Cloud Fabric Data Center Basic Network Solution Design — TRILL-based Implementation
Contents

- Data Center Network Design Requirements and Objectives
- Data Center Network Fabric Design
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
What Is Cloud Computing

Definition: Provide services for a large number of users (tenants) uniformly by consolidating, managing, and scheduling computing resources in a network.

I. Measured Service
- 8-Core CPU
- 100 GB disk
- 2 NICs
- 8GB database
- 200 Gbps traffic
- Computing performance

II. On-Demand Self Service
- Self-service, on-demand service subscription anytime

III. Resource Pooling

IV. Rapid Elasticity
Great Changes in IT Applications in Cloud Computing Era

Cloud computing core values
- Economies of scale
- Shorter service provisioning period
- Multi-tenant service model

Served objects and delivery mode
- Public cloud
- Private cloud
- Hybrid cloud
- Software as Service (SaaS)
- Platform as Service (PaaS)
- Infrastructure as Service (IaaS)

Cloud computing infrastructure
- Cloud security
- Cloud host
- Cloud storage
- Cloud computing
- Cloud management
Features of Cloud-Computing Data Center Networks

Cloud computing core values

- Economies of scale
- Shorter service provisioning period
- Multi-tenant service model

Cloud network key features

- Elastic architecture
  - Large-scale deployment of 10G/40G servers
  - Wide of collaborative computing, sharp increase of east-west traffic, frequent virtual machine migrations

- O&M automation
  - Service visualization
  - Dynamic service provisioning
  - Uniform management

- Resource virtualization
  - Dynamic resource allocation
  - Tenant network security and QoS guarantee

Cloud network key requirements
All physical & virtual resources visible
E2E automatic deployment and O&M
Open, interoperable multi-service platform, uniform management

Decoupled logical and physical structure
Dynamic allocation of network, security, and IT resources, on-demand isolation and communication between tenants
Hierarchical QoS guarantee for tenants
Elastic non-blocking architecture, large-scale resource access
Flexible access to resource points (subnets and value-added services), free from restriction of physical networks
Uniform IT resource management
Contents

- Data Center Network Design Requirements and Objectives
- Data Center Network Fabric Design
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - Security Design
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - Design References
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
Transparent Interconnection of Lots of Links (TRILL) is a link state routing protocol running on Layer 2 networks. It is implemented based on IS-IS extensions. Devices running the TRILL protocols are called routing bridges (RBs). RBs form a TRILL network.

**TRILL characteristics:**
- Loop free, high link bandwidth efficiency
- Support for equal-cost multipath (ECMP)
- Real-time entire network topology awareness, subsecond-level failover
- Automatic path discovery, simple deployment

**TRILL Overview**

**Encapsulation**
- Forward data according to outer nickname
- Outer MAC changes hop by hop

**Control**
- IS-IS as control plane
- SPF auto path discovery

**Forwarding**
- ECMP per flow load balancing
- RPF loop prevention in TRILL network

**MAC learning**
- Learning on data plane
- MAC sensing only on edge RBs

**Failover**
- Failover on control plane
- No MAC flooding

**Edge access**
- Active-standby access based on AF

**TRILL**
- Transit RB
- Edge RB
- Edge access
- Edge RB
- Data MA MB
- STP
- IS-IS as control plane
- SPF auto path discovery
- ECMP per flow load balancing
- RPF loop prevention in TRILL network
- Active-standby access based on AF
- Learning on data plane
- MAC sensing only on edge RBs
- Failover on control plane
- No MAC flooding
- Loop free, high link bandwidth efficiency
- Support for equal-cost multipath (ECMP)
- Real-time entire network topology awareness, subsecond-level failover
- Automatic path discovery, simple deployment
TRILL fabric Architecture

TRILL fabric increases the number of access nodes supported to 512, establishing a large-scale Layer 2 network.

Flattened network structure and ECMP realize non-blocking switching.

TRILL fabric makes the entire Layer 2 network reachable, allowing for flexible subnet deployment:

1. Subnet deployment is not restricted by physical topology. Traffic can be transmitted in the entire Layer 2 network regardless of physical zones. Virtual machines (VMs) can be migrated on demand.

2. Subnet scale can be flexibly expanded within a zone or across zones.

3. Data forwarding configuration only needs to be configured on edge RBs, and the TRILL fabric can automatically discover forwarding topologies, saving the need for hop-by-hop configuration. This makes subnet access much simpler.
TRILL fabric Design Points

- **Security design**
  - TRILL fabric gateway design
  - TRILL fabric multi-active gateway design: VRRP

- **Network reliability design**
  - Data center interconnection design: EVN/VPLS

- **TRILL fabric access design**:
  - Traditional Layer 2 network converged access
  - Active-active server access

- **Storage network convergence design**: FCOE, IP SAN
Contents

- Data Center Network Design Requirements and Objectives
- **Data Center Network Fabric Design**
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - Security Design
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - Design References
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
Gateway Design

TRILL packets are terminated on edge nodes and change into Ethernet packets, which are then forwarded to the IP network through the gateway devices. TRILL sets up a non-blocking fabric network on which all nodes are reachable at Layer 2. The gateway can be deployed in any position in the TRILL campus:

1. Deployed on TRILL backbone nodes (use backbone switches as gateway or deploy independent gateway devices). This is similar to traditional gateway deployment. It is currently the recommended deployment.

2. Deployed on TRILL leaf nodes to form a resource pool. This is an evolved fabric deployment. The spine nodes work as switching nodes and do not process any services. Gateway is deployed on leaf nodes as a service to create a resource pool. This deployment enables flexible gateway expansion without any change on spine nodes. However, this deployment differs a lot from traditional gateway deployment. It is used only as a reference solution now.

The gateway implements communication between data center subnets. The scale of a subnet determines the traffic volume and required device performance in the subnet. Select gateway device based on service distribution, scale, and interactions. Consider the following:

1. Whether the number of MAC address and ARP entries supported by the gateway device meet service requirements.
2. Whether gateway device performance meets service requirements. If not, more gateway devices can be deployed.
3. To enhance gateway reliability, gateway devices can run the Virtual Route Redundancy Protocol (VRRP) for gateway redundancy and load balancing.
Gateway Design Scheme 1: Configuring VSs on Switches (Recommended)

Each two TRILL backbone switches set up a cluster, and then the cluster is virtualized into two VSs, one as the TRILL backbone node and the other as the L2/L3 gateway. The four backbone devices are virtualized into two logical devices, which run VRRP to implement gateway redundancy and load balancing.

The cluster +VS gateway deployment further improves gateway reliability and reduces the number of managed nodes.

1. Clear managed topology, TRILL network separated from gateway devices
2. Flexible resource re-allocation using VSs
Gateway Design Scheme 2: Deploying Independent Gateway Devices

Routers provide large ARP table capacities and support a large number of access devices. This deployment applies to scenarios with many access hosts and requiring a large number of ARP entries.

Deploy independent routers or firewalls as gateway devices. The gateway devices run VRRP and exchange heartbeat packets through the TRILL network.

When four devices are used as gateway devices, you can configure them in one VRRP group to increase gateway capacity and redundancy. To fully use gateway capacities, you can configure two VRRP groups and add two gateway devices to each VRRP group. Distribute gateways equally on the two groups of devices based on service traffic in different VLANs.
Gateway Design Scheme 3: Integrating Gateway on TRILL RBs

After TRILL packets are converted into common Ethernet packets on edge RBs, the edge RBs use loopback interfaces as gateways to process the packets. This deployment can be used when gateway devices do not support VS (such as CE6800 and CE5800) and independent gateway devices cannot be deployed.

1. This deployment is flexible and has no special requirements for gateway devices. Any TRILL device can act as a gateway.
2. Fewer nodes need to be managed.
3. Gateway load balancing can be implemented using VRRP and VRRP+.

To prevent loops, configure VLAN mapping on the loopback inbound interfaces. For example, VLANs 1-100 must be mapped to VLANs 101-200.
VRRP implements gateway redundancy and load balancing. VRRP packets traverse the TRILL network over downlinks. VRRP supports only VLAN-based link load balancing.

To improve VRRP switchover performance, configure VRRP and BFD association between gateways 192.168.1.1 and 192.168.1.2.

The number of VRRP groups equals the number of gateways (VLANIF). A large number of gateways produce too many VRRP packets. Number of VRRP packets can be reduced by binding a management VRRP (mVRRP) to multiple service VRRP groups. Failover and notification are completed by mVRRP.

VRRP+ is an extension of VRRP. One VRRP+ group can contain four gateway devices, which are virtualized into one gateway for a VLAN for gateway redundancy. Traffic in this VLAN is equally distributed to the four gateways based on MAC addresses for load balancing.
Contents

- Data Center Network Design Requirements and Objectives
- **Data Center Network Fabric Design**
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - Security Design
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - Design References
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
Access Design

Server active-active access design:
To improve server access reliability and link efficiency, each server connects to two access devices through double active NICs. The two uplinks of a server set up an inter-device link aggregation group (LAG) to load balance traffic from the server.

Traditional Layer 2 access design:
The traditional Layer 2 network (xSTP) of the data center connects to the TRILL fabric. Measures need to be taken to prevent loops after Layer 2 traffic traverses the TRILL fabric network.

SVF access design:
Super virtual fabric (SVF) simplifies access layer management. Access switches set up an SVF system to connect to the TRILL fabric network.
Server Active-Active Access Design: E-Trunk (1)

Problems in active-active access:
A server connects to two devices using double active NICs (NIC teaming enabled, load balancing based on MAC addresses). Therefore, the two devices learn the same MAC address (MAC address flapping).

Note: MAC address flapping is considered an exception on network devices and can be prevented by changing MAC address priorities on interfaces. However, MAC address flapping is inevitable in VM migration scenarios. Besides, if one of interfaces is blocked because of MAC address flapping, the server cannot connect to the network in active-active mode.

E-Trunk solution:
**TRILL campus:** RBs in a TRILL campus forwards data traffic based on nicknames. Two access devices connecting to the same server are configured the same nickname and act as one logical forwarding node. Traffic is load balanced on link 1 and link 2.

**Ethernet network:** Inter-device E-Trunk links are established between the access devices. Member links synchronize MAC address entries and link states, and load balance traffic.

Topology:
**Peer-link:** The two access devices connecting to active-active server are connected using an Ethernet link. This link is used to transmit protocol packets, including but not limited to TRILL active-active packets, E-Trunk synchronization packets, MAC address synchronization packets, and multicast information synchronization packets, as well some data packets.

**Nickname:** The access devices connecting to the dual-homed server use a virtual nickname as the forwarding identifier, and the devices connecting to single-homed servers use physical nicknames.

(Note: If an access device is a CE12800 switch, it still uses a physical nickname as the forwarding identifier when it connects to a single-homed server. In this case, the forwarding model is similar to that in the dual-homing scenario. In the following pages, access devices connecting to single-homed servers use physical nicknames as forwarding identifiers.)
Server Active-Active Access Design: E-Trunk (2)

MAC address learning is disabled on the peer-link, to prevent the MAC address of a dual-homed server from being learned by both the access interface and the peer-link. Otherwise, MAC address flapping will occur. Therefore, RB2 and RB3 need to synchronize MAC addresses they learn.

**MAC address learning on devices connecting to dual-homed servers:**
RB-2 and RB-3 synchronize the MAC addresses learned on the dual-homing links, to ensure that they can find MAC address entries for packets sent from the network side, reducing broadcast of unicast packets.

**MAC address learning on devices connecting to single-homed servers:**
The MAC address entry learned from a single-homed server is synchronized to the peer RB. The source interface of the MAC address entry is the peer-link interface.
Server Active-Active Access Design: E-Trunk (3)

Devices connecting to single-homed servers use physical nicknames as forwarding identifiers in the TRILL campus. North-south unicast traffic is transmitted on path 1. East-west unicast traffic is transmitted on path 2 across the peer-link. MAC addresses of the single-homed servers are learned on peer-link interfaces.

Devices connected to dual-homed servers use a virtual nickname as the forwarding identifier in the TRILL campus. North-south unicast traffic is load balanced on path 1 and path 2. East-west unicast traffic is forwarded locally.
Multicast traffic from the network side to server side is processed in the same way in single-homing and dual-homing scenarios.

To prevent dual-homed servers from receiving the same multicast flow from the network side, RB-1 selects one path to send multicast packets to RB-2. RB-2 forwards multicast packets to MAC1, MAC3, and peer-link. Peer-link replicates and forwards the packets to MAC2. (Port isolation is configured on the peer-link interfaces, so multicast packets are not sent to MAC3 repeatedly or forwarded to the TRILL network.

When multicast traffic is sent from a server, for example, MAC1, RB-2 replicates the multicast packets and send the copy to MAC3, the peer-link, and the TRILL network. RB-1 does not forward the packets to RB-3. Peer-link replicates and forwards the packets to MAC2, and does not forward the packets to MAC3 or the TRILL network.

**Limitations:** Peer-link transmits multicast traffic of single-homed devices and east-west traffic. If many devices are single-homed, the peer-link may be overloaded. Therefore, this design is used in scenarios where most servers are dual homed.
Traditional Layer 2 Access Design

TRILL edge node stacking:
TRILL edge nodes are stack systems. For example, devices A and B set up a stack and act as one device. These stack systems are the boundary of the TRILL and MSTPs. Stacking and MSTP technologies prevent loops on the entire network.

When a link/device failure occurs, TRILL or MSTP can complete network topology convergence automatically.

TRILL edge nodes as xSTP root bridges:
Traditional Layer 2 networks connect to the TRILL campus and run xSTP to prevent loops. Switch C (TRILL edge node) is a root bridge. Loops are only prevented on the traditional Layer 2 networks, but new loops are created between the TRILL and xSTP networks, for example, path 1.

All TRILL edge nodes are stand-alone devices. After edge node D is configured as a virtual root node, all edge nodes participate in spanning tree calculation as root bridges. Another port is blocked (blocking point 2), and no loop exists on the entire network.

When a link/device failure occurs, TRILL or MSTP can complete network topology convergence automatically.

Appointed forwarder AF configured on edge node:
When downstream devices are dual-homed to the TRILL campus, you can set a higher priority for edge node H than edge node G. (Set priorities for the edge nodes based on their VLAN settings.) Another port is blocked (blocking point 3), and no loop exists on the entire network.

When device H or the link between I and H fails, priority of device G is raised, and G becomes the AF.
# Comparison of Traditional Access Designs

<table>
<thead>
<tr>
<th>Design</th>
<th>Applicable Scenario</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRILL edge node stacking</td>
<td>TRILL edge nodes are stack systems, each of which acts as one logical device in the TRILL and xSTP networks. TRILL and xSTP prevent loops in the respective networks.</td>
<td>Prevents loops between the Layer 2 network and TRILL network. MSTP and TRILL need to be associated to speed up topology convergence.</td>
</tr>
<tr>
<td>TRILL edge nodes as xSTP root bridges</td>
<td>TRILL edge nodes are stand-alone devices. Two edge nodes act as xSTP root bridges and have the same bridge MAC address and root priority.</td>
<td>Prevents loops between the Layer 2 network and TRILL network. Edge nodes must be able to act as xSTP root bridges. MSTP and TRILL need to be associated to speed up topology convergence.</td>
</tr>
<tr>
<td>AF configured on TRILL edge node</td>
<td>TRILL edge nodes are stand-alone devices. The network does not run xSTP and uses AF to prevent loops.</td>
<td>Prevents loops between the Layer 2 network and TRILL network. The topology convergence is slow, usually takes 15s to 30s.</td>
</tr>
</tbody>
</table>
SVF – a Technology to Simplify Access Layer Management

SVF, a simplified access switch management and O&M solution
Simple deployment: Only one node needs to be managed, saving the trouble of assigning a management IP address to each access device and configuring all the access devices one by one.
Low maintenance cost: All access devices can be deployed or upgraded using one software package, saving the workload on separate device upgrade and inspection.

SVF configuration:
1. Specify software package for the spine node and configure fabric ports (callout 1).
2. Leaf nodes automatically discover the spine node and start link negotiation. A super virtual fabric (SVF) system is set up after successful link negotiation.
3. Leaf nodes obtain configuration files to complete configuration.

SVF system upgrade:
1. Upgrade the standby device. Traffic is forwarded by the master device. The standby device becomes the master after the upgrade.
2. Access devices download software package and configuration files from the spine node to complete upgrade.
3. Upgrade the original master device and switch traffic to the new master device. The original master device joins the iStack/CSS system after the upgrade. Traffic interruption time is about 2s.
SVF System Access Design

Multiple access switches set up an SVF system and connect to the TRILL fabric network as one logical node.

SVF systems greatly increase the number of ports at the access layer, so SVF and TRILL technologies can be used together to establish a large Layer 2 network.

Each SVF system connects to the TRILL fabric network as one logical node (logically EOR deployment and physically TOR deployment). This simplifies cable connections.

The simple logical networking simplifies network management.

Servers can connect to an SVF system in active-active mode.

Note:

an SVF system can connect to a TRILL network only when the spine and leaf nodes all support TRILL.

This design is applicable only when the leaf nodes are CE6810 switches.
Contents

- Data Center Network Design Requirements and Objectives
- **Data Center Network Fabric Design**
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - **Reliability Design**
  - Security Design
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - Design References
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
A server is **dual-homed** to the TRILL network in active-standby mode. When the active device/link fails, traffic is automatically switched to the standby link.

TRILL gateway devices run VRRP or VRRP+ for gateway redundancy. If the active gateway device fails, traffic is automatically switched to the standby gateway.
Reliability Design: TRILL Link Protection

**Enhance link reliability between neighboring devices:**
Multiple physical links between neighboring devices can be bundled into a logical link to ensure uninterrupted traffic forwarding upon a single link failure. The link bundle implements load balancing while providing link redundancy.

**Enhance link reliability using ECMP:**
TRILL supports equal-cost multipath (ECMP). As shown in the figure on the left, four equal-cost paths are available from RB-5 to RB-10:
- RB-5->RB-1->RB-10
- RB-5->RB-2->RB-10
- RB-5->RB-3->RB-10
- RB-5->RB-4->RB-10

Once any path fails, traffic on this path is switched to the other three paths. This implements link redundancy and load balancing.

**Link switching time:**
- Switchover time upon a link failure: about 10 ms.
- Switchback time after fault recovery: about 200 ms
Contents

- Data Center Network Design Requirements and Objectives
- **Data Center Network Fabric Design**
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - **Security Design**
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - Design References
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
Security Deployment

**Perimeter security:** Deploy FW/VPN and anti-DDoS devices at the border between external and internal networks.

**TRILL fabric security:** Deploy FW/IPS and other high-performance security devices to form a security resource pool.

**Host/VM security:** Redirect traffic of VMs to physical firewall devices for traffic isolation. Deploy virtualized security suites (such as DeepSecurity) on servers to protect security of hosts and VMs.
Redirect and clean traffic only when attacks occur. The detection and cleaning devices are attached to the egress router in bypass mode. The splitters copy all traffic to the detection device. After the detection device detects malicious traffic, the management center dynamically delivers BGP routes to redirect malicious traffic to the cleaning device. After being cleaned, the traffic is re-injected to the egress router through static routing or policy-based routing.

Redirect and clean protected traffic no matter whether attacks occur. Only a cleaning device is attached to the egress router in bypass mode. Traffic is redirected to the cleaning device through policy-based routing or BGP routing. After being cleaned, the traffic is re-injected to the egress router through static routing or policy-based routing.
TRILL Fabric Security Design – Transparent FW Deployment

Scenario: Gateway for the network is deployed on VS1, and the firewalls are deployed between VS1 and VS2. The firewalls work in hot standby mode and run VRRP.

Characteristics:
1. Transparent firewall deployment is simple.
2. The network has a clear boundary between Layer 2 and Layer 3.
3. Even when firewalls both fail, traffic can still be forwarded in bypass mode. This design provides high network reliability.
**Scenario:** Gateway of the network is deployed on VS1, and the firewalls are attached to VS1 in bypass mode. The firewalls work in hot standby mode and run VRRP. When the firewalls fail, low-priority routes take effect to ensure uninterrupted service traffic forwarding. Inbound and outbound routes passing through the firewalls are isolated by VRF instances to prevent routing loops.

**Characteristics:**
1. Complex VRF configuration is required on the gateway.
2. Bypass firewall deployment is flexible and makes network expansion easy.

---

**Scenario:** Gateway of the network is deployed on VS1, and the firewalls are attached to VS1 in bypass mode. The firewalls work in hot standby mode and run VRRP. Traffic is redirected to firewalls through policy-based routing. Upon a firewall failure, policy-based routing fails and service traffic is forwarded normally without being redirected. If high security is required, you can configure the egress devices to drop packets after failure of the firewalls. Service will be interrupted in this case.

**Characteristics:**
1. Complex policy-based routing configuration is required on the gateway.
2. Bypass firewall deployment is flexible and makes network expansion easy.
Host/VM Security Design – Deploying Physical FW

**Scenario:**
1. A VM may need to communicate with other VMs in the same VLAN or in other VLANs.
2. To save computing resources on servers, all VM communication traffic is redirected to external physical firewalls for traffic filtering.
3. Gateway for VMs is deployed on TRILL switches and redirects VM traffic exchanged between VLANs to firewalls through policy-based routing.
4. For communication between VMs in the same VLAN, traffic in the server is redirected to the physical network: Traffic of different VMs is isolated by PVLAN (supported by VMware) enabled on vSwitches and MUX VLAN enabled on physical switches. Traffic exchanged between VMs is redirected to firewalls through policy-based routing.

**Characteristics:**
1. Access control is performed on the external network, reducing loads on servers.
2. Redirecting VM communication traffic to the external network makes it easier to monitor the traffic.

Physical firewall devices provide high performance and comprehensive functions, and can implement all-round VM traffic monitoring. Physical firewalls can be virtualized to isolate different services. Firewall virtualization realizes profound service isolation to enhance security and simplifies service planning.
Host/VM Security Design – Deploying Virtual Security Gateway

**Scenario:**
1. External firewalls can be used with virtual security gateways in the virtualized network to protect VM security.
2. A Trend Micro DeepSecurity virtual security gateway is deployed in each VMware ESX/ESXi host to provide FW/IPS/WAF and other security functions for VMs.
3. DeepSecurity is logically deployed in front of VMs to provide real-time security protection.

**Characteristics:**
1. VM communication traffic is filtered in the virtualized network without being redirected to the physical network. This improves traffic processing efficiency.
2. Trend Micro’s unified security management platform can configure and deliver required security functions for VMs, simplifying security management.
3. When a VM migrates, DeepSecurity can migrate security policies for the VM accordingly.

Virtual security gateway consumes server resources and cannot provide high performance.
Contents

- Data Center Network Design Requirements and Objectives

- Data Center Network Fabric Design
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - Security Design
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - Design References

- Data Center Network Intelligent O&M Design

- Data Center Network Multi-tenant Design
Service and Storage Network Convergence Design
— Overview

Service and storage network convergence enables service and storage devices to be deployed on the same network. Storage networks include direct attached storage (DAS) network, network attached storage (NAS) network, and storage area network (SAN). Network convergence applies only to SAN networks, which are Fibre Channel (FC) or IP based.

Convergence technology for FC-based SAN networks is Fibre Channel over Ethernet (FCoE), which applies to LANs. Convergence technology for IP-based SAN networks is IP SAN, which applies to LANs and WANs. The convergence technologies need to be used with QoS.

Network convergence trend:
Simplified network structure, LAN/SAN convergence, unified switching. Full use Ethernet's advantages of high-speed forwarding and high bandwidth. Lower investment, simpler cabling in data centers.
Service and Storage Network Convergence Design — FCoE

Networking:
1. Servers and storage devices are configured with double converged network adapters (CNAs), which connect to two switches to provide double planes.
2. Access switches connected to server work as FIP snooping bridges (FSBs), and FC SAN switches work as Fibre channel forwarders (FCF). Data center bridging (DCB) is configured between server - >FSB->FCF to ensure lossless forwarding of FC traffic over the Ethernet network.
3. If access switches set up a stack system, you can configure `fcoe dual-fabric enable` in the stack system to separate the double SAN planes.

Remarks:
1. This networking converges the FC SAN and Ethernet networks. If the FC SAN network has been deployed, customers do not need to configure CNAs in storage devices, and only need to connect FCF switches to the FC switches.
2. The converged network reduces cable connections, but also increases difficulties in locating service and storage faults. Consider the risks of the convergence solution when deciding to recommend this solution to customers.
3. Huawei CE series can be used as FSB switches, and Broadcadc VDX6730 switches can be used as FCF switches.
Service and Storage Network Convergence Design — IP SAN

IP SAN realizes complete convergence of storage and service networks. Only servers need to use host bus adapters (HBAs), and no other additional devices are needed on the converged network.

Networking:
1. Servers are configured with independent storage network adapters and connected to access switches through bundled uplinks.
2. Service and storage traffic and traffic of different storage devices are isolated by VLANs.
3. Switches have DCB enabled and use priority flow control (PFC) and enhanced transmission selection (ETS) to ensure bandwidth for storage traffic.

Remarks:
NICs of servers can be used exclusively for storage or shared by storage and service traffic.
Priorities are configured on the network boarder and mapped between different networks.

1. Network boarder devices: identify services (storage, VM migration, streaming media, and data services) or trust service priorities. They map all data traffic (including storage traffic) to eight priorities.

   If the service systems (server VMs, physical servers or vSwitches, NAS & ISCSI storage arrays, backup media servers & tape libraries) can set 802.1p priorities for packets, configure access switches to trust 802.1p priorities.

   If service systems cannot set 802.1p priorities for packets, access switch can identify packet types and set 802.1p priorities for packets.

2. Priority mapping is performed when packets are transmitted between traditional Layer 2 network and the TRILL network.

3. When packets are transmitted from across network segments (L2 to L3), their 80.21p priorities are mapped to DSCP values.
Contents

- Data Center Network Design Requirements and Objectives
- **Data Center Network Fabric Design**
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - Security Design
  - Service and Storage Network Convergence Design
  - **Data Center Interconnection Design**
  - Design References
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
DC Interconnection Requirements and Scenarios

Active-active DC scenario:
Service access through two active DCs
Flexible service resource scheduling
Inter-DC VM scheduling (virtual resource pooling)
Inter-DC app cluster (such as RAC)

Active-active DC key requirements:
L2 interconnection between DCs
L2 domain isolation between DCs
Higher bandwidth utilization on interconnection links
Simplified deployment

Active-active DC interconnection solutions:
Bare fiber connection on Ethernet network
IVPLS over GRE on IP network
EVN on IP network (recommended)
L2 DC Interconnection over an OTN Network

Application scenario:
1. DCs in the same city, less than 100 km from each other, connected using leased bare fibers
2. DCs connected using enterprise's own optical transport network (OTN)

Deployment:
1. Core network gateway devices are directly connected to OTN devices.
2. Four core switches are connected in full mesh or square topology.
3. TRILL RBs set up inter-chassis Eth-Trunk with active-active member links to prevent loops on the entire network.
4. Traffic is transmitted between DCs over the Ethernet network, without a need for overlay.

Characteristics:
1. Active-active interconnection link, high link utilization
2. Interconnection without overlay, lower latency
3. 1+1 backup to ensure highly reliable transmission; failover on transport layer undetectable to upper layer
4. Independent TRILL networks in the two DCs for fault isolation
L2 DC Interconnection Through VPLS over GRE

Application scenario:
1. L2 transmission required between DCs
2. Inter-DC transmission over an IP network, no need to deploy MPLS on the entire network

Deployment:
1. VPLS over GRE is deployed for DC interconnection over an IP network.
2. CE12800 switches function as PE devices and set up stacks to prevent loop on the entire network.
3. One GRE tunnel is deployed as outer tunnel.
4. One PW tunnel is deployed as inner tunnel to implement L2 transmission through VPLS.
5. Broadcast and unknown unicast traffic suppression is configured on DC interconnection devices.

Characteristics:
1. No need to deploy MPLS on the backbone network
2. Independent TRILL networks in the two DCs for fault isolation

Product support: CE12800
Next-Generation L2 Interconnection Technology – EVN

**Ethernet Virtual Network (EVN):** Optimization on the control, data, and management planes

**Control plane:** support for larger L2 network
1. Over 1M MAC address entries supported, higher L2 scalability
2. Broadcast reduction (ARP flooding suppression by selectively flooding ARP packets with unknown addresses) to improve WAN link bandwidth efficiency

**Data plane:** fine-grained load balancing
1. UDP encapsulation to facilitate load balancing
2. Support for 16M tenants
3. Distributed gateway

**Management plane:** simpler deployment and O&M
1. Unaware of BGP configuration, no need to configure AS numbers
2. Automatic discovery of VPN sites and tunnels
L2 DC Interconnection Through EVN

Application scenario:
1. L2 transmission required between DCs
2. Inter-DC transmission over an IP network (no need to deploy MPLS on the entire network), simple deployment and O&M
3. Control plane optimized to improve WAN link bandwidth efficiency, more suitable for scenarios with insufficient WAN link bandwidth (private MSTP lines or leased MPLS VPN links)

Deployment:
1. EVN deployed over an IP network (reachable routes available on the IP network)
2. EVN PE deployed on gateway VS1, isolated from RB VS2
3. VLAN access: 1:1 binding of EVN instances (EVIs) and VLANs

Characteristics:
1. EVN overlaid on IP, no need for VPLS tunnel
2. MAC address synchronization between DCs through the EVN control layer
3. L2 broadcast domain isolation between DCs, greatly reducing broadcast and unknown unicast traffic
4. Flexible multi-homing access
EVN and TRILL Fabric Interconnection

**Deployment:** Two EVN PE devices in the same site are configured the same Ethernet segment identifier (ESI). Packets of VLANs with odd and even VLAN IDs are transmitted by different PE devices. The PE forwarding a VLAN's traffic is the designated forwarder (DF) of the VLAN, and the non-DF PE blocks traffic of this VLAN.

1. TRILL and EVN are both L2 overlay technologies, so loops may occur over the tunnels when TRILL and EVN networks are connected. PE A is the DF for VLAN100, and PE B blocks traffic of VLAN 100. Packet loops are prevented in this way.
2. Non-DF devices for a VLAN block traffic of the VLAN at the sender and receiver ends to prevent packet loops, without a need for loop detection packets.
3. Non-DF devices block VLAN traffic to reduce broadcast traffic.

**Characteristics:**
1. PE devices are selected as DFs for VLANs based on VLAN ID parity to implement load balancing among VLANs.
2. Non-DF devices for a VLAN block traffic of the VLAN at the sender and receiver ends to prevent packet loops, without a need for loop detection packets.
3. Non-DF devices block VLAN traffic to reduce broadcast traffic.
## L2 Interconnection Solution Comparison and Selection

<table>
<thead>
<tr>
<th></th>
<th>Traditional L2 Interconnection</th>
<th>VPLS over GRE</th>
<th>EVN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Applicable bearer network</strong></td>
<td>Bare fiber + OTN</td>
<td>Private MSTP/IP line</td>
<td>Private MSTP/IP line</td>
</tr>
<tr>
<td><strong>WAN traffic optimization</strong></td>
<td>Rate limiting for unknown unicast and ARP broadcast traffic</td>
<td>Rate limiting for unknown unicast and ARP broadcast traffic</td>
<td>Unknown unicast traffic blocking, local ARP proxy</td>
</tr>
<tr>
<td><strong>MAC address learning between DCs</strong></td>
<td>At data layer</td>
<td>At data layer</td>
<td>At control layer</td>
</tr>
<tr>
<td><strong>CE multi-homing</strong></td>
<td>Use inter-chassis Eth-Trunk links between PE and CE devices to load balance traffic and prevent loops</td>
<td>Use inter-chassis Eth-Trunk links between PE and CE devices to load balance traffic and prevent loops</td>
<td>Use EVN DFs to load balance traffic and prevent loops</td>
</tr>
<tr>
<td><strong>Application scenario</strong></td>
<td>Two centers in the same city, with bare fibers available</td>
<td>Two centers in different cities, requiring open standards and protocols</td>
<td>Multiple data centers in different cities, requiring WAN link bandwidth optimization</td>
</tr>
</tbody>
</table>

**Note:**
EVN separates the control plane from the forwarding plane. It prevents broadcast flooding on the IP backbone and avoids the need to deploy a spanning tree across the backbone, greatly improving stability of the entire network. EVN is deployed over an IP network and does not need dedicated physical links, which are required in the traditional L2 interconnection solution. The EVN solution is also easier to configure and maintain than the VPLS and VPLS over GRE solutions. Therefore, EVN is the recommended solution for L2 interconnection between DCs.
Contents

- Data Center Network Design Requirements and Objectives
- **Data Center Network Fabric Design**
  - TRILL fabric Architecture
  - Gateway Design
  - Access Design
  - Reliability Design
  - Security Design
  - Service and Storage Network Convergence Design
  - Data Center Interconnection Design
  - **Design References**
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
TRILL Fabric Scale Design Considerations

**Network scale**
1. Total number of VMs/hosts
   - Determine the computing scale and capability
2. Total number of access ports for VMs/hosts
   - Determine sizes of MAC table and ARP table
3. Fault domain size
   - Determine network reliability and availability

**Device performance**
1. MAC table size
   - Determine the L2 domain scale
2. ARP table size
   - Determine the gateway performance
3. Oversubscription
   - Determine uplink and downlink port density and throughput
4. Number of ECMP links or Eth-Trunk member links
   - Determine physical connection design and affect total interconnection link bandwidth

Gateway’s MAC and ARP table size > number IP addresses
(Number of IP addresses is calculated based on number of vNICs and active NICs on physical hosts)
Small-sized DC Networking and Network Scale Estimation

Application scenario:
Small-sized new DCs of enterprises require cloud computing, resource pooling, and capability of resource expansion for long-term evolution.

Solution design:
TRILL deployed on the entire network, L2/L3 gateway deployed on core switches
Low oversubscription ratio or non-blocking design on the entire network

Network design:
1. 240 physical servers: 2*10GE NICs/server
2. 4800 VMs (1:20 virtualization): 1 vNIC/VM

L2 network architecture:
1. Access layer: 10*CE6850
   480*10GE access ports
   40*40GE uplink ports (3:1)
2. Core layer: 2*CE12804

Number of entries:
1. ARP: 4k (CE12800)
2. MAC: 4k (CE12800, CE6800)
Medium-sized DC Networking and Network Scale Estimation

Application scenario:
Medium-sized DCs are generally built by medium enterprises or medium/small ISPs. They require cloud computing, resource pooling, and capability of resource expansion for long-term evolution.

Solution design:
TRILL deployed on the entire network, L2/L3 gateway deployed on core switches
Low oversubscription ratio or non-blocking design on the entire network

Network design:
1. 1920 physical servers: 2*10GE NICs/server
2. 38400 VMs (1:20 virtualization): 1 vNIC/VM

L2 network architecture:
1. Access layer: 40*CE6850
   1920*10GE access ports
   160*40GE uplink ports (3:1)
2. Core layer: 2*CE12804

Number of entries:
1. ARP: 16k (CE12800)
2. MAC: 16k (CE12800, CE6800)
Large-sized DC Networking (TRILL Fabric in L2 Service Areas)

Application scenario:
DCs of large enterprises or ISPs have specific requirements on cloud computing and resource pooling. Such a DC consists of multiple large service areas (PODs), among which VM migration and communication are required.

Solution design:
TRILL is deployed in each service area. The core layer runs IGP for L3 communication and has EVN deployed to implement L2 interconnection over the L3 network.

Network design:
1. 19200 physical servers: 2*10GE NICs/server
2. 384000 VMs (1:20 virtualization): 1 vNIC/VM

L2 network architecture (per POD)
1. Access layer: 40*CE6850
   1920*10GE access ports
   160*40GE uplink ports (3:1)
2. Aggregation layer: 2*CE12808
   32*40GE uplink ports (5:1)
3. Core layer: 2*CE12812

Number of entries at the aggregation layer:
1. ARP: 16k (CE12800)
2. MAC: 16k (CE12800, CE6800)
Large-sized DC Networking (TRILL Fabric for L2 Service Area Expansion)

Application scenario:
In a traditional large-sized DC, service areas of different security levels are isolated. As enterprise business develops rapidly, the originally isolated areas require L2 interconnection.

Solution design:
1. Link type between core and aggregation layers is set to hybrid, so that the links support both IP and TRILL transmission.
2. North-source traffic and inter-VLAN traffic are transmitted between core and aggregation layers over through IP routing.
3. Intra-VLAN traffic is transmitted through TRILL, but user VLANs do not need to be configured on the core switches. TRILL leverages the IS-IS mechanism to implement multipath load balancing and support smooth network expansion.

Network design:
The network of a large enterprise is usually large and all traffic exchanged between service areas are forwarded by the core switches; therefore, the bandwidth oversubscription of the network should be low. It is recommended to use CE12808 switches at the aggregation layer and CE12812 switches at the core layer.
Contents

- Data Center Network Design Requirements and Objectives
- Data Center Network Fabric Design
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
Intelligent DC Network O&M: Automatic, Visualized, Open and Collaborative

Physical network automation: zero touch provisioning (ZTP)
Logical network automation: TRILL networking, automatic service provisioning with VM migration
Integrated with the cloud computing platform through REST over HTTP on northbound interfaces
Support third-party Apps on the open programmability system (OPS)

Automatic
Visualized
Open & collaborative

Provide visualized traffic analysis through sFlow and NetStream functions
Physical Network Automation: Zero Touch Provisioning

Solution:
1. Aggregation switches work as DHCP servers and have DHCP address pools, Option 67, and IP address of the file server (NMS server). The gateway is deployed on the core switches.

2. Unconfigured access switches automatically download and parse .ini or .py file after startup. After obtaining names of the software package and configuration file from the .ini or .py file, access switches download the files from the file server, and then start with the files.

3. When an access is replaced, the NMS can automatically discover the associated software version and configuration based on the location and model of the replaced device.

Value:
Automatic device deployment, plug-and-play
Logical Network Automation: E2E Automatic Service Provisioning on TRILL Network

Solution:
1. Access switches connect to the TRILL network as RBs.
2. CE VLANs of RBs in the TRILL campus are set to service VLANs of VMs.
3. When a VM migrates between servers connected to different servers, the destination RB automatically sets the CE VLAN to the service VLAN of the VM, enables TRILL on the destination interface, and sets the link type of the interface to access.

Value:
CE VLANs are automatically configured, saving the workload on manual service VLAN configuration.
Visualized Traffic Analysis: NetStream

Solution:
1. NetStream is enabled in the outbound direction on the outbound interfaces of core switches (connected to egress routers), or enabled in the inbound direction of the inbound interfaces of aggregation switches.
2. Core/Aggregation switches sample traffic for analysis based on NetStream original flow (V5), aggregation flow (V8), or flexible flow (V9) such as the protocol port number and flexible flow statistics template.
3. eSight works as the NetStream Collector and Analyzer and shows the traffic analysis results in charts and tables, to help in network optimization.

Value:
North-south and east-west IPv4 traffic is shown in charts and tables, enabling administrators to know details about traffic and discover abnormal traffic.

Remarks:
All the CE series products support the NetStream feature.

The traffic volume within a data center is huge. It is recommended to deploy NetStream at the DC egress.
Visualized Traffic Analysis: sFlow

Solution:
1. Network devices that need to collect traffic statistics are configured as sFlow Agent and the NMS server (eSight) functions as the sFlow Collector.
2. The sFlow sampling mode is set to flow sampling (based on data packets) or counter sampling (based on time period) in appropriate direction on corresponding interfaces.
3. eSight shows the traffic analysis results in charts and tables, to help in network optimization.

Value:
Sampled flow (sFlow) is a standard network traffic analysis protocol.
North-south and east-west traffic is shown in charts and tables, enabling administrators to know details about traffic and discover abnormal traffic.

Remarks:
All the CE series products support the sFlow feature.

The traffic volume within a data center is huge, but the current traffic sampling performance of data center devices is not high enough to handle the traffic. It is recommended to deploy sFlow at the DC egress.
Open and Collaborative Management Platform

Solution:
1. **Seamless connection with cloud management platforms:** nCenter and vCenter can connect to mainstream cloud management platforms through open northbound API – REST. (Cloud management platforms all provide standard REST interfaces.)

2. **Management of multi-vendor network devices:** nCenter can manage network devices of different vendors using protocols such as NetConf and SNMP.

3. **Virtual network device management:** vCenter can manage VMware vSwitches.

Value:
- Seamless integration with mainstream cloud management platform through REST interface
- All-round management from VMs to network devices
- E2E automatic service provisioning during VM migration
Contents

- Data Center Network Design Requirements and Objectives
- Data Center Network Fabric Design
- Data Center Network Intelligent O&M Design
- Data Center Network Multi-tenant Design
Public cloud data center lease computing resources to customers. Enterprise data centers need to allocate and adjust computing resources based on services. These are the drive forces for multi-tenant data center construction.

Tenants of a cloud data center shares hardware systems or program components to deploy their own services. Virtual resources of tenants are independent and isolated.
Multi-tenant Model and Network Requirements

Large and medium enterprises may use many computing resources to support different services. Different tenants or services within the same enterprise must be isolated, and security policies must be deployed for tenants and services.

Small enterprises and individual users only lease several or even one VM to support their services. The data center network must implement isolation and communication between VMs of a large number of individual users (much more than 4K) according to their requirements.

To adapt to dynamic service adjustment of tenants, VMs of the same service or tenant must be able to migrate in the same data center or across data centers.

Network service quality must be ensured according to requirements of different tenants and services.

Resources are dynamically created for tenants according to their service requests. Tenants can apply for or cancel use of computing resources anytime, anywhere. Data center resources must be dynamically deployed to meet requirements of tenants.

Tenants can determine the number of VMs they release depending on their service requirements.
Each tenant/service has an independent virtual network and security devices.

1. Each tenant network maps to a VRF instance.
2. Each VRF instance has one or multiple subnets, and each subnet is assigned a VLAN. VMs of different tenants and services of the same tenant belong to different VLANs.
3. A virtual firewall (vFW) is assigned to each tenant/VRF instance.

Tenant networks are isolated, and security policies of tenants are deployed on their vFWs.

Each service of a tenant is assigned a VLAN for L2 isolation. L3 isolation and communication are implemented using security policies on vFWs.

Tenants connect to the data center through MPLS VPN, IPSec VPN/IP, or other access methods.

Restricted by the number of VLANs (4K), a data center supports a maximum of 4K tenants or services.
VM traffic isolation within a VLAN through MUX VLAN

1. Enable private VLAN (PVLAN) vSwitches and add all VMs to an isolated VLAN.
2. Enable MUX VLAN on access switches. Set the link type of their uplink and downlink interfaces to trunk and configure the interfaces to allow the principal VLAN and subordinate VLANs.
3. Enable MUX VLAN on the core switches, and create the VLANIF interface in the principal VLAN. The VLANIF interface functions as the gateway to reply to all ARP requests from hosts in the corresponding subnets.
4. Deploy a vFW for each VLAN to isolate traffic of different hosts in the same VLAN.

In this deployment, all traffic transmitted between VMs pass through the firewalls. The firewalls implement fine-grained access control for VMs.
Tenant VM Migration Solution

When services of a tenant change, some VMs may need to migrate within the L2 domain.

1. TRILL and EVN are deployed to build a large L2 transparent network.
2. The network has two management platforms: vCenter (VM management platform) and nCenter (network management platform).
3. vCenter detects a need for VM migration and sends a VM migration notification to nCenter.
4. nCenter delivers the network configuration and policies for the VM to the destination network device.
5. The VM migrates to the destination network device. Then vCenter sends a migration-complete notification to nCenter.
6. nCenter deletes the configuration and policies for the VM from the original network device.
Differentiated network services can be provided based on types of tenant and services. Egress of a data center is a bandwidth bottleneck, and classes of services are marked on access switches. Therefore, the QoS control points for tenant networks are located on access switches and egress routers.

1. Different services for different department of a tenant are marked different priorities on access switches.
2. Access switches map priorities of CE VLANs to priorities of carrier VLANs.
3. Egress routers perform hierarchical QoS scheduling (different service flows -> VLANs (assigned based on subnets) -> VRF-based rate limiting -> traffic classification and queuing -> queue scheduling)

Map network quality to service level for tenants:
Services for a department of a tenant (high priority for delay sensitive services such as voice service) -> different departments of a tenant (high priority for key departments) -> traffic shaping for different tenants (drop excess traffic of a tenant when committed bandwidth is exceeded) -> queuing based on tenants’ service priorities -> queue scheduling based on queue priorities
Thank you

www.huawei.com