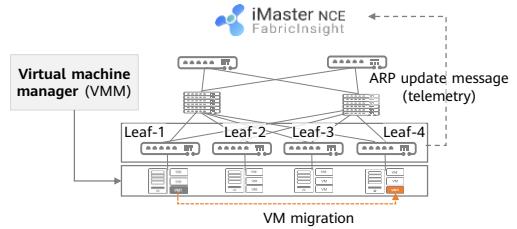
IP Address Visualization: IP Address

- The **IP Address** tab page displays IP address information about devices and VMs, including the IP address, name, MAC address, fabric, VLAN, access device, access IP address, gateway interface, first discovery time, latest discovery time, active status, and discovery mode. IP addresses that frequently migrate are displayed on the top of the list.
- In addition, you can filter items by the IP address, MAC address, fabric, access device, access interface, virtual routing and forwarding (VRF), VLAN ID, active status, and access type of a VM.

Filter	Usage Equals Host IP	X	Search								
	IP Address	MAC	Name	IP Subnet	Usage	Fabric	Status	VLAN	First Discovery Time	Last Discovery Time	Remark
	192.202.16.36	68:CC:6E:FA:45:55	~		Host IP	openlab	● Online	~	03-29 15:28:50	-04-07 15:43:50	
Position1	Connected Device: ServerLeaf1				Connected Device Port: MEIn0/G0	Gateway Port: MEIn0/G0					VRF: _public_
Position2	Connected Device: ServerLeaf2/2				Connected Device Port: MEIn0/G0	Gateway Port: ~					VRF: _public_
View Historical IP Address Access											
>	192.202.16.160	00:08:61:F1:29:12	~		Host IP	openlab	● Online	~	03-29 15:28:50	-04-07 15:43:50	
>	192.202.16.80	F4:E5:F2:20:EB:91	~		Host IP	openlab	● Online	~	03-29 15:28:50	-04-07 15:43:50	
>	165.1.1.184	00:00:00:11:01:84	~		Host IP	openlab	● Online	3000	03-29 15:28:50	-04-07 15:43:50	
>	192.202.16.200	E4:68:A3:F6:D0:7A	~		Host IP	openlab	● Online	~	03-29 16:13:50	-04-07 15:43:50	
>	130.1.1.10	00:50:56:81:C7:CD	~		Host IP	openlab	● Online	2001	03-29 15:28:50	-04-07 15:43:50	
>	165.1.1.100	00:00:00:50:01:00	~		Host IP	openlab	● Online	3000	03-29 15:28:50	-04-07 15:43:50	
>	165.1.1.183	00:00:00:11:01:83	~		Host IP	openlab	● Online	3000	03-29 15:28:50	-04-07 15:43:50	
>	165.1.1.10	00:50:56:81:26:8A	~		Host IP	openlab	● Online	2000	03-29 15:28:50	-04-07 15:43:50	
>	30.1.1.100	00:50:56:8B:EC:94	~		Host IP	openlab	● Offline	2004	03-30 01:13:50	-04-06:19:58:50	

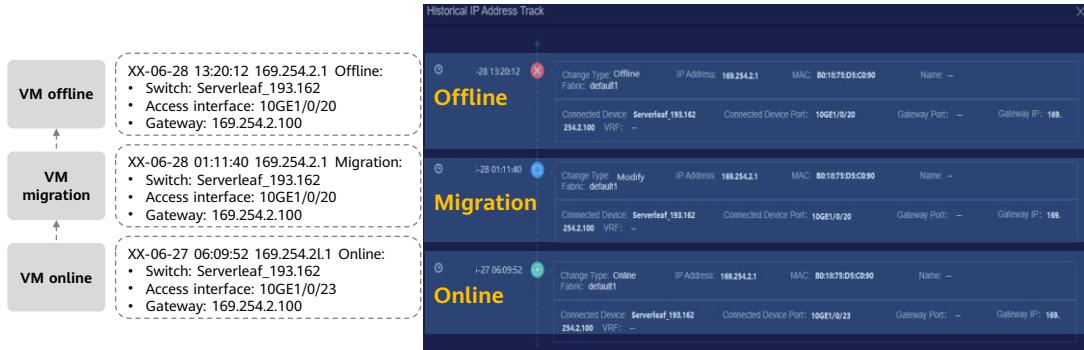
IP Address Visualization: Full-Lifecycle VM Management (1)

- Scenario:
 - During dynamic VM migration, the network team cannot determine whether the location of the switch to which the VM is connected has changed. As a result, the VM cannot be accessed to the network before dynamic VM migration and it is difficult to locate the fault.
- Solution:
 - iMaster NCE-FabricInsight uses telemetry to collect ARP update information (including the added, deleted, and modified information) of network-wide devices, and supports full-lifecycle visualization of VM login, logout, and migration records based on fabric information.



IP Address Visualization: Full-Lifecycle VM Management (2)

- On the **IP Address** tab page, click **View Historical IP Address Access** to view the historical information about IP addresses. As such, you can implement full-lifecycle VM management and view details about VM login, logout, and migration records.



Full Log Analysis Principles

- The log module of the system software logs events occurring during system operation. Logs provide reference information for system diagnosis and maintenance, and help you check the device running status, analyze network conditions, and locate faults. There are eight levels based on the severity, each identified by a number. A smaller number indicates higher log severity levels. The detailed definition of log levels is listed in the following table.

Level	Severity	Description
0	Emergencies	A fault that makes the device unable to run normally unless it is restarted. For example, the device restarts due to a program exception or an error in memory usage.
1	Alert	Major device fault, which requires an immediate solution. For example, the device memory usage reaches the upper limit.
2	Critical	A fault that needs to be analyzed and processed. For example, the memory usage of the device falls below the lower limit, or BFD detects that a device is unreachable.
3	Error	An incorrect operation or service processing exception that does not affect services but needs to be analyzed. For example, users enter incorrect commands or passwords, or error protocol packets are received.
4	Warning	An exception that occurs when a device is operating and requires attention because it may cause service processing faults. For example, a routing process is disabled, BFD detects packet loss, or error protocol packets are detected.
5	Notification	A key operation that is performed to ensure normal operations of the device, such as the interface shutdown, the neighbor discovery (ND), or the status change of the protocol state machine.
6	Informational	A common operation that is performed to ensure normal operations of the device. For example, the display command is run.
7	Debugging	Common information that is generated during normal operations of the device, which requires no attention.

- iMaster NCE-FabricInsight monitors all logs from level 0 to level 4, with statistics collected in different dimensions, such as the device name, IP address, module, severity level, and type, so as to quickly master the distribution of abnormal logs on the network.
- The system analyzes and displays fault logs on devices and allows you to filter logs by the device name, device IP address, module, severity, type, and details. Log severities include **Emergencies**, **Alert**, **Critical**, **Error**, and **Warning**.

Full Network Log Visualization Enables Intelligent Analysis on Abrupt and Occasional Exceptions

Intelligent identification of abrupt log changes and exceptions with proactive warning



Intelligent exception identification

Detect the abrupt change of network-wide logs based on machine learning and provide warnings in time.



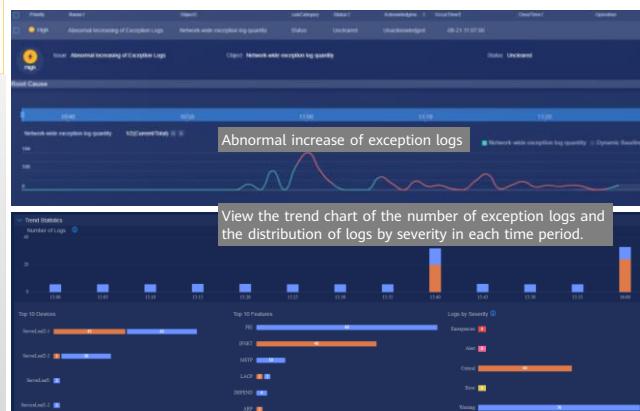
Occasional/New log analysis

Count the abrupt change of logs and the type, module, level, and quantity of new logs to quickly identify key check points.



Network-wide log event visualization

Display the trend, distribution statistics, and details of logs from level 0 to level 4 in multiple dimensions to present intuitive insights into network-wide log events.



- Application scenario:

- iMaster NCE-FabricInsight identifies exception logs that increase sharply on a network. By performing dynamic baseline exception detection, compressing logs, and comparing logs generated before and after exceptions, iMaster NCE-FabricInsight helps O&M personnel to quickly identify root causes of exceptions.

- Exception identification principles:

- iMaster NCE-FabricInsight checks whether the number of exception logs on the entire network increases sharply based on the dynamic baseline.
- It then analyzes the logs that increase sharply by log type and frequency to identify log distribution and check whether there are occasionally generated logs.
- iMaster NCE-FabricInsight performs multi-dimensional clustering analysis on the analysis result and automatically generates an issue, prompting users to solve the issue in a timely manner.

- View the trend of the number of exception logs and details about exception logs: The trend chart displays the trend of exception logs in the current time window, top 10 devices and features by the number of exception logs, and log distribution by severity.

- Move the pointer to a time period in the trend chart and view data in the time period.
- Move the pointer to a device, feature, or log severity to view the corresponding statistics.
- Click **Top 10 Devices**, **Top 10 Features**, or **Logs by Severity** to display the corresponding exception log statistics and log list.

Network Traffic Analysis Overview



NetStream flow analysis

Network traffic composition analysis based on traffic sampling

Analyze the network traffic, traffic trend statistics, and traffic characteristics from multiple dimensions based on the NetStream traffic sampling technology to identify abnormal network traffic and allocate resources properly.



ERSPAN (TCP)

Connectivity fault analysis based on service flow and network association

Implement one-click troubleshooting of connectivity issues based on correlation analysis between TCP services and networks through flow path visualization, hop-by-hop latency awareness (feature packets), and abnormal traffic analysis.



Edge intelligence

SLA detection based on TCP, UDP, and multicast flows

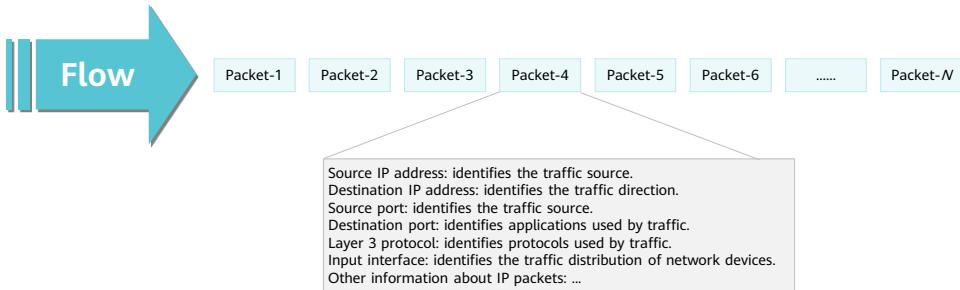
Monitor major services based on quality analysis on the connectivity and packet loss/latency of specified services, and quickly locate fault points after poor-QoE issues such as packet loss occur.



- Note:
 - ERSPAN: Encapsulated Remote Switched Port Analyzer

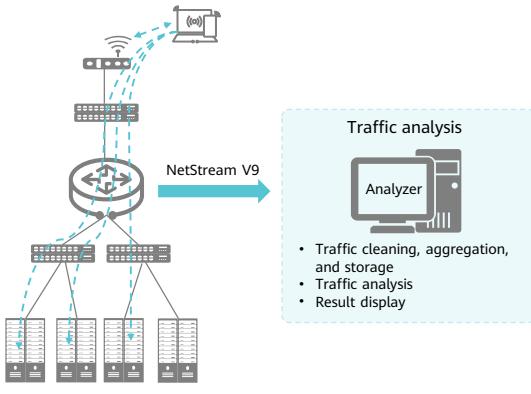
NetStream Packet Statistics Collection Principle

- A service flow is a flow of unidirectional data packets transmitted from a source IP address to a destination IP address. The packets in this service flow have the same attributes: source IP address, source port, destination IP address, destination port, IP protocol, and inbound and outbound interfaces. When a device receives the first IP data packet, a flow is initialized. All data packets that meet the characteristics of the flow are included into the byte count and packet count of the flow, and the information about the flow is uploaded through UDP for analysis.



- Packet sampling is enabled on an interface by running the **NetStream sampler random-packets packet-interval { inbound | outbound }** command.
- That is, packets are periodically sampled within the specified packet interval (1–65535). For example, if the interval is 100 packets, one random packet is sampled from every 100 packets.

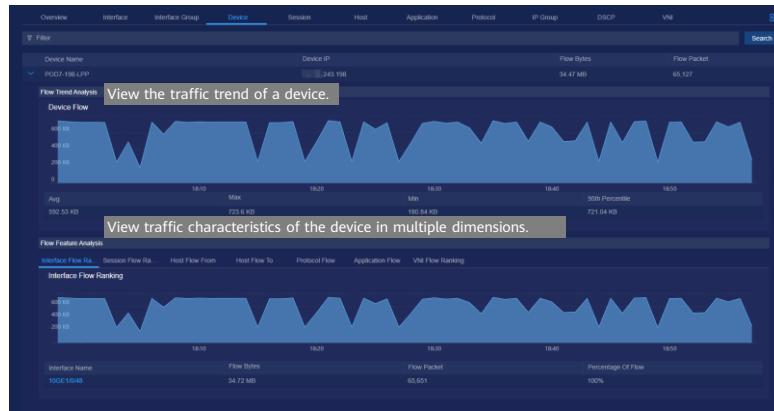
Refined Network Traffic Analysis Based on NetStream, Gaining Insights into Network Traffic Composition



- User-defined application visualization, supporting traffic analysis on 500 user-defined applications.
- Multi-dimensional traffic analysis and multi-dimensional drill-down correlation analysis, understanding the detailed traffic distribution and trend to properly allocate resources and identify network capacity expansion points.
- Customization of IP groups, port groups, overall analysis on subnets, domain names, and private line traffic, identifying the traffic proportion and analyzing abnormal traffic and interface usage to quickly locate network exceptions.

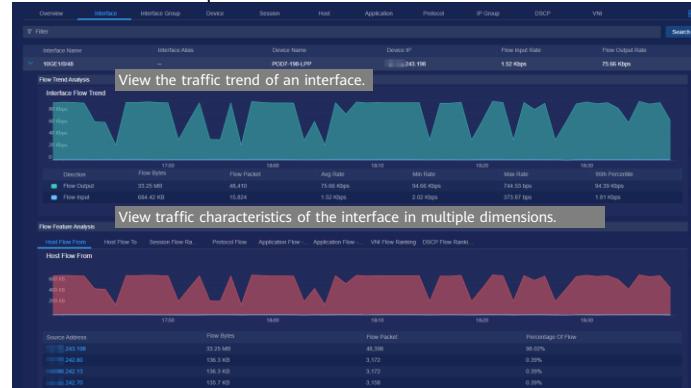
Network Traffic Analysis: Device Traffic

- iMaster NCE-FabricInsight provides a traffic statistics list by device. You can click a row to view the traffic trend of a device and multi-dimensional analysis results of traffic characteristics.



Network Traffic Analysis: Interface Traffic

- iMaster NCE-FabricInsight provides a traffic statistics list by interface. By default, interfaces are sorted in descending order of traffic volume. You can click a row to view the traffic trend of the current interface and the main traffic components of the interface from multiple dimensions.



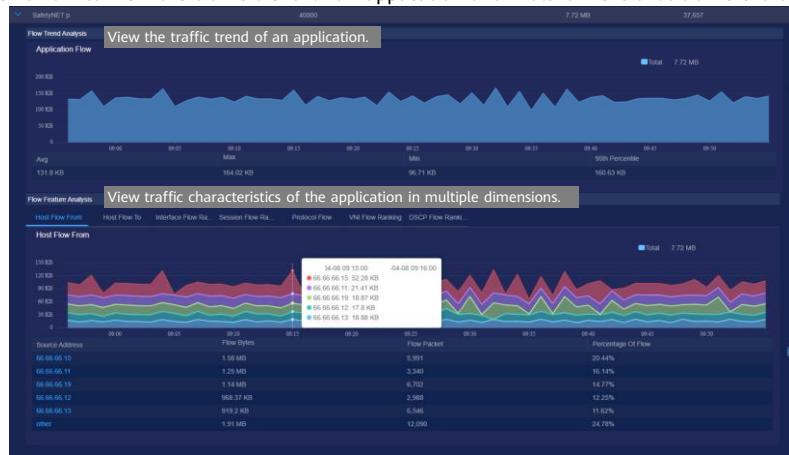
Network Traffic Analysis: Application Traffic (1)

- iMaster NCE-FabricInsight provides a traffic statistics list by application.

Overview	Interface	Interface Group	Device	Session	Host	Application	Protocol	IP Group	DSCH	VNI	
<input type="button" value="Search"/>											
▼ Filter											
						Application Name	Ports		Flow Bytes	Flow Packet	
>	Unknown App					--			45.9 MB	82,000	
>	app2					any			20.33 MB	58,891	
>	app242					any			13.05 MB	33,097	
>	app224					any			1.03 MB	25,362	
>	SafetyNET p					40000			423.89 KB	2,664	
>	BOOTP-server/DHCP					67			518.27 KB	980	
>	BOOTP-client/DHCP					68			316.71 KB	975	
>	SNMP					161			134.32 KB	730	
>	SNMP-Trap					162			85.18 KB	623	
>	SSDP					1900			61.24 KB	312	
											Show More

Network Traffic Analysis: Application Traffic (2)

- You can click a row to view the traffic trend of an application and multi-dimensional traffic characteristic analysis results.



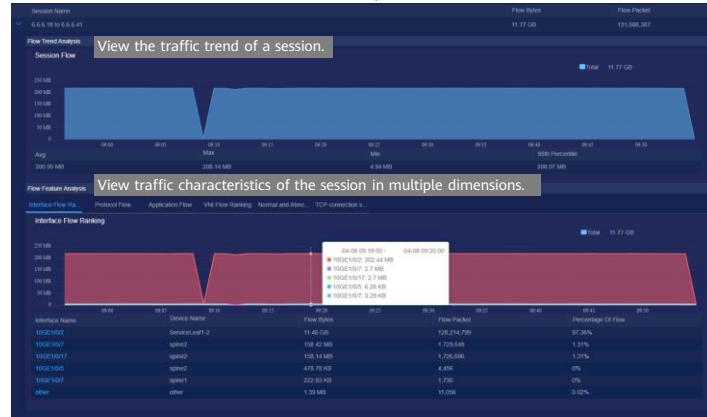
Network Traffic Analysis: Host Traffic

- iMaster NCE-FabricInsight provides a traffic statistics list by host. You can click a row to view the traffic trend of a host and multi-dimensional traffic characteristic analysis results.

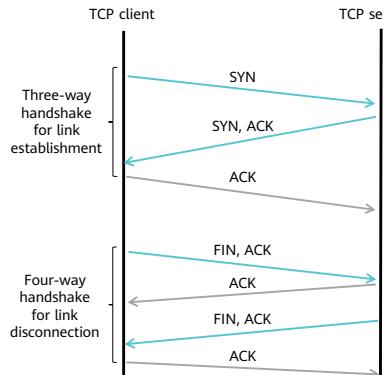


Network Traffic Analysis: Session Traffic

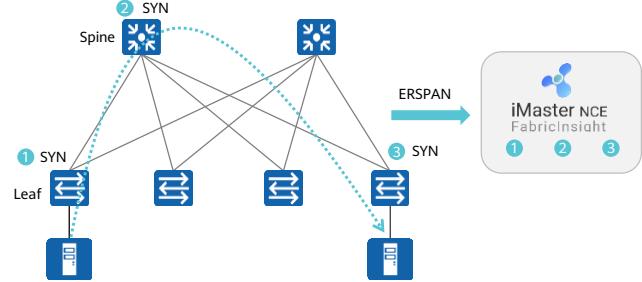
- iMaster NCE-FabricInsight provides a traffic statistics list by session. You can click a row to view the traffic trend of a session and multi-dimensional traffic characteristic analysis results.



ERSPAN Flow Analysis: TCP Control Packet Collection Principle



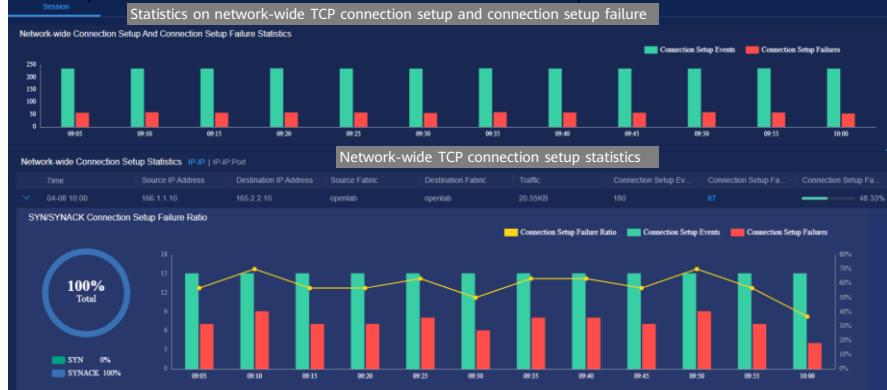
TCP flows on the network can be collected through ERSPAN mirroring of TCP feature packets (SYN, FIN, and RST packets).



- iMaster NCE-FabricInsight can obtain the following information about a TCP flow:
 - Packet forwarding route information.
 - TCP start and end time.
 - Transmitted bytes. (FIN serial number minus SYN serial number.)
 - SYN route latency and FIN route latency.
 - Exception: latency >1 ms, TCP Flags exception (RST), TCP retransmission, TTL < 3, etc.

ERSPAN Flow Visualization: TCP Flow Analysis

- The **Dashboard** page provides the statistics of top N information in a session from multiple dimensions and session visualization statistics analysis pages to display the trends of network-wide TCP connection setup times and TCP connection setup failures as well as network-wide TCP connection setup statistics.

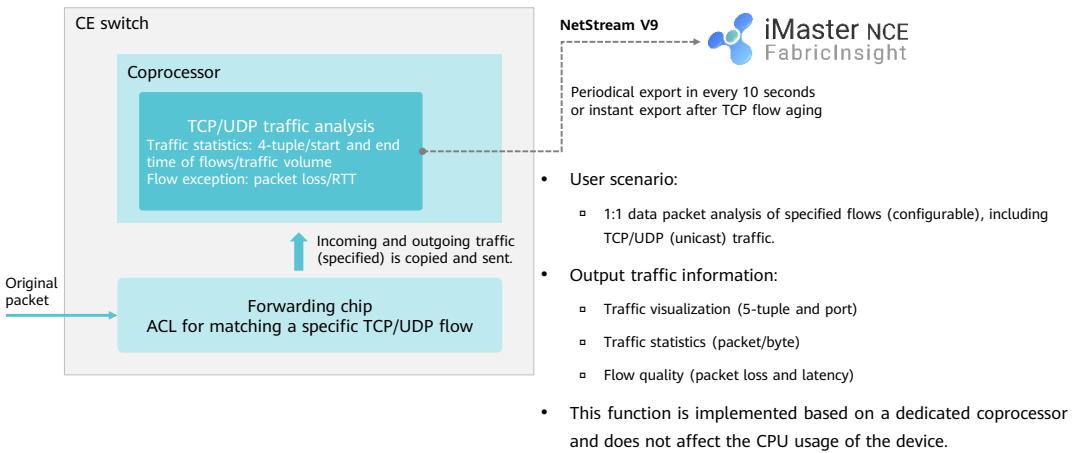


71 Huawei Confidential

HUAWEI

- Statistics on network-wide TCP connection setup and connection setup failure: This portlet displays the trend chart of the number of connection setup times and number of connection setup failures. By default, the system filters data by the "session status = connection setup failure" condition.
- Network-wide connection setup statistics: This portlet displays network-wide connection setup statistics. You can switch statistics between the 2-tuple (source IP address and destination IP address) and 3-tuple (source IP address, destination IP address, and destination port). You can click a row to view details about connection setup failures, including the ratio of SYN and SYN ACK connection setup failures, number of connection setup failures, statistics chart of the number of connection setup failures, and trend chart of the connection setup failure rate.

Edge Intelligence Solution

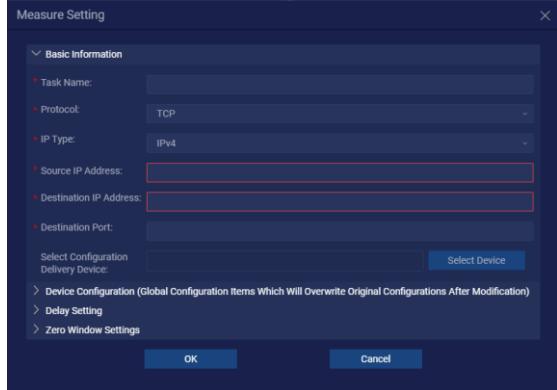


- Note:

- RTT: Round-Trip Time, indicating the total latency from the time when the transmit end sends data to the time when the transmit end receives an acknowledgment from the receive end (the receive end sends an acknowledgment immediately after receiving the data).

Edge Intelligence: Setting a Specified Flow Analysis Task

- Edge intelligence performs full-packet analysis on specific flows and proactively identifies information such as packet loss, latency, and zero windows.
- Set a flow analysis task:
 - You can set the source IP address, destination IP address, destination port, and protocol of a flow analysis task to measure and analyze specified flows. The configuration can be automatically delivered to selected devices.
 - When the protocol is TCP, IPv4 and IPv6 (IPv6 overlay) are supported and the device configuration items include ACL number, whether to match VXLAN packets, whether to match packets containing one-layer VLAN tags, whether to configure aging of TCP termination packets, aging time of active items, aging time of inactive items, unidirectional flow matching sequence number, and unidirectional flow matching mask. Latency settings include the RTT threshold. Zero window settings include the zero window threshold.



- When the protocol is UDP, only IPv4 is supported and the device configuration items include ACL number, whether to match VXLAN packets, whether to match packets containing one-layer VLAN tags, and aging time of inactive items. Latency settings include the latency threshold.

Edge Intelligence: Specified TCP Flow Analysis (1)

- Quality analysis of specified TCP flows:

The screenshot shows the Edge Intelligence interface with the following details:

- Overview Portlet:**
 - Task Name: Test
 - Protocol: TCP
 - Total Session Count: 1
 - Number of Packet Loss ...: 1
 - Enabled Measuring Point: 0/0
 - RTT Threshold (us): 5
 - Zero Window Threshold: 10
 - Device Configuration
- Basic Information Portlet:**
 - Number of Packet Loss Sessions: 1/1
 - Request packet loss/Total num.: 566/570
 - Response packet loss/Total num.: 564/570
 - Total request flow: 556.64KB
 - Total response flow: 556.64KB
 - Zero Windows in Request Direct...: 140
 - Zero Windows in Response Direct...: 0
- Analysis Conclusion Portlet:**
 - Request direction packet loss: Packet loss rate: 49.82%. The packet loss occurs in the downstream of device pod1-spine1-144 for request direction.
 - Response direction packet loss: Packet loss rate: 49.74%. The packet loss occurs in the downstream of device pod1-spine2-145 for response direction.
- Event List Portlet:**

IP-IP/Port	Application interac...	Total Traf...	Total Lost...	(Average/Maximu...	Zero Win...	First Zero Wind...	Last Zero Wind...	Total Traf...	Total Lost...	(Average/Maximu...	Zero Win...	First Zero Wind...	Last Zero Wind...
1.1->2.2:809	ads->Undefined App	234.37KB	236/540	0us/0us	120	3-03 15:1...	04-22 07:2...	234.37KB	236/540	0us/0us	0	-	-

74 Huawei Confidential



- Quality analysis of specified TCP flows:

- Basic information: displays the number of sessions with packet loss, total number of sessions, number of lost packets in the request directions, total number of packets in the request directions, number of lost packets in the response directions, total number of packets in the response directions, total traffic in the request direction, total traffic in the response direction, number of zero windows in the request direction, and number of zero windows in the response direction.
- Analysis conclusion: displays the packet loss rate analysis results, packet loss node analysis results, average RTT analysis results in the request and response directions, and analysis results for the maximum number of zero windows in the request and response directions.
- Event list: displays the source IP address, destination IP address, and destination port number for the measurement task, and allows you to view the topology and metrics for the session with the specified source IP address, destination IP address, and destination port number.

Edge Intelligence: Specified TCP Flow Analysis (2)



Edge Intelligence: Specified TCP Flow Analysis (3)

Measurement point-based flow quality analysis											
▲ (Average/Maximum) RTT of the Measurement Point (and Extension Group at the Same Level) in the Request Direction: 0us/0us ▲ (Average/Maximum) RTT in the Response Direction: 0us/0us											
Time	Device	Session Request Direction					Session Response Direction				
		Total Tr.	Total Lost Pa...	Measurement P...	RTT(Me...	Zero ...	First Zero ...	Last Zero ...	Total Tr...	Total Lost Pa...	Measurement P...
I-03-02 16:35:00	pod12-spine2-145	0B	0/0	-	0us	130	-03:02...	04:21...	556.64KB	964/570	▲ Downstream pac 0us
I-03-02 16:28:10	pod12-spine2-145	0B	0/0	-	0us	120	-03:02...	04:21...	273.44KB	277/280	▲ Downstream pac 0us
I-03-02 16:28:00	pod12-spine2-145	0B	0/0	-	0us	110	-03:02...	04:21...	253.91KB	257/260	▲ Downstream pac 0us
I-03-02 16:27:50	pod12-spine2-145	0B	0/0	-	0us	100	-03:02...	04:21...	234.37KB	237/240	▲ Downstream pac 0us
I-03-02 16:27:40	pod12-spine2-145	0B	0/0	-	0us	90	-03:02...	04:21...	214.84KB	217/220	▲ Downstream pac 0us
I-03-02 16:27:30	pod12-spine2-145	0B	0/0	-	0us	80	-03:02...	04:21...	195.31KB	197/200	▲ Downstream pac 0us
I-03-02 16:27:20	pod12-spine2-145	0B	0/0	-	0us	70	-03:02...	04:21...	175.78KB	177/180	▲ Downstream pac 0us
I-03-02 16:27:10	pod12-spine2-145	0B	0/0	-	0us	60	-03:02...	04:21...	156.25KB	157/160	▲ Downstream pac 0us
I-03-02 16:26:50	pod12-spine2-145	0B	0/0	-	0us	50	-03:02...	04:21...	117.19KB	118/120	▲ Downstream pac 0us
I-03-02 16:26:40	pod12-spine2-145	0B	0/0	-	0us	40	-03:02...	04:21...	97.66KB	98/100	▲ Downstream pac 0us

More

Overall Network Health Check, Systematically Evaluating the DCN Quality



Key assurance & network SLAs

You can view key metrics such as the network-wide key assurance object, intent verification result, transmission latency, and packet loss rate.



Network-wide resource status check

You can view network-wide underlay/overlay network resources and collected KPI metric data, and compare the data collected yesterday with that collected today.



Five-layer health evaluation system

You can view detailed analysis from dimensions such as device, network, protocol, service, and overlay to check whether the network health status is normal.



77 Huawei Confidential

HUAWEI

- Health evaluation refers to the evaluation on the overall health status of the current network based on identified network issues, helping you quickly and accurately identify and rectify faults.
 - This portlet displays the overall health status of the network based on multiple metrics such as the number of service assurance objects, network connectivity intent verification, average transmission latency, and packet loss rate. It also displays the distribution and growth of abnormal data from dimensions such as device and telemetry.
 - This portlet displays the number of pending issues, events, and resources from dimensions such as device, network, protocol, overlay, and service. You can click each layer to view the total number of events, resources, and events unassociated with issues of network entities.

Issues That Can Be Identified Based on Health Evaluation (1)

Category	Issue
Performance	Switch CPU threshold exceeded, switch memory threshold exceeded, service affected by switch interface congestion, firewall CPU or IPv4 session threshold exceeded, abnormal switch CPU usage increase, abnormal switch memory usage increase, abnormal firewall CPU usage increase, abnormal firewall memory usage increase, abnormal drop packet increase, abnormal error packet increase, abnormal unicast packet increase, abnormal multicast packet increase, abnormal broadcast packet increase, abnormal bandwidth usage change, abnormal huge page memory usage increase, and abnormal forwarding core usage increase
Capacity	Switch ARP entry threshold exceeded, switch ND entry threshold exceeded, switch MAC entry threshold exceeded, switch storage space threshold exceeded, switch ACL resource threshold exceeded, switch SFU forwarding performance insufficiency, switch FIB4 entry threshold exceeded, switch FIB6 entry threshold exceeded, number of routes received from a BGP peer exceeding the limit, abnormal switch ARP entry increase, abnormal switch ND entry increase, abnormal switch FIB4 entry increase, abnormal switch FIB6 entry increase, abnormal switch MAC entry increase, predicted traffic threshold exceeding, switch BD threshold exceeded, switch VRF entry threshold exceeded, switch Layer 2 sub-interface threshold exceeded, abnormal TCAM rule usage increase, predicted forwarding core usage threshold exceeding, abnormal EMC entry usage increase, abnormal ND-suppress entry usage increase, abnormal ARP-suppress entry usage increase, and abnormal virtual port usage increase

Issues That Can Be Identified Based on Health Evaluation (2)

Category	Issue
Status	Switch LPU exception, repeated switch LPU exception, switch MPU exception, switch SFU exception, repeated switch SFU exception, switch fan exception, switch power exception, link port status flapping, unidirectional link connectivity fault on the network side of a switch, routing loop, switch port Error-Down, suspected subhealthy optical link, suspected switch entry change, switch ARP entry loss, switch routing table loss, BGP peer status flapping, access-side IP address conflict on the VXLAN network, suspected Layer 2 loop, optical module type mismatch, repeated switch MPU exception, repeated switch restart, switch fault, switch disconnection, switch M-LAG dual-active state, switch chip soft failure, VXLAN tunnel interruption, license file expiration, OSPF router ID conflict, physical switch port suspension, OSPF DR IP address conflict, OSPF neighbor status change, BGP peer status change, stack fault, host IP address conflict, IP address conflict on the network side, access-side port blocked by STP, license file about to expire, and abnormal increase of exception logs
Policy	TCP SYN flood attack, ARP attack, ND attack, and invalid ARP packet received by a switch
Connection	Single IP address fault on the access side, server access fault, TCP service port not enabled, TCP service port fault, and service interruption caused by BD deletion, sub-interface shutdown, or sub-interface deletion
Intent	Inconsistent link and port metrics, routing loop on the entire network, routing blackhole on the entire network, service reachability intent verification failure, and service isolation intent verification failure

Real-Time or Periodical Push of Health Reports, Providing References for Optimization

Network overview

I. Network Overview:

Network health status check is performed in terms of Device, Network, Protocol, Overlay, Service. The check covers the resource running State, Performance, Capacity, Policy, Intent. Connection to intuitively display the overall experience quality of the entire network, helping O&M personnel quickly identify issue objects and driving continuous improvement of network quality.

1.Resource Overview:

This test covers 14 Fabric Details:..

default1:

Device: 20 device(s), 21 board(s), 37 power(s), and 57 fan(s)..

Network: 17 link(s), 10131 instance(s), 114 optical module(s) and 9 intent(s)..

Protocol: 17 OSPF instance(s), 22 OSPF Peer(s), 14 BGP instance(s), 26+ BGP Peer(s), and 3 M-LAG instance(s)..

Overlay: 24 NVE instance(s) and 110 VAP instance(s)..

Service: 9 host IP address(es) and 15 intent(s)..

DC11:

Device: 8 device(s), 13 board(s), 18 power(s), and 33 fan(s)..

Network: 50 link(s), 2101 instance(s), 88 optical module(s) and 9 intent(s)..

Protocol: 10 OSPF instance(s), 28 OSPF Peer(s), 30 BGP instance(s), 56+ BGP Peer(s), and 4 M-LAG instance(s)..

Overlay: 10 NVE instance(s) and 57 VAP instance(s)..

KPI details

II. Metric Details:

. 1. DeviceHealth Check:

. Overview:

No resource detected.No issue is detected..

default1:

Detected resource: 24 device(s), 27 board(s), 48 power(s), and 73 fan(s). Detected 57 issues. 1 issue(s) are unclerked, 56 issue(s) are cleared..

. KPI Information:

Default:

1.CPU/Memory Usage:

Top 5 CPU Usage:

Device	CPU Usage	Detail	Memory Usage
Argo	Max.: 100% (1m)	Min.: 0% (1m)	Max.: 100% (1m)
Max.: 100% (1m)	Min.: 0% (1m)	Max.: 100% (1m)	
Min.: 0% (1m)	Max.: 100% (1m)	Min.: 0% (1m)	
Max.: 100% (1m)	Min.: 0% (1m)	Max.: 100% (1m)	
Min.: 0% (1m)	Max.: 100% (1m)	Min.: 0% (1m)	

Top 5 Memory Usage:

Device	CPU Usage	Detail	Memory Usage
Argo	Max.: 100% (1m)	Min.: 0% (1m)	Max.: 100% (1m)
Max.: 100% (1m)	Min.: 0% (1m)	Max.: 100% (1m)	
Min.: 0% (1m)	Max.: 100% (1m)	Min.: 0% (1m)	
Max.: 100% (1m)	Min.: 0% (1m)	Max.: 100% (1m)	
Min.: 0% (1m)	Max.: 100% (1m)	Min.: 0% (1m)	

2.PIB/ARP Resource:

Top 5 PIB Usage:

Device	Argo	Detail	Max.: 100% (1m)
Argo	Max.: 100% (1m)	Detail	Max.: 100% (1m)
Detail	Max.: 100% (1m)	Argo	Max.: 100% (1m)
Max.: 100% (1m)	Detail	Argo	Max.: 100% (1m)
Detail	Argo	Max.: 100% (1m)	Max.: 100% (1m)

Report details

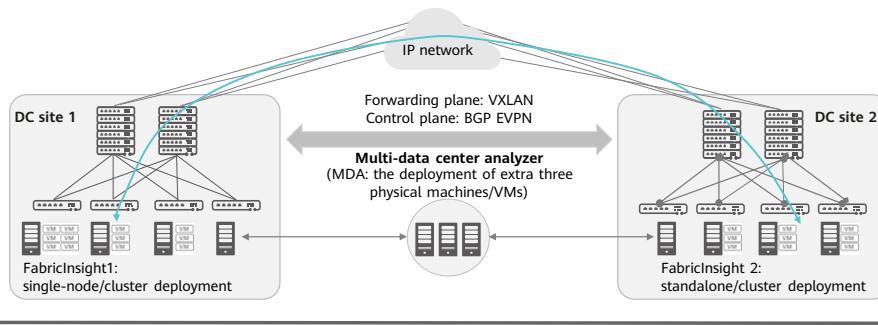
.Check Report Details:

Name	Objects	Status	Description	Detected Time	Vibe	Details
Switch Health	23 devices	Normal	No device health issue	2023-01-20 09:00:00	-	
Network Health	23 devices	Normal	All resources are normal	2023-01-20 09:00:00	-	
Switch Disconnection	23 devices	Normal	Switch disconnected	2023-01-20 09:00:00	-	
Switch MPPs	23 devices	Normal	No MPP connection issue	2023-01-20 09:00:00	-	
Network Health	23 devices	Normal	No network connection issue	2023-01-20 09:00:00	-	
Switch IP Exception	23 devices	Normal	No switch IP exception	2023-01-20 09:00:00	-	
Network Health (IP)	23 devices	Normal	No network IP exception	2023-01-20 09:00:00	-	
Network Health (IP)	23 devices	Normal	No repeated IP	2023-01-20 09:00:00	-	
Switch IP Exception	23 devices	Normal	No switch IP exception	2023-01-20 09:00:00	-	
Repetitive Switch (IP)	23 devices	Normal	No IP exception over redundant	2023-01-20 09:00:00	-	
Switch IP Exception	23 devices	Normal	No switch IP exception	2023-01-20 09:00:00	-	
Switch Power	23 devices	Normal	No power consumption issue	2023-01-20 09:00:00	-	
Diagnostic Loop	23 devices	Normal	No loop issue	2023-01-20 09:00:00	-	
Switch Health	23 devices	Normal	No switch health issue	2023-01-20 09:00:00	-	
Minimum Remaining LifeTime	23 devices	Normal	Minimum remaining life time	2023-01-20 09:00:00	Green	
The primary system	23 devices	Abnormal	Primary system abnormal	2023-01-20 09:00:00	Red	
Unused Allocated	23 devices	Normal	No IP for the unallocated resources	2023-01-20 09:00:00	-	
Unused IP	23 devices	Normal	No unused IP	2023-01-20 09:00:00	-	
Unused Memory	23 devices	Normal	No unused memory usage over threshold	2023-01-20 09:00:00	-	
Unused Memory	23 devices	Normal	No unused memory usage over threshold	2023-01-20 09:00:00	-	

80 Huawei Confidential

- Network overview:
 - Display different key information, including resource overview, client overview, and quality overview.
- KPI details:
 - Identify network quality issues based on five dimensions of the network health evaluation system.
- Report details:
 - Display the health status from five dimensions in detail and identify exceptions in a timely manner to provide optimization suggestions.

Multi-Cloud Multi-DC Analysis, with Unified Cross-Domain Health Evaluation



- Unified O&M portal:
 - Scenario example: a data center has 47 PoDs, with over 10 O&M portals before and different login passwords for each O&M system.
 - Solution: network-wide O&M requiring only one login based on single sign-on (SSO).
- Multi-domain application mutual access traffic visualization:
 - Scenario example: bandwidth costs are allocated and applications consuming a large number of bandwidth resources should be counted.
 - Solution: cross-DC/-fabric application interaction traffic and trend visualization, enabling fast identifications of abnormal burst traffic.
- Multi-domain interconnection traffic visualization:
 - Scenario example: private line bandwidth should be scaled in/out based on service changes, requiring evaluations on the inter-domain interconnection traffic.
 - Solution: traffic visualization of the Internet, VPN, and private line on the Fabric egress.
- Multi-domain network health evaluation:
 - Scenario example: the overall health status of the network should be evaluated in routine inspections to check whether the network traffic increases or decreases sharply.
 - Solution: network-wide health condition interpretation from dimensions of north-south DC traffic, east-west DC traffic, and intra-DC traffic.

Health Evaluation by MDA



Cross-DC/-fabric network access traffic

Cross-DC interconnection network traffic is identified based on egress links to analyze the traffic of Internet, VPNs, and private lines at the fabric egress.



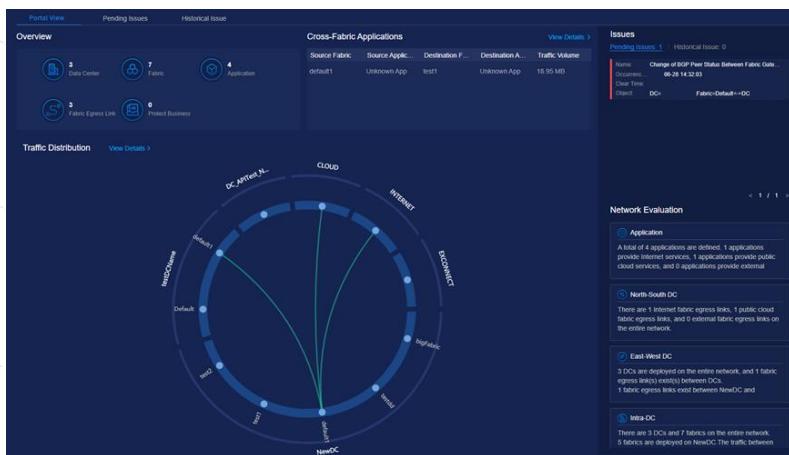
Cross-DC/-fabric application access traffic

Cross-fabric application interaction traffic and trend are displayed to quickly identify abnormal traffic changes, facilitating fault locating and capacity expansion.



Cross-DC/-fabric network evaluation

The composition of north-south DC, east-west DC, and intra-DC traffic at peak hours is analyzed to identify the applications with high traffic at peak hours, and evaluate the overall network health status.



- The health evaluation function evaluates the overall health status of the current network based on identified network issues, helping customers quickly and accurately identify and rectify faults.
 - Network evaluation: This portlet analyzes the composition of north-south and east-west traffic at peak hours, identifies the applications with high traffic at peak hours, and evaluates the overall network health status.
 - Traffic distribution: You can click **View Details** to view details about the composition of north-south and east-west traffic on the entire network, including the traffic statistics, bandwidth usage, and health evaluation statistics.

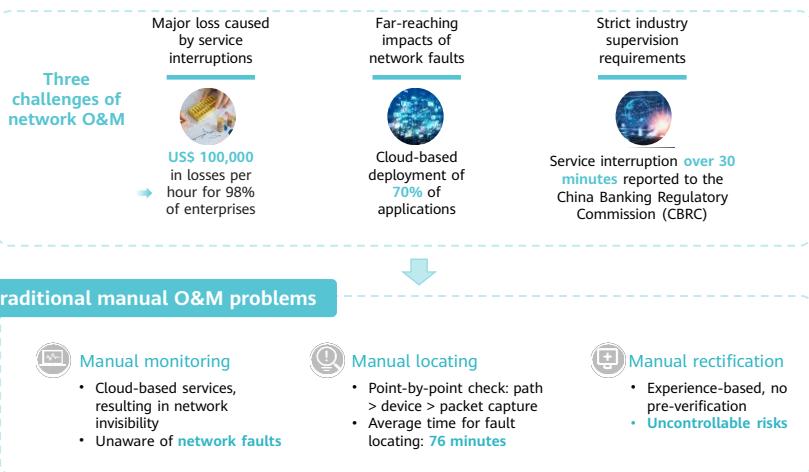
Issues That Can Be Identified Based on Health Evaluation by MDA

Category	Issue
Capacity	Cross-fabric routes received from BGP peers exceeding the threshold
Status	Change of BGP peer status between fabric gateways, BGP peer relationship flapping between fabric gateways, cross-fabric host IP address conflict, and VXLAN tunnel interruption
Intent	Link port metrics inconsistency, routing loop on the entire network, routing blackhole on the entire network, service reachability intent verification failure, and service isolation intent verification failure

Contents

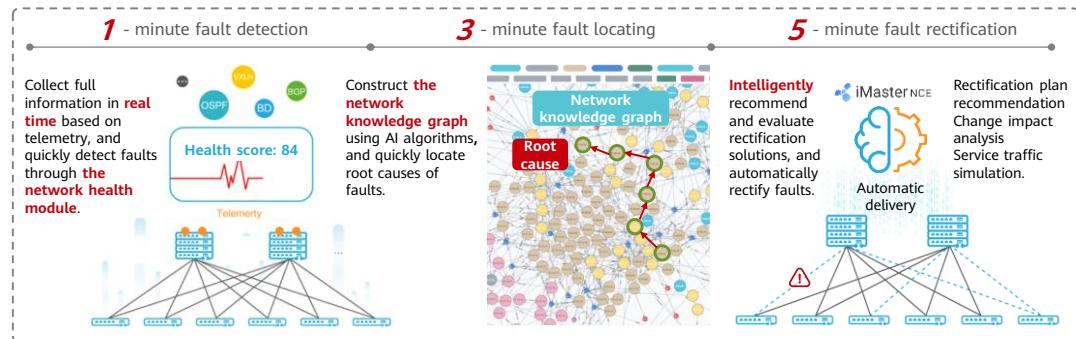
1. DCN O&M Challenges and CloudFabric Intelligent DCN O&M Solution
2. iMaster NCE-Fabric
- 3. iMaster NCE-FabricInsight**
 - Overview
 - Network Visualization and Health Evaluation
 - Fault Locating**
 - Change Assurance

"1-3-5" Troubleshooting (1)

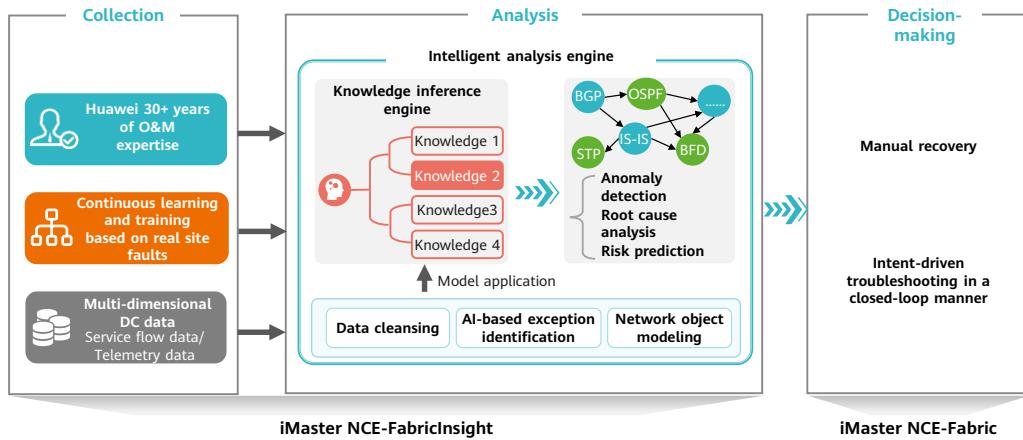


- The passive troubleshooting based on personal experience evolves into the AI-based automatic locating of closed loops.

"1-3-5" Troubleshooting (2)



AI-based Knowledge Inference, Achieving Fast Fault Locating



"1-3-5" Troubleshooting Scope (1)

Dimension	Issue	Number of Issues	SDN Network	Non-SDN Network
Device	Switch faults, repeated switch restarts, switch disconnections, switch MPU exceptions, repeated switch MPU exceptions, switch LPU exceptions, repeated switch LPU exceptions, switch SFU exceptions, repeated switch SFU exceptions, switch fan exceptions, switch power exceptions, switch CPU threshold crossing/abnormal increase of the switch CPU usage, switch memory usage threshold-crossing/abnormal increase of the switch memory usage, switch ACL resource threshold crossing, switch FIB4 entry threshold crossing/abnormal increase of switch FIB4 entries, switch FIB6 entry threshold crossing/abnormal increase of switch FIB6 entries, switch ND entry threshold crossing/abnormal increase of switch ND entries, switch ARP entry threshold crossing/abnormal increase of switch ARP entries, switch MAC entry threshold crossing/abnormal increase of switch MAC entries, switch SFU forwarding performance insufficiency, switch storage space threshold crossing, stack faults, suspicious Layer 2 loops, abnormal increase of exception logs, firewall CPU usage or IPv4 session resource usage threshold crossing, abnormal increase of the firewall CPU usage, abnormal increase of the firewall memory usage, license file expiration, license about to expire, switch BD/VRFL2 sub-interface threshold crossing, traffic exceptions caused by lost switch routing hardware tables, flow exceptions caused by CE switch chip soft failures, traffic exceptions caused by switch entry inconsistencies between the software table and hardware table, and traffic exceptions caused by lost switch ARP entries	43	✓	✓

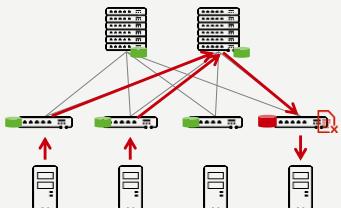
"1-3-5" Troubleshooting Scope (2)

Dimension	Issue	Number of Issues	SDN Network	Non-SDN Network
Network	Access-side interfaces blocked by STPs, suspected sub-healthy optical links, services affected by switch interface congestions, link interface status flapping, switch interface error-down, switch physical interface suspended, unidirectional link connectivity faults on the network side of a switch, host IP address conflicts, IP address conflicts on the network side, predicted traffic threshold crossing, link interface metric inconsistency, routing loops on the entire network, routing blackholes on the entire network, abnormal increase of drop packets, abnormal increase of error packets, invalid ARP packets received by switches, optical module type mismatch, ARP attack, and ND attack	19	✓	 The following issues are excluded: Link interface metric inconsistency Routing loops on the entire network Routing blackholes on entire network
Protocol	OSPF router ID conflicts, changes of the OSPF neighbor status, OSPF DR IP address conflicts, changes of the BGP neighbor status, BGP peer relationship flapping, two master switches in M-LAG, and routes received from BGP peer threshold exceeding the limit	7	✓	✓
Overlay	VXLAN tunnel interruptions, IP address conflicts on the VXLAN network access side, service interruptions due to BD deletion, and service interruptions due to sub-interface deletion	5	✓	✗
Service	Access-side single IP address exceptions, server access exceptions, TCP service interface exceptions, TCP service interface disabled, TCP SYN flood attacks, service reachability intent verification failures, and service isolation intent verification failures	7	✓	✗

Use Case: Millisecond-Level Queue Detection and Proactive Identification of Service Packet Loss

Service interruption caused by packet loss due to microbursts, resulting in difficult fault locating.

Big data services require a large number of servers to form a cluster and work together. Once the traffic of multiple nodes is sent to the same compute node, **packet loss due to transient congestion** may occur on the network and services are interrupted.



- The traditional NMS collects data every 5 minutes, unable to identify microbursts.
- Issues occur irregularly, which are difficult to trace and reproduce.

Timely display of service packet loss based on millisecond-level detection, enabling fast fault locating.

- Interface buffer size is detected at a **100-ms interval** based on telemetry.
- For example, when packets are discarded due to queue congestion, the 5-tuple details (port-queue-discarded packet) should be proactively detected.
- Faults are discovered based on interfaces.



90 Huawei Confidential

HUAWEI

- Note:
 - NMS: Network Management System

Case: Root Cause Analysis of Service Packet Loss



Case: Handling Suggestions for Service Packet Loss

Repair Advice

Repair advice

Step 1: Run the **display qos queue statistics interface port-type port-number** command on the switch (CE8861-4C-EI is used as an example) to check whether packet loss of each queue on the interface always occurs.

Queue	CIR-FIR	Passed	Pass Rate	Dropped	Drop Rate
(% or bytes)	(packets/bytes)	(pps/bps)	(packets/bytes)	(pps/bps)	
0	0	0	0	0	0
0	25000000	0	0	0	0
1	0	36489	0	0	0
0	25000000	2335296	32	0	0
2	0	0	0	0	0
0	25000000	0	0	0	0
3	0	0	0	0	0
0	25000000	0	0	0	0

Step 2: Choose Telemetry > Interface on FabricInsight, check whether the traffic trend of the interface complies with the historical trend.

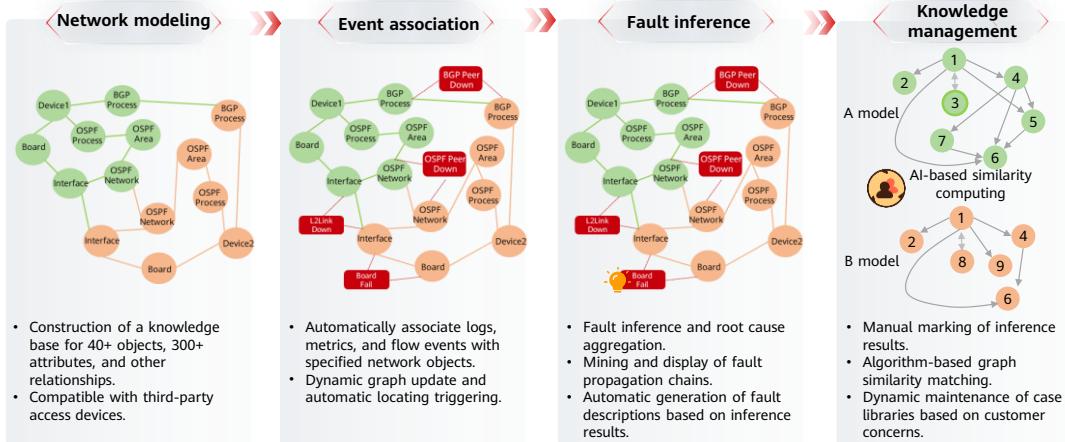
Interface Received Packets Trends

Interface: ge0/14_ServLeaf-vmg serc=1 o... maximum : 216,177,141 minimum : 125 average : 62,677,371
216,177,141
100,000,000

Step 3: If the traffic trend does not comply with the historical trend and the traffic increases sharply, wait for 30 minutes and check whether any application fault is reported. If no fault is reported, close this issue.

Step 4: If a service fault is reported, notify the corresponding service team. If packet loss due to congestion always occurs, migrate some high-traffic VMs or use switches with higher bandwidth.

Unknown Fault Inference: Network Fault Inference and Source Tracing Based on Knowledge Graphs



Unknown Fault Inference: Exception Analysis

Exception Analysis

- Total Events: 1272
- Events Unassociated with Issues: 1271
- Events in Fault Exploration: 60
- Possible Root Cause: 4

Select a root cause to view the fault propagation paths.

Event Name	Occurrence Time	Possible Root Cause
bgpBackwardTransition_active	12-23 09:48:16	The BGP Peer transition from higher numbered state A to lower numbered state. (bgpPeerNumber<=19) ...
hwBoardResThresholdExceeded	12-23 09:32:19	Possible Root Cause
hwCLogFileAging	12-23 09:12:05	Possible Root Cause
hwBoardResThresholdExceeded	12-23 09:34:55	Possible Root Cause

Fault Propagation Paths

Number of paths: 5

```

graph TD
    A[hwBoardResThresholdExceeded_active] --> B[bgpBackwardTransition_active]
    B --> C[ospfNbrStateChange_active]
    C --> D[hwBoardResThresholdExceeded]
    D --> E[hwBoardResThresholdExceeded]
    E --> F[hwBoardResThresholdExceeded]
  
```

View fault propagation paths by alarm.

Display fault source tracing by network object in a graph.

Event list

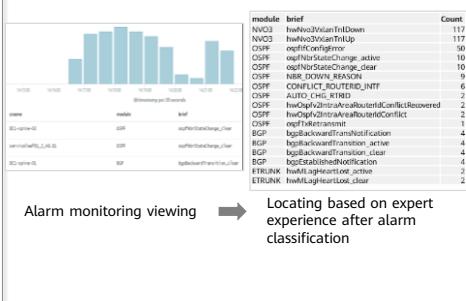
Event Level	Event Name	Object Name	Occurrence Time
Warning	SNMP_MIB...	POD7-board...	12-23 09:30:00
Warning	CONFIGMIS...	POD7-board...	12-23 09:30:00
Warning	SNMP_MIB...	POD7-board...	12-23 09:30:00
Critical	bgpBackward...	POD7-board...	12-23 09:30:00
Critical	linkDown_acti...	POD7-board...	12-23 09:30:00
Warning	hwEthernetA...	POD7-board...	12-23 09:30:00
Warning	hwAppPacker...	POD7-board...	12-23 09:30:00
Warning	TransceiverTy...	POD7-board...	12-23 09:30:00
Alert	hwGInReach...	POD7-board...	12-23 09:30:00
Critical	hwFlagPeerR...	POD7-board...	12-23 09:30:00
Critical	hwHostIPConf...	POD7-board...	12-23 09:30:00
Warning	INVO3_TUNN...	POD7-board...	12-23 09:30:00
Warning	hwStorageBL...	POD7-board...	12-23 09:30:00
Warning	hwAppCallow...	POD7-board...	12-23 09:30:00
Warning	hwAppPacker...	POD7-board...	12-23 09:30:00
Warning	hwMtpSdStar...	POD7-board...	12-23 09:30:00

Use Case: Minute-level Root Cause Locating Based on Knowledge Graphs

AS-Is:

Manual locating based on alarms and experience

Scenario: Services are interrupted in a bank and 300+ alarms are reported within a short period. As a result, the one-hour manual troubleshooting shows that the interruption is mainly caused by the router ID conflict.



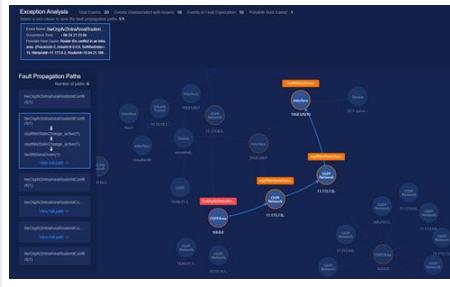
Alarm monitoring viewing

VS.

TO-Be:

Automatic exploration of root causes based on knowledge graphs

Root causes of fault issues are proactively reported, enabling minute-level automatic locating of fault points.



HUAWEI

95 Huawei Confidential

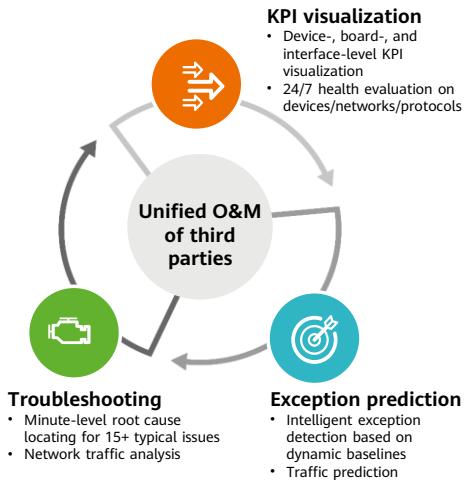
- Challenges to traditional O&M:

- Massive alarms: hundreds of alarms triggered by a fault, resulting in difficult locating.
- Manual fault locating: locating based on expert experience, requiring a lot of time.
- Source tracing failure: unable to identify the fault impact scope, resulting in difficult root cause tracing.

- Intelligent O&M:

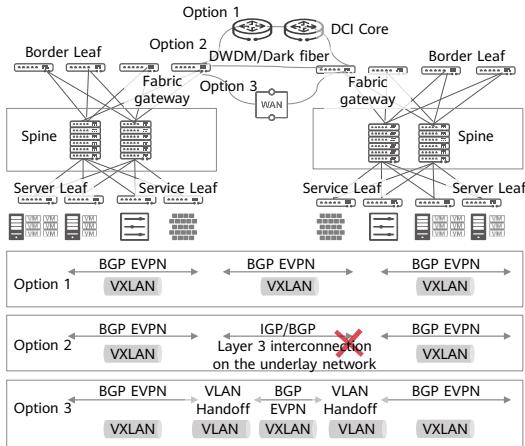
- Fault aggregation: only root causes of fault issues are reported.
- Automatic locating: AI-based intelligent inference, not relying on personnel skills.
- Path visualization: fault propagation path display and fault impact scope identification.

Unified O&M of Multiple Vendors



Function Category	Detailed List
Troubleshooting	<ul style="list-style-type: none"> • Switch fault • Repeated switch restart • Abnormal increase of exception logs • Switch LPU fault • Suspicious Layer 2 loop • IP address conflict on the network side • EtherChannel interface down • Switch interface error-down • Link interface status flapping • OSPF router ID conflict • Predicted traffic threshold exceeding • Abnormal increase of the switch CPU usage • Abnormal increase of the switch memory usage • Abnormal increase of discarded packets on an interface • Abnormal increase of error packets on an interface
KPI visualization	<ul style="list-style-type: none"> • CPU usage • Memory usage • Number of received/sent packets, number of received/sent broadcast packets, number of received/sent multicast packets, number of received/sent unicast packets, number of received/sent bytes, number of received/sent discarded packets, and number of received/sent error packets

Minute-level Proactive Detection and Locating of Inter-DC or Inter-Fabric Faults



- MDA identifies inter-DC or inter-fabric network issues in minutes through the knowledge graph modeling and analysis of cross-domain networks to quickly locate root causes.
- The following inter-DC or inter-fabric network issues can be identified currently:

Issue List
BGP peer status changing between fabric gateways
BGP peer status flapping between fabric gateways
Cross-fabric routes received from BGP peers threshold exceeding the limit
VXLAN tunnel interruption
Cross-fabric host IP address conflict
Link interface metric inconsistency
Routing loop on the entire network
Routing blackhole on the entire network
Service reachability intent verification failure
Service isolation intent verification failure

Cross-DC or Cross-Fabric Pending Issues and Historical Issues

- Issues are identified using certain rules based on original exception information such as exception logs reported by devices and detected KPI exceptions.

The screenshot shows two tables of issues:

Pending Issues:

Priority	Name	Object	Type	Clearance Status	Acknowledgment Status	OccurTime	ClearTime	Operation
High	Change of BGP Peer Status Between DC=DC1.Fabric<default2><->DC=DC2@get%&Y	DC=DC1.Fabric<default2>	Status	Uncleared	Unacknowledged	05-08 18:35:37		<input checked="" type="checkbox"/>
High	Change of BGP Peer Status Between DC=DC1.Fabric<default2><->DC=DC1.Fabric<data...>	DC=DC1.Fabric<default2>	Status	Uncleared	Unacknowledged	05-08 17:13:59		<input checked="" type="checkbox"/>
High	Cross-fabric Host IP Address Conflict L2 VNI=5141, Host IP=188.122.1.3	L2 VNI=5141	Status	Uncleared	Unacknowledged	05-08 14:25:42		<input checked="" type="checkbox"/>
Medium	Service reachability intent verification failed Fabric=DC2.default.DC1.Gateway	Fabric=DC2.default.DC1.Gateway	Intent	Uncleared	Unacknowledged	05-08 11:35:42		<input checked="" type="checkbox"/>

Historical Issues:

Priority	Name	Object	Type	Clearance Status	Acknowledgment Status	OccurTime	ClearTime	Operation
High	Change of BGP Peer Status Between DC=DC1.Fabric<default2><->DC=DC1...	DC=DC1.Fabric<default2>	Status	Cleared	Acknowledged	05-08 16:51:36	05-08 16:46:33	
High	Cross-fabric Routes Received from BG Local DC=DC1.Local Fabric=<defa...	Local DC=DC1.Local Fabric=<defa...	capacity	Cleared	Acknowledged	05-08 16:46:11	05-08 17:06:02	
High	Change of BGP Peer Status Between DC=DC1.Fabric<default2><->DC=DC1...	DC=DC1.Fabric<default2>	Status	Cleared	Acknowledged	05-08 16:46:10	05-08 16:43:51	
High	Change of BGP Peer Status Between DC=DC1.Fabric<default2><->DC=DC1...	DC=DC1.Fabric<default2>	Status	Cleared	Acknowledged	05-08 15:54:00	05-08 15:52:58	
High	Change of BGP Peer Status Between Local DC=DC1.Local Fabric=<defa...	Local DC=DC1.Local Fabric=<defa...	Status	Cleared	Acknowledged	05-08 15:25:04	05-08 15:44:51	

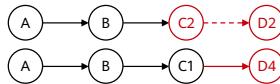
- The **Pending Issues** list displays all issues that are not cleared or acknowledged. The **Historical Issue** list displays all cleared and acknowledged issues. You can view issue details, including basic issue information and issue impact scope.
- Issues of the MDA health evaluation function and issues of the iMaster NCE-FabricInsight health evaluation function are independent of each other. The two kinds of issues cannot be cleared at the same time or the acknowledgment status of them cannot be conducted at the same time.

Flow Analysis: Path Comparison

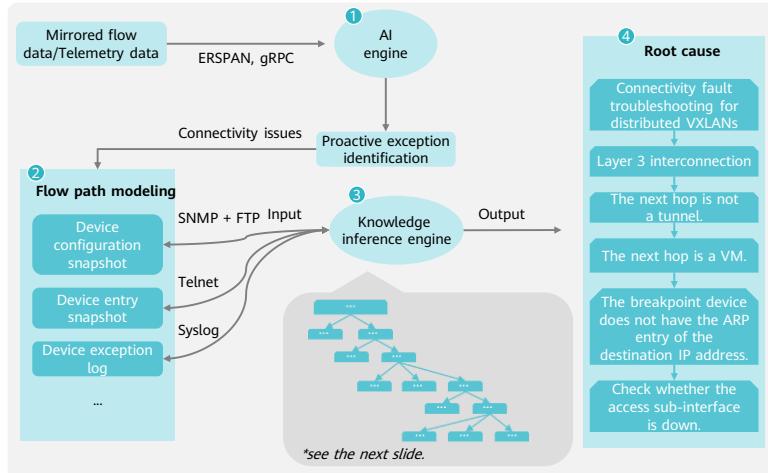
- Implementation of path comparison:
 - Automatic learning of network forwarding paths between VMs is enabled based on the VXLAN overlay forwarding model. For instance, four paths are available for IP1 to access IP2.



- When an exception occurs, the forwarding path of an abnormal packet and those of a normal one are compared to quickly detect their differences.
 - Scenario 1: The path is incomplete.
 - Scenario 2: The path changes.

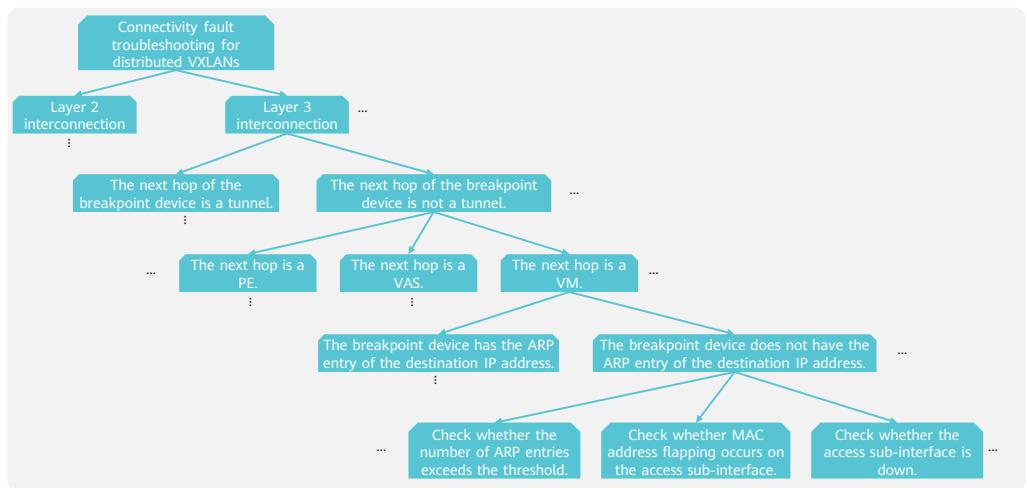


Flow Analysis: Troubleshooting Based on the Rule Engine



- ① Clustering analysis is conducted on the mirrored flow data on the network through the AI engine to proactively detect abnormal connectivity.
- ② Flow paths with abnormal connectivity are modeled to identify the logical topology of the event flow. Current abnormal flow paths and previous normal flow paths are compared to identify breakpoint devices of flow paths.
- ③ The fault inference is performed based on the fault inference rule library and the context of breakpoint devices to identify possible faults.
- ④ The detailed troubleshooting process is displayed.

Knowledge Inference Engine Example



ERSPAN TCP Flow Analysis: Flow Troubleshooting (1)

- The flow troubleshooting function displays ERSPAN mirrored packets that have undergone packet combination and request direction identification and allows you to query the ERSPAN mirrored packets by multiple dimensions (including the source IP address, source port, destination IP address, and destination port).
- For SYN/SYN ACK packets whose Status is displayed as TCP Retransmission or Flow events in which packet status is abnormal TTL, you can switch to the fault inference diagram as well as automatic and intelligent troubleshooting.

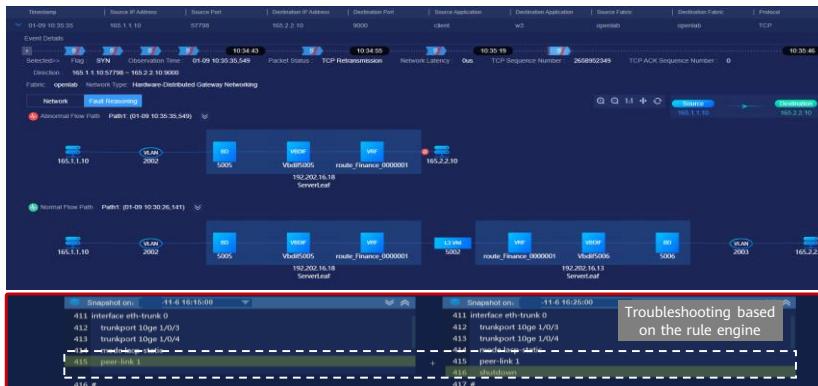


102 Huawei Confidential

HUAWEI

- By default, only abnormal flow events (TCP retransmission, abnormal TTL, TCP RST, and abnormal TCP flag) and long flows (TCP flows that are not terminated within 10 seconds) are displayed.

ERSPAN TCP Flow Analysis: Flow Troubleshooting (2)

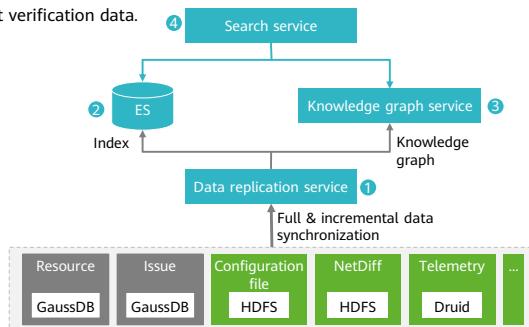


- Based on the expert experience library and troubleshooting process, iMaster NCE-FabricInsight summarizes a unified troubleshooting model and provides an automatic troubleshooting framework that can be orchestrated and requires no manual perception.
- Troubleshooting actions involve network checks. Users can perform one-click troubleshooting, improving the troubleshooting efficiency.

- Click **Fault Reasoning** to view the logical topology passed by the event. Compare the current abnormal flow paths with the previous normal flow paths.
- After the paths are calculated for normal flows and abnormal flows, you can click **No Troubleshooting** or **Timely Troubleshooting** to select a troubleshooting mode. Troubleshooting can be performed only when abnormal flow paths exist. By default, **No Troubleshooting** is used, indicating that troubleshooting is not performed. If normal flow paths exist and **Timely Troubleshooting** is selected, the system performs troubleshooting for log issues based on the timestamps of normal flow paths and abnormal flow paths, and performs troubleshooting for entry, configuration change, and firewall policy blocking issues based on the timestamps of normal flow paths and current timestamps. If no normal flow path exists and **Timely Troubleshooting** is selected, the system performs troubleshooting based on the timestamps of abnormal flow paths and current timestamps. Before performing troubleshooting for configuration change and firewall policy blocking issues, the system automatically synchronizes the latest configurations.

Overview and Principle of Intelligent Network Search

- The intelligent network search engine is an application constructed by iMaster NCE-FabricInsight based on the knowledge graph to quickly search the real-time network information.
- In the traditional O&M system, information is isolated and no association exists between configurations, entries, KPIs, and logs, requiring multiple searches and manual matching. Intelligent network search improves the searching efficiency of network data, including network resource data, device configuration files, device forwarding entries, "1-3-5" troubleshooting issues, device exception logs, KPIs, and intent verification data.



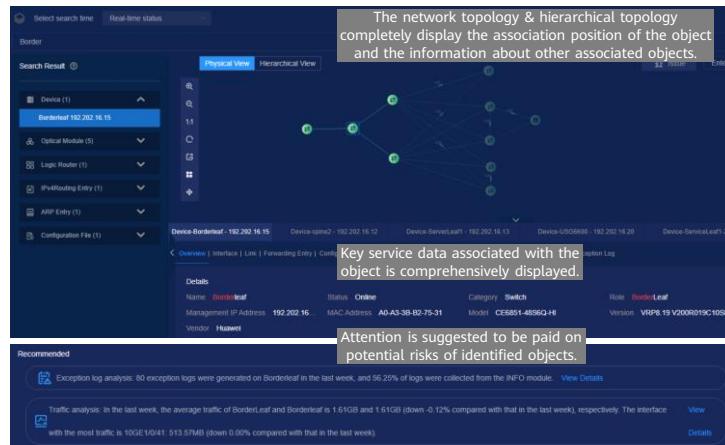
104 Huawei Confidential

HUAWEI

- Technical principles:

- Data replication service:
 - Multiple data source access mechanisms (Gauss DB, Druid, and HDFS).
 - Multi-source multi-mode data increment change awareness and incremental synchronization.
 - Full & incremental data synchronization efficiency and data compression algorithm.
 - Search service and elastic search:
 - Multi-source multi-mode data source mapping.
 - Indexed incremental change and remote data search efficiency.
 - Knowledge graph service:
 - Multi-source multi-mode data access and performance for querying TDBs.
 - Search performance and optimization in big data scenarios.
- Note:
 - HDFS: Hadoop distributed file system
 - TDB: trivial database

Minute-level Fault Demarcation Based on Intelligent Network Search



105 Huawei Confidential

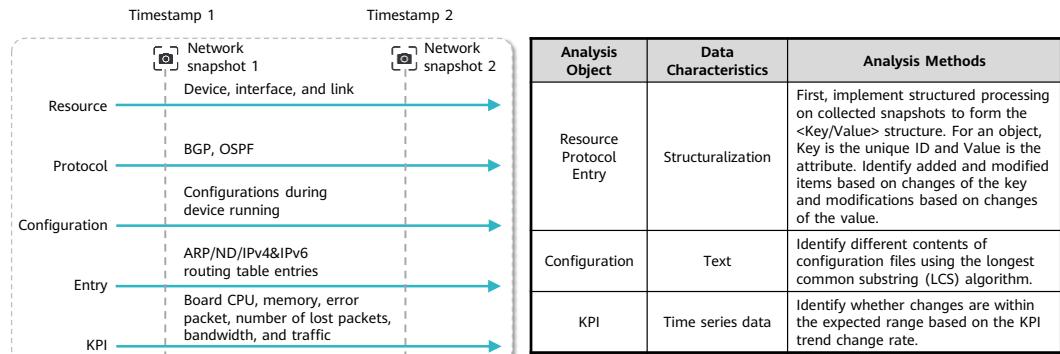
HUAWEI

- The network search function searches for resources, entries, configuration files, issues, and other objects on the network in a unified manner and displays information such as metrics, associated applications, and entries of target objects, as well as the recommended correlation analysis result. The search function can efficiently search for target objects and their associated data, improving O&M efficiency.
 - Searches for objects such as devices, boards, interfaces, power modules, fan modules, optical modules, ARP entries, routing table entries, configuration files, and issues.
 - Displays the physical topology and hierarchical topology of target objects.
 - Displays the recommended correlation analysis result of target objects.
 - Searches for issues and displays issue details.
 - Searches for entries and configuration files.
- Scenario:
 - In an enterprise, the network department receives a fault report from the service department that an IP service is interrupted, requiring joint troubleshooting.
- Solution: Search the VM IP address and obtain the comprehensive information about the VM to quickly locate the failure point. The information includes:
 - VM access location.
 - VM access interface status.
 - Whether congestions occur on the interface connected to the VM.
 - Whether changes occur in the configurations of gateways connected to the VM.
 - Whether the incoming and outgoing traffic of the VM changes sharply.
 - Whether the VM frequently goes online and offline.

Contents

1. DCN O&M Challenges and CloudFabric Intelligent DCN O&M Solution
2. iMaster NCE-Fabric
3. **iMaster NCE-FabricInsight**
 - Overview
 - Network Visualization and Health Evaluation
 - Fault Locating
 - **Change Assurance**

Network Snapshot Analysis Principles



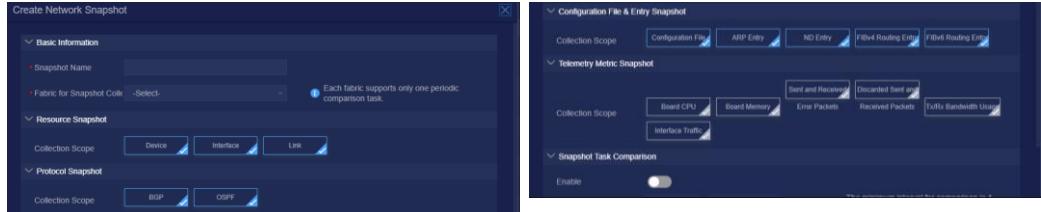
- The network snapshot refers to the data backup file running on a device at a specified time point. The first snapshot is the synchronization of the full device data. A new snapshot is created based on incremental changes.
- Changes of resources, protocols, configurations, entries, and KPI trends are managed in real time based on telemetry and network changes are rapidly detected based on the comparison between snapshots at different time points.

Automatic Check on 16 Changes in Five Dimensions Based on Snapshot Analysis (1)

Four steps of the DCN cutover tool

Step 1: information collection before a change

Create snapshot collection tasks before a change and provide snapshots of multiple performance metrics, such as device configurations, ARP entries, ND entries, RIB entries, CPU usage, memory usage, and interface bandwidths.



- Five dimensions: Configurations, Entries, Topologies, Capacities, and Performances

Automatic Check on 16 Changes in Five Dimensions Based on Snapshot Analysis (2)

Four steps of the DCN cutover tool

Step 2: information collection after a change

Automatically synchronize and analyze device configurations and entry snapshots after a change, supporting manual snapshot synchronization.

The screenshot shows a table of snapshots with columns: Creation Time, Name, Fabric, Type, Latest Collection Time, Status, and Operation. One row is highlighted in blue, showing '03-09 14:39:49' as the creation time, '0309' as the name, 'openlab' as the fabric, 'Manual' as the type, '03-15 20:16:11' as the latest collection time, and 'Collection completed' as the status. The 'Operation' column contains icons for edit, delete, and sync. A modal dialog box titled 'Information' displays a green checkmark icon and the message 'Snapshot tasks collected successfully.' with an 'OK' button at the bottom.

Creation Time	Name	Fabric	Type	Latest Collection Time	Status	Operation
03-15 20:14:52	0315	openlab	Manual		Collecting	
03-09 14:39:49	0309	openlab	Manual	03-15 20:16:11	Collection completed	
03-08 09:10:51	0308	openlab	Manual		To be collected	
03-03 11:01:14	0303	openlab			cted	
02-26 16:42:33	0226	openlab			cted	
02-22 16:06:50	0219	openlab			cted	
02-14 03:10:08	Test1	openlab			cted	

Total records: 7

Automatic Check on 16 Changes in Five Dimensions Based on Snapshot Analysis (3)

Four steps of the DCN cutover tool

Step 3: automatic analysis on change results

Compare and analyze data snapshots before and after a change, visualizing the differences of each device.



Automatic Check on 16 Changes in Five Dimensions Based on Snapshot Analysis (4)

Four steps of the DCN cutover tool

Step 4: comparison details about change differences

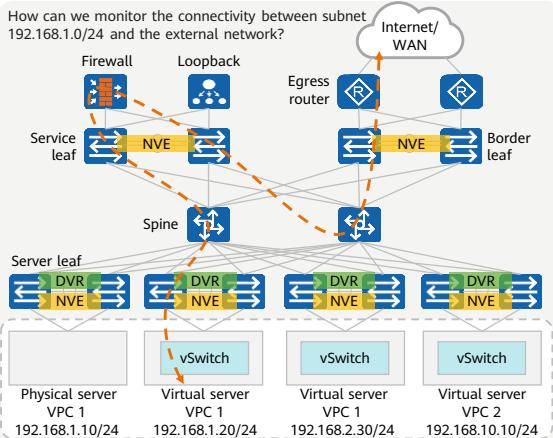
Display detailed comparisons of configured entries and other dimensions of snapshots before and after a change, identifying configuration changes.

Basic Snapshot Comparison Information									
Creation Time: 03-09 14:43:10		Progress: Success		Fabric: openlab		Before: 03-08 09:18:51		After: 03-09 14:39:49	
								Difference: 2	
Resource		New Entry: 0	Deleted Entry: 2	Modified Entry: 9	Same Entry: 170	Display by Device	Display by Entry	Select...	Search...
Device	0								
Interface	0								
Link	0								
Protocol									
BGP	0								
OSPF	0								
Configuration File & Entry									
configuration File	0								
ARP Entry	2								
IPv4 Routing Entry	0								

111 Huawei Confidential



Difficulties in Providing Service-oriented Assurance Capabilities



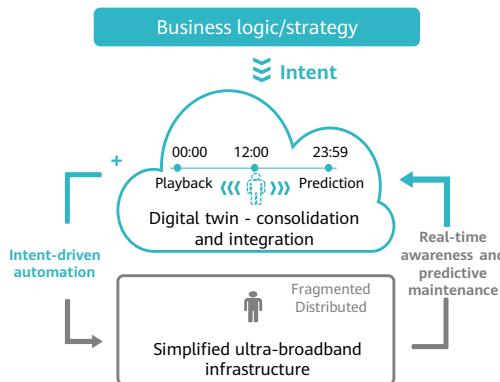
112 Huawei Confidential

DCN services are frequently changed and are manually verified, leading to low efficiency.

- Traditional O&M and verification methods:
 - Ping/Traceroute:
 - Unpredictable result: In SDN networking, it is hard to predict which gateway on a leaf node is pinged.
 - Incomplete path coverage: Not all ECMP paths can be covered. The ping test passes but service packets cannot be transmitted. In addition, it is difficult to traverse network-wide services within a limited change time window.
 - Packet mirroring:
 - High deployment cost: The mirroring mode needs to be enabled on each device on the entire network to implement full-flow mirroring analysis.



Intent Verification Overview

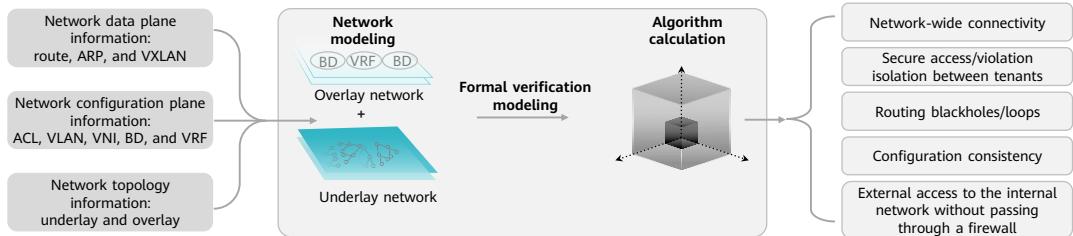


Passive and responsive maintenance -> Data-based AI-driven maintenance

- The concept of **Intent-Based Networking (IBN)** emphasizes the service-driven perspective. Intent is the input of users to the system, aiming to convert service intents into network configurations. As such, iMaster NCE-FabricInsight can obtain the network status through data collection and analysis during the entire running process, and perform closed-loop dynamic adjustment to ensure that the actual behavior of the system is consistent with the service intents.
- Data plane verification (DPV)** is used to verify IBN changes. By collecting network data after configuration changes, iMaster NCE-FabricInsight creates a model to check whether the actual network forwarding behavior is consistent with service intents. Based on the verification result, you can check whether the changes meet the expectation and causes issues. If an intent verification fails, you can locate the root cause, greatly improving the O&M efficiency in network change scenarios.

- DPV builds a model based on the data plane information on the DCN. The data plane information includes forwarding entries of network devices, such as routing forwarding entries, ARP entries, VXLAN tunnel connection relationships and status, VXLAN peer connection relationships and status, as well as physical link relationships and status on the underlay network. This information reflects the actual forwarding behavior on the DCN.
- When service configurations change on a network, data on the forwarding plane changes accordingly. As a result, service forwarding behavior is affected.
- This is where post-event verification comes in. This function is implemented based on the collection, modeling, and analysis of network data plane information, as well as service intents input by users.

Data Plane Simulation and Verification



Data collection

iMaster NCE-FabricInsight collects information about topologies, configurations, and forwarding entries of the current data center network at a high speed.

114 Huawei Confidential

Network modeling

iMaster NCE-FabricInsight models the underlay and overlay networks based on collected information of the live network. It also conducts intent calculations by transforming network models into transfer functions.

Intent verification

Solution models returned by the algorithm are displayed as the verification result and root cause of issues in terms of reachability, consistency, isolation, and existence, and are integrated with network health evaluations to notify users of the intent verification status in a timely manner.



Service Functions Supported by Intent Verification (1)

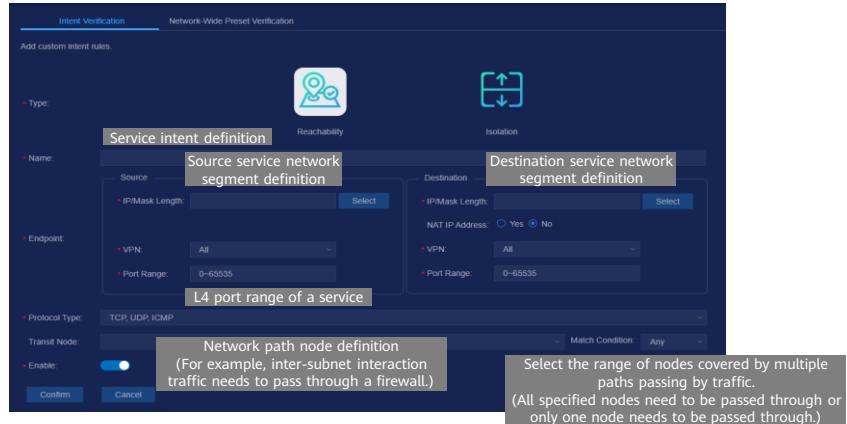
Intent Category	Intent Subcategory	Source
Reachability	[Overlay] East-west reachability verification within a PoD or across PoDs on the same subnet	Customized
	[Overlay] East-west reachability verification within a PoD or across PoDs within a VPC on different subnets	Customized
	[Overlay] East-west reachability verification within a PoD or across PoDs between different VPCs, without passing through a firewall	Customized
	[Overlay] East-west reachability verification within a PoD or across PoDs between different VPCs, passing through a firewall	Customized
	[Overlay] North-south reachability verification within a PoD or across PoDs: communication between IP addresses of hosts on a fabric and external IP addresses of a fabric	Customized
	[Underlay] Communication between IP addresses within a fabric or across fabrics	Customized
	[Underlay] Traffic forwarding according to underlay routes within a fabric, such as communication between BGP peers and between VTEPs of a VXLAN tunnel	Customized
	Constraint-based forwarding path passing through one node to N nodes	Customized
	Verification and display of ECMP reachability	Customized
	Verification of route reachability between BGP peers on the entire network	Preset
	Verification of route reachability between VTEPs of VXLAN tunnels on the entire network	Preset

Service Functions Supported by Intent Verification (2)

Intent Category	Intent Subcategory	Source
Isolation	Verification of whether two subnets (or IP addresses) are isolated from each other	Customized
Existence	Verification of whether routing loops occur on the network	Preset
	Verification of whether routing blackholes exist on the network	Preset
Consistency	Verification of whether interface configurations on both sides of a link are the same, including the maximum transmission unit (MTU) information, rate, duplex mode, auto-negotiation mode, working mode, VLAN ID, and IP subnet	Preset

Intent Verification: Service Assurance Intent Input

- Reachability and existence intents are preset in iMaster NCE-FabricInsight and user-defined intent rules are also supported.

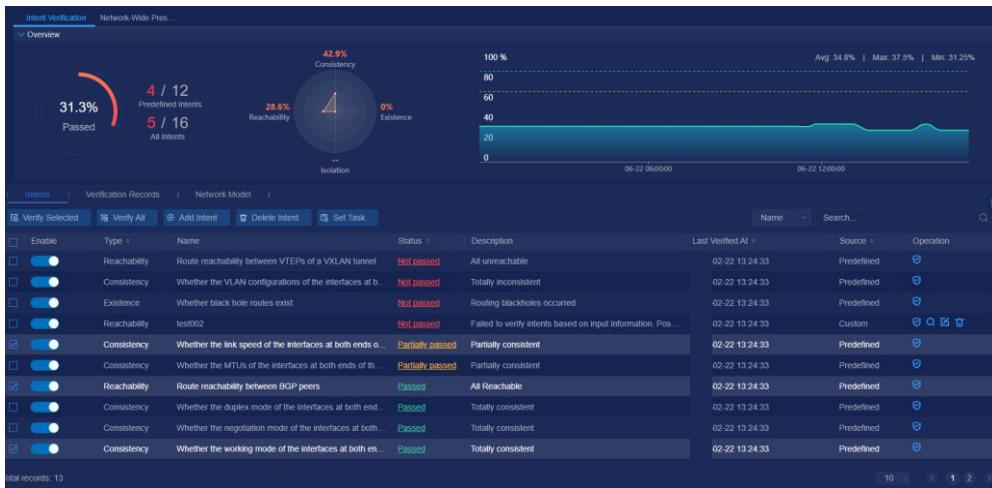


117 Huawei Confidential

HUAWEI

- Reachability intent: checks the network connectivity between source and destination IP addresses. On the rule creation page, you can specify verification rules such as the source IP network segment, destination IP network segment, protocol type, port number, and devices through which traffic passes. From the perspective of access direction, DPV can be used to verify east-west and north-south access traffic in a single fabric and across fabrics. From the perspective of the network plane, DPV can be used to verify underlay route reachability and overlay service reachability in a single fabric and across fabrics.
- Isolation intent: checks whether the source and destination IP addresses are isolated. Isolation intents are generally used for verifying the network policy compliance. For example, they can be used to check whether the security policies of firewalls are as expected. The page for creating an isolation intent rule is similar to that for creating a reachability intent rule, except that you do not need to set the transit node on the former page.

Intent Verification: Status Query



118 Huawei Confidential

HUAWEI

- Intent verification result overview: The overview area on the **Intent Verification Overview** tab page displays the intent pass rate, distribution, and trend. The intent pass rate distribution is displayed in terms of reachability, isolation, existence, and consistency. You can switch the time range to view the intent pass rate trend in a specified time range.
- User-defined reachability intent: In the **Intents** list, you can click a reachability intent verification result link to view the detailed verification result.

Quiz

1. (Multiple-answer question) Which of the following are covered by "1-3-5" troubleshooting? ()
 - A. Service
 - B. Device
 - C. Network
 - D. Interface
 - E. Protocol

1. ABCE

Summary

- Currently, with the rapid increase of services and traffic, it is a must-have to implement effective, flexible, and fast O&M. CloudFabric intelligent DCN O&M solution enables O&M engineers to implement O&M in an intelligent way rather than by themselves.
- This course describes the multi-dimensional, refined, and visualized O&M capabilities provided by iMaster NCE-Fabric, helping to solve the problems of mixed physical and virtual devices, blurred O&M boundaries of network and IT devices, and decoupling of physical and logical networks. Various intelligent O&M functions provided by iMaster NCE-FabricInsight are also introduced, including network visualization, network health evaluation, "1-3-5" troubleshooting, service flow analysis and troubleshooting, and intent verification, with an aim to solve problems such as traditional passive O&M and difficult fault locating, and provide ubiquitous application and network assurance.

Thank you.

把数字世界带入每个人、每个家庭。

每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

Copyright©2023 Huawei Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technologies, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

