



## Recommendations

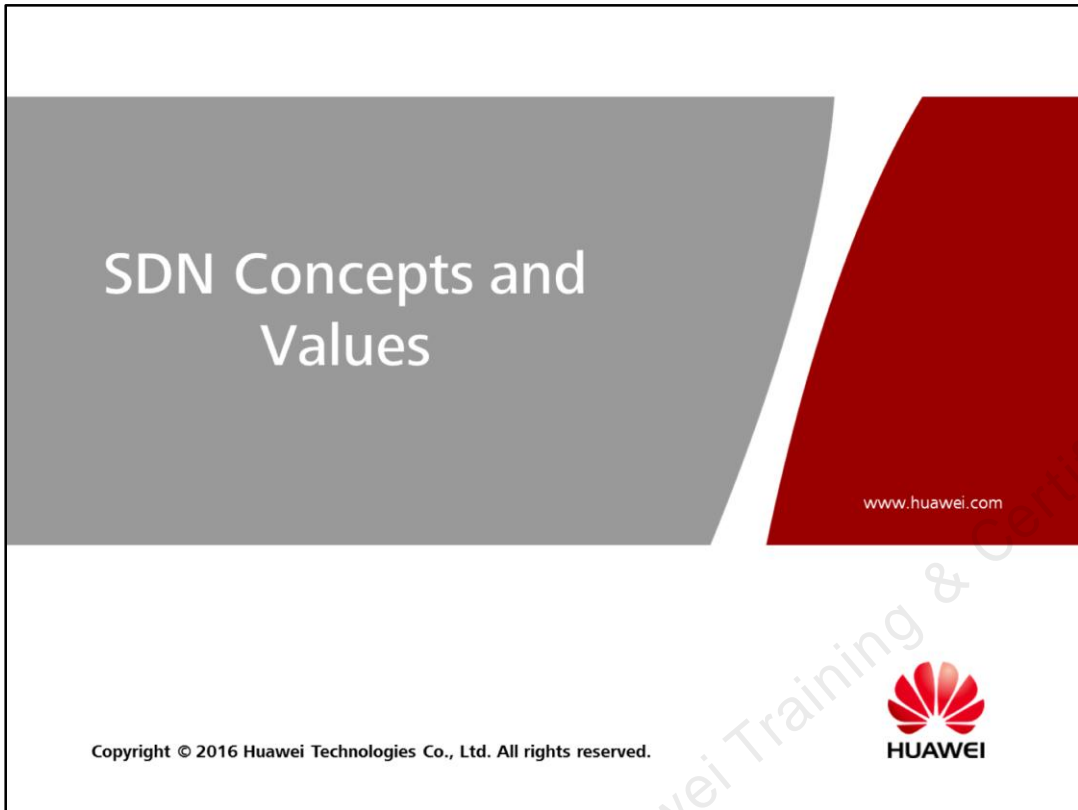
- Huawei Learning Website
  - <http://learning.huawei.com/en>
- Huawei e-Learning
  - <https://ilearningx.huawei.com/portal/#/portal/ebg/51>
- Huawei Certification
  - [http://support.huawei.com/learning/NavigationAction!createNavi?navId=\\_31&lang=en](http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=en)
- Find Training
  - [http://support.huawei.com/learning/NavigationAction!createNavi?navId=\\_trainingsearch&lang=en](http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=en)



## More Information

- Huawei learning APP






SDN Concepts and Values

[www.huawei.com](http://www.huawei.com)

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.



HUAWEI

Huawei Training & Certification





## Objectives

- Upon completion of this course, you will be able to:
  - Understand the limitations on the conventional network and how SDN network can help in these limitations.
  - Understand what is SDN and history timeline of SDN
  - Understand SDN architecture and basic working principles.
  - Understand SDN values from technical and operator perspectives.
  - Understand SDN challenges and the proposed solutions
  - Understand SDN related organizations and
  - Understand differences between SDN and NFV
  - Understand SDN influences in telecommunication networks.



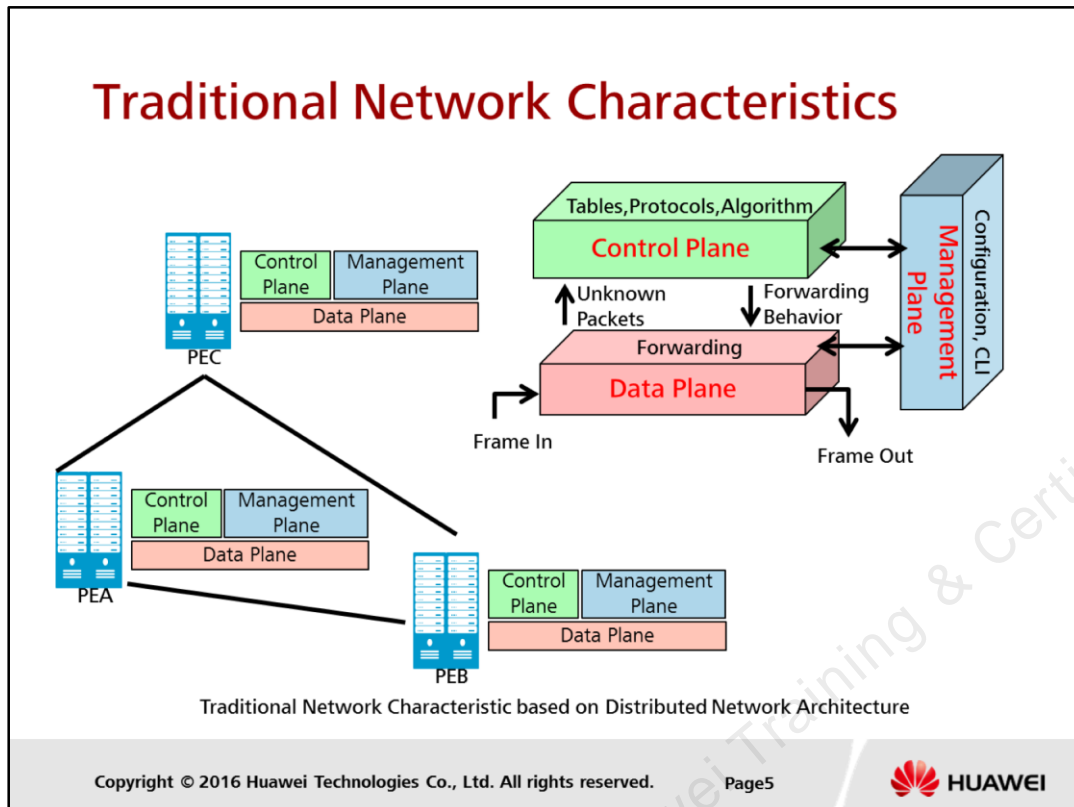
## Contents

1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. SDN Value Proposition
5. SDN Challenges and Solutions
6. SDN Related Concepts and Organizations
7. SDN Influences to Current Telecom Network

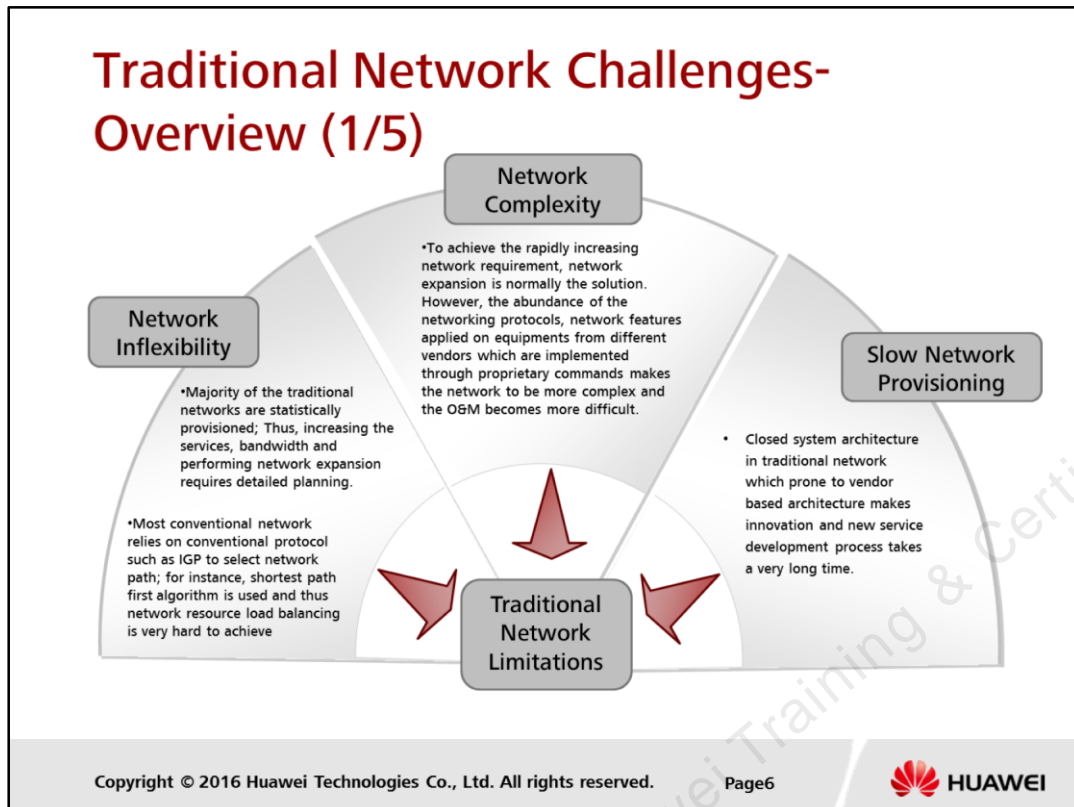


## Contents

- 1. Traditional Network Limitations**
- 2. SDN Overview and History**
- 3. SDN Network Architecture**
- 4. SDN Value Proposition**
- 5. SDN Challenges and Solutions**
- 6. SDN Related Concepts and Organizations**
- 7. SDN Influences to Current Telecom Network**



- Due to reliability and high availability, the current network are based on distributed networking approach, where every device is configured independently; each of them performs calculation independently and administratively.
- For example, imagine a traffic enter from PEA and exit at PEB. When PEA receive the traffic, PEA will check according to routing table. Based on routing table, it decides that in order to reach PEB, it have to go through PEC as next hop. Then PEA forward the traffic to PEC. PEC does exactly the same thing as PEA; check routing table, finding the next hop and forward to PEB. This forwarding manner called per-hop forwarding. Information inside routing table is collected and built through static routing or dynamic routing. In large scale network, normally routers are running dynamic routing protocol such as OSPF and ISIS. Every router in the routing domain collect link state information and then perform independently by using the same routing algorithm, for example, Shortest Path First Algorithm to find the shortest route. This networking called as distributed networking.
- Traditionally, control plane and data plane reside in the same physical hardware.
  - The internal architecture of a network devices has three planes of operation:-
    - **Management Plane** handles external user interaction and administrative tasks like authentication, logging, and configuration via a Web interface or CLI
    - **Control Plane** performs the internal device operations, provides the instructions used by the silicon engines to direct the packets; it also runs the routing and switching protocols and feeds operational data back to the management plane.
    - **Data Plane** is the engine room that moves packets through the device, using the forwarding table supplied by the control plane to determine the output port.



- Before we go into the details of SDN basics, we need to understand first how has the network been existing currently before SDN deployment. As per introduced in the previous slide, traditional network is working based on distributed networking with 3 differentiated planes including control plane, data plane and management plane. Every device contains its own control plane and forwarding plane and the network service management might be unified in NMS in management plane; for instance, if there are some changes in the network, the status changes or updates will automatically delivered and flooded within the network, and each router will perform recalculation based on new status information, update routing table and forwarding table at the end. This kind of networking method has been used in the current network for more than 30 years. However, this networking method is proven to have significant limitations that must be overcome in order to meet IT requirements and the rapidly growing network requirements nowadays.
- Diagram above generally concludes the 3 main challenges and limitations faced by conventional network nowadays, listed as:-
  - Network inflexibility
  - Network complexity
  - Slow network provisioning and innovation

## Traditional Network Challenges- Network Inflexibility (2/5)

**Issue:** The link between router A and router B is the shortest path and will face congestion soon. The other links are idle.

**Discussion:** Why is it impossible to transfer some traffic over link A-C-B?

**Service Requirement**  
 1:A->E 6G, 2: C->G 4G, 3: C->D 8G.  
 All link bandwidth is 10G. Service is built based on sequence.  
 Issue Link 3 failed to be established.

**Discussion:** Why a global calculation method cannot be used to ensure the establishment of all links?

\*Used BW/Total BW

\*each link =10G

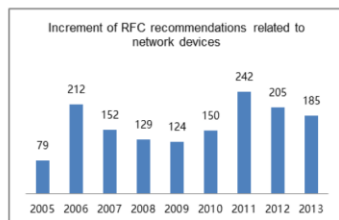
Network Inflexibility due to shortest path first calculation mechanism

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page7

- For the first part of the figure, based on shortest path first algorithm, traffic from Router A to Router B will take link between A-B. At the moment, link bandwidth usage has been occupied around 6G over 5G links, in which are link bandwidth usage are over-utilized, whereas link A-C-B is under-utilized. Even though MPLS TE has been introduced to solve under-utilized link usage, but most of MPLS TE planning are pre-configured and does not solve real-time or sudden burst traffic, hence causing network congestion.
- For second part of figure, there are some certain requirement. For example, there are 3 tunnel to be establish in sequence. First tunnel requirement needs 6G from A to E, hence after calculation or explicit configuration, it might be take A-B-C-D-E path. Followed by second tunnel with requirement 4G from C to G, eventually take the path C-B-A-F-G. For the third tunnel with requirement 8G from C to D. As a result, due to insufficient bandwidth, it fails to be established.

## Traditional Network Challenges- Network Complexity (3/5)

- If you want to become an expert in IP field, you must read 2500 RFC documents. You need more than 6 years to finish reading all the documents even if you read one every day. However, these documents are only 1/3 of the total RFC recommendations. In addition, the number of RFC recommendations is still increasing.
- If you want to skillfully operate the devices of a vendor, you must master more than 10000 commands. The number of commands available on each device is still increasing.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page8



- Traditional distributed networking approach causes many control plane protocols to be deployed and configured on a devices, including IGP protocol, BGP protocol, MPLS protocol, ipv6 protocol ,etc.
- IETF has produced thousands of protocol standardization to describe various network protocol features and the numbers of standardizations are still increasing from time to time when there are new features and network functions being developed and implemented.
- This makes a networking engineer has to learn complicated technology and need to master certain knowledge in order to perform network operation and maintenance.
- On the other hands, some vendors may deploy private proprietary protocol in operator network, causing a further difficulty in operation and maintenance. Difference vendor networking devices such as Huawei, CISCO, juniper provide different type of GUI, causing networking engineer has to learn multi-vendor on how to operate the networking devices.

## Traditional Network Challenges- Network Complexity (4/5)

Enterprise L3VPN service Provisioning

- Configure IGP
- Configure MPLS
- Configure VPN instance
- Configure PE-CE protocol
- Configure MP-BGP

**Minimum 50 CLI on each device**

```

1 Configure IGP
2 Configure IGP
3 Configure MPLS
4 Configure PE-CE part
5 Configure BGP
6 Configure BGP
7 Configure BGP
8 Configure BGP
9 Configure BGP
10 Configure BGP
11 Configure BGP
12 Configure BGP
13 Configure BGP
14 Configure BGP
15 Configure BGP
16 Configure BGP
17 Configure BGP
18 Configure BGP
19 Configure BGP
20 Configure BGP
21 Configure BGP
22 Configure BGP
23 Configure BGP
24 Configure BGP
25 Configure BGP
26 Configure BGP
27 Configure BGP
28 Configure BGP
29 Configure BGP
30 Configure BGP
31 Configure BGP
32 Configure BGP
33 Configure BGP
34 Configure BGP
35 Configure BGP
36 Configure BGP
37 Configure BGP
38 Configure BGP
39 Configure BGP
40 Configure BGP
41 Configure BGP
42 Configure BGP
43 Configure BGP
44 Configure BGP
45 Configure BGP
46 Configure BGP
47 Configure BGP
48 Configure BGP
49 Configure BGP
50 Configure BGP

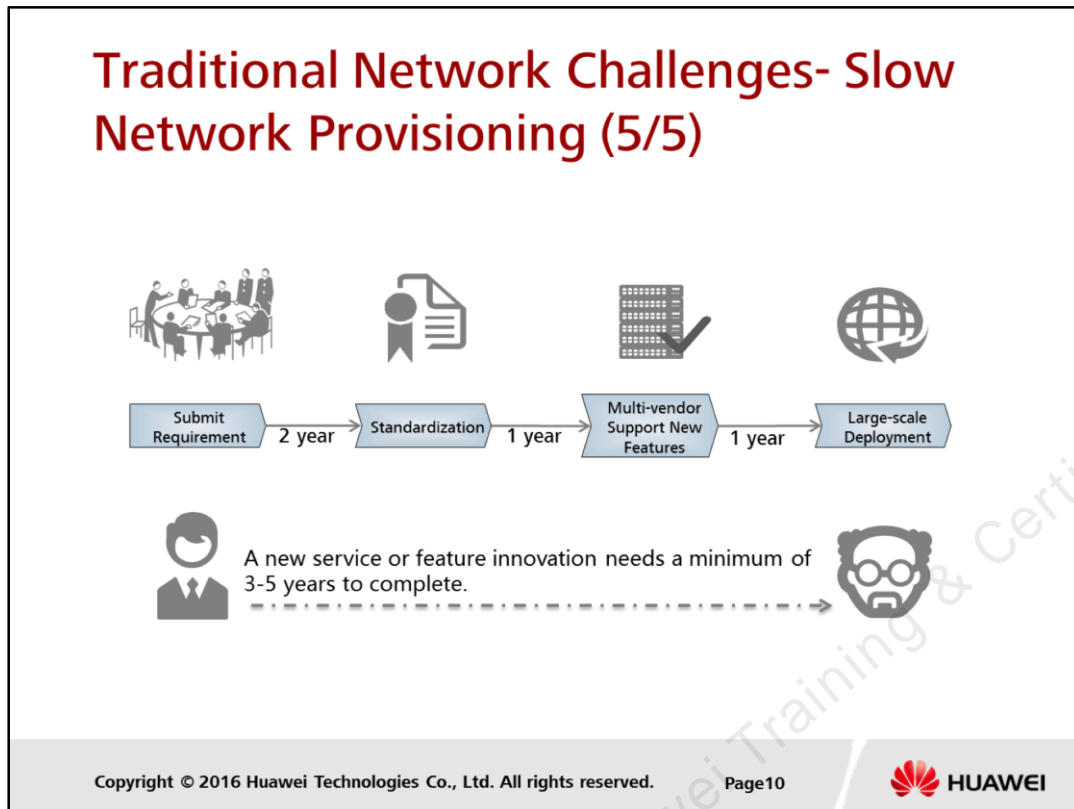
```

Network complexity due to network protocol complexity and different vendor device operability

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page9 HUAWEI

- In order to cope with the aggressively increasing network requirement, network expansion and increasing numbers of protocols and features deployed in the network will definitely increasing; Various types of network protocols and features have to be deployed in order to cater for different network requirements.
- The example shown on the slide above clearly shows the complexity of network O&M and configurations in traditional network; for instance, to configure a L3VPN service in network, the steps of configuring IGP protocol, configuring MPLS, configure L3VPN instance, configure routing protocol between PE and CE and configuring MP-BGP peer; these series of configurations need to be done on all provider edge (PE) routers which are connected to customer edge (CE) routers. The configuration scripts shown on the right shows a complete configurations command deployed on PE3, just to configure a L3VPN service, and the command involved is a lot.
- The example given above is based on the assumption that PE devices connected are all Huawei devices. Configuration work will become even more complicated if it is applied to the scenario that different vendors devices are used in the network; for examples, some PE devices are belonged to Cisco while some belongs to Huawei's. The network engineer needs to be familiar on Cisco configuration platform and also Huawei configuration platform in order to complete the end to end configuration in this case.



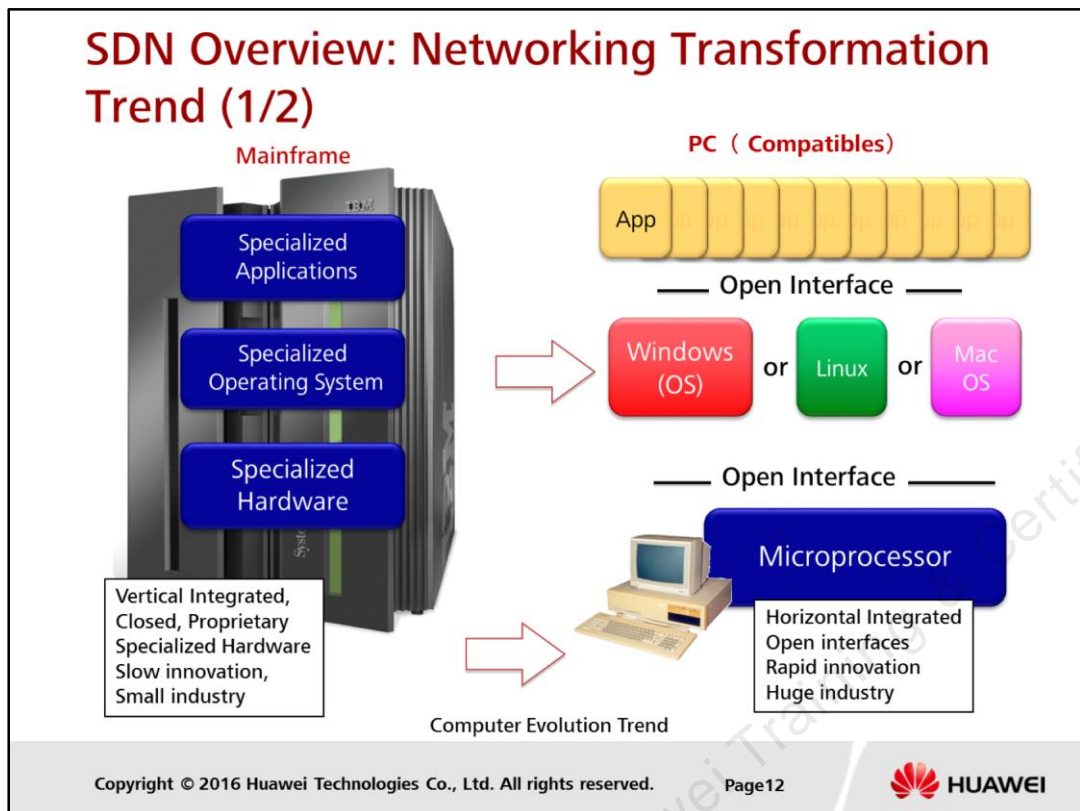


- Another major issue brought by the distributed networking architecture is the slow network innovation and slow network provisioning. This can be proven to see that there are very less new service innovation is completed since the past 30 years. The service features, such as L3VPN and L2VPN are the conventional features which have been developed long ago. Why is this scenario happening?
- In the process of defining a new service feature or service type, this service features must be standardized first before proceeding with further development. Thus, the service requirement will first needs to be sent to IETF, to standardize the feature standard. This process normally takes up 1 to 2 years for definition discussion before the standards about this feature is released officially; Standardization is very crucial in the step to ensure multi-vendor inter-operability; as all vendors will do research and survey on this features based on the standard released.
- Each vendor will then need at least one year to embed this features into vendors' devices once R&D has successfully developed the features based on IETF standards; Devices then can be upgraded to support this particular features; after this, deployment and configurations need to be planned and performed too, to make this particular service goes online.
- The whole process will require 3 to 5 years duration and this long duration definitely cannot fulfill the service requirement from the network operator; this needs to be solved as soon as possible to guarantee operator's satisfaction



## Contents

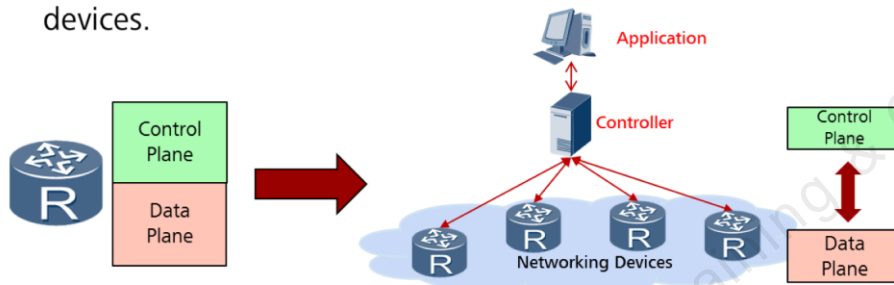
1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. SDN Value Proposition
5. SDN Challenges and Solutions
6. SDN Related Concepts and Organizations
7. SDN Influences to Current Telecom Network



- To better explain SDN for easier understanding, the analogy between computer evolution and network evolution is normally used as the example of discussion.
- Above showing analogy comparison when buying computer with specialized hardware, specialized operating system and specialized applications, all-in-one package, in which manufactured by vendor or computer with microprocessor with open interfaces published, in which leads to many operating system that able run on top of it and also equipped with open interfaces in which leads to many applications that able installed on top of operating system
- For the first part, is called vertical integrated, mostly closed and proprietary, lead to slow innovation and small industry.
- Latter part, is called horizontal integrated, very rapid innovation and became huge industry.
- Reference from Slides titled "How SDN will shape Networking" by Nick McKeown in Open Networking Summit 2011.

## SDN Overview: SDN Definition

- “Software Defined Networking”
- Introduced by Project Clean Slate - “Redesign the Network”
- The physical separation of the network control plane from the forwarding plane, and where a control plane controls several devices.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

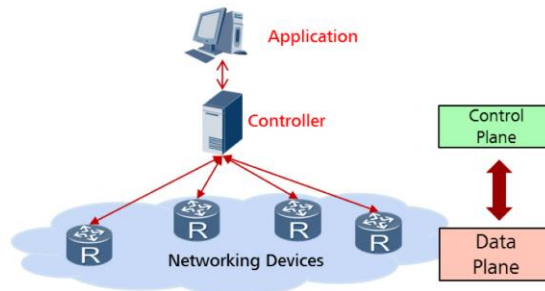
Page13



- Software defined networking is a new networking approach where physically separation of the network control plane from the forwarding plane. SDN introduces new component called controller, in which manage several devices in centralized manner.
- It is reconstruction on the current networks. In future, new services are deployed by programming on the SDN controller and adding or upgrading the software programs on the SDN. Customer requirements can be met quickly.

## SDN Overview: Characteristics of SDN Controller

- There are three main characteristics of SDN Controller
  - Separation of Forwarding Plane and Control Plane
  - Centralized Networking Approach
  - Open Interfaces



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page14



- Separation of Forwarding Planes and Control Plane
  - The separation of control plane and forwarding plane is one of fundamental of SDN.
  - The control plane establishes the local data set used to create the forwarding table entries, which are in turn used by the data planes to forward the traffic between ingress and egress ports on a device. The data set used to store the network topology is called the routing information base (RIB). The RIB is often kept consistent (i.e. Loop free) through exchange of information between other instances of control planes within the network. Forwarding table entries are commonly called as forwarding information base (FIB) and are often mirrored between the control and data planes of a typical devices. The FIB is programmed once the RIB is deemed consistent and stable.
  - In a typical SDN, the network intelligence is logically centralized in controllers (software-based), which enables the control logic to be designed and operated on a global network view, as a centralized application, rather than a distributed system.

## SDN History: Origin of SDN



Stanford University  
**CLEAN SLATE**  
An Interdisciplinary Research Program


- Clean Slate Program
  - “.. explore what kind of Internet we would design if we were to start with a clean slate and 20-30 years of hindsight...”



Nick McKeown



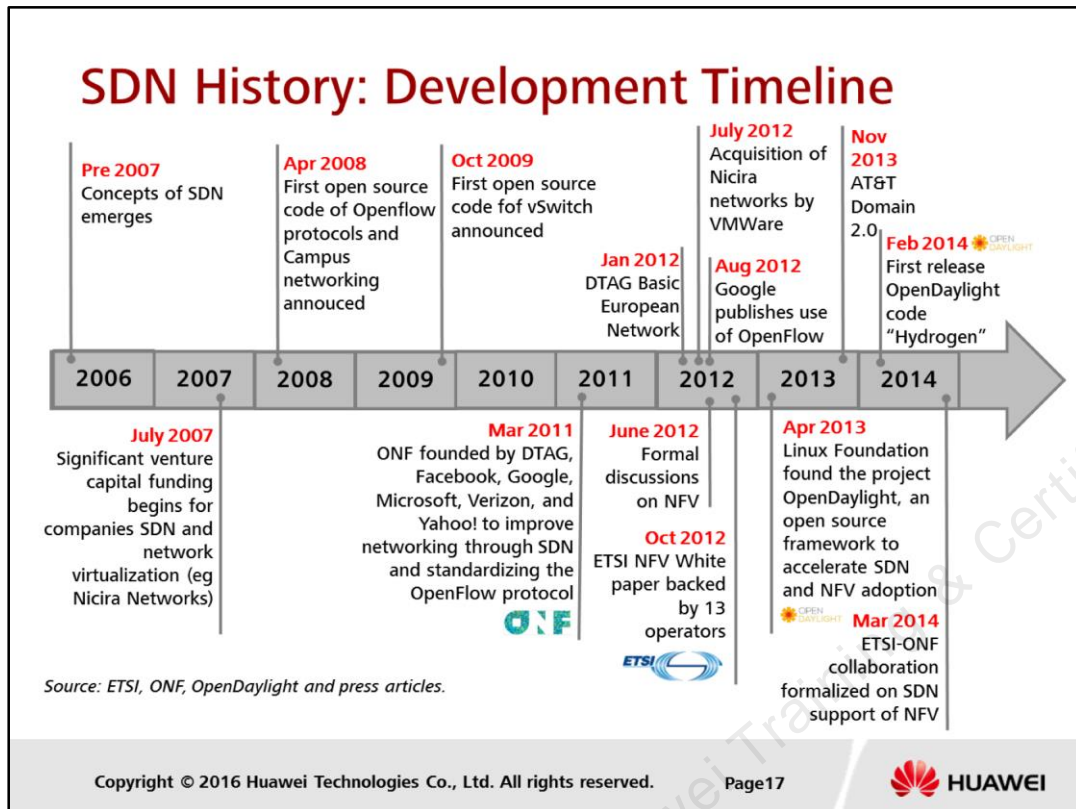
Martin Casado



Scott Shenker

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page16 

- In year 2006, SDN is originated from the Clean Slate Program in Standford University which is funded by GENI in United States. Led by Nick McKeown, the processor of the research team in Stanford University
- McKeown is active in the software-defined networking (SDN) movement, which he helped start with Scott Shenker and Martin Casado. SDN and OpenFlow arose from the PhD work of Casado at Stanford University, where he was a student of McKeown. OpenFlow is a novel programmatic interface for controlling network switches, routers, Wi-Fi access points, cellular base stations and WDM/TDM equipment. OpenFlow challenged the vertically integrated approach to switch and router design of the past twenty years. McKeown works closely with Guru Parulkar, Executive Director of the Stanford Open Network Research Centre (ONRC) and the Open Networking Lab (ON.Lab).[8]
- In 2007, Casado, McKeown and Shenker co-founded Nicira Networks, a Palo Alto, California based company working on network virtualization, acquired by VMWare for \$1.26 billion in July 2012. In 2011 McKeown and Shenker co-founded the Open Networking Foundation (ONF) to transfer control of OpenFlow to a newly created not-for-profit organization. Above are reference from Wikipedia.



- Diagram shown on the slide describes some important events on SDN development timeline.



## Contents

1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. SDN Value Proposition
5. SDN Challenges and Solutions
6. SDN Related Concepts and Organizations
7. SDN Influences to Current Telecom Network





## Contents

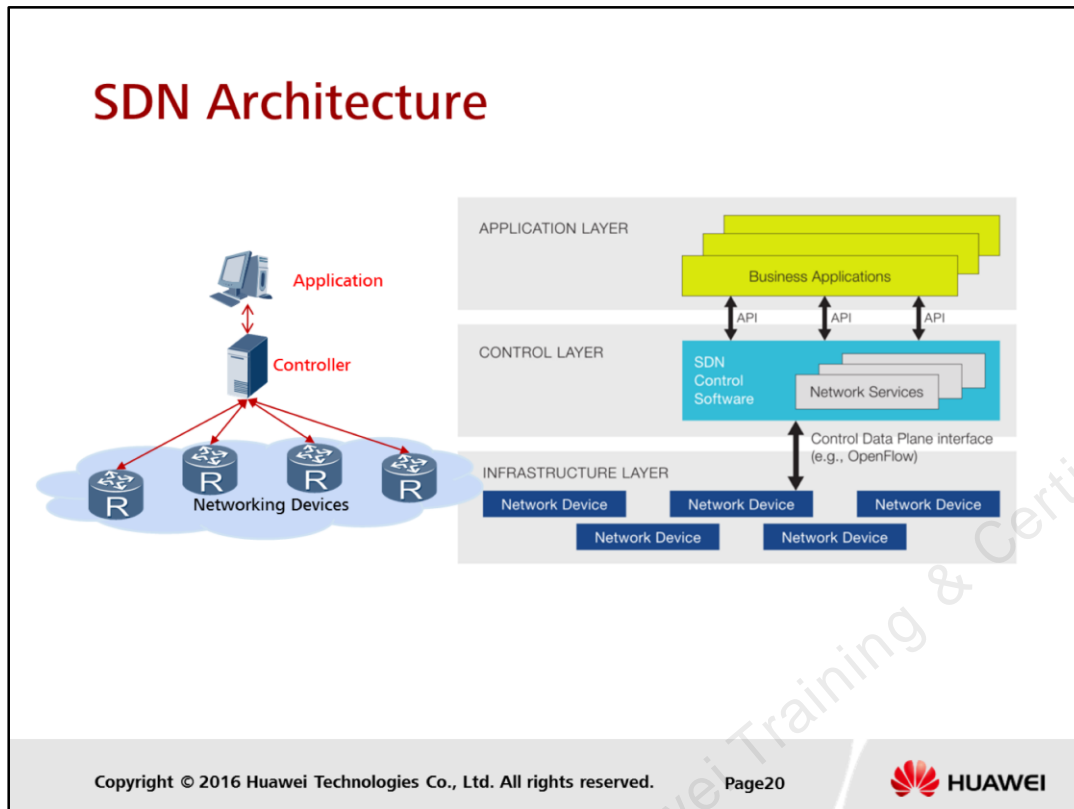
### 3. SDN Network Architecture

#### 3.1 SDN Network Architecture Overview

#### 3.2 SDN Infrastructure Layer

#### 3.3 SDN Control Layer

#### 3.4 SDN Application Layer



- As shown in the figure, there are three different layers:
  - **Application Layer:** Encompasses solutions that focus on the expansion of network services. These solutions are mainly software applications that communicate with the controller.
  - **Control Plane Layer:** Includes a logically-centralized SDN controller, which maintains a global view of the network. It also takes requests through clearly defined APIs from application layer and performs consolidated management and monitoring of network devices via standard protocols.
  - **Infrastructure or Data-plane Layer:** Involves the physical network equipment, including Ethernet switches and routers. Provides programmable and high speed hardware and software, which is compliant with industry standards.
  
- In a software defined network, the control plane and the data plane are separated. The OpenFlow protocol is a foundational element for building SDN solutions. The SDN Architecture is :
  - **Directly programmable :** Network control is directly programmable because it is decoupled from forwarding functions.
  - **Agile:** Abstracting control from forwarding lets administrators dynamically adjust network-wide traffic flow to meet changing needs.
  - **Centrally managed:** Network intelligence is (logically) centralized in software-based SDN controllers that maintain a global view of the network, which appears to applications and policy engines as a single, logical switch.
  - **Programmatically configured:** SDN lets network managers configure, manage, secure, and optimize network resources very quickly via dynamic, automated SDN programs, which they can write themselves because the programs do not depend on proprietary software.
  - **Open standards-based and vendor-neutral:** When implemented through open standards, SDN simplifies network design and operation because instructions are provided by SDN controllers instead of multiple, vendor-specific devices and protocols.



## Contents

### **3. SDN Network Architecture**

#### 3.1 SDN Network Architecture Overview

#### **3.2 SDN Infrastructure Layer**

#### 3.3 SDN Control Layer

#### 3.4 SDN Application Layer

## SDN Infrastructure Layer

- Infrastructure layer mainly consist of forwarding devices.
- Responsible for forwarding traffic.
- Forwarding table can be layer 2 forwarding table or layer 3 forwarding table. The entries is decided or calculated by the controller, not the forwarding devices
- Infrastructure layer and Control layer are communicate using southbound interface. Infrastructure layer basically report network resource status and information to the control layer and receive forwarding information from Control layer.

- At the bottom layer, the physical network consists of the hardware forwarding devices which store the forwarding information base (FIB) state of the network data plane (e.g., TCAM Entries and configured port speeds), as well as associated meta-data including packet, flow, and port counters. The devices of the physical network may be grouped into one or more separate controller domains, where each domain has at least one physical controller. OpenFlow plane interface or standards-based protocols, typically termed as 'southbound protocols', define the control communications between the controller platform and data plane devices such as physical and virtual switches and routers. There are various southbound protocols such as OpenFlow, PCEP, SNMP, OVSDb, etc.



## Contents

### **3. SDN Network Architecture**

#### 3.1 SDN Network Architecture Overview

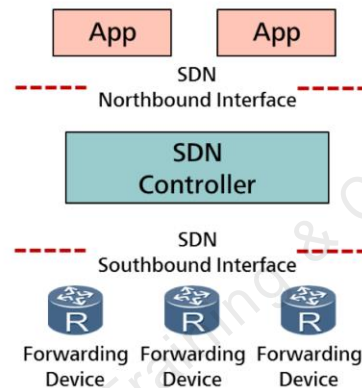
#### 3.2 SDN Infrastructure Layer

#### **3.3 SDN Control Layer**

#### 3.4 SDN Application Layer

## SDN Control Layer

- Core component of the SDN Architecture
- Collect physical network state, calculating routing algorithm and deliver routing entries to forwarder.
- Realized by Controller.
- Provide northbound interface connecting to application layer and southbound interface connecting to infrastructure layer.



- The control-plane layer is the core of the SDN, and is realized by the controllers of each domain, which collect the physical network state distributed across every control domain. This component is sometimes called the “Network Operating System” (NOS), as it enables the SDN to present an abstraction of the physical network state to an instance of the control application (running in Application Layer), in the form of a global network view.



## Contents

### **3. SDN Network Architecture**

#### 3.1 SDN Network Architecture Overview

#### 3.2 SDN Infrastructure Layer

#### 3.3 SDN Control Layer

#### **3.4 SDN Application Layer**

## SDN Application Layer

- Application Layer covers an array of applications such as OSS, Openstack, etc.
- OSS handles network service provisioning; for example, Openstack used in Data Center Network.
- Application layer also can be consisting of some other applications such as security applications such as Security Apps, Service Provision App, etc.
- Control Layer provides Northbound Interface, for example, RestFul, Netconf , other open APIs to application layer.

- Northbound open APIs refer to the software interfaces between the software modules of the controller and the SDN applications. These interfaces are published and open to customers, partners, and the open source community for development. The application and orchestration tools may utilize these APIs to interact with the SDN Controller.
- Application layer covers an array of applications to meet different customer demands such as network automation, flexibility and programmability, etc. Some of the domains of SDN applications include traffic engineering, network virtualization, network monitoring and analysis, network service discovery, access control, etc. The control logic for each application instance may be run as a separate process directly on the controller hardware within each domain.





## Contents

1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. **SDN Value Proposition**
5. SDN Challenges and Solutions
6. SDN Related Concepts and Organizations
7. SDN Influences to Current Telecom Network

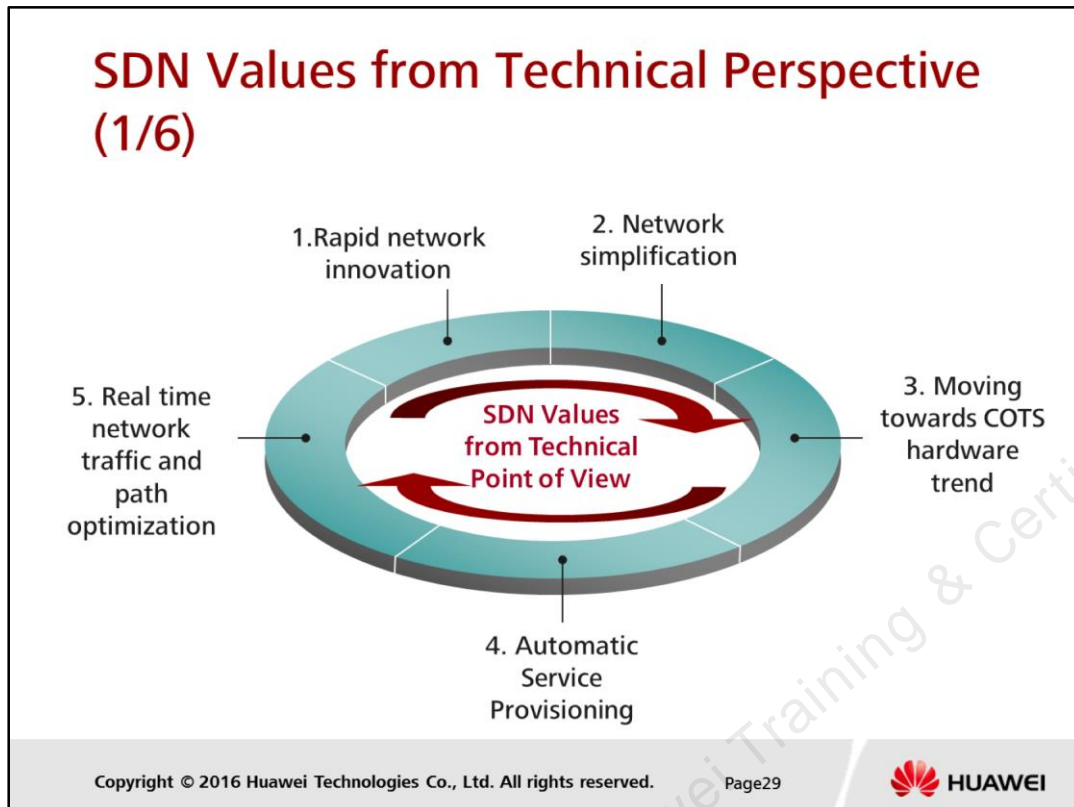


## Contents

### 4. SDN Value Proposition

#### 4.1 SDN Values from Technical Perspective

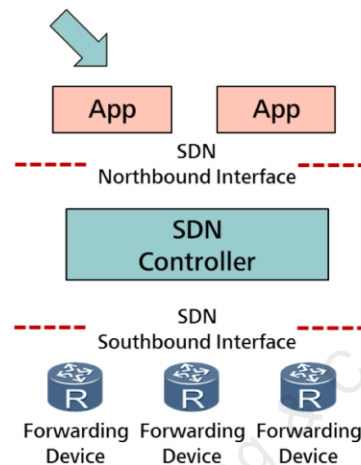
#### 4.2 SDN Values from Operator Perspective



- SDN network is able to bring various values to operators and customers, including real time traffic optimization, network simplification, service automation, and rapid service innovation. The main factor for realizing all these crucial values is related to SDN main characteristics, which are control and forwarding plane separation, centralized control and open programmability.
- Once control and forwarding plane is separated and centralized control is realized, SDN controller is based on a software calculation to calculate internal network routes, generates forwarding table and forward the forwarding entry to the forwarders. Hence, the original mass number of routing protocols usage is replaced by software calculation and this simplifies network protocols indirectly. In other words, SDN characteristic of control and forwarding plane separation brings network simplification!
- Through SDN controller, network can be treated as a software-based logical router. SDN controller can directly provide corresponding service interfaces to end users regardless the technologies used behind. Through SBI, users can automate services to be deployed directly. Unlike traditional network, users need to know the details of the technologies. For instance, in traditional network, to realize a PW service, users need to know the concepts of MPLS tunneling technology, LDP protocols, and PW parameters etc.; However, SDN controller NBI is dealing with software models, which can be directly deployed for service automation, and this centralized control characteristic of SDN controller helps to realize service automation and rapid service provisioning.

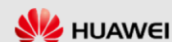
## Rapid Network Innovation (2/6)

- SDN open source and direct programmability contributes to rapid network innovation
- It can be achieved because the network architecture is centrally managed by a controller; controller has a global view of topology.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

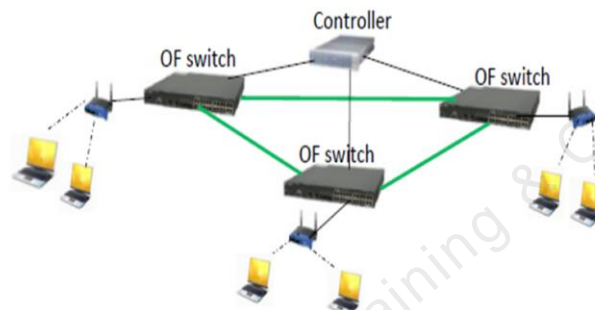
Page31



- One of the characteristics of SDN Network Architecture is providing open interface and direct programmability, in which gives network industry a rapid service provisioning as well as innovation. When new services are required to be deployed, you can just modify the software for SDN or improve the agility of programmability of the software, hence providing a faster way for service online.
- Unlike the traditional network service provisioning process discussed in the earlier section, to make a newly developed service go online, processes such as service requirement submission to IETF, service requirement standardization, service requirement research, service requirement publishing need to be gone through and the whole process will take up 3 to 5 years!
- SDN network provides a faster network innovation capability. If the new service has a value, it will be maintained in the network. If the new service is no longer needed, it can be made offline by means of software. SDN brings the time of new service online from number of years into few months.

## Network Simplification (3/6)

- SDN characteristic of control and forwarding plane separation simplifies network by reducing a huge number of IETF standardization protocol deployed. It means that OpEx has been drastically reduced, and this further improves service provisioning rapidly.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page32

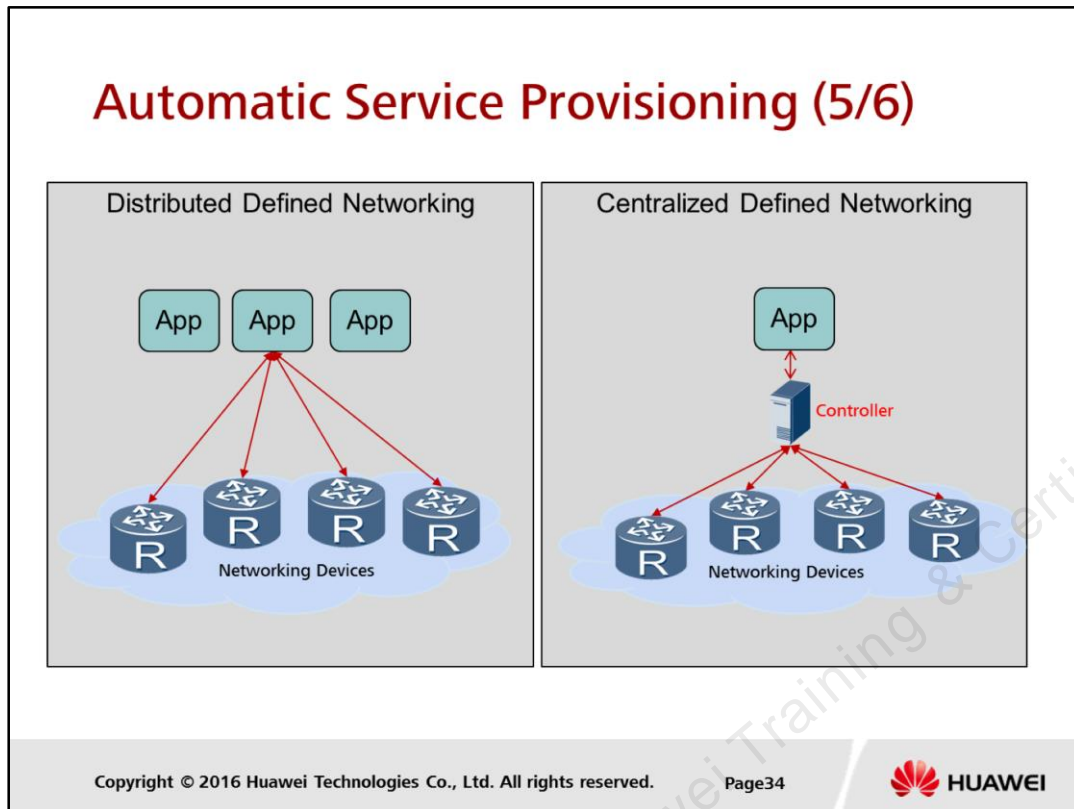


- Once control and forwarding plane is separated and centralized control is realized, SDN controller is based on a software calculation to calculate internal network routes, generates forwarding table and forward the forwarding entry to the forwarders. Hence, the original mass number of routing protocols usage is replaced by software calculation and this simplifies network protocols indirectly. In other words, SDN characteristic of control and forwarding plane separation brings network simplification!
- Due to physically separation of control plane and data plan, all the control operation handled by the controller, which means that protocols such as LDP, RSVP, MBPG, PIM, etc are no longer required. It is because all the computational routing algorithm is handled by the controller, and pass the result to the forwarding devices. In future, all those traditional protocol will be no more required, and replaced by southbound interface and northbound interface.

## COTS Hardware Implementation (4/6)

- COTS (Commercial Off The Shelf) hardware is referring to the ability of using 'generic,' off-the-shelf switching (or white box switching) and routing hardware, in the forwarding plane of a software-defined network (SDN).
- COTS switches rely on an operating system (OS), which may come already installed or can be purchased from a software vendor and loaded separately

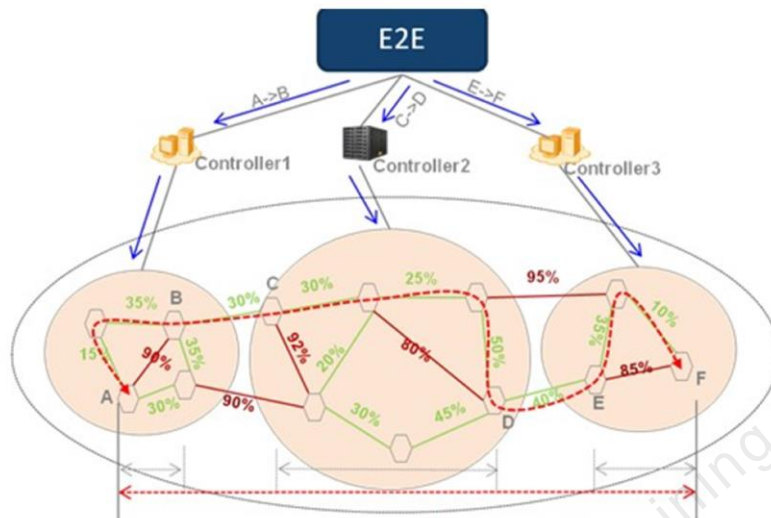
- COTS, Commercial Off The Shelf hardware is referring to the ability of using 'generic,' off-the-shelf switching (or white box switching) and routing hardware, in the forwarding plane of a software-defined network (SDN). COTS switches are really just that – 'blank' standard hardware. They represent the foundational element of the commodity networking ecosystem required to enable organizations to pick and choose the elements they need to realize their SDN objectives.
- COTS switches rely on an operating system (OS), which may come already installed or can be purchased from a software vendor and loaded separately, to be integrated with the deploying organization's Layer 2/Layer 3 topology and support a set of basic networking features. A common operating system for white box switches is Linux-based because of the many open and free Linux tools available that help administrators customize the devices to their needs. Traditional switches and routers generate and maintain their own forwarding and routing tables that can, generally speaking, broadcast to neighboring switches and routers. A COTS switch may come pre-loaded with minimal software or it may be sold as a bare metal device. The advantage of this approach is that switches can be customized to meet an organization's specific business and networking needs



- Within SDN environment, the controller has the global view of network topology. Application running on top of can implement automation service provision easily.
- Through SDN controller, network can be treated as a software-based logical router. SDN controller can directly provide corresponding service interfaces to end users regardless the technologies used behind. Through SBI, users can automate services to be deployed directly. Unlike traditional network, users need to know the details of the technologies. For instance, in traditional network, to realize a PW service, users need to know the concepts of MPLS tunneling technology, LDP protocols, and PW parameters etc.; However, SDN controller NBI is dealing with software models, which can be directly deployed for service automation, and this centralized control characteristic of SDN controller helps to realize service automation and rapid service provisioning.



## Real Time Network Traffic and Path Optimization (6/6)



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page35



- In fact, there are some traditional network traffic engineering techniques to solve such shortest path congestion problems, such as traffic engineering MPLS TE is one of example of technology, but the technology is similar to other traditional protocols - fully distributed. It will lead to previously described business order dependency problems; another aspect of traditional traffic engineering protocols RSVP soft state because of its mechanism, it can not lead to large-scale deployment. In order to adopt SDN architecture, business can focus directly path computation and directly establish a tunnel, no RSVP protocol, not only can solve the real-time traffic paths dynamically adjust capacity to enhance network utilization, and also depend on the order of business to solve the problem.
- As SDN controller is able to obtain real time network status information such as interface bandwidth utilization rate, service traffic rate etc, SDN controller is able to perform traffic optimization and adjustment when there is traffic congestion in the network, thanks to SDN ability of performing centralized control.



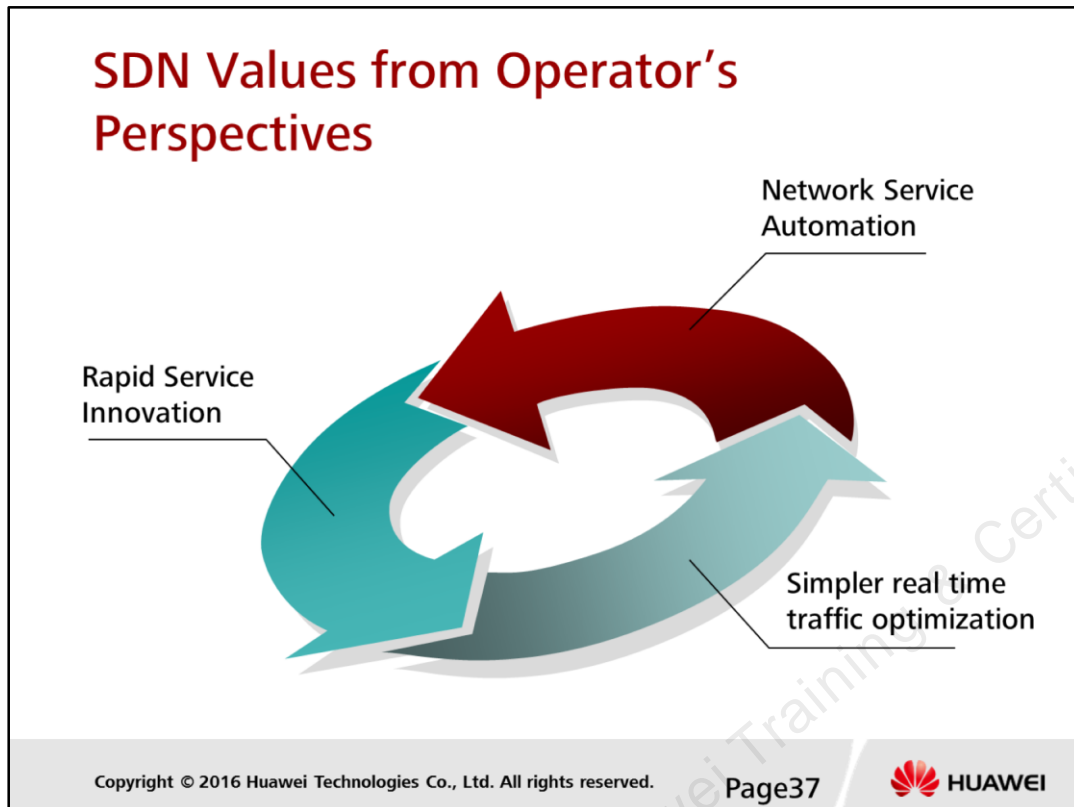


## Contents

### 4. SDN Value Proposition

#### 4.1 SDN Values from Technical Perspective

#### 4.2 SDN Values from Operator Perspective



- From operator's point of view, SDN achieves 3 main values, as per listed below:-
  - **Network Service Automation.**
    - In traditional distributed networking, network service automation can only be realized through OSS or NMS. Operators perform service configurations on the NMS, and distributes the dedicated configuration script to each and every NE; once the configuration on all NE is completed, service is finished provisioning and can be made online. However, the potential risk and issue is on the single point failure issue; for instance, if configuration on one NE has some mistakes, this might cause network loop or service interruption. In traditional method, it is going to be a huge effort to perform network automation through NMS with minimum errors, as there are enormous numbers of different devices from different vendors existing in the network!
    - Besides, this SDN value of network service automation would even be more highly appreciated in the data center industry. Data center tenants hope to rent the computing resources, storage resources etc within minutes or seconds, and at the same time these rented resources should be isolated from other tenants. As virtual network is widely deployed in data center, and some data center service providers have already started with the service of providing IAAS (Infrastructure as a service), network service automation is the most needed characteristics in data center, which this is something that is hardly to be achieved by traditional network.



## Contents

1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. SDN Value Proposition
5. **SDN Challenges and Solutions**
6. SDN Related Concepts and Organizations
7. SDN Influences to Current Telecom Network

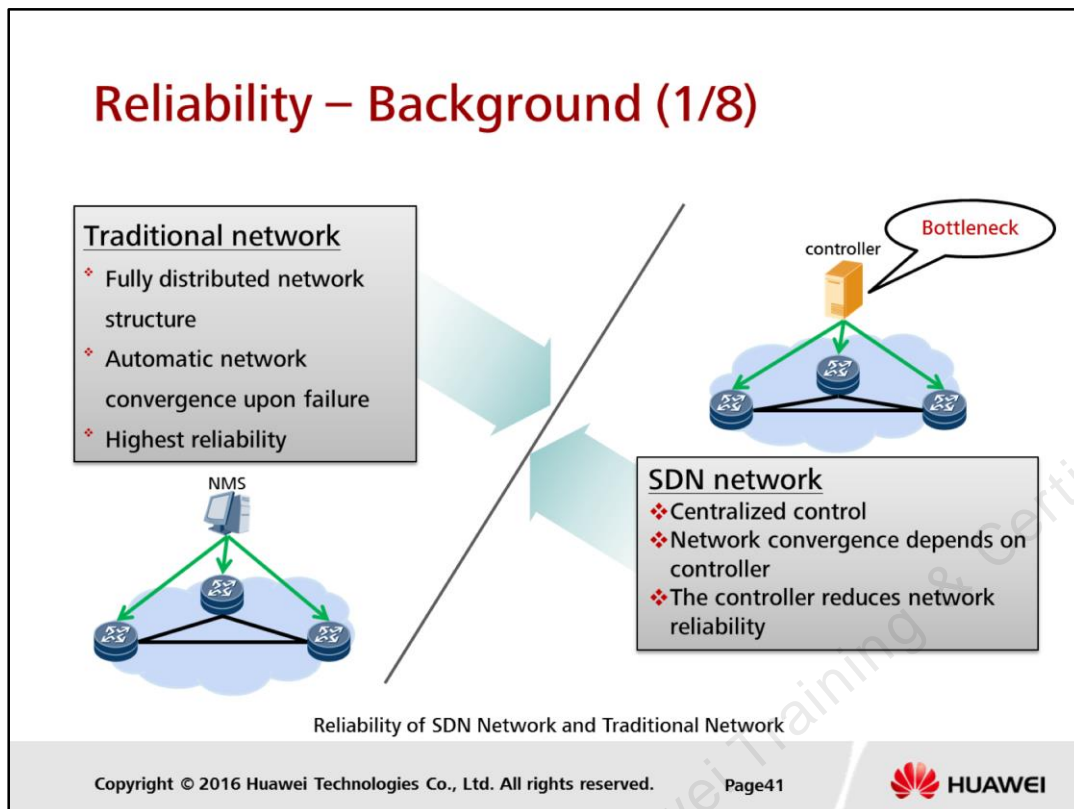
 **Contents**

**5. SDN Challenges and Solutions**

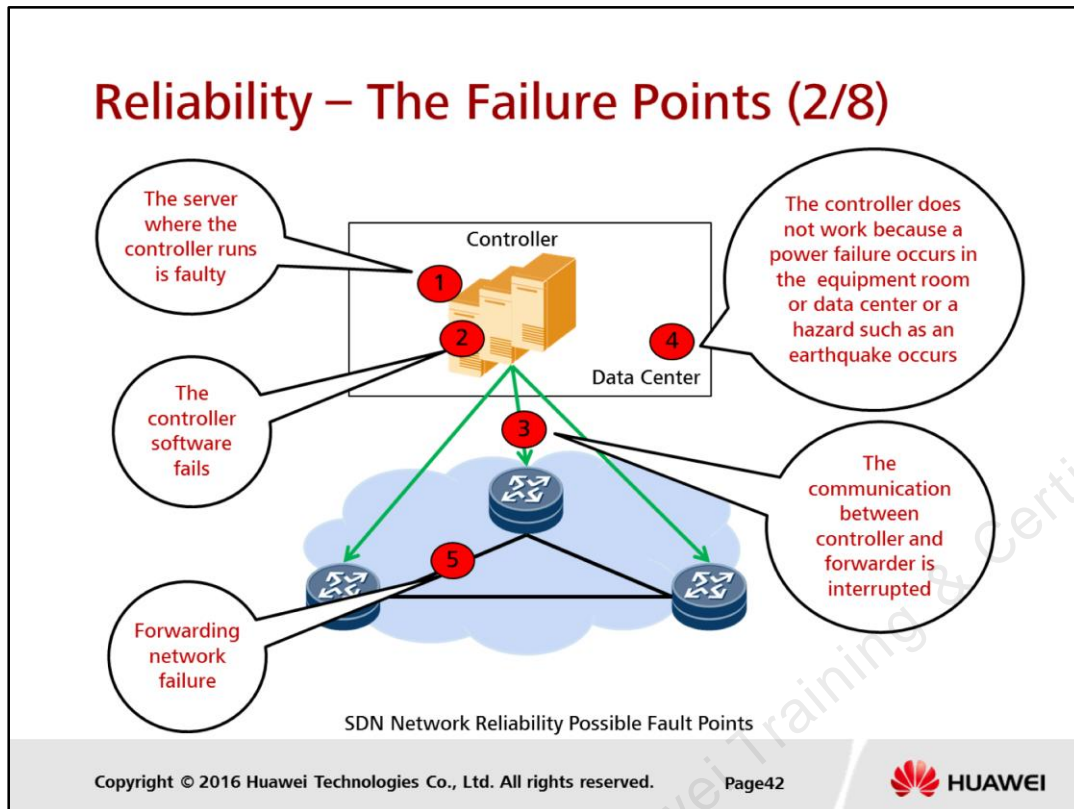
**5.1 Reliability**

5.2 Performance

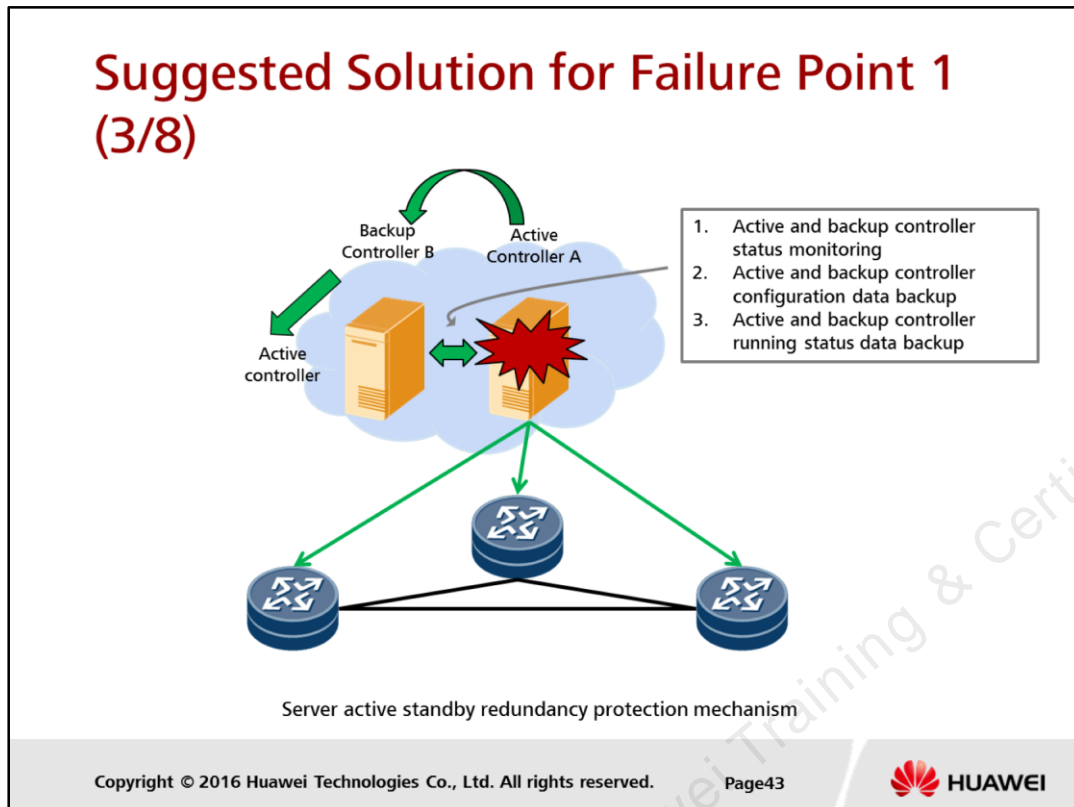
5.3 Openness Capability



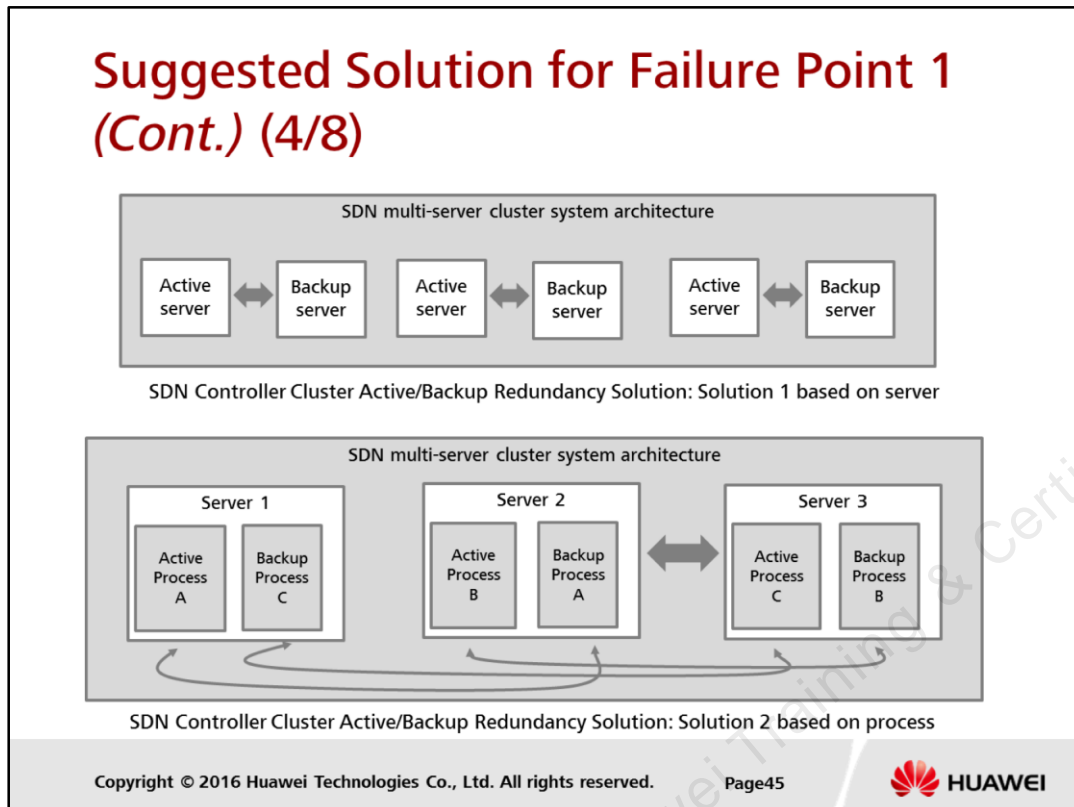
- Difference between distribute networking in traditional network and centralized networking with SDN controller as the core of the network are as follow:
  - Traditional Network
    - Fully distributed networking structure where each routers are calculating routing entries independently. Each routers in the network collect link state information as material for the routing algorithm to calculates to shortest path to certain destination. Any changes of network topology will trigger the routers to flood the new link state information and recalculates the new route independently. Hence, traditional network has an automatic network convergence upon failure. This also has the highest of reliability.
  - In SDN network, the network convergence depends on the SDN controller as SDN controller are the core component and manage the networking devices in centralized manner. As a result , It may introduce single point of failure.



- Basically in SDN network, there are 5 failure points that contribute to the factors of reliability, as per listed below:-
  1. Server where controller runs may experience faulty , for example, server shutdown due to power down.
  2. The controller software fails
  3. The communication between SDN and forwarder fails
  4. Some disasters happen in data center that consists of servers.
  5. Forwarding network failure, such as links down between forwarders.
  
- The first 4 failure points out of the 5 listed above will not immediately affect services in the forwarding network; in other words, forwarding network is still working properly, and forwarders can perform normal forwarding using the forwarding entries that are distributed before failures happen. However, services might be interrupted if there is some changes in the forwarding network, for instance link down between forwarders, because controller cannot send new calculation forwarding entry if any failure point from 1 to 4 happens. Thus these failures should be solved as sooner possible as well. As for failure number 5, that involves forwarding network topology changes and controller should trigger recalculation to establish new forwarding entry for the affected path.

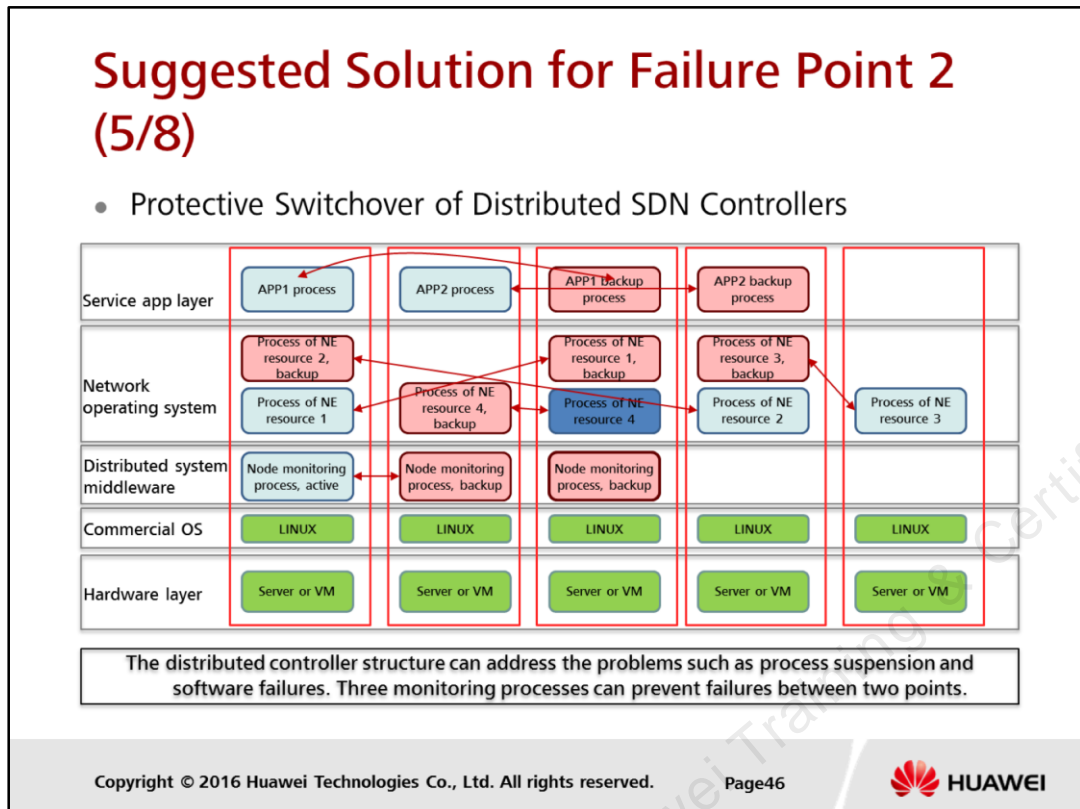


- It is suggested to use server redundancy as a protection measure for failure point 1, which is the server hardware failure.
- Dual-device hot backup is implemented based on the following mechanisms:
  - Active/Standby control: Two controllers are specified for a forwarder. When priorities are configured for the two controllers or an election mechanism is used, the forwarder selects a controller as the active controller for service forwarding. The other controller is in the standby state.
  - Service control: Service control data on the active and standby controllers must be the same in dual-device hot backup scenarios. After an active/standby controller switchover is performed, the new active controller takes over without the need of reconfiguring service control data.
- There must be a heartbeat connection established between active and backup controller. Backup controller should meet the few requirements as shown below to enable backup controller is able to take over the active controller task when it is detected that an active/backup status happens:-
  - Backup controller software should be same or compatible with active controller
  - All necessary configuration data should be loaded to backup router; normally configuration synchronization is performed between master and backup controller
  - Backup controller must be connected to the forwarding network.

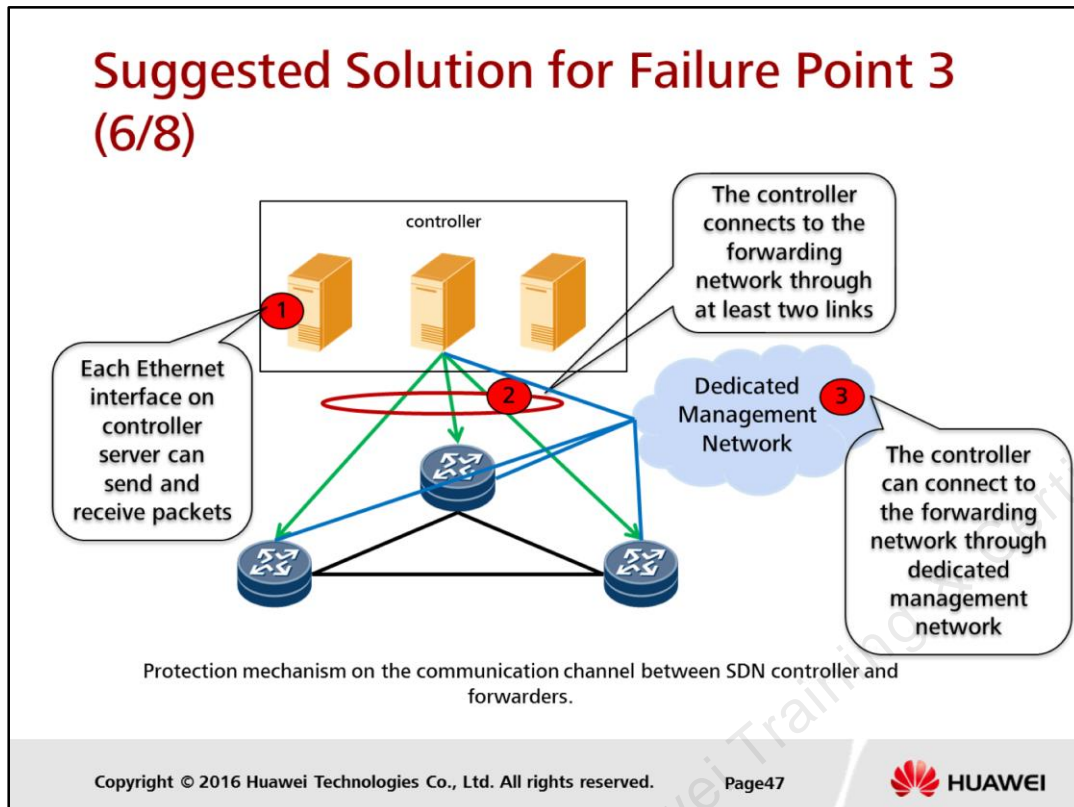


- As we know, a SDN controller hardware might be a standalone server or it might be consisting of a server cluster. Normally, server cluster solution is deployed if SDN is deployed to manage a mass numbers of forwarders. A controller cluster can be understood as a SDN controller software running on multiple physical servers in a server cluster. However, OSS or forwarders will treat this controller cluster as 1 controller, regardless its physical realization.
- In the SDN controller cluster solution, how are we going to establish server redundancy in this case? There are 2 possible solutions, as per listed below:-
  - Solution 1: All servers in cluster are divided into 2 groups, one for active and one for backup. Each active server is mapped to a backup server; if active server is down, the backup server will switchover to become active. This solution is simple and easy to be deployed.
  - Solution 2: Servers provide active/ backup protection switchover based on process level. Every server will simultaneously running the active process and backup process. All process is in charge of different functions; for example, process A, B and C has different functions. Each server serves as the active process for certain processes and at the same time, serves as backup for another process, as shown on the diagram above. For example, server 1 serves as the active server for process A, and backup server for process B; Process A is protected by backup server 2, while process B is protected by the backup server 3.

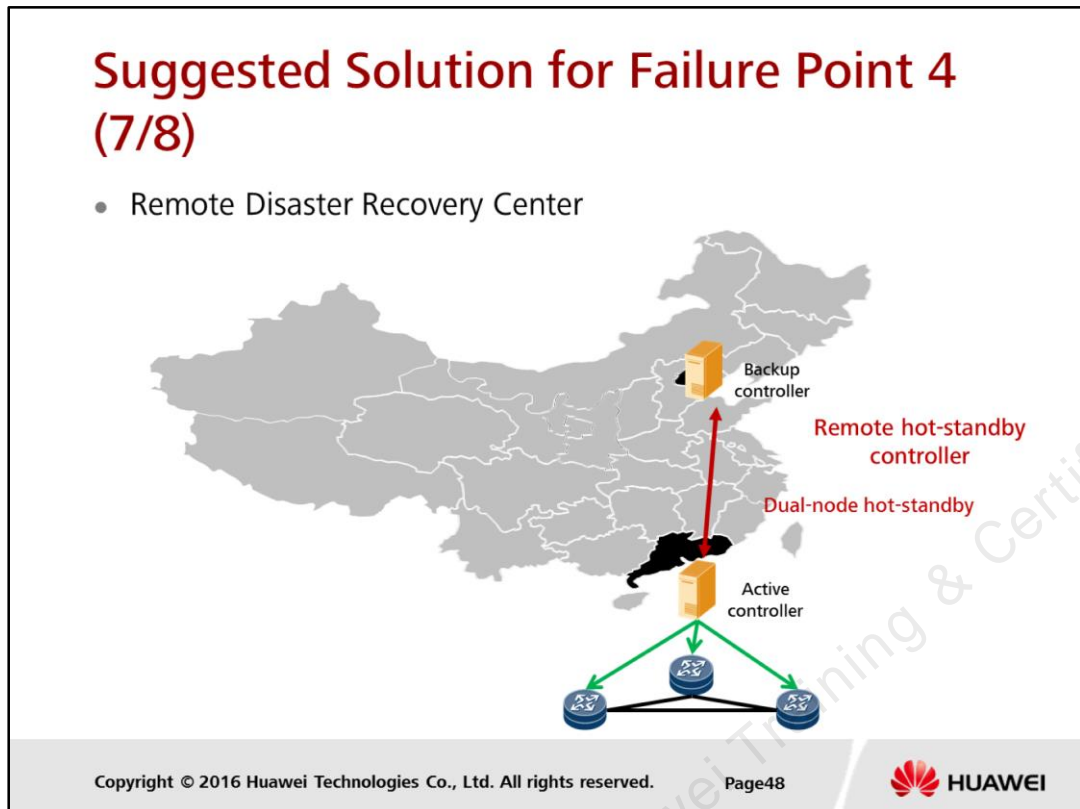




- In a cluster SDN architecture , there are many components comprise the system, such as netconf, openflow, pce, restful, snmp, telnet, etc. Some other components such as BGP , ARP also contribute to the distributed architecture. We do not hope that those component to cause problems and to the worse, causing the whole system down.
- Software component failures have many factors, such as program deadlock, or some certain unpredictable events. Those problem might affect system tp operate normally. How do we avoid those problems from happening? Below is explaining how Huawei controller realizes those reliability
- As shown as figure above:-
  - Each process must have equipped with some detection mechanism. Without detection mechanism on the component, we not able to detect the fault. For example, there is extra one layer called Distributed system middleware to do monitoring or detection purpose to monitor software status, process status, and other node in cluster system status.
  - More than one monitoring system running; other than primary monitoring system running on main server, it may have other monitoring system running on backup server(normally controller are installed on different physical server). It is possible to realize dual-device hot backup. By integrating NSR and NSF technology, once any component detect fault, NSR and NSF may come into play, to solve distributed SDN controller failure.



- When the connection channel between controller and forwarding plane is broken, this is similar to the condition that the brain is not able to send signal and control over hands and legs of our body system. To solve this, different redundancy mechanism can be introduced. The control channel normally can be established through either outband networking, which is a dedicated control channel, or inband networking, which is the shared channel between control channel and forwarding channel.
- Normally, inband networking is deployed due to its practicability and cost saving factor. Normally, controller will not be connected to every single forwarders, but just to few which are nearest to the controller. The directly connector forwarders will help to forward the control packets from controller to forwarder; However, who is going to generate these control packet path flow between forwarders? The controller? There comes a problem here, if the controller is the one generating these route, this might come to the problem that if there is network changes happen, controller will need to generate flow tables entries to forwarders before generating the control channels? This is not possible because control channel needs to be established first. This explains why the control channel established between forwarders cannot be relied on controller (for example using openflow), but forwarders need to perform the table calculation by their own using traditional routing methods such as OSPF or ISIS.



- To prevent on the data center damage due to natural disaster such as earthquakes, it is recommended to perform remote server backup in different locations, or even different countries. The issue might be raised here is, how are we going to realize data backup between active server and backup server? There are 2 solutions for this:-
  - Directly establish connection between active and backup server and backup the data directly to the backup server. However, this design is not so good; if there are multiple backup server existing in the network, the active server needs to establish connection to each backup server and change of configuration might be involved.
  - Copy the backup data into a database server and multiple backup servers obtain the backup data from the database server. This is better solution in term of scalability and extensibility.

## Suggested Solution for Failure Point 5 (8/8)

- There are a few possible solutions for SDN architecture to detect and switchover over these kind of link failure:-
  - Under SDN architecture, it is possible that SDN controller can forward **FRR routes** (2 routes, active and backup) to forwarders so that during link failure, fast switchover can take place in forwarding plane without notifying controller.
  - SDN controller can automatically deploy **OAM technologies** on forwarding plane for failure detection. For instance, in MPLS network, SDN controller can automatically deploys MPLS TP OAM on forwarding network to increase O&M efficiency

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page49



- In traditional network, if there are some link failures between forwarders, the distributed control plane on each devices will automatically trigger route recalculation and trigger network convergence to recover network service. However in SDN scenario which is deploying centralized control, when there is a link failure between forwarders, controllers need to be triggered to perform recalculation. The issue here is who is performing the link failure detection and who is the one performing the route recalculation.
- There are a few possible solutions for SDN architecture to detect and switchover over these kind of link failure:-
  - Under SDN architecture, it is possible that SDN controller can forward FRR routes (2 routes, active and backup) to forwarders so that during link failure, fast switchover can take place in forwarding plane without notifying controller.
  - SDN controller can automatically deploy OAM technologies on forwarding plane for failure detection. For instance, in MPLS network, SDN controller can automatically deploys MPLS TP OAM on forwarding network to increase O&M efficiency

 **Contents**

**5. SDN Challenges and Solutions**

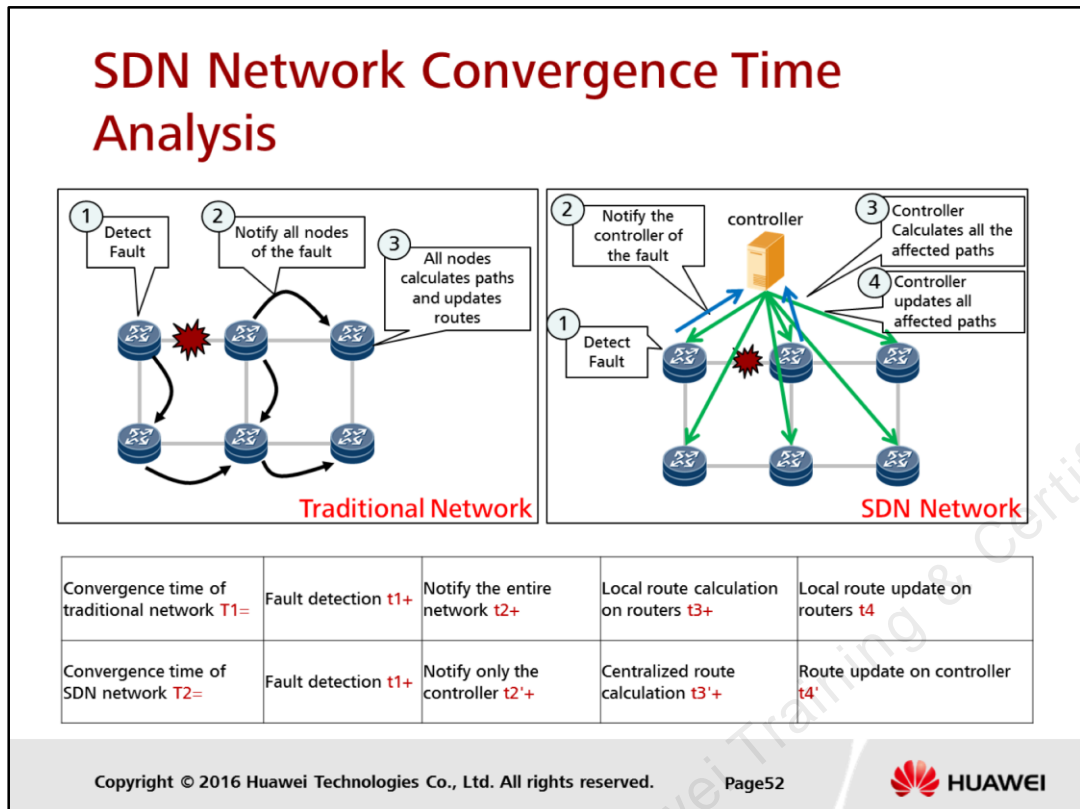
5.1 Reliability

**5.2 Performance**

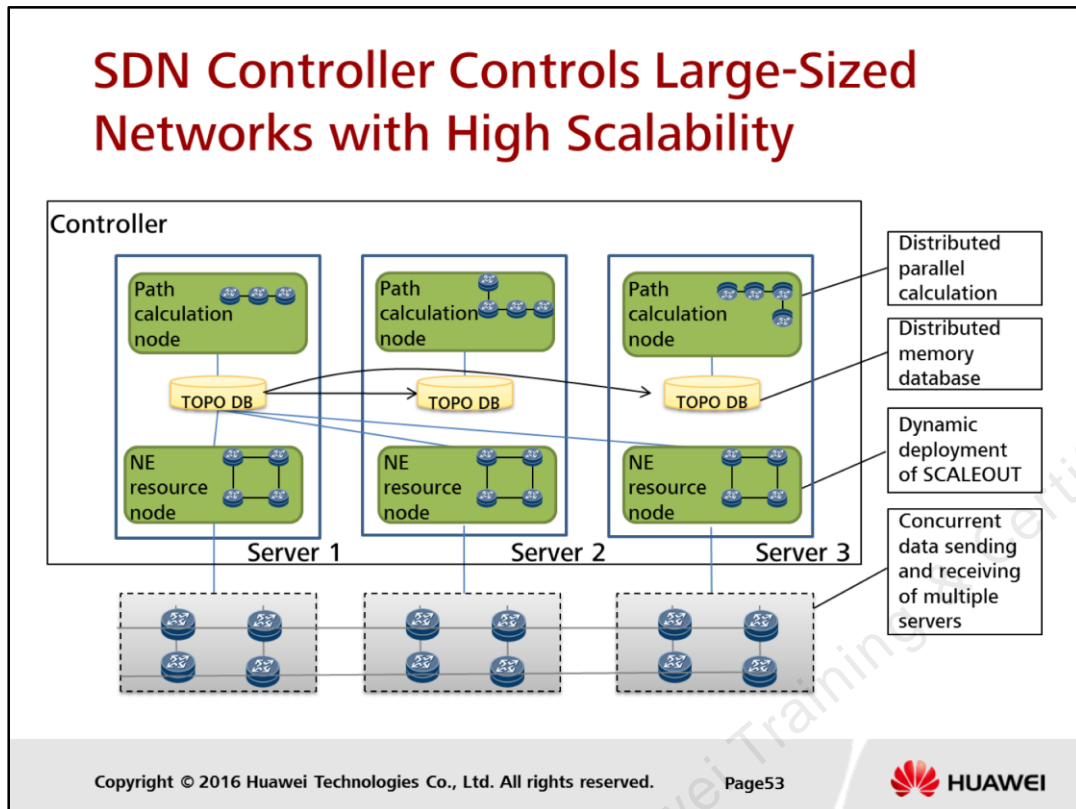
5.3 Openness Capability

## Performance Requirements on SDN Controller Structure

- Time
  - The failure convergence time of a network with an SDN controller deployed must be close to that of a traditional network.
- Space
  - The DC must have the ability to support millions of OVSs.
  - On the DCI/metro/core NETWORK, each controller needs to manage 2000 devices.
  - In the IPRAN access scenario, each controller needs to control 20000 devices.



- There are some difference in the convergence time between traditional network and SDN network
  - Take an example of traditional network as shown as figure above, when a fault occur on the network. The total convergence time would be the sum of Fault detection time, Notification time for entire network, time for local route calculation on routers and time for local routes updates on routers
  - On the other hands, in SDN network, all forwarders register and communicate through southbound interface such as OpenFlow to the controller, hence the controller have the global view of the topology. Once the fault occur, the total convergence time are the sum of fault detection time, the time for forwarder to notify fault to controller, Calculation for new route and time for routes updates to forwarder.
- To shorten the SDN network convergence time, the centralized route calculation time  $t3'$  and route update time on controller  $t4'$  must be shortened. The fault notification time  $t2'$  is shorter than  $t2$ , so the key to shorten SDN network convergence time is the algorithm, hardware performance, and distributed computing capability of the controller.



- SDN controller is able to control large size network by scaling the server clusters to handle different parts of forwarding devices.



 **Contents**

**5. SDN Challenges and Solutions**

5.1 Reliability

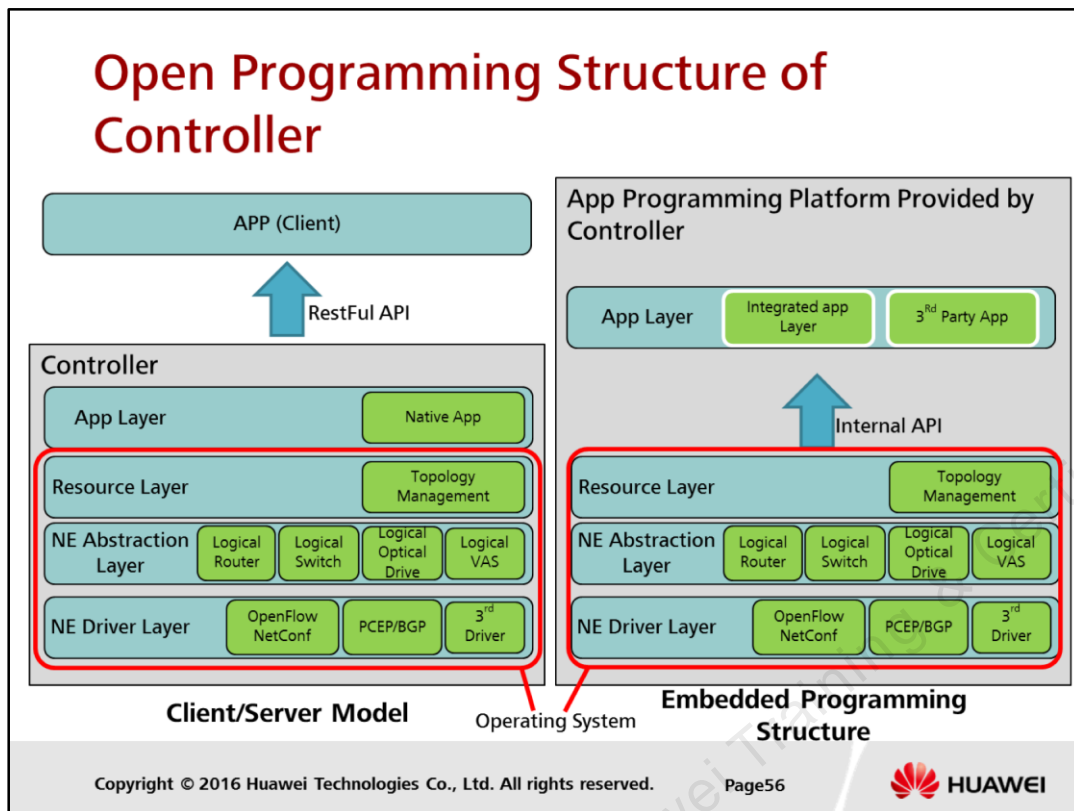
5.2 Performance

**5.3 Openness Capability**

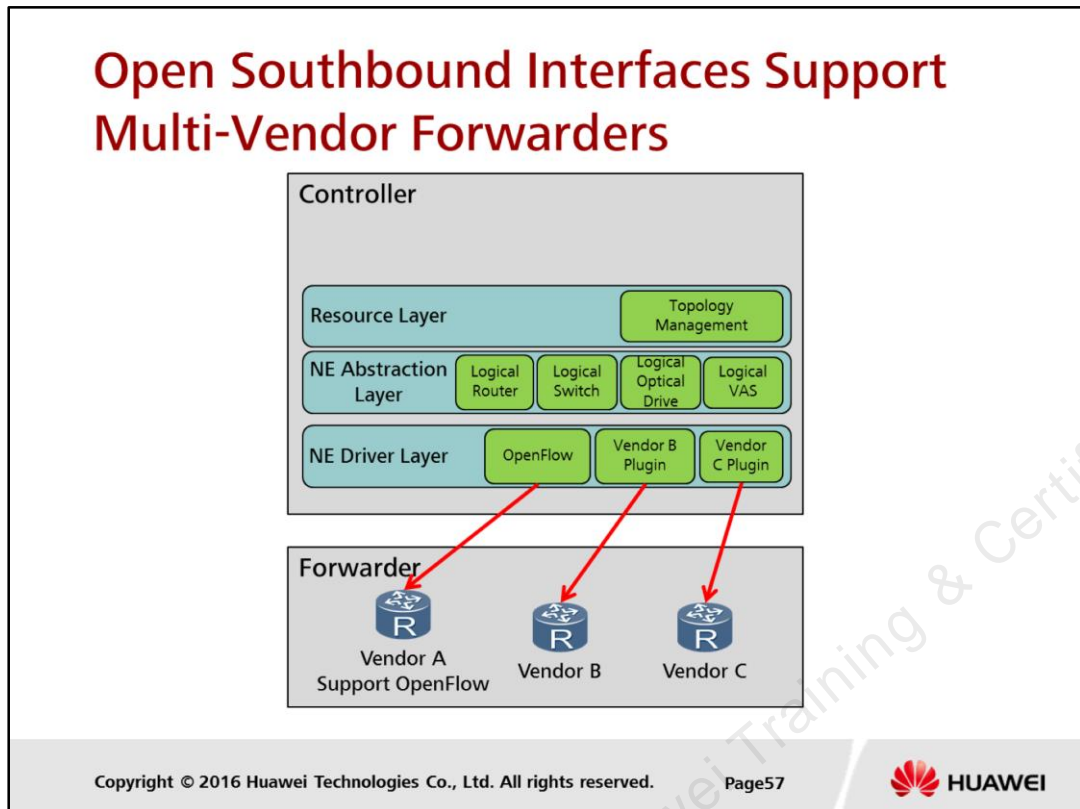
## SDN Provide Openness Capability

- Open standards-based and vendor-neutral: When implemented through open standards, SDN simplifies network design and operation because instructions are provided by SDN controllers instead of multiple, vendor-specific devices and protocols.
- Directly programmable : Network control is directly programmable because it is decoupled from forwarding functions.

- A characteristic of Open SDN is that its interfaces should remain standard, well documented, and not proprietary. The APIs that are defined should give software sufficient control to experiment with and control various control plane options. The premise is that keeping open both the northbound and southbound interfaces to the SDN controller will allow for research into new and innovative methods of network operation. Research institutions as well as entrepreneurs can take advantage of this capability in order to easily experiment with and test new ideas. Hence the speed at which network technology is developed and deployed is greatly increased as much larger numbers of individuals and organizations are able to apply themselves to today's network problems, resulting in better and faster technological advancement in the structure and functioning of networks. The presence of these open interfaces also encourages SDN-related open source projects. In addition to facilitating research and experimentation, open interfaces permit equipment from different vendors to interoperate. This normally produces a competitive environment which lowers costs to consumers of network equipment. This reduction in network equipment costs has been part of the SDN agenda since its inception.



- Within an SDN environment, the apps running on top of the SDN Controller are what provide the higher level orchestration and programmability of the network.
- Client/Server Mode : Recommended using this C/S Model to provide open source and programmable API .
  - Loosely coupled of applications and controllers provide opportunity to deploy independently, and no restriction or limitation on programming language/platform/deployment location. Short, APP and the controller are independently of each other.
  - Most of controller support this mode.
  - Performance lower than Embedded Programming Structure
- Embedded Programming Structure :
  - Restrict by programming language such as JAVA/C, depend on Platform, deployment location and run together with controller.
  - Example Apps that run on MS WIN/MAC OS etc is based on this structure.
  - Performance much better than C/S.



- SDN not only provide open Northbound interface support variety of application, it also provide Open Southbound interfaces to support multi-vendor forwarders.
- In order to solve the multi-vendor hardware compatibility issues:
  - Vendor can adopt Standard Openflow protocol, so that controller able to communicate with forwarder and advertise flow table directly into the devices
  - Or, controller supports vendor specific plugin function.



## Contents

1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. SDN Value Proposition
5. SDN Challenges and Solutions
6. SDN Related Concepts and Organizations
7. SDN Influences to Current Telecom Network



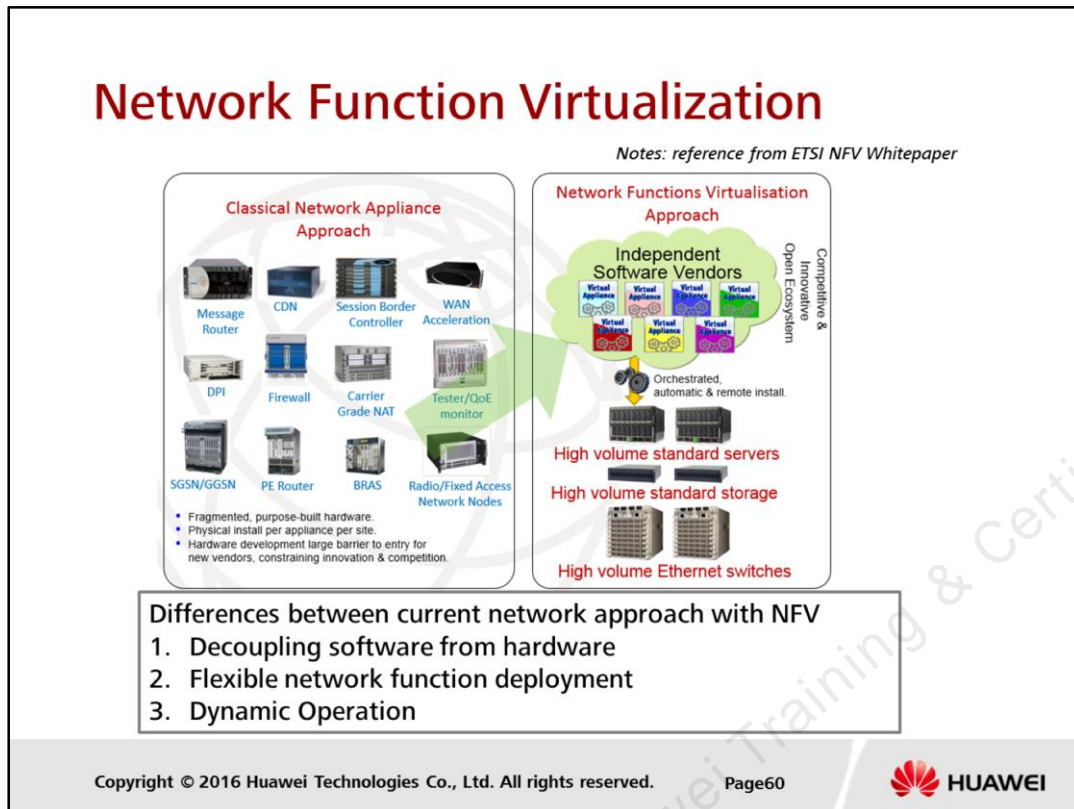
## Contents

### **6. SDN Related Concepts and Organizations**

#### **6.1 SDN and NFV**

#### 6.2 SDN Related Organizations

#### 6.3 ODL and ONOS



- Network Function Virtualization Offer a new network architecture concept that decouples the network functions, such as Network Address Translation (NAT), Firewall, Intrusion Detection, domain name services (DNS), SGSN/GGSN/IMS, to name a few, from proprietary hardware appliances so they run in software.
- It utilizes standard IT virtualization technologies that run on high-volume service, switch and storage hardware to virtualizes entire classes of network node functions into building blocks that may be connected, or chained, together to create communication services.
- Concept of NFV originated fro SDN
  - First ETSI white paper showed overlapping Venn Diagram, but it was removed in the second version of the white paper
- In non-virtualized network, NFs are implemented as a combination of vendor specific software and hardware, often referred to as network nodes or network elements. Network Function Virtualization represents a step forward for the diverse stakeholders in the telecommunication network environment. As such, NFV introduces a number of differences in the way network service provisioning is realized in comparison to current practice. In summary, these differences can be listed below:-

## NFV Basic Concepts

**VNF:** a virtualized network element, composed by VMs

vMME

vSGW

vPGW

vPCRF

vOCS

DNS

vHSS

**Network Service:** a group of VNFs that cooperate to deliver services

vEPC

VoLTE

M2M

**VNFD:** VNF descriptor, a set of files that describes the VM topology and resource requirements during the lifecycle of the VNF

VNF Template

VNF Deployment Plan

**NSD:** NS descriptor, a set of files that describes the types of VNFs, topology between VNFs and resource requirements during the lifecycle of the NS

NS Template

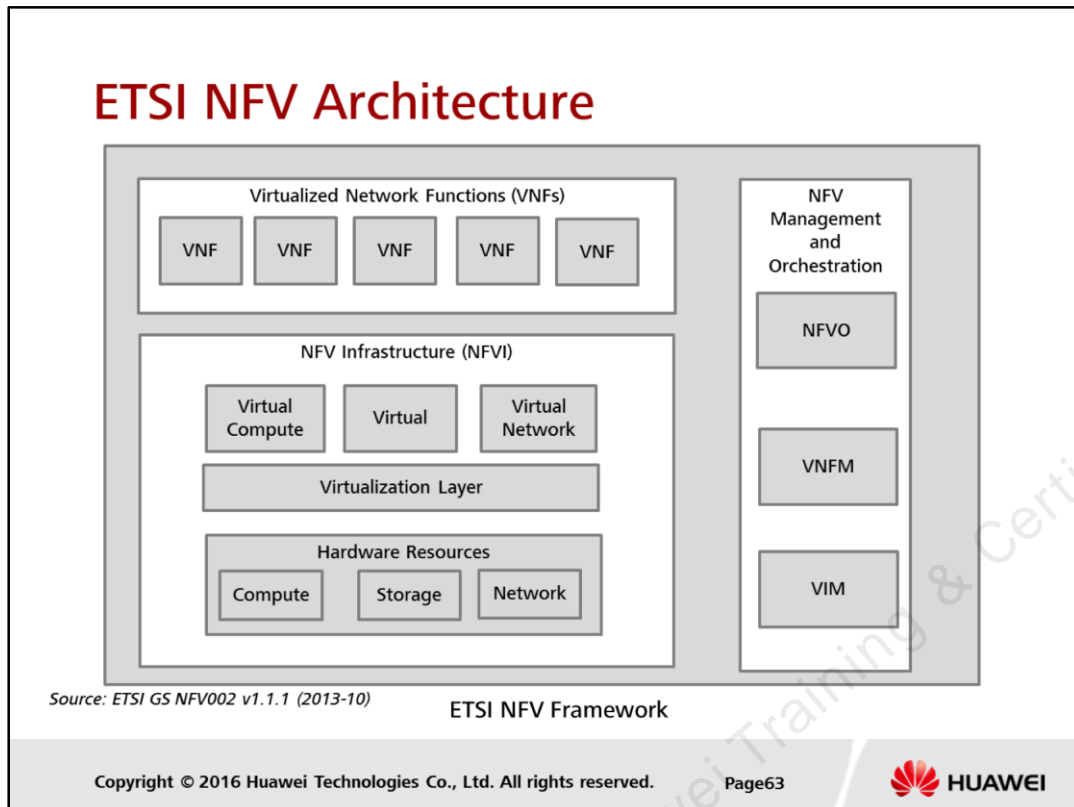
NS Deployment Plan

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page62

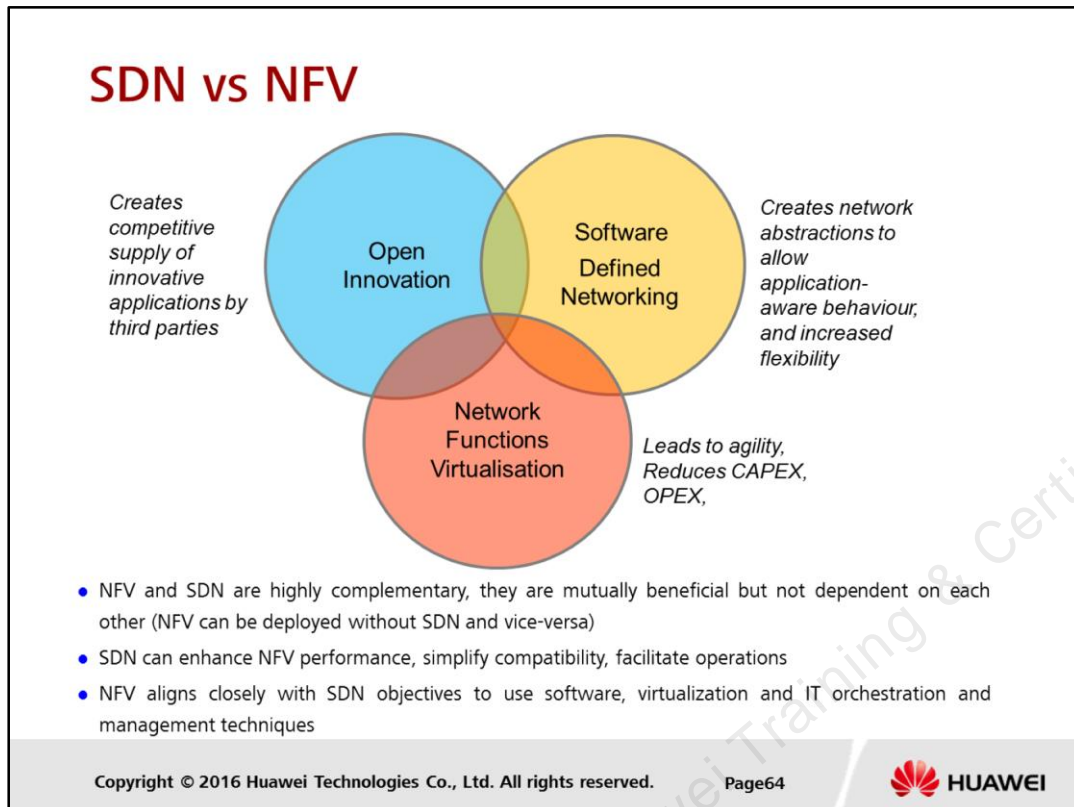
**HUAWEI**

- To understand about NFV, there are some basic terms and concepts we need to know, as shown in the slide above.

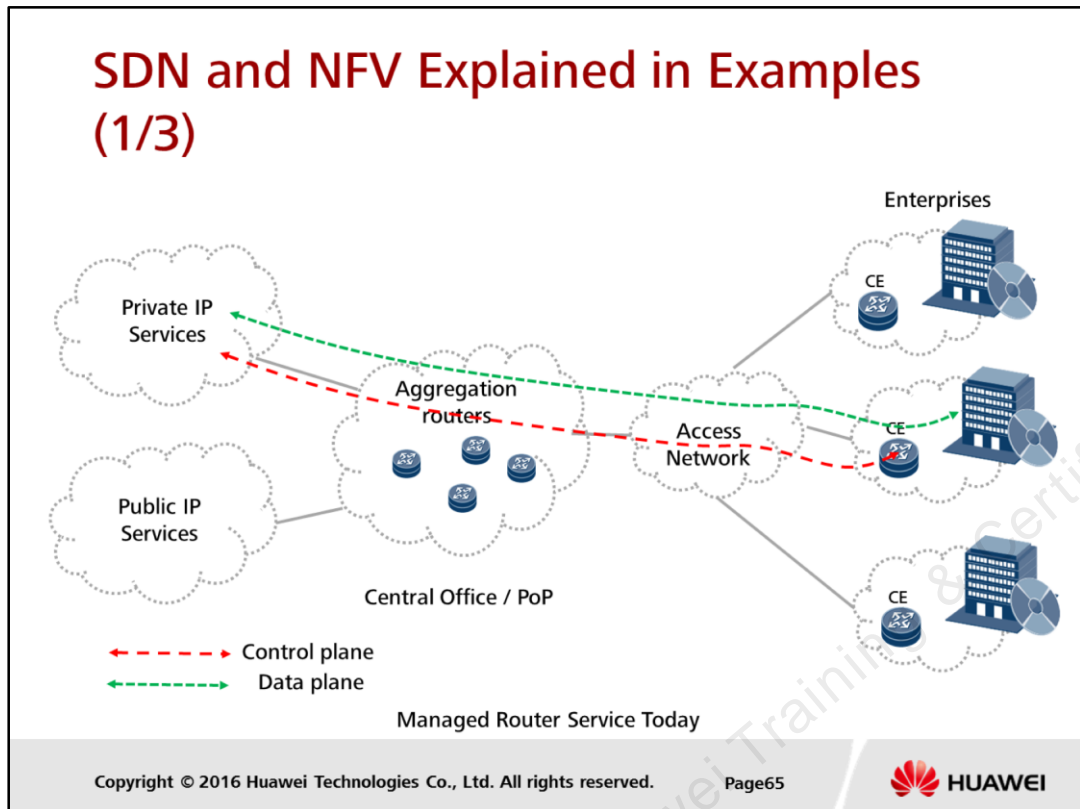




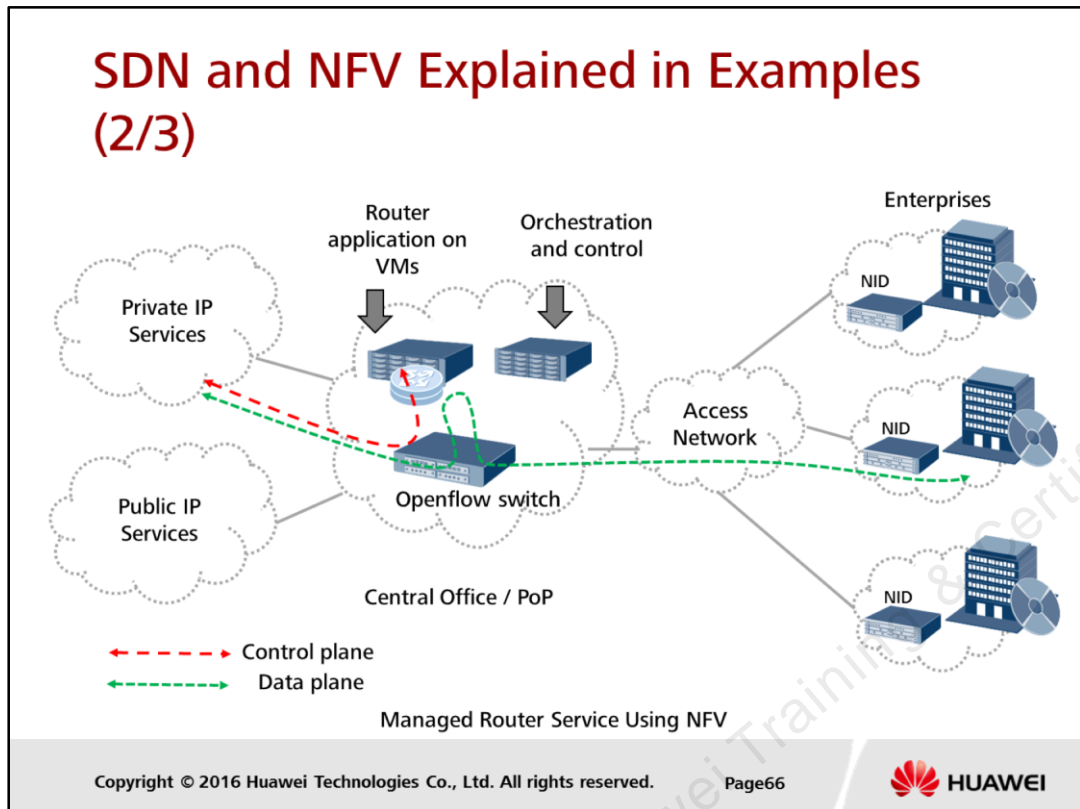
- Network Function Virtualization envisages the implementation of network function as software-only entities that run over the NFV Infrastructure. Diagram above illustrates the high level NFV framework. As such, three main working domains are identified in NFV
  - Virtualized Network Function (VNF) : the software implementation of a network function which is capable of running over NFVI
  - NFV Infrastructure (NFVI) : including the diversity of physical resources and how these can be virtualized. NFVI supports the execution of VNFs.
  - NFV Management and orchestration: covers the orchestration and lifecycle management of physical and/or software resources that support the infrastructure virtualization, and the lifecycle management of VNFs. NFV Management and orchestration focuses on all virtualization specific management task necessary in the NFV framework
- The NFV framework enables dynamic construction and management of VNF instances and relationships between them regarding data, control, management, dependencies and other attributes. To this end, there are at least three architectural views of VNFs that are centered around different perspective and contexts of a VNF. These perspectives include:-
  - A virtualization deployment/ on-boarding perspective where the context can be a VM
  - A vendor- developed software package perspective where the context can be several inter-connected VMs and a deployment template that describes their attributes
  - An operator perspective where the context can be operation and management of a VNF received in the form of a vendor software package.
- *p/s: Information above is quoted from ETSI GS NFV002 v1.1.1 (2013-10)*



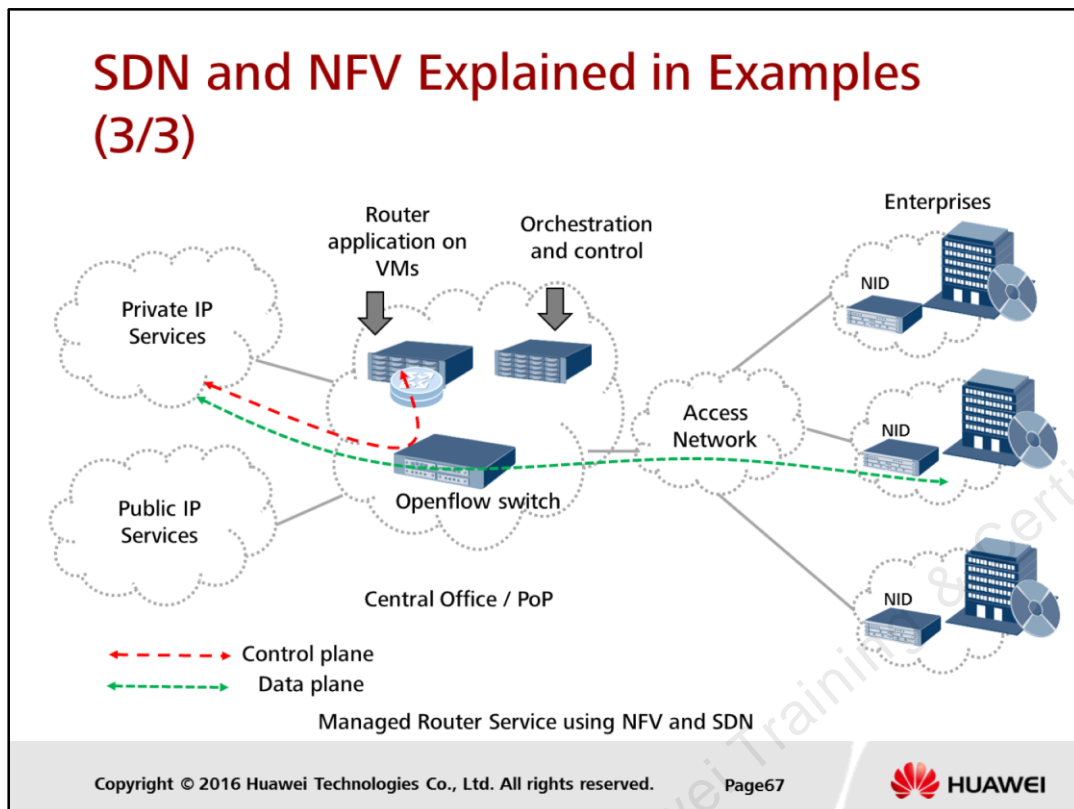
- NFV and SDN are highly complementary, they are mutually beneficial but not dependent on each other (NFV can be deployed without SDN and vice-versa)
- SDN can enhance NFV performance, simplify compatibility, facilitate operations
- NFV aligns closely with SDN objectives to use software, virtualization and IT orchestration and management techniques
- Network Functions Virtualization (NFV) is a network architecture concept that proposes using IT virtualization related technologies, to virtualize entire classes of network node functions into building blocks that may be connected, or chained, together to create communication services.
- SDN enables the forwarding layer to be defined through software. NFV enables the network device role and quantity to be defined through software.
- SDN and NFV both use cloud and Internet-related technologies to reconstruct carrier networks.
- **Conclusion: In Mobile Network, NFV is the major architecture for network evolution. In some certain scenarios, SDN shall be introduced.**



- To have a better understanding on the complementary concepts between SDN and NFV, we can use an example to see the different between the current network deployment, network deployment with only NFV, and network deployment using both NFV and SDN.
- Figure above shows the example of how a managed router service is implemented today in the current conventional network, using a router at the customer site. You can find out that both control plane and data plane traffic takes the same path.



- NFV is applied in the example of illustration above, where it is used in the situation by virtualizing the router function. On the customer side, a NID device, which is known as Network Interface Device is used for providing a point of demarcation, and in addition to measure network performance in the network.
- Compared to the previous example, router application is now applied on virtual machines which might be purchased from a cloud service providers. Control traffics is now controller by the router application on the VMs. However, you can see here, for the data plane, the traffic is still going router applications on VMs before going to the end user destination.



- Finally, SDN is introduced together with NFV in order to completely separate data plane from control plane. From the diagram shown above, you can see that the data packets are forwarded by an optimized data plane, while the routing function which is running on the control plane is running in a virtual machine running in a servers.
- The combination of SDN and NFV solution offers an optimal solutions with advantages listed below:-
  - A costly and dedicated appliance is now replace by generic or COTs hardware with advanced software
  - The software control plane is shifted from a dedicated platform to a better location which is more cost friendly, that is in a data center or POP
  - The control of data plane has been unified and standardized, allowing network and application evolution without the need of upgrading all the existing network devices.

## SDN vs NFV

SDN	Aspect	NFV
Separates control and data plane, centralized control and programmability	Basic Concept	Relocate network functions from dedicated appliances to generic servers
Campus, data center & cloud	Target Location	Service provider network
Commodity servers and switches	Target Devices	Commodity servers and switches
Cloud orchestration and networking	Initial Applications	Routers, firewalls, gateways, CDN, WAN accelerators, SLA assurance
OpenFlow	New Protocols	None
Open Networking Foundation (ONF)	Formalization	ETSI NFV Working Group

Comparisons between SDN and NFV

- Table above shows the comparisons between SDN and NFV.

## Contents





### **6. SDN Related Concepts and Organizations**


#### 6.1 SDN and NFV

#### **6.2 SDN Related Organizations**

#### 6.3 ODL and ONOS

## SDN Relevant Organizations (1/3)

Organization	Mission	Effort
	An industry consortium dedicated to the promotion and adoption of SDN through open standard development	OpenFlow
	The Internet's technical standards body. Produces RFCs and Internet standards.	Interface to routing systems (I2RS) Service function chaining
	An EU-sponsored standards organization that produces globally applicable standards for information and communications technologies	NFV Architecture
	United Nations agency that produces Recommendations with a view to standardizing telecommunications on a worldwide basis.	SDN functional requirements and architecture

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page70 






● ONF is a user-driven organization dedicated to the promotion and adoption of Software-Defined Networking (SDN) through open standards development. ONF emphasizes an open, collaborative development process that is driven from the end-user perspective. Our signature accomplishment to date is introducing the OpenFlow® Standard, which enables remote programming of the forwarding plane. The OpenFlow® Standard is the first SDN standard and a vital element of an open software-defined network architecture.

- IETF(Internet Engineering Task Force)

- Larger and open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. It is open to any interested individual.
- Compare to ONF, this stands much toward to networking vendor sides.
- In SDN environment, all networking vendors in IETF discussing on how to implement SDN at their existing hardware
- In early stages, there is two group formed under IETF : ForCES, Forwarding and Control Element Separation and ALTO, Application-layer traffic Optimization.
  - ForCES : involving in standardizing SDN requirement, framework, protocol, forwarding element, MIB etc.
  - ALTO mainly discussing on implementation of traffic optimization by providing more network information to application layer.

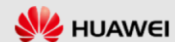


## SDN Relevant Organizations (2/3)

Organization	Mission	Effort
	Research group within IRTF. Produces SDN-related RFCs.	SDN Architecture
	Industry consortium that promotes the use of Ethernet for metropolitan and wide-area applications.	Defining APIs for service orchestration over SDN and NFV
	Industry consortium developing broadband packet networking specifications	Framework for SDN in telecommunications broadband networks
	An IEEE committee responsible for developing standards for LANs.	Standardize SDN capabilities on access networks.
	Industry consortium promoting development and deployment of interoperable networking solutions and services for optical networking products.	Requirements of transport networks in SDN architectures


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page72



- Table above shows some other related organizations that assists in SDN developments.

## SDN Relevant Organizations (3/3)

Organization	Mission	Effort
	Consortium of leading IT organizations developing interoperable solutions and services for cloud computing.	SDN Usage Model
	A standards organization that develops standards for the unified communications (UC) industry.	Operational opportunities and challenges of SDN/NFV programmable infrastructure
	An open source project focused on accelerating the evolution of NFV.	NFV Infrastructure
	An open source project focused on creating the most efficient DC hardware	Compute & Storage

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page73 

- Table above shows some other related organizations that assists in SDN developments.



## Contents

### **6. SDN Related Concepts and Organizations**

6.1 SDN and NFV

6.2 SDN Related Organizations

**6.3 ODL and ONOS**

## ODL and ONOS in a Glance



OpenDayLight <http://www.opendaylight.org/>

- On Apr 8, 2013, [The Linux Foundation](#), announced the founding of the OpenDaylight Project as a community-led and industry-supported open source framework to accelerate adoption, foster new innovation and create a more open and transparent approach to Software-Defined Networking (SDN) and Network Functions Virtualization (NFV).
- The project's founding members—[Arista Networks](#), [Big Switch Networks](#), [Brocade](#), [Cisco](#), [Citrix](#), [Ericsson](#), [HP](#), [IBM](#), [Juniper Networks](#), [Microsoft](#), [NEC](#), [Nuage Networks](#), [PLUMgrid](#), [Red Hat](#) and [VMware](#)—committed to donating software and engineering resources for OpenDaylight's open source framework to help define the future of an open source SDN platform.
- On Feb, 2014 First Release "Hydrogen"



ONOS <http://onosproject.org/>

- The Open Network Operating System (ONOS) is the first open source SDN network operating system targeted specifically at the Service Provider and mission critical networks.
- ONOS has created useful Northbound abstraction and APIs to enable easier application development and Southbound abstractions and interfaces to allow for control of OpenFlow-ready and legacy devices.
- ONOS has been developed in concert with leading service providers ([AT&T](#), [NTT](#)), with demanding network vendors ([Ciena](#), [E//](#), [Fujitsu](#), [Huawei](#), [Intel](#), [NEC](#)), R&E network operators ([Internet2](#), [CNIT](#), [CREATE-NET](#)), collaborators ([SRI](#), [Infoblox](#)), and with ONF to validate its architecture.
- First release "Avocet" was released on Dec, 2014.

 **Contents**

**6.3 ODL and ONOS**

**6.3.1 ODL**

**6.3.2 ONOS**

**6.3.3 ODL vs ONOS**

## ODL Project - Overview



- Hosted by the Linux Foundation, Open source SDN project featuring major vendors involvements.
- Focuses on having an open framework for building upon SDN/NFV innovations.
- Open to anyone, including end users and customers
- It is not only adopting OpenFlow as southbound interface but also including other interface protocols such as BGP, PCEP, OVS-DB, etc.
- Exposes open northbound interfaces, which are used by applications

## SDN Needs a Common Platform

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page77

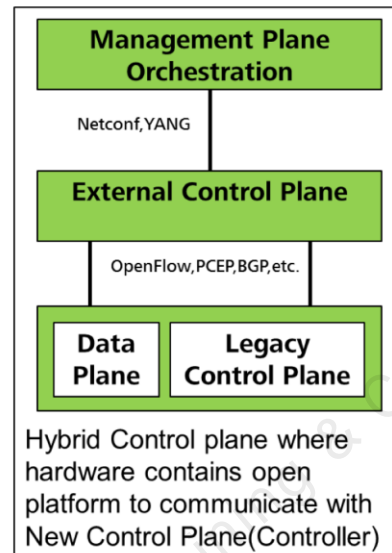


- OpenDaylight Project is a collaborative open source project hosted by The Linux Foundation. The goal of the project is to accelerate the adoption of software-defined networking (SDN) and create a solid foundation for Network Function Virtualization (NFV). The software is written in Java.
- OpenDaylight officially started on April 8<sup>th</sup> 2013.
- Southbound interface is not limited to OpenFlow innovations, but in fact decoupled from it allowing the two to evolve independently.

## ODL Project – The SDN Position in ODL



- SDN is a new approach to separate control plane and data plane to realize a centralized control plane
- Hybrid Approach is supported in ODL
- Multiple southbound interface protocols such as OpenFlow, BGP-LS, PCEP, etc are supported.
- Supports Open Northbound interface such as YANG modeled Netconf



- Software-Defined Networking (SDN) is an industry movement for building programmable networks that are flexible and responsive to organizations' and users' needs. OpenDaylight, the largest open source SDN controller, is helping lead this transition. By uniting the industry around a common SDN platform, the OpenDaylight community -- solution providers, individual developers, and users working together -- is delivering interoperable, programmable networks to service providers, enterprises, universities and a variety of organizations around the globe. *(Quoted from [www.opendaylight.org](http://www.opendaylight.org))*

## ODL Project – The Objectives

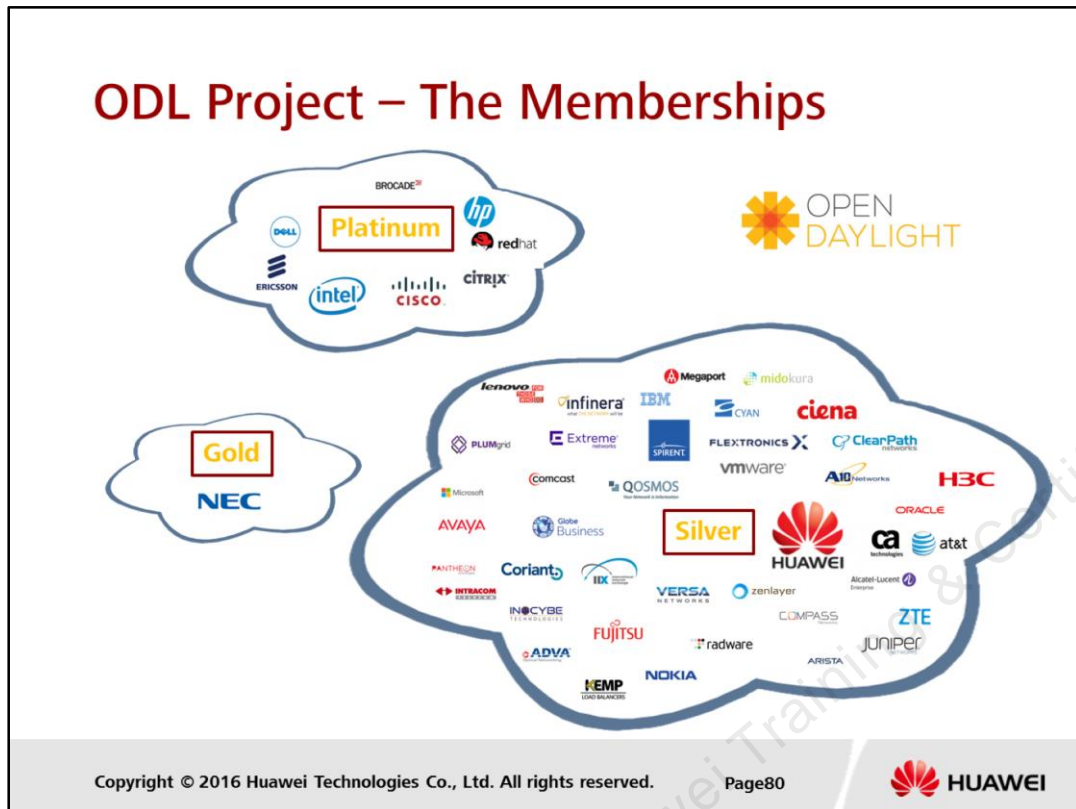
Code	Acceptance	Community
To create a robust, extensible, open source code base that covers the major common components required to build an SDN solution	To get broad industry acceptance amongst vendors and users <ul style="list-style-type: none"> <li>• using OpenDaylight code directly or through vendor products</li> <li>• Vendors using OpenDaylight code as part of commercial products</li> </ul>	To have a thriving and growing technical community contributing to the code base, using the code in commercial products, and adding value above, below and around.

**Open, Transparent, Fair**

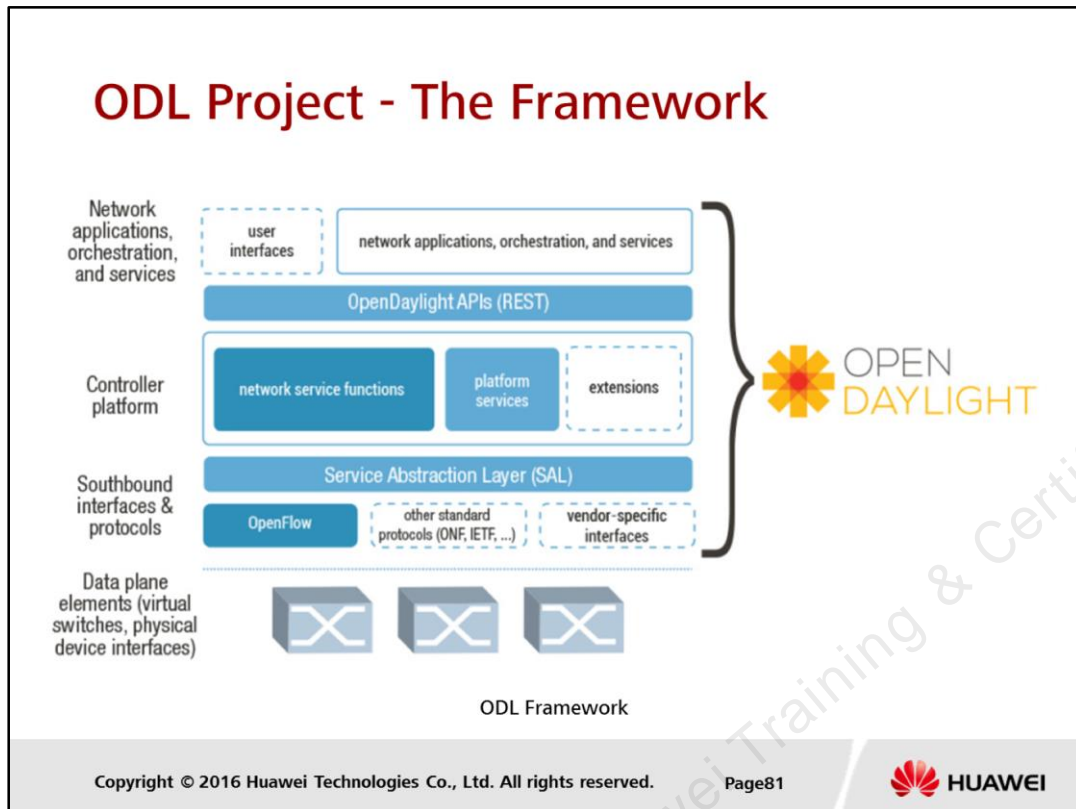


- The benefits of OpenDaylight can speed up of service provision and deploy SDN in operator network with low OpEx.
- Customers can participate and gain access new technologies more quickly without changing current structure of networks.
- Other than that, It also enable faster innovation by vendors.

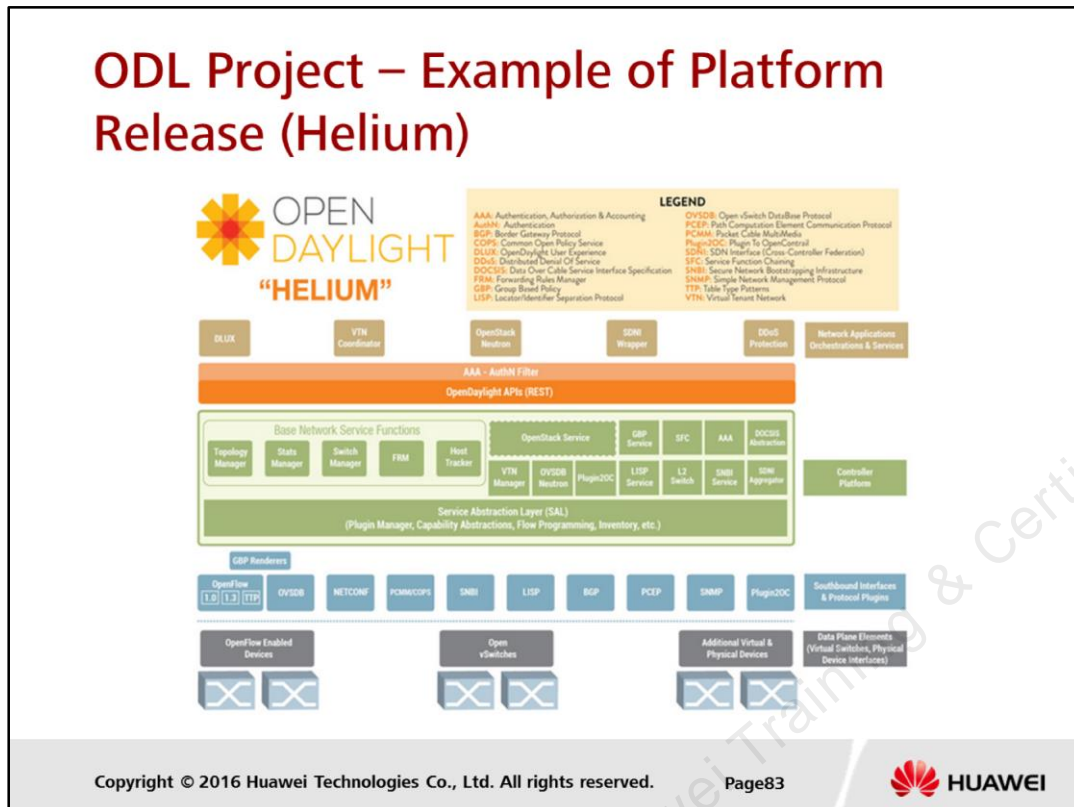




- Above shown member as Feb 2, 2016
- In terms of member support, grown from 18 to over 50 members today including Comcast etc. and Continue to see widespread interest and support from the industry



- The Open Daylight Controller is a pure software and as a JVM it can be run on any OS and Metal as long as it supports Java.
- The OpenDaylight controller platform is designed as a highly modular and plugin based middleware that serves various network applications in a variety of use-cases. The modularity is achieved through the Java OSGi framework. The controller consists of many Java OSGi bundles that work together to provide the required controller functionalities.



- The first software code release for the OpenDaylight Controller is Hydrogen. It was the first simultaneous release of OpenDaylight, and features three different editions to help users get started: the Base Edition, the Virtualization Edition, and the Service Provider Edition. The three types of the software ensure a wide array of users can implement Hydrogen.
- The second code release for OpenDaylight Controllers is Helium. It features a new user interface, and a more simplified and customizable installation process, due to the use of the Apache\_Karaf container. This code release also has deeper integration with OpenStack, including improvements in the Open\_vSwitch Database Integration project, as well as other features like Security Groups, Distributed Virtual Router, and Load Balancing-as-a-Service.
- Currently, the OpenDaylight Project is working on the third software release, Lithium, set for a summer 2015 release.



## Contents

### 6.3 ODL and ONOS

#### 6.3.1 ODL

#### 6.3.2 ONOS

#### 6.3.3 ODL vs ONOS

## ONOS project - Overview

- Is introduced by ON.Lab
- ONOS requirements are mainly from **Carrier** targeting carrier networks due to its policy-driven network programmability and operator friendly characteristic.
- Intent-based networking
- Based on distributed controller platform, which is designed specifically for scalability and high-availability
- Enable Web style agility
- Open Platform, with no vendor lock-in risk

ON.LAB



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page85



- The Open Networking Lab (ON.Lab), a non-profit organization founded by SDN inventors and leaders from Stanford University and UC Berkeley to foster an open source community for developing tools and platforms to realize the full potential of SDN, today (Dec 5,2014) introduced the open source SDN Open Network Operating System (ONOS).

## ONOS Project – The Memberships

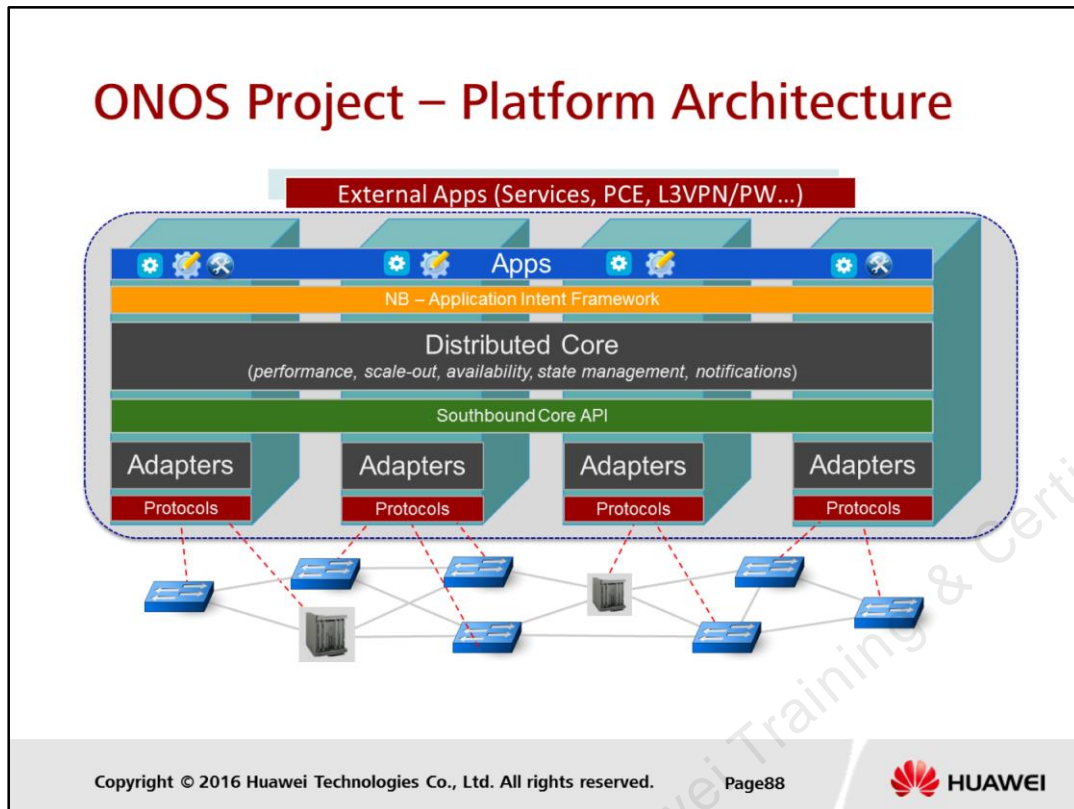


- **Partner**
  - Alcatel Lucent, AT&T, China Unicom, Ciena, Cisco, Ericsson, Fujitsu, Huawei, Intel, NEC, NTT Communications, SK Telekom, Verizon.
- **Collaborator**
  - AARNET, Adara, Airhop Communications, Akamai, AmLight, BlackDuck, BTI Systems, Beijing University of Posts and Telecommunications, Cavium, ClearPath Networks, CNIT, CREATE-NET, Criterion Networks, CSIRO, ECI Telecom, ETRI, Consortium GARR, GEANT, Happiest Mind, Internet2, KAIST, KREONET, KISTI, NAIM Networks, NetCracker, OpenFlow Korea, Oplink Communications, Open Networking Foundation, Postech, Radisys, SRI International

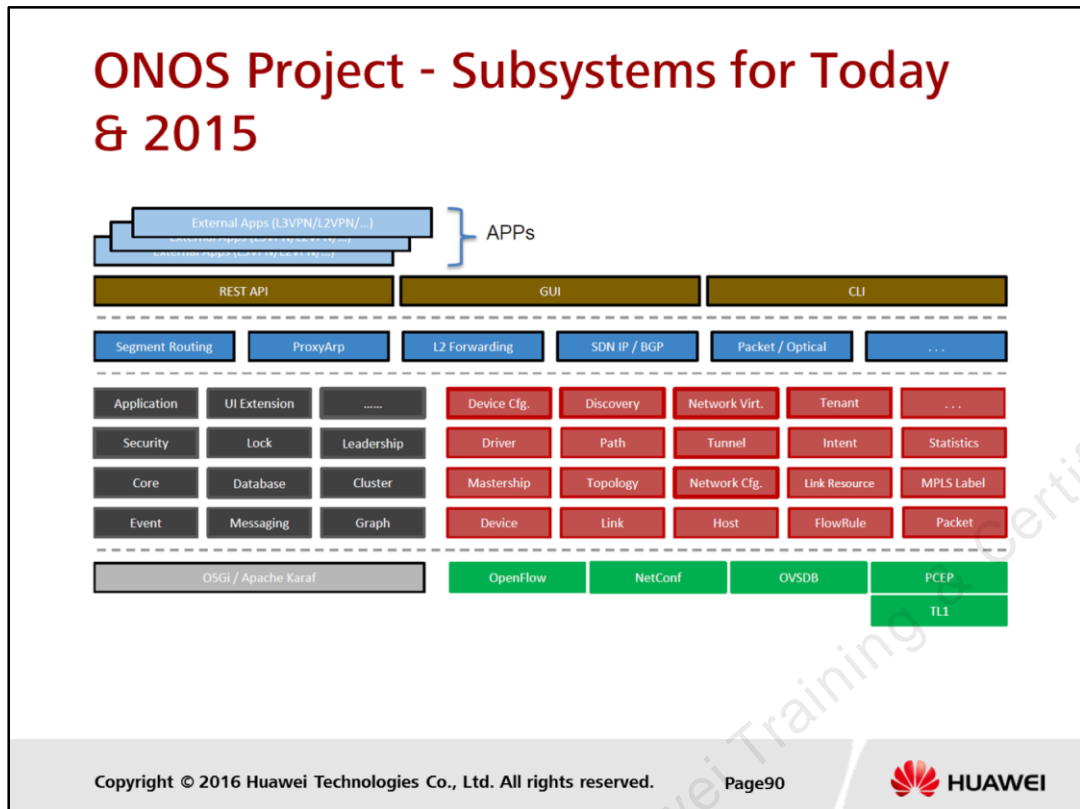
Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page87



- Here is the community for ONOS Community, it is included a set of leading service provider, leading vendors, and a large group of collaborating organizations, and the larger community.
- Huawei is one of the nine ONOS founding members and plays a crucial role in development of architecture and eco-system.



- High Availability, Scale-out, and Performance are the fundamental , as are powerful abstractions at the Northbound and Southbound interfaces.
- Layer Description:
  - Distribute Core
    - Deployed as a service on a cluster of server, and the same ONOS software runs on each server.
    - Provides scalability, high availability and performance
    - Bring carrier grade features to the SDN control plane. The ability of ONOS to run as a cluster is one way that ONOS brings web style agility to the SDN control plane and to service provider.



- There are numerous features or resources supported for operator Network
  - Multi-layer SDN control of packet-optical core
    - Optical Topology
    - Layer 2 Topology
    - Layer 3 Topology
    - Overlay Tunnel as link
  - Standardized Southbound interface and flexible plugins
    - Compatible and supported many protocols such as BGP-LS/PCEP/Openflow
    - Supported data modeling protocols such as OpenFlow and Netconf
  - Multi-layer network resource unified management
    - Interface resource management
    - Lamda wavelength management
    - MPLS nodes/Globally label management
    - VLAN resource management
    - IP address management
  - Network performance statistic and fault management
    - Statistic
    - SLA monitor deployment
  - Support interoperation between non-SDN network and SDN network during SDN migration from traditional network
- Notes: Reference from





## Contents









### 6.3 ODL and ONOS

#### 6.3.1 ODL

#### 6.3.2 ONOS



#### 6.3.3 ODL vs ONOS

## ODL vs ONOS (1/2)

ODL 	Aspect	ONOS 
	Organization	
 	Founder	 
Primary focus on NBI/SBI and DC scenarios; Maintain control plane on legacy devices	Controller Ideology	Primary focus on Core components and WAN scenarios; Vendor has to follow up unified SBI framework/abstraction
Support multiple NBI for network function and application	Northbound Framework	Lack of NBI, no commercial use till now. Network abstraction: Support global view on Topology/Network status/Resource

- Table above shows the comparisons between ODL and ONOS.

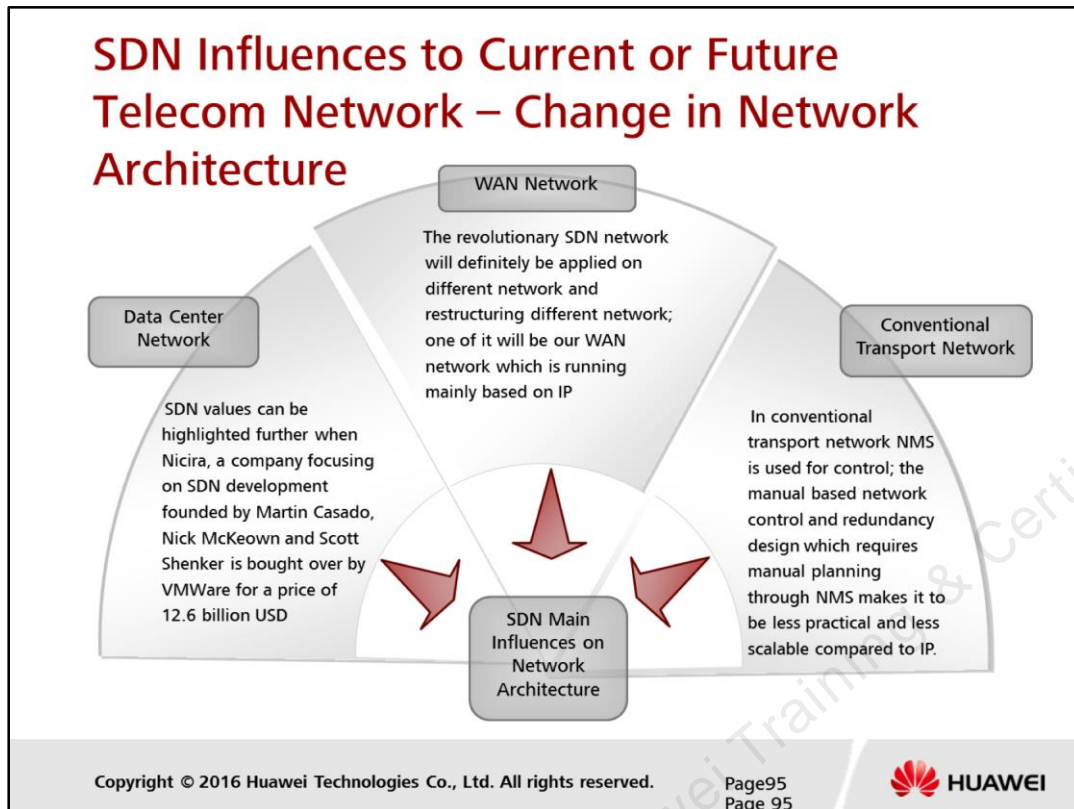
## ODL vs ONOS(Cont.) (2/2)

ODL 	Aspect	ONOS 
Network abstraction: MD-SAL; Rich SBI: OpenFlow, Netconf, BGP/PCEP, LISP, OpFlex	Southbound Framework	Unified SBI abstraction to support OpenFlow and a diversity of interfaces and devices
41 Vendors ( includes Brocade, Cisco, Citrix, Ericsson, HP, IBM, Juniper Microsoft etc).	Major Contributors	Research: Stanford, UC Berkeley Vendor: Intel, TI, Huawei, E///, HP, Ciena, Cisco, NEC, Vm ware Operator: Docomo, Google, Geni, CableLabs
Founded on Feb, 2013. Announced on Apr.2013. First Release "Hydrogen" on Feb,2014. Second Release "Helium" on Oct,2014 Third Release "Lithium" is on roadmap. Amount of Code: 1.8M	Achievement and Progress	First release "Avocet" was released on Dec, 2014. Second release "Blackbird" will be released on Feb, 2015. Amount of Code: around 100K

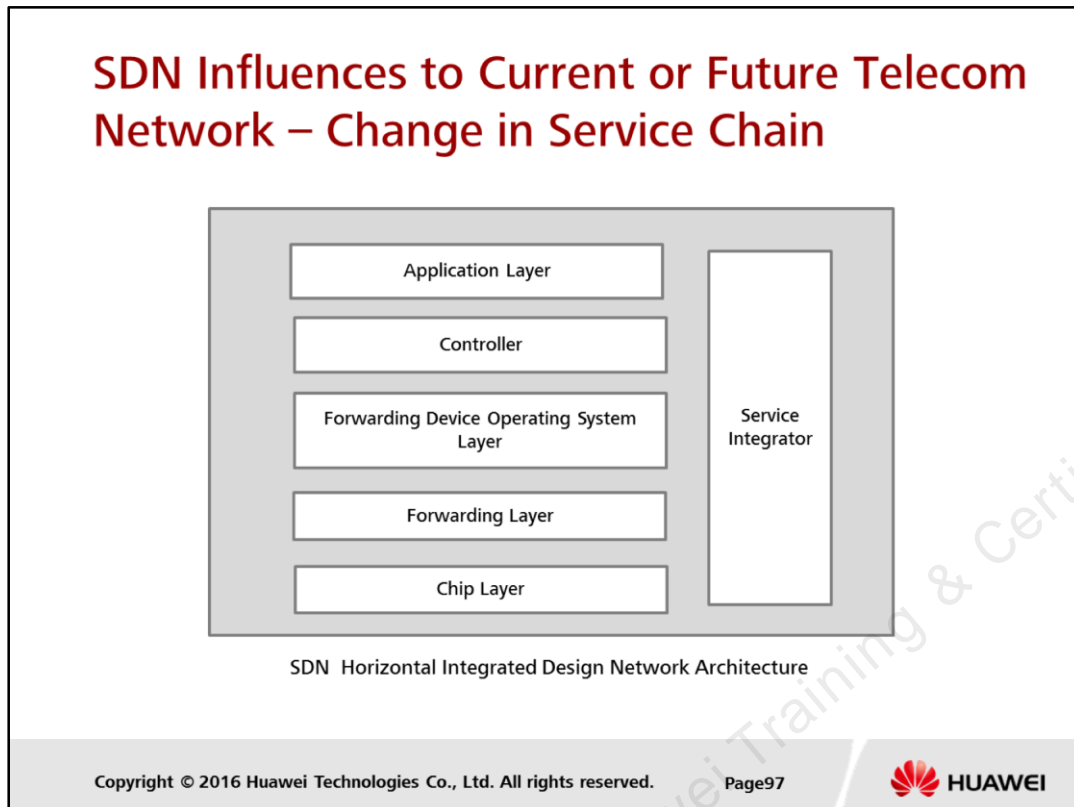
- Table above shows the comparisons between ODL and ONOS.

## Contents

1. Traditional Network Limitations
2. SDN Overview and History
3. SDN Network Architecture
4. SDN Value Proposition
5. SDN Challenges and Solutions
6. SDN Related Concepts and Organizations
7. **SDN Influences to Current or Future Telecom Network**



- SDN is able bring a wave of network transformation to the current traditional network due to its 3 main characteristics, which are open programmability, control and data plane separation and centralized control.
- The traditional network is exiting and developing for more than 30 years; almost all requirements have been met and achieved through distributed networking in traditional network. Thus, we cannot say that SDN solves the unsolved problems of traditional network; however, SDN is able to solve those issues in a much simpler and easier attempt which makes SDN to be more welcoming compared to traditional dedicated network design. All these advantages brought by SDN are actually realized by the powerful open programmability available in SDN; this allows SDN to be able to modify and alter a network just like how IT can achieve in fast network service provisioning and modification. To be more exact, SDN is not introducing any new feature or function, but it is restructuring the whole network architecture to be simpler.



- The emergence of the evolutionary SDN network is going to transform the vertical networking architecture into horizontal networking architecture. This also causes the COTS hardware trend becoming more and more popular in conjunction of SDN development. In other words, the traditional network equipment providers will be affected in this case. It is definitely that the traditional big vendors such as Cisco or Huawei will invest in the SDN infrastructure development based on COTS trend, there are some newly established startups are interested to be part of this evolution too.
- Other than Huawei is putting a lot of efforts in SDN research and development, Cisco too develops multiple SDN solutions such as ONEPK, WAE, ACI, XNC etc; HP and IBM establishes ODL and gained interests from different vendors too. All traditional equipment vendors and newly established startups are in the midst of finding their own positions in this new service chain; for instance, some chips manufacturers are putting efforts in developing Openflow components to lead the markets; Huawei, Cisco and others are developing controllers; besides, as SDN architecture is going to impact on OSS companies too, companies like HP and IBM too involves actively in SDN development progress.
- The new service chain developed from vertical integration design into horizontal design might be consisting of several layers as shown in the diagram above.
  - For chip layer, there will be a lot of chip manufacturing companies competing to develop chips running on Openflow protocols.
  - Forwarding layer is referring to forwarders which is referring to COTS devices now.
  - Forwarding device operating system layer is referring to the software kernel running on forwarders; for instance, Linux developed Cumulus; Conventional vendors have their own traditional network OS, for example, Huawei VRP, Cisco IOS, Juniper Junos etc.
  - Application layer is referring to network service realization such as Openstack etc.



## Summary

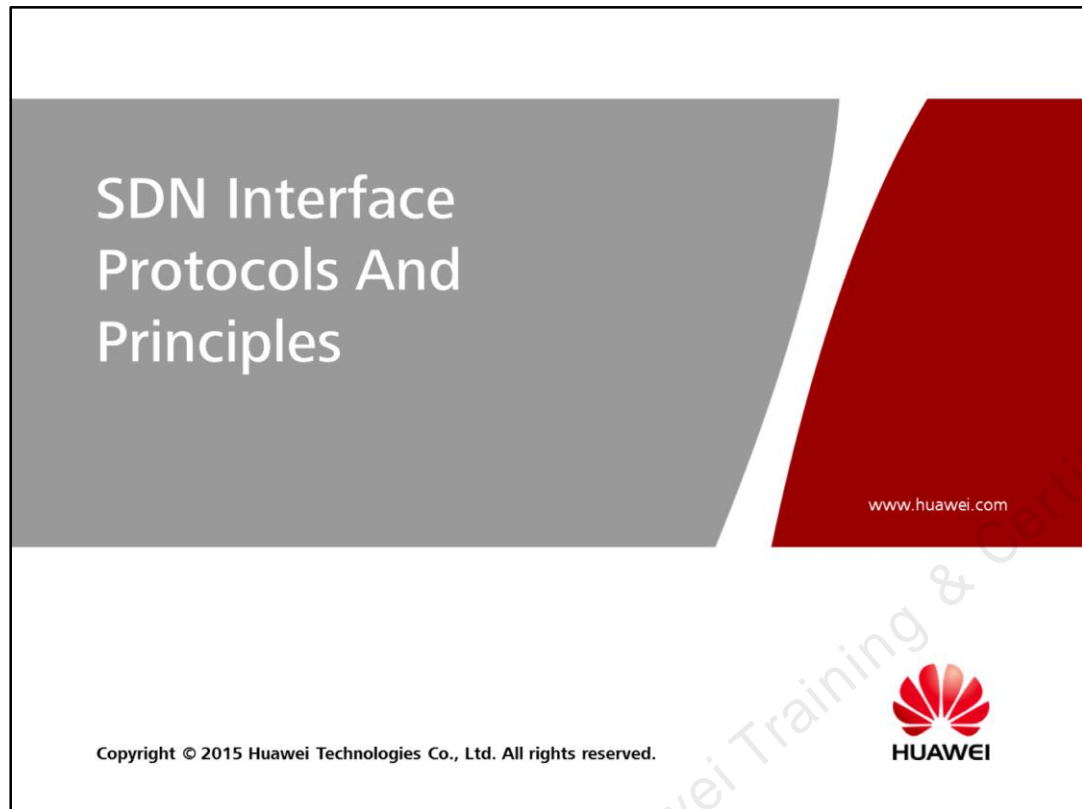
- Discuss traditional network limitations and how SDN is able to resolve them
- Discuss SDN basic concepts and working principles
- Discuss SDN values from both technical and operator perspectives.
- Discuss challenges faced by SDN and how to solve them
- Discuss SDN related organizations
- Discuss differences between SDN and NFV
- Discuss SDN influences in the current and future telecommunication network

**Thank you**

[www.huawei.com](http://www.huawei.com)

Huawei Training & Certification Huawei Training & Certification






SDN Interface  
Protocols And  
Principles

[www.huawei.com](http://www.huawei.com)

Copyright © 2015 Huawei Technologies Co., Ltd. All rights reserved.



HUAWEI

The slide features a white background with a large grey shape on the left and a red shape on the right. The title 'SDN Interface Protocols And Principles' is centered in the grey area. The website 'www.huawei.com' is in the red area. The copyright notice and Huawei logo are at the bottom.



## Objectives

- Upon completion of this course, you will be able to:
  - Describe SDN Openflow protocol
  - Describe SDN NETCONF protocol
  - Describe SDN SNMP protocol
  - Describe SDN RESTful protocol
  - Describe SDN NETSTREAM protocol

 **Contents**

1. Introduction of SDN OpenFlow Protocol
2. Introduction of SDN Netconf Protocol
3. Introduction of SDN SNMP Protocol
4. Introduction of SDN RESTful Service
5. Introduction of SDN Netstream Protocol



## Contents

- 1. Introduction of SDN OpenFlow Protocol**
2. Introduction of SDN Netconf Protocol
3. Introduction of SDN RESTful Protocol
4. Introduction of SDN SNMP Protocol
5. Introduction of SDN Netstream Protocol



## Contents

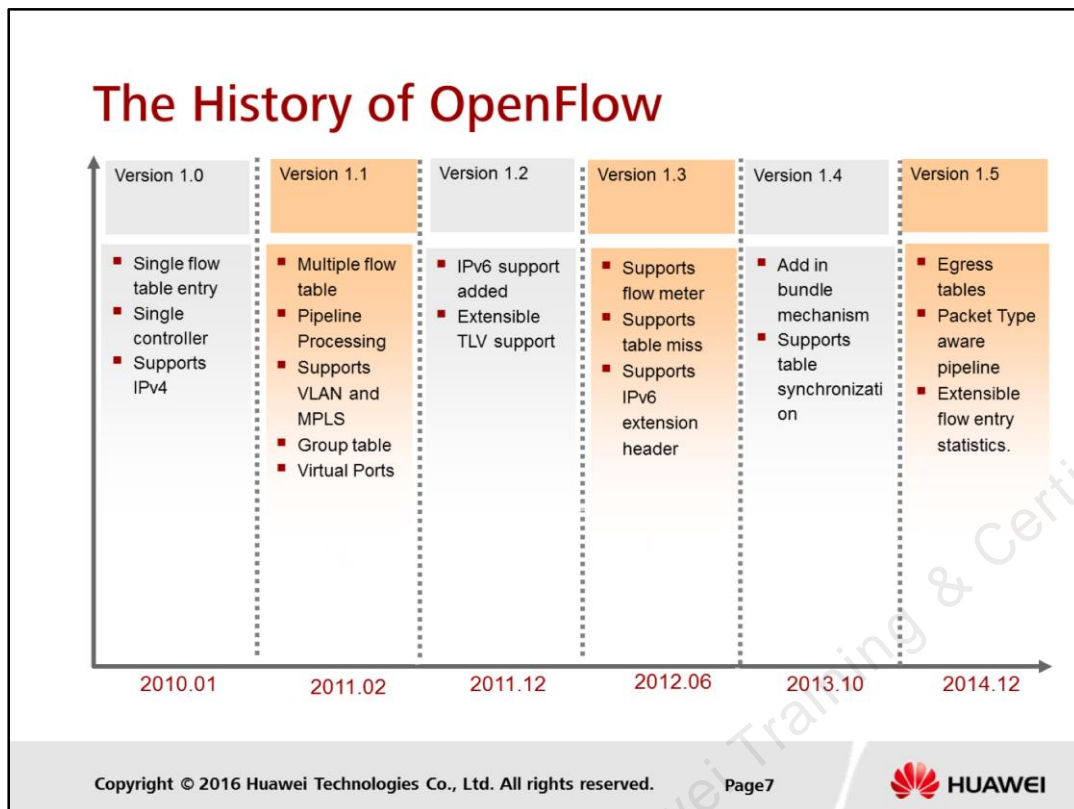
1. Introduction of SDN OpenFlow Protocol
  - 1.1 **OpenFlow History and Version Evolution**
  - 1.2 OpenFlow Basic Architecture
  - 1.3 OpenFlow Table Description
  - 1.4 OpenFlow Working Principle
  - 1.5 OpenFlow Application in SDN

## The History of OpenFlow

OpenFlow in SDN Framework

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page6

- OpenFlow protocol is the first standard southbound interface protocol used in standard SDN network architecture, which is considered as a vital element in SDN network.
- The OpenFlow protocol was first originated from the Stanford University Clean Slate Research Project, led by Nick McKeown, the professor of Stanford University, with the aim to “reinvent the internet” to overcome the current internet limitations and improve the network service innovation.
- OpenFlow protocol was first proposed by Nick McKeown in April 2008 on his published thesis “OpenFlow: enabling innovation in Campus Network”.
- OpenFlow protocol undergoes rapid growth and development after the Open Network Foundation (ONF) was founded by McKeown and Scott Schenker in March 2011, due to its advantages of high programmability and innovation.
- In the recent years, different versions of OpenFlow protocols have been released with different features and functions.



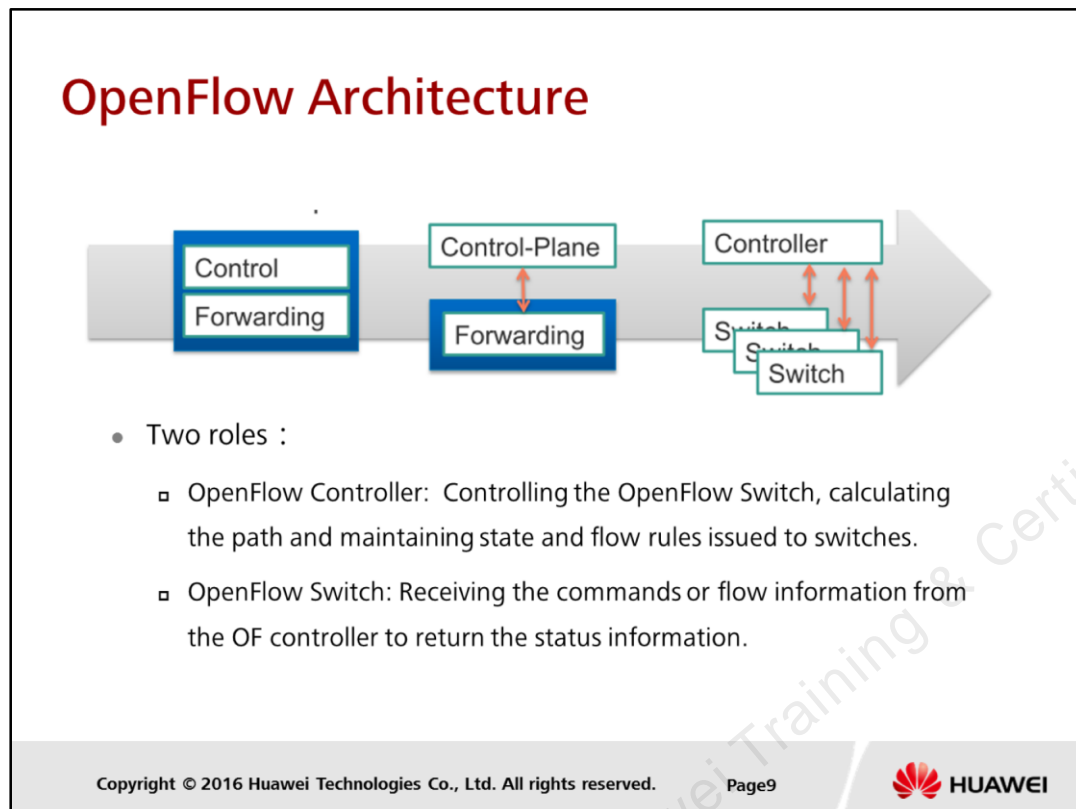
- Since OpenFlow was first introduced, different versions of OpenFlow has been introduced with different features and enhancements added from version 1.0 in year 2010, until version 1.5 in year 2014.
- The first version of OpenFlow is the only version which supports only single flow table entry. Version afterwards can supports multiple flow table.
- Features and additional functions have been added in conjunctions with different versions, and the different features added are summarized on the figure shown above.



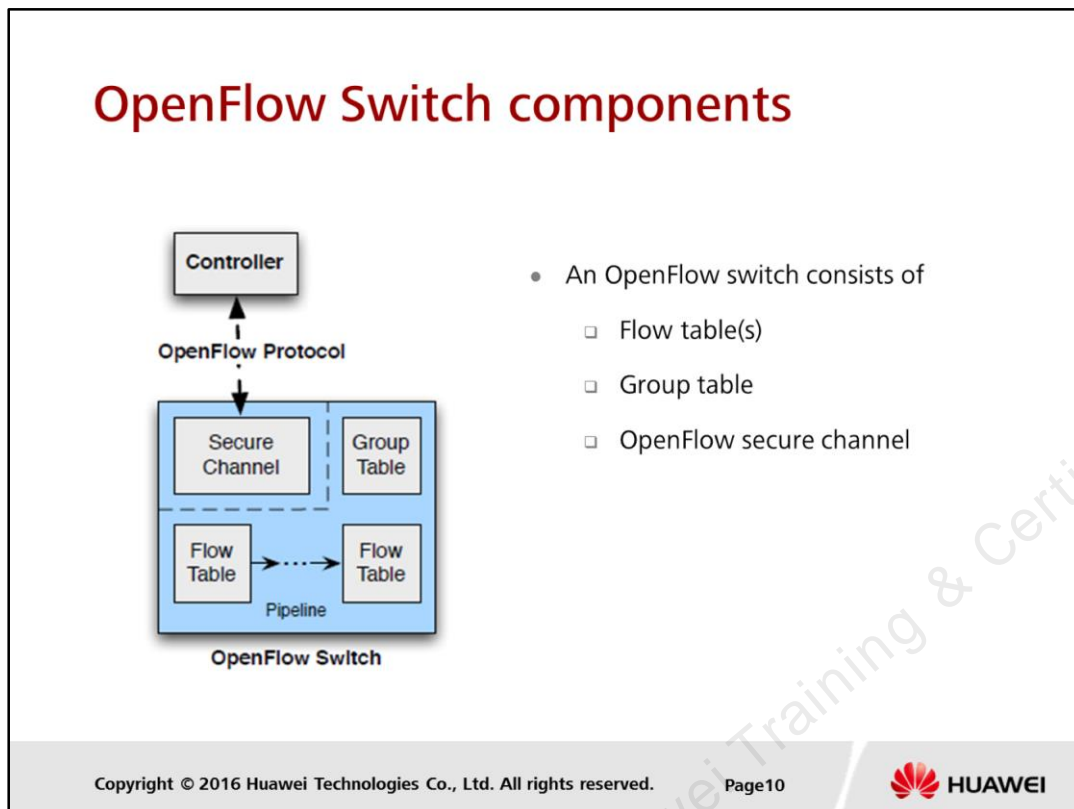
## Contents

1. Introduction of SDN OpenFlow Protocol
  - 1.1 OpenFlow History and Version Evolution
  - 1.2 OpenFlow Basic Architecture**
  - 1.3 OpenFlow Table Description
  - 1.4 OpenFlow Working Principle
  - 1.5 OpenFlow Applications





- One of the important characteristics of SDN architectures is the separation of control plane and forwarding plane. In conjunction with this characteristics, two hardware roles have been introduced in OpenFlow architecture, which are the OpenFlow controller and OpenFlow Switch.
- OpenFlow controller, serves as the control unit to perform path calculation and distributes OpenFlow entries to the forwarder which is responsible of performing packet forwarding based on the received flow entries. Huawei controller is called SNC, Smart Network Controller.
- OpenFlow switch is the device which receives the command or flow table information from controller to return the status information. Based on the flow entries information generated and sent from controller, OpenFlow switch serves as the forwarder to perform data packet physical forwarding.



- A flow table in an OpenFlow switch consists of a sets of flow table entries which contain multiple information such as match fields, counters and a sets of instructions to be applied for matching packets.
- There might be multiple flow tables existing in an OpenFlow switch and the flow tables will be handled through pipeline processing. For example, when a packet is forwarded and matched with the flow entry in the first flow table, the corresponding instruction will be associated with pipeline processing instruction which allows packets to be sent to subsequent tables for further processing.
- The table pipeline processing stops if the instructios associated wit a matching flow entry does not specify the following flow table.
- The group table contain a list of group entries and each group entry is associated with action buckets or action groups; Packets defined in group and matched with the group entries will be matched with one or more action buckets.
- The secure channel is the component used to connect to external controller and establishes communication channel through openFlow protocol.

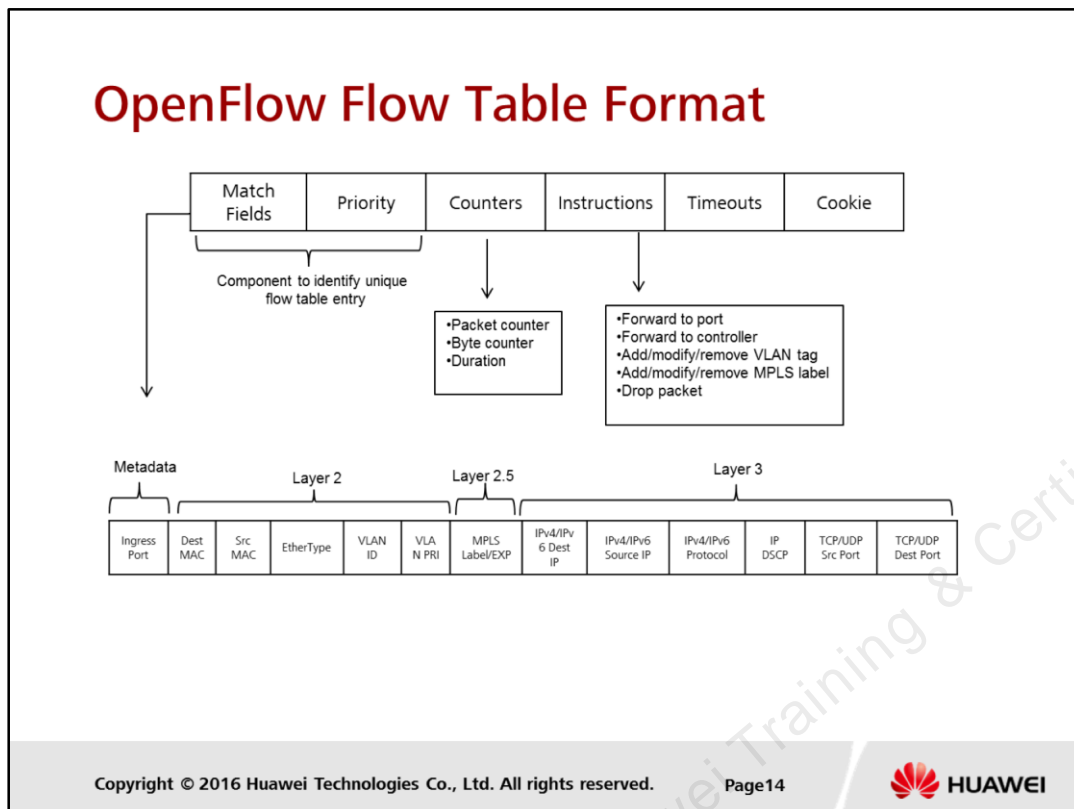
## OpenFlow Ports

- An OpenFlow switch must be able to support 3 types of OpenFlow standard ports:-
  - **Physical ports**
    - One to one hardware interface on switch
    - In virtualized switch hardware, physical ports are referred to a virtual slice of the hardware interface on the switch.
  - **Logical ports**
    - Higher level ports which do not associate directly with the physical interfaces and maybe realized through technologies such as link aggregation groups, tunnels or loopback interface.
  - **Reserved ports**
    - Defined by OpenFlow specification to specify forwarding actions, for instance, sending to controller, flooding, or forwarding without using OpenFlow method etc.
    - An OpenFlow switch must be able to supports the port functions ALL, CONTROLLER, TABLE, IN\_PORT, and ANY. However, there are some other optional port functions such as LOCAL, NORMAL and FLOOD which can be optionally supported.

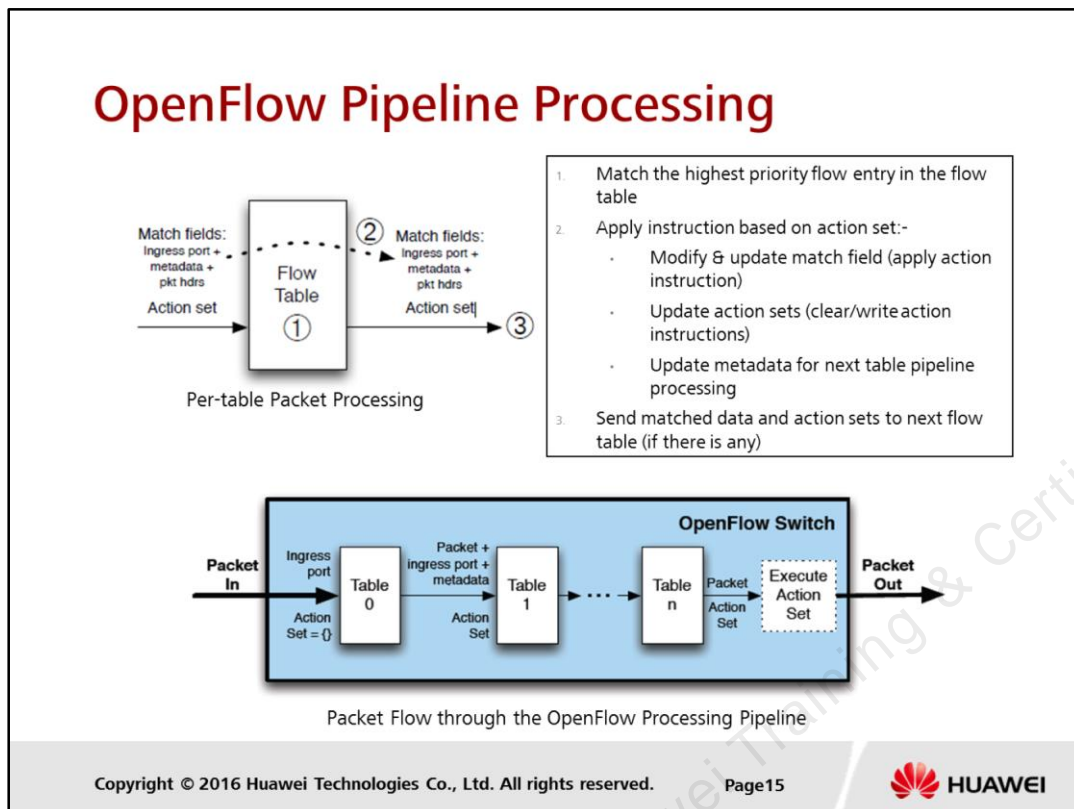
- OpenFlow ports are used to connect an OpenFlow switch to another OpenFlow switch and also connect to external network.
- OpenFlow switch connect to each other logically via OpenFlow ports. OpenFlow packets are received from the ingress port and the packet be processed and matched through pipeline processing, and may be forwarded out to an Output port.
- The OpenFlow reserved ports are also considered as OpenFlow specification defined ports, used to specify forwarding actions once a flow entry is matched.
- The compulsory required reserved ports are listed below:-
  - **ALL:** Represents every port on switch can be used to forward packet and work as output ports.
  - **CONTROLLER:** Represents the port used for communication channel between switch and controller; it can work as a ingress or output port.
  - **TABLE:** Represents the start of the OpenFlow pipeline processing, to submit the received packets to the first flow table for pipeline processing; it can work as a output port.
  - **IN\_PORT:** Represents the packet ingress port and work as output port to forward packets out from the ingress port where it receives the packet.
  - **ANY:** A special port where no handling action will be done; it can neither be an ingress or output port.

 **Contents**

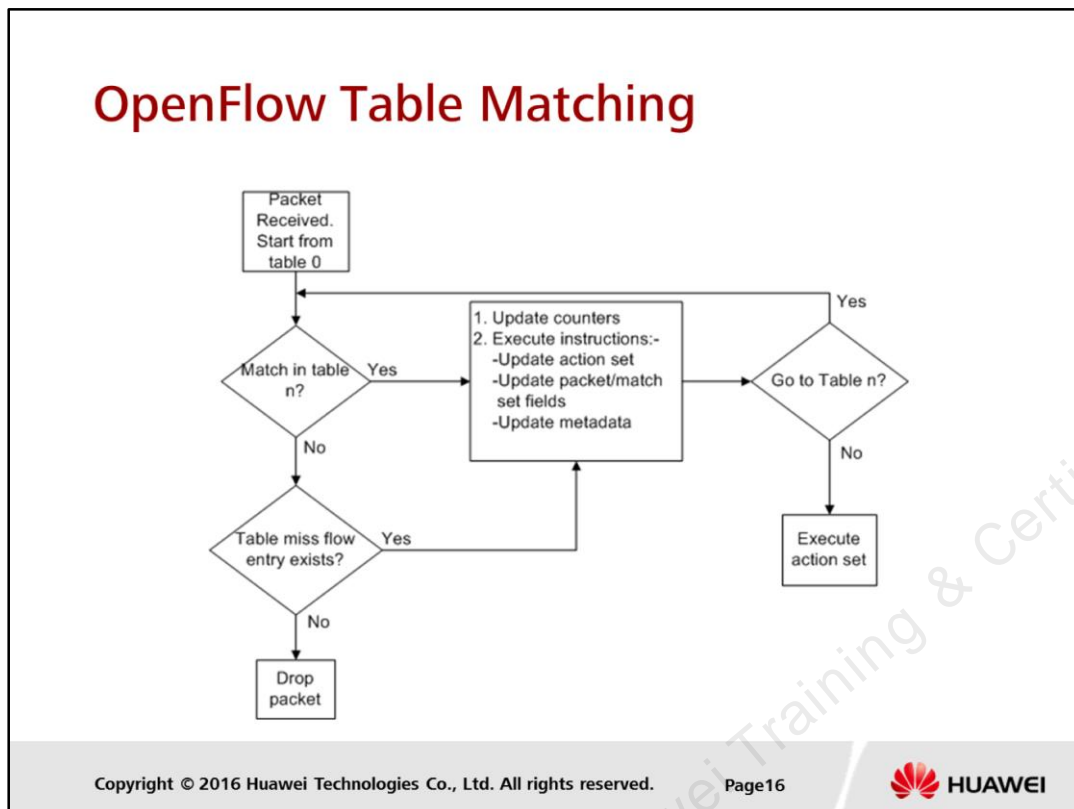
1. Introduction of SDN OpenFlow Protocol
  - 1.1 OpenFlow History and Version Evolution
  - 1.2 OpenFlow Basic Architecture
  - 1.3 OpenFlow Table Description**
  - 1.4 OpenFlow Working Principle
  - 1.5 OpenFlow Application in SDN



- OpenFlow 1.4 defines 3 types of table, including flow table, group table and meter table. Each type of table will be discussed in this section.
- An OpenFlow flow table is the fundamental element in Openflow protocol. Each flow table entry contains elements below:-
  - Match fields: for packet matching which consists of ingress ports and optional matched fields or metadata specified from the previous table
  - Priority: is defined the precedence of flow table entry
  - Counters: updated per packet matching
  - Instructions: to modify action sets or pipeline processing
  - Timeouts: maximum time for to define a flow expiration
  - Cookie: is the opaque data value chosen by the controller which maybe used by the controller to perform flow filtering, flow modification or flow deletion; it is not used during packet processing
- A match field and priority is used to identify a flow table entry matched.
- Required match field based on OpenFlow 1.4:-
  1. Ingress Port
  2. Destination MAC address
  3. Source MAC address
  4. Ethernet Type
  5. IPv4/IPv6 protocol number
  6. IPv4/IPv6 source address
  7. IPv4/IPv6 destination address
  8. TCP/UDP source port
  9. TCP/UDP destination port
- After a packet is matched with flow entry with highest priority, packets will be handled based on the associated instructions such as modifying packets, modifying action sets and/ or perform pipeline processing.
- Example of instructions include:-
  - Meter: to direct packets to particular meter to perform meter band type such as dropping or DSCP remarking
  - Apply-action: to immediately apply specific action without any Action set modification.
  - Clear-action: to immediately delete all actions in action set
  - Write-actions: to immediately add or overwrite action set
  - Goto-Table: to direct packets to be processed by the subsequent table.



- Starting from OpenFlow version 1.1 and above, OpenFlow protocol supports pipeline processing in conjunction with the features of supporting multiple flow tables. Compared with OpenFlow version 1.0 which supports single flow table, pipeline processing reduces the size for flow table match entries and improve the flow table checking efficiency by dividing the flow table match fields to multiple flow tables.
- The flow table for an OpenFlow switch is sequentially numbered, starting from 0. A OpenFlow switch must have at least 1 flow table and pipeline processing will be required to process packets if there are more than 1 flow tables in the OpenFlow switch. When packet is forwarded into the first flow table, the highest priority flow entry is matched and instruction set will direct the packet to the next flow table and the per-table packet processing process repeats again. If instruction does not direct packet to the following table, the pipeline processing will stop and processed with if associated action set.
- In the case that if a packet does not match with any flow entry in the flow table, this is called table miss. If there is particular configuration done for table miss entry, the table miss flow entry can be handled with different action sets, such as discarding, forwarding them transparently to controller etc.



- Upon receiving a packet, an OpenFlow switch will start packet processing by looking up at the first flow table. Based on the match fields information in the packets such as source address, IPv4 address or metadata fields, the flow table counters will be updated and corresponding instructions will be performed, such as performing or updating action set, updating packet match fields or updating metadata for next flow table processing. If there is a metadata defined for next flow table processing, the packet will be forwarded to the following flow table, performing match field checking and instruction checking again. If packets have already been sent to last flow table with the largest sequence number, it means that there is no more flow table to be checked, and thus action set defined in the last flow table will be executed accordingly.
- In the case if there is no matched flow entries in the flow table, the packet will be dropped if there is no table miss entry configured in the switch. If the table miss function is used, the packet will be handled with the instructions defined for table miss option.

 **Contents**

1. Introduction of SDN OpenFlow Protocol
  - 1.1 OpenFlow History and Version Evolution
  - 1.2 OpenFlow Basic Architecture
  - 1.3 OpenFlow Table Description
  - 1.4 OpenFlow Protocol Working Principles**
  - 1.5 OpenFlow Application in SDN



## OpenFlow Protocol Message Types

- The Openflow protocol supports 3 messages types:-
  - **Controller-to-switch message**
    - Initiated by the controller and may or may not require response message from switch
  - **Asynchronous message**
    - Initiated by the OpenFlow switch mainly to acknowledge on a packet arrival or switch state change.
  - **Symmetric message**
    - Sent in either direction from controller to switch, or from switch to controller

- **Controller-to switch message**

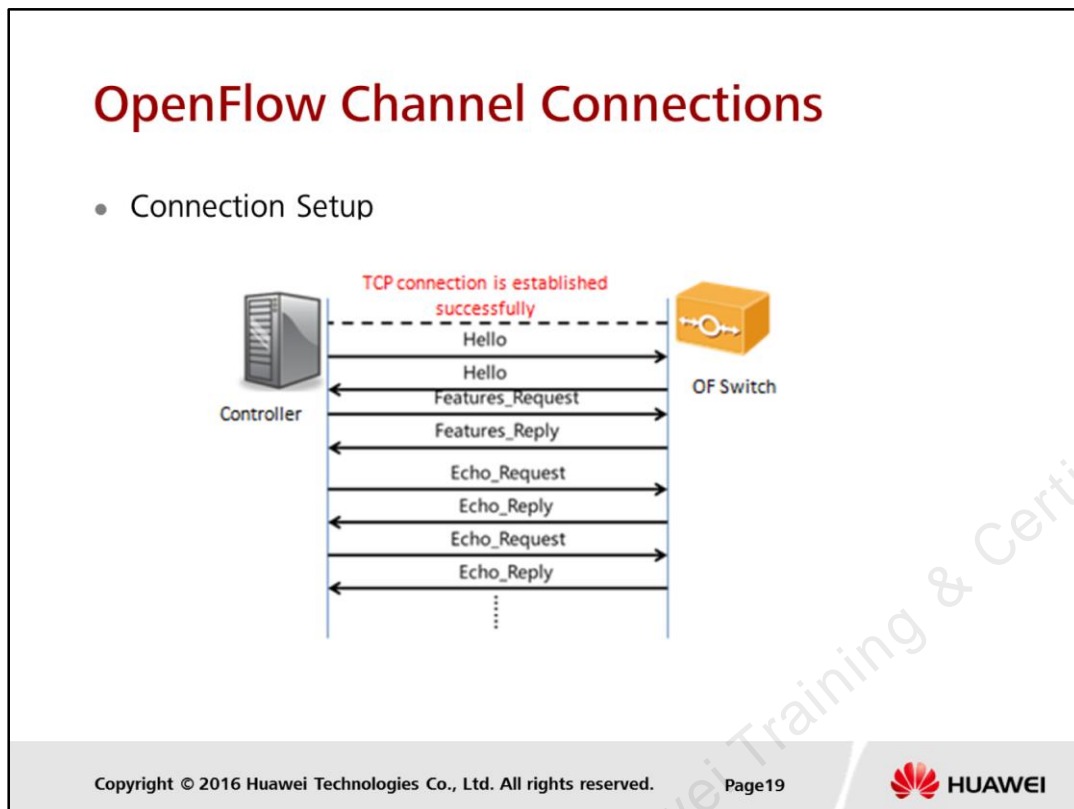
- This message is initiated by controller, mainly to request for certain identities or basic capabilities functions from switch, which might be required during the establishment of the Openflow channel.
- The examples of this message are Modify-State messages used to add, delete or modify flow entries in Flow tables, Read-State message used to collect information such as configuration data from switch, Packet-out message used to forward packets received via packet-in message etc. Some other message examples include Barrier, Role Request and Asynchronous-Configuration message.

- **Asynchronous Message**

- This message is sent from Openflow switch to controller, in order to denote a packet arrival or notify switch state change.
- The examples of this message include Packet-In message used to transfer a packet to the controller, Flow-Removed message used to inform about deletion of flow table entry in flow table, and port-status message used to notify about port status change on the switch.

- **Symmetric Message**

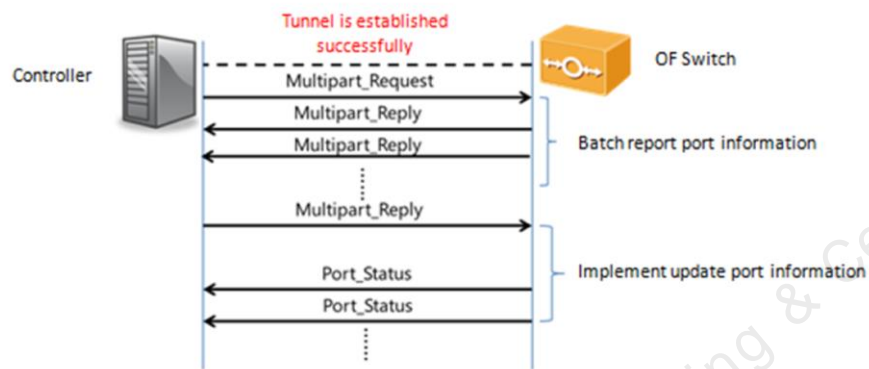
- Symmetric message might be sent from a switch to controller or vice versa.
- The examples of the symmetric messages include Hello messages used during connection startup, Echo request/reply messages to verify the connection status, and error messages to notify the peer side on the problems.



- An OpenFlow channel is established over a TCP connection. Devices at both ends of a channel exchange heartbeat packets to maintain the connection. Figure above illustrates the process of establishing and maintaining an OpenFlow channel:
  1. The controller and forwarder establish a TCP connection.
  2. The controller and forwarder exchange Hello packets to negotiate parameters, such as version numbers.
  3. After negotiation is successful, the controller sends a Features\_Request packet to query attributes on the forwarder.
  4. Upon receipt of the Features\_Request packet, the forwarder sends the controller a Features\_Reply packet carrying the forwarder's attributes, such as the supported flow table format and buffer size.
  5. After receiving the Features\_Reply packet, the controller succeeds in establishing an OpenFlow channel with the forwarder.
  6. The controller and forwarder exchange Echo packets (heartbeat packets) to monitor the status of each other. One end periodically sends an Echo\_Request packet to the other end. Upon receipt of this packet, the other end responds with an Echo\_Reply packet. If one end fails to receive an Echo\_Reply packet after sending a specified number of Echo\_Request packets, it considers the other end faulty and terminates the connection to the other end.

## OpenFlow Channel Connections

- Reporting OpenFlow port information

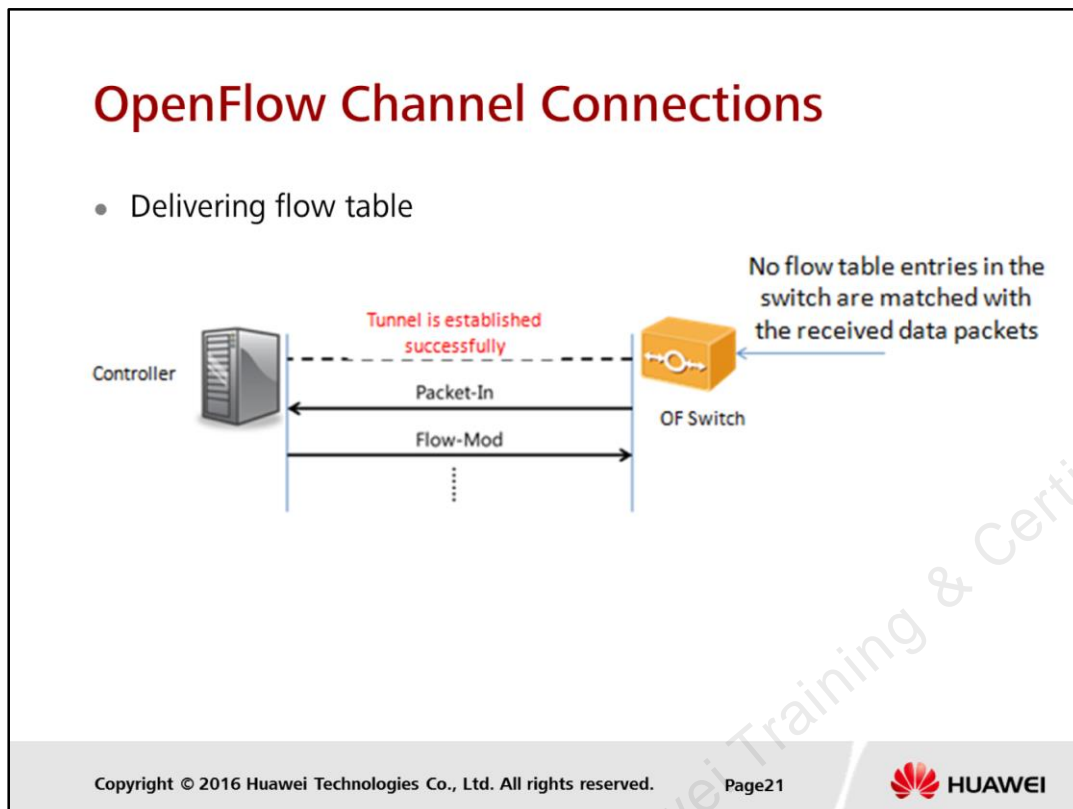


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page20



- A forwarder can report its port information to a controller along an OpenFlow channel.
- In figure above, after an OpenFlow channel is established between a controller and a forwarder, the controller queries port information on the forwarder. The detailed process is as follows:
  - The controller sends a Multipart\_Request message to query the forwarder's port information.
  - Upon receipt of this message, the forwarder sends a Multipart\_Reply message carrying its port information in batches to the controller. If the port status changes, the forwarder sends a Port\_Status message to notify the controller of the change.



- If a packet is sent to an Openflow switch and there is no flow table entry in the switch are matched, the packet will send a Packet-in message to controller.
- A controller sends a Flow\_Mod message carrying flow table information to a forwarder. The Flow\_Mod packet primarily carries basic flow table information (such as the table ID and priority), a match attribute set, and an instruction set. The match attribute set contains information (such as MAC and IP addresses) used to match packets against entries in the flow table. Matching packets are processed using a specific instruction set. The instruction set contains various instructions, such as modifying packet attributes and forwarding packets through a specific outbound interface.

## OpenFlow Table Distribution Modes

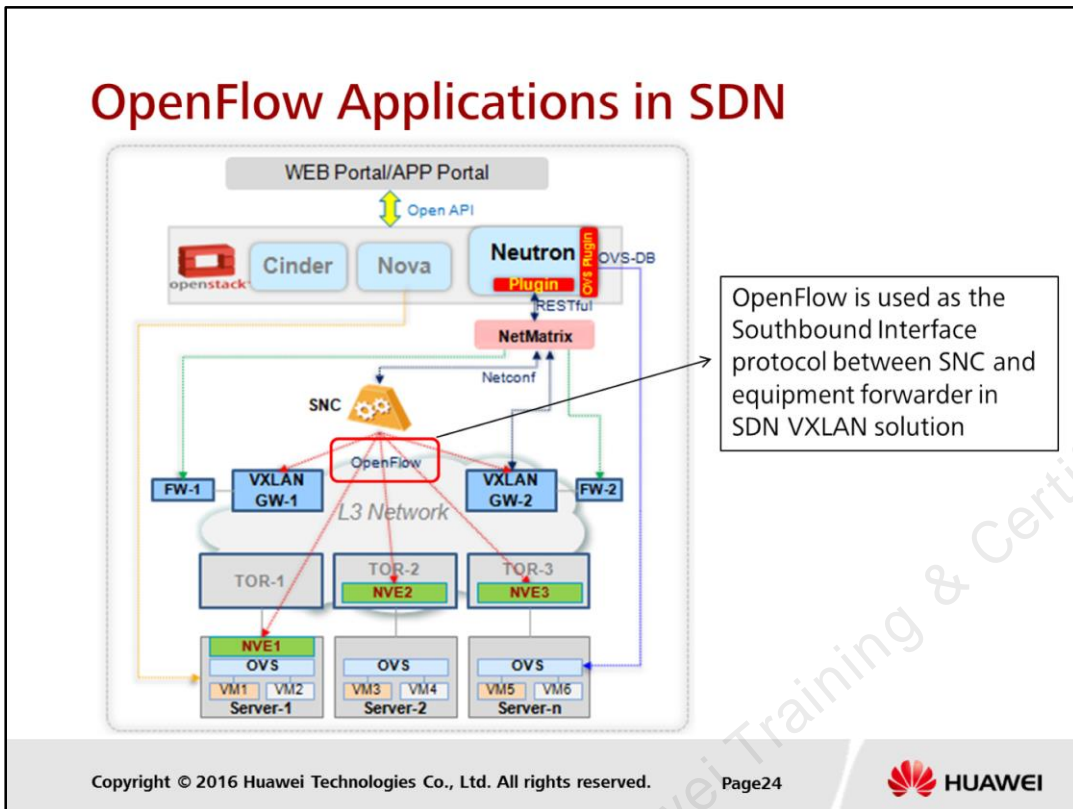
- The OpenFlow table supports 2 distribution modes:-
  - **Active mode**
    - Controller automatically update switch with the its collected flow entry information so that OpenFlow switch can forward a packet directly by using its flow table.
    - If there is no flow entry matched, the packet can be handled in 2 ways: discard or use Packet-In message to forward to packet back to controller
  - **Passive mode**
    - Controller passively distributes flow table information to switch upon receiving a Packet\_In message containing packet from switch.

- The passive openflow table distribution is better in term of TCAM capacity saving as switch or network devices need maintain the flow table when there is a real-time packet flows in the network.



## Contents

1. Introduction of SDN OpenFlow Protocol
  - 1.1 OpenFlow Version Evolution and History
  - 1.2 OpenFlow Basic Architecture
  - 1.3 OpenFlow Table Description
  - 1.4 OpenFlow Protocol Message Types
  - 1.5 OpenFlow Application in SDN**



- OpenFlow is used in one of the Huawei SDN solution, VxLAN as the communication protocol between the SDN controller- SNC and the forwarder which is the forwarding switch connected to the server or server itself.

 **Contents**

1. Introduction of SDN OpenFlow Protocol
- 2. Introduction of SDN Netconf Protocol**
3. Introduction of SDN RESTful Protocol
4. Introduction of SDN SNMP Protocol
5. Introduction of SDN Netstream Protocol





## Contents

### 2. Introduction of SDN NETCONF Protocol

#### **2.1 NETCONF Introduction**

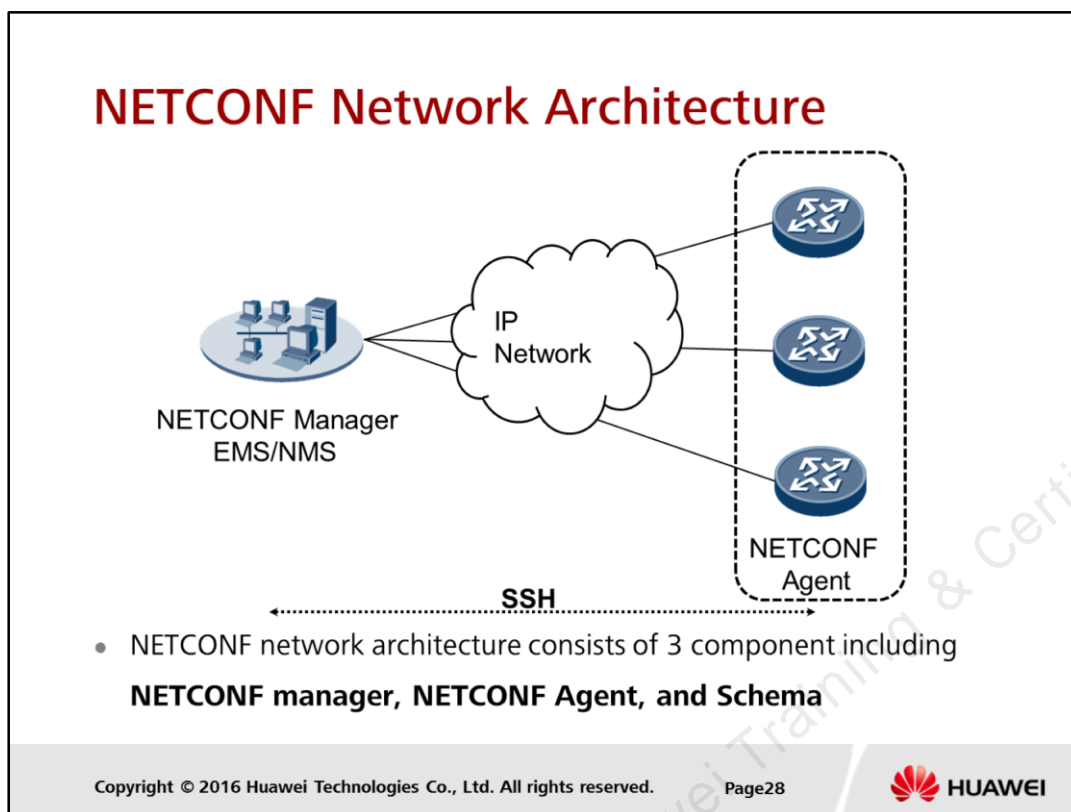
#### 2.2 NETCONF Protocol Layers

#### 2.3 NETCONF Application in SDN

## NETCONF Overview

- The Network Configuration Protocol (NETCONF) is an **extensible markup language (XML)** based network configuration and management protocol.
- NETCONF provides mechanisms to install, manipulate and delete configurations on network devices; it is commonly use by NMS to remotely manage and monitor network devices..
- NETCONF uses a simple **remote procedure call (RPC)** mechanism to implement communication between a client and server.
- NETCONF is usually transported over **SSH** protocol.

- NETCONF is a network management protocol developed and standardized by IETF; it was first published in RFC4741 in December 2006 and was later revised in RFC6241 in June 2011.
- NETCONF is a session-based network management protocol, which uses XML- encoded remote procedure calls (RPC) and configuration data to manage network devices.
- The mandatory transport protocol for NETCONF is Secure Shell Transport Layer Protocol (SSH). The default TCP port assigned is 830. NETCONF server must listen for connections to the NETCONF subsystem on this port.
- NETCONF offers the following benefits:
  - Facilitates configuration data management and interoperability between different vendors' devices using XML encoding to define messages and the RPC mechanism to modify configuration data.
  - Reduces network faults caused by manual configuration errors.
  - Improves the efficiency of system software upgrade performed using a configuration tool.
  - Provides high extensibility, allowing different vendors to define additional NETCONF operations.
  - Improves data security using authentication and authorization mechanisms.



- NETCONF network architecture consists of the following components:

- NETCONF manager**

- A NETCONF manager functions as a client that uses NETCONF to manage devices.
- A NETCONF manager can send `<rpc>` elements to the NETCONF agent of a managed device to query or modify the configuration data on the managed device.
- A NETCONF manager can learn the status of a managed device based on the alarms and events actively reported by the NETCONF agent of the managed device.

- NETCONF agent**

- A NETCONF agent functions as a server that maintains the configuration data on the managed device, responds to the `<rpc>` elements sent by a NETCONF manager, and sends the requested information to the NETCONF manager.
- After a NETCONF agent receives an `<rpc>` element, the NETCONF agent parses the element, processes the element based on the Configuration Management Framework (CMF), and sends an `<rpc-reply>` element to the NETCONF manager.
- NETCONF agent reports the alarm or event to a NETCONF manager for the NETCONF manager to learn the status of the managed device.

- Schema**

- A schema file is a data model file that defines a set of rules used to describe an XML document. A schema file defines all the management objects on a managed device, the constraints and hierarchical relationships between these management objects, and the read and write permissions of these management objects.
- A schema file functions in a way similar to a Simple Network Management Protocol (SNMP) management information base (MIB) file.

- The information that can be retrieved from a running device is as follows:
  - Configuration data:** a set of writable data that is required to transform a device from its initial default state into its current state
  - State data:** the additional non-configuration data on a device, such as read-only status information and collected statistics
- NETCONF deals with configuration data operations performed by the NETCONF manager and is not involved with how configuration data is stored.



## Contents

### 2. Introduction of SDN NETCONF Protocol

#### 2.1 NETCONF Introduction

#### **2.2 NETCONF Protocol Layers**

#### 2.3 NETCONF Application in SDN

## NETCONF Protocol Layers

Layer	Description	Example
Layer 4: Content	•Describe the configuration data; the configuration data depends on vendors' devices	Configuration data, notification data
Layer 3: Operations	The operation layer defines a series of basic operations used in RPC..	<get-config>, <edit-config>, <notification>
Layer 2: Messages	The message layer provides a simple and transport-independent framing mechanism for encoding RPCs.	<rpc>, <rpc-reply>
Layer 1: Secure Transport Protocol	Provides communication channel between NETCONF client and server	SSH, SSL, BEEP

- Like the open systems interconnection (OSI) model, the NETCONF protocol framework also uses a hierarchical structure. A lower layer provides services for the upper layer.
- The hierarchical structure enables each layer to focus only on a single aspect of NETCONF and reduces the dependencies between different layers.
- The four layers of NETCONF are described as listed below:-
  1. Secure Transport Protocol
    - NETCONF can be layered over any transport protocol that meets the following requirements:
      - The transport protocol is connection-oriented. A permanent link is established between the NETCONF manager and agent. This link provides reliable and sequenced data delivery.
      - The transport protocol provides authentication, data integrity, and confidentiality for NETCONF.
      - The transport protocol provides a mechanism to distinguish the session type (client or server) for NETCONF.
    - Currently, the VRP can use only SSH as the transport protocol for NETCONF.



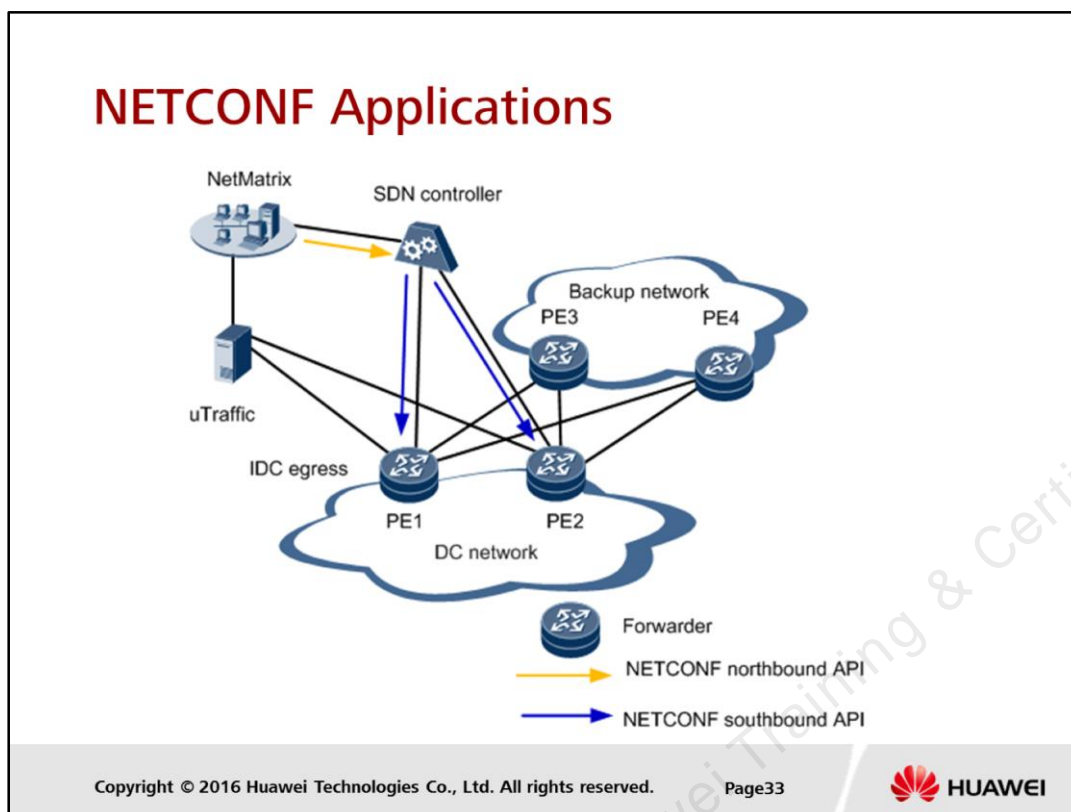
## Contents

### 2. Introduction of SDN NETCONF Protocol

2.1 NETCONF Introduction

2.2 NETCONF Protocol Layers

**2.3 NETCONF Application in SDN**



- The SDN controller is the core of the intelligent traffic control solution. NETCONF runs on the NMS and controller to transmit XML remote procedure call (RPC) packets. The northbound API function enables you to use an NMS to log in to and configure the controller. The southbound API function enables you to use the controller to log in to and configure a forwarder.
- An NETCONF connection is established between the SDN controller and a forwarder as follows:
  1. NETCONF is enabled on the SDN controller, and forwarder information is configured, including the IP address, port number, user name, password, and connection status.
  2. The SDN controller sends an RPC request packet to the forwarder for an NETCONF connection. The RPC request packet carries the IP address of the forwarder.
  3. Upon receipt of the RPC request packet, the forwarder checks whether the carried IP address is its own IP address. If the carried IP address is its own IP address, the forwarder responds with an RPC reply packet to establish an NETCONF connection.

 **Contents**

1. Introduction of SDN OpenFlow Protocol
2. Introduction of SDN Netconf Protocol
- 3. Introduction of SDN SNMP Protocol**
4. Introduction of SDN RESTful Protocol
5. Introduction of SDN Netstream Protocol





## Contents

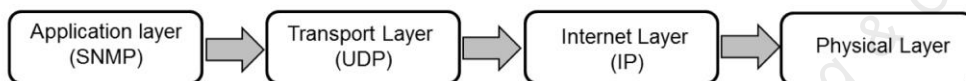
### 3. Introduction of SDN SNMP Protocol

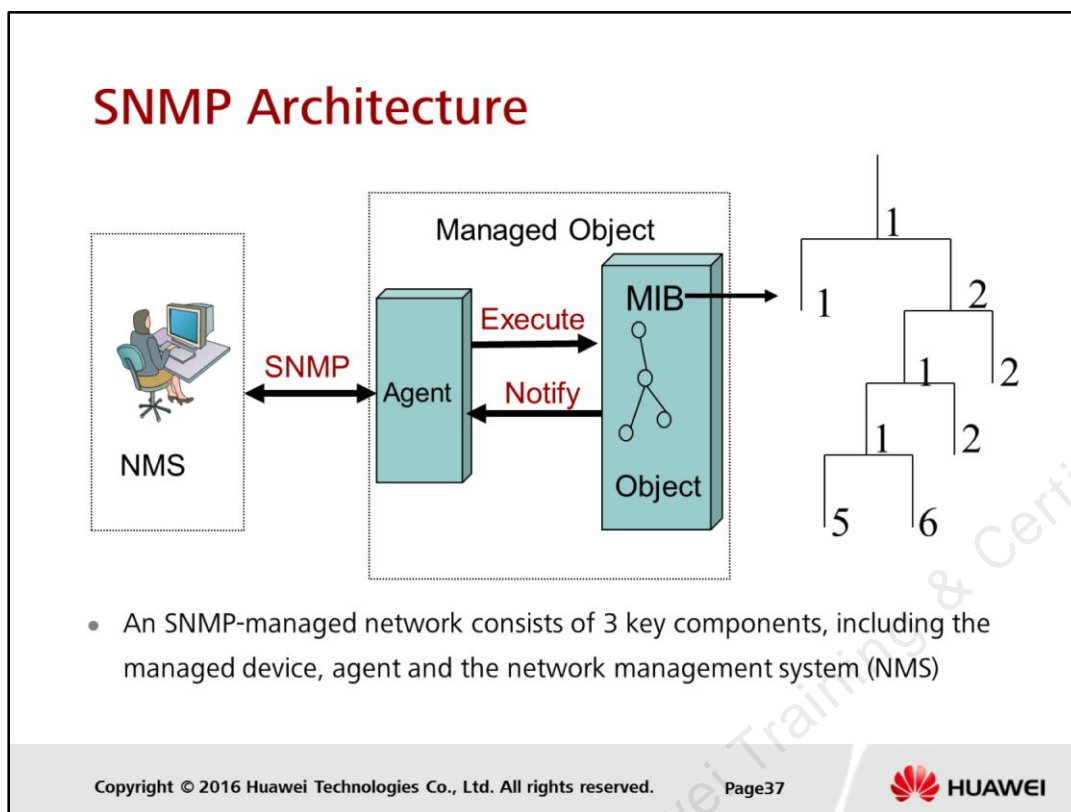
#### **3.1 SNMP Overview and Basic Concepts**

#### 3.2 SNMP Version and Protocol Packet Structures

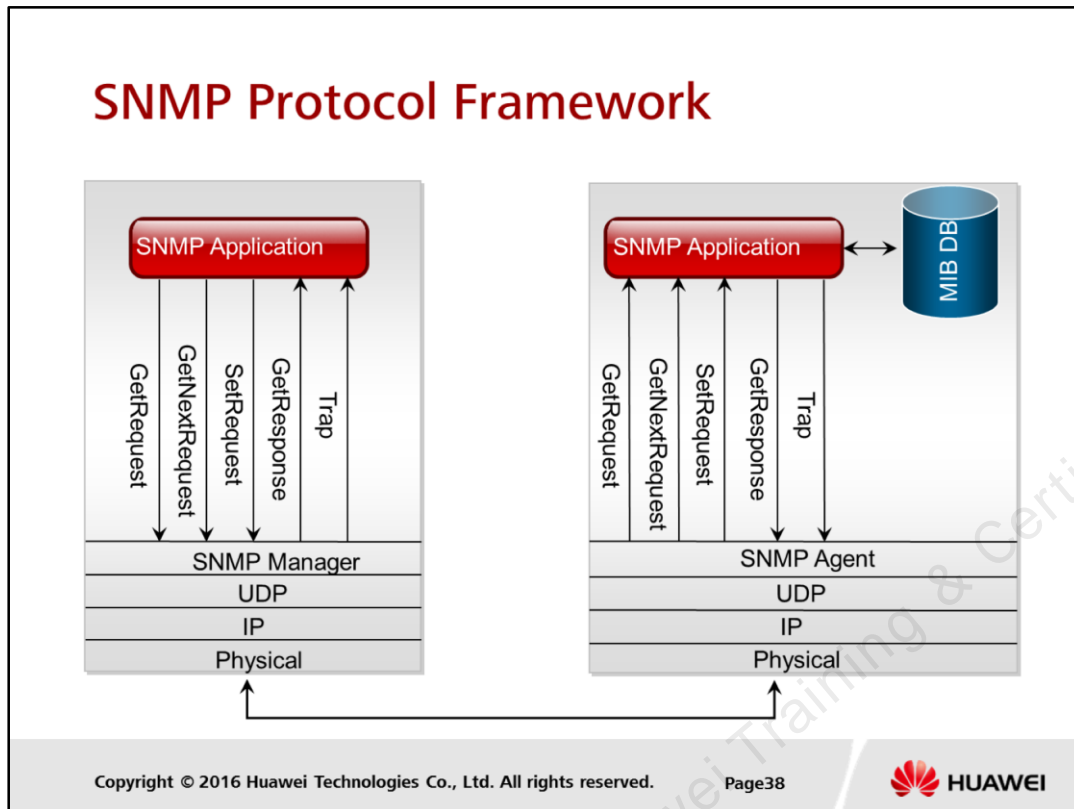
## SNMP Overview

- Simple Network Management Protocol (SNMP) is an application layer protocol widely used in TCP/IP network for collecting, managing and modifying information of managed devices
- Being the part of TCP/IP protocol suite, the SNMP messages are wrapped as User Datagram Protocol (UDP) and transmitted in the Internet Protocol.

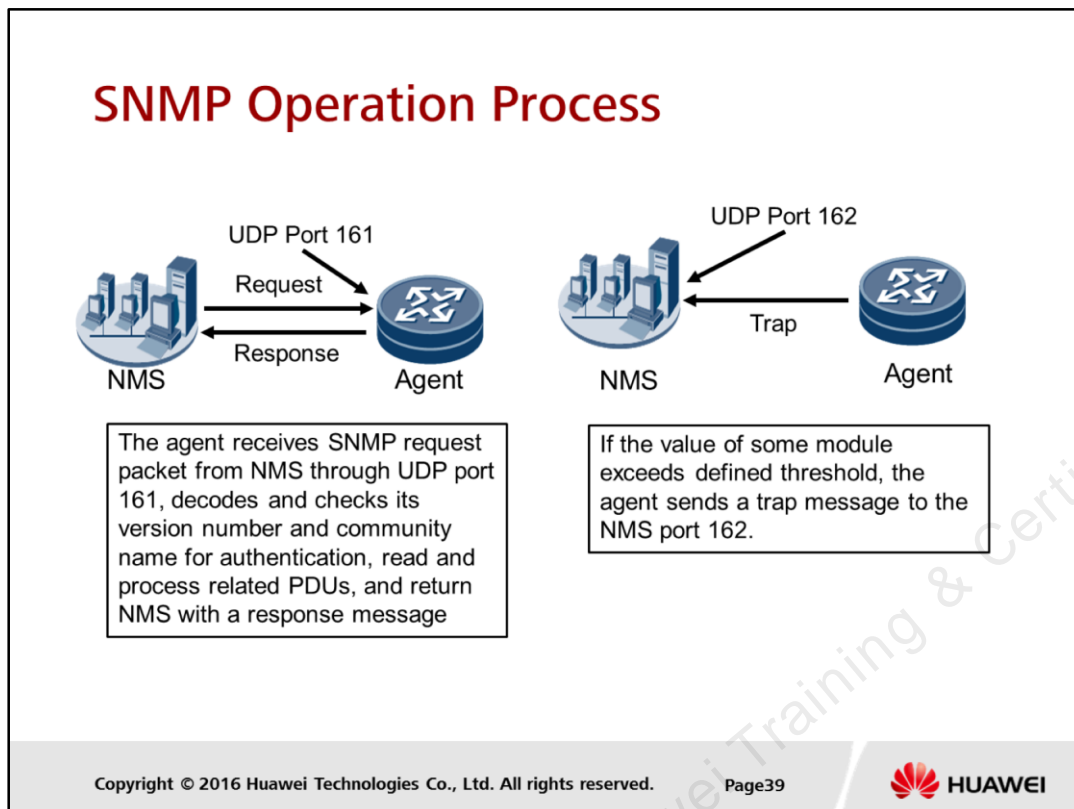




- An SNMP-managed network consists of 3 key components, as per listed below:-
  1. **Managed object:** devices or network elements to be monitored.
  2. **Agent**
    - An agent is the process run on the managed devices. After the managed device receives the request sent by the NM station, the agent is responsible for responding to the request. The agent has the following functions:
      - Collects the information about device status.
      - Supports remote operations on the device through NMS.
      - Sends trap messages to the NM station.
  3. **NMS**
    - Network Management System (NMS) is the network management software run on the Network Management station (NM station). The network manager sends requests to the managed devices and monitors and configures the network devices through NMS.
- In SNMP, the NMS and the agent communicate through packet exchanging.
- The NMS acts as a manager and sends an SNMP request packet to the SNMP agent.
- The agent searches in the MIB on the device for the required information and sends a response packet to the NMS.
- The agent sends a trap message to the NMS when the value of some module on the managed device exceeds the defined threshold. According to the trap message, the network manager can process the occurred event in time



- To simplify the development of the Agent side, SNMP only defines two kinds of operations --- Get and Set. Get is used to obtain management information from managed equipment. And Set is used to configure managed equipment via setting the value of variable.
- NMS and Agent transfer management information to each other via packet. And SNMP V1 only defines five kinds of packets:
  - Get Request packet: Used to get the value of specified management variable.
  - GetNext Request packet: Used to continuously get the values of a group of variables.
  - GetResponse packet: Used to respond request, return value for request or error type, etc.
  - Set Request packet: Used to set the specified management variable.
  - Trap packet: Used for managed equipment to send information to NMS initiatively in urgent cases.
- GetRequest and GetNextRequest are used to obtain information of the managed object in NM. SetRequest is used to configure the managed object. These three kinds of requests correspond with three kinds of SNMP messages. Agent responds them via sending GetResponse message.
- Trap is generated by Agent. It is used to report abnormal event of the managed equipment to the NM. Agent will send Trap to notify NM when equipment gives alarm or important data is changed by user/console/other NMs. When SNMP Manager receives the Trap, relevant actions will be initiated, such as diagnosing fault via polling, adopting recovery measures, modifying relevant database of the NM.



- SNMP request and response message operation are described in details as shown below:-
  1. The agent receives an SNMP request packet from the NMS through UDP port 161.
  2. The agent decodes the packet based on ASN.1 basic coding rules and represents it in an internal data structure. The agent discards the packet if there is a decoding failure.
  3. The agent gets the version number from the packet. The agent discards the packet if the version is inconsistent with the SNMP version it supports.
  4. The agent gets the community name from the packet. The community name is filled by the NMS that sends the request. If the community name is inconsistent with that of the agent, the packet is discarded. A trap message or an Inform packet is generated simultaneously.
  5. The agent gets PDUs from the authenticated ASN.1 object. If failed to get the PDUs, the agent discards the packet; otherwise, the agent processes the PDUs.
  6. The agent processes PDUs differently and gets the management variables of the corresponding protocol modules by searching nodes that correspond to management variables in the MIB.
  7. The agent encapsulates the values of management variables in a PDU, uses the source IP address and port of the request packet as the destination IP address and port, and adds the SNMP version number. A response packet is then generated. After being coded, the response packet is sent to the NMS
- Trap is an unprompted behavior of the managed device. It does not belong to the basic operations on the NMS. As shown in the figure on the right above, if the value of some module exceeds the defined threshold, the agent sends a trap message to the NMS. The NMS receives the trap message from UDP port 162, according to which the manager can process the network abnormality in time. In the case that the interface status changes, the agent also sends a trap message to the NMS. The manager can then diagnose and rectify the fault according to the trap message.

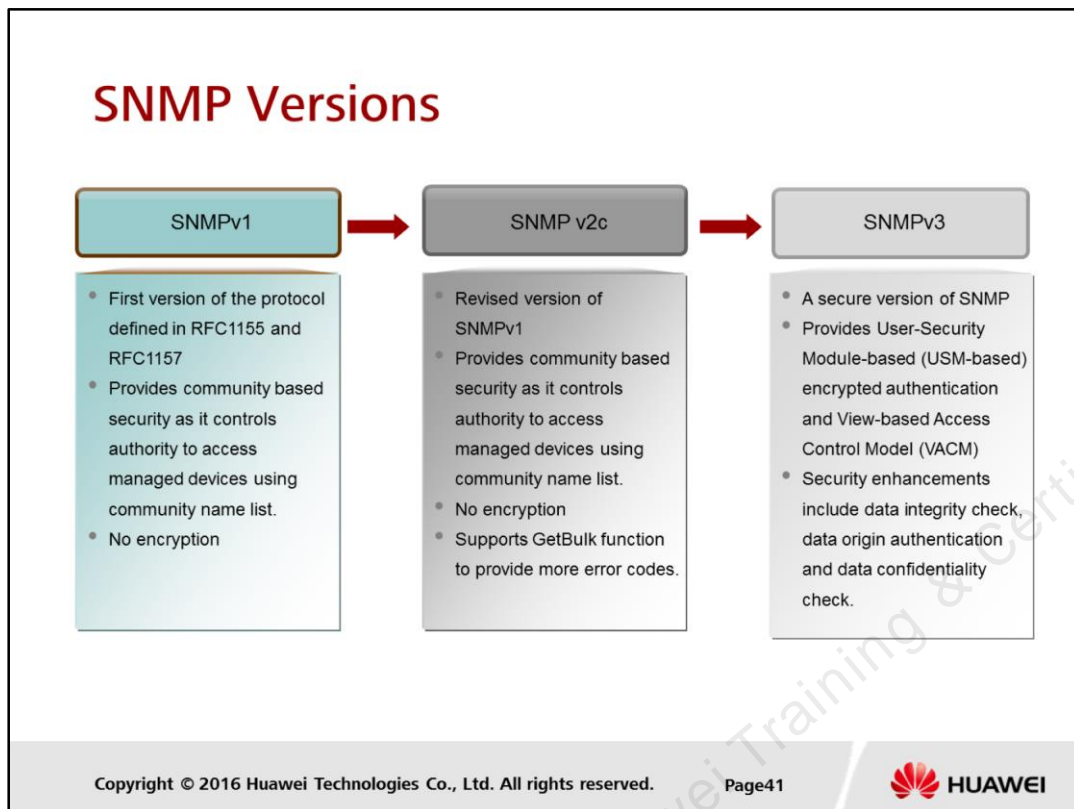


## Contents

### 3. Introduction of SDN SNMP Protocol

#### 3.1 SNMP Overview and Basic Concepts

#### **3.2 SNMP Version and Protocol Packet Structures**



- There are three versions of SNMP: SNMPv1, SNMPv2c, and SNMPv3.
- In SNMPv1 and SNMPv2c, the NMS controls the authority to access managed nodes by using the community name list. The agent does not check the validity of the community name. SNMP packets are transferred without encryption. That is, security is not guaranteed for authentication and confidentiality.
- Compared with SNMPv1, SNMPv2c supports:
  - More operations and data types
  - Plenty of error codes
  - Multiple transport layer protocols
- SNMPv3 provides all the functions of SNMPv1 and SNMPv2, and features a security mechanism that authenticates and encrypts SNMP packets. In terms of security, SNMPv3 emphasizes security of data and access control.
- SNMPv3 ensures the security for SNMP packets in the following ways:
  - Data integrity check
    - The data cannot be modified in a unauthorized manner. The change of the data sequence is limited to the allowed extent.
  - Data origin authentication
    - SNMPv3 authenticates the managed node from which the received packet originates and not the application that generates the packet.
  - Data confidentiality
    - When the NMS or the agent receives a packet, it checks the time at which the packet is generated. If the difference between the creation time and the system time exceeds the threshold, the packet is discarded. In this way, the packets that are modified by malicious users are not accepted.

- SNMPv3 control the access to the MOs by the operations of the protocol.

Huawei Training & Certification Huawei Training & Certification



 **Contents**

1. Introduction of SDN OpenFlow Protocol
2. Introduction of SDN Netconf Protocol
3. Introduction of SDN SNMP Protocol
- 4. Introduction of SDN RESTful Service**
5. Introduction of SDN Netstream Protocol



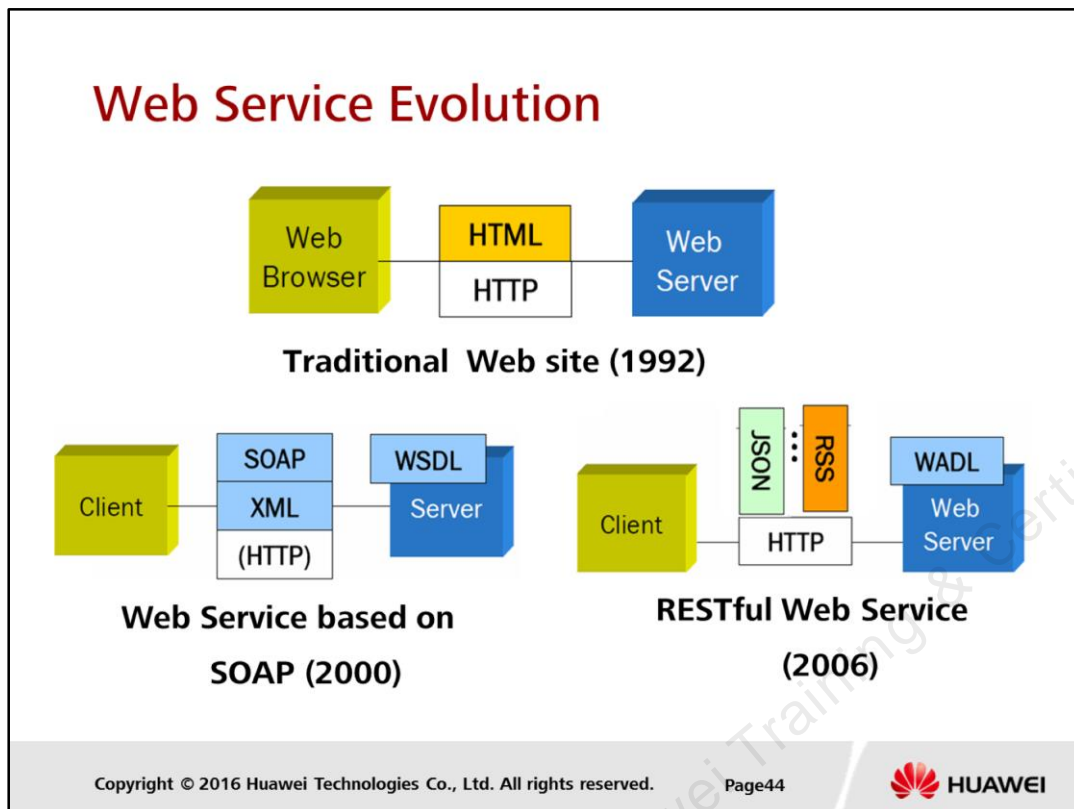
## Contents

### 4. Introduction of SDN RESTful Service

#### **4.1 REST and RESTful Basic Concepts**

#### 4.2 REST Architectural Elements

#### 4.3 RESTful Service Application in SDN



- In traditional websites of World Wide Web (WWW) when the web service was just developed during year 1992, the client only served as a web browser, which was used to access the web server for information exchange. A web server is only a platform used for providing information resources for web client, and no other services could be provided.
- In conjunction with the web service growth till year 2006, web server now is no longer serve for information exchange only, but other additional functions and services have been added, such as remote application service etc.
- SOAP (Simple Object Access Protocol) is the common method used for computer network in web service implementation. It uses XML Information Set for its message format, and depends on application layer protocols such as the most common HTTP protocol for message transmission between client and web server.
- RESTful Web service is a new type of web service built based on REST software architectural style. It uses JSON (Javascript Object Notation) as its message format and also relies on HTTP protocol as application layer message transmission medium.
- RESTful Web service has soon grown as a preferable method for web service owing to its advantages in term of simplicity and flexibility

## REST and RESTful Service

- **What is REST?**
  - Representational State Transfer (REST)
  - REST is a **software architectural style** for distributed system such as World Wide Web (WWW)
  - REST is NOT a standard or protocol
- A service based on REST architecture is called a **RESTful** service.

- In computing point of view, Representational State Transfer (REST) is the software architectural style of the World Wide Web (WWW). To explain it in a more layman term, the architectural style decides how a client can access to a web service. For example, most of our web service in computer networks nowadays used SOAP (Simple Object Access Protocol) information by using XML information set as its message format. REST is just another different architectural style to exchange information between client and a web server by using JSON (JavaScript Object Notation) as a message format.
- REST was defined by Roy Thomas Fielding in his 2000 PhD dissertation "Architectural Styles and the Design of Network-based Software Architectures". He developed REST architecture to describe a desired web architecture, to identify existing web problems and provide an alternative solution to make web more successful. Fielding used REST to design HTTP and URI (Uniform Resource Identifier)



## Contents

### 4. Introduction of SDN RESTful Service

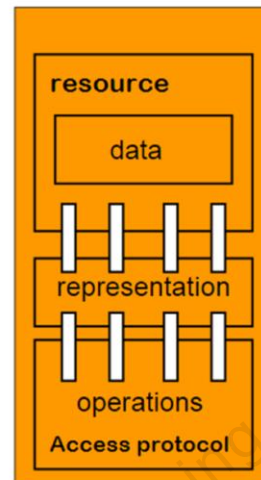
#### 4.1 REST and RESTful Basic Concepts

#### **4.2 REST Architectural Elements**

#### 4.3 RESTful Service Application in SDN

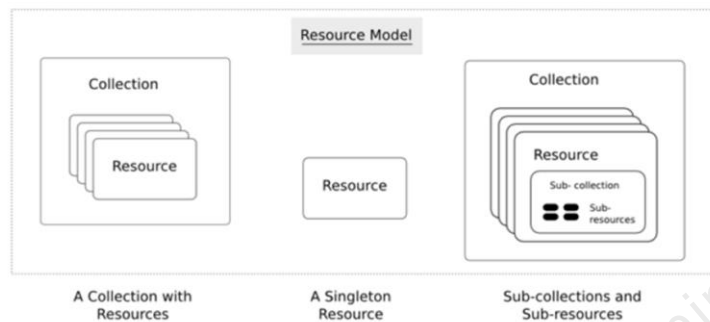
## REST Architectural Element (1/5)

- The crucial elements in REST architectures consists of:-
  1. Resources & Resource Identifiers
  2. Representation
  3. REST State Transfer Operation



## Resource (2/5)

- A resource is an object with a type, associated data, relationships to other resources, and a set of methods that operate on it.



### RESTful API Resource Model

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

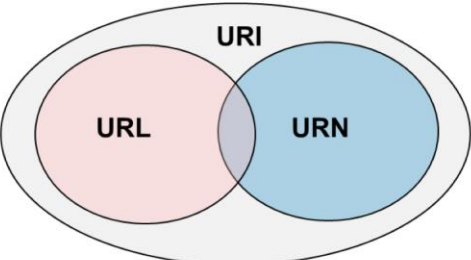
Page48



- A resource can be defined as a conceptual mapping to a set of entities or values.
- A resource is an object with a type, associated data, relationships to other resources, and a set of methods that operate on it.
- Any information that can be named can be a resource; for instance, a document or an image can be a resource.
- Resources in RESTful API can be grouped into collections; each collection is homogenous, contains only one type of resource and unordered. Resources can also stand alone and exist outside any collection, which is called singleton resources. Besides, collection can exist globally at the top level of an API, but can also be contained in a single resource, which is referred to as sub-collections.
- Applications process singleton and collection resources differently. For example, applications can apply paging and filtering techniques to collection resources, but not to singleton resources. Another difference is that some HTTP operations are not allowed on collection resources.

## Uniform Resource Identifier (URI) (3/5)

- In RESTful web service, Uniform Resource identifier (URI) is used to identify the particular resource.



**URI= Uniform Resource Indicator**  
**URL = Uniform Resource Locator**  
**URN= Uniform Resource Name**

**URI = URL + URN**

**http ://example.org/wiki/Main\_Page**

⏟

**Scheme**

⏟


**URL**

⏟

**URN**

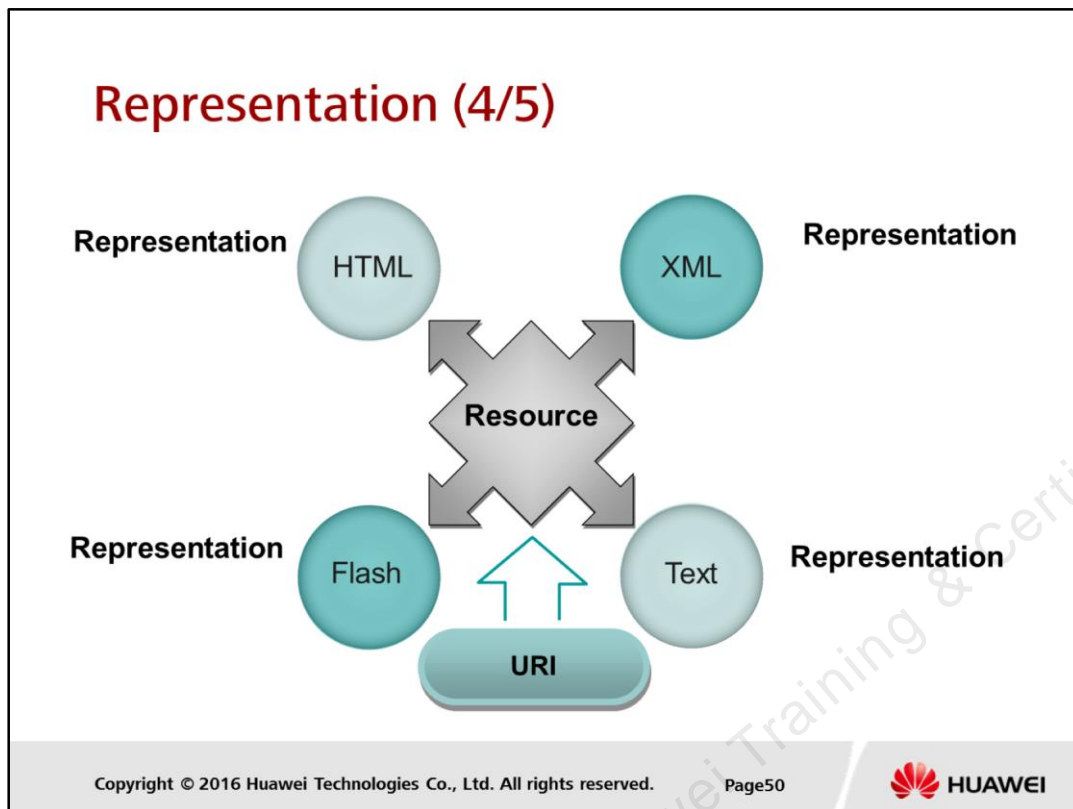
Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page49

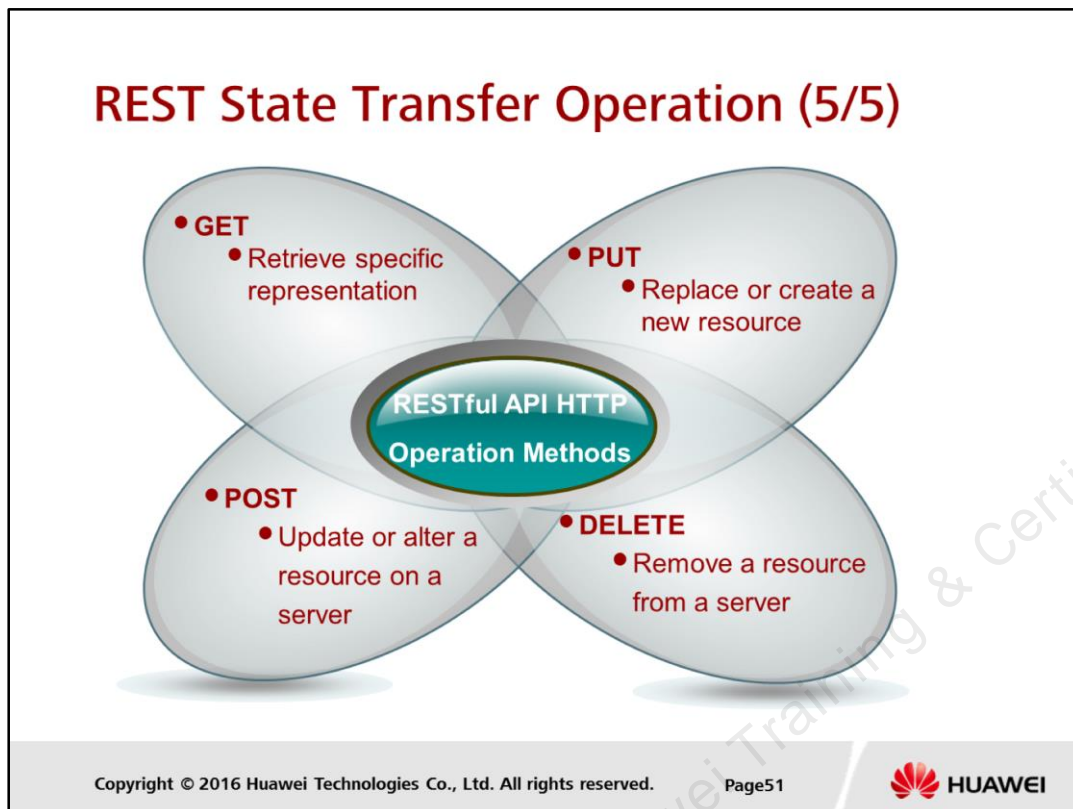


- In RESTful web service, Uniform Resource identifier (URI) is used to identify the particular resource. In other words, a resource identifier is like an unique ID assigned for respective resource.
- The Uniform Resource Identifier (URI) standard is defined by the IETF, and provides a syntax that includes a list of allowed and prohibited characters, a generic structure for identifiers, and further defines the concepts of Uniform Resource Locators (URLs) and Uniform Resource Names (URNs).
- A URL is a URI that, in addition to identifying a web resource, specifies the means of acting upon or obtaining the representation, specifying both its primary access mechanism and network location.
- A URN is a URI that identifies a resource by name in a particular namespace. A URN can be used to talk about a resource without implying its location or how to access it.





- In the previous slides, we have discussed about resource and resource identifier. However, they are still abstract entities. Before they can be communicated to clients over a HTTP connection, they need to be serialized to a textual representation. This representation can then be included as an entity in an HTTP message body. The message format representation can be in HTML, XML, flash or text representation.



- In RESTful API, a client and server exchange representations of a resource, which reflect its current state or its desired state. In other words, REST is the way for two machines to transfer the state of a resource via representation.
- There are four RESTful API HTTP operation methods, as per listed below:-
  - PUT: Used to replace the addressed member of the collection or if it does not exist, create it; it is idempotent
  - GET: Retrieve a representation of the addressed member of the collection, expressed in an appropriate internet media type; it is safe and idempotent
  - POST : Not so commonly used; treat the addressed member as a collection in its own right and create a new entry on it; it is idempotent
  - DELETE: Delete the addressed member of the collection; it is idempotent
- Idempotent means that the operation will produce the same result no matter how many times it is repeated.



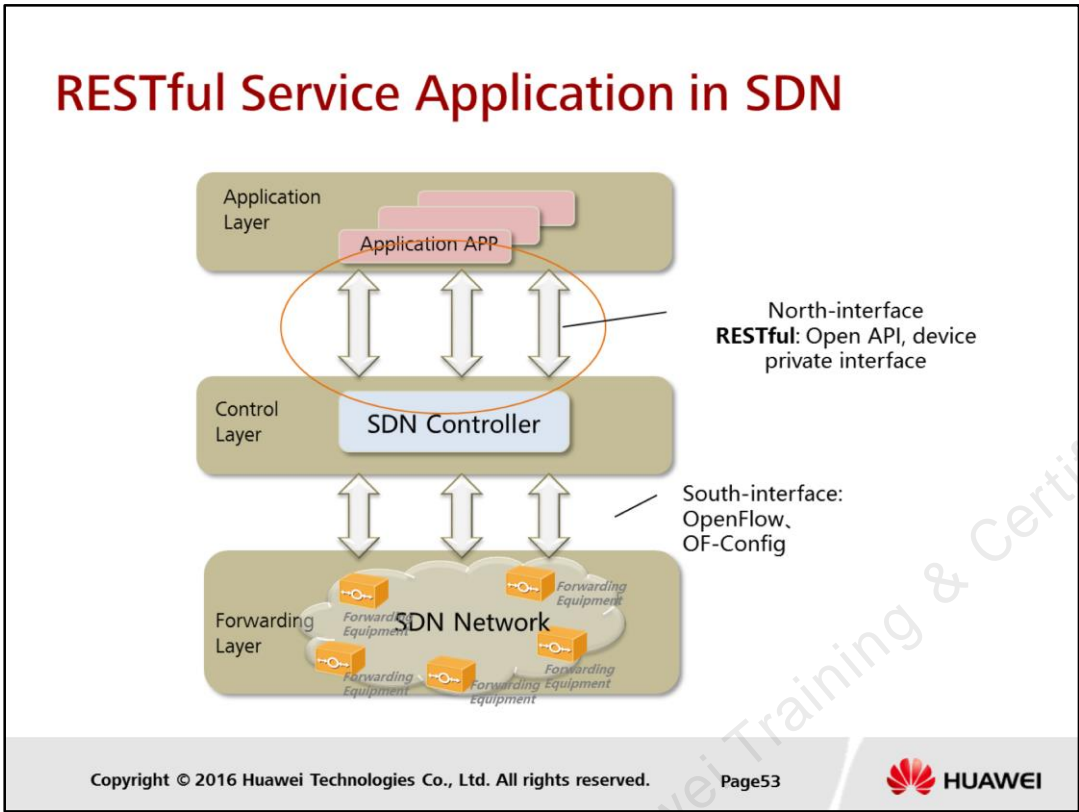
## Contents

### 4. Introduction of SDN RESTful Service

4.1 REST and RESTful Basic Concepts

4.2 REST Architectural Elements

**4.3 RESTful Service Application in SDN**



 **Contents**

1. Introduction of SDN OpenFlow Protocol
2. Introduction of SDN Netconf Protocol
3. Introduction of SDN SNMP Protocol
4. Introduction of SDN RESTful Protocol
- 5. Introduction of SDN Netstream Protocol**



## Contents

### 5. Introduction of SDN NetStream Protocol

#### 5.1 NetStream Overview

#### 5.2 NetStream Implementation Principle

## NetStream Overview

- NetStream is a network stream measurement technique by categorizing and measuring traffic on the traffic and utilization of the resources.
- NetStream is normally applied in the network for:-
  - Network charging
  - Network planning and analysis
  - Network monitoring
  - Application monitoring and analysis
  - User monitoring and analysis
- NetStream is based on a “stream”, which is referring to the packets from the same quintuple. NetStream will record the statistics of a stream, including the time stamp, packets, and bytes

- The rapid development of Internet offers users with larger bandwidth and predictable QoS. On the part of the users, they need more careful network management and charging. Therefore, there must be an appropriate technique to support such needs. NetStream is just such a measurement and release technique based on network stream information. It can categorize and measure the traffic on the network and the utilization of resources, and it performs management and charging for various services and based on different QoS. Thus, the following applications are provided:
  - **Charging:** NetStream provides accurate data for charging based on resources (for example, line, bandwidth, and time segment) utilization. These data include IP addresses, packets, bytes, time, TOS and application types. Internet service providers can use such information to enforce flexible charging strategies, for example, based on time, bandwidth, application, and QoS. Enterprise customers can use such information to calculate the expenses of each department or amortize the costs, for more efficient use of resources.
  - **Network planning and analysis:** NetStream can provide advanced network management tools with key information, to achieve the best network performance and reliability at the lowest operation cost by optimized network design and planning.
  - **Network monitoring:** NetStream can provide nearly real-time network monitoring. RMON, RMON-2 and stream-based analysis techniques can be used to visually represent the traffic model of a single router and the whole network, and provide proactive fault detection, efficient troubleshooting and rapid problem solution.
  - **Application monitoring and analysis:** With NetStream, detailed network application information can be obtained. For example, the network administrator can view the percentages of the traffic occupied by Web, FTP, Telnet and other well-known TCP/IP applications. Internet contents and service providers can plan and allocate the network resources based on such information to meet the users' needs.
  - **User monitoring and analysis:** NetStream enables the network operator to obtain the detailed information regarding the users' utilization of the network and application resources, and uses such information to effectively plan and allocate resources and guarantee safe operation of the network.

- NetStream is based on “stream”. A stream is composed of the packets from the same sub-interface, with the same source and destination IP address, protocol type, and same source and destination protocol port, and same ToS (usually referred to as the quintuple). NetStream will record the statistics of a stream, including the time stamp, packets, and bytes.

Huawei Training & Certification Huawei Training & Certification





## Contents

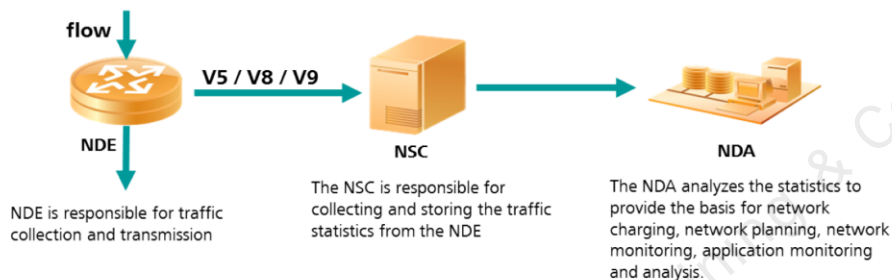
### 5. Introduction of SDN NetStream Protocol

#### 5.1 NetStream Overview

#### **5.2 NetStream Implementation Principle**

## NetStream Implementation Principles

- NetStream composes of 3 components:-
  - NetStream Data Exporter (NDE)
  - NetStream Collector (NSC),
  - NetStream Data Analyzer (NDA)

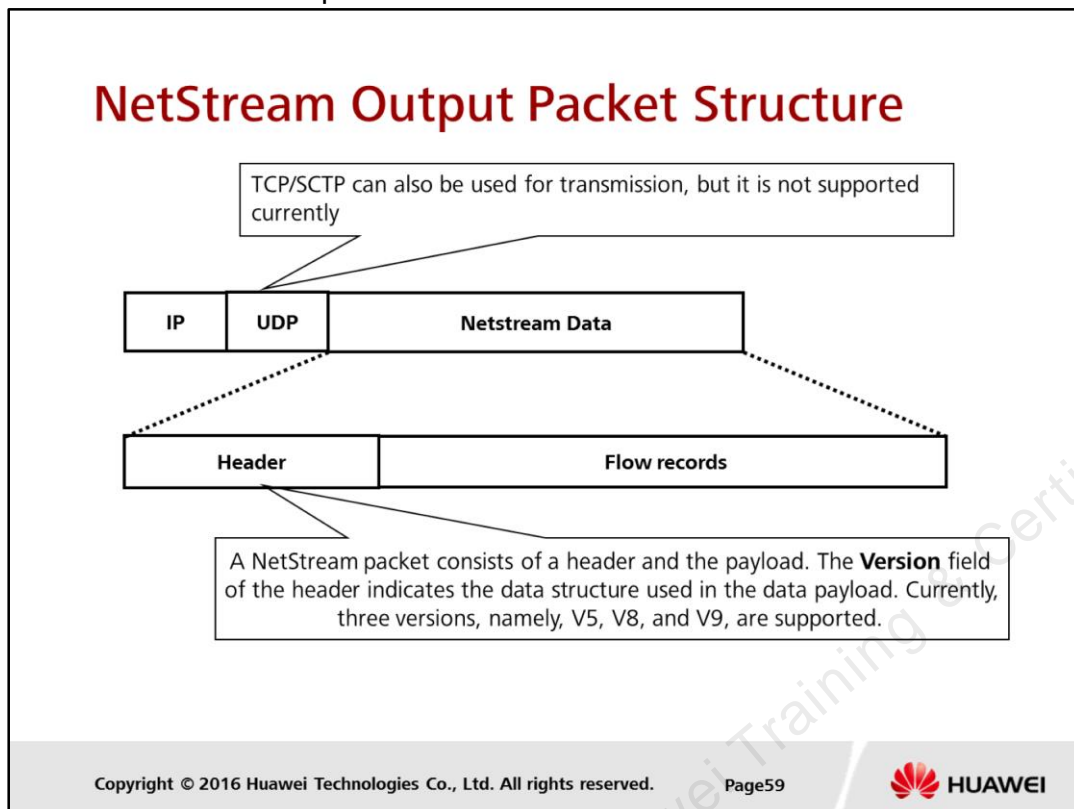


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page58



- The NDE is responsible for traffic collection and transmission. The NSC is responsible for collecting and storing the traffic statistics from the NDE. The NDA analyzes the statistics to provide the basis for network charging, network planning, network monitoring, application monitoring and analysis.
- **NetStream Data Exporter (NDE)**
  - It analyzes and processes the network stream, takes the stream statistics that meets the specified conditions, and outputs it to the NSC. Before the output, the NDE also perform some processing to the data, for example, aggregation.
- **NetStream Collector (NSC)**
  - As a UNIX application running on Solaris, it parses the packets of the NDE and stores the statistics to the database, for further analysis by the NDA. The NSC can collect the data outputted from multiple NetStream devices, and filter and aggregate them.
- **NetStream Data Analyzer (NDA)**
  - As a network traffic analyzer, it takes the statistics from the NDC for subsequent processing, and provides the basis for various services (for example, network planning and attack monitoring). With the graphical user interface, it is easy to use, allowing the users to easily obtain, display and analyze the data collected by the NSC.



- The statistics of the network stream collected by the NDE is encapsulated in the UDP packets for output to the NSC/NDA. One UDP packet can carry multiple statistics records. The formats of these statistics records are determined by the NDE equipment. After the NSC/NDA receive the statistics messages from the NDE, they first check the version of the records. Records of different versions have different formats. Currently, three versions of records are supported: V5, V8, and V9.

## NetStream Implementation Principles - Comparison of Packet Formats

V9

**V5**

- The packet format is fixed. The original data flow is generated on the basis of the septet.
- Advantages: The output fields are diverse. All the fields of the flow records before aggregation can be output to the NSC. In addition, the workload of the equipment is low.
- Disadvantages: The format is fixed and expansion is difficult. The NSC cannot store the large volume of generated data for a long time. The NSC and NDA are under great pressure.

**V8**

- The packet format is fixed and thus expansion is difficult.
- Advantages: The traffic is low. The contents carried are relatively simple. This format is suitable for specific analysis. New aggregation modes can be added.
- Disadvantages: The format is fixed and is not expandable. Only aggregated flow records can be output to the NSC. The equipment completes the aggregation and thus the workload of the equipment is high. Before a new aggregation mode can be added, the software versions of the NDE and NSC need to be upgraded.

- This format is template-based for easy expansion. Two types of data, namely, statistics data and option data, can be output.
- Advantages: The output mode is the most flexible and the format is changeable. The flow records before aggregation or those after aggregation can be output.

NetStream packet format

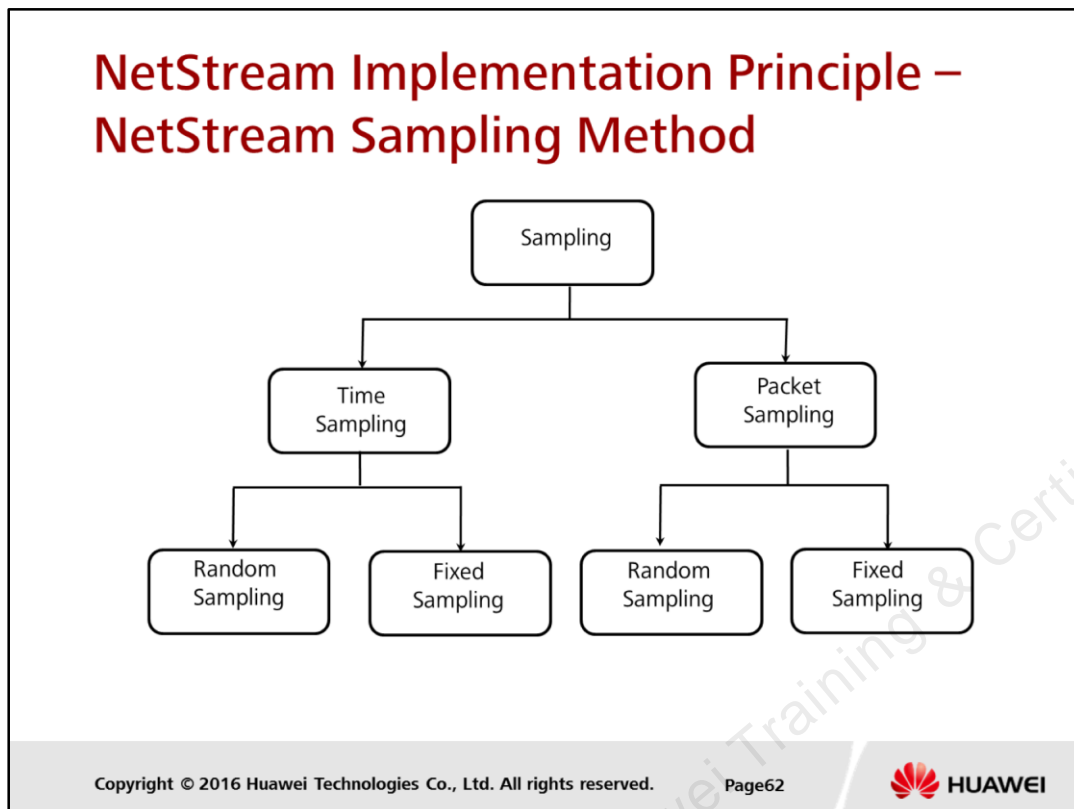
Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page60

- V5 is used to output flow details to the NSC/NDA.
- V8 is used to output aggregated flow information to the NSC/NDA. The common disadvantage of V5 and V8 is as follows. As new user requirements emerge, the NDE needs to expand output information. In this case, the software of the NSC/NDA needs to be modified to adapt to NDE changes. Thus, the NSC/NDA software of different vendors or even different versions of the NSC/NDA software of the same vendor fail to parse the statistics packets sent by the NDE.
- The most noticeable difference between V9 and earlier versions is that V9 is template-based. A template provides a flexible and expandable packet output format, which allows new flow statistics services to be added easily without changing the basic record format.

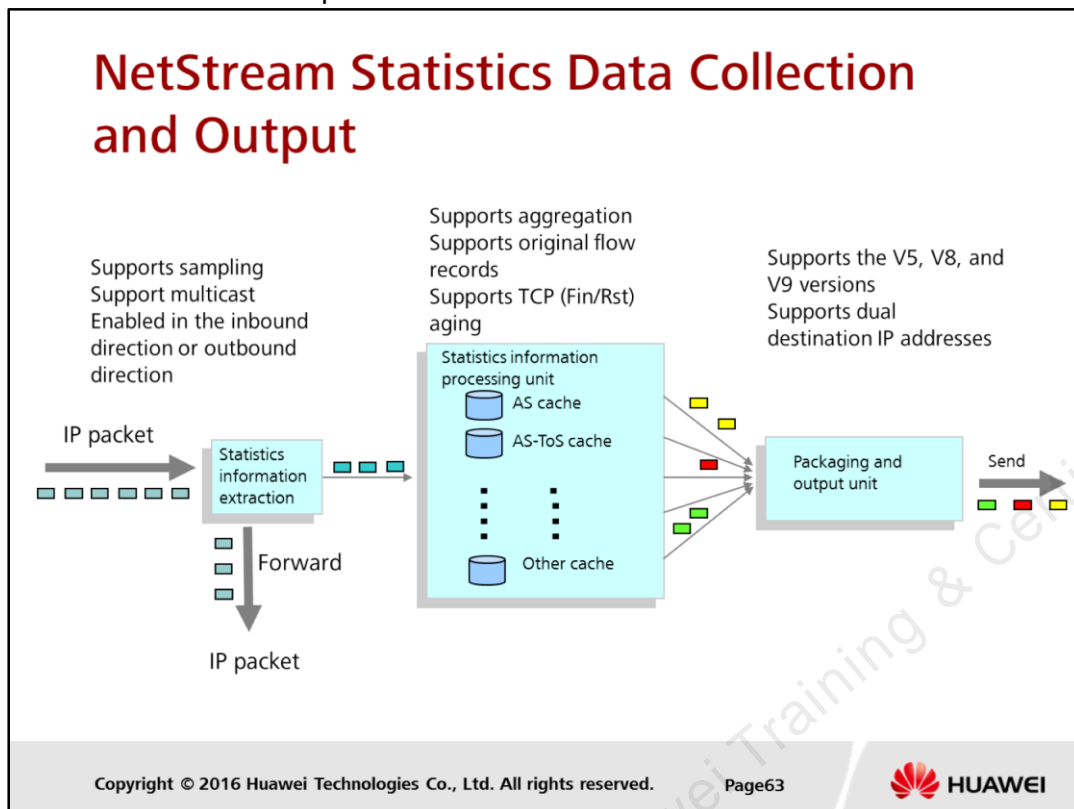
## NetStream Implementation Principle

- The NDE can output streams through 3 methods, as per listed below:-
  1. **Stream-by Stream:** Output the information of each stream to NSC equipment
  2. **Sampling:** Output the stream statistics information to NSC by ratio and NSC will restore the information to the original statistics based on the sampling rate.
  3. **Aggregation:** Combine the statistics information of the streams with the same attributes, for example, all packets between two autonomous domains, and the packets to the same destination etc.

- The NDE can output streams in one of the three methods: Stream-by-stream, sampling and aggregation. Stream-by-stream is the process to output the information of each stream to the NSC equipment. Sampling is the process to output the stream statistics information to the NSC by ratio, and the NSC will restore the information to the original statistics based on the sampling rate.
- The advantages of the stream-by-stream method are that the NSC can obtain the details of the stream, and can perform more flexible subsequent processing to these stream records. However, its disadvantage is as obvious in that the network bandwidth and CPU utilization are increased, enormous storage media space is required to store such information, and probably the users do not need the information in such great detail. Sampling can reduce the pressure on the storage media space, but the statistics information will suffer a certain degree of distortion.
- Is there is solution that solves the information explosion of the stream-by-stream method, while reducing distortion? Thus, aggregation is proposed. Aggregation is the process to combine the statistics information of the streams with the same attributes, for example, all the packets between two autonomous domains, and the packets to the same destination. On the NDE equipment, first the streams are aggregated, and then outputted to the NSC. This way, the network bandwidth, CPU utilization and storage media space are reduced greatly.



- The 4 types of NetStream sampling methods are explained in details below:-
  1. **Fixed packet sampling:** One packet is sampled out of every *fix-packets-number* packets. For example, if the *fix-packets-number* value is N, a NetStream-enabled interface samples every Nth packet that passes through a NetStream-enabled interface.
  2. **Random packet sampling:** One packet is randomly sampled out of every *random-packets-number* packets. For example, the sample ratio configured on a NetStream-enabled interface is M:1, and N packets out of N x M packets passing through the interface are sampled.
  3. **Fixed interval sampling:** One packet is sampled every *fix-time-value* ms.
  4. **Random interval sampling:** One packet is sampled every *random-time-value* ms, based on the sampling ratio.



- The figure above shows how a NDE collects and output a statistics information to NSC.
- The statistics information collection unit identifies the streams and outputs those that meet the pre-set conditions to the statistics information processing unit, which can aggregate the statistics information, and obsolete the stream records. The obsolete streams are assembled and outputted by the assembly and output unit.



## Summary

- OpenFlow protocol was first originated in Stamford University and developed together with Open Networking Foundation. It is used as the communication protocol between OpenFlow controller and forwarder.
- NETCONF is an extensible markup language (XML) based network configuration and management protocol.
- SNMP is an application layer protocol widely used in TCP/IP network for collecting, managing and modifying information of managed devices.
- REST is the software architectural style of the World Wide Web.
- NetStream is a network stream measurement technique by categorizing and measuring traffic on the traffic and utilization of the resources.

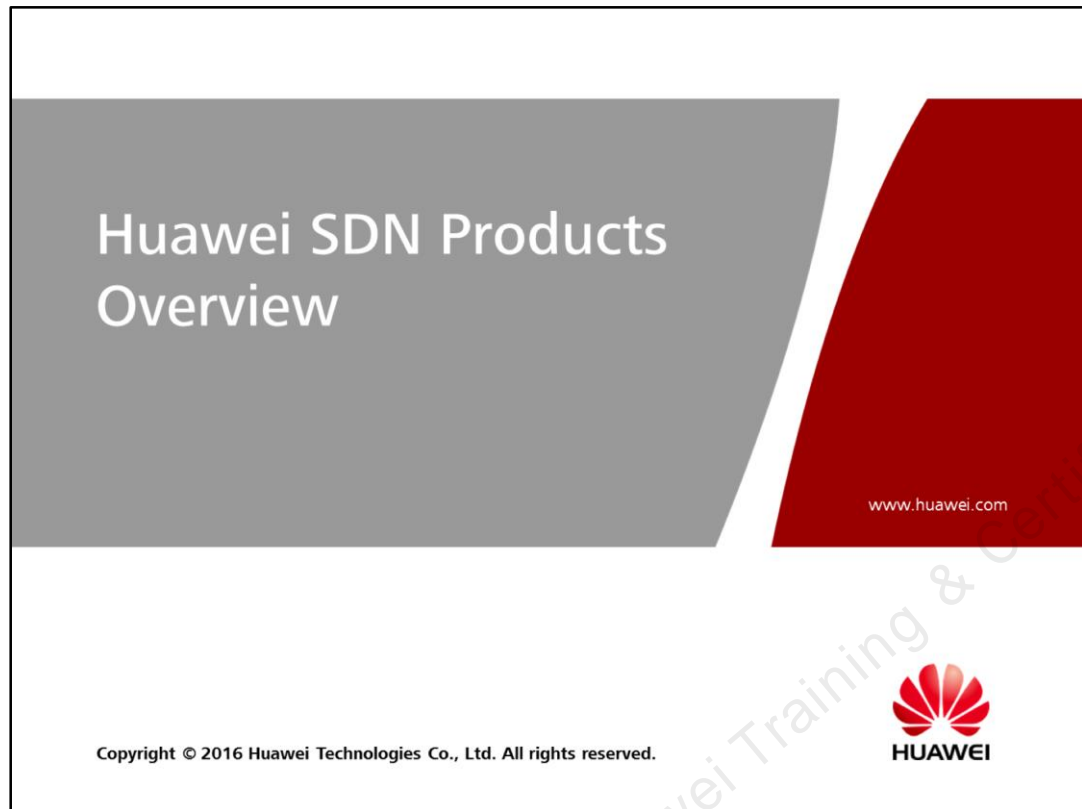


**Thank you**

[www.huawei.com](http://www.huawei.com)




Huawei Training & Certification Huawei Training & Certification



Huawei SDN Products  
Overview

www.huawei.com

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.



HUAWEI

Huawei Training & Certification



## Foreword

- The traditional Internet is consisting of switches, routers, terminals and the other devices. There are multiple issues existing in the traditional network, for example, slow service innovation, distributed network control etc. SDN is a technology which is blossoming in recent years to solve the existing issues in traditional network.



## Objectives

- Upon completion of this course, you will be able to:
  - Understand SDN Routers
  - Understand SDN CloudEngine Switches
  - Understand SDN Controller - Agile Controller



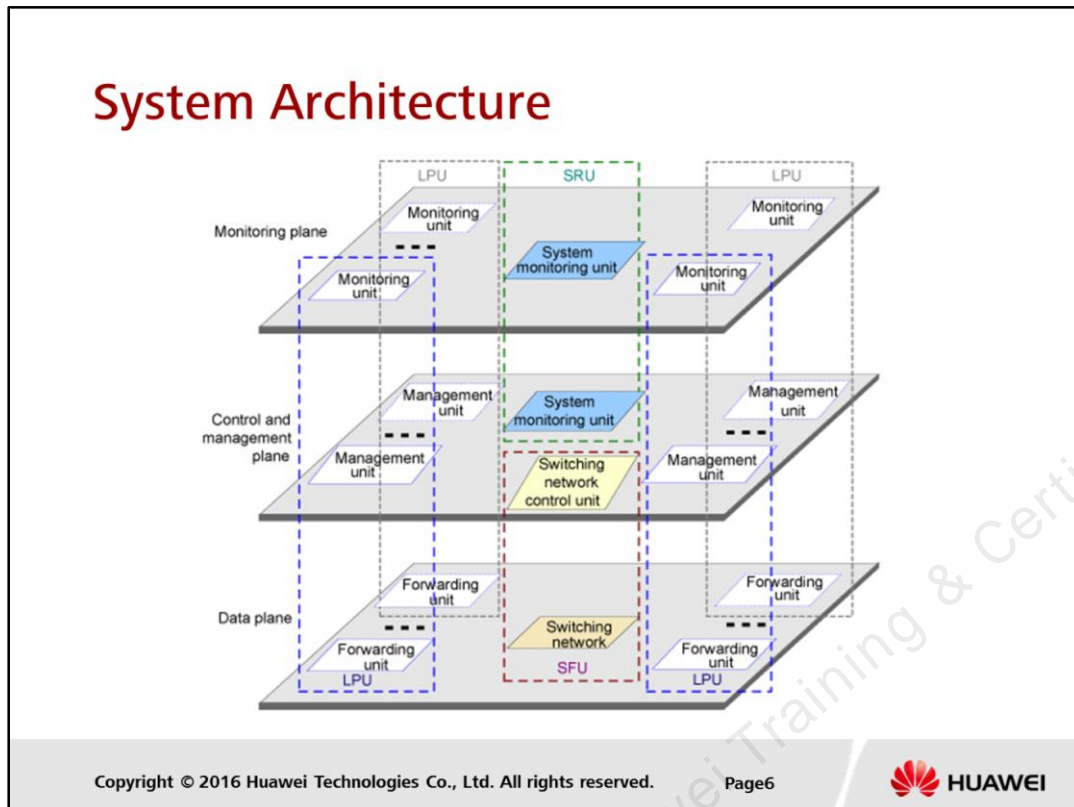
## Contents

1. Huawei SDN Routers Introduction
2. Huawei SDN CloudEngine Switches Introduction
3. Huawei SDN Controller Introduction

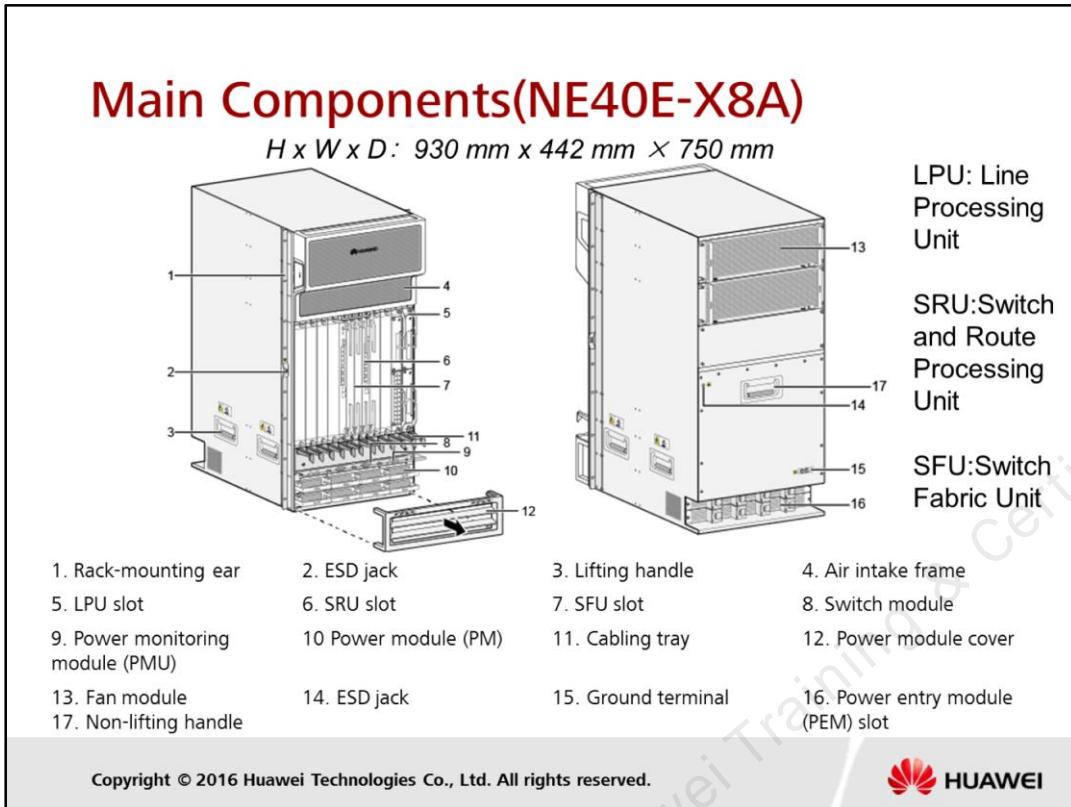


## Contents

- 1. Huawei SDN Routers Introduction**
2. Huawei SDN CloudEngine Switches Introduction
3. Huawei SDN Controller Introduction



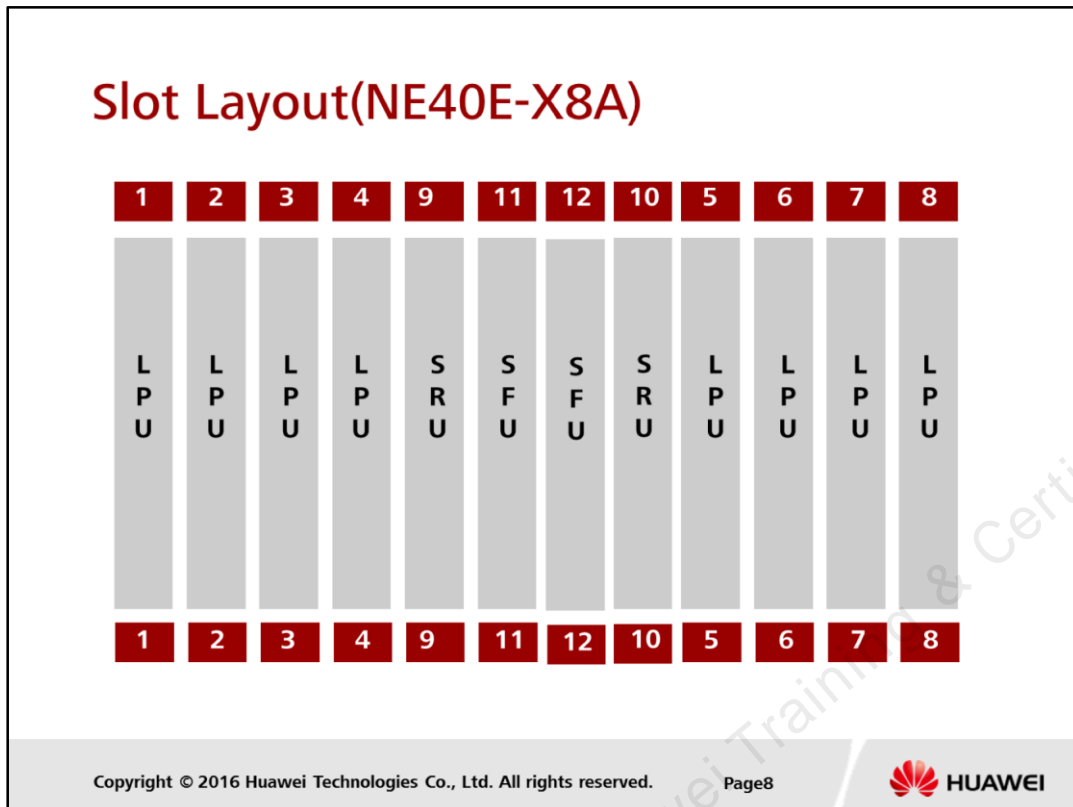
- The NE40E-X8A consists of the data plane, control and management plane, and monitoring plane, as shown in the figure above. This architecture improves system reliability and facilitates upgrades on each plane.
- The data plane is responsible for high speed processing and non-blocking switching of data packets. It encapsulates or decapsulates packets, forwards IPv4/IPv6/MPLS packets, performs QoS as well as scheduling and internal high-speed switching, and collects statistics.
- The control and management plane completes all control and management functions for the system and is the core of the entire system. Control and management units process protocols and signals, configure and manage the system, and display the system status.
- The monitoring plane monitors the ambient environment to ensure the secure and stable operation of the system. It detects voltage levels, controls system power-on and power-off, monitors the temperature, and controls fan modules. If a unit fails, the monitoring plane isolates the faulty unit promptly so the other units remain unaffected.



- Dimensions (H x W x D)

- 930 mm x 442 mm × 650 mm (chassis main body dimensions)
- (36.64 in. x 17.40 in. x 25.59 in.) Chassis main body dimensions
- 930 mm x 442 mm × 750 mm (chassis dimensions including the chassis's front and rear assembly and cable racks)
- (36.64 in. x 17.40 in. x 29.53 in.) Chassis dimensions including the chassis's front and rear assembly and cable racks





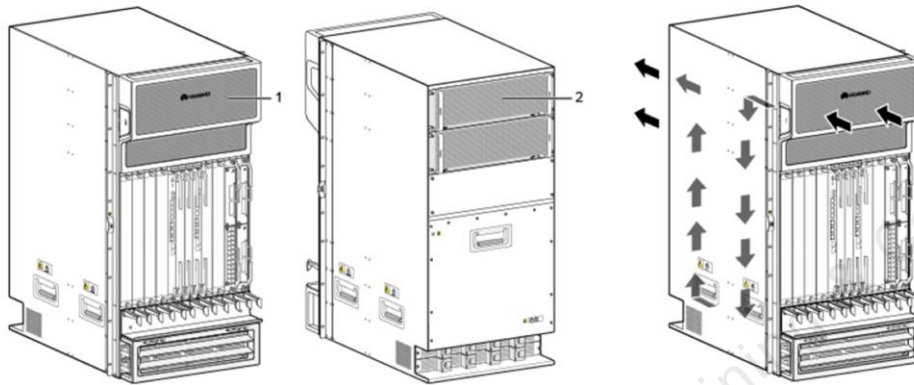
- SRUs work in 1:1 backup mode
- SFUs work in 3+1 backup mode

## System configuration of the NE40E-X8A

Description	Typical Configuration	Remarks
Processor	Dominant frequency: 2.0 GHz	Four core.
Boot ROM	16 MB	-
SDRAM	8 GB	Can be extended with an additional 16 GB.
NVRAM	512 K	-
Flash	32 MB	-
SSD	8 GB	-
Switching capacity	12.58 T (bidirectional)	-
Backplane capacity	50 Tbps	-
Interface capacity	7.68 T (bidirectional)	-
Number of LPU slots	8	-
Number of SRU slots	2	-
Number of SFU slots	2	-

## System Air Channel(NE40E-X8A)

- The NE40E-X8A supports a maximum of two fan modules. Each fan module consists of six fans.



1. Air intake vent

2. Fan module

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page10



- The system draws air from the front and discharges air from the back. The air intake vent resides above the board area on the front chassis; the air exhaust vent resides above the board area on the rear chassis.
- The air filters on the air intake vents are vertically installed and featured with the curved face, large area, and small windage resistance, helping to improve the heat dissipation efficiency.

## Fan Module and Air Filter(NE40E-X8A)

- The fan module is located on the air exhaust vent. Each fan module contains six fans. When a single fan fails, the heat dissipation system enables the system to work at an ambient temperature of 40° C (104° F) for a short period.

Fan module



Air filter



- The black sponge air filter on the air intake vent is used to prevent dust from getting into the system.
- To ensure good heat dissipation and ventilation for the system and to prevent the accumulation of dust on an air filter, you need to clean the air filter regularly. It is recommended that an air filter be cleaned at least once every three months and be replaced once every year. When an air filter is placed in the dusty environment, it needs to be cleaned more frequently.

## Fan Speed Adjustment(NE40E-X8A)

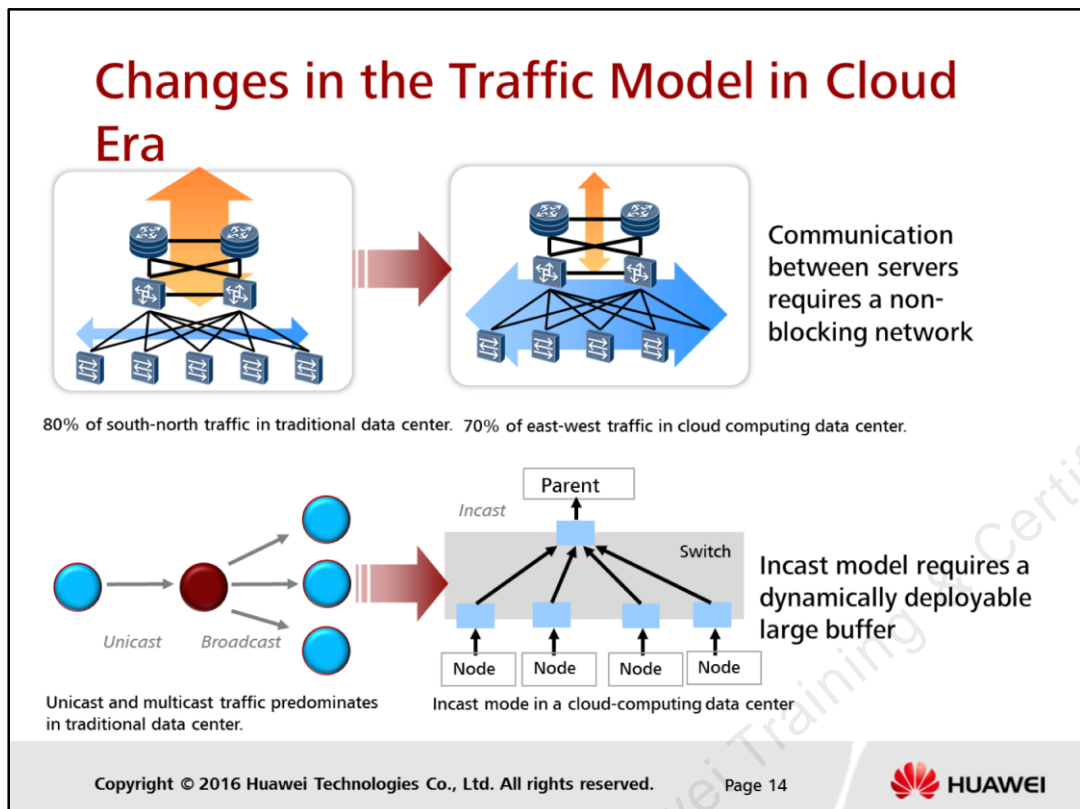
- When the system is configured to the largest capacity, the fan rotation speeds are adjusted based on the temperatures reported by the sensors on the SRUs, LPUs, and SFUs.

Ambient Temperature	Rotational Speed	Noise/Heat Control
Below 27°C (75.2°F)	Low rotation speed (50%)	When the ambient temperature is below 27°C (75.2°F), fans rotate at a fixed low speed, which meets the ETSI noise requirement and the heat dissipation requirement.
27°C-45°C (75.2°F-113°F)	Linear speed adjustment	When the ambient temperature is between 27°C (75.2°F) and 45°C (113°F), the fan rotation speeds are adjusted smoothly in linear mode, and the fan noise does not change violently.
Above 45°C (113°F)	High rotation speed (100%)	When the ambient temperature is above 45°C (113°F), fans rotate at a fixed high speed, which meets heat dissipation requirements.



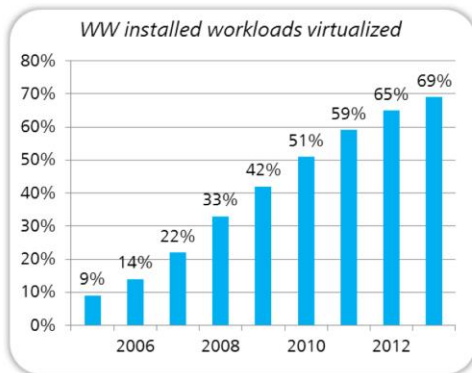
## Contents

1. Huawei SDN Routers Introduction
- 2. Huawei SDN CloudEngine Switches Introduction**
3. Huawei SDN Controller Introduction



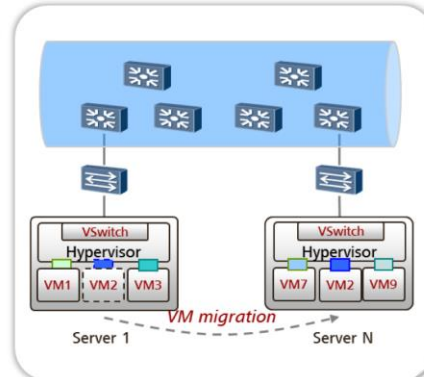
- Communication between servers requires a non-blocking network
  - Parallel computing, 3D rendering, and search services require collaboration between server clusters, resulting in a large amount of east-to-west traffic.
  - VMs need to synchronize a large amount of data in real time to support flexible deployment and dynamic migration.
- Incast model requires a dynamically deployable large buffer
  - In collaborative computing, multiple nodes in a cluster send TCP data to the master node simultaneously.
  - This multipoint-to-point communication causes traffic bursts and congests the network.

## Server Virtualization in Cloud Era



### Era of server virtualization is coming

- By 2013, **69%** of computing will be completed on VMs
- Customers start to lease **virtual servers** or even application programs instead of physical servers.

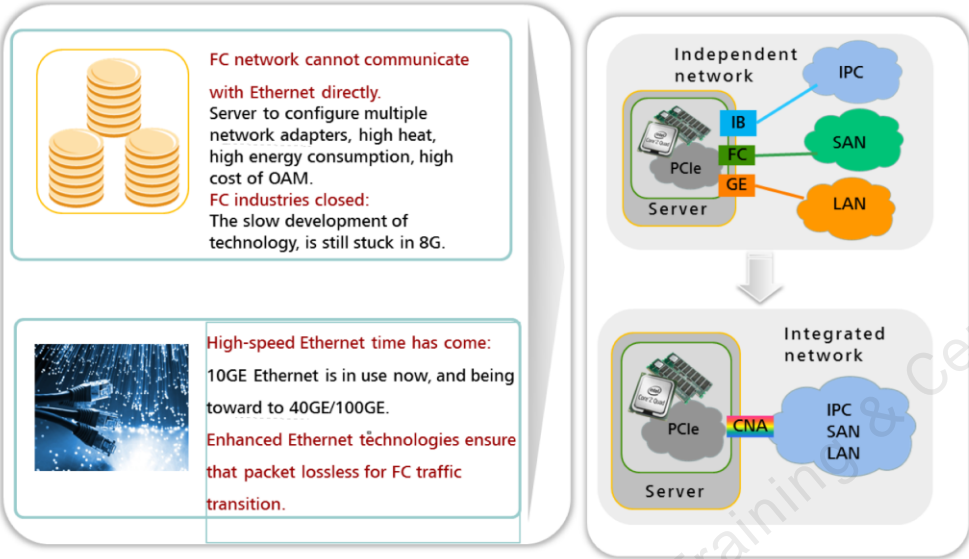


### Challenges of VMs maintenance

- Vswitch in the server based on software, consume CPU resource, and difficult to position the failure.
- By the growth of VM number, need larger L2 network, VM migration need the L2 communication in multi-sites.



# Network Merge Requirement in Cloud Era



Huawei Training & Certification

# Cloud Fabric Data Center Network Solutions

## Scalable Fabric

- High speed line cards:
  - 4 x 100 GE (Q2 2013) and 8 x 100 GE (Q4 2013)/
  - 24 x 40 GE/96 x 10 GE
- Super large capacity, three times the industry average
- 360 Tbps non-blocking network
- Support four generations of servers in a 10 year lifecycle

## Virtualized Fabric

- Network virtualization: 1:8 VS (Q2 2013) and 4:1 CSS (Q2 2013) reduces investment
- TRILL: Large L2 network for VM migration
- nCenter for rapid batch VM migration

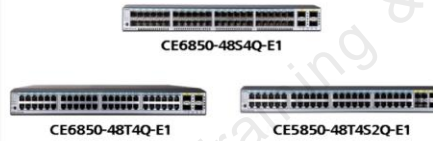
## Converged Fabric

- FCoE helps build a converged network to reduce TCO
- Centralized FCoE gateway deployment simplifies network management and design
- DCB helps build a lossless Ethernet network

### Core Switches



### TOR Switches



## Industry Leading Data Center Switch



8\*100GE CXP line card

24\*40GE/96\*10GE line card

	Huawei CE12800 series switch		Industry Avg.
Switch Capacity	48Tbps	3 <sup>x</sup>	18Tbps
Line card capability	1Tbps	2 <sup>x</sup>	480Gbps
Throughput per slot	2Tbps	4 <sup>x</sup>	500Gbps



Industry leading performance, meet future 10 years requirement

Huawei Training & Certification

## Flexible Access CE Series TOR Switches

### All 40GE Uplink TOR Switches

•40GE Uplink 10GE Base-T  
TOR  
48\*10GE Base-T, and 4\*40GE  
QSFP+



CE6850-48T4Q-EI

•40GE Uplink 10GE SFP+/SFP  
TOR  
48\*10GE SFP/SFP+ and 4\*40 QSFP+



CE6850-48S4Q-EI

### GE TOR: 40GE per port non-blocking stack

•40GE Uplink GE TOR  
48\*GE Base-T, 4\*10GE  
SFP/SFP+ ,and 2\*40GE QSFP+

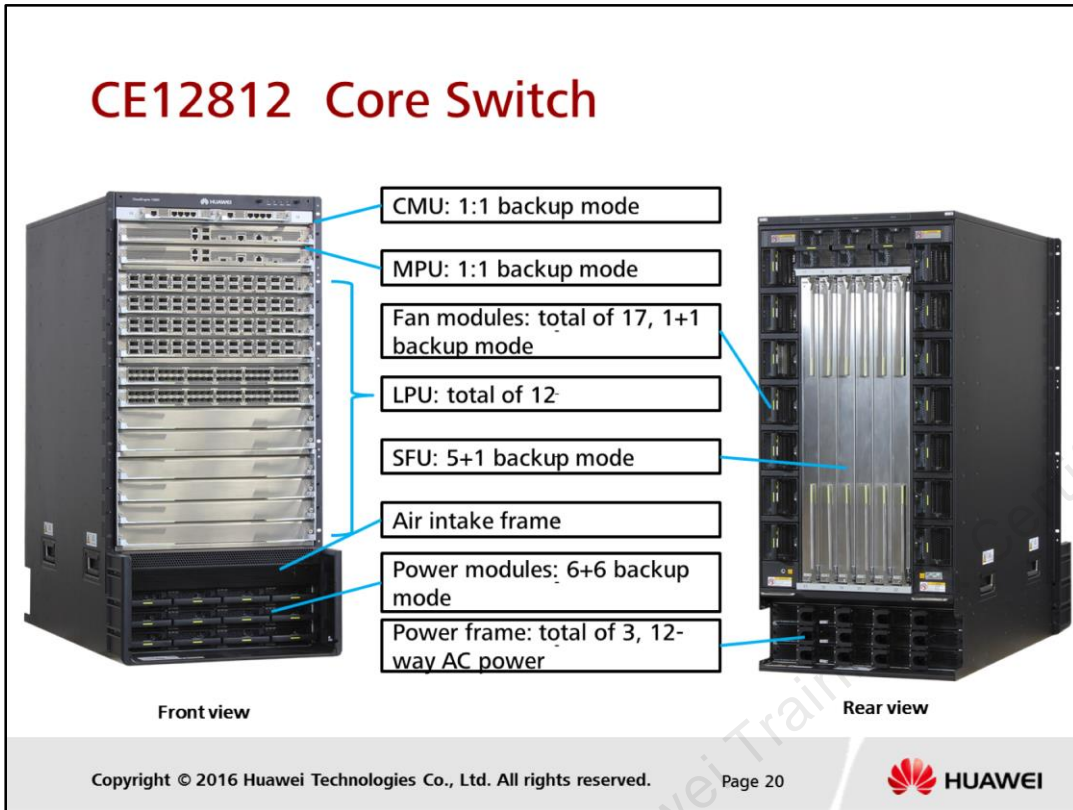


CE 5850-48T4S2Q-EI

1\*40GE = 4\*10GE

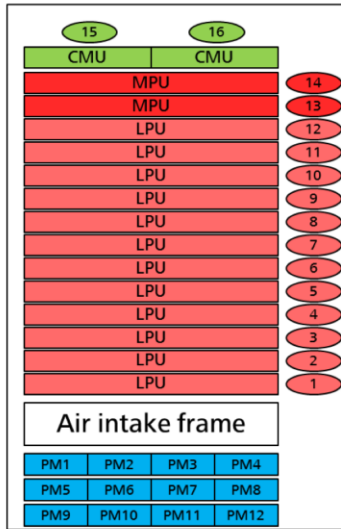


QSFP+  
1:4 breakout

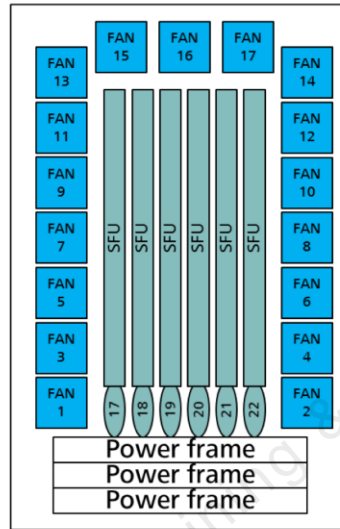


- CMU: Centralized Monitoring Unit
- MPU: Main Processing Unit
- LPU: Line Process Unit
- SFU: Switch Fabric Unit

## CE12812 Slot Distribution Diagram

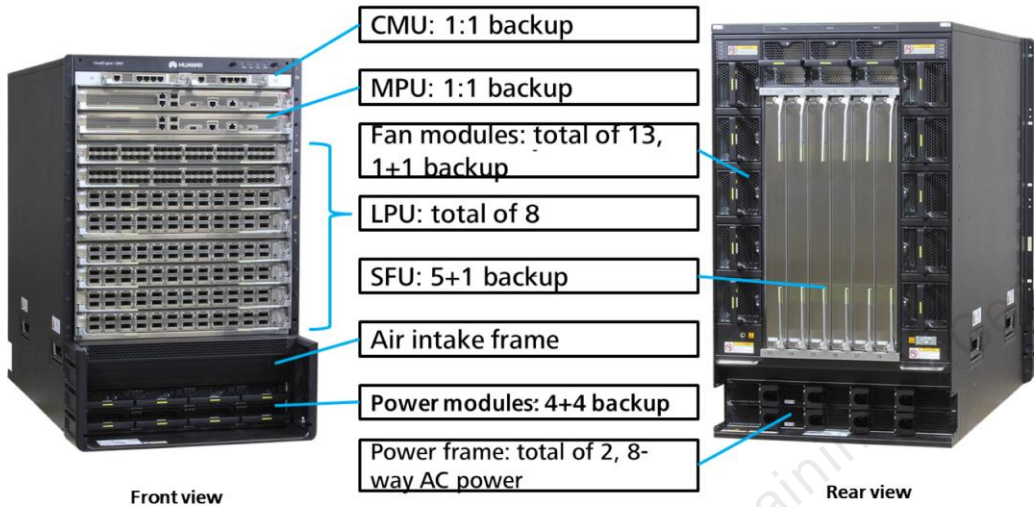


CE12812 front view



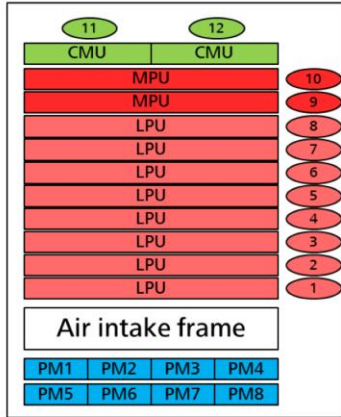
CE12812 rear view

## CE12808 Core Switch

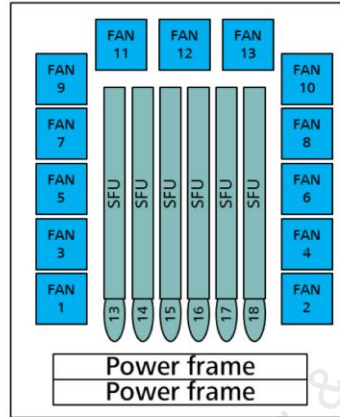


Huawei Training & Certification Huawei Training & Certification

## CE12808 Slot Distribution Diagram



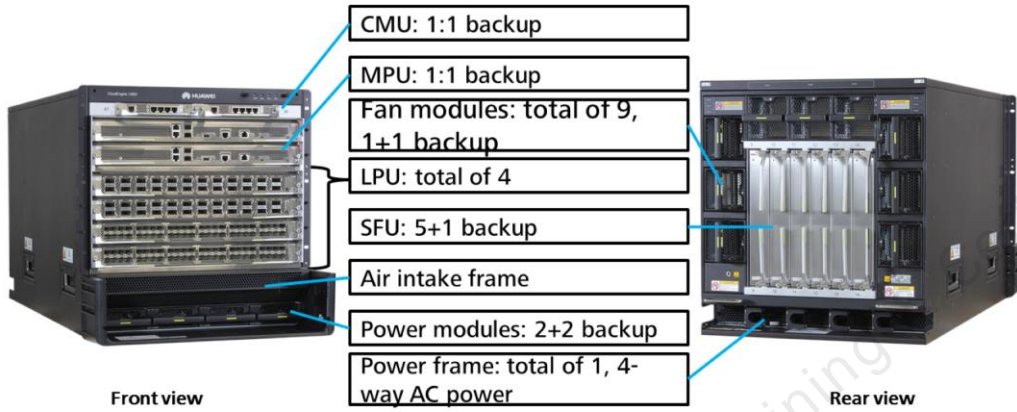
CE12808 front view



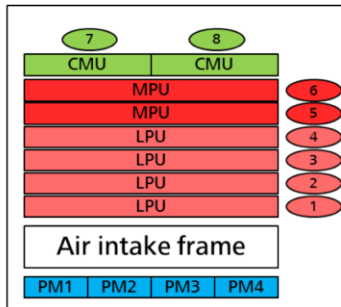
CE12808 rear view



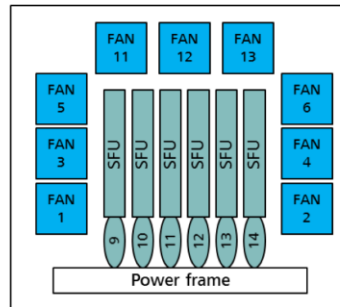
## CE12804 Core Switch



## CE12804 Slot Distribution Diagram



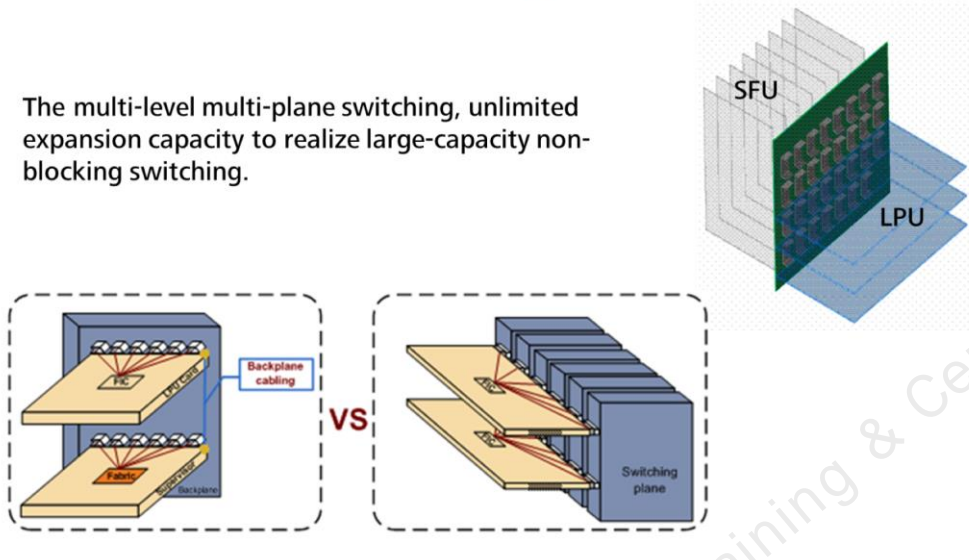
CE12804 front view




CE12804 rear view

## CE12800 Non-blocking CLOS Architecture

The multi-level multi-plane switching, unlimited expansion capacity to realize large-capacity non-blocking switching.



Traditional Architecture      **VS**      Orthogonal Architecture


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.      Page 26       **HUAWEI**

- On core switches, cabling between line cards and switch fabric units is an important factor that determines per slot bandwidth. A longer backplane cable and a higher rate indicate a greater loss.
- The CE12800 uses an orthogonal architecture, which does not require backplane cables. This architecture greatly increases system bandwidth and improves the evolution capability. With an orthogonal design (three-level Clos architecture), service line cards and switch fabric units of the CE12800 constitute a multi-level multi-plane switch fabric. This switch fabric allows for unlimited capacity expansion, helping implement large-scale non-blocking switching in data centers.
- Clos architecture has multiple levels, at each of which a switch unit is connected to all switch units at a lower level. Clos architecture is non-blocking, re-arrangeable, and scalable.


## CE12800 Series Switches Main Parameters

Items	CE12804	CE12808	CE12812
Switching capability	16Tbps	32Tbps	48Tbps
Packet forwarding	4800Mpps	9600Mpps	14400Mpps
Number of slot	4	8	12
Size (W×D×Hmm)	442×938×486	442×938×752	442×938×975
Weight (chassis)	<75Kg	<90Kg	<110Kg
Operating voltage	AC: 90V~290V		
Maximum power	≤5400W	≤10800W	≤16200W


## CE Series TOR Switches




**CE6850-48S4Q-EI**



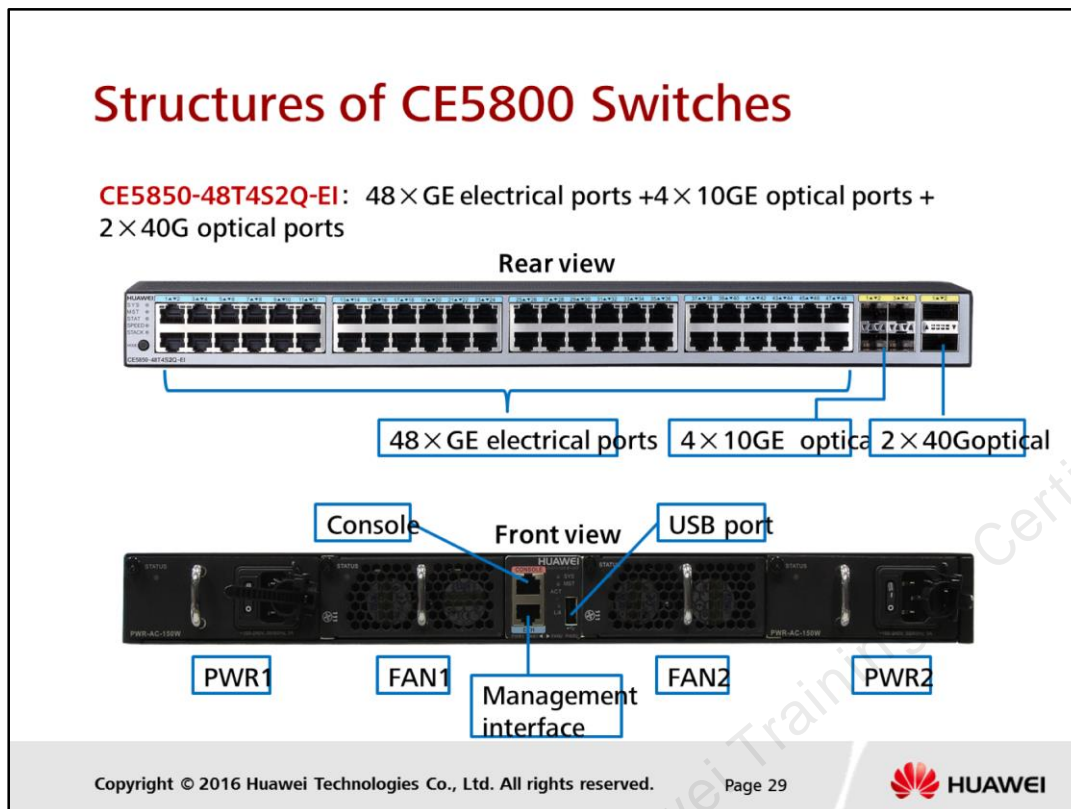
**CE6850-48T4Q-EI**



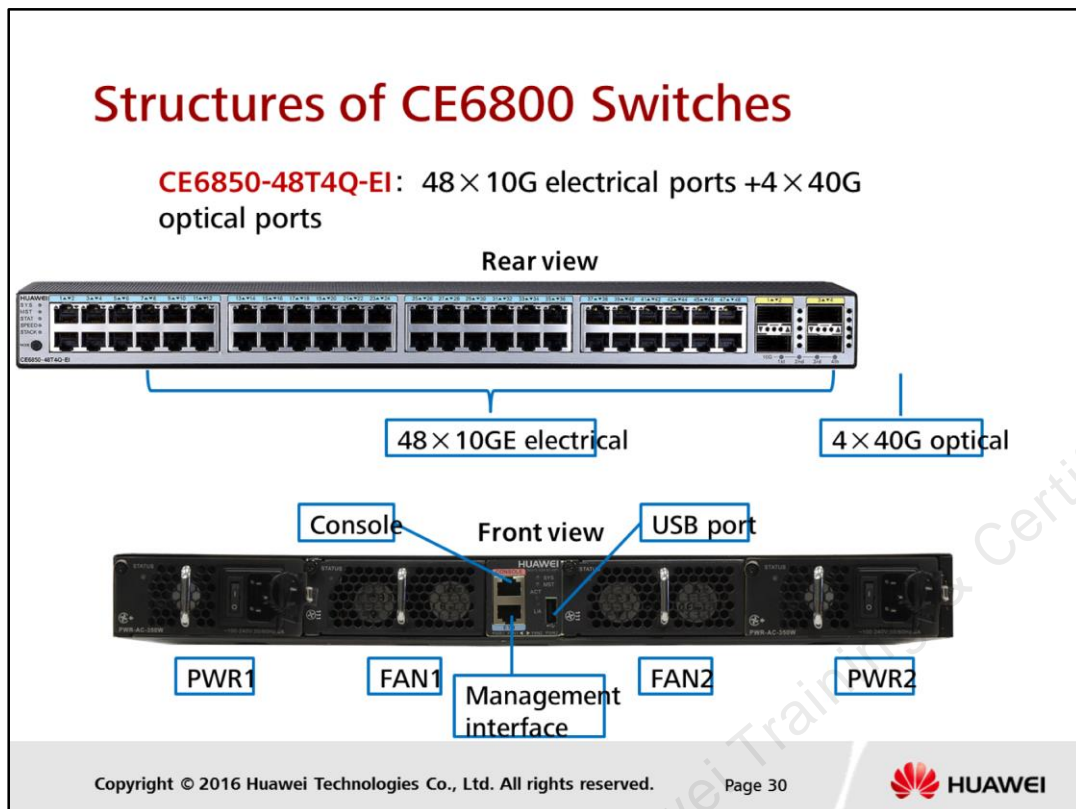
**CE5850-48T4S2Q-EI**

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page 28 

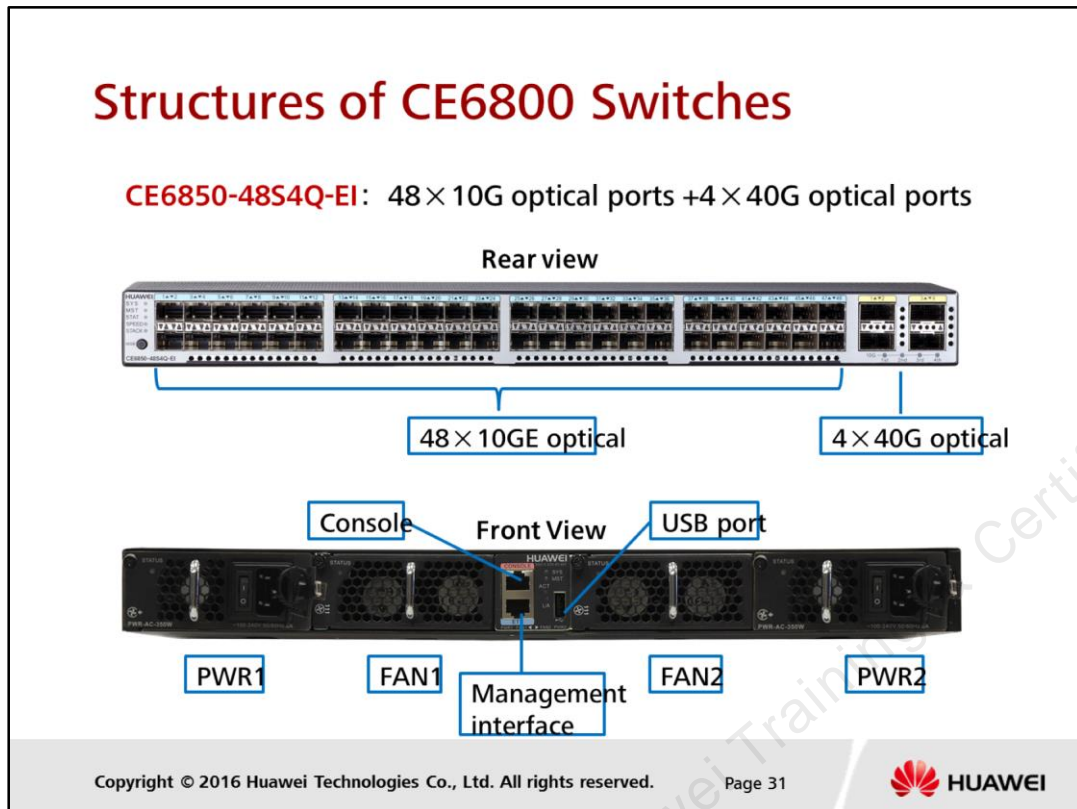
- Huawei CE6800/5800 series switches are next-generation data center switches designed for high-performance data centers. The CE6800 series switches support 10GE access, while the CE5800 series switches support GE access.
- The CE6800/5800 series switches function as access switches, servers connect to CE6800/5800 switches through GE/10GE uplinks; and CE6800/5800 switches connect to core switches CE12800 through 10GE/40GE uplinks.
- CE6800/5800 switches provide high-performance 40GE ports, which can connect to high-density 40GE line processing units (LPUs) on CE12800 switches to construct full-40G data center networks.



- CE6800/5800 switches use cutting-edge hardware platforms in the industry. By using a 1 U (1 U = 44.45 mm) box, CE6800/5800 switches provide high port densities and line-rate forwarding capabilities. Next-generation, high-performance servers in super high density arrangements can easily connect to CE6800/5800 switches.
- CE5850-48T4S2Q-EI: Provides forty-eight 10/100/1000BASE-T Ethernet ports, four 10G SFP+ Ethernet optical ports, and two 40G QSFP+ Ethernet optical ports.



- CE6850-48T4Q-EI: Provides forty-eight 10G BASE-T Ethernet ports and four 40G QSFP+ Ethernet optical ports
- The 1.28Tbps exchange capacity, the industry's highest performance; 960Mpps forwarding performance, L2/L3 wire-speed forwarding. 64 10GE interfaces at most, the industry's highest density, to meet Gigabit server density access needs. Supports four 40G high-performance QSFP+ interface, and the QSFP+ can be used as four 10GE interfaces to allows flexible networking capability; 40GE uplink and CE12800 series work together to build a non-blocking network platform.



- CE6850-48S4Q-EI: Provides forty-eight 10G SFP+ Ethernet optical ports and four 40G QSFP+ Ethernet optical ports.



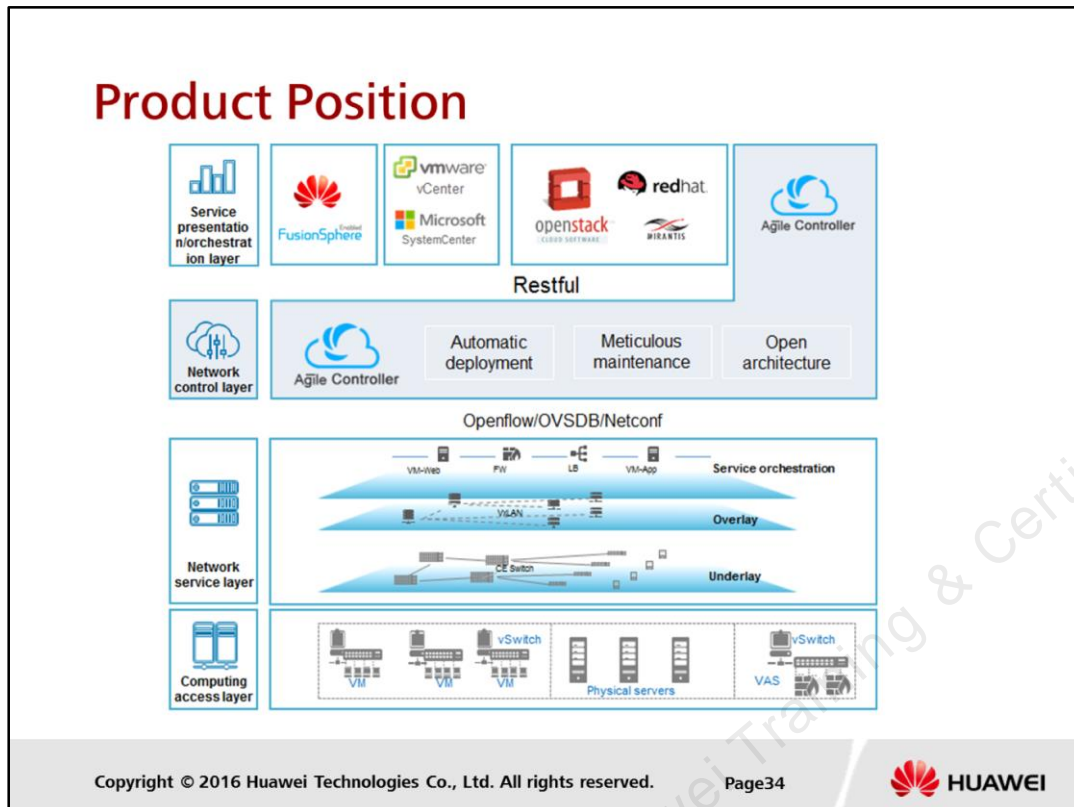
## CE5000/6800 Main Parameters

Items	CE6850-48T4Q-EI	CE6850-48S4Q-EI	CE5850-48T4S2Q-EI
Port describe	48个10GE Base-T	48个10GE SFP+	48个10/100/1000BASE-T
	4个40GE QSFP+	4个40GE QSFP+	4个10GE SFP+, 2个40GE QSFP+
Switching capability	1.28Tbps		336Gbps
Packet forwarding	960Mpps		252Mpps
Size mm (WxDxH)	440*600*43.6	440*600*43.6	440*420*43.6
Weight	10Kg	10Kg	8Kg
Operating voltage	Rated voltage range : 100V ~ 240V AC ; 50~60Hz		Rated voltage range : 100V ~ 240V AC ; 50~60Hz
	Maximum voltage range : 90V ~ 290V AC ; 45~65Hz		Maximum voltage range : 90V ~ 290V AC ; 45~65Hz
Maximum power	≤320W	≤350W	≤150W



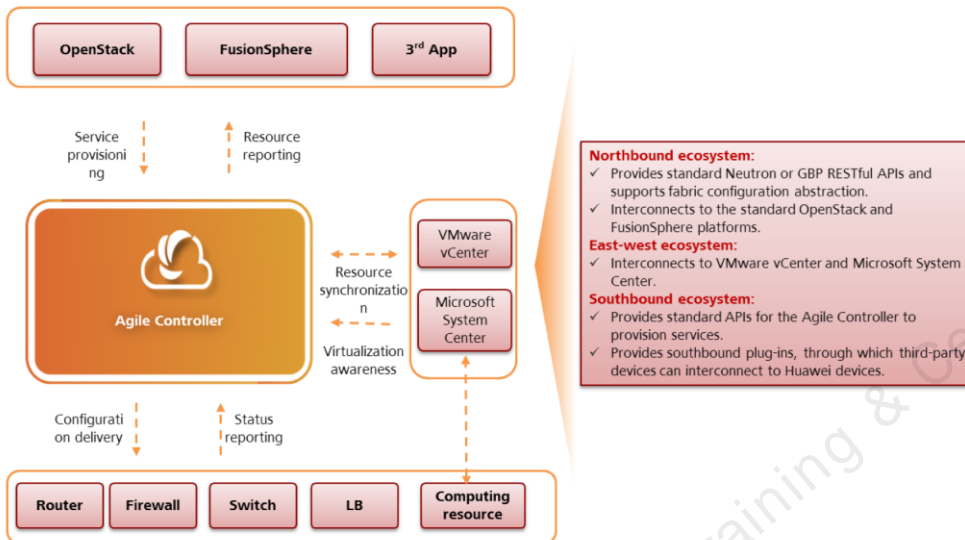
## Contents

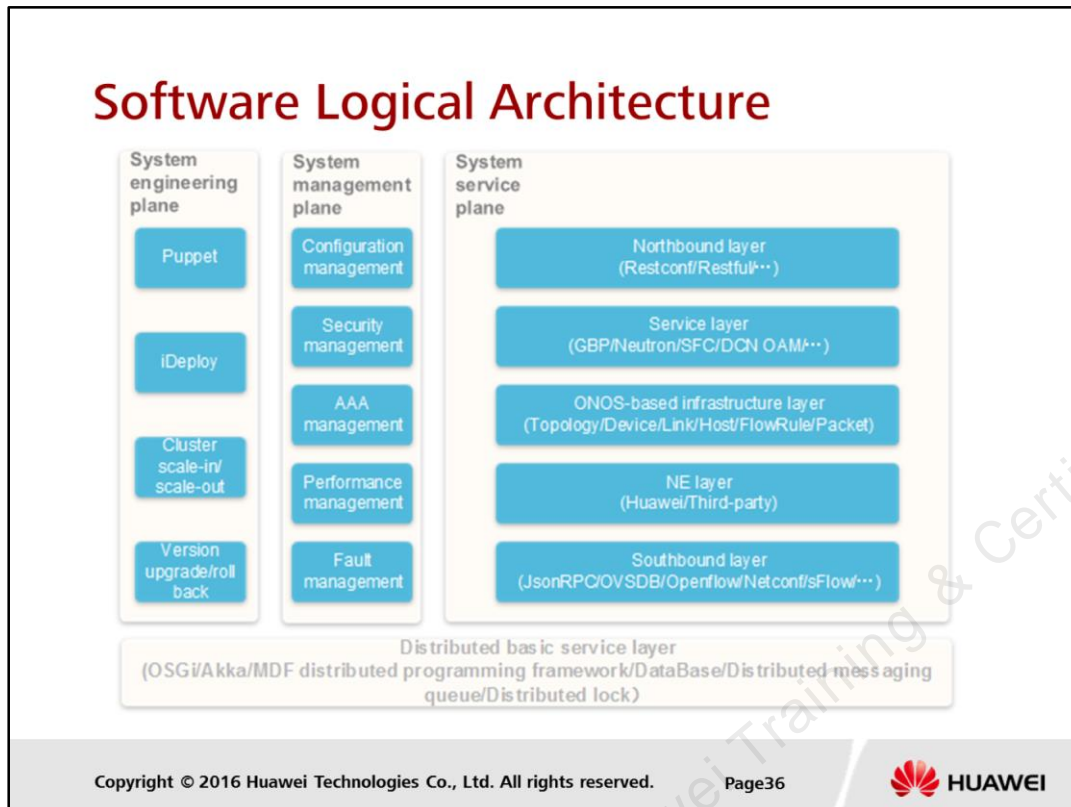
1. Huawei SDN Routers Introduction
2. Huawei SDN CloudEngine Switches Introduction
- 3. Huawei SDN Controller Introduction**



- The Agile Controller-DCN is deployed at the network control layer of the Huawei CloudFabric DCN solution. Based on the open and high reliable distributed cluster architecture, the Agile Controller-DCN can interconnect with mainstream cloud platforms and computing management platforms to provide capabilities such as automatic deployment, refined O&M, and multi-DC disaster recovery (DR).

## Open Architecture for Building an Open-Source SDN Network Ecosystem



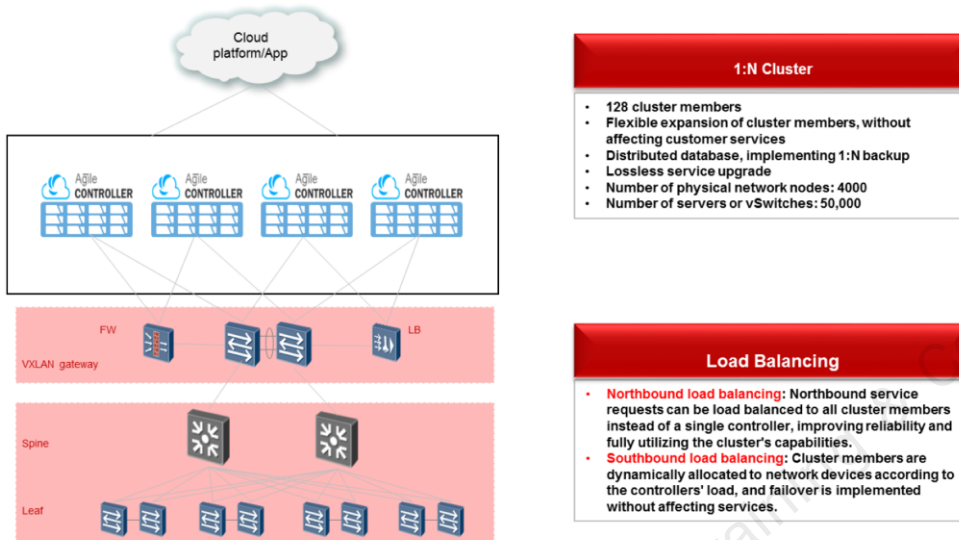


- Distributed basic service layer
  - This layer provides basic middleware services for SDN-based distributed programming, including the OSGi container, Akka cluster management, distributed cache, distributed database cache, and distributed lock. The OSGi container is provided by the Open Network Operating System (ONOS) platform; Akka cluster management is implemented on the OpenDaylight (ODL) platform; other distributed services are enhanced commercial functions developed using mainstream open-source components. This layer provides reliability, performance, and security.
  - The distributed model driven framework (MDF) provides a modular service architecture based on ODL model-driven service abstraction layer (MD-SAL) to ensure separated running and scheduling of processes and threads of various service protocols. This framework is compatible with MD-SAL interfaces to support enhanced functions, such as synchronous/asynchronous RPC encapsulation, routed RPC performance optimization, and high-performance DOM storage. The MDF framework integrates Kafka-based distributed messaging service bus and distributed event management capability, providing reliability and performance.
- System engineering plane: This plane provides functions such as the Agile Controller-DCN cluster installation, deployment, scale-in, scale-out, upgrade.
- System management plane: This plane provides system management capabilities, including configuration management, security management, AAA management, service performance monitoring, and fault management.
- System service plane: This plane is the key for Agile Controller-DCN service implementation. It collects network resources in the southbound and abstracts them for unified display and provides open northbound interfaces to provision services to SDN networks. This plane can be further divided into the following layers:
  - Northbound layer: This layer processes access protocols, manages security of northbound interfaces, and completes automatic service provisioning in the background.
  - Service layer: Various service applications required in DCN SDN solutions are deployed in this layer to provide network-level service capabilities, such as DCN network service provisioning, Neutron, service function chain (SFC), and DCN operation, administration, and maintenance (OAM).
  - Infrastructure layer: This layer is built based on the ONOS core to provide management and control services on standard or abstracted SDN core resources for upper-layer apps. The resources include devices, links, topology, hosts, packet transmit and receive.
  - NE layer: This layer provides the device abstraction level (DAL) capability to the upper layer and uses the NE driver and device driver plug-in to realize adaptation and conversion of driver models.
  - Southbound layer: This layer supports comprehensive standard DCN SDN southbound protocols, including configuration management protocols NETCONF, JSON RPC, and SNMP, the device control protocol OpenFlow, and flow data collection protocol sFlow.

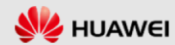
## Performance Specifications

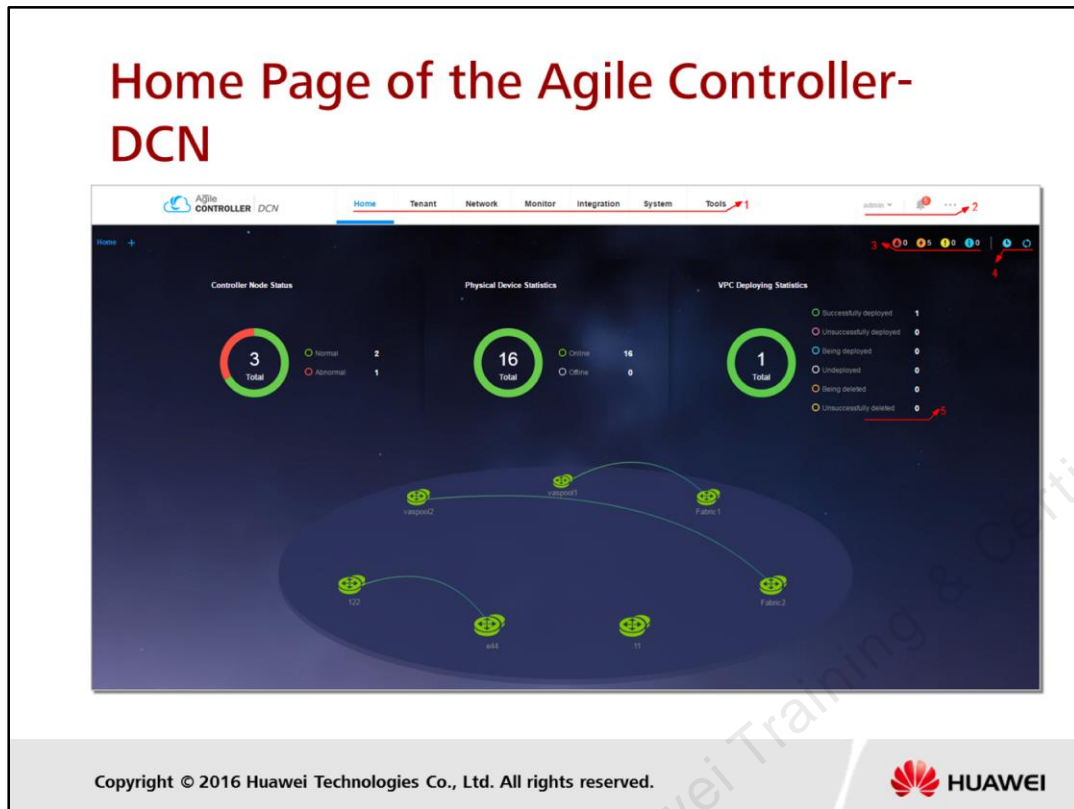
Item	Capability
Number of Agile Controller-DCN cluster nodes	128
Number of physical servers in an Agile Controller-DCN cluster	64K
Number of VMs in an Agile Controller-DCN cluster	300K
Number of fabrics in an Agile Controller-DCN cluster	32
Number of VAS pools in an Agile Controller-DCN cluster	32 Supported types: FW, LB, and DHCP
Number of tenants supported by an Agile Controller-DCN cluster	30K
Number of VPC instances supported by an Agile Controller-DCN cluster	30K
Bandwidth between the Agile Controller-DCN and clients	3Mbit/s/Client
Bandwidth between the Agile Controller-DCN and NEs	1G

## Industry's Largest Controller Cluster Capacity (128 Nodes)



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.





- 1. Main menu area: Function as a main entrance to different functions of the Agile Controller-DCN.
- 2. General information area: Displays the following common information about the Agile Controller-DCN from left to right
  - **User Name:** Indicates the user name for logging in to the Agile Controller-DCN.
  - **Logout:** used for logging out of the Agile Controller-DCN or switching login account.
  - **Notification Icon:** Click the notification icon to display the Agile Controller-DCN alarms.
  - **Others**
- 3. Alarm indicator area: Shows severity and entries of uncleared alarms of the Agile Controller-DCN. They are critical alarm number, major alarm number, minor alarm number, and warning number, from left to right.
- 4. Statistics information fresh area: Displays refreshed device status and services on the statistic graph.
- Statistics area: Shows device statuses and services, including measurement of network device statuses and physical network device statuses.



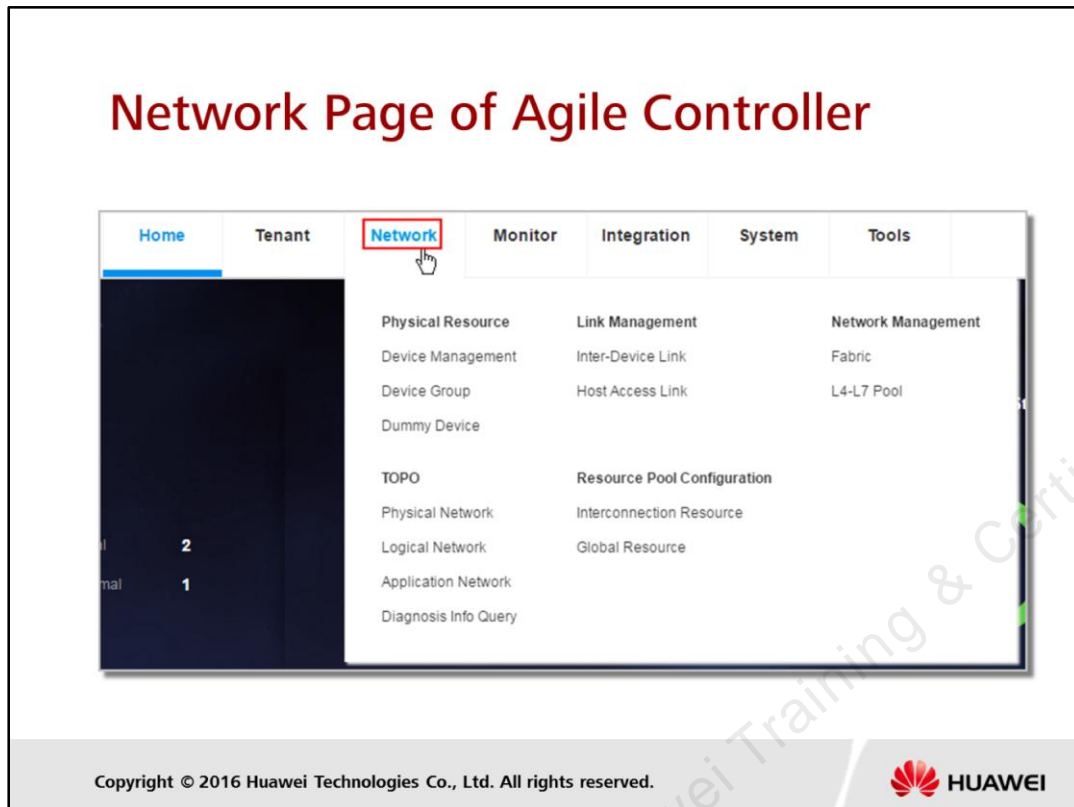
## Tenant Page of Agile Controller

The Second-level Menu	Menu Description
Tenants Management	You can manage tenant information. You can create a local tenant or synchronize the tenant list form a cloud management platform.
Public Service	You can create a public service. Public service indicates services defined by a system administrator and can be used by multiple tenants.
Public Template	You can create a public profile. Public profile is a service parameter profile defined by a system administrator to facilitate service provisioning for tenants.

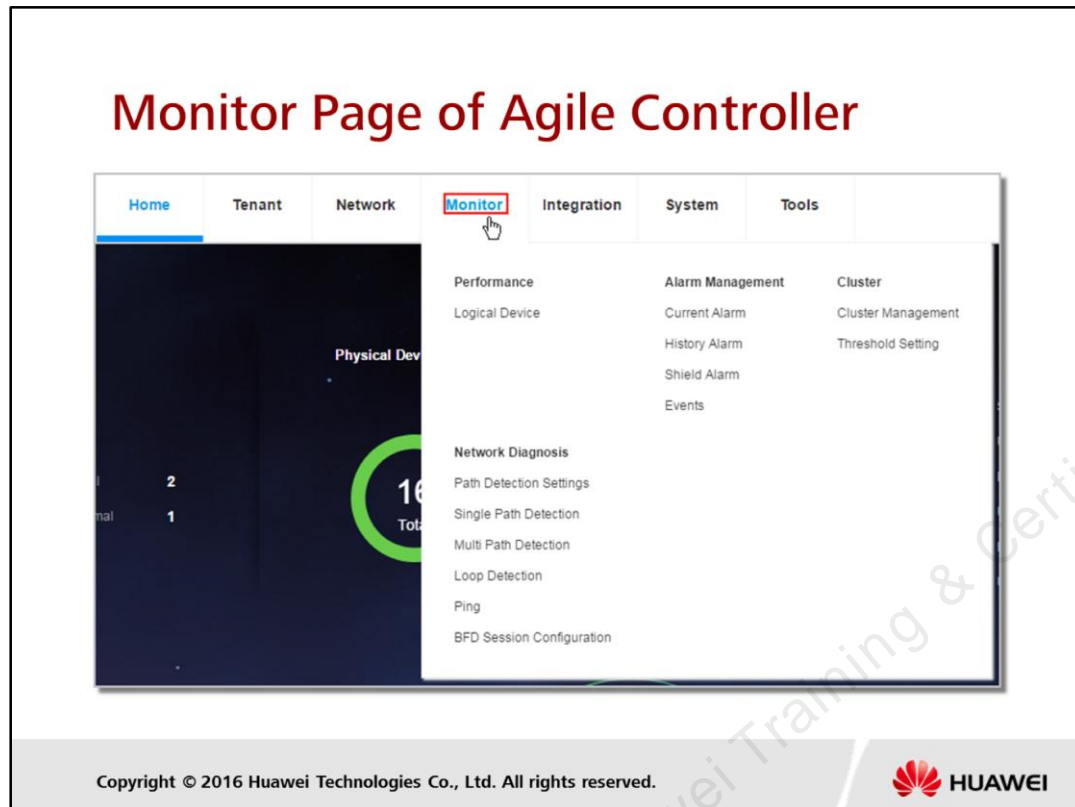


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

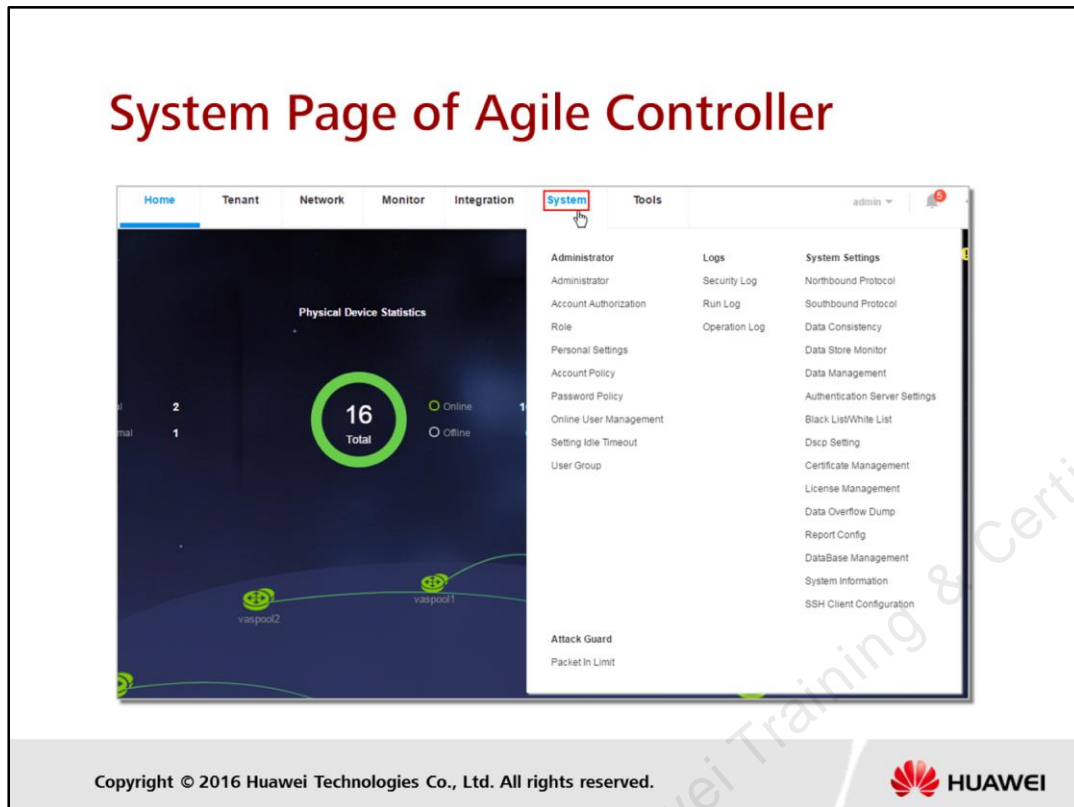




- Physical Resource:
  - You can manage physical resources in a Fabric network. Network devices and third-party devices are automatically discovered or manually imported to the Agile Controller-DCN and then network devices, servers, and third-party devices are added to resource pools.
- Link Management:
  - You can manage device links. Links are automatically discovered to the Agile Controller-DCN through automatic link discovery or you can add links to the Agile Controller-DCN by manually creating links. You can set device balancing by managing device links.
- Network Management:
  - You can create and configure resource pools to manage Fabric network resources. You can add fabric network devices to resource pools, synchronize networks, set device roles, and allocate network resources to complete interconnection between the Agile Controller-DCN and peripheral systems.
- TOPO:
  - Network topologies are displayed, including physical topologies, logical topologies, and application topologies.
- Resource Pool Configuration:
  - You can configure network resources. When planning a network, you need to preconfigure not only fixed resources such as egress public IP and VPC communication IP but also public resources shared by tenants such as bridge domains (BDs), global VNI, global VLAN, interconnection VLAN, and interconnection IP. The Agile Controller-DCN can allocate the resources to tenants.



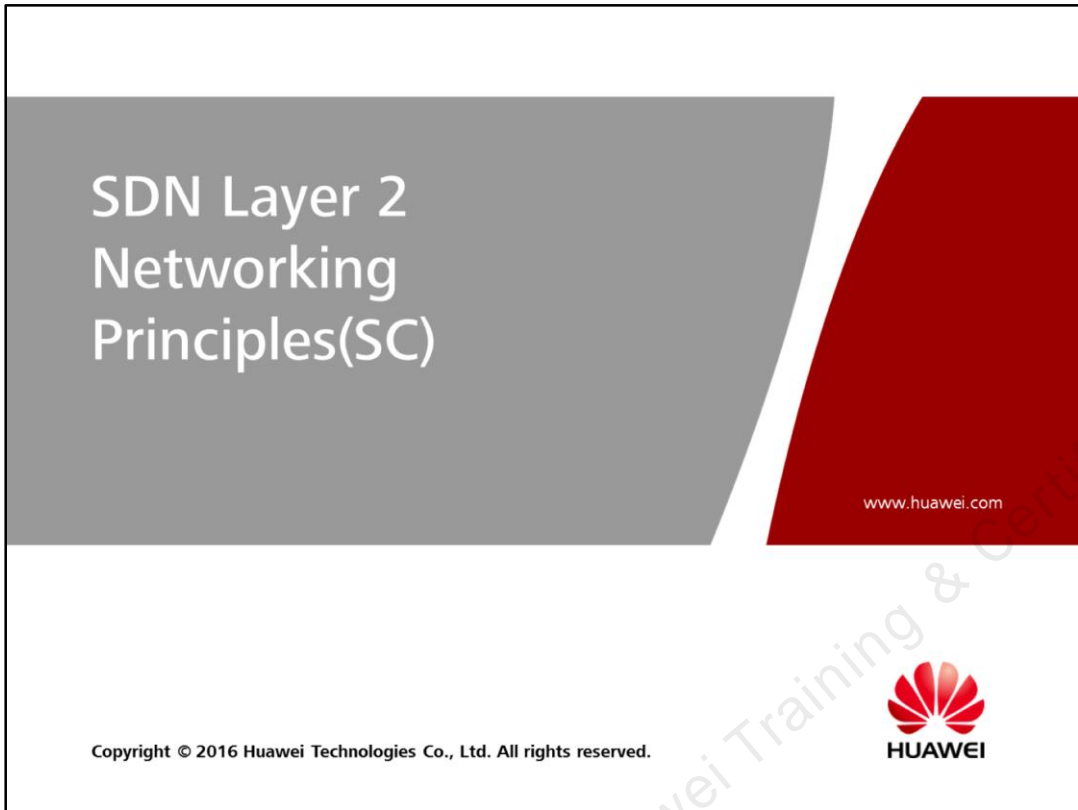
- Performance Monitor:
- Physical devices can be monitored. Status of fabric network devices and traffic statistics on the entire network are displayed.
- Alarm Management:
- The Agile Controller-DCN information is displayed. Alarm information can be monitored in real time. You can clear alarms according to the alarm information and handling suggestions to ensure proper service running.
- Cluster:
- You can set cluster thresholds, including CPU, memory, and hard disk thresholds on the cluster host.
- Diagnosis:
- Logical resources are monitored. You can view the port, subnet, network, and route information.



- Administrator:
- An administrator manages the account list, authorization list, role list, password, account, and users.
- Log:
- The Agile Controller-DCN records logs including the administrator operation logs on the Agile Controller-DCN and logs generated during Agile Controller-DCN running for auditing and fault location.
- System Settings:
- You can configure the authentication server, email server, data dumping, and logs and alarms reporting, and set system whitelist.
- You can apply for a license or perform routine inspections on the license to ensure that the licence meets service requirements.
- Attack Guard:
- Each OVS device can only send the packet-in packets with the maximum rate according to packet rate limit. The OVS devices discard the excessive packets to save the Agile Controller-DCN resources and improve performance.

## Summary


- Huawei SDN Routers
- Huawei SDN CloudEngine Switches
- Huawei SDN Controller



SDN Layer 2  
Networking  
Principles(SC)

[www.huawei.com](http://www.huawei.com)

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.



HUAWEI

Huawei Training & Certification



## Introduction

- In SDN Network Architecture, SDN controller is the core component and acts as controlling element , where as , networking devices act as forwarding elements.

Communication between SDN controller and the forwarders are very important aspect in term of operation, maintenance, and troubleshooting. If communication fault happen between controller and forwarders, it will cause whole network down.



## Objectives

- Upon completion of this course, you will able to:
  - Understand the principles of layer 2 networking method between controller and forwarders.
  - Understand the principles of Ethernet networking





## Contents

1. Control Channel Overview
2. Ethernet Networking Principles



## Contents

1. **Control Channel Overview**
2. Ethernet Networking Principles



## Contents

1. Control Channel Overview
  1. Control Channel Principles Overview
  2. Control Channel Process Establishment Overview
  3. SDN Layer 2 Networking Principles
  4. Control Channel Function



## Contents

1. Control Channel Overview
  - 1. Control Channel Principles Overview**
  2. Control Channel Process Establishment Overview
  3. SDN Layer 2 Networking Principles
  4. Control Channel Function

## Control Channel Principles Overview

- Channel between SDN controller and Forwarders called Control Channel.
- Basically, there are two types of control channel that can be built between Controller and Forwarders
  - Outband Networking
    - Independent network between Controller and Forwarders. Control traffic and Service traffic isolated each other physically.
  - Inband Networking
    - Control Channel network and Service network are shared.

## Comparison Between Outband Networking and Inband Networking

Type	Inband Networking	Outband Networking
Cost	Cost lower, control channel shared with service channel	Cost higher due to construction of another network in which only consists of Control traffic.
QoS	Provide and set high priority for control traffic over service traffic	Control traffic has dedicated bandwidth, service traffic does not affect the control traffic
Fault Effect	Depend on networking convergences. Due to higher priority of control traffic over service traffic, control traffic will preempt low priority service traffic during network convergence	Depend on networking convergence. During network convergence, control traffic will not bring any effect to the service traffic, and vice versa.



## Contents

1. Control Channel Overview
  1. Control Channel Principles Overview
  - 2. Control Channel Process Establishment Overview**
  3. SDN Layer 2 Networking Principles
  4. Control Channel Function

## Control Channel Process Establishment Overview

- Either outband networking or inband networking, the key point is reachability between controller and forwarders. Traditional Networking can be deployed: Layer 2 Networking or Layer 3 Networking.
- Based on different type of networking, protocols used also different.
  - Layer 2 Networking : VLAN, MSTP
  - Layer 3 Networking : RIP, OSPF, ISIS, BGP.





## Contents

1. Control Channel Overview
  1. Control Channel Principles Overview
  2. Control Channel Process Establishment Overview
  - 3. SDN Layer 2 Networking Principles**
  4. Control Channel Function

## SDN Layer 2 Networking Principles

- Normally, SDN controller such as SNC, and forwarders such as NE40E product series are located in the same network segment.
- Forwarding based on layer 2 methods, for example, MAC address-based forwarding.
- In order to isolate control traffic and service traffic, VLAN is used.
- In layer 2 networking, when number of forwarders increased, layer 2 loop prevention technologies is needed, for example, MSTP



## Contents

1. Control Channel Overview
  1. Control Channel Principles Overview
  2. Control Channel Process Establishment Overview
  3. SDN Layer 2 Networking Principles
  - 4. Control Channel Function**

## Control Channel Functions

- Once the control channel has been established, SDN controller will start :-
  - Collect network topology and link state information.
    - Device Information
    - Label information
    - Interface information
    - Topology Information

- Once Control channel has been established between SDN controller and Forwarders, SDN controller start to collect network element information, network topology and link state information. SDN controller collect vendor device information, device type, device version and device ID information. The purpose of those information is to allow SDN controller able to manage multi-vendor networking devices. Different vendor might have different type of interface protocol between SDN controller and forwarders. In order to solve this issues, SDN controller need to be installed different type of driver programs that supported by different type of devices. In order to do that, SDN controller need to collect device information such as vendor information, device type, device version, and device ID.
- SDN controller also need to collect label information. Recommended SDN network deploy using MPLS switching, because MPLS networking is mature networking, mostly of vendor devices support MPLS features. Compare other tunneling protocols such as IPSec, GRE, MPLS uses only 4 bytes of overhead, and able to implement traffic engineering.
- Interface information such as interface name, interface ID number, interface media types, interface bandwidth , etc. are also gathered by SDN controller through control channel. Those information help controller to understand how many interconnection of devices and able to use those information to build the network topology.
- Once network resources has been collected, controller also need network topology information. Network topology information is described network nodes, network interconnection, and interconnection between network nodes. There are few protocols able to help SDN controller to collect network topology information, for examples, Link Layer Discovery Protocols (LLDP), suitable for Data Center network. Layer 3 Protocols also able to help to discover network topology such as ISIS/OSPF. Those traditional routing protocol at first every networking devices advertise their own link state information and then build LSDB. However, ISIS or OSPF still not able to solve cross-AS domain topology. So far, IETF has one protocol that able to solve cross-AS domain topology, BGP-LS. BGP-LS can collect cross-AS domain link state information to help SDN controller to have global view of the network topology.

## Control Channel Function (*Cont*)

- Once information has been collected by SDN controller, SDN controller will start to do flow calculations, and then updates the result into flow table on each forwarders.
- The algorithm is the same as traditional IP routing algorithm.
- To updates flow table on each forwarders can be done by deploying Southbound interface, such as OpenFlow, PCE protocol, BGP protocol, NetConf protocol.
- Forwarders forwarding traffic based on this flow table.



## Contents

1. Control Channel Overview
- 2. Ethernet Networking Principles**



## Contents

2. Ethernet Networking Principles
  1. Ethernet Technology Overview
  2. The Basic Principles of Ethernet
  3. Layer 2 Switching
  4. Ethernet Port Technologies
  5. VLANs and Layer 3 Switching



## Contents

2. Ethernet Networking Principles
  - 1. Ethernet Technology Overview**
  2. The Basic Principles of Ethernet
  3. Layer 2 Switching
  4. Ethernet Port Technologies
  5. VLANs and Layer 3 Switching



## Ethernet Technology Overview

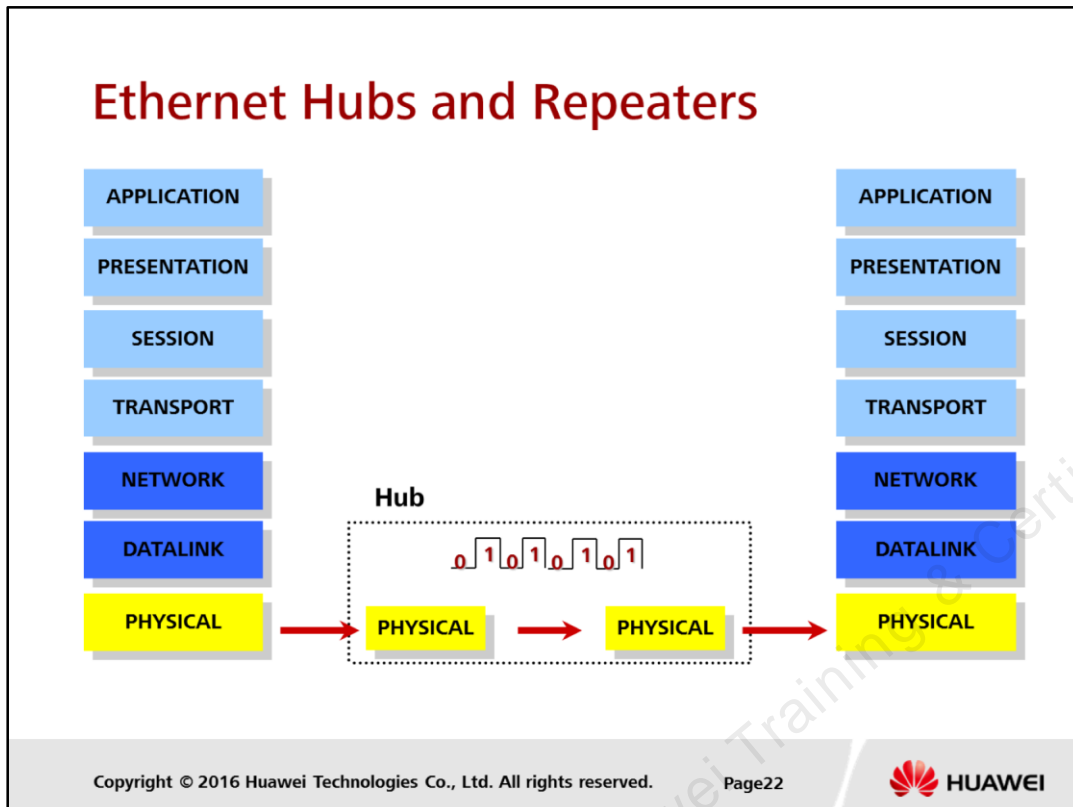
- Most popular LAN technologies used until now.
- Developed by Bob Metcalfe and colleagues at Xerox's PARC
- IEEE 802.3 Standardized
- Numerous of standard has been released: 10Base-5, 100Base-T, 10 Gigabit Ethernet, 100G Ethernet.

- Ethernet had its 30<sup>th</sup> birthday in 2003 and has seen many changes since its inception. Ethernet has constantly been reinvented as computing technology has developed over the years. The first part of this course will look at the origins of Ethernet and the developments it has undergone over a 30 year period.
- The birthplace of Ethernet is Xerox's Palo Alto Research Centre. PARC is famous for, amongst other things, its invention of the Xerox Alto, the first personal computer with graphical user interfaces and mouse pointing devices. PARC's inventions also included the first laser printers for personal computers. The developments in computing at PARC has made remarkable changes to the computing environment. In the 1970's computers were large expensive mainframes and the idea of a personal computer was revolutionary.
- Ethernet was developed over a number of years, the first sign of this new technology appeared on May 22<sup>nd</sup> 1973, Bob Metcalfe, who is now widely recognised as being the father of Ethernet, wrote a memo describing the network system he had invented for interconnecting the Alto workstations and high speed laser printers. His memo was based on an earlier experiment in computer networking called the Aloha network. This was a network which was designed and built by Norman Abramson and colleagues at the university of Hawaii from 1966 to 1970. Aloha was a radio network for communicating between the Hawaiian islands. The idea on which this experimental network was based was the use of a common channel for communicating, which in this case was a common radio channel. The access protocol which allowed computers to transmit data on this channel is now called "Pure Aloha" and was the original "Carrier Sense Multiple Access" protocol.
- Metcalfe's original network was called the Alto Aloha Network, i.e. a network for interconnecting Alto computers using the Aloha protocol. Later in 1973 the name was changed to Ethernet. The term "ether" was an idea in physics that was disproved by Michelson and Morley in 1887. The ether was thought to carry electromagnetic signals through space. Metcalfe thought that this would be a good name for his technology.
- When building the first network, the Ethernet interface was driven by the clock on the Alto workstations this gave the transmission speed on 75 Ohm coaxial cable of 2.94 Mb/s.



## Contents

2. Ethernet Networking Principles
  1. Ethernet Technology Overview
  - 2. The Basic Principles of Ethernet**
  3. Layer 2 Switching
  4. Ethernet Port Technologies
  5. VLANs and Layer 3 Switching




- To interconnect all workstations and servers on a small network we may use a single hub. This will create a star shaped network with the hub at the centre (or hub!) of the network.
- Networks that involve repeaters or hubs are often called shared networks in this case “shared Ethernet” because all devices are sharing the same media.
- A hub’s basic function is to repeat any datagrams received on a port out of every other port. Any devices attached to these “hub” ports will receive all datagrams and check the address to see if it is destined for them.

## The working principles of a Hub

- A hub works in half duplex mode
- If one device speaks all other devices listen.

The diagram shows a network hub with five ports. A red arrow labeled 'IN' points to the first port, indicating data input. Four blue arrows originate from the first port and point to the second, third, fourth, and fifth ports, illustrating that data received at one port is broadcasted to all other ports. The hub is represented as a single rectangular box containing five RJ45 port symbols.

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page23  HUAWEI

- The working principle of a hub is very simple: The data received from any of the ports will be forwarded to all the other ports, except the port received the data.
- A hub cannot receive and send data at the same time, this mode of working is called Half-duplex. When all devices are contending for use of the same media this is known as a collision domain, devices will listen until the network is quiet then transmit.

## Issues with Hub-based Networks

- Hub-based networks may have drawbacks:
- The media is still shared between all devices – shared Ethernet
- Only one device can talk at a time
- Severe collisions
- No security

- Shared Ethernet networks have a number of drawbacks:
  - Collisions: if there is more than one device transmitting data at one time ,then a collision occurs.
  - No security: All data is transmitted across the whole network it is up to the end station whether or not it discards the packet. There are many available software packages which will turn the NIC into promiscuous mode to enable the “sniffing” of all network data.

## Types of Ethernet Addresses

- Unicast
- Broadcast
- Multicast

- The destination address in the Ethernet header can be one of three forms:

- Unicast address

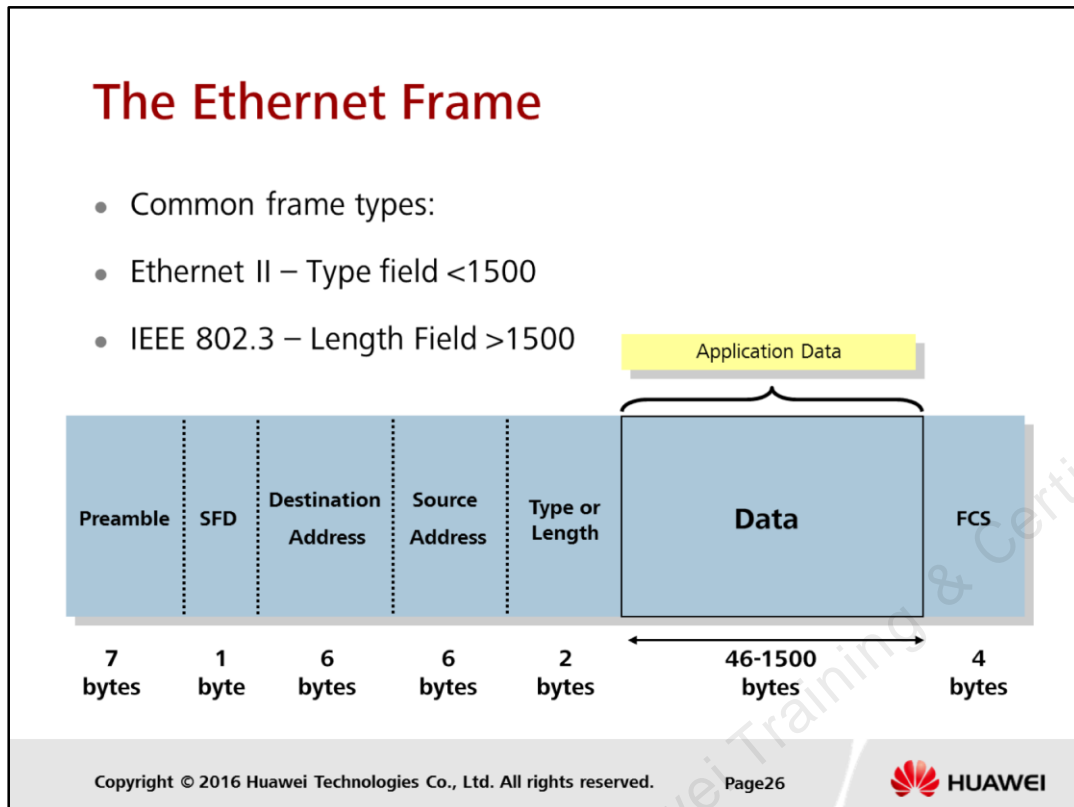
Only the specified host will process the frame e.g. a frame sent from PC to Server

- Broadcast address

This frame will be processed by every host (PC, Printer, server & Mainframe)

- Multicast address

All hosts that are members of the specified multicast group will process this frame e.g. a router to other routers



### • Ethernet Frame Fields

- **Preamble:** 7 bytes of 10101010 to allow timing synchronisation between sender and receiver
- **SFD:** Start Frame Delimiter – 10101011 to tell the receiver the next byte is the start of the frame
- **DMAC:** destination MAC address, 6bytes
- **SMAC:** source MAC address, 6bytes
- **Length/Type:** 2bytes, it has the following meanings:
  - if Length/Type > 1500 ,this field indicates the type of the upper protocol of an Ethernet\_II frame
  - if Length/Type < 1500, the field indicates the length of the frame, and it is a 802.3 frame.
- **DATA:** 0~1500bytes , If the data field is less than 64 bytes padding is used to fill up the frame to make sure that the length of frame must be at least 64 bytes.
- **FCS:** 4bytes
- The whole length of the fame is 64~1518 bytes.

## Ethernet MAC Address

- Globally Unique 48-bit address

7	0	7	0	7	0	7	0	7	0	7	0
00000001	00000000	01011110	XXXXXXXX	XXXXXXXX	XXXXXXXX						

- Represented by a 12 digit hexadecimal number

01	00	5E	XX	XX	XX
----	----	----	----	----	----

Organizationally Unique Identifier (OUI)

Assigned by each organization

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page27

- MAC: media access control
- The MAC address is the physical address of each device connected to the network. MAC address assignments are administrated by IEEE and are be globally unique. Each address is composed of two parts, the OUI which is the Organisation Unique Identifier which is the address space assigned to the provider, the remainder of the address space is assigned by each organization.
- The first 24 bits denotes the provider code and the last 24 bit is assigned at the provider's discretion.



## MAC Broadcast Address

- Broadcasts use a MAC address of FF-FF-FF-FF-FF-FF.
- If the eighth bit is set to 1, it represents a multicast address

7	0	7	0	7	0	7	0	7	0
00000001	00000000	01011110	0XXXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXXX	XXXXXXXXX	0

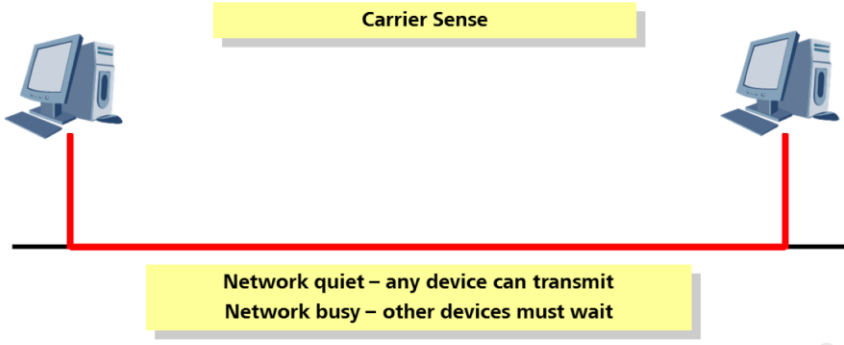
Broadcast / Multicast Bit (Bit 0)

Locally Administered address Bit (Bit 1)

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page28


- All 1's in the address otherwise written as FF-FF-FF-FF-FF-FF in hexadecimal denotes a broadcast address, which will be read by all attached stations.
- If bit 0 is set to 1, it represents a multicast address
- If bit 1 is set to 1 it denotes that this is a locally set MAC address that does not follow the international standard scheme, this however is almost never used

## Carrier Sense Multiple Access

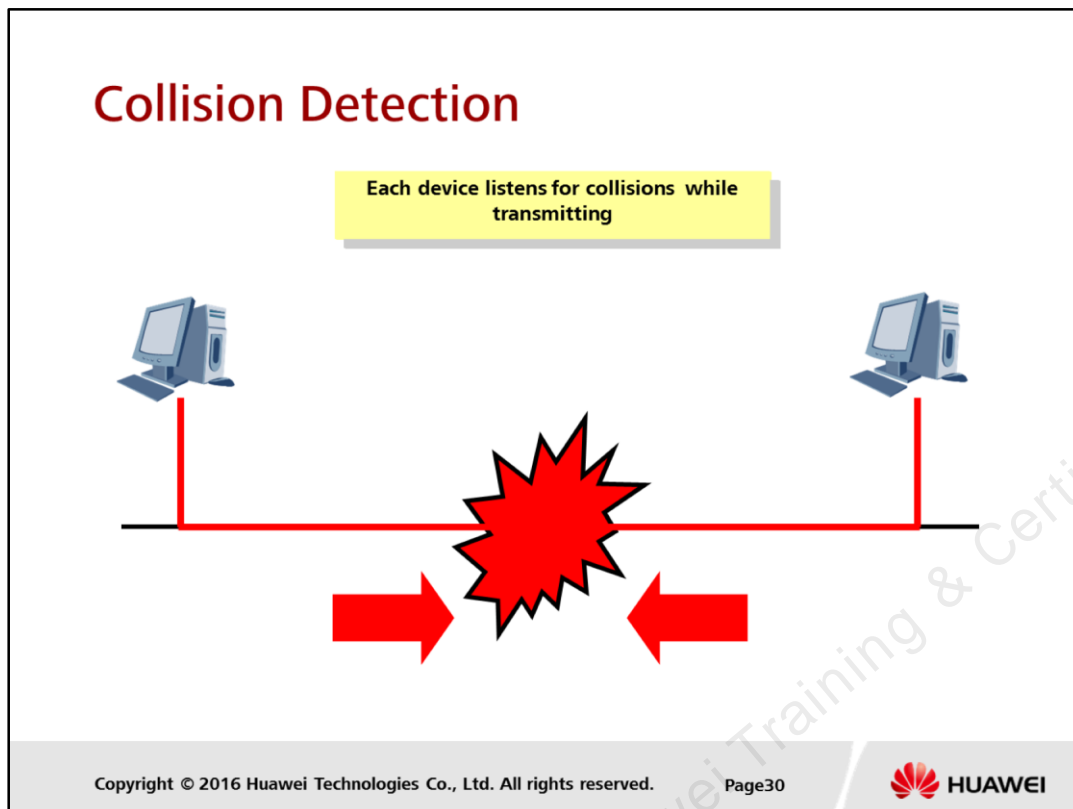


- Multiple Access – multiple devices connected to the same media

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page29



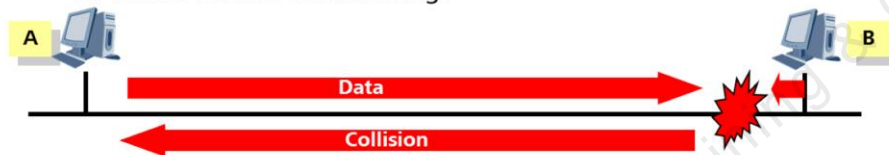
- CSMA/CD is the media access protocol that Ethernet uses.
- CSMA/CD stands for **C**arrier **S**ense **M**ultiple **A**ccess with **C**ollision **D**etection.
- Taking each part of the protocol in turn, “multiple access” simply means that multiple devices have access to the same network. The “carrier sense” part, well, Ethernet has no actual carrier signal the term is a hangover from the Aloha protocol, the carrier in Ethernet terms simply means the presence of traffic on the network. An Ethernet NIC senses whether the media has any signals on it, if it has another device on the network is transmitting, every device must wait for a period of quiet before transmitting.



- The CSMA part of the protocol was based on the Aloha protocol, however to make it more efficient collision detection was added. If two devices both sense that there is no one else transmitting they may both simultaneously decide to send data, when this occurs there is a collision and the data transmitted by both parties will be corrupted.
- If a collision is detected by the senders they will both know that they need to retransmit their data
- This random time is selected according to the following rule:
  - If there are  $n$  collisions ( $n < 16$ ) of a packet, the node will select a number  $K$  randomly with equal possibility from  $0, 1, \dots, 2^{n-1}$  and wait for  $K * 512$  bit time (for example: in 10Mbps Ethernet, 1 bit time =  $10^{-7}$  seconds).
  - If  $n > 15$ , the node will give up the transmission.

## Collision Issues

- Collision detection has an issue
  - All devices need to be able to detect a collision while they are transmitting – otherwise they will not know who caused the collision.
  - So data sent by “A” must be capable of getting to the far end of the network and the collision has to have time to return to sender all while A is still transmitting.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

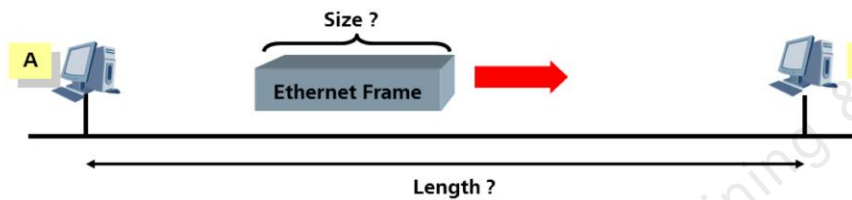
Page31



- To be sure that collisions can be detected and that devices know that they are part of a collision the collision must occur while they are still transmitting. Therefore data must be capable of “filling” the network. This means that the data must have time to reach from one end of the network to the other and a collision returning to the originating station while that station is still transmitting.

## Collision Decisions

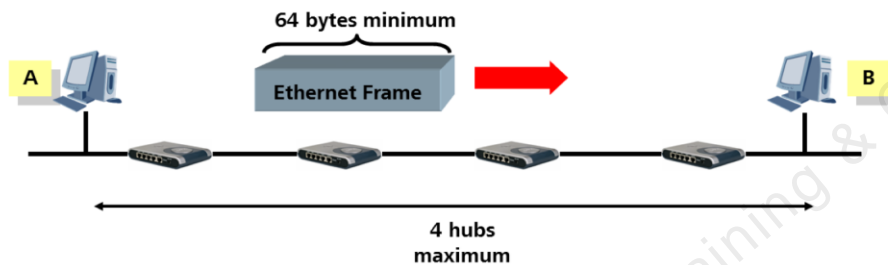
- There are 3 factors to consider:
- Network size – how far does the data have to travel?
- Network speed – how fast does the data travel?
- Minimum packet size - How long will each device transmit for?



- The three factors which make up the decision on what collision domain parameters were selected by Ethernet designers are as follows
  - The size of the network, which determines how far the traffic will have to travel
  - The speed of transmission over the network
  - The smallest size of packet that will be transmitted

## Collision Solutions

- The trade-off:
  - Network size – at 10 Mbps up to 4 hubs
  - Network size – at 100 Mbps up to 2 hubs
  - Minimum packet size – 64 bytes minimum



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

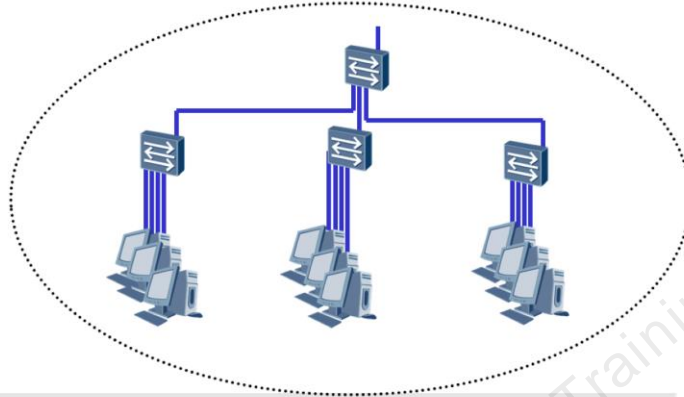
Page33



- The parameters in use for Ethernet networks were based on a minimum packet size of 64 bytes this leads to the trade off that we may have a network size with 10Mb Ethernet that allows us to span a network of 4 hubs (which will introduce a little delay) and 100m between each one. For 100Mb Ethernet this shrinks the network down to 2 hubs.
- With 1G Ethernet the network can only support one hub, and we have to increase the minimum packet size, however very few vendors supplied Gigabit Ethernet hubs and have mainly supplied switched Ethernets products.

## Collision Domains

- All devices contend for the same media and only one device can transmit at a time
- This is called a collision domain.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page34



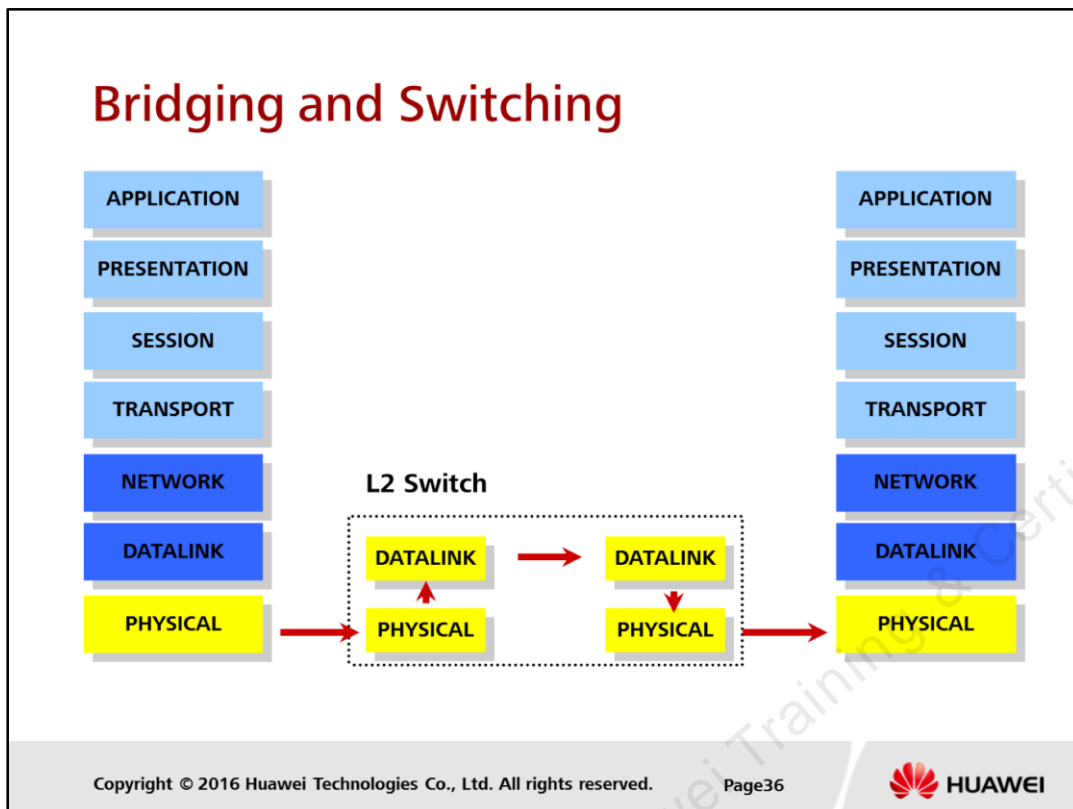
- A collision domain is a logical network segment where data packets can "collide" with one another for being sent on a shared medium.
- As each device that connects to the network has to wait for a period of quiet before transmitting the more devices that connect to the network the more collisions there are likely to be. The more collisions that occur in the domain the less efficient it will be.
- As we add more workstations the lower the efficiency of the network becomes.



## Contents

1. Ethernet Networking Principles
  1. Ethernet Technology Overview
  2. The Basic Principles of Ethernet
  - 3. Layer 2 Switching**
  4. Ethernet Port Technologies
  5. VLANs and Layer 3 Switching

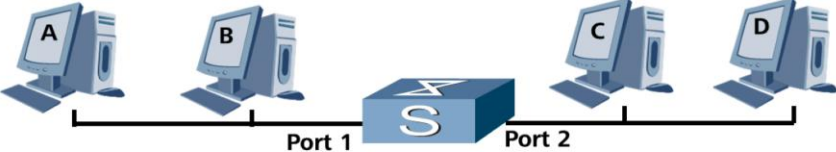




- An Ethernet switch or bridge has the following functions:
  - Source MAC address learning.
  - Forwarding based on the destination address.
  - Filtering based on the destination address
  - Flooding based on the destination address


## MAC Forwarding

MAC Address	Port
MAC A	1
MAC B	1
MAC C	2
MAC D	2



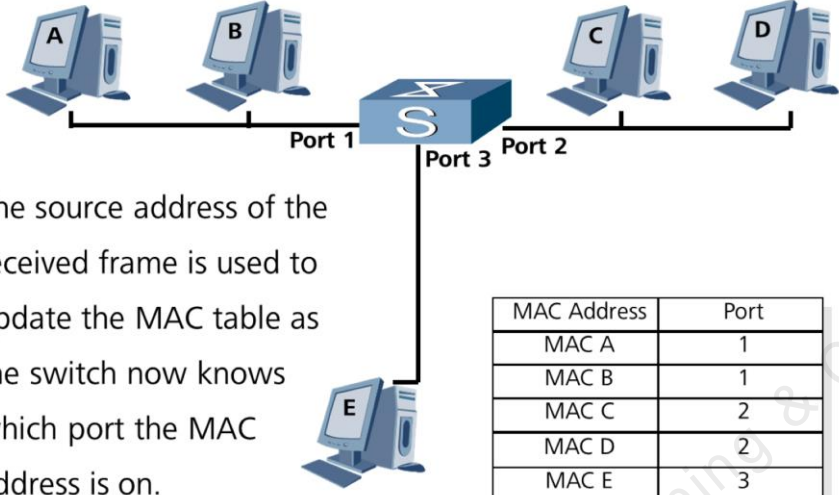
- Layer 2 Ethernet switches perform 3 operations on traffic:
- Forward - send frame out of port towards known destination device
- Filter - if known device is on same port where traffic is received
- Flood – device not in table send frame out of all ports

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page37




- Every Ethernet switch has a MAC address table, which contains the mapping of MAC address to port number. This enables the switch to forward Ethernet frames to the correct device.
- Other names for MAC Address tables:
  - FDB – Forwarding Database
  - SDB – Switch Database
  - CAM – Content Addressable Memory
- Once the table is populated it contains the mappings of MAC address to port number so the switch knows where each device is connected.
- It can then **forward** traffic out of the port to which a known destination device is connected. i.e. one that is in the table.
- If traffic is received on the same port to which the destination device is connected the it will **filter** the traffic.
- If when receiving an Ethernet frame and there is no entry in the MAC table the frame will be **flooded** out of all ports except the incoming port
- The next question we have is how is the address table populated?
-

## Source MAC Learning

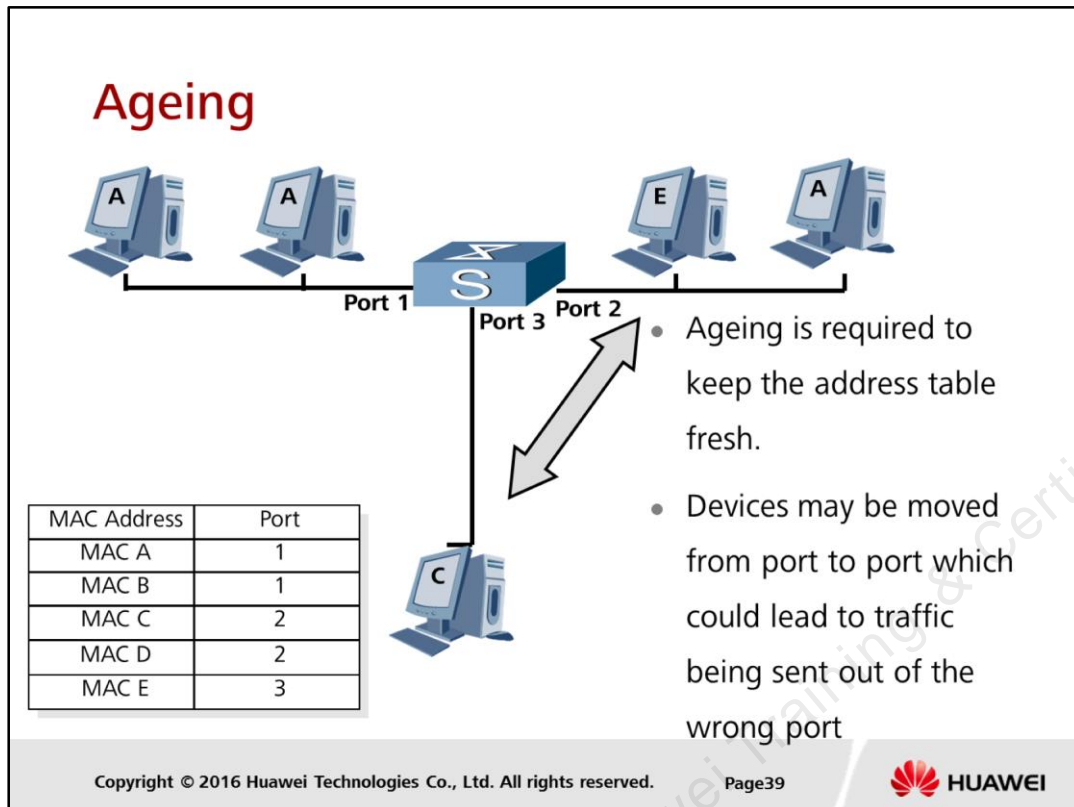


- The source address of the received frame is used to update the MAC table as the switch now knows which port the MAC address is on.

MAC Address	Port
MAC A	1
MAC B	1
MAC C	2
MAC D	2
MAC E	3

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page38 

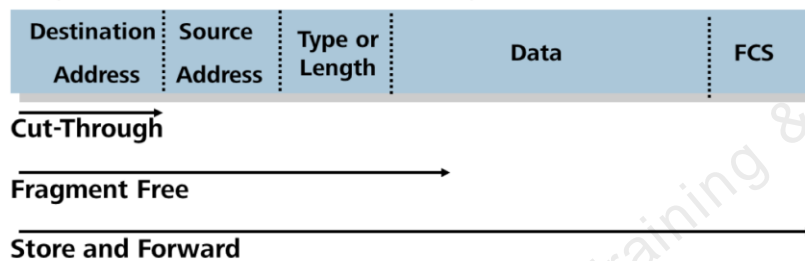
- Source address learning is the process of obtaining the MAC address of devices. When a bridge is first turned on, it has no entries in its bridge table. As traffic passes through the bridge, the sender's MAC address is stored in a table along with the associated port on which the traffic was received.
- As we see in the diagram - If PC A sends a frame to PC D, the Switch receives the frame on port 1, first, it looks at the destination MAC address, and checks the MAC address table, if the table has no entry which matches the destination MAC address the LAN Switch sends the frame to all the other ports, it will then write the source MAC address of the received frame into the table. This establishes the mapping relationship between the port 1 and the MAC address of PC A. Using this method, each switch will establish the mapping relationship and the MAC address table can be populated.
- When a bridge does not have an entry in its bridge table for a specific address, it will transparently forward the traffic through all its ports except the source port. This is known as flooding. The source port is not "flooded" because the original traffic came in on this port and already exists on that segment. Flooding allows the bridge to learn, as well as stay transparent to the rest of the network, because no traffic is lost while the bridge is learning. After the bridge learns the MAC address and associate port of the devices to which it is connected, the benefits of transparent bridging can be seen by way of filtering. Filtering occurs when the source and destination are on the same side (same bridge port) of the bridge.
- Forwarding is simply passing traffic from a known device located on one bridge port to another known device located on a different bridge port.



- In addition to the MAC address and the associated port, a bridge also records the time that the device was learned. Ageing of learned MAC addresses allows the bridge to adapt to the movement, addition, and change of devices in the network. After a device has been learned by the switch, the bridge starts an aging timer. Each time the bridge forwards or filters a frame from a device, it restarts that device's timer. If the bridge doesn't hear from a device in a preset period of time, the aging timer expires and the bridge removes the device from its table.
- The IEEE 802.1D default is 300 seconds (five minutes).

## Three Switching Modes

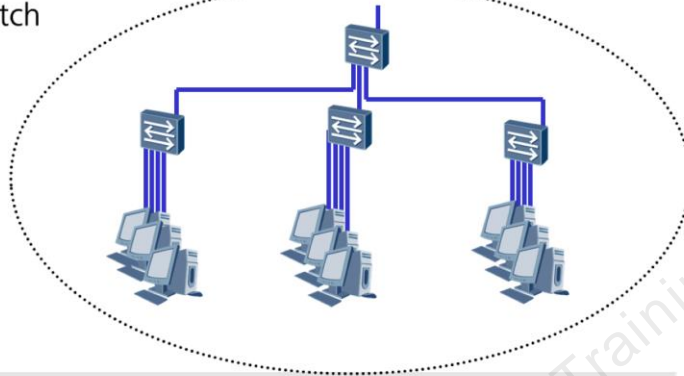
- There are 3 forwarding models which can be used to improve switching performance.
  - Cut-through – The fastest
  - Store & forward – The safest – The whole frame
  - Fragment free - Fast and ensures fragments are not forwarded



- Cut-Through Mode
  - Switches operating in cut-through mode receive and examine only the first 6 bytes of a frame. These first 6 bytes cover the destination MAC address, which has sufficient information to make a forwarding decision. Although cut-through switching offers the least latency when transmitting frames, it may transmit fragments created during Ethernet collisions, corrupted frames.
- Fragment-Free Mode
  - Switches operating in fragment-free mode receive and examine the first 64 bytes of frame. Why examine 64 bytes? In a properly designed Ethernet network, collision fragments must be detected in the first 64 bytes.
- Store-and-Forward Mode
  - Switches operating in store-and-forward mode receive and examine the entire frame, resulting in the most error-free type of switching, however this is also the slowest type of switching.

## Broadcast Domain

- Introducing a Layer 2 switch will improve network performance by containing unicast traffic within each collision domain
- Broadcast traffic however will still be forwarded across the switch



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page41



- Layer 2 switches separate the network into collision domains, so if a hub is attached to each port there will be a collision domain on each port.
- But the switch and the LANs connected to the switch form a Broadcast Domain. Any broadcast packets will be flooded across the whole domain, and all connected devices will receive the broadcast packet. If there are many broadcast packets in a domain, it may occupy a large part of the available network bandwidth and because these broadcasts may go to areas of the network where they are not needed this will reduce the efficiency of the network.

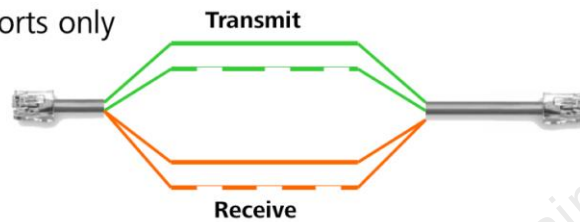


## Contents

2. Ethernet Networking Principles
  1. Ethernet Technology Overview
  2. The Basic Principles of Ethernet
  3. Layer 2 Switching
  - 4. Ethernet Port Technologies**
  5. VLANs and Layer 3 Switching

## Full Duplex

- Now we have reached the point where we can have a separate switched port for each user we can use full duplex communications.
- This allows full speed in each direction simultaneously, therefore twice the throughput.
- Switch ports only



- If we connect a single end station to an Ethernet switch port the port will function in full-duplex mode this allows a more efficient mode of working where data can be transmitted and received at the same time.

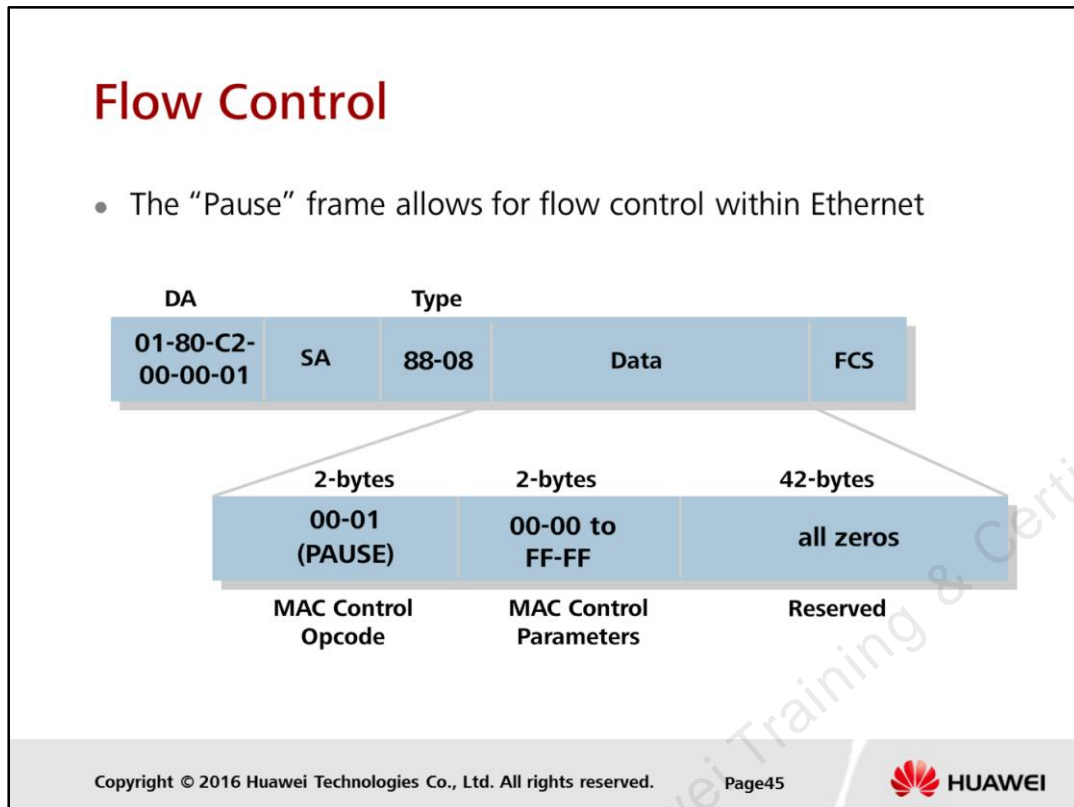


## Auto Negotiation

- Auto-negotiation uses pulses similar to link integrity pulses to advertise capability in the form of a 16 bit message
- Auto-negotiation priority:
  - 1000BASE-T full duplex
  - 1000BASE-T half duplex
  - 100BASE-TX half duplex
  - 10BASE-T full duplex
  - 10BASE-T half duplex
  - 100BASE-TX full duplex



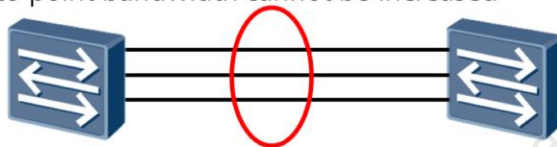
- Auto-negotiation was originally defined in the IEEE 802.3u standard in 1995. It was introduced into the fast Ethernet part of the standard but is also backwards compatible to 10BASE-T. This was later updated in 1999, the negotiation protocol was significantly extended by IEEE 802.3ab, which specified the protocol for Gigabit Ethernet, making auto-negotiation mandatory for Gigabit Ethernet.
- Auto-negotiation is used by devices that are capable of different transmission rates (such as 10Mbit/sec and 100Mbit/sec), and different duplex modes (half duplex and full duplex). Every device will declare its possible modes of operation. The two devices then choose the best possible mode of operation that are shared by the two devices, where higher speed (100Mbit/sec) is preferred over lower speed (10Mbit/sec), and full duplex is preferred over half duplex at the same speed.
- Parallel detection is used when a device that is capable of auto-negotiation is connected to one that is not. This happens if one device does not support auto-negotiation or it is disabled via software. In this condition, the device that is capable of auto-negotiation can determine the speed of the other device. This procedure cannot determine the presence of full duplex, so half duplex is always assumed.
- A duplex mismatch will result if the other device is in full duplex mode, that is, one device is using full duplex while the other one is using half duplex. The typical effect of duplex mismatch is that the connection is working but at a very low speed.
-



- The Ethernet standard includes an optional flow control operation known as "PAUSE" frames. PAUSE frames permit one end station to temporarily stop all traffic from the other end station.
- For example, if we have a full-duplex link that connects two devices called "Device A" and "Device B". If Device A transmits frames at a faster rate than Device B can process because there is no buffer space remaining to receive additional frames. Device B can transmit a PAUSE frame to Device A requesting that Device A stop transmitting frames for a specified period of time.
- The format of a PAUSE frame conforms to the standard Ethernet frame format but includes a unique type field with additional parameters.
- The MAC Control Parameters field contains a 16-bit value that specifies the duration of the PAUSE in units of 512-bit times. Valid values are 00-00 to FF-FF (hex).
- A 42-byte reserved field (transmitted as all zeros) is required to pad the length of the PAUSE frame to the minimum Ethernet frame size.

## Link Aggregation

- 802.3ad Link Aggregation allows more than one link to be used between two switches to increase bandwidth and add resilience. Link selected based on source and/or destination MAC address
  - Increases Aggregate bandwidth
  - All traffic from A to B will always use the same link
  - Point to point bandwidth cannot be increased



- 802.3ad Link aggregation (sometimes known as port trunking) allows more than one full duplex link to be used between two Ethernet switch devices. In a single trunk all links must be of the same speed.
- The use of link aggregation between two switches will increase the total bandwidth between two switches. However a hashing algorithm is used to select which link is used for each conversation to ensure Ethernet frames arrive at their destination in the right order. So traffic between two devices on the network will always select the same link. This means that for point to point communications, for example between a particular PC and a particular server, the maximum bandwidth available will be limited to that of a single link in the trunk.

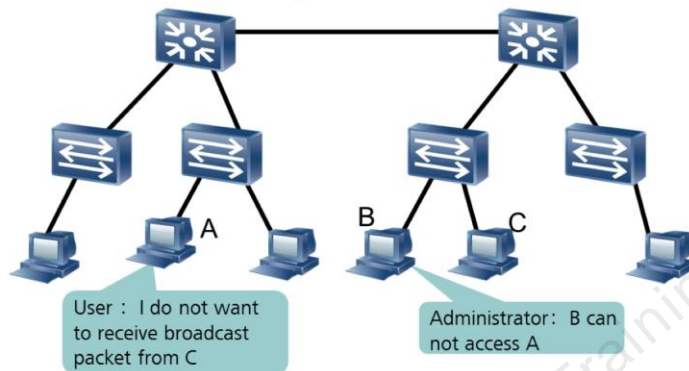


## Contents

2. Ethernet Networking Principles
  1. Ethernet Technology Overview
  2. The Basic Principles of Ethernet
  3. Layer 2 Switching
  4. Ethernet Port Technologies
  5. **VLANs and Layer 3 Switching**

## Layer 2 Switch Limitations

- Layer 2 switches create collision domains
- Broadcast traffic however will still be flooded across the switch
- Traffic control and security issues

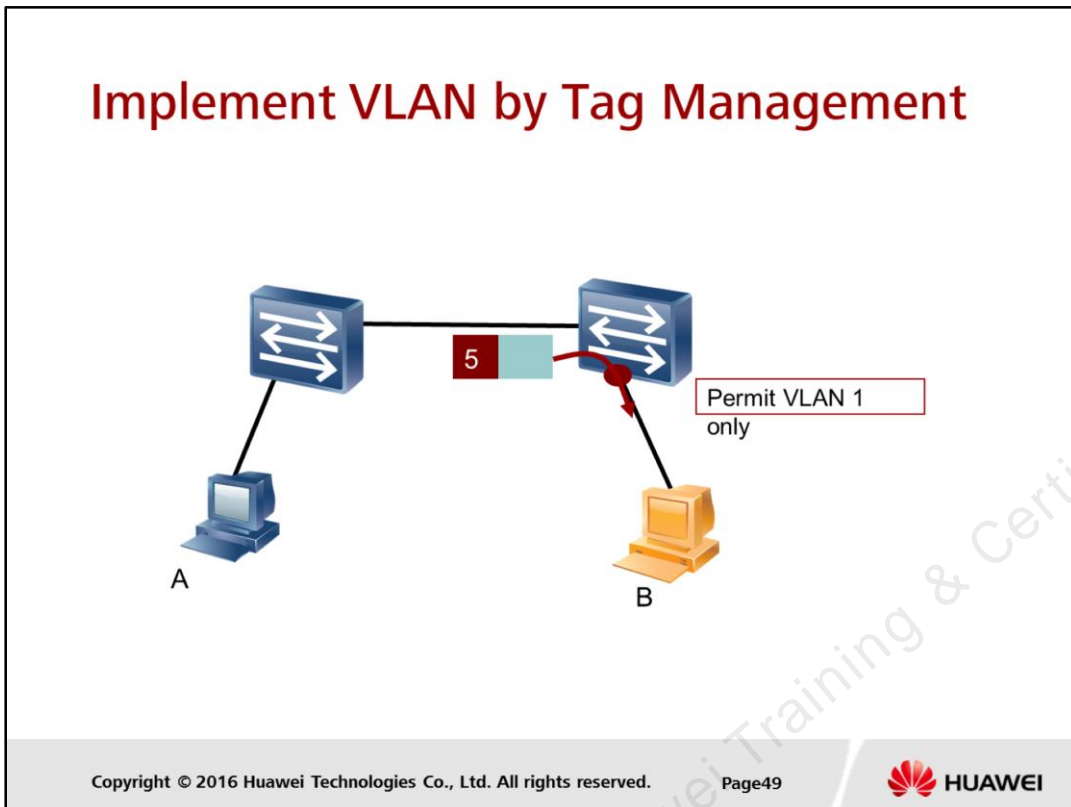


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

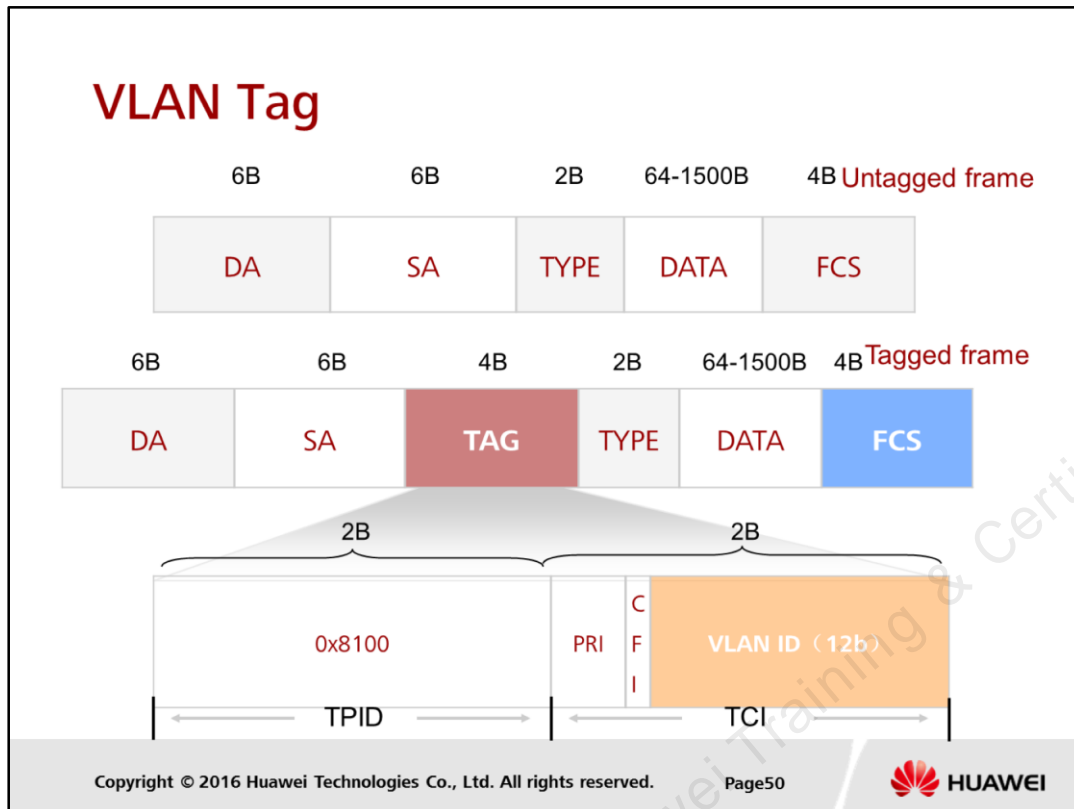
Page48



- Due to all broadcast traffic being flooded across a broadcast domain this gives us issues over both traffic control and security



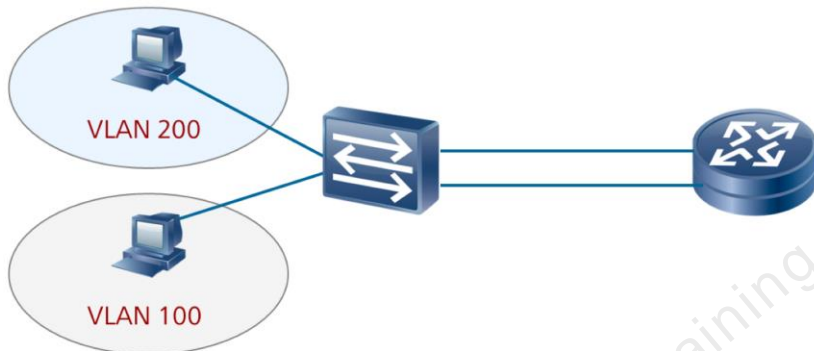
- In order to control the forwarding, the switch will add VLAN tag to Ethernet frame before forwarding it, then decide how to deal with the tag and frame, including discarding frame, forwarding frame, adding tag and removing tag.
- Before forwarding the frame, the switch will check the VLAN tag of the packet, whether the tag is allowed to pass the port, so as to decide whether the frame can be forwarded from the port. In the figure above, if the switch adds tag 5 to all the frames sent from A, and then look up the layer-2 forwarding table, and according to destination MAC address forward them to the port connected to B. But this port is configured to only allowed VLAN 1 to pass, so the frames sent by A will be discarded.
- Hence, switch supporting VLAN will forward Ethernet frame not only according to destination MAC address but also VLAN configuration of the port, so as to implement layer-2 forwarding control.



- 4-byte VLAN tag is added to Ethernet frame header directly. Document IEEE802.1Q describes VLAN tag.
- TPID: Tag Protocol Identifier, 2 bytes, fixed value, 0x8100, new type defined by IEEE, it indicates that it is a frame with 802.1Q tag.
- TCI: Tag Control Information, 2 bytes.
- • Priority: 3 bits, the priority of Ethernet frame. It has 8 priorities, 0 – 7, is used to provide differential forwarding service.
- • CFI: Canonical Format Indicator, 1 bit. Used to indicate bit order of address information in token ring or source route FDDI media access, namely, whether the low bit is transmitted before high bit.
- • VLAN Identifier: VLAN ID, 12 bits, from 0 to 4095. Combined with VLAN configuration of port, it can control the forwarding of Ethernet frame.
- Ethernet frame has two formats: the frame without tag is called untagged frame; the frame with tag is called tagged frame.
- This course will only discuss VLAN ID of VLAN tag.

## Inter-VLAN communication

- To communicate between VLANs we will need to use a router to route between them. If we use the method shown below then we will be using up an additional port in each VLAN.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page51

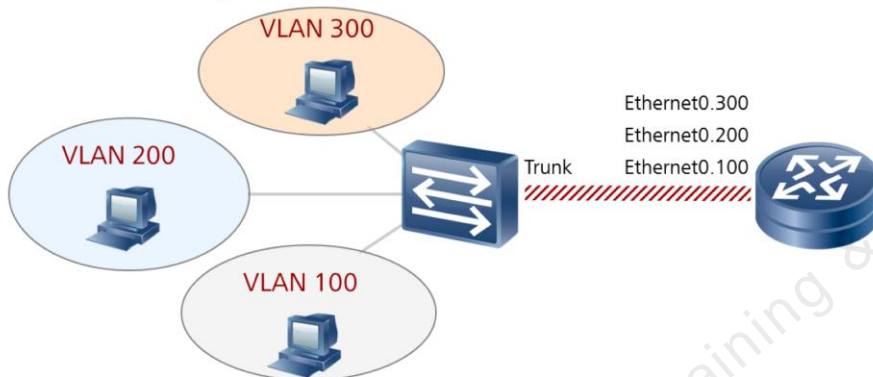


- To communicate between VLANs we need to use a router to process messages at layer 3 as communication between VLANs at layer 2 is not possible. Each VLAN occupies one interface of the router as well as one port in each VLAN on the switch.
- Example :
- If the Engineering department wants to communicate with the Marketing department. It will send the packet to the router, after the router receives the packet ,it will check the destination IP address of the packet and the IP routing table and forward the packet to the correct VLAN.



## Using VLAN Trunking

- Configure the two ports on the link that is between switch and router as VLAN Trunking, then multiple VLANs in the network can share only one physical link.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page52



- To solve the problem of physical interface shortage, another router appears---single-arm router. One Ethernet interface can bear all VLAN gateway by creating sub-interface.
- As shown above, the router only provides one Ethernet interface, and provides three sub-interfaces as default gateways for three VLAN users. When users in
- VLAN100 need to communicate with users in other VLAN, the user only needs to send the data packet to the default gateway, and the default gateway will modify
- VLAN tag of data frame and then send it to the VLAN of destination host. Hence, the communication between VLANs is accomplished.

## Layer 3 Switch

- Functional integration of layer 2 switches and routers forms the layer 3 switch; the layer 3 switch functionally realizes VLAN classification, VLAN internal layer 2 switching and inter-VLAN route functions.

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page53

- The third method is to use layer-3 switch, it is the integrated device of layer-2 switch and router, and it combines the advantages of layer-2 switch and router.



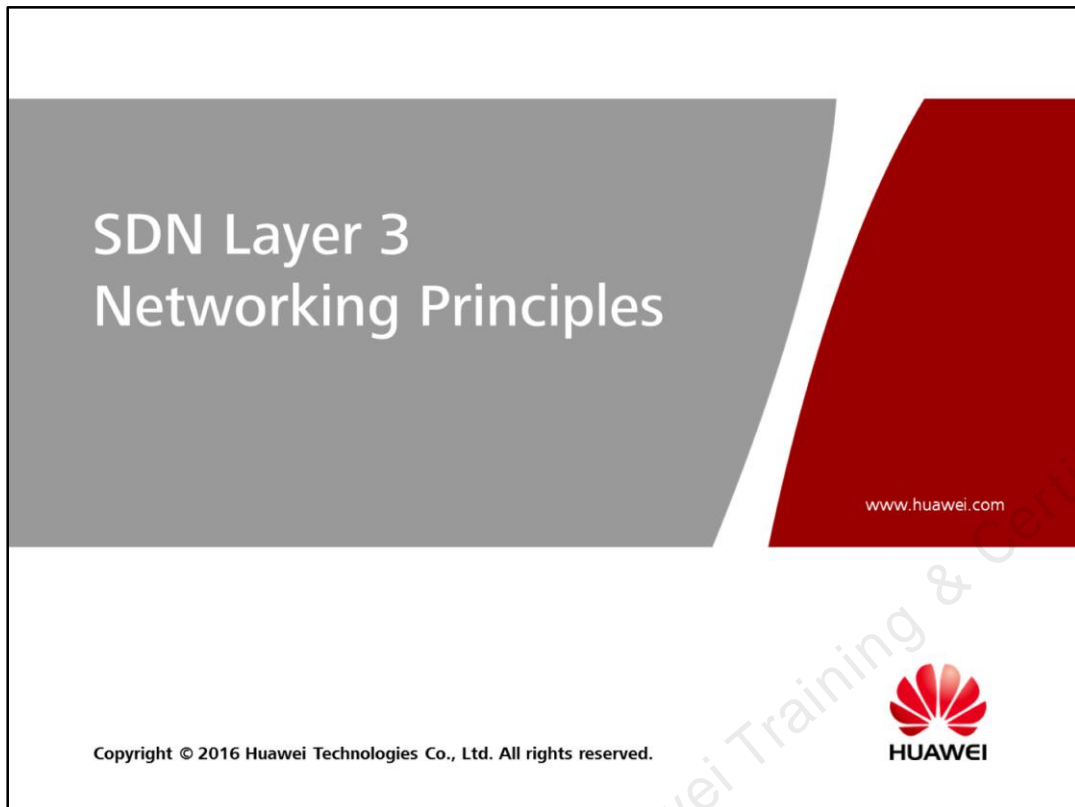
## Summary

- SDN Control Channel Overview
- SDN Layer 2 Networking
- Ethernet Networking

**Thank you**

[www.huawei.com](http://www.huawei.com)

Huawei Training & Certification Huawei Training & Certification



 **Objectives**

- Upon completion of this course, you will be able to:
  - Understand the SDN control channel establishment in layer 3 networking scenario.
  - Understand the basic concepts of OSPF.
  - Understand the basic concepts of ISIS.
  - Understand the basic concepts of BGP.



## Contents

1. SDN Layer 3 Control Channel Establishment
2. SDN Layer 3 Commonly Used Protocols



## Contents

- 1. SDN Layer 3 Control Channel Establishment**
2. SDN Layer 3 Commonly Used Protocols



## SDN Layer 3 Control Channel Establishment

- Control channel, which is referring to the channel reach-ability between SDN controller and forwarder can be realized through Layer 2 or Layer 3 networking method.
- To establish layer 3 control channel, traditional Interior Gateway Protocol (IGP), such as OSPF or ISIS is run between controller and forwarder; controller can be technically considered as a normal router in this case.
- It is important to assure the user data traffic is isolated from the control plane to avoid congestion in the control plane link.

- Control channel, which is referring to the channel reach-ability between SDN controller and forwarder can be realized through Layer 2 or Layer 3 networking method.
- To establish layer 3 control channel, traditional Interior Gateway Protocol (IGP), such as OSPF or ISIS is run between controller and forwarder; controller can be technically considered as a normal router in this case.
- However, SNC controller is normally installed on a server hardware. It does not have a router hardware which is equipped with the dedicated forwarding engine module that is used to forward data traffic. SNC is just a distributed control software running on a server, used for only control packet between SNC and forwarder; Thus, a good SDN network design needs to have the mechanism to avoid user traffic to flow through the SNC controller, which might cause link congestion on the control channel.
- In conjunction of Layer 3 control channel establishment through traditional routing protocols such as OSPF or ISIS, these protocols have features to avoid user traffic to flow through the control channel; for instance, the stub router features in OSPF and overload bit feature in ISIS.
- However, this problem no longer exists if the SNC controller software is installed directly on a router, which is one of the SDN solutions supported by Huawei SNC controller.

## SDN Control Channel Route Establishment: Pre-routing & Flow Triggered Routing

Pre-routing	Flow Triggered Routing
In traditional IP Network, routers use the pre-routing methods to prepare routes for traffic forwarding. Routes are prepared before traffic forwarding and packets with mismatched destination will be discarded.	Flow triggered routing is a route establishment method that can generate routes upon request. When a devices receive a user packet and find out that there is no matching entry in its flow table, the packets will be forwarded to controller; controller will generate a route entry particularly for this packet.
Mature in application	Still not mature in application
Higher security	Lower security as it is fragile for malicious attacks.

- In traditional IP Network, routers use the pre-routing methods to prepare routes for traffic forwarding. Different kinds of routing protocols are used to generate routing entries in routing table in preparation for traffic forwarding to certain destination. If a packet is forwarded through the router and is not matched with any routing entry, it will be discarded.
- Flow triggered routing is a route establishment method that can generate routes upon request. When a devices receive a user packet and find out that there is no matching entry in its flow table, the packets will be forwarded to controller; controller will generate a route entry particularly for this packet.
- In term of security, flow trigger routing is more fragile to malicious attacks. Hackers can attack the controller by sending a huge amount of packets with unknown destination to the forwarder. When forwarder does not has the destination exists in its entry table, the unknown packets will be sent to the controller; when a large amount of packets reaches to the controller, this might cause burden to the controller and causes controller failure.
- Although flow triggered routing is more spontaneous and suitable in SDN applications, it is still not as mature as pre-routing method which is widely used in traditional IP network; thus, only certain SDN applications are deploying flow triggered routing, while some is still using pre-routing as the control channel route establishment.



## Contents

1. SDN Layer 3 Control Channel Establishment
- 2. SDN Layer 3 Commonly Used Protocols**

 **Contents**

## 2. SDN Layer 3 Commonly Used Protocols

**2.1 OSPF Routing Protocol**

## 2.2 ISIS Routing Protocol

## 2.3 BGP Routing Protocol

 **Contents**

## 2. 1. OSPF Routing Protocol

**2.1.1 OSPF Overview**

## 2.1.2 Basic OSPF Concepts

## 2.1.3 OSPF Neighbor and Adjacency Establishment

## 2.1.4 OSPF Route Calculation

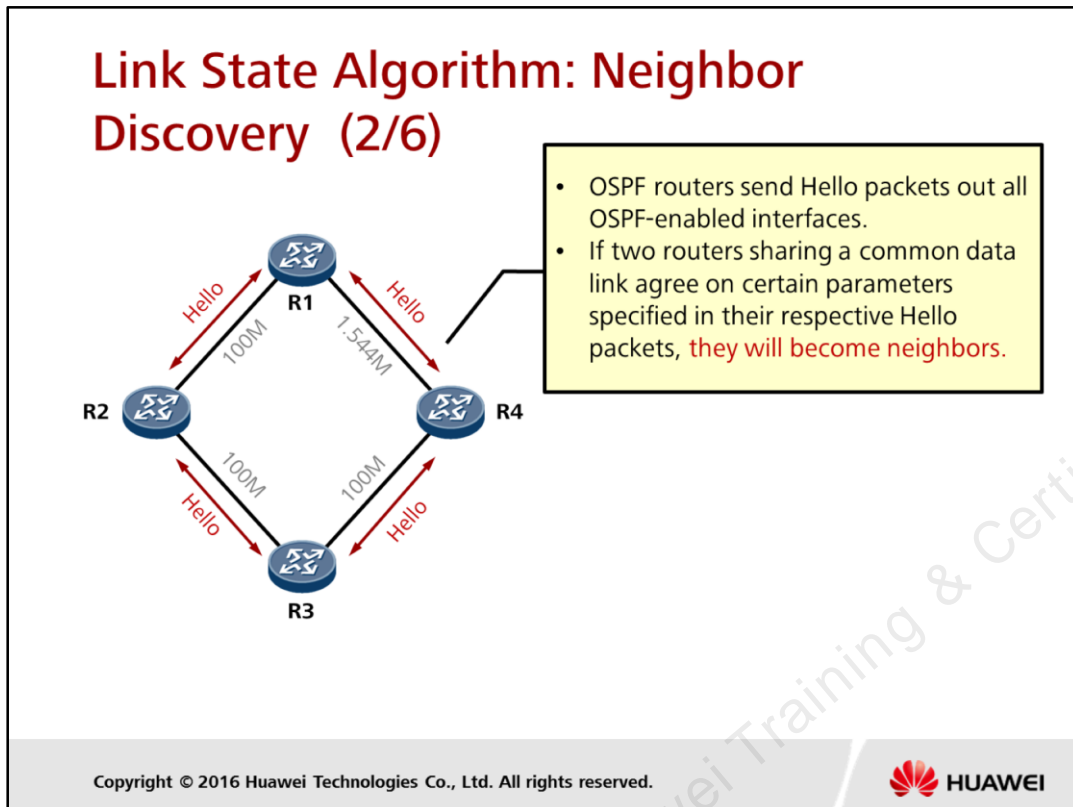
## OSPF Overview (1/6)

- The Open Shortest Path First (OSPF) protocol, developed by the Internet Engineering Task Force (IETF), is a link-state Interior Gateway Protocol (IGP).
- At present, OSPF Version 2, defined in RFC 2328, is intended for IPv4, and OSPF Version 3, defined in RFC 2740, is intended for IPv6.
- OSPF features the following advantages:
  - Receives or sends packets in multicast mode to reduce load on the Router that does not run OSPF.
  - Supports Classless Inter domain Routing (CIDR).
  - Supports load balancing among equal-cost routes.
  - Supports packet encryption
  - Loop free.

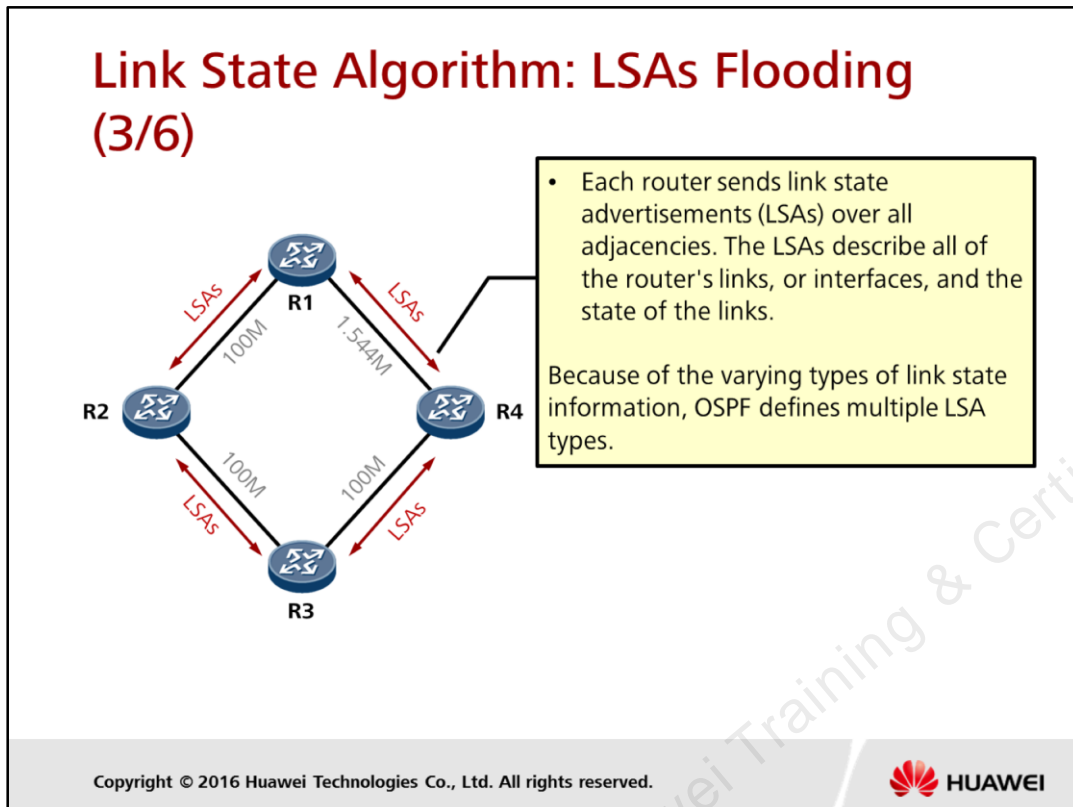
Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.



- The characteristics of OSPF is listed as below:-
  1. **Sending and receiving protocol data by using IP multicast:** OSPF routers send and receive protocol data by using multicast and unicast. Therefore, the network traffic generated is very low.
  2. **Supporting Classless Inter-Domain Routing (CIDR):** As a routing protocol specially developed for TCP/IP environments, OSPF explicitly supports CIDR and Variable-Length Subnet Mask (VLSM).
  3. **Supporting equal-cost routes:** When multiple equal-cost paths exist to the same destination address, the traffic is evenly distributed on these equal-cost paths.
  4. **Free of routing loops:** OSPF calculates routes based on detailed link state information, namely, network topology information, to generate a shortest path tree (SPT) rooted on the local router. Therefore, the routes calculated by OSPF are loop-free.
  5. **Supporting protocol packet authentication:** All the packets exchanged between OSPF routers are authenticated. This ensures network security at the protocol level.

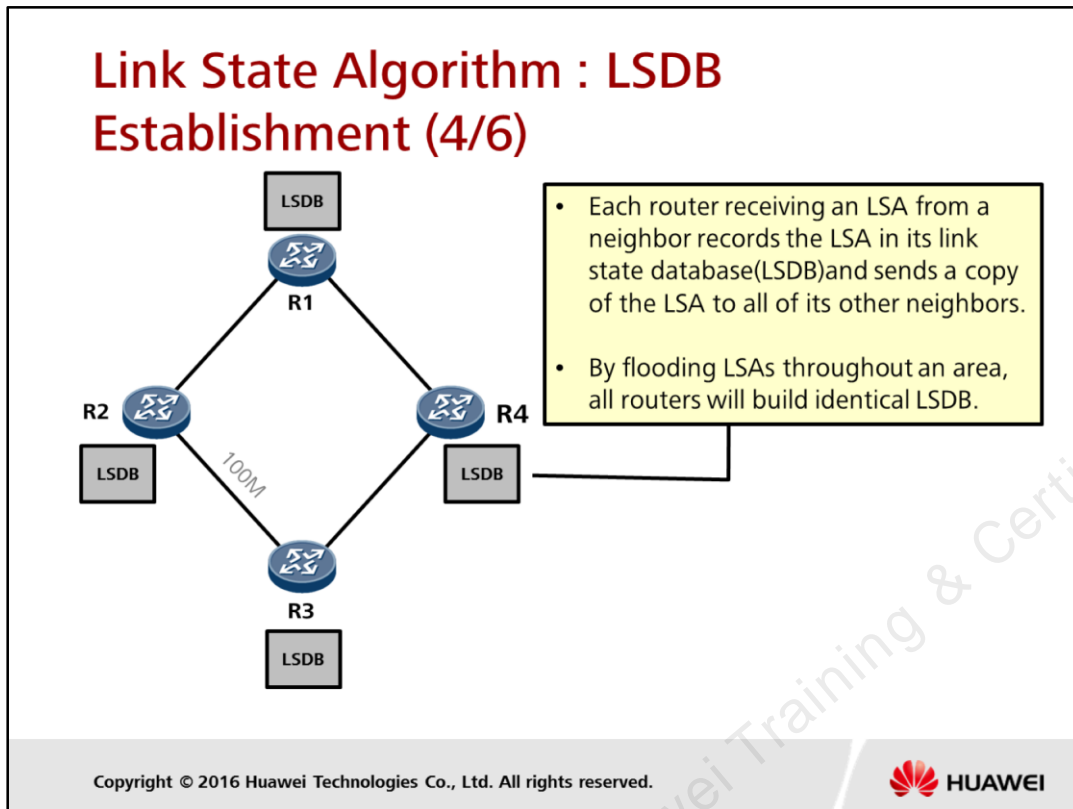


- Neighbor discovery is the initial step of setting up a link state environment before discovering the topology.
- Once the routers are enabled with OSPF, Hello packets will be sent out from all the OSPF-enabled interfaces.
- Information carried in the Hello packets such as router ID, area id, Network masks etc will be used as the parameters to determine whether a neighbor relationships to be established.
- Adjacency can be established after a neighbor relationships is established. More detailed descriptions about OSPF neighbor and adjacency establishment, and their differences will be described in the next chapter in this topic.

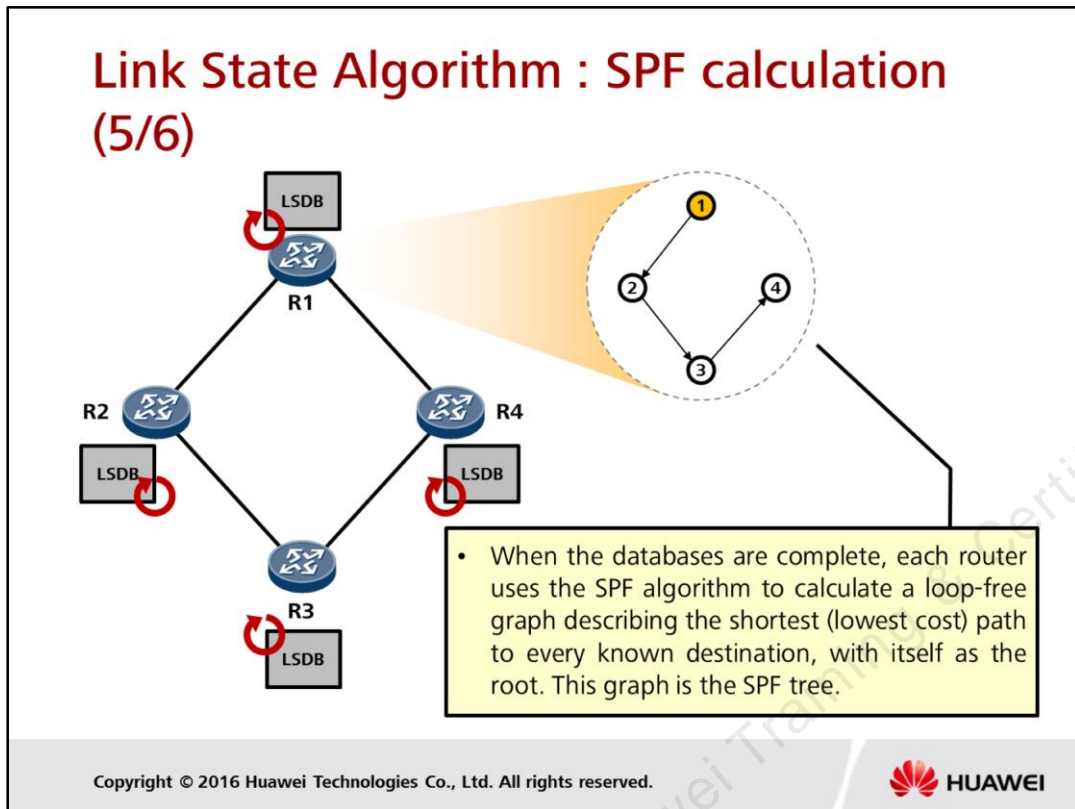


- Upon the adjacency establishment, second step in OSPF link state algorithm is LSA (Link state advertisement) flooding. The routers begin to send out LSA to every neighbor. There are different types of LSA generated, used to describe different information such as routers' links, interfaces, link status etc.
- As per implied in "LSA flooding", LSA is flooded to every OSPF-enabled routers in the network. In other words, a router generates LSA and sends to its directly connected neighbors, and the neighbors will copy and forward the received LSA to all its neighbors, except the one that sent the originating LSA.

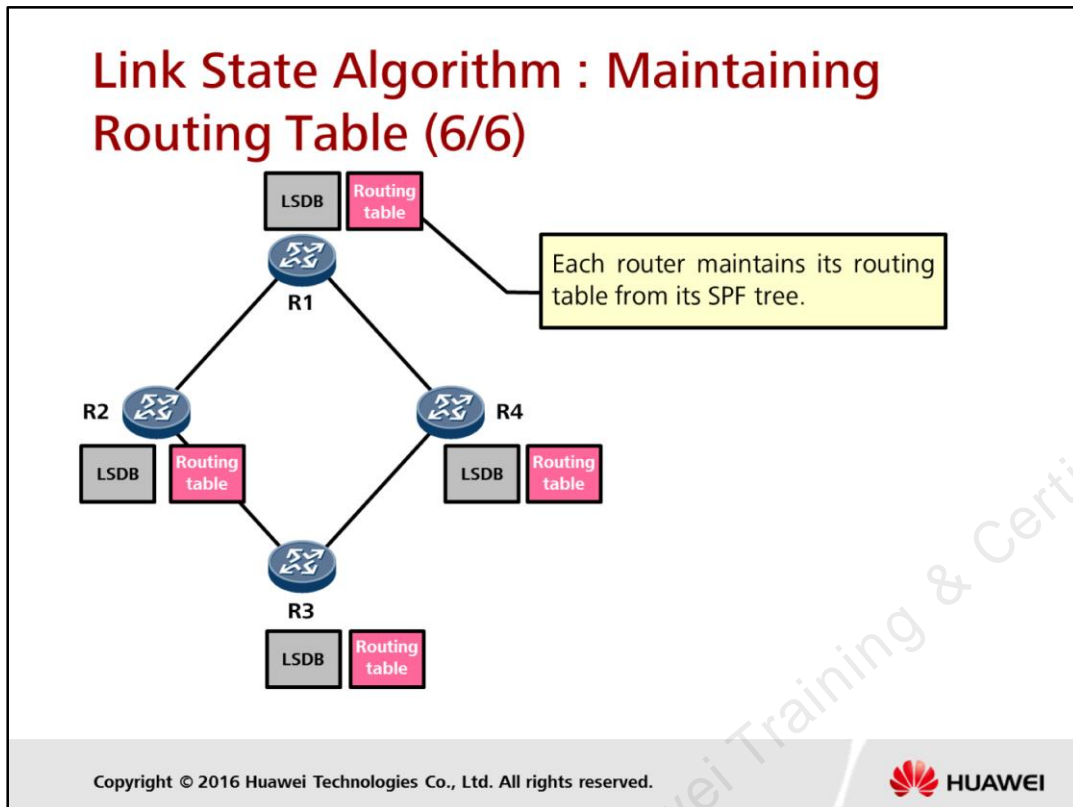




- After performing LSA flooding, all the LSA will be stored in link state database. The LSDB serves as a database to keep and collect all the self-originated LSA and also the LSA received, in order to gather a complete network topology information, as the preparation of route calculation in the later stage.



- After having a complete LSDB, each router will start to perform Shortest Path First (SPF) calculation. The shortest path in link state algorithm is determined by the path metric. Each router will perform SPF to calculate its own shortest path tree; the local router works as the root of the tree, and other nodes serve as the leaves of the SPT.
- As the SPF route calculation algorithm is a loop-free algorithm, a loop-free topology can be guaranteed from the shortest path tree (SPT) calculated.



- From the shortest path tree calculated, each router then converts the SPT into routing table entries and maintain routing table in each router. This is how an OSPF route can be generated and kept in the routing tables. The routes inserted into the routing table will be the shortest routes calculated through OSPF shortest path first algorithm.



## Contents

### 2. 1. OSPF Routing Protocol

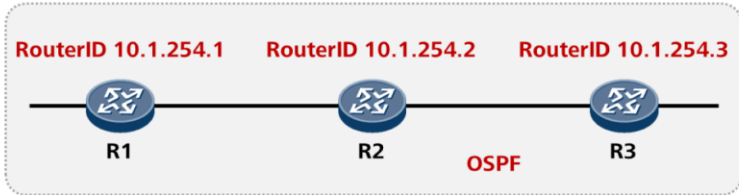
#### 2.1.1 OSPF Overview

#### **2.1.2 Basic OSPF Concepts**


#### 2.1.3 OSPF Neighbor and Adjacency Establishment

#### 2.1.4 OSPF Route Calculation

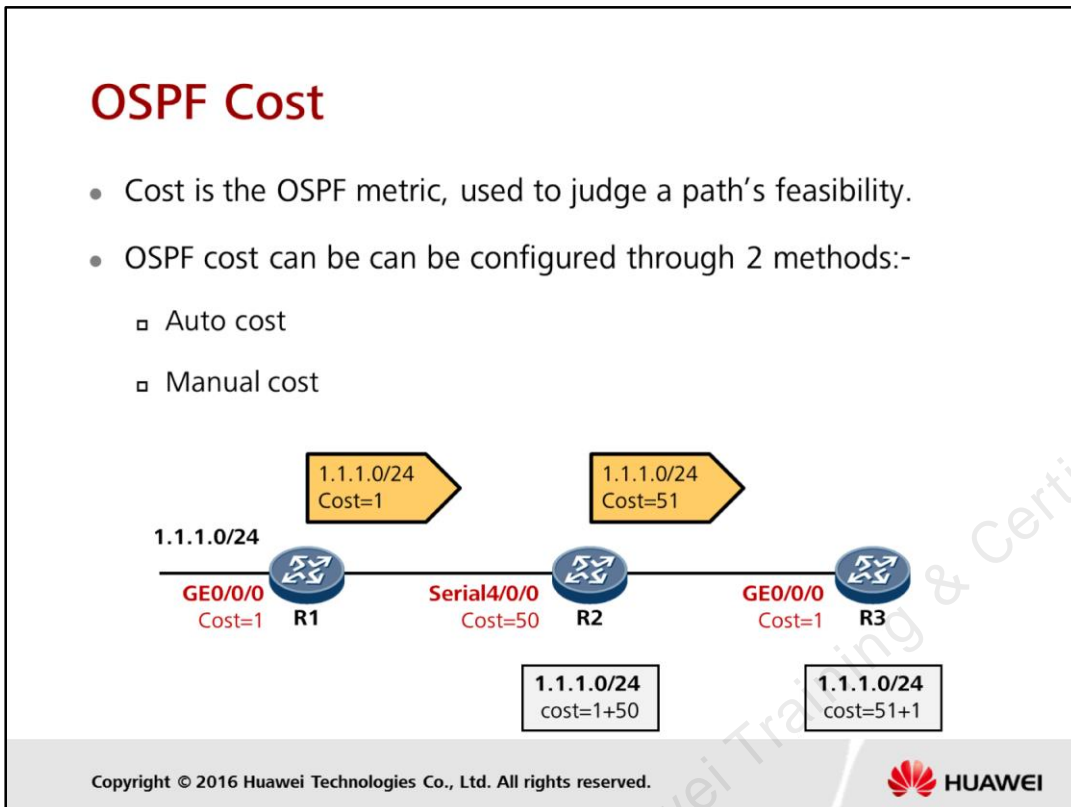
## OSPF Router-ID



The Router-ID is a 32-bit number assigned to each OSPF enabled router, which is used to uniquely identify the router within an autonomous system(or OSPF Domain).

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.  HUAWEI

- During OSPF route calculation, each OSPF router needs to save the link state information about all the routers on the network. To distinguish the link state information about different routers in an LSDB, each router on the network is uniquely identified by a route ID in the LSDB.
- A router ID can be configured manually. If no router ID is specified by using a command, the system automatically selects one of the existing interface IP addresses as the router ID.
- The principle of selecting a router ID is as follows:
  1. The highest IP address among the loopback addresses is preferentially selected as the router ID.
  2. If no loopback interface is configured, the highest IP address among the physical interface addresses is selected as the router ID.
- In any of the following situations, the router ID is re-selected:
  1. The **ospf** command is used to re-configure an OSPF router ID.
  2. The system router ID is re-configured, and then the OSPF process is restarted.
  3. The IP address of the original system router ID is deleted, and then the OSPF process is restarted.



- OSPF cost can be configured through 2 methods, as per listed below:-

### 1. Auto cost

- By default, OSPF automatically calculates the cost of the interface according to the bandwidth of the interface using the formula :  $\text{Cost} = \frac{\text{Bandwidth reference value}}{\text{interface bandwidth}}$ ; default interface bandwidth is  $10^8$ , which is 100MB.
- If the integer result obtained is less than 1, the interface cost is rounded to 1; this causes any interfaces with bandwidth higher than 100MB will have a automatic cost of 1, making the link is not at their optimal usage.
- Thus, the auto cost calculation result can be altered by modifying the bandwidth reference value to a bigger value than 100MB.

### 2. Manual cost

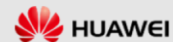
- We can manually set a manual cost for a specific interface by configuring the command "***ospf cost <1-65535>***" under the particular interface view.

## OSPF Area Concept: Why Need Area? (1/4)




- Frequent calculations of the shortest path first (SPF)
- Large link-state table
- Large routing table

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.




- If OSPF is deployed in large network with a large numbers of routers is enabled with OSPF, a few problems will be encountered, as per listed below:-
  1. The running of the SPF algorithm is more complicated and occupies more CPU resources.
  2. All the routers generate LSAs respectively and the LSDBs become very large. Therefore, LSDB synchronization takes long and occupies much memory space. Besides, when the network size grows, the probability of topological changes also increases. As a result, a large number of OSPF packets are transferred on the network. This lowers the bandwidth utilization of the network.
  3. The size of routing table will be increasing in proportion with the increasing route numbers in OSPF.
- OSPF resolves this problem by partitioning an AS into different areas. An area is regarded as a logical group and each group is identified by an area ID.
- An area is a logical group of routers and is identified by an area ID. A network segment (link) belongs to only one area. In other words, the area to which an OSPF interface belongs must be specified.

## OSPF Area: Area Division Concept (2/4)



- Reduce frequency of SPF calculations
- Reduce link-state update (LSU) overhead
- Smaller routing tables

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. 

- An area is a logical collection of OSPF networks, routers, and links that have the same area identification.
- Areas are identified through a 32-bit Area ID expressed in dotted decimal notation like 0.0.0.1 (or decimal integer like area 1).
- All routers within an OSPF area keep a link state database. Each router within the area builds a topology tree of the area, with shortest paths to all other links/routers with itself as the root.
- A router within an area must maintain a topological database for the area to which it belongs.
- The router doesn't have detailed information about network topology outside of its area, thereby reducing the size of its database.

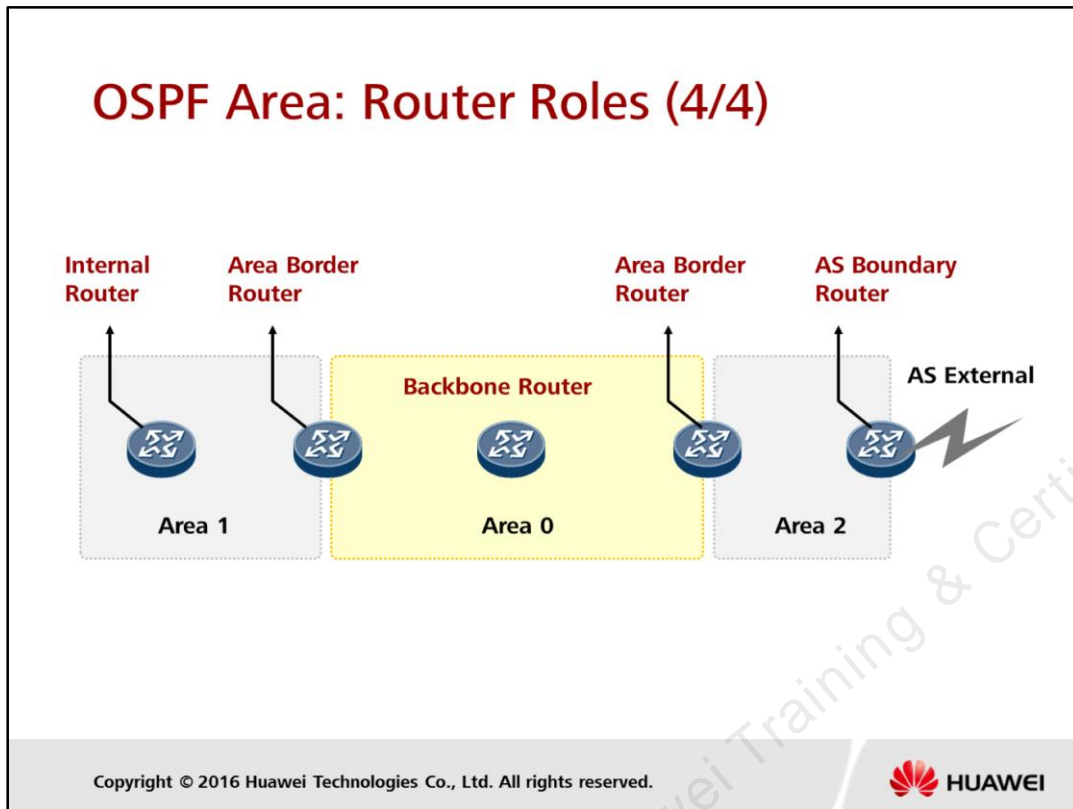


## OSPF Area: Multi-area Design (3/4)

- A backbone area (Area0) connects all the other OSPF areas. .
- The backbone area is responsible for forwarding inter-area routing information. The routing information between the non-backbone areas must be forwarded through the backbone area. So the non-backbone areas must be connected to backbone area directly.

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. HUAWEI

- Area 0 is the backbone area. The backbone area is responsible for advertising the routing information (not detailed link state information) summarized by area border routers (ABRs) between non-backbone areas.
- To prevent inter-area routing loops, OSPF disallows direct inter-area routing information advertisement between non-backbone areas. Therefore, an ABR must have at least one interface to the area 0. That is, each non-backbone area must be connected to the backbone area.
- Each area has an LSDB unique to the area. A router maintains a separate LSDB for each area to which the router is connected. Detailed link state information is not advertised outside any area. Therefore, LSDB sizes are greatly reduced.

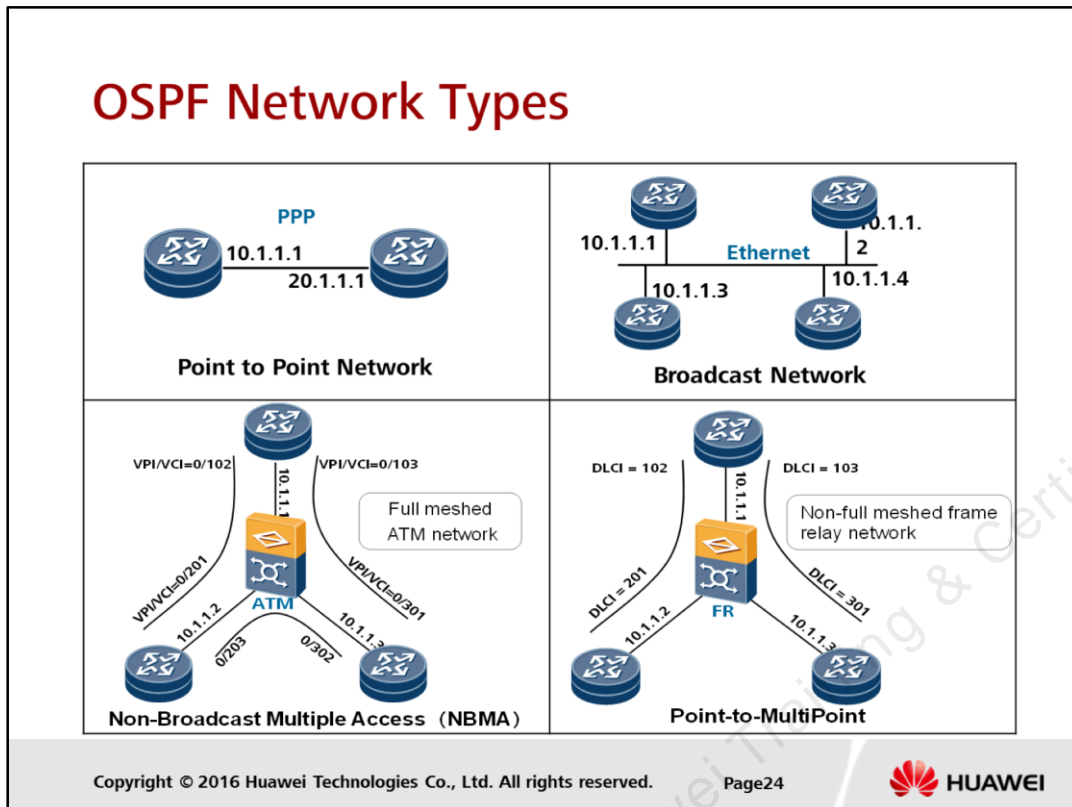


- Due to the area division concept deployed in OSPF, OSPF-enabled router can be classified into different router roles, based on the location of the routers in the OSPF domains and areas, as per listed below:-
  1. **Internal router (IR):** An IR is a router with all directly connected networks belonging to the same area. IRs that belong to the same area maintain the same LSDB.
  2. **Area border router (ABR) :** An ABR is a router directly connected to multiple areas. An ABR maintains an LSDB for each area to which the ABR is directly connected.
  3. **Backbone router:** A backbone router is a router that has at least one interface (or virtual link) to the backbone area. Backbone routers include all the ABRs and the routers with all their interfaces directly connected to the backbone area.
  4. **AS boundary router (ASBR) :**An ASBR is a router that exchanges routing information with routers belonging to other ASs. An ASBR advertises AS-external routes throughout the entire AS. An ASBR can be an IR or ABR. An ASBR can belong to or does not belong to the backbone area.

## OSPF Packet Types

Packet Type	Function
Hello packet	Sent periodically to discover and maintain OSPF neighbor relationships.
Database Description (DD) packet	Contains brief information about the local link-state database (LSDB) and synchronizes the LSDBs on two devices.
Link State Request (LSR) packet	Requests the required LSAs from neighbors. LSR packets are sent only after DD packets are exchanged successfully.
Link State Update (LSU) packet	Sends the required LSAs to neighbors.
Link State Acknowledgement (LSAck) packet	Acknowledges the receipt of an LSA.

- There are five types of OSPF packets. By exchanging protocol packets, OSPF routers establish neighbor relationships among them and exchange link state information to complete route calculation. This section describes the functions of OSPF packets, as per listed below:-
  1. **Hello packets** discover neighbors and maintain neighbor relationships.
  2. **Database description (DD) packets** summarize link states by carrying LSA header information.
  3. **Link state (LS) request packets** are used to request the LSAs that are discovered by receiving DD packets but not available on the local router.
  4. Detailed LSAs are sent in **LS Update packets** to synchronize LSDBs.
  5. **LS Ack packets** are flooded to guarantee reliable routing information exchange.



- In OSPF, four network types are defined: point-to-point (P2P), broadcast, non-broadcast multi-access (NBMA), and point-to-multipoint (P2MP).
1. **P2P:** A P2P network is a network where two routers are directly interconnected.
  2. **Broadcast:** A broadcast network is a network that supports the interconnection of more than two routers and has broadcast capabilities.
  3. **NBMA:** An NBMA network is a network that supports the interconnection of more than two routers but does not have any broadcast capability. On an NBMA network, OSPF simulates the operations performed on a broadcast network, but the neighbors of each router need to be configured manually. All the routers on an NBMA network must be fully-meshed.
  4. **P2MP:** An entire non-broadcast network is considered as a group of P2P networks. The neighbors of each router can be discovered by using a lower-layer protocol, for example, inverse address resolution protocol (ARP).

 **Contents**

## 2. 1. OSPF Routing Protocol

## 2.1.1 OSPF Overview

## 2.1.2 Basic OSPF Concepts

**2.1.3 OSPF Neighbor and Adjacency Establishment**

## 2.1.4 OSPF Route Calculation

## OSPF Neighbor & Adjacency: Overview (1/ 3)

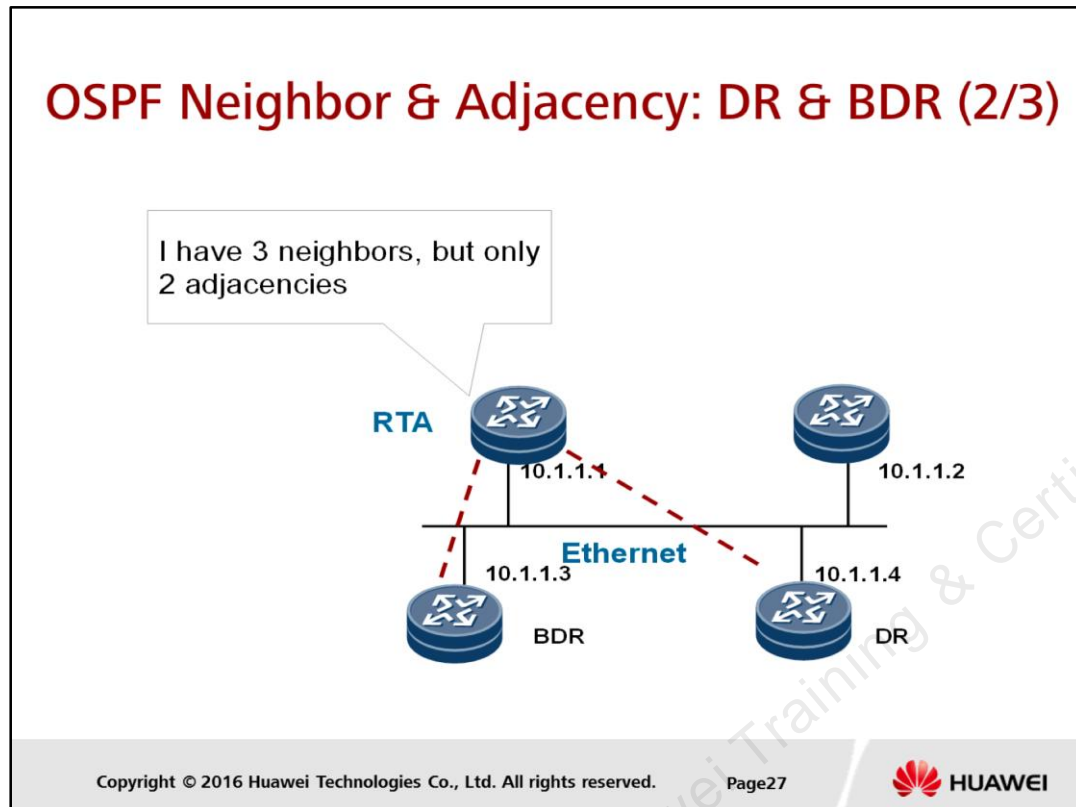
I have 3 neighbors, but only 2 adjacencies

Neighbor	Routers that are directly connected to the same network segment. Neighbor relationships are maintained through Hello packets
Adjacency	An adjacency is a neighbor relationship selected to exchange routing information. An adjacency relationships involve the exchange of DD, LSR, LSU and LS Ack packet

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page26

**HUAWEI**

- Not all the neighbor relationships can become adjacencies. Whether an adjacency is established also varies with network types.
- In this example, RTA and the other three routers are directly connected to the same network segment. As shown in the preceding figure, OSPF runs on all the interfaces of all the routers. RTA establishes neighbor relationships with the other three routers. However, RTA only forms the adjacency relationships with 2 routers which work as Designated Router (DR) and Backup Designated Router (BDR).



- In a broadcast network or NBMA network, routing information needs to be transferred between any two routers. If  $n$  routers exist in the network,  $n(n-1)/2$  adjacencies need to be established. As a result, a route change on any router needs to be transferred for multiple times and bandwidth resources are wasted. To solve this problem, DR is defined in the OSPF protocol and all the routers only need to send information to the DR for broadcasting the network link states.
- If the DR fails due to a fault, all the routers in the network must re-elect the DR and be synchronized to the new DR. During this process, which takes quite long, route calculation may be incorrect. To shorten this process, the BDR concept is defined in OSPF.
- The router roles in OSPF network is listed below:-
  1. **DR:** A DR is the router that maintains adjacencies with all the other OSPF routers on the same network segment and exchanges LSAs with these routers.
  2. **BDR:** A BDR is a backup DR.
  3. **DR Other:** A router that is neither a DR nor BDR is a DR Other. DR Others do not form adjacencies between themselves or exchange routing information. Therefore, the number of adjacencies formed between the routers on the broadcast network or NBMA network is reduced.



## OSPF Neighbor & Adjacency : DR & BDR Election (3/3)

- Instead of being manually specified, the DR and BDR are elected among all the routers on the local network segment. The DR priority of a router interface determines the eligibility of the interface in the DR election and BDR election.

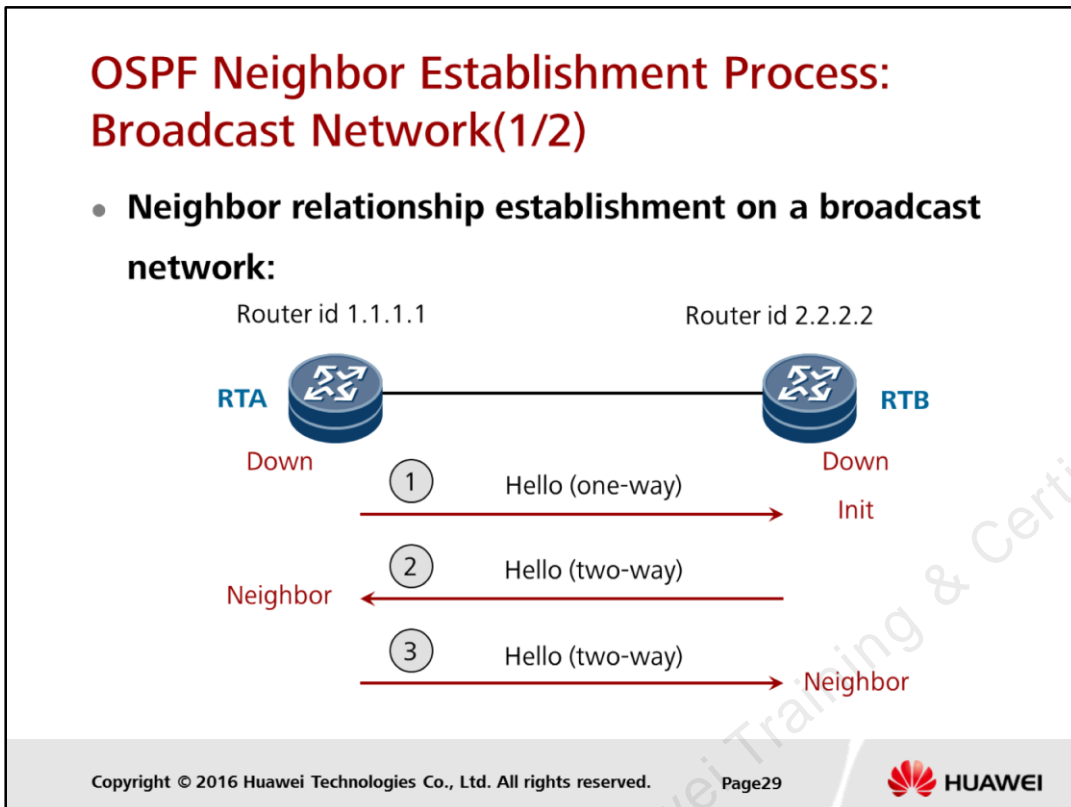
Router with highest Router Priority may not be DR/BDR

Red numbers indicate Router Priority of interface

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page28 HUAWEI

- The DR and BDR are elected by the Hello protocol. Each router writes the DR it votes for into a Hello packet advertised to other routers on the same network segment. The selection criteria of DR as BDR is listed as below:-
  1. When two routers on the same network segment declare themselves the DR, the router with a higher DR priority wins.
  2. If the DR priorities are the same, the router with a larger router ID wins.
  3. A router with the priority 0 is not elected as DR or BDR.
- DR is elected only on broadcast or NBMA interfaces. No DR is elected on P2P or P2MP interfaces.
- DR is based on the network segment and relative to a router interface. A router that functions as the DR on an interface may be a BDR or DR Other on another interface.
- If the DR and BDR are elected, a newly added router, regardless of its DR priority, does not become the DR of the network segment immediately.
- The DR is not necessarily the router with the highest DR priority. Likewise, the BDR is not necessarily the router with the second highest DR priority.
- On the Ethernet shown in the preceding figure, the DR is 10.1.1.1 and the BDR is 10.1.1.2. If a router is added to the network, configure the priority of the added router as 120, which is greater than the priority of the original DR, 100, and the priority of the original BDR, 90. The added router does not become the new DR though it has the highest priority. This maintains the network stability.

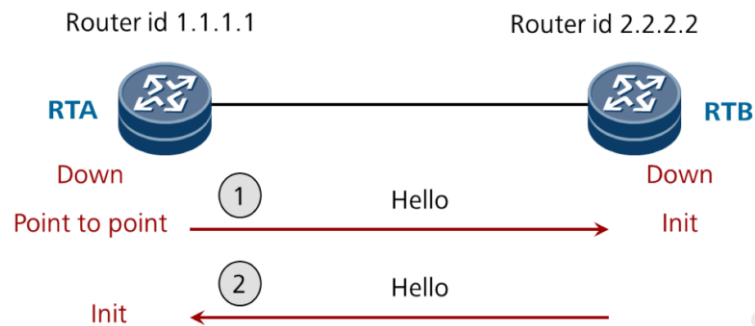




- When the OSPF state becomes "Two-way" on a broadcast network, it means this Router's neighbor relationship is established.

## OSPF Neighbor Establishment Process: P2P Network(2/2)

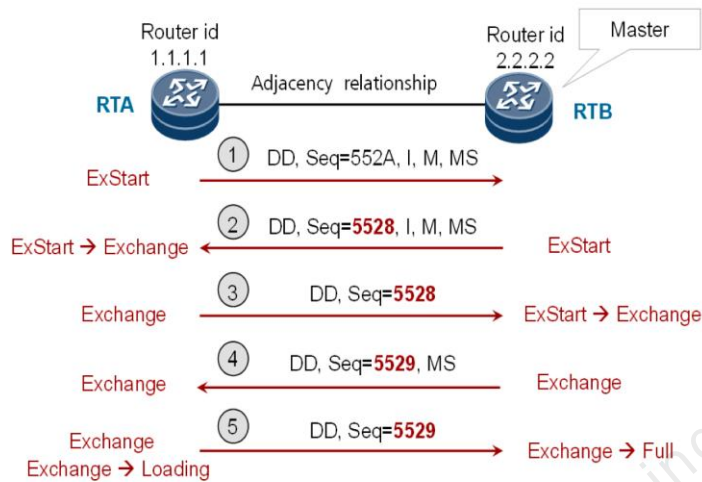
- **Neighbor relationship establishment on a P2P network:**



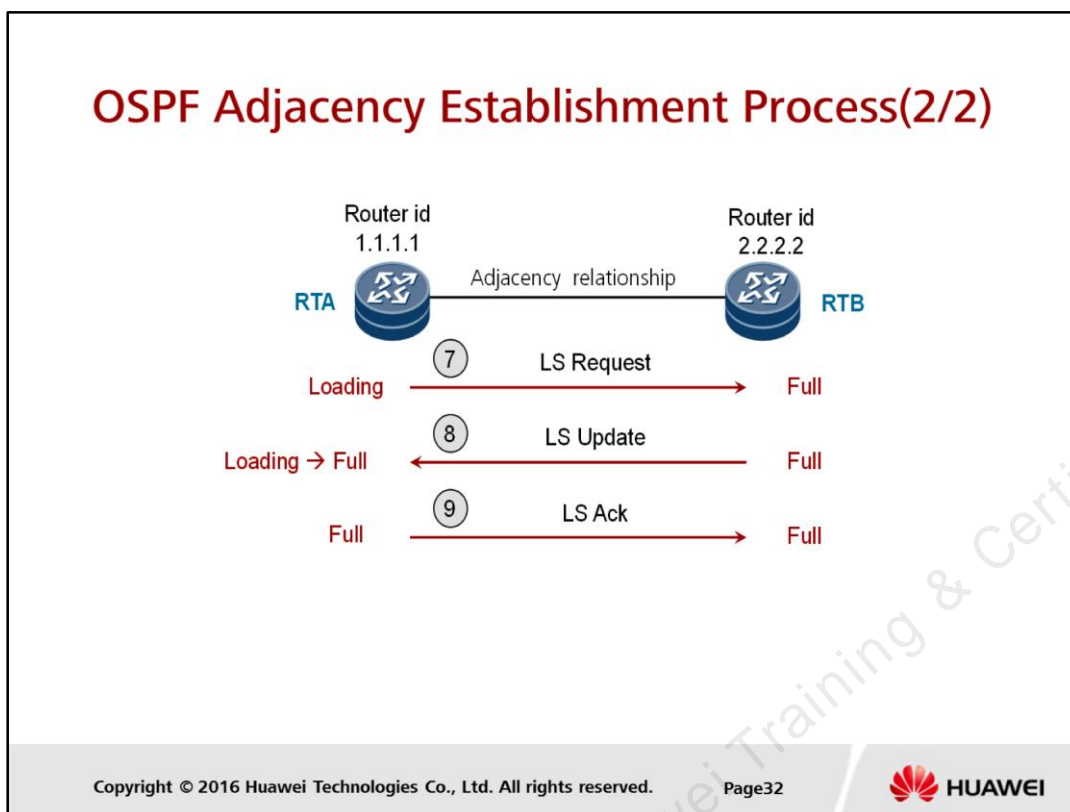
- No DR or BDR needs to be elected on a P2P link, P2MP link, or virtual link.

- No DR or BDR needs to be elected on a P2P link, P2MP link, or virtual link.
- After the interface goes up, the router enters the point-to-point state and attempts to establish a neighbor relationship with its neighbor.
- After the interface receives a Hello packet, the router enters the Init state. This process is different from that on a broadcast link or NBMA link.

## OSPF Adjacency Establishment Process(1/2)



- After the neighbor state becomes ExStart, RTA sends RTB the first DD packet in which the DD sequence number is set to 552A (assumed). The Init bit is set to 1, indicating that this packet is the first in the sequence of DD packets. The More bit is set to 1, indicating that more DD packets are to follow. The Master/Slave bit is set to 1, indicating that the router is the master during the database exchange process.
- After the neighbor state becomes ExStart, RTB sends RTA the first DD packet in which the DD sequence number is set to 5528 (assumed). As the router ID of RTB is larger than that of RTA, RTB should function as the master. After the router ID comparison is complete, RTA generates a NegotiationDone event. Therefore, the RTA neighbor state transitions from ExStart to Exchange.
- After the neighbor state becomes Exchange, RTA sends a new DD packet carrying the LSDB summary information. The DD sequence number is set to that used by RTB in step 2. The More bit is set to 0, indicating that no more DD packet is needed to describe the LSDB. The Master/Slave bit is set to 0, indicating that RTA asserts itself as the slave. On receiving the DD packet, a NegotiationDone event is generated on RTB. Therefore, the state of RTB changes to Exchange.
- After the neighbor state changes to Exchange, RTB sends a new DD packet that carries LSDB description information, with the DD sequence number set to 5529 (the previously used DD sequence number increase 1).
- RTA does not need any new DD packet to describe its LSDB. Functioning as the slave, however, RTA needs to acknowledge each DD packet sent by RTB, which is the master. Therefore, RTA sends RTB a new and empty DD packet whose sequence number is 5529.
- After sending the last DD packet, RTA generates an ExchangeDone event to change the neighbor state to Loading. After RTB receives the last DD packet, if the LSDB of RTB is the most recent and RTB does not need to send any update request to RTA, RTB transits to the Full state.



7. After the neighbor state becomes Loading, RTA starts sending LSRs to RTB, asking for the link state information that is discovered by DD packets in the Exchange state but is not found in the local LSDB.
8. After receiving the LSR, RTB sends an LSU carrying the detailed information about the requested link state to RTA. On receiving the LSU, RTA changes the neighbor state from Loading to Full.
9. RTA sends LS Ack packets to RTB to ensure reliable information transmission. LS Ack packets are used to flood the acknowledgments of received LSAs.
10. The neighbor state becomes Full, indicating that an adjacency is fully established.

 **Contents**

## 2. 1. OSPF Routing Protocol

## 2.1.1 OSPF Overview

## 2.1.2 Basic OSPF Concepts

## 2.1.3 OSPF Neighbor and Adjacency Establishment

**2.1.4 OSPF Route Calculation**

## OSPF LSA Types (1/2)

LS Type	LSA Type	Function
Type1	Router-LSA	Describes the link status and cost of a router. Router-LSAs are generated by a router and advertised within the area to which the router belongs.
Type2	Network-LSA	Describes the link status of all routers on the local network segment. Network-LSAs are generated by a DR and advertised within the area to which the DR belongs.
Type3	Network-Summary-LSA	Describes routes on a network segment. Network-summary-LSAs are generated by an ABR and are advertised within the non-totally stub area or NSSA.
Type4	ASBR-Summary-LSA	Describes routes to an ASBR in an area. ASBR-summary-LSAs are generated by an ABR and are advertised to the areas except the area to which the ASBR belongs.

- LSAs contain the topology information required for OSPF route calculation and provide basis for OSPF route calculation. Each router in an AS generates one type or multiple types of LSAs depending on router types. In OSPF, routing information description is encapsulated in advertised LSAs.

## OSPF LSA Types (2/2)

LS Type	LSA Type	Function
Type5	AS-external-LSA	Describes AS external routes, which are advertised to all areas except stub areas and NSSAs. AS-external-LSAs are generated by an ASBR.
Type7	NSSA LSA	Describes AS external routes. NSSA-LSAs are generated by an ASBR and advertised only within NSSAs.
Type9/ Type10/ Type11	Opaque LSA	Opaque LSA Provides a general mechanism for OSPF extension : <ul style="list-style-type: none"> <li>●Type 9 LSAs are advertised only on the network segment where the interface advertising the LSAs resides.</li> <li>●Type 10 LSAs are advertised within an OSPF area.</li> <li>●Type 11 LSAs are advertised within an AS but have not been used in practice.</li> </ul>

 **Contents**

## 2. SDN Layer 3 Commonly Used Protocols

## 2.1 OSPF Routing Protocol

**2.2 ISIS Routing Protocol**

## 2.3 BGP Routing Protocol



 **Contents**

## 2.2. IS-IS Routing Protocol

**2.2.1 IS-IS Overview**

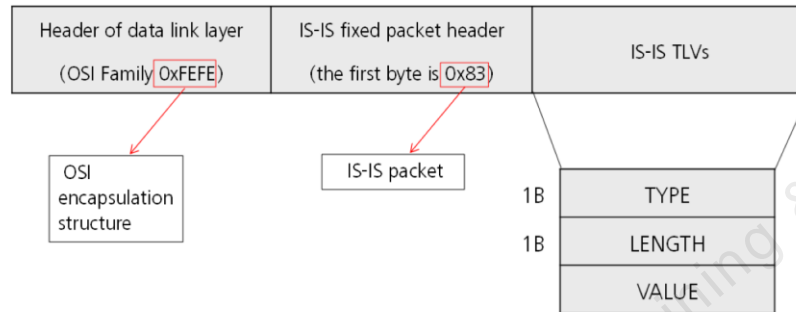
## 2.2.2 Basic Concepts of IS-IS

## 2.2.3 IS-IS Route Calculations

## 2.2.4 Route Leaking

## IS-IS Characteristics

- IS-IS runs at the link layer directly, it transfers link information by sending protocol data units (PDUs), thereby completing LSDB synchronization.
- IS-IS protocol packets adopt the type-length-value (TLV) format to support new feature extension easily.



Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page38



- As an interior gateway protocol (IGP) based on the link state algorithm, IS-IS resembles OSPF in many ways. IS-IS, however has two characteristics different from those of OSPF.
- Packets constructed by using the TLV structure allow high flexibility and scalability. The overall structure of a packet adopting the TLV format is fixed. The only difference of packets is their TLV parts. In addition, multiple TLV structures can be used in one packet and TLVs can be nested. A new feature can be supported by adding the corresponding TLV structure type without changing the entire packet structure.
- The TLV design allows IS-IS to easily support new technologies such as traffic engineering (TE) and IPv6.

 **Contents**

## 5.2. IS-IS

## 5.2.1 IS-IS Overview

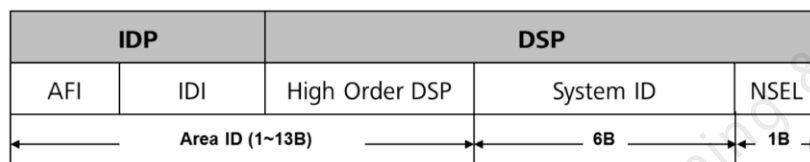
**5.2.2 Basic Concepts of IS-IS**

## 5.2.3 IS-IS Route Calculations

## 5.2.4 Route Leaking

## ISIS Basic Concepts: Network Entity (1/2)

- NSAP: Network Service Access Point, it's used to locate resources.
- NET: Network Entity Titles, A NET can be regarded as a special NSAP
  - In IP, n-selector=0
  - A NET uniquely identifies a network device in the OSI




Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page40



- An NSAP is composed of the Initial Domain Part (IDP) and the Domain Specific Part (DSP). IDP is the counterpart of network ID in an IP address, and DSP is the counterpart of the subnet number and host address in an IP address.
- IDP consists of the Authority and Format Identifier (AFI) and Initial Domain Identifier (IDI). AFI specifies the address assignment mechanism and the address format; the IDI identifies a domain.
- DSP consists of the High Order DSP (HODSP), system ID, and NSAP Selector (SEL). The HODSP is used to divide areas; the system ID identifies a host; the SEL indicates the service type.
- The lengths of the IDP and DSP are variable. The length of the NSAP varies from 8 bytes to 20 bytes.
- A NET can be regarded as a special NSAP. The length of the NET field is the same as that of an NSAP, varying from 8 bytes to 20 bytes.

## ISIS Basic Concepts: Network Entity (2/2)




```
[Quidway]isis
[Quidway-isis]network-entity 49.0021.1921.6800.1001.00
```

<b>49.0021</b>	<b>. 1921.6800.1001</b>	<b>. 00</b>
AreaID	SystemID	N-SEL

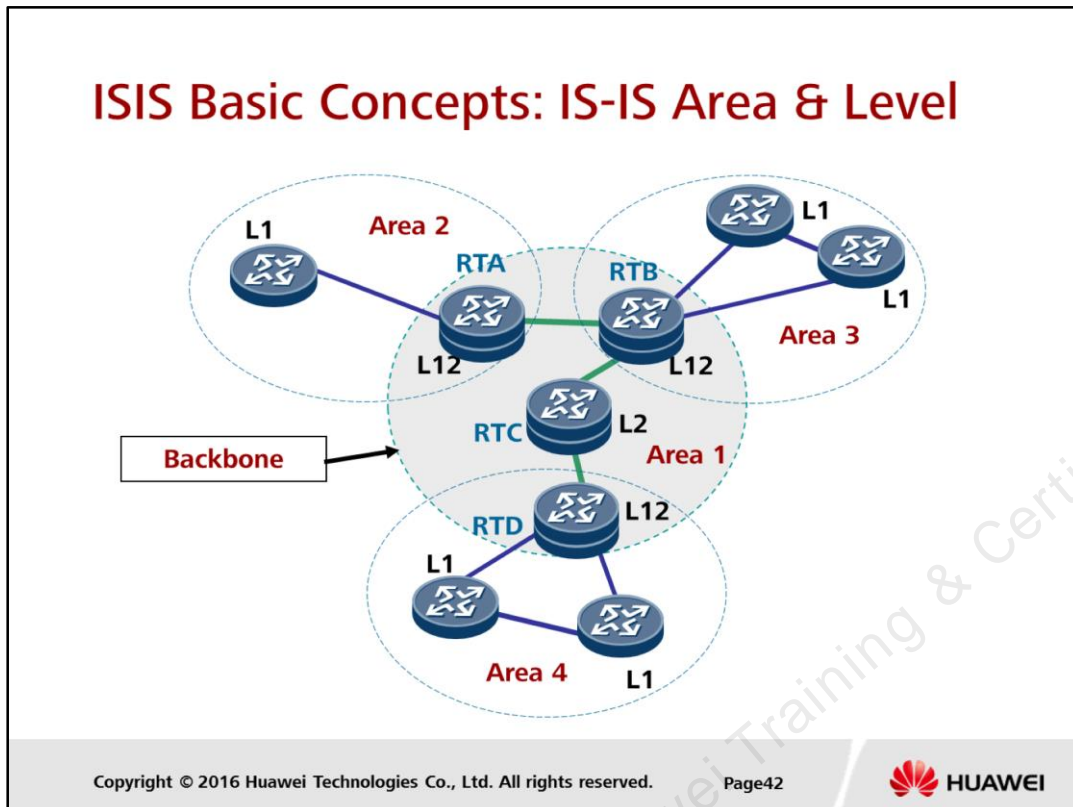
  

<b>88.0001.0755</b>	<b>. 000f.e225.da08</b>	<b>. 00</b>
AreaID	SystemID	N-SEL

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page41



- A NET address length ranges from 8 octets to 20 octets. A NET can be divided into two parts: area ID and system ID.
- A system ID length ranges from 1 byte to 8 bytes. The lengths of all the ISs in the entire routing domain, however, must be the same. The Huawei versatile routing platform (VRP) specifies a system ID length as 6 bytes. A system ID can be converted from a MAC address or IP address.
- A system ID must be unique on the network. In the current VRP version, a router can be configured with a maximum of three NETs that can be switched to adapt to network changes.
- In actual application, router IDs are often used to generate system IDs. Suppose that the IP address of Loopback0, 192.168.1.1, is used as the router ID of a router. Add 0s to 192.168.1.1 so that each part of the address contains three digits, namely, 192.168.001.001. Then, divide these 12 digits into three parts, namely, 1921.6800.1001, which is the system ID.



- To support large-scale routing networks, IS-IS adopts a two-level structure in a routing domain. A large domain is divided into areas. Three types of devices on the IS-IS network are described as follows:
- **Level-1 device:** A Level-1 device manages intra-area routing. It establishes neighbor relationships with only the Level-1 and Level-1-2 devices in the same area and maintains a Level-1 LSDB. The LSDB contains routing information in the local area. A packet to a destination beyond this area is forwarded to the nearest Level-1-2 device.
- **Level-2 device:** A Level-2 device manages inter-area routing. It can establish neighbor relationships with all Level-2 devices and Level-1-2 devices, and maintains a Level-2 LSDB which contains inter-area routing information.
- **Level-1-2 device:** A device, which can establish neighbor relationships with both Level-1 devices and Level-2 devices, is called a Level-1-2 device. A Level-1-2 device can establish Level-1 neighbor relationships with Level-1 devices and Level-1-2 devices in the same area. It can also establish Level-2 neighbor relationships with Level-2 devices and Level-1-2 devices in other areas. Level-1 devices can be connected to other areas only through Level-1-2 devices.

## ISIS Basic Concepts: IS-IS Authentication

Based on packet types, the authentication is classified as follows :

- **Interface authentication:** authenticate Level-1 and Level-2 IS-to-IS Hello PDUs (IIHs).
- **Area authentication :** authenticate Level-1 CSNPs, PSNPs, and LSPs.
- **Routing domain authentication :** authenticate Level-2 CSNPs, PSNPs, and LSPs.

- As the Internet develops, more data, voice, and video information are exchanged over the Internet. New services, such as e-commerce, online conferencing and auctions, video on demand, and distance learning, emerge gradually. The new services have high requirements for network security. Carriers need to prevent data packets from being intercepted or modified by attackers or unauthorized users. IS-IS authentication applies to the area or interface where packets need to be protected. Using IS-IS authentication enhances system security and helps carriers provide safe network services.

- 3 types of ISIS authentication methods are listed below:-

### 1. Interface Authentication

- Authentication passwords for IIHs are saved on interfaces. The interfaces send authentication packets with the authentication TLV. Interconnected router interfaces must be configured with the same password.

### 2. Area Authentication

- Every router in an IS-IS area must use the same authentication mode and have the same key chain.

### 3. Routing Domain Authentication

- Every Level-2 or Level-1-2 router in an IS-IS area must use the same authentication mode and have the same key chain.

 **Contents**

## 5.2. IS-IS

## 5.2.1 IS-IS Overview

## 5.2.2 Basic Concepts of IS-IS

**5.2.3 IS-IS Route Calculations**

## 5.2.4 Route Leaking



## ISIS Protocol Packets (1/2)

PDU Type	Type Value	Packet Function
L1 LAN IIH	15	Hello packets are used to set up and maintain neighbor relationships.
L2 LAN IIH	16	
P2P IIH	17	
L1 LSP	18	Link State PDU (LSP)s are used to exchange link state information.
L2 LSP	20	

- Hello packets, also called the IS-to-IS Hello PDUs (IIHs), are used to set up and maintain neighbor relationships. Level-1 LAN IIHs are applied to the Level-1 routers on broadcast LANs. Level-2 LAN IIHs are applied to the Level-2 routers on broadcast LANs. P2P IIHs are applied to non-broadcast networks.
- LSPs are used to exchange link-state information. There are two types of LSPs: Level-1 and Level-2. Level-1 IS-IS transmits Level-1 LSPs. Level-2 IS-IS transmits Level-2 LSPs. Level-1-2 IS-IS can transmit both Level-1 and Level-2 LSPs.

## ISIS Protocol Packets (2/2)

PDU Type	Type Value	Packet Function
L1 CSNP	24	CSNPs carry summaries of all LSPs in LSDBs, which ensures LSDB synchronization between neighboring routers.
L2 CSNP	25	
L1 PSNP	26	PSNPs list only the sequence numbers of recently received LSPs. A PSNP can acknowledge multiple LSPs at a time. If an LSDB is not updated, PSNPs are also used to request a new LSP from a neighbor.
L2 PSNP	27	

- CSNPs carry summaries of all LSPs in LSDBs, which ensures LSDB synchronization between neighboring routers. On a broadcast network, the designated intermediate system (DIS) sends CSNPs at an interval. The default interval is 10 seconds. On a P2P link, neighboring devices send CSNPs only when a neighbor relationship is established for the first time.
- PSNPs list only the sequence numbers of recently received LSPs. A PSNP can acknowledge multiple LSPs at a time. If an LSDB is not updated, PSNPs are also used to request a new LSP from a neighbor.

## IS-IS Network Types

- In IS-IS, two network types are defined: point-to-point and broadcast.
  - Point-to-point: A point-to-point network is a network where two routers are directly interconnected; common link layer protocol: PPP, HDLC, it is recommended to configure a non-broadcast multi-access (NBMA) as a point-to-point network.
  - Broadcast: A broadcast network is a network that supports the interconnection of more than two routers and has broadcast capabilities; common link layer protocol: Ethernet, Token Ring.

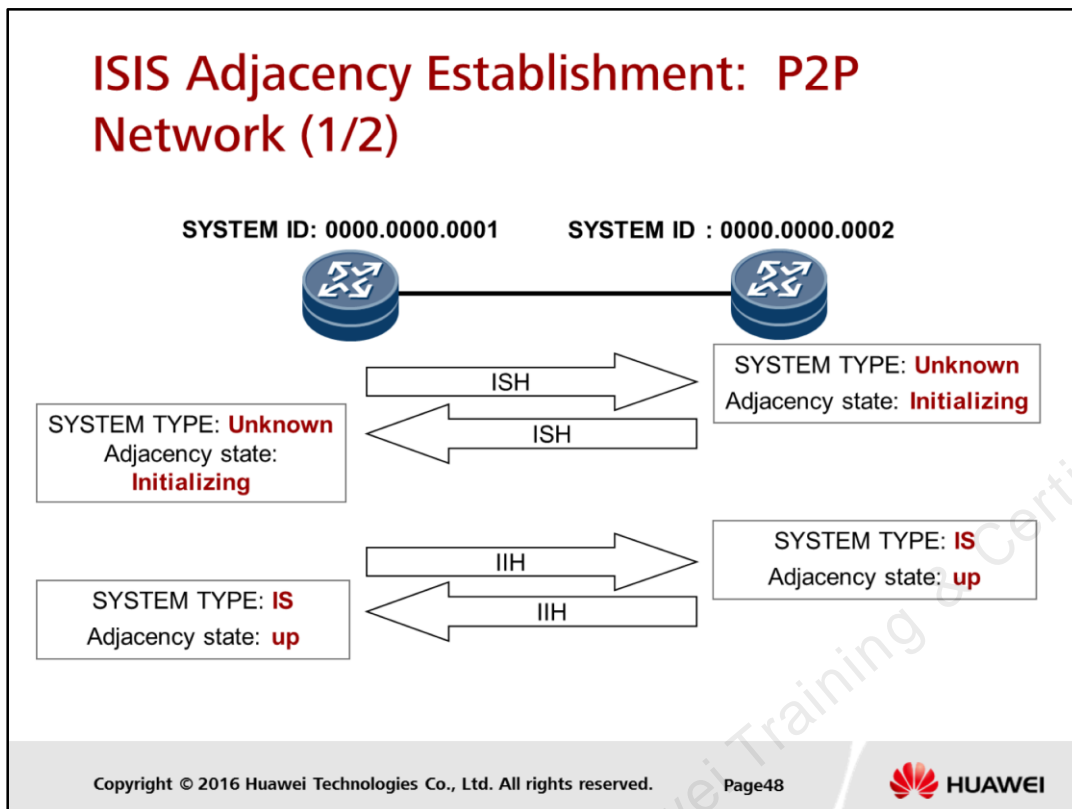


Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

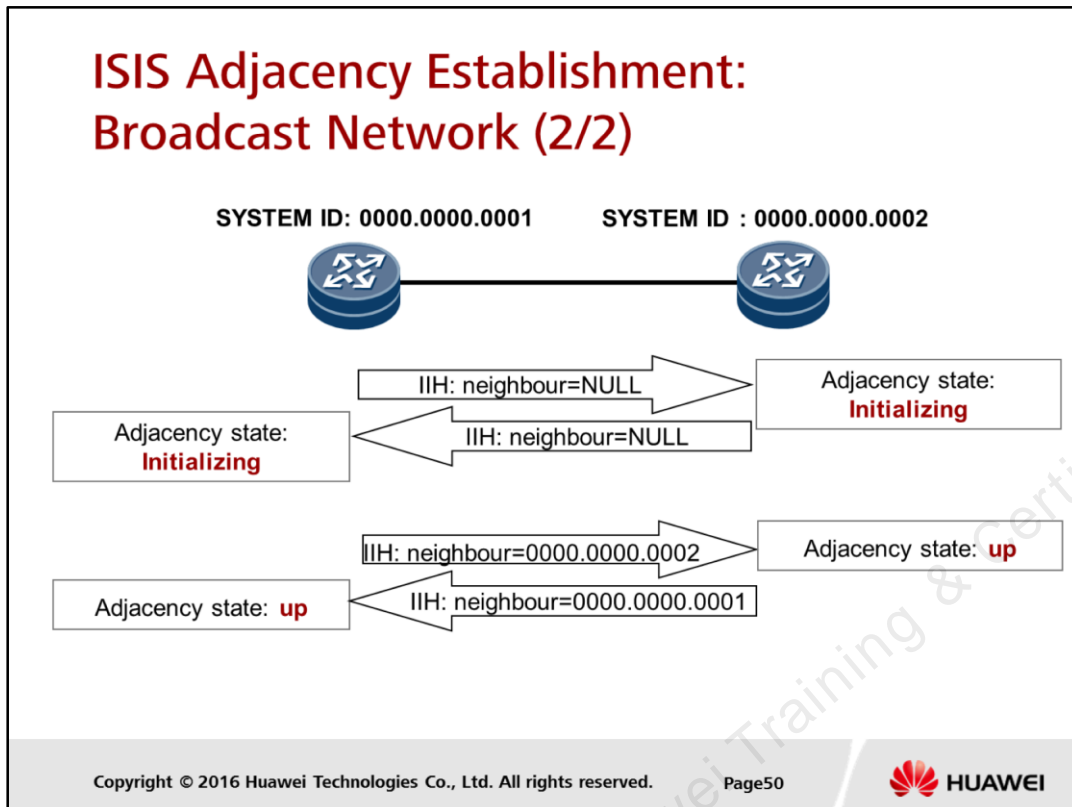
Page47



- IS-IS can run on the P2P network and broadcast network.
- Adjacency establishment mechanisms vary with network types.
- Before two IS-IS routers can exchange protocol packets to implement routing functions, a neighbor relationship must be established between them. IS-IS neighbor relationship establishment modes vary with network types.
- In theory, IS-IS does not support NBMA or point-to-multipoint (P2MP) networks. If IS-IS needs to be deployed on an NBMA or P2MP network, convert the NBMA or P2MP network into a point-to-point network by configuring sub-interfaces.

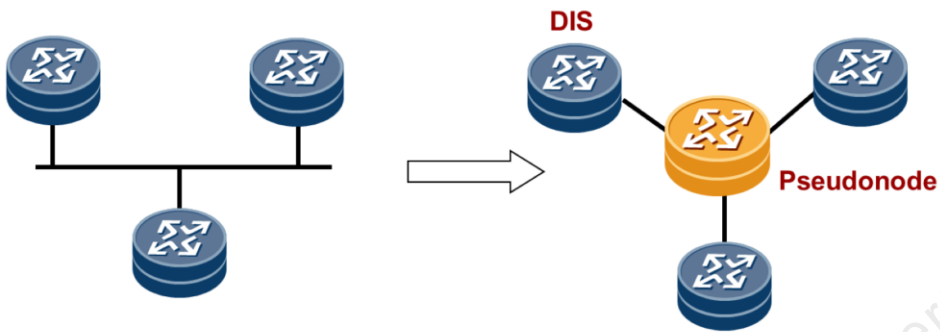


- Adjacency relationship on Point-to-Point links are initialized by the receipt of ISH packets through the ES-IS protocol. ISH packet is HELLO packet sent by IS, this type of packet is defined in ES-IS protocol and we will not discuss it in detail here.
- When an ISH is received on a newly enabled point-to-point link, the router verifies whether an adjacency already exists with the sender by checking the source SysID in the ISH against its adjacency database. The ISH is ignored if an adjacency exists. If not, the receiving router creates a new adjacency and sets its state to "initializing" and the system type to "unknown".
- The router then sends the new neighbor an IIH in response. Upon receiving a subsequent IIH from the new neighbor, the router will check the parameters inside the packet. If the checking is passed, the router then moves the adjacency to an 'up' state and changes the neighbor's system type to IS. Thus, Point to point adjacency relationship formation is a " 2 way handshake process".
- In this process, the local router is unable to determine whether its hellos reach the remote end. This might lead to situations where one end of an adjacency is up but the other end is not. Later, RFC has defined a more reliable way to form point-to-point IS-IS adjacencies - by using a three-way handshake process. The three-way handshake process for reliably forming point-to-point adjacencies introduces a new type length value field (Type 240), known as Point-to-Point Adjacency State TLV. It records the remote ID in Point to Point IIH packet.




- At broadcast network, the adjacency relationship is formed via 3 ways handshake process. IIH packets need to carry the neighbor identifier in order to accomplish the 3 ways handshake process. In LAN IIH packets, TLV 6 is used to carry neighbor identifier information.
- In Point to Point IIH, TLV 6 doesn't exist. The absence of information on neighbors identifier in point-to-point IIHs as specified in the original hello format caused reliability issues in forming point-to-point adjacencies. This is due to 3 ways handshake can not be realized. Later, TLV Type 240 is proposed to address this problem by recording the remote ID in Point to Point IIH packet. Obviously, 3 ways handshake is much more reliable compared to 2 ways handshake. The situations will not happen where one end of an adjacency is up but the other end is not.
- In broadcast network, LAN IIH packet is used to form the adjacency relationship. There are 2 types of LAN IIH namely L1 LAN IIH ( with multicast MAC address 01-80-C2-00-00-14) and L2 LAN IIH ( with multicast MAC address 01-80-C2-00-00-15). Level-1 ISIS routers exchange L1 IIH with each other in order to form adjacency relationship. Level-2 ISIS routers exchange Level-2 IIH with each other to form adjacency relationship. Level-1-2 ISIS routers exchange Level-1 LAN IIH and Level-2 LAN IIH simultaneously to form adjacency relationship.

## ISIS LSDB Synchronization: DIS and Pseudonode Concept (1/4)



- DIS: Designated Intermediate System
- Function: Create and renew the pseudonode on broadcast network

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page52  HUAWEI

- In IS-IS protocol, a DIS will be elected on the broadcast network. The DIS generates a pseudonode to interact with other routers. A pseudonode is not a real router, but it occupies an extra LSP. Actually, the pseudonode LSP is created by the DIS.
- DIS: Designated IS in the broadcast network, similar to the DR in OSPF.
- Pseudonode: Pseudonode is not an actual router, but it occupies an extra LSP.

## ISIS LSDB Synchronization: DIS Election Process (2/4)

- In LAN, one of the routers is elected to be the DIS
  - DIS is elected based on the highest interface priority
  - The highest subnetwork point of attachment (SNPA) is used as tie-breaker when interfaces priority are the same.
    - In LAN, SNPA is referred to MAC address
    - In Frame Relay network, SNPA is referred to local data link connection identifier (DLCI)
  - The router with the highest System ID is elected to be the DIS when the SNPAs are the same
  - DIS can be preempted at any time.

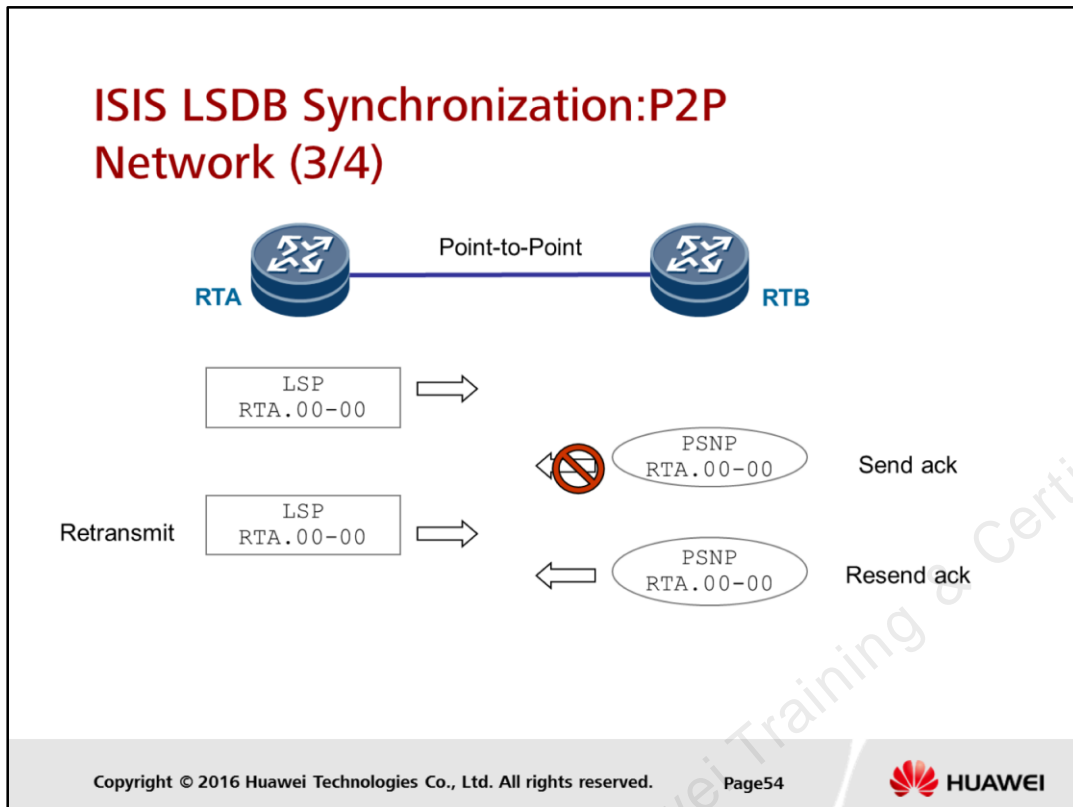
Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page53



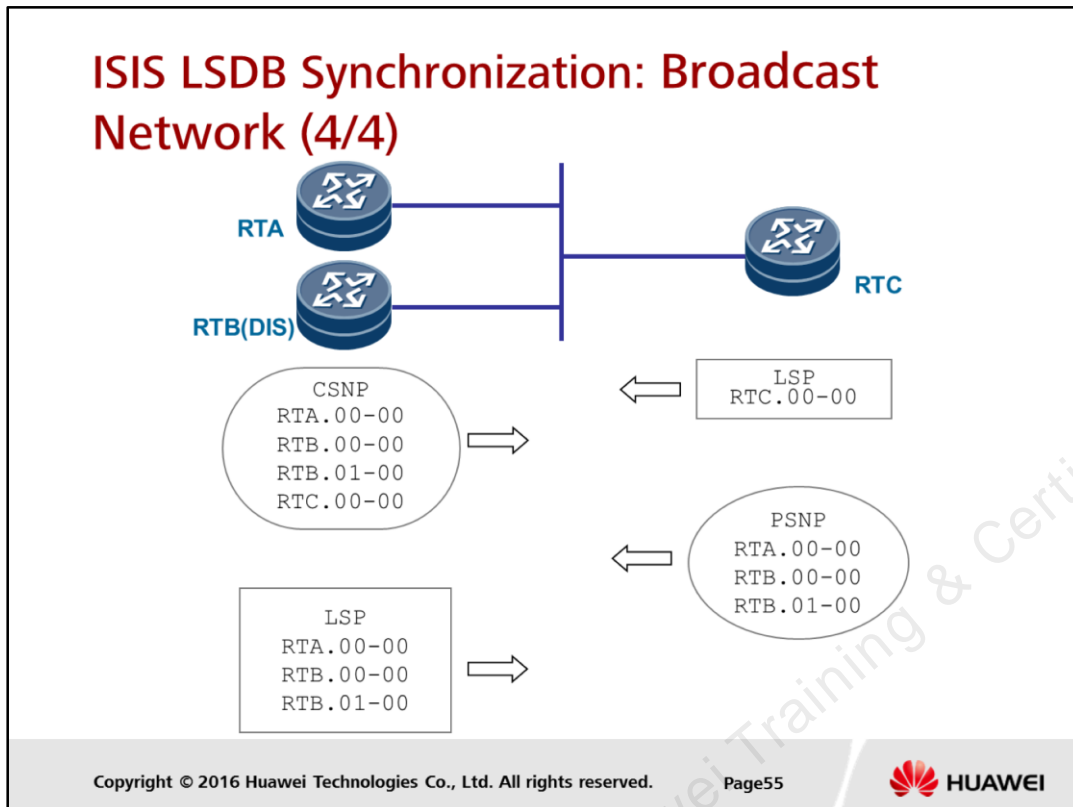
• In LAN, the DIS is elected based on the interface priority. The priority value ranges from 0 to 127 and its default interface priority is 64. The higher the value, the higher the priority is. Separate DISs are elected for level 1 and level 2 routing. The highest subnetwork point of attachment (SNPA) is used as tie-breaker when the interfaces priorities are the same. In LAN, SNPA is referred to MAC address. In frame relay network, SNPA is referred to local data link connection identifier (DLCI). If DLCI is used as SNPA in frame relay network, the SNPA might be the same for 2 ends of the link. In this scenario, the router with the highest System ID is elected to be the DIS. The interface of every IS-IS routers can be configured with either Level 1 or Level 2 priorities ranging from 0 to 127.

• Unlike OSPF, the IS-IS DIS can be preempted at any time by any eligible router connecting to the LAN. The newly elected DIS is responsible for purging the old pseudonode LSP and flooding the network with new LSP. The DIS concept is applicable to broadcast network only. The DIS election is not required on point-to-point network. Separate DISs are elected for level 1 and level 2 routing on broadcast network. No backup DIS is elected for either level-1 or level-2. This doesn't turn out to be a problem because DIS transmits hello packets 3 times faster compare to the other routers on the LAN. This allow for fast detection of DIS failure and immediate replacement.



- After the adjacency relationship has been formed on the point-to-point link, routers will exchange the CSNPs that describe the contents of the local link state database. Next, routers will compare the received CSNP with the content of local link state database. At the same time, the routers will send LSP to the neighbor that send out PSNP packet to request the transmission of current or missing LSPs. The neighbor that has received the LSP will send a reply by using a PSNP packet for acknowledgement purpose.
- If the acknowledgement is not received within a specified period, referred to as retransmission interval, the LSP is assumed lost during the transmission and is retransmitted. This retransmission process is repeated on the point-to-point link until the PSNP acknowledgement is received. This can ensure the integrity of the LSDB.
- CSNP contains only the summaries of the known LSP in the local database and this can simplify the process of database synchronization. Furthermore, a router will proactively send a copy of the LSP to its neighbor when it discovers that its neighbor doesn't have any of the LSP's in its local database. This can accelerate the database synchronization process. Please notice that CSNPs are sent only once when the IS-IS adjacency is initialized preceding the exchange of LSPs over the link on point-to-point link. After the adjacency has been formed, only PSNP packets are used to either request LSP or to acknowledge LSP. If there are any changes in the LSDB (for example the interface cost has been changed), router will directly send out a LSP packet regarding the changes to the neighbor. The neighbor will then check the sequence number of the received LSP to determine whether need renew the LSP. Then, the PSNP is used for acknowledgement. At the same time, retransmission mechanism is used to ensure the integrity of the database.





- In the diagram shown, RTC is the last router to join the broadcast network. RTA and RTB are already connected to the broadcast network and RTB is selected as the DIS. After establishing the adjacencies with RTA and RTB, RTC creates an LSP, RTC.00-00. RTB (DIS) advertises a CSNP by multicast over the link. RTC receives a copy of the CSNP, checks it against the local LSDB, and found 3 missing LSPs: RTA.00-00, RTB.00-00, and RTB.01-00. At this moment, RTC has only its own LSP, RTC.00-00. in its local link state database. RTC then sends out a PSNP to request the complete copies of RTA.00-00, RTB.00-00, and RTB.01-00. RTB floods RTA.00-00, RTB.00-00 and pseudonode LSP RTB.01-00 through multicast. And then RTC receives the copies.

- Please note that RTB is the DIS. Thus RTB will generate a router LSP and a pseudonode LSP.

- On broadcast network, no retransmission mechanism is adopted and the flooding over broadcast link is unreliable. IS-IS routers rely on periodic multicast of CSNPs from the DIS to ensure the database synchronization over the broadcast links.

 **Contents**

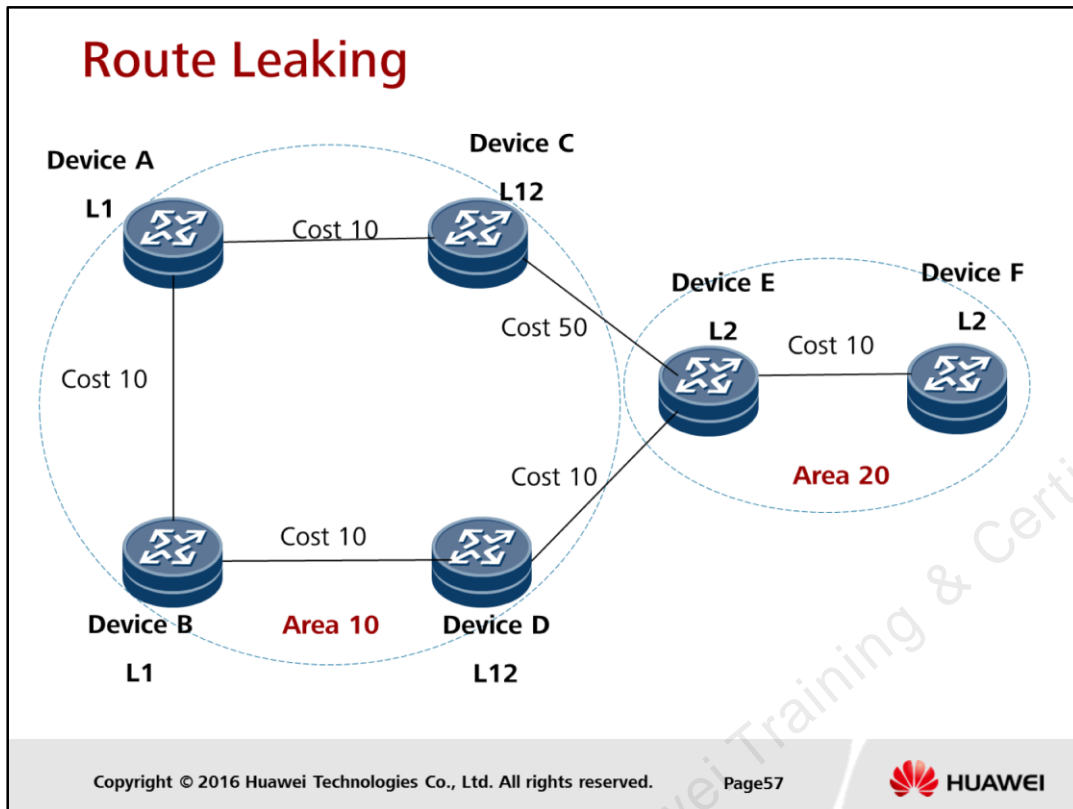
## 5.2. IS-IS

## 5.2.1 IS-IS Overview

## 5.2.2 Basic Concepts of IS-IS

## 5.2.3 IS-IS Route Calculations

**5.2.4 Route Leaking**



- When Level-1 and Level-2 areas both exist on an IS-IS network, Level-2 routers do not advertise the learned routing information about a Level-1 area and the backbone area to any other Level-1 area by default. Therefore, Level-1 routers do not know the routing information beyond the local area. As a result, the Level-1 routers cannot select the optimal routes to the destination beyond the local area.
- With route leaking, Level-1-2 routers can select routes using routing policies, or tags and advertise the selected routes of other Level-1 areas and the backbone area to the Level-1 area.
- If Device A sends a packet to Device F, the selected optimal route should be Device A -> Device B -> Device D -> Device E -> Device F because its cost is 40 ( $10 + 10 + 10 + 10 = 40$ ) which is less than that of Device A -> Device C -> Device E -> Device F ( $10 + 50 + 10 = 70$ ). However, if you check routes on Device A, you can find that the selected route is Device A -> Device C -> Device E -> Device F, which is not the optimal route from Device A to Device F.
- This is because Device A does not know the routes beyond the local area, and therefore, the packets sent by Device A to other network segments are sent through the default route generated by the nearest Level-1-2 device.
- In this case, you can enable route leaking on the Level-1-2 devices (Device C and Device D). Then, check the route and you can find that the selected route is Device A -> Device B -> Device D -> Device E -> Device F.

 **Contents****2. SDN Layer 3 Commonly Used Protocols**

2.1 OSPF Routing Protocol

2.2 ISIS Routing Protocol

**2.3 BGP Routing Protocol**

 **Contents**

## 5.3. BGP Routing Protocol

**5.3.1 BGP Overview**

## 5.3.2 BGP Basic Principles

## 5.3.3 BGP Route Control and Selection

## BGP Basic Characteristics

- Unlike an IGP, such as OSPF, BGP is an EGP which controls route advertisement and selects optimal routes between ASs rather than discovering or calculating routes.
- BGP uses TCP as the transport layer protocol, which enhances BGP reliability.
- BGP supports Classless Inter-Domain Routing (CIDR).
- When routes are updated, BGP transmits only the updated routes, which reduces bandwidth consumption during BGP route distribution.
- BGP is a distance-vector routing protocol, it's designed to prevent loops.
- BGP provides many routing policies to flexibly select and filter routes.

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved.

Page60



- BGP has the following characteristics:
  1. Unlike an Interior Gateway Protocol (IGP), such as Open Shortest Path First (OSPF) and Routing Information Protocol (RIP), BGP is an Exterior Gateway Protocol (EGP) which controls route advertisement and selects optimal routes between ASs rather than discovering or calculating routes.
  2. BGP uses Transport Control Protocol (TCP) as the transport layer protocol, which enhances BGP reliability.
    - BGP selects inter-AS routes, which poses high requirements on stability. Therefore, using TCP enhances BGP's stability.
    - BGP peers must be logically connected through TCP. The destination port number is 179 and the local port number is a random value.
  3. BGP supports Classless Inter-Domain Routing (CIDR).
  4. When routes are updated, BGP transmits only the updated routes, which reduces bandwidth consumption during BGP route distribution. Therefore, BGP is applicable to the Internet where a large number of routes are transmitted.
  5. BGP is a distance-vector routing protocol.
  6. BGP is designed to prevent loops.
    - Between ASs: BGP routes carry information about the ASs along the path. The routes that carry the local AS number are discarded to prevent inter-AS loops.

- Within an AS: BGP does not advertise routes learned in an AS to BGP peers in the AS to prevent intra-AS loops.
1. BGP provides many routing policies to flexibly select and filter routes.

Huawei Training & Certification Huawei Training & Certification

 **Contents**

## 5.3. BGP Routing Protocol

## 5.3.1 BGP Overview

**5.3.2 BGP Basic Principles**

## 5.3.3 BGP Route Control and Selection

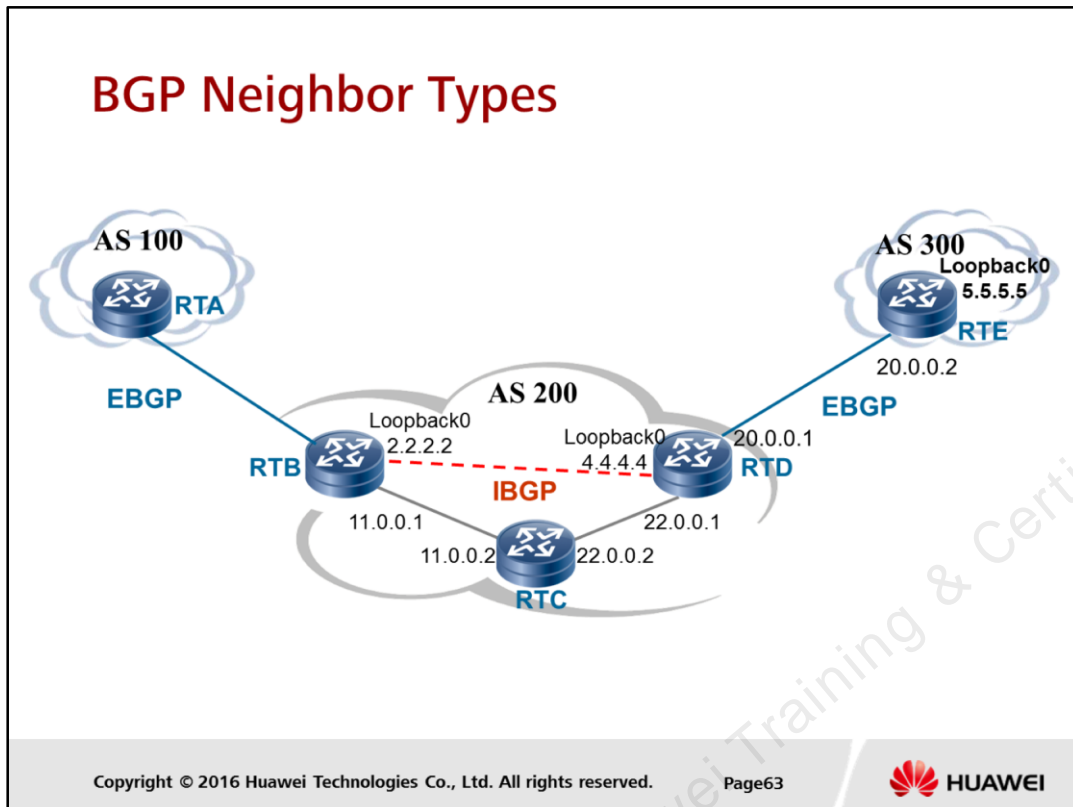


## BGP Messages

Packet Type	Description
Open	The first message sent after a TCP connection is set up is an Open message, which is used to set up BGP peer relationships.
Keep alive	Maintain peer relationships.
Update	Exchange routes between BGP peers.
Notification	When BGP detects an error, it sends a Notification message to its peer.
Route-refresh	Request that the peer resend all reachable routes.

Copyright © 2016 Huawei Technologies Co., Ltd. All rights reserved. Page62 HUAWEI

- BGP runs by sending five types of messages: Open, Update, Notification, Keepalive, and Route-refresh.
  - The descriptions for the BGP messages are listed as below:-
1. **Open:** The first message sent after a TCP connection is set up is an Open message, which is used to set up BGP peer relationships. After a peer receives an Open message and the peer negotiation is successful, the peer sends a Keepalive message to confirm and maintain the peer relationship. Then, peers can exchange Update, Notification, Keepalive, and Route-refresh messages.
  2. **Keepalive:** BGP periodically sends Keepalive messages to peers to maintain peer relationships.
  3. **Update:** This type of message is used to exchange routes between BGP peers.
    - An Update message can advertise multiple reachable routes with the same attributes. These route attributes are applicable to all destination addresses (expressed by IP prefixes) in the Network Layer Reachability Information (NLRI) field of the Update message.
    - An Update message can be used to delete multiple unreachable routes. Each route is identified by its destination address (using the IP prefix), which identifies the routes previously advertised between BGP speakers.
    - An Update message can be used only to delete routes. In this case, it does not need to carry the route attributes or NLRI. In addition, an Update message can be used only to advertise reachable routes. In this case, it does not need to carry information about the deleted routes.
  4. **Notification:** When BGP detects an error, it sends a Notification message to its peer. The BGP connection is then torn down immediately.
  5. **Route-refresh:** This type of message is used to request that the peer resend all reachable routes.
    1. If all BGP routers are enabled with the Route-refresh capability and the import policy of BGP changes, the local BGP router sends a Route-refresh message to its peers. After receiving the Route-refresh message, the peers resend their routing information to the local BGP router. In this manner, BGP routing tables are dynamically refreshed and new routing policies are used without tearing down BGP connections.



- BGP runs in the following two modes: IBGP (Internal BGP), EBGP (External BGP)
  - If two peers that exchange BGP messages belong to the same AS, they are Internal BGP (IBGP), such as RTB and RTD.
  - If two peers that exchange BGP messages belong to different AS, they are External BGP (EBGP), such as RTD and RTE.

## BGP Route Advertisement Principles

1. When multiple paths exist, BGP speaker only selects the best route from the BGP for its own use.
2. BGP speaker advertises only the routes used by itself to its peers.
3. For the routes obtained from EBGP, BGP speaker will advertise them to all its neighbors (including EBGP and IBGP)
4. For the route obtained from IBGP, the BGP speaker will not advertise them to its IBGP neighbors.
5. For the routes obtained from IBGP, whether the BGP speaker will advertise them to its EBGP neighbors depends on the synchronization state of IGP and BGP

- There are 5 BGP route advertisement principles, as per listed in the slide:-
- Under the normal circumstance, when there is more than one alternatives route to the same IP subnet, the BGP speaker will select the best route for its own use. The best route is the candidate for installation in the IP routing table. However, before a route can be installed, the router will check if there is there is any other routing protocol that has information about the same subnet. If the subnet is known via different sources, the router uses the route preference to determine which source is more trustworthy. The router will install the route with smaller route preference value. That is to say the router will select also the best route for its own use and the best route of BGP speaker might not be the best route for the router.
- BGP speaker advertises only the best routes used by itself to its peers. This means that it only advertises the BGP routes which are installed in the IP routing table to its peers. Once the best route (">") that has been selected by BGP is installed in the IP routing table, BGP will send Update message which consists of that best route entry to other BGP peer.
- For the routes obtained from EBGP, BGP speaker will advertise them to all its neighbors (including EBGP and IBGP)
- For the route obtained from IBGP, the BGP speaker will not advertise them to its IBGP neighbors. This rule is used to prevent routing loop inside an autonomous system. However, the enforcement of this rule introduces a new problem to the network, where some IBGP peers might not be able to get the Update messages from other IBGP peer. Thus, full-meshed IBGP peer connections should be established to solve this problem.
- For the routes obtained from IBGP, whether the BGP speaker will advertise them to its EBGP neighbors depends on the synchronization state of IGP and BGP. The concept of synchronization between BGP and IGP: BGP speaker will not advertise the routing information learnt from the IBGP peer to its EBGP peer unless all routers within the AS had learned about that route through the IGP. This is known as synchronization. If a router knows about these destinations via an IGP, it assumes that the route has already been propagated inside the AS, and internal reach-ability is assured. In Huawei, this synchronization feature is disabled by default.

 **Contents**

## 5.3. BGP

## 5.3.1 BGP Overview

## 5.3.2 BGP Basic Principles

**5.3.3 BGP Route Control and Selection**

## BGP Attributes – Classification (1/6)

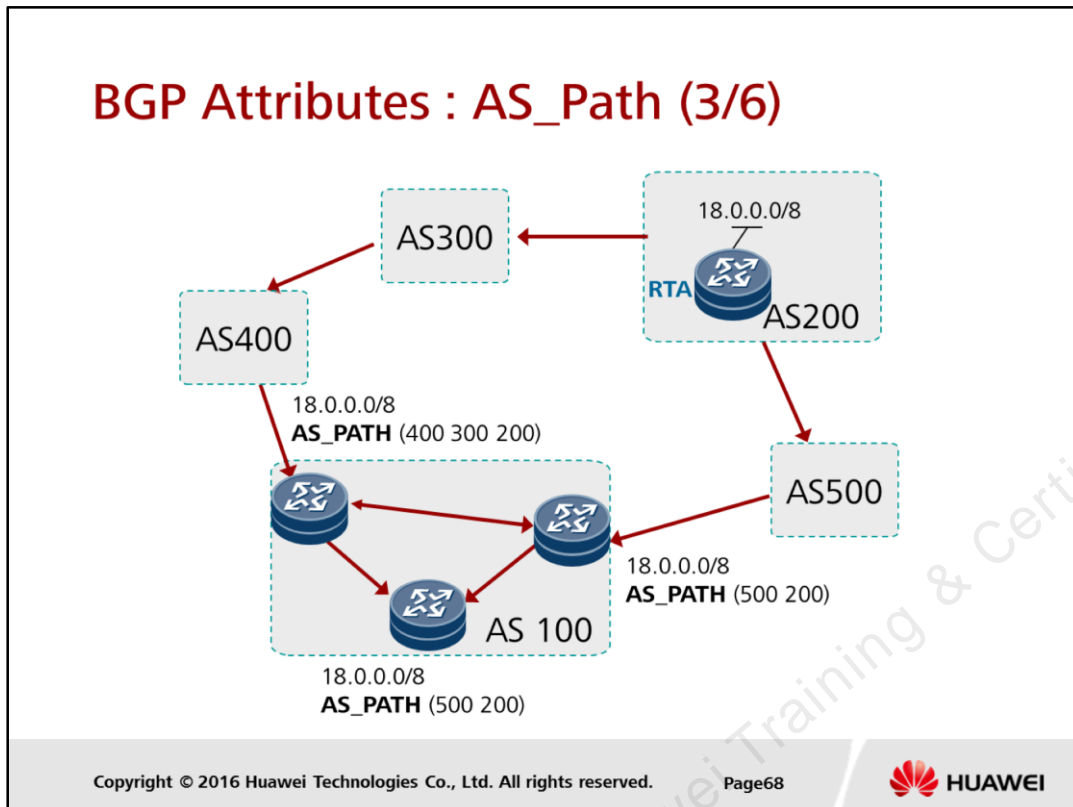
BGP Route Attributes	Description
Well known mandatory	This type of attribute can be identified by all BGP routers and must be carried in Update messages. Without this attribute, errors occur in the routing information.
Well known discretionary	This type of attribute can be identified by all BGP routers. This type of attribute is optional and, therefore, is not necessarily carried in Update messages.
Optional transitive	This indicates the transitive attribute between ASs. A BGP router may not recognize this attribute, but the router still receives it and advertises it to other peers .
Optional non-transitive	If a BGP router does not recognize this type of attribute, the router does not advertise it to other peers.

- Compared to other routing protocols, one of the most powerful strength of BGP is that BGP is rich with a lot of BGP attributes, which can be used to control, filter and select routes.
- The table above shows 4 categories of BGP attributes types.
- Below listed the examples of each BGP attributes:-
  - Well known mandatory: Origin, AS-PATH, Next-hop
  - Well known discretionary: Local-Preference, Atomic-Aggregate
  - Optional transitive: Aggregator, Community
  - Optional non-transitive: MED, Cluster-List, Originator-ID

## BGP Attributes : Origin (2/6)

- ORIGIN specifies the origin of the routing update. When BGP has multiple routes, it uses ORIGIN as one factor in determining the preferred route.
- Three Origin attributes are listed as below:-
  - IGP
  - EGP
  - Incomplete
- The precedence of the Origin attributes in descending order is IGP > EGP > Incomplete

- Origin attribute specify the origin of the BGP path information. In fact, it is the methods for BGP speaker to generate the BGP route. BGP considers three types of origins:
  - IGP: The route with origin IGP is marked with "i" in BGP routing table (by using the "display bgp routing-table" command).The origins are IGP for the routes internal to the AS and are advertised via the network command. This method is also called as semi dynamic redistribution of BGP information. The network advertised via the network command is dynamically discovered and calculated by IGP (including static route). Some of the routing information is selectively redistributed into the BGP system via network command. That's why it is called as "semi dynamic".
  - EGP: The route with origin EGP is marked with "E" in BGP routing table. The origin "EGP" was used when the Internet when the routes are redistributed from EGP into the BGP routing table. It is used when the Internet was migrating from EGP to BGP. It is rather difficult to encounter the route with origin EGP in the real network. This is because EGP protocol is basically obsolete and not used anymore.
  - Incomplete: The route with origin Incomplete is marked with "?" in BGP routing table. The route with origin incomplete is learned by some other means. It means that the information for determining the origin of the route is incomplete. Routes that BGP learnt through redistribution from IGP or static route carry the incomplete origin attribute. Injecting the IGP routes into BGP dynamically or semi dynamically is based on the dependency of the BGP routes on the IGP routes.
- The precedence order of the 3 origin values are IGP>EGP>INCOMPLETE.
- These 3 origin values are used to control the selection of BGP routes.



- The AS-Path attribute records all ASs through which a route passes from the local end to the destination in distance-vector (DV) order.
- When a BGP speaker advertises a local route:
  - When advertising the route beyond the local AS, the BGP speaker adds the local AS number to the AS\_Path list and then advertises it to the neighboring routers through Update messages.
  - When advertising the route within the local AS, the BGP speaker creates an empty AS\_Path list in an Update message.
- When a BGP speaker advertises a route learned from the Update messages of another BGP speaker:
  - When advertising the route beyond the local AS, the BGP speaker adds the local AS number to the left of the AS\_Path list. From the AS\_Path attribute, the BGP router that receives the route learns the ASs through which the route passes to the destination. The number of the AS that is nearest to the local AS is placed on the left of the list, while other AS numbers are listed in sequence.
  - When advertising the route within the local AS, the BGP speaker does not change the AS\_Path attribute.

## BGP Attributes : Next Hop (4/6)

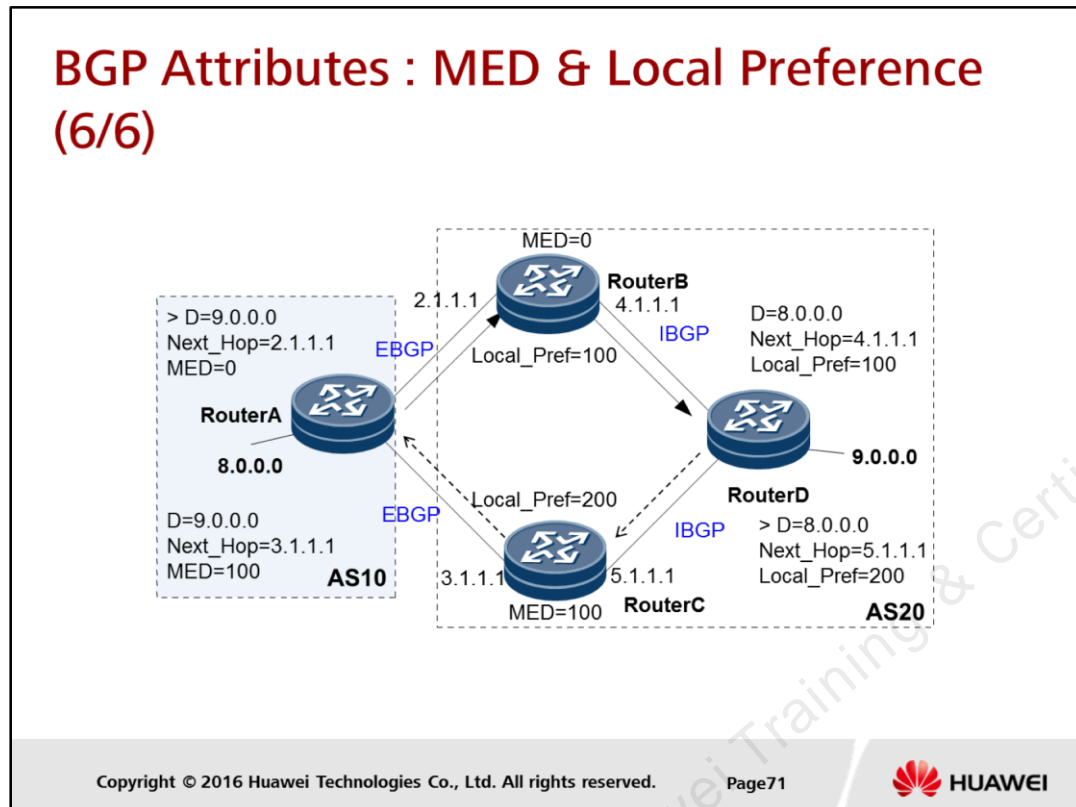
- In most cases, the Next\_Hop attribute in BGP complies with the following rules:
  - When advertising a route to an EBGP peer, a BGP speaker sets the Next\_Hop of the route to the address of the local interface through which the BGP peer relationship is established.
  - When advertising a locally generated route to an IBGP peer, a BGP speaker sets the Next\_Hop of the route to the address of the local interface through which the BGP peer relationship is established.
  - When advertising a route learned from an EBGP peer to an IBGP peer, the BGP speaker does not change the Next\_Hop of the route.



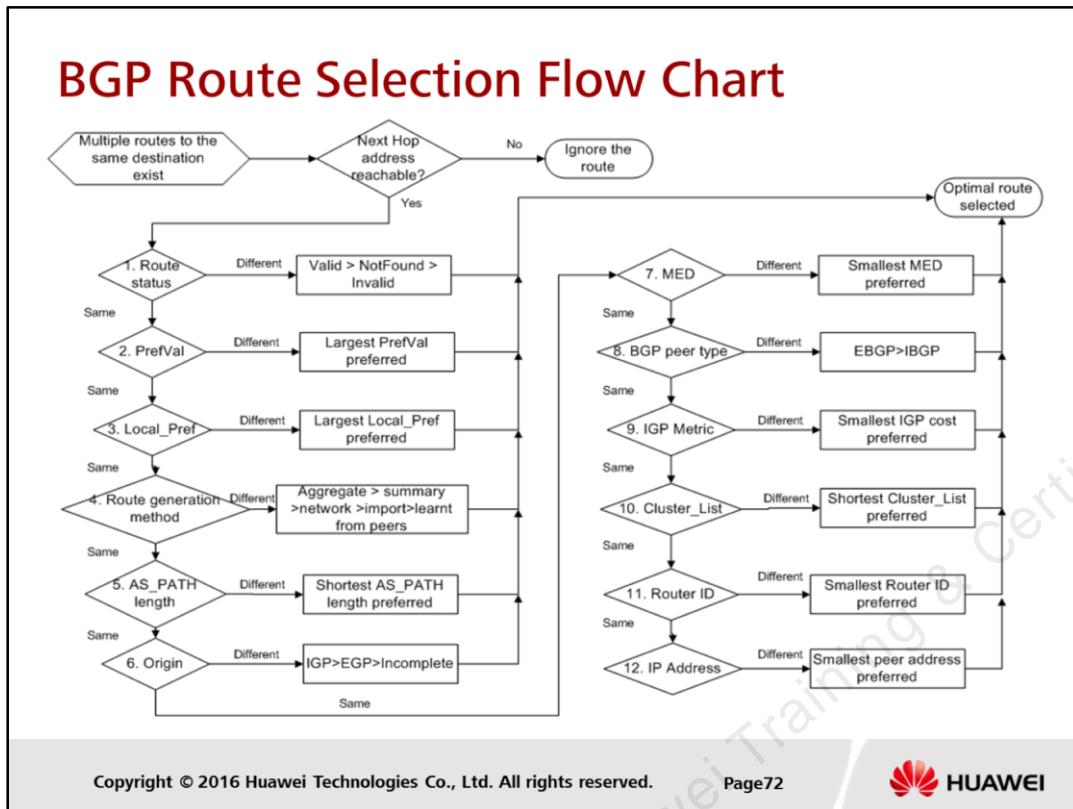
## BGP Attributes : MED & Local Preference (5/6)

- The MED is used to determine the optimal route when traffic enters an AS.
- The Local Preference attribute is used to determine the optimal route when traffic leaves an AS.

Attribute	Application Scope	Value	Impact on Traffic
MED	Between two neighboring ASs	The smaller the value, the better	determine the optimal route when traffic enters an AS
Local_Pref	Within an AS	The larger the value, the better.	determine the optimal route when traffic leaves an AS



- Normally, BGP compares only the MED values of the routes from the same AS. If router A in AS 10 receives the same route from AS 20 and another AS, router A does not compare the MED values of the routes.
- In the implementation on VRP5, the compare-different-as-med command can be run to force BGP to compare the MED values of the routes learnt from different ASs.



When multiple routes to the same destination are available, BGP selects routes based on the following rules:

1. Prefers routes in descending order of Valid, Not Found, and Invalid after BGP origin AS validation states are applied to route selection in a scenario where the device is connected to an RPKI server.
2. Prefers the route with the largest PreVal value. PrefVal is Huawei-specific. It is valid only on the device where it is configured.
3. Prefers the route with the highest Local\_Pref. If a route does not carry Local\_Pref, the default value 100 takes effect. To change the value, run the **default local-preference** command.
4. Prefers a locally originated route to a route learned from a peer. Locally originated routes include routes imported using the **network** or **import-route** command, as well as manually and automatically summarized routes. Prefers a summarized route over a non-summarized route.
5. Prefers a route obtained using the **aggregate** command over a route obtained using the **summary automatic** command.
6. Prefers a route imported using the **network** command over a route imported using the **import-route** command.



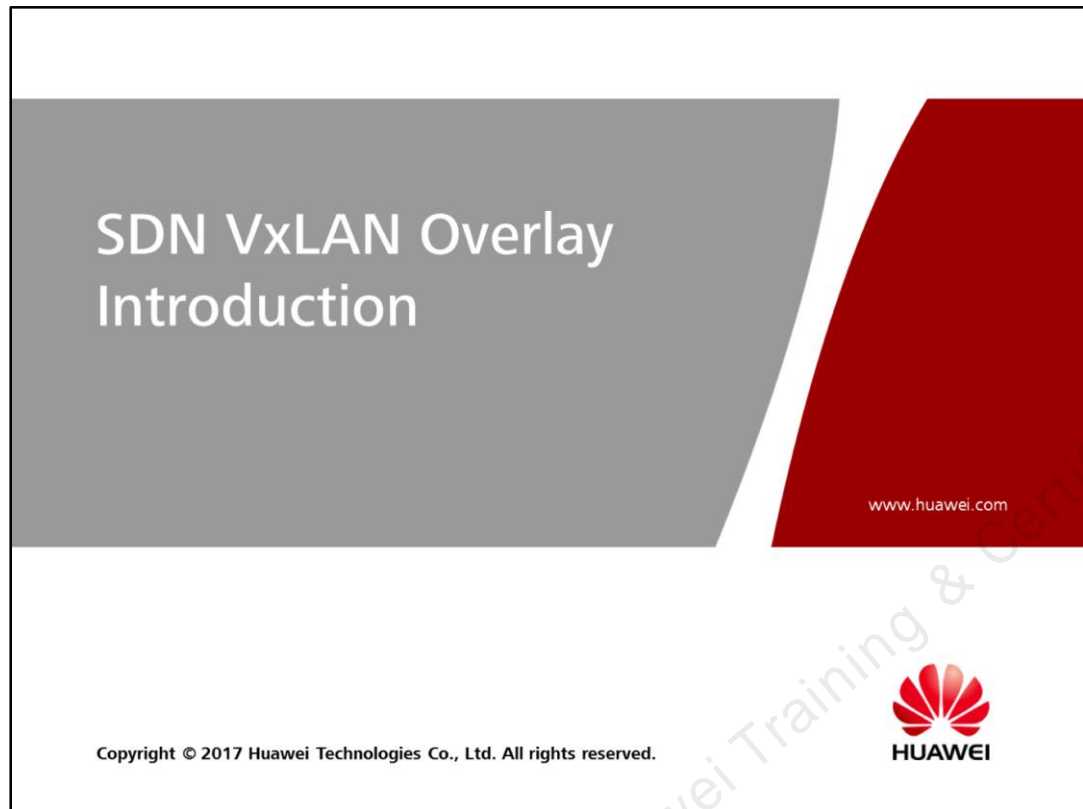
## Summary

- There are 2 different methods to establish the SDN control channels between the controller and forwarder, which are layer 2 networking and layer 3 networking scenarios.
- For layer 3 networking scenarios, traditional routing protocols such as OSPF, ISIS or BGP can be deployed.
- Basic concepts of OSPF, ISIS and BGP were discussed and included in this chapter in order to build up L3 networking control channel between SDN controller and forwarder

**Thank you**

[www.huawei.com](http://www.huawei.com)


Huawei Training & Certification Huawei Training & Certification



SDN VxLAN Overlay  
Introduction

[www.huawei.com](http://www.huawei.com)

Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.



HUAWEI

The slide features a white background with a large grey shape on the left and a red shape on the right. The title 'SDN VxLAN Overlay Introduction' is centered in the grey area. The website 'www.huawei.com' is in the red area. The copyright notice and Huawei logo are at the bottom.



## Foreword

- VxLAN is a very important overlay technology used throughout the whole AC-DCN network; thus, understanding the concepts of VxLAN and its applications and configuration in AC-DCN network is crucial before we go into the end-to-end service deployment through the AC-DCN underlay deployment.



## Objectives

- Upon completion of this course, you will be able to:
  - Understand why VXLAN is needed in DCN
  - Understand VxLAN basic concepts
  - Understand VxLAN application in SDN AC-DCN network
  - Understand VxLAN configuration in SDN AC-DCN network





## Contents

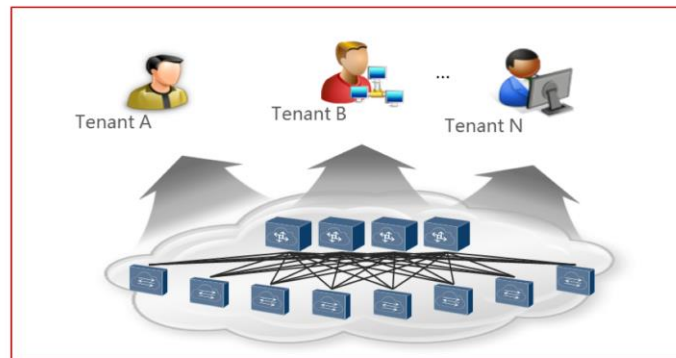
1. VxLAN Overlay Overview
2. VxLAN Basic Concepts
3. VxLAN Applications in SDN AC-DCN Cloud Fabric Network
4. VxLAN Configuration Examples in SDN AC-DCN Cloud Fabric Network



## Contents

1. **VxLAN Overlay Overview**
2. VxLAN Basic Concepts
3. VxLAN Applications in SDN AC-DCN Cloud Fabric Network
4. VxLAN Configuration Examples in SDN AC-DCN Cloud Fabric Network

## Why VxLAN?



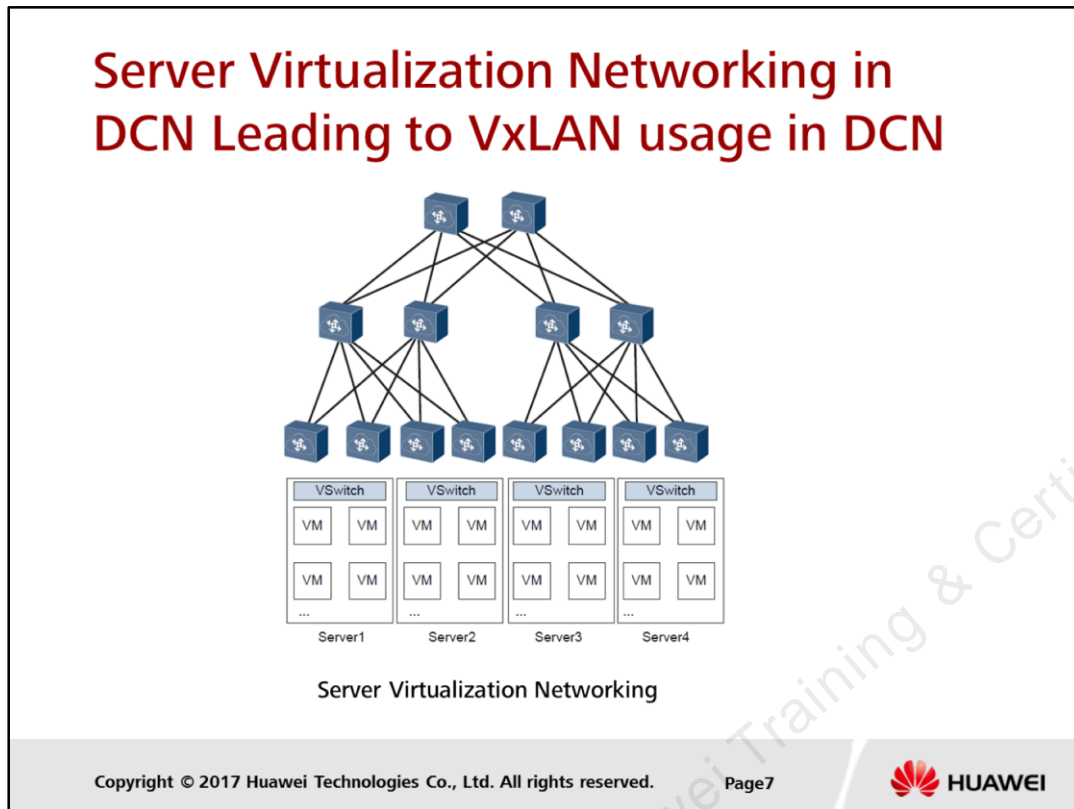
- VxLAN is a technology which is much needed in the DCN network nowadays prior to the demand of data center multi-tenants scenarios. The traditional VLAN technology which can only support up to maximum 4096 VLANs is definitely not sufficient in identifying user on Layer network in DCN. Thus, VxLAN, serves as a overlay tunnel technology is implemented to solve this requirement.

Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

Page6



- The VLAN tag field, as defined in IEEE 802.1Q, has only 12 bits, and can only identify a maximum of 4096 VLANs, making it insufficient for identifying users on large Layer 2 networks;
- VXLAN uses a VXLAN network identifier (VNI) field similar to the VLAN ID field defined in IEEE 802.1Q. The VNI field has 24 bits and can identify a maximum of 16M VXLAN segments.



- On the network shown in the diagram above, one server is virtualized into multiple virtual machines (VMs), each of which acts as a host. However, the exponential increase in the number of hosts leads to the following problems on a virtual network:

- The number of VMs is limited by network performance.**

- On a large Layer 2 network, data packets are forwarded based on MAC address entries. Therefore, the number of VMs supported on the network depends on the MAC address table size.

- Network isolation capabilities are limited.**

- Most networks use VLANs or virtual private networks (VPNs) for network isolation. However, these two network isolation technologies have the following limitations on large scale virtualized networks:
  - The VLAN tag field, as defined in IEEE 802.1Q, has only 12 bits, and can only identify a maximum of 4096 VLANs, making it insufficient for identifying users on large Layer 2 networks.
  - VLANs or VPNs cannot support dynamic network adjustment on traditional Layer 2 networks.

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

- After VMs are started, they may need to be migrated from one server to another due to server resource problems (for example, CPU overload or insufficient memory). To ensure uninterrupted services during VM migration, the IP and MAC addresses of VMs must remain unchanged. To meet this requirement, the service network must be a Layer 2 network that provides multipath redundancy and reliability.

## VxLAN Benefits

- The implementation of VxLAN helps to solve the problems on Layer 2 DCN, as shown below:-
  - VM scale limitations imposed by network performance
  - Limited network isolation capabilities
  - VM migration scope limitations imposed by network architecture
- The advantages of VxLAN is listed below:-
  - Supports maximum 16M VxLAN segments contributing to large number of tenants allowed in DCN
  - Reduces the MAC address learnt in network devices
  - Extends L2 networks using MAC in UDP encapsulation and decouples physical with virtual network; thus simplifies the network management.

- VXLAN addresses the above problems on large Layer 2 networks as follows:

### 1. VM scale limitations imposed by network performance

- VXLAN encapsulates data packets sent from VMs into UDP packets and encapsulates IP and MAC addresses used on the physical network into outer headers. The network is only aware of the encapsulated parameters. This greatly reduces the number of MAC addresses required on large Layer 2 networks.

### 2. Limited network isolation capabilities

- VXLAN uses a VXLAN network identifier (VNI) field similar to the VLAN ID field defined in IEEE 802.1Q. The VNI field has 24 bits and can identify a maximum of 16M  $[(2^{24}-1)/1024^2]$  VXLAN segments.

### 3. VM migration scope limitations imposed by network architecture

- When VXLAN is used to construct a large Layer 2 network, VM IP and MAC addresses can remain unchanged after VM migration.

- The advantages of VxLAN are listed below:-

1. Supports a maximum of 16M VxLAN segments with 24-bit VNIs, so a data center can accommodate a large number of tenants.
2. Reduces the number of MAC addresses that network devices need to learn and enhances network performance because only devices at the edge of the VxLAN network need to identify VM MAC addresses.
3. Extends Layer 2 networks using MAC-in-UDP encapsulation and decouples physical and virtual networks. Tenants can plan their own virtual networks, without being limited by the physical network IP addresses or broadcast domains. This greatly simplifies network management.



## Contents

1. VxLAN Overlay Overview
2. **VxLAN Basic Concepts**
3. VxLAN Applications in SDN AC-DCN Cloud Fabric Network
4. VxLAN Configuration Examples in SDN AC-DCN Cloud Fabric Network



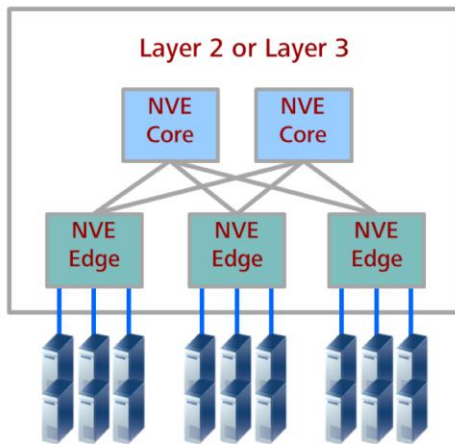
## Contents

### 2. VxLAN Basic Concepts

#### **2.1 VxLAN Basic Principles**

#### 2.2 VxLAN Forwarding Models

## NVO3 and VxLAN Overview



### NVO3

NVO3 (Network Virtualization over Layer 3) is a general term for IP overlay network virtualization technology based on Layer 3. The famous NVO3 virtualization technology examples include, VxLAN, NVGRE and STT.

### VXLAN

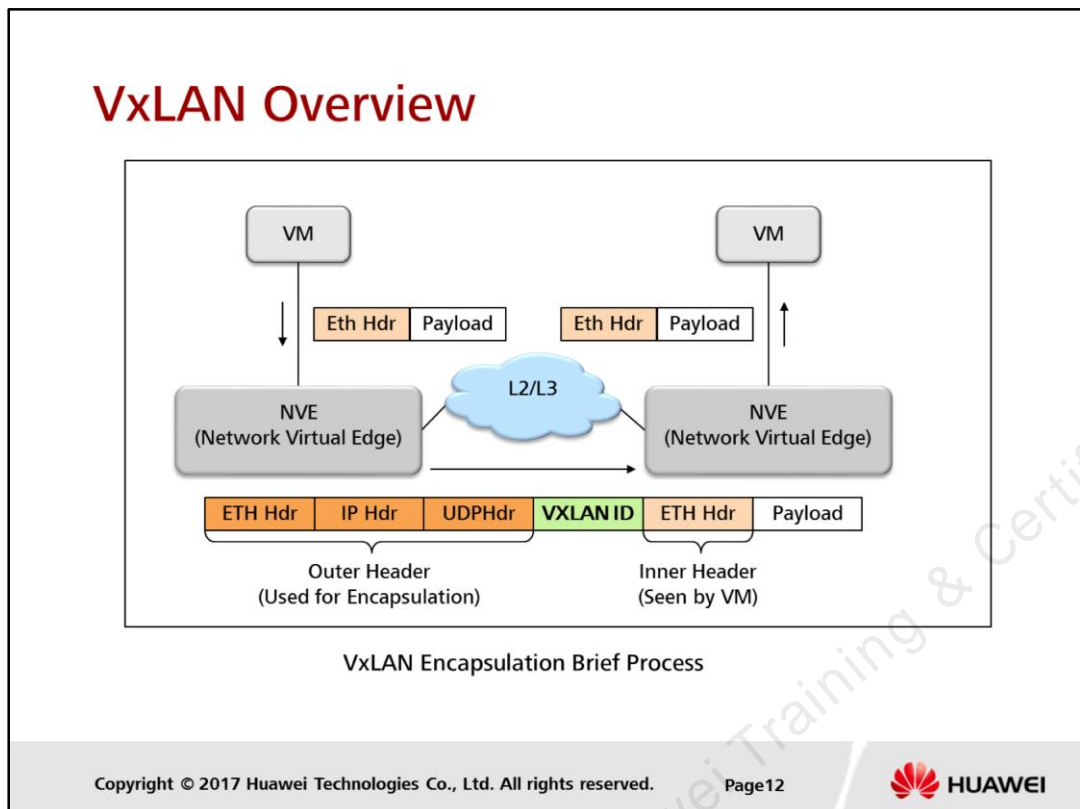
VXLAN is a Network Virtualization over Layer 3 (NVo3) technology that uses MAC in User Datagram Protocol (MAC-in-UDP) to encapsulate packets.

Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

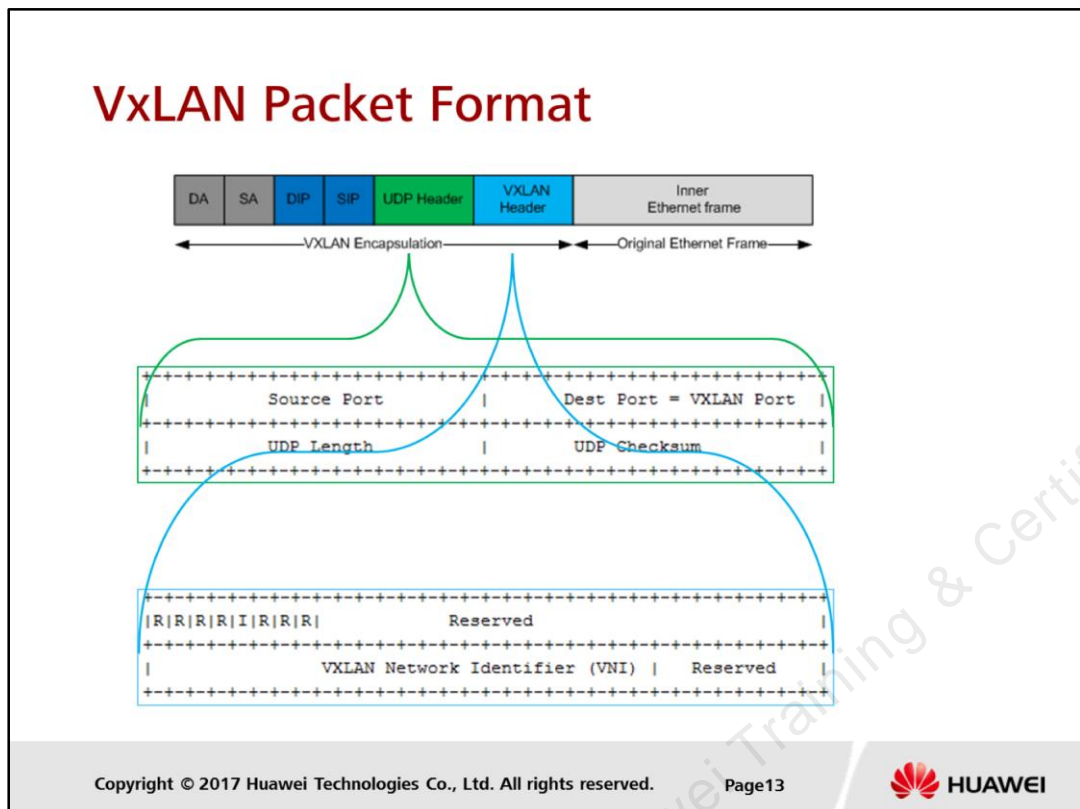


- NVO3 (Network Virtualization over Layer 3) is a general term for IP overlay network virtualization technology based on Layer 3. The famous NVO3 virtualization technology examples include, VxLAN (Virtual extensible Local Area Network) , NVGRE (Network Virtualization using Generic Routing Encapsulation) and STT (Stateless Transport Tunneling Protocol).
- Devices running NVO3 is called NVE (Network Virtualization Edge), which is located at the edge of the overlay network to realize L2 and L3 virtualization function.
- VxLAN is considered as the most widely used NVO3 technologies nowadays.



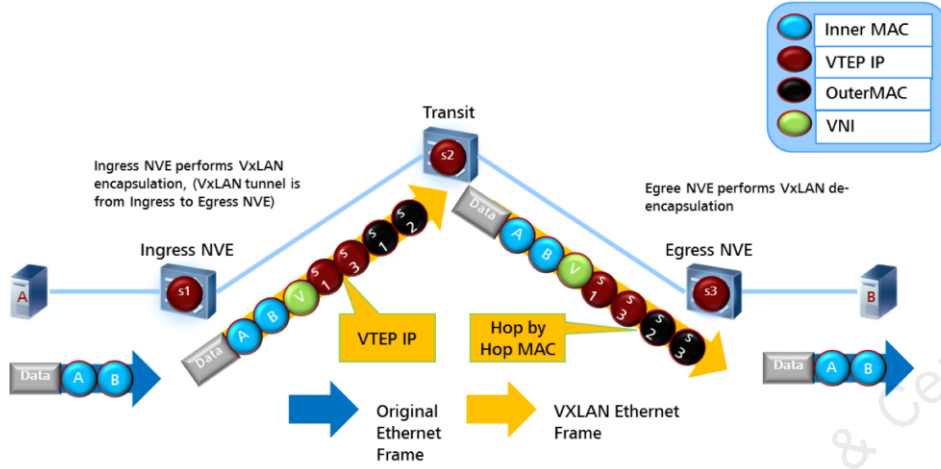


- VxLAN (Virtual Extensible LAN) is a Network Virtualization over Layer 3 (NVo3) technology that uses MAC in User Datagram Protocol (MAC-in-UDP) to encapsulate packets.
- In SDN Cloud Fabric solution, VxLAN technology is realized through AC in order to build the overlay network over L3 fabric. It uses MAC over UDP encapsulation to achieve the requirements of multi-tenants scenarios in data center network virtualization solution.



- The explanation of the VxLAN packet format is as below:-
- **VXLAN header**
  - Flags (8 bits): The value is 00001000. (R flag must be set to 0 and I flag must be set to 1)
  - VNI (24 bits): used to identify a VXLAN segment.
  - Reserved fields (24 bits and 8 bits): must be set to 0.
- **Outer UDP header**
  - The destination UDP port number is 4789. The source port number is the hash value calculated using parameters in the inner Ethernet frame header.
- **Outer IP header**
  - In the outer IP header, the source IP address is the IP address of the VTEP where the sender VM resides; the destination IP address is the IP address of the VTEP where the destination VM resides.
- **Outer Ethernet header**
  - SA: specifies the MAC address of the VTEP where the sender VM resides.
  - DA: specifies the next-hop MAC address in the routing table of the VTEP where the destination VM resides.
  - VLAN Type: This field is optional. The value of this field is 0x8100 when the packet has a VLAN tag.
  - Ethernet Type: specifies the type of the Ethernet frame. The value of this field is 0x0800 when the packet type is IP.

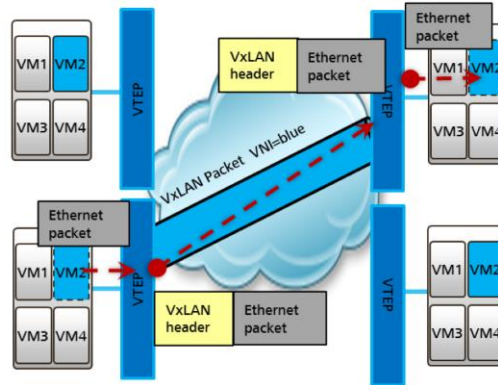
## VxLAN Packet Encapsulation Process



The Layer 2 Ethernet frame can be sent through IP network transparently on top of L3 IP network; VxLAN network is similar to Bridge Fabric for end terminal.

## VxLAN Concepts - VTEP

- **VTEP** -A VXLAN tunnel endpoint that encapsulates and decapsulates VXLAN packets. It is represented by an NVE. VTEP can be realized on physical switches (hardware overlay scenario) or logical vSwitches (software overlay scenario).



Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

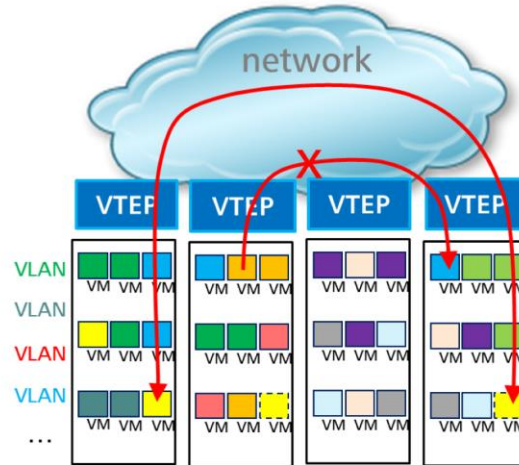
Page15



- VTEP stands for Virtual Tunnel End Point.
- VXLAN tunnel endpoints that are deployed on NVE nodes and responsible for VXLAN packet encapsulation and decapsulation. VTEPs are connected to the physical network and assigned IP addresses (VTEP IP) of the physical network. VTEP IP addresses are independent of the virtual network. A local VTEP IP address and a remote VTEP IP address identify a VXLAN tunnel.
- There are 2 types of configuration for VTEP configuration, as per listed below:-
  - **VTEP on physical TOR (Hardware overlay)**
    - Advantage : High performance, line speed forwarding capability
    - Disadvantage : limited by chip manufacturers; TOR switch needs virtual perceptions.
  - **VTEP on vSwitch (Software overlay)**
    - Advantage : TOR switch does not need to virtual perception, achieve simply
    - Disadvantage: Consume the host hardware resources, impact the host performance

## VxLAN Concepts - VNI

- VNI- A VXLAN segment identifier similar to a VLAN ID. VMs on different VXLAN segments cannot communicate directly at Layer 2.

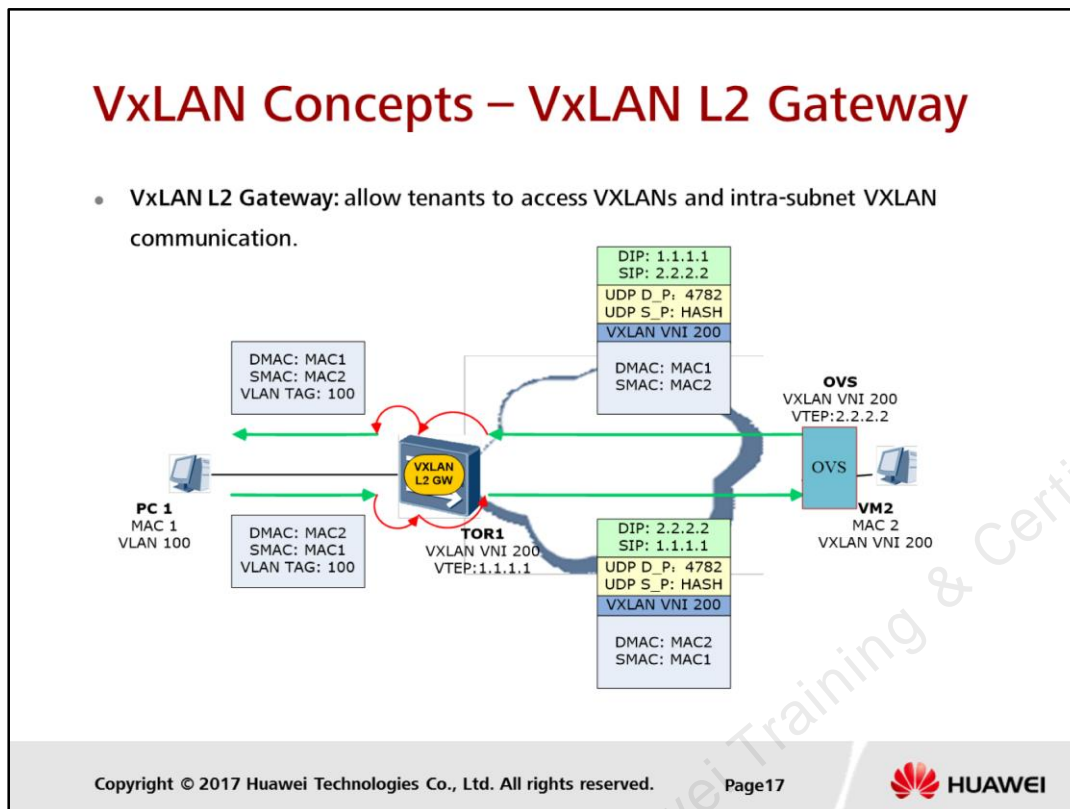


Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

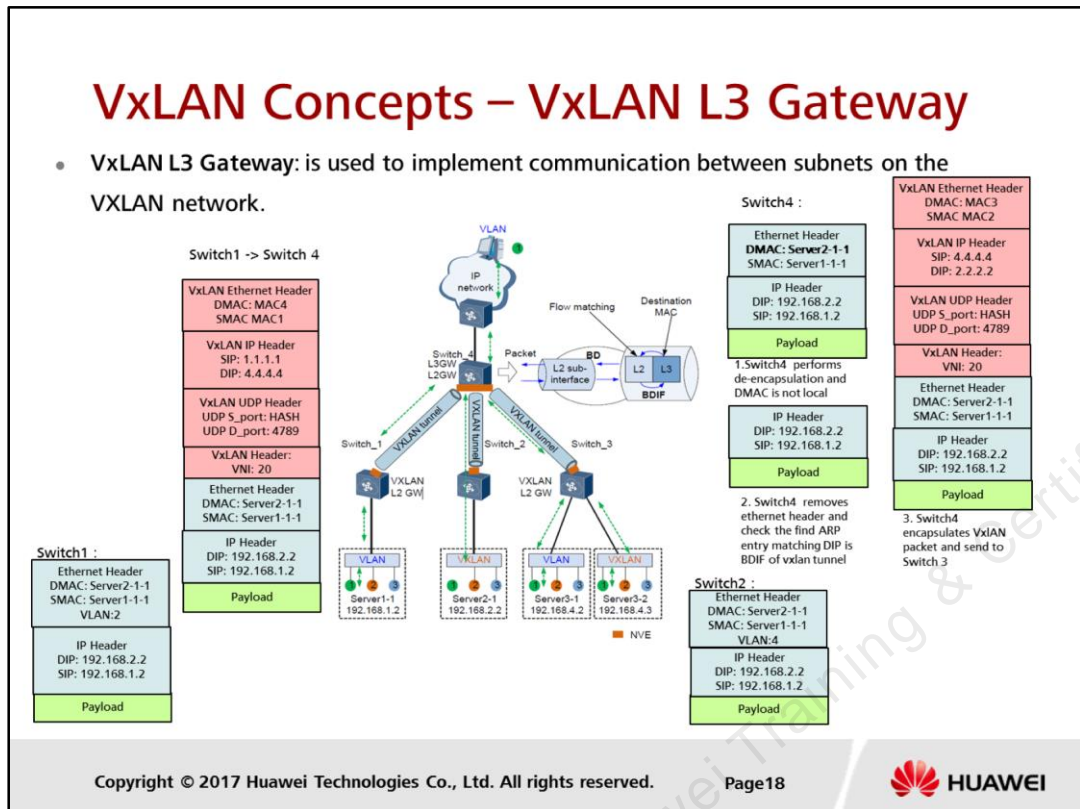
Page16



- VNI is VXLAN network identifier that identifies a VXLAN segment.
- ◆ VNI segment : 24 bits;
- ◆ Within the same VM VNI can communicate directly;
- ◆ Different VM VNI cannot communicate directly, must use **Layer 3 Gateway** to communicate.

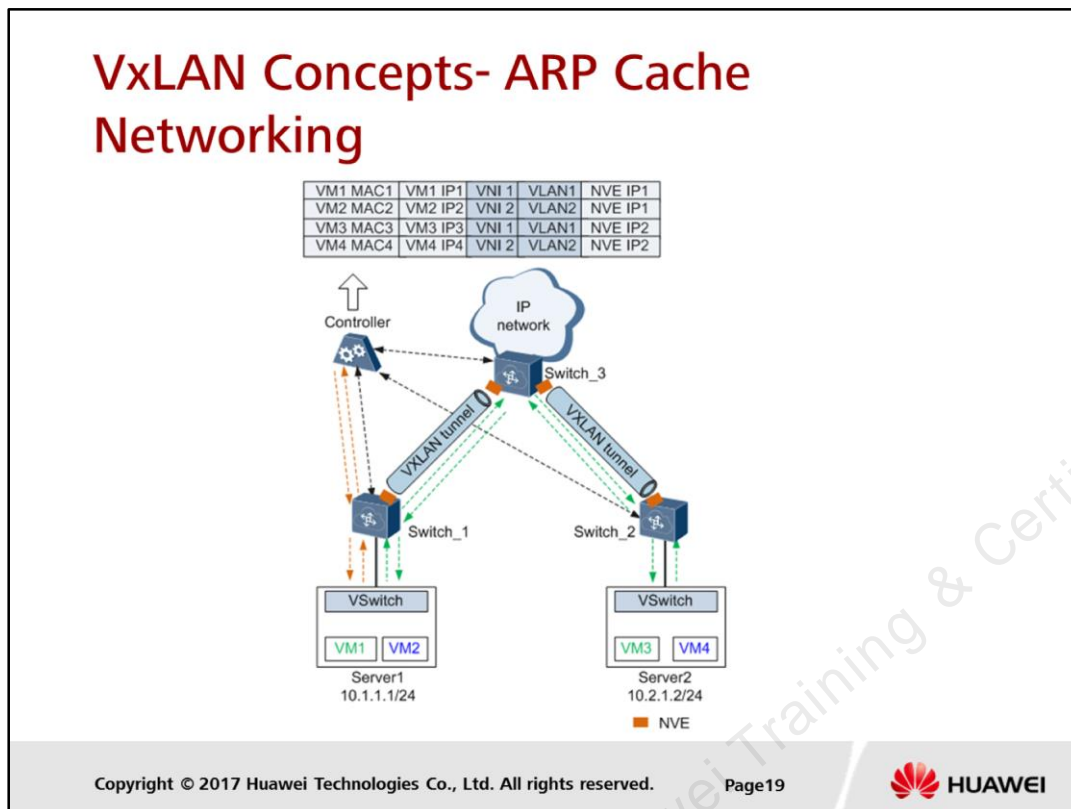


- **Layer 2 gateway:**
- After a VxLAN Layer 2 gateway receives user packets, it forwards the packets in different processes based on the packets' destination MAC address type:
  - If the MAC address is a broadcast, unknown unicast, and multicast (BUM) address, the Layer 2 gateway follows the [BUM packet forwarding process](#).
  - If the MAC address is a unicast address, the Layer 2 gateway follows the [unicast packet forwarding process](#).



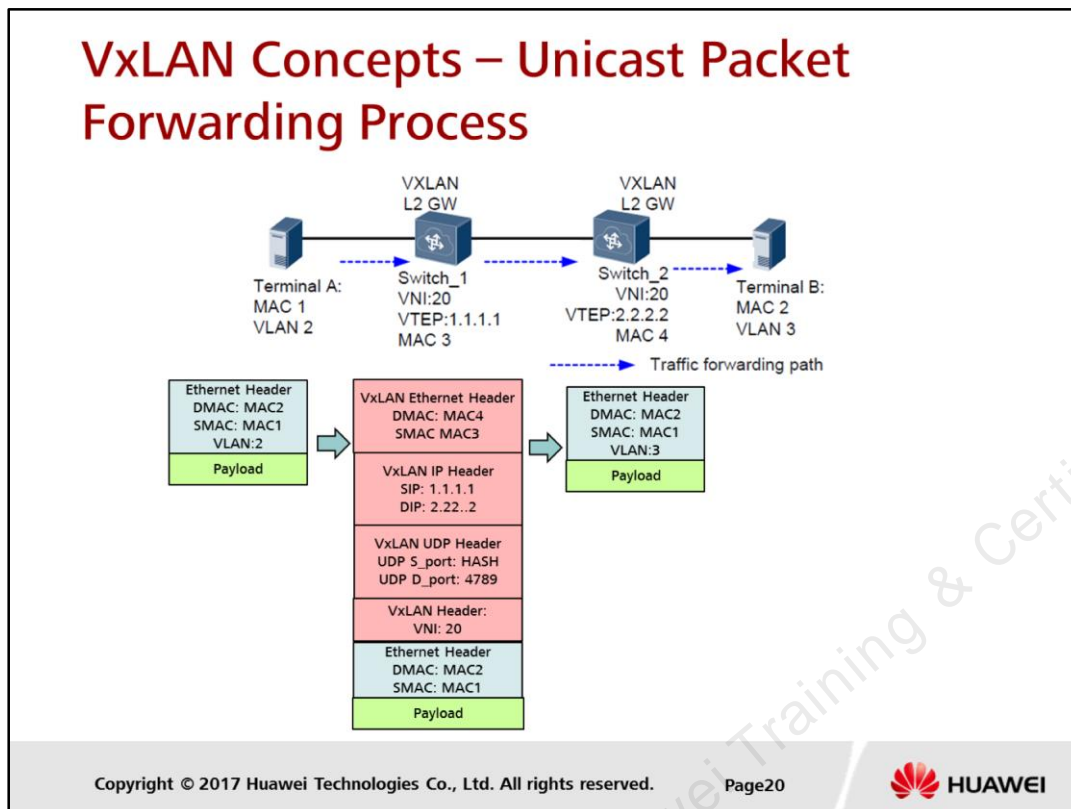
- VxLAN communication between different network segment, and the communication between VxLAN network and non VxLAN network requires IP routing. Thus, Bridge Domain (BD) needs to be established on L3 gateway; VNI is mapped to a BD in 1:1 ratio; BDIF interface is configured based on different BD to allow L3 communication. BDIF is similar to Vlanif in VLAN concept.
1. After Switch\_4 (VxLAN Layer 2 gateway) receives a VxLAN packet, it decapsulates the packet and checks whether the destination MAC address in the inner packet is the gateway MAC address.
    - If so, Switch\_4 forwards the packet to the Layer 3 gateway on the destination network segment. Go to Step 2.
    - If not, Switch\_4 searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain.
  2. Switch\_4 functions as a VxLAN Layer 3 gateway to remove the Ethernet header of the inner packet and parse the destination IP address. Switch\_4 searches for the ARP entry matching the destination IP address and checks the destination MAC address, VxLAN tunnel's outbound interface, and VNI.
    - If the VxLAN tunnel's outbound interface and VNI are not found, Switch\_4 forwards the packets at Layer 3.
    - If the VxLAN tunnel's outbound interface and VNI are found, go to Step 3.
  3. Switch\_4 functions as a VxLAN Layer 2 gateway to encapsulate the VxLAN packet again by adding the gateway's MAC address as the source MAC address in the Ethernet header.



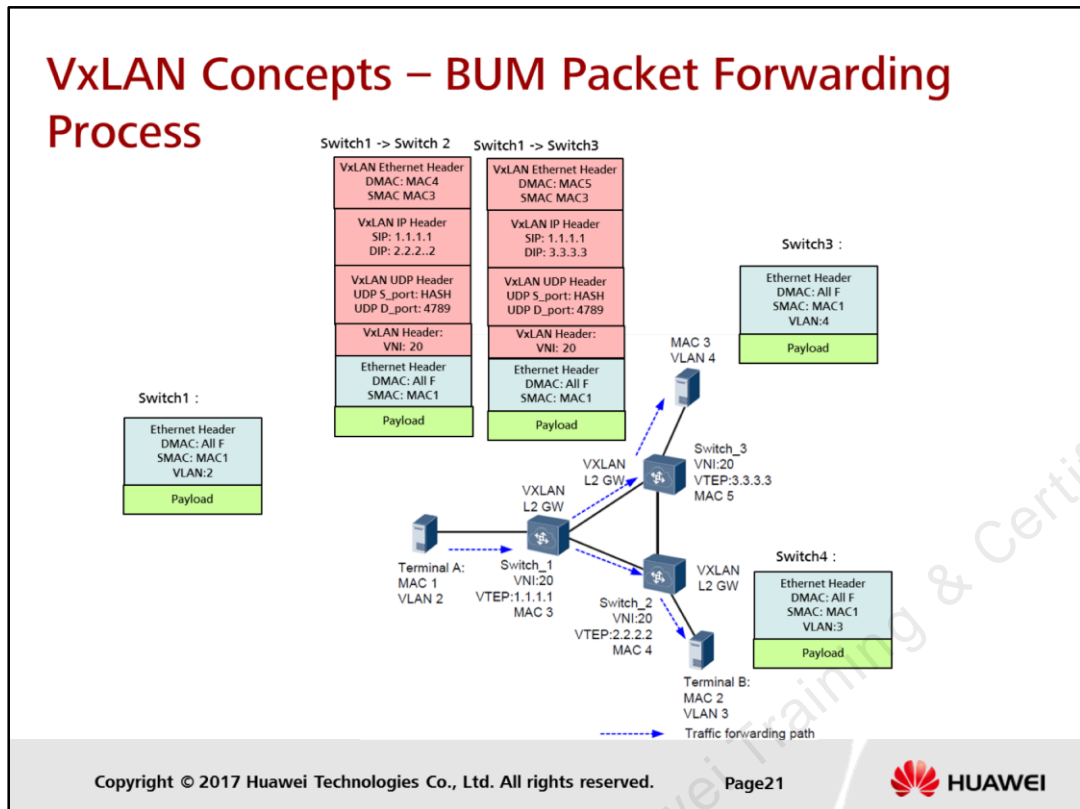


- Traditionally, a terminal needs to send a broadcast ARP request message before it communicates with another terminal for the first time. For example, on the network shown in Figure above, VM1 needs to send an ARP request message to VM3 when it needs to communicate with VM3 for the first time. The ARP request message is broadcast on the Layer 2 network. After receiving the ARP request message, VM3 sends a unicast ARP reply message to VM1.
- To prevent broadcast storms caused by broadcast ARP request messages, ARP cache can be enabled on the controller, as shown in the figure above. After that, the following process occurs when VM1 sends an ARP request message to request VM3's MAC address:
  - VM1 sends an ARP request message, with the source MAC address MAC1, source IP address IP1, destination MAC address FF-FF-FF, and destination IP address IP3.
  - After receiving the ARP request message, the NVE1 sends the message to the controller through an OpenFlow channel.
  - The controller searches the user information database based on IP3 and obtains the MAC address (MAC3) of VM3.
  - The controller sends an ARP reply message to the NVE1 through the OpenFlow channel.
  - After receiving the ARP reply message, the NVE1 sends the message to VM1 through the outbound interface specified by the controller, which is the inbound interface that receives the ARP request message

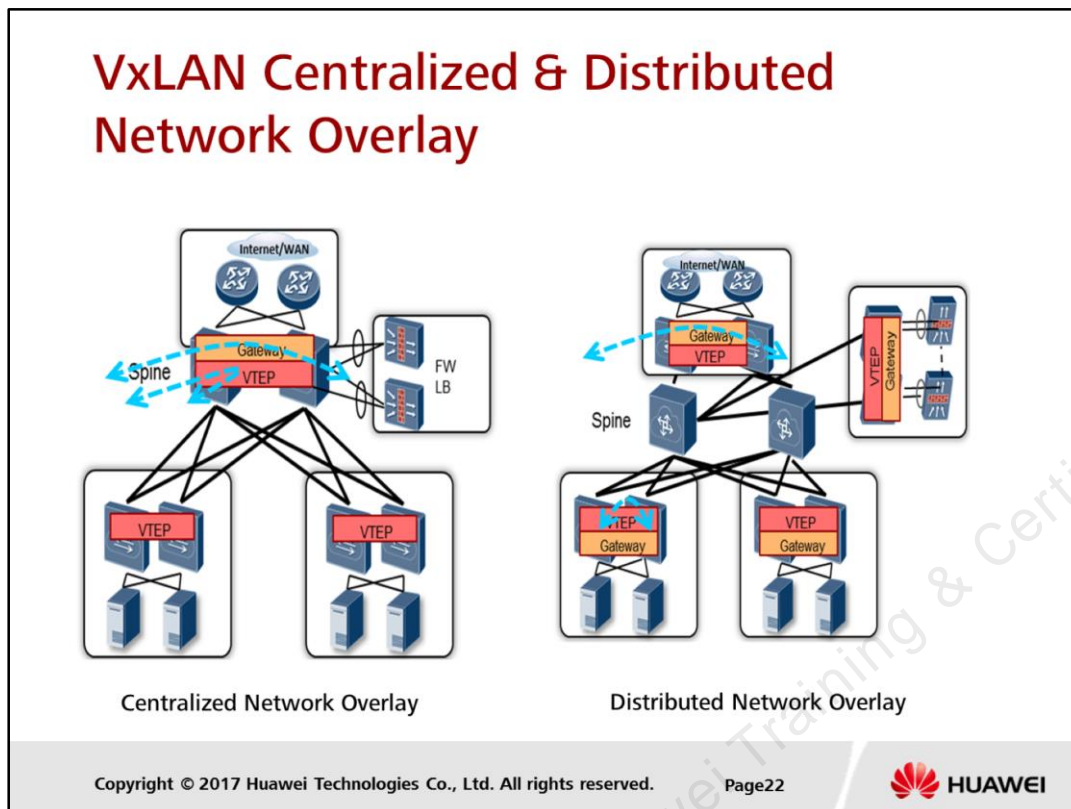




1. After Switch\_1 receives a packet from terminal A, Switch\_1 determines the Layer 2 broadcast domain of the packet based on the access interface and VLAN information carried in the packet, and checks whether the destination MAC address is a known unicast address.
  - If the destination MAC address is a known unicast address, Switch\_1 checks whether the destination MAC address is a local MAC address.
    - If so, Switch\_1 processes the packet.
    - If not, Switch\_1 searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain. Go to Step 2.
  - If the destination MAC address is not a known unicast address, Switch\_1 broadcasts the packet in the Layer 2 broadcast domain. Go to Step 2.
2. The VTEP on Switch\_1 performs VXLAN tunnel encapsulation based on the outbound interface and encapsulation information, and forwards the packet.
3. After the VTEP on Switch\_2 receives the VXLAN packet, it checks the UDP destination port number, source and destination IP addresses, and VNI of the packet to determine its validity. The VTEP obtains the Layer 2 broadcast domain based on the VNI and performs the destination MAC address is a known unicast address.
  - If the destination MAC address is a known unicast address, the VTEP searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain. Go to Step 4.
  - If the destination MAC address is not a known unicast address, the VTEP checks whether the destination MAC address is a local MAC address.
    - If so, the VTEP sends the packet to Switch\_2.
    - If not, the VTEP forwards the packet according to the BUM Packet Forwarding Process.
4. Switch\_2 adds a VLAN tag to the packet based on the outbound interface and encapsulation information, and forwards the packet to terminal B.



- BUM = Broadcast Unknown Unicast and Multicast
1. After Switch\_1 receives a packet from terminal A, Switch\_1 determines the Layer 2 broadcast domain of the packet based on the access interface and VLAN information carried in the packet, and checks whether the destination MAC address is a BUM address.
    - If the destination MAC address is a BUM address, Switch\_1 broadcasts the packet in the Layer 2 broadcast domain. Go to Step 2.
    - If the destination MAC address is not a BUM address, Switch\_1 forwards the packet according to the **Forwarding Process of Known Unicast Packets**.
  2. The VTEP on Switch\_1 obtains the ingress replication list for the VNI based on the Layer2 broadcast domain, replicates the packet based on the list, and performs VXLAN tunnel encapsulation by adding the VXLAN header and outer IP header. Switch\_1 then forwards the packet through the outbound interface.
  3. After receiving the VXLAN packet, the VTEP on Switch\_2 or Switch\_3 checks the UDP destination port number, source and destination IP addresses, and VNI of the packet to determine its validity. The VTEP obtains the Layer 2 broadcast domain based on the VNI and performs VXLAN tunnel decapsulation to obtain the inner Layer 2 packet. The VTEP then determines whether the destination MAC address is a BUM address.
    - If the destination MAC address is a BUM address, the VTEP broadcasts the packet in the Layer 2 broadcast domain.
    - If the destination MAC address is not a BUM address, the VTEP checks whether the destination MAC address is a local MAC address.
      - If so, the VTEP sends the packet to Switch\_2 or Switch\_3.
      - If not, the VTEP searches for the outbound interface and encapsulation information in the Layer 2 broadcast domain. Go to Step 4.
  4. Switch\_2 or Switch\_3 adds a VLAN tag to the packet based on the outbound interface and encapsulation information, and then forwards the packet to terminal B or terminal C.



- **Centralized mode**

- In VxLAN network, L3 gateway function is centralized on one or one group of switches
  - Leaf switches who is connected to firewall, load balancer, and various servers only serves as L2 gateway.

- **Distributed Mode:**

- In Network Overlay distributed VxLAN network, all leaf physical switches are equipped with L3 gateway functions; Spine functions as traffic forwarding node, and does not function as VTEP;

- The disadvantages of centralized network overlay is that traffic might be passing through the sub optimal path as all inter network routes must be passed through spine. Distributed mode can solve this problem as leaf is serving as L3 gateway too.
- For V3R3 AC DCN solutions, both hardware overlay in centralized mode and distributed mode are supported. Underlay physical design is same for both centralized and distributed gateway deployment; However, the overlay configuration might be different.
- Centralized hardware overlay will be used for the following discussion in the following slides.



## Contents

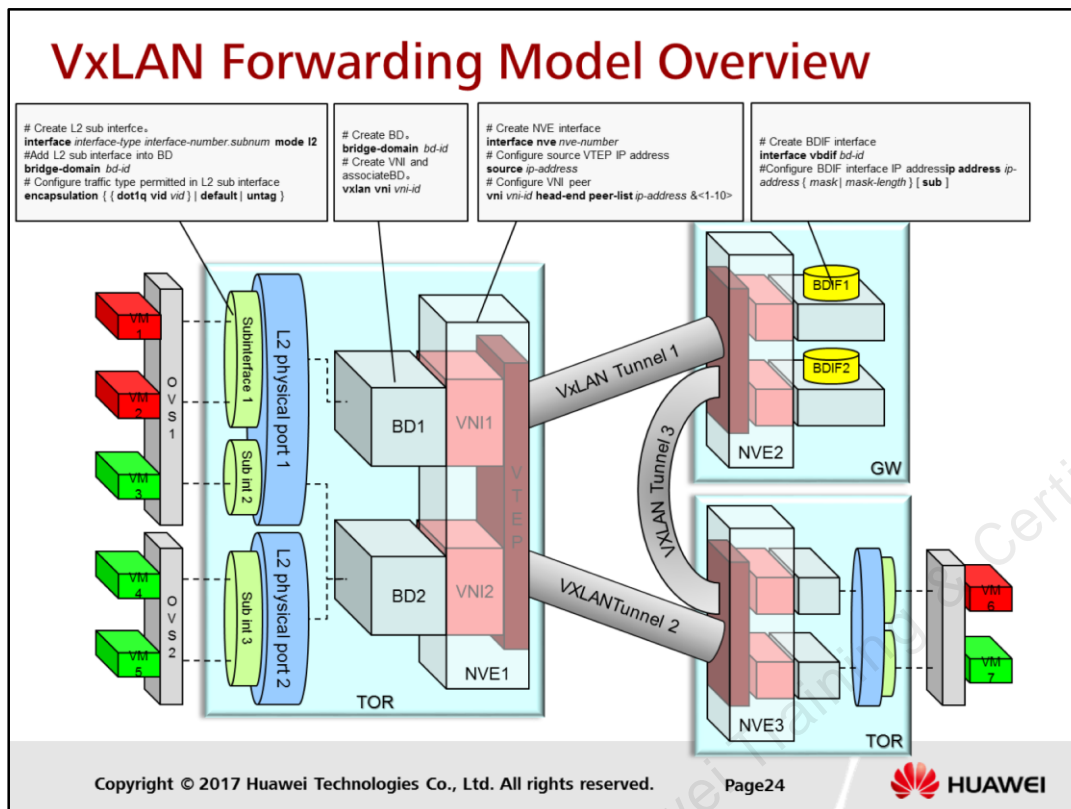
### 2. VxLAN Basic Concepts

#### 2.1 VxLAN Basic Principles

#### **2.2 VxLAN Forwarding Models**

##### 2.2.1 VxLAN Forwarding Models for VMs in same subnet

##### 2.2.2 VxLAN Forwarding Models for VMs in different subnet



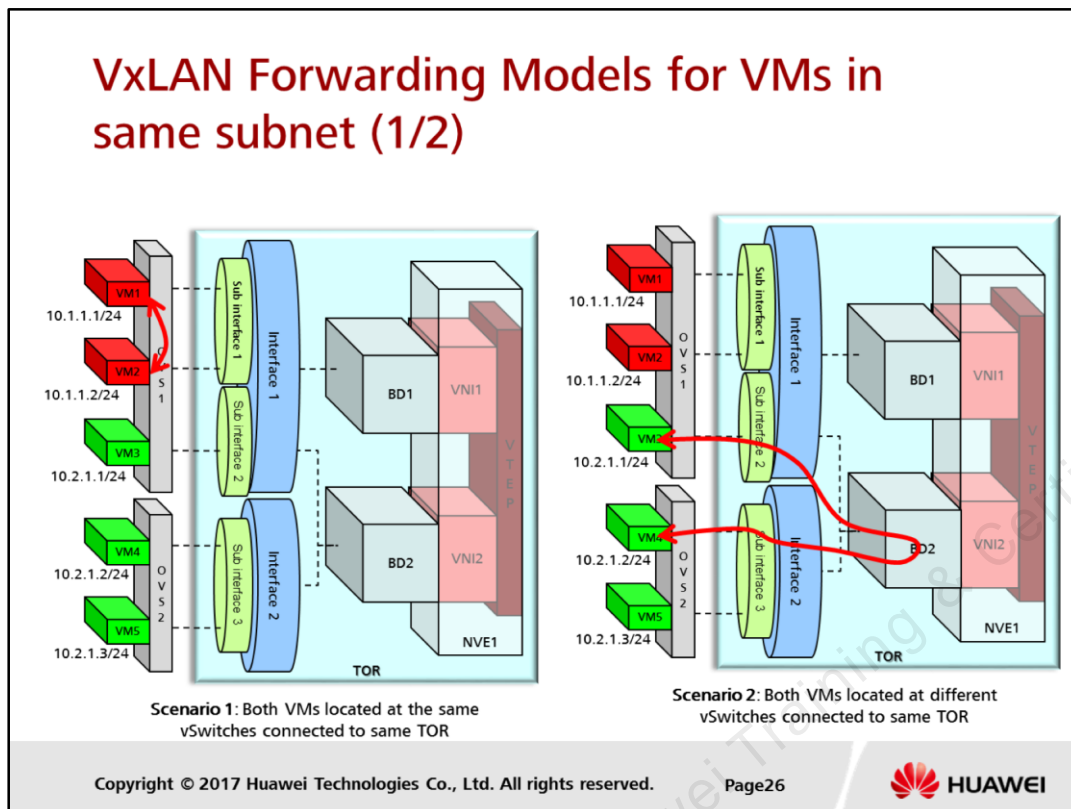
- The VxLAN forwarding model here is only discussing on the centralized gateway mode; the distributed gateway mode is not discussed in this slide here. Diagram above shows the basic configuration to be done on VxLAN configuration to prepare for VxLAN forwarding model.

## Contents

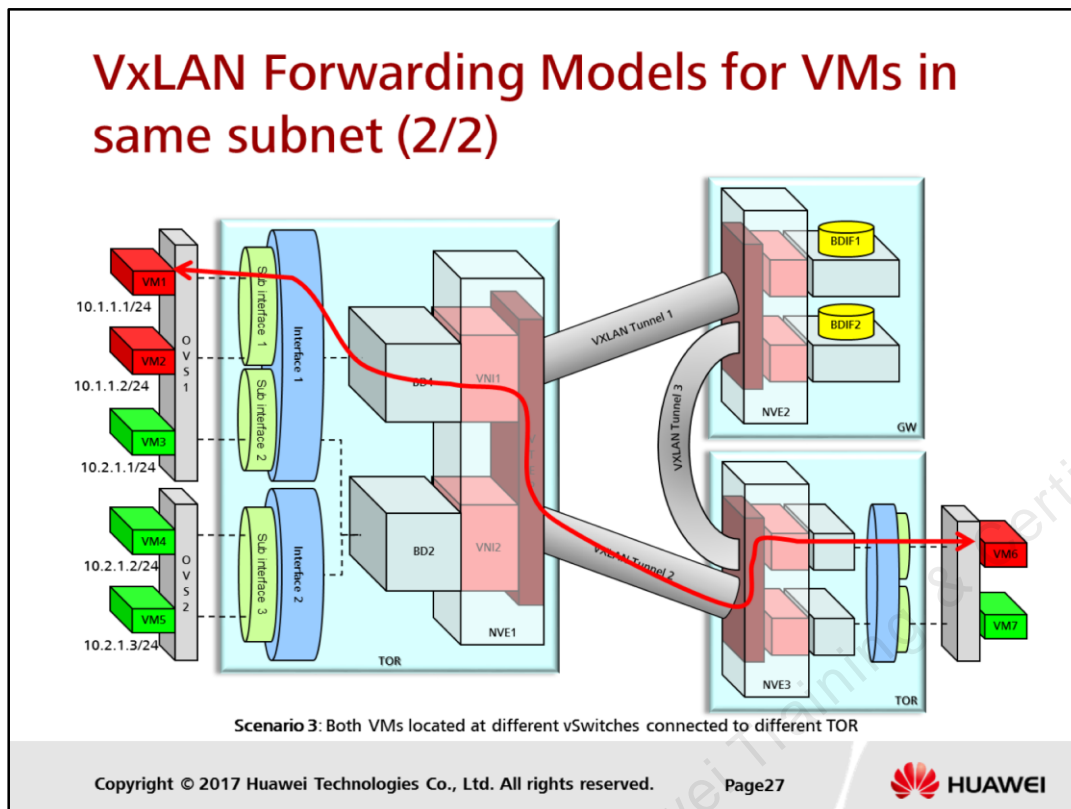
### 2.2 VxLAN Forwarding Models

#### **2.2.1 VxLAN Forwarding Models for VMs in same subnet**

#### 2.2.2 VxLAN Forwarding Models for VMs in different subnet



- Communication for VMs in the same subnet can be further divided into 3 scenarios, depending on the locations of VMs and connections to TOR, as per listed below:-
  1. **Both VMs are located in the same vSwitches connected to the same TOR.**
    - In this case, the communication of the 2 VMs in the same network segment is just through L2 in OVS
  2. **Both VMs are located in different vSwitches but connected to the same TOR.**
    - TOR, serving as the NVE binds VM in the same network segments in the same bridge domain mapping to the same VNI (VNI is mapped to the user access VLAN as well). Thus, inter vSwitches communication can be achieved on TOR in the same bridge domain as they are mapping to the same VLAN.



### 3. Both VMs re located at different vswitches connected to different TOR.

- As this is intra-segment communication, the communication does not need to pass through gateway but can be achieved by using the VxLAN tunnel built between 2 VxLAN L2 gateway, which is on both TOR serving as VTEP.

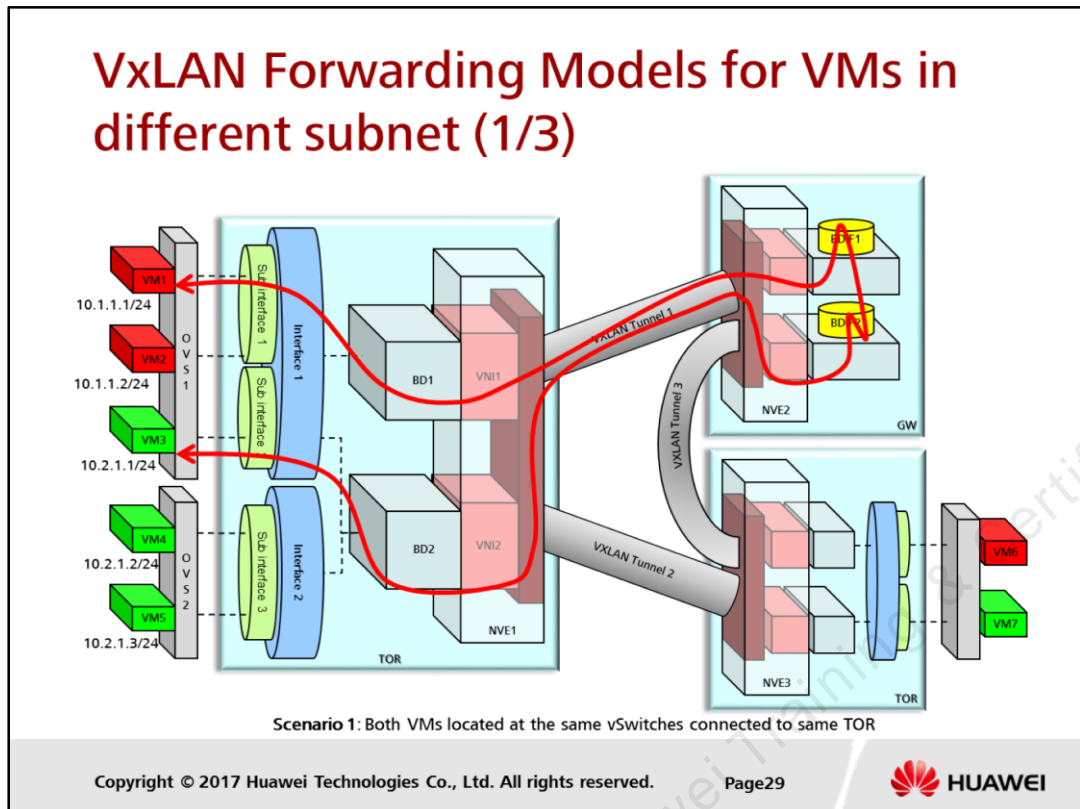


## Contents

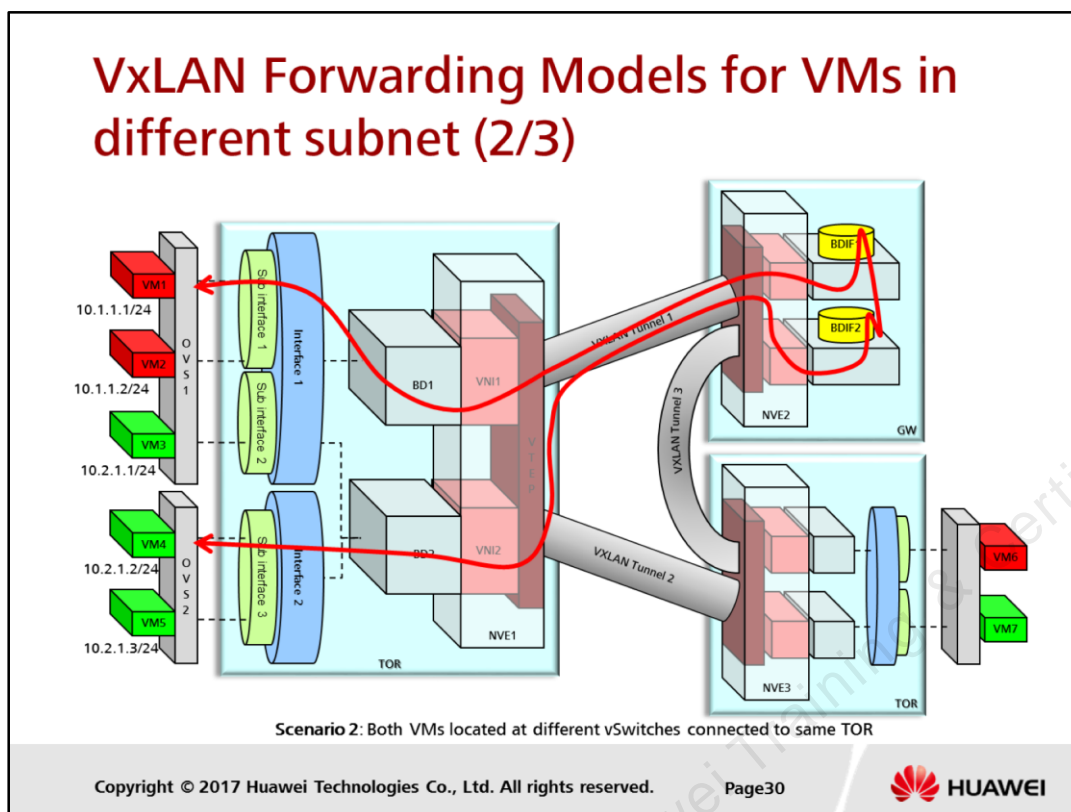
### 2.2 VxLAN Forwarding Models

#### 2.2.1 VxLAN Forwarding Models for VMs in same subnet

#### **2.2.2 VxLAN Forwarding Models for VMs in different subnet**

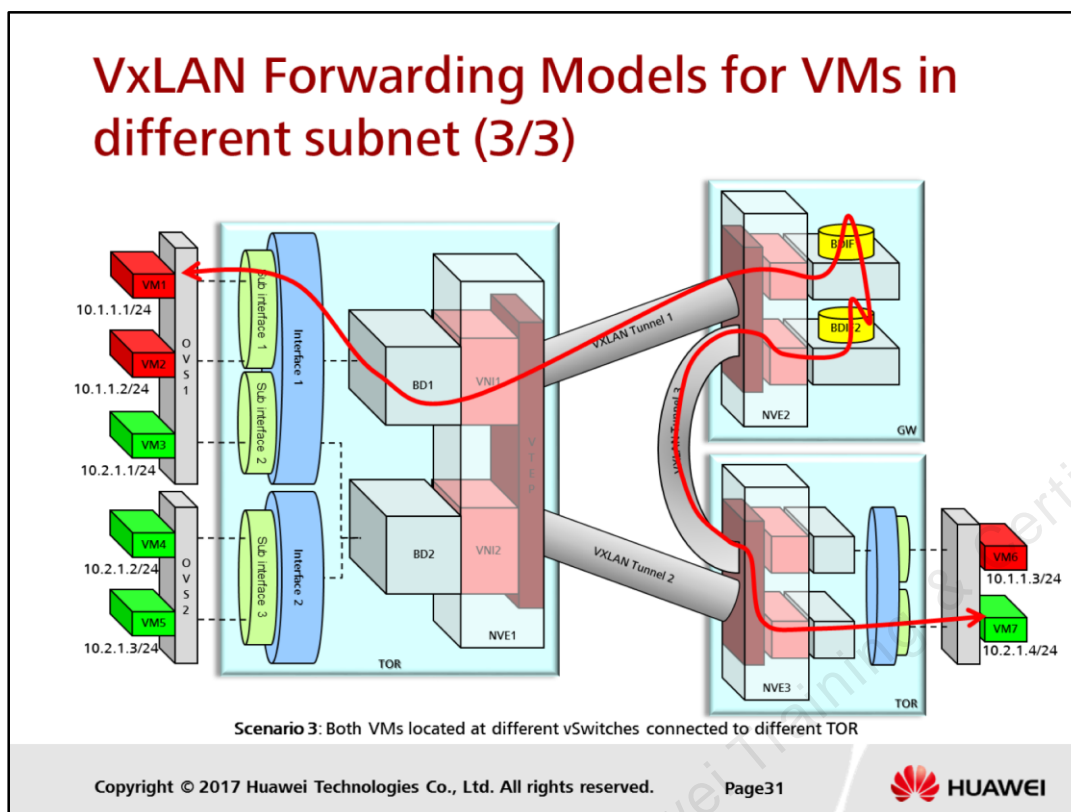


- Communication for VMs in the different subnet can be further divided into a 3 scenarios, depending on the locations of VMs and connections to TOR, as per listed below:-
  1. **Both VMs are located in the same vSwitches connected to the same TOR.**
    - As both VMs are in different network segments, VM1 will forward its data to the L3 gateway, the L3 gateway will route it to the VM3 through another BDI interface to reach another network.



## 2.. Both VMs are located in different vSwitches but connected to the same TOR.

- As both VMs is in different network segment, VM1 will forwards it data to the L3 gateway, L3 gateway will route it to the VM3 through another Bdinterface to reach another network.



### 3. Both VMs are located at different vswitches connected to different TOR.

- As both VMs is in different network segment on different TOR, VM1 will forwards it data to the L3 gateway through VxLAN tunnel, L3 gateway will route it to the VM3 through another Bdinterface to reach another network located on another switches through another VxLAN tunnel.

## Contents

1. VxLAN Overlay Overview
2. VxLAN Basic Concepts
3. **VxLAN Applications in SDN AC-DCN Cloud Fabric Network**
4. VxLAN Configuration Examples in SDN AC-DCN Cloud Fabric Network

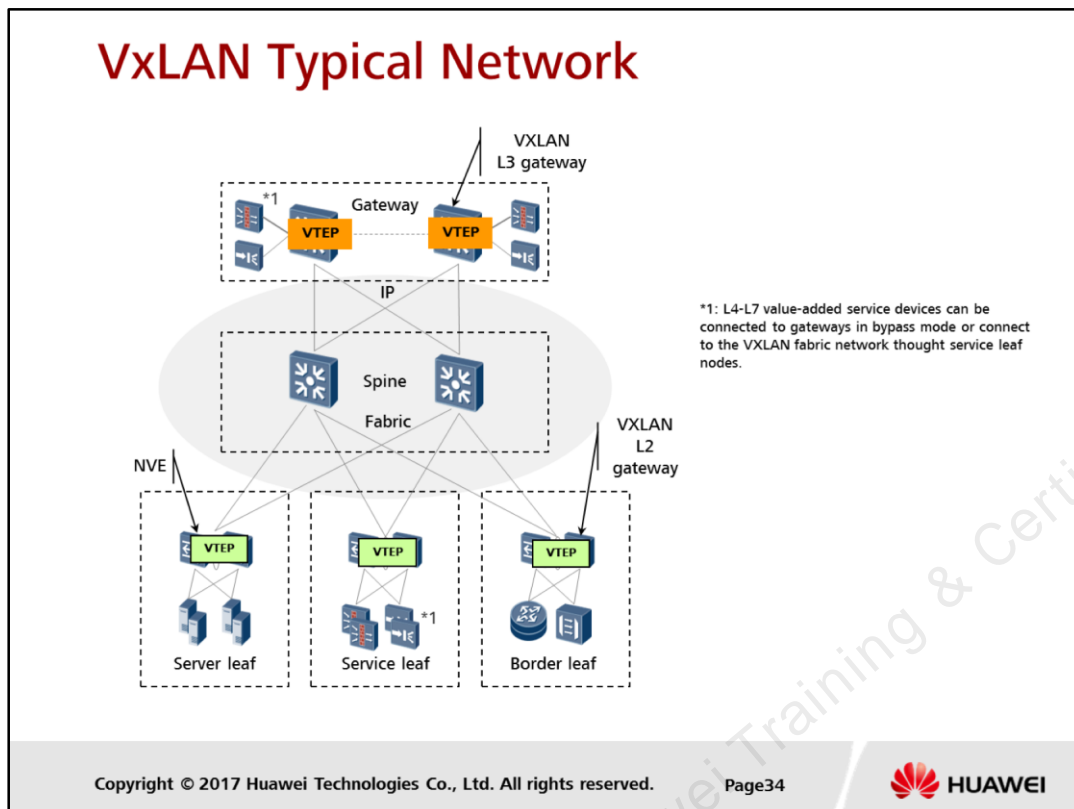


## Contents

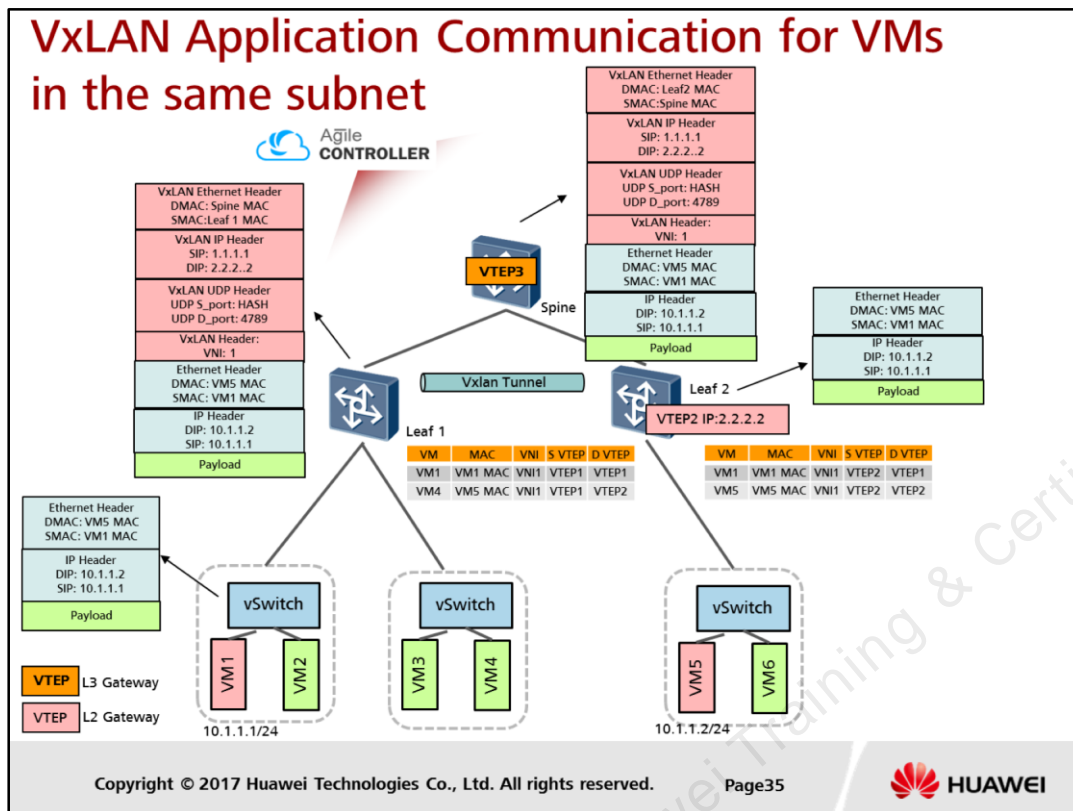
### 3. VxLAN Applications in SDN Cloud Fabric DCN

#### **3.1 VxLAN VM Communication in Cloud Fabric DCN**

#### 3.2 VxLAN Fabric Deployment in Cloud Fabric DCN

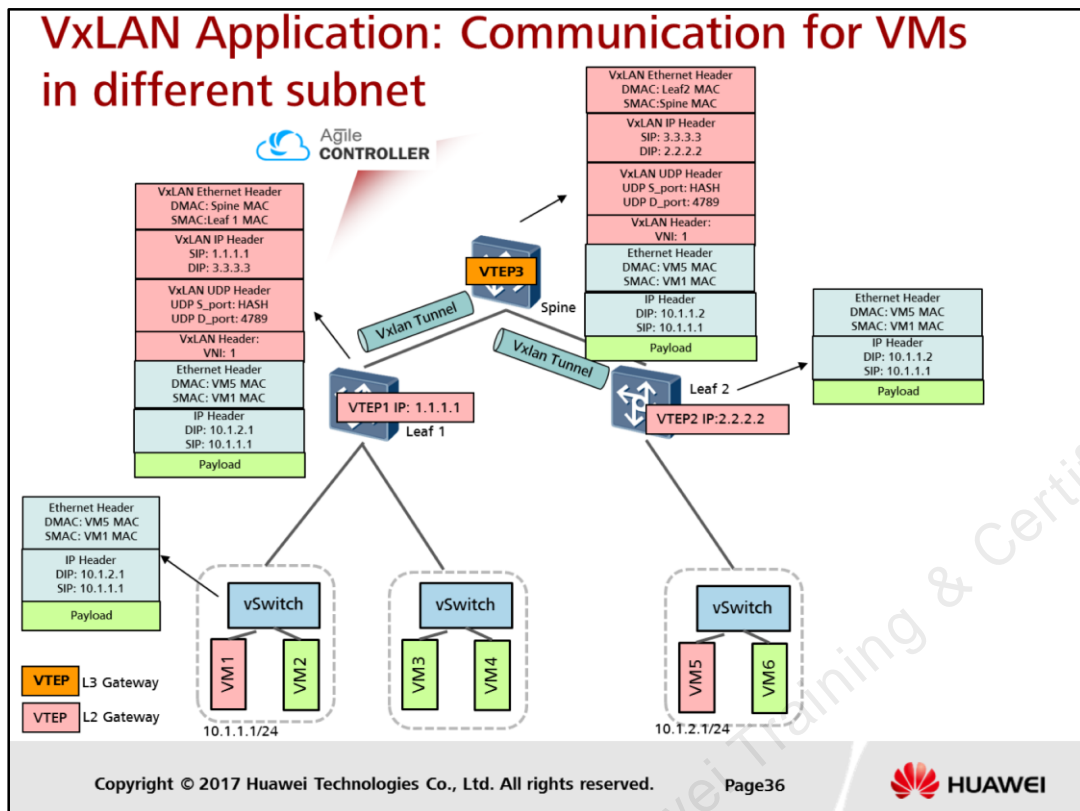


- The explanation and general terms of VxLAN network in DCN is listed below:-
1. **Fabric:** A basic physical network for a data center, which is composed of a group spine and leaf nodes.
  2. **Spine:** Core node of a VXLAN fabric network, which uses high-speed interfaces to connect to functional leaf nodes and provides high-speed IP forwarding.
  3. **Leaf:** An access node that is deployed on a VXLAN fabric network to connect various network devices to the VXLAN network.
  4. **Service leaf:** A functional leaf node that connects L4-L7 value-added service devices, such as firewall and LB, to the VXLAN fabric network.
  5. **Server leaf:** A functional leaf node that connects computing resources (virtual or physical servers) to the VXLAN network.
  6. **Border leaf:** A functional leaf node that connects to a router or transmission device and forwards traffic sent from external networks to the data center.
  7. **NVE:** Network virtualization edge, a network entity that implements network virtualization. NVE nodes establish an overlay virtual network on the underlay Layer 3 basic network.
  8. **VTEP:** VXLAN tunnel endpoints that are deployed on NVE nodes and responsible for VXLAN packet encapsulation and decapsulation. VTEPs are connected to the physical network and assigned IP addresses (VTEP IP) of the physical network. VTEP IP addresses are independent of the virtual network. A local VTEP IP address and a remote VTEP IP address identify a VXLAN tunnel.
  9. **VNI:** VXLAN network identifier that identifies a VXLAN segment. Traffic sent from one VXLAN segment to another must be forwarded by a VXLAN L3 gateway.

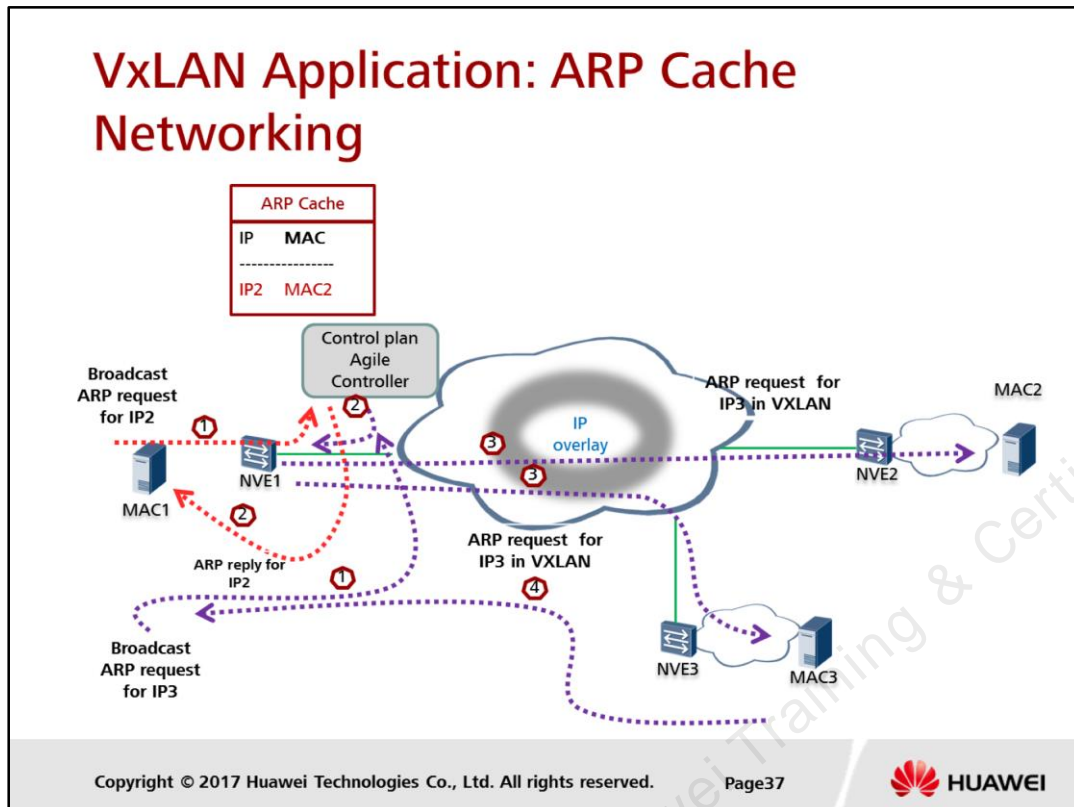


- As shown in the diagram above, Leaf switches serve as VxLAN L2 gateway and VxLAN tunnel is established between both Leaf1 and Leaf2 switches. Openflow channel is established between AC and forwarders through openflow protocol. VxLAN configuration is performed by administrator on AC using Netconf protocol, and the VxLAN configuration is deployed to forwarders through VxLAN channel configured.
- Through the VxLAN tunnel established, same segment VM communication will be performed through VTEP by using the MAC address mapping table.





- As shown in the diagram above, Spine switch serves as VxLAN L3 gateway while leaf1 and leaf2 switches serves as VxLAN L2 gateway; VxLAN tunnel is built between switches and realizes inter-segment VM communication (VM1 to VM5) through L3 gateway. Openflow channel is established between AC and forwarder using Openflow protocol.
- Through Netconf protocol, administrator performs VxLAN configuration on AC and AC deploys VxLAN information to forwarder through Openflow channel.
- Logical BDIF interface configuration is performed on L3 gateway and ARP cache function is enabled on AC. Inter segment VM communication can be achieved by VxLAN L3 gateway and ARP cache proxy function.
- For example based on diagram above, VM1 is to communicate with VM5 in different network segment connected to different leaf switches. All VxLAN configuration has been completed by admin and configuration has been deployed from AC to switches using Openflow. When the L2 Ethernet frame reaches VTEP1, VTEP1 serves as L2 gateway will perform VxLAN encapsulation and perform forwarding based on VxLAN Ethernet Header. When VTEP3 receives the VxLAN frames, it performs VxLAN de-encapsulation and find out that the DMAC is not on local segment; It serves as VxLAN L3 gateway, removes the ethernet header, check the ARP entry matching entry of BDIF interface, and encapsulates back VxLAN and send it to VTEP2. When VxLAN frame reaches VTEP2, VxLAN header is removed and frame is forwarded based on the original ethernet frame.



- AC obtains the VM detailed information and keeps ARP cache in the controller. When MAC1 broadcast an ARP request for IP2, ingress NVE sends this broadcast message will be sent to AC for processing; AC analyzes the ARP request and searches the ARP cache based on mapping of MAC address entry; if The MAC entry is existing in the cache, AC will reply with ARP unicast reply and the broadcast message is terminated.
- Else if the MAC entry it is not existing in ARP cache, AC will send to ARP request back to NVE1, NVE1 broadcasts to everyone to get reply from the real host. As shown in example, when MAC1 sends a broadcast ARP request for IP3 and the corresponding MAC address does not exist in ARP cache of AC, AC will send back to NVE1 and NVE1 broadcast to NVE2 and NVE3 to reach all hosts. MAC3 will reply in this case.

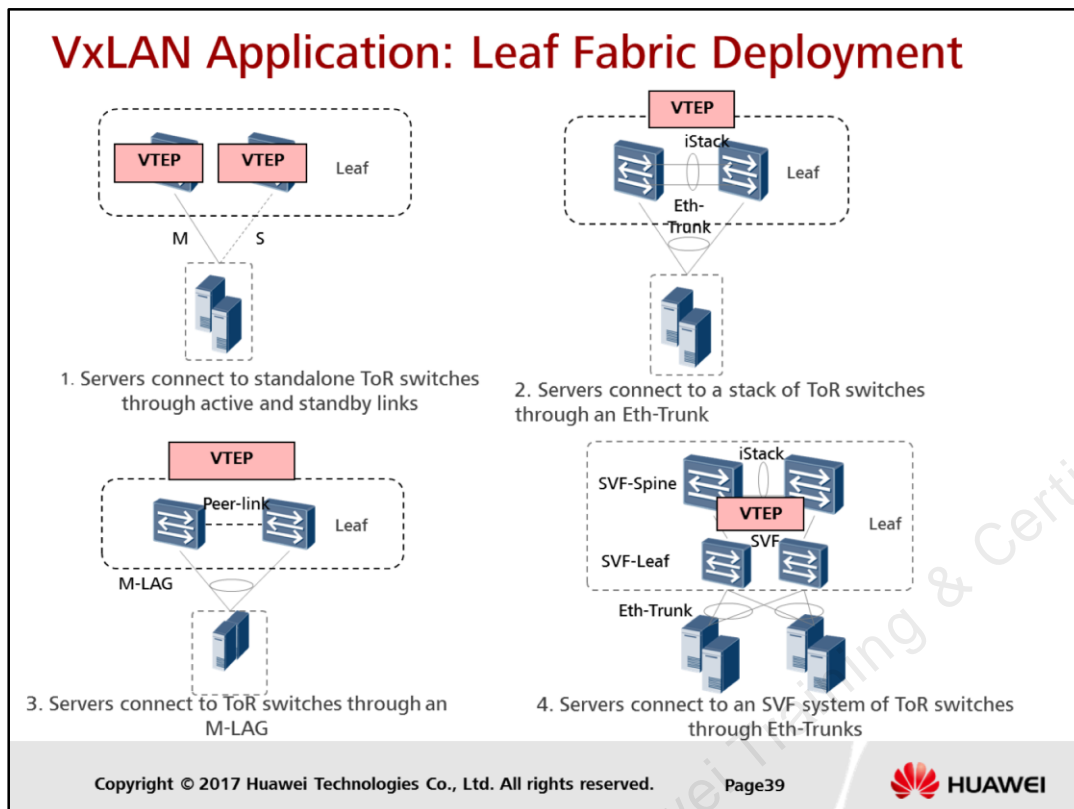


## Contents

### 3. VxLAN Applications in Cloud Fabric DCN

#### 3.1 VxLAN VM Communication in Cloud Fabric DCN

#### **3.2 VxLAN Fabric Deployment in Cloud Fabric DCN**



- There are basically 4 different types of VxLAN leaf fabric deployment in Cloud DCN, considering redundancy and protection level. Thus, a server is normally dual-homed to 2 leaf fabric physically. The 4 types of VxLAN leaf fabric deployment is listed below:

1. **Servers connect to standalone ToR switches through active and standby links**

- A standalone ToR switch acts as a leaf node. Each server is connected to two ToR switches using active-standby NICs (NIC bonding). Only one NIC in a server sends and receives packets at a time, resulting in a low bandwidth efficiency. The VTEP IP address will change after an active/standby NIC switchover. In this case, the upstream VTEP needs to learn the forwarding entry from the BUM traffic sent from the server.

2. **Servers connect to a stack of ToR switches through an Eth-Trunk**

- iStack technology virtualizes two ToR switches into one logical switch with a single control plane, which simplifies device management. The NICs of a server work in active-standby/load-balancing mode to improve bandwidth efficiency. However, the upgrade and maintenance operations for the logical device are complex.

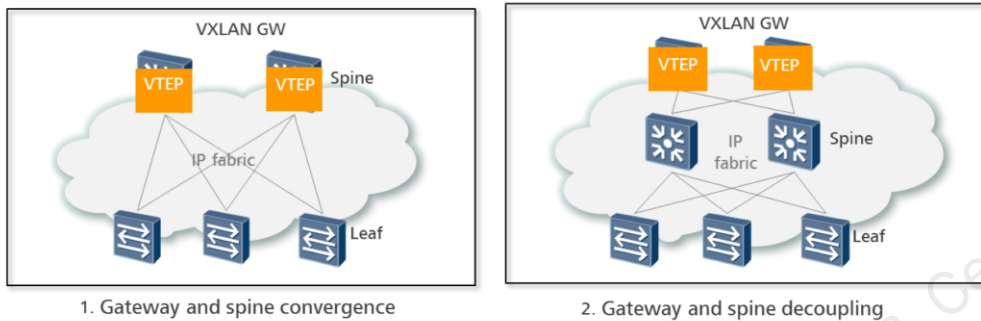
3. **Servers connect to ToR switches through an M-LAG.**

- Two ToR switches are connected using a peer-link and set up a Dynamic Fabric Service (DFS) group. The two switches act as one logical device but have their own independent control planes, simplifying upgrade and maintenance while improving system reliability. Their downlink ports set up an M-LAG for dual-homing of servers. NICs of a server work in active-standby/load-balancing mode. The configuration is complex because each ToR switch has an independent control plane.

4. **Servers connect to an SVF system of ToR switches through Eth-Trunks**

- Two high-performance ToR switches supporting VXLAN use iStack technology to set up a stack. Cost-effective ToR switches with SVF configured are connected to the stack to provide cost-effective 1G/10G VXLAN network access capability. Each server is dual-homed to two SVF leaf nodes, with the NICs working in active-standby/load-balancing mode.

## VxLAN Application: Gateway Deployment Mode



Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

Page40



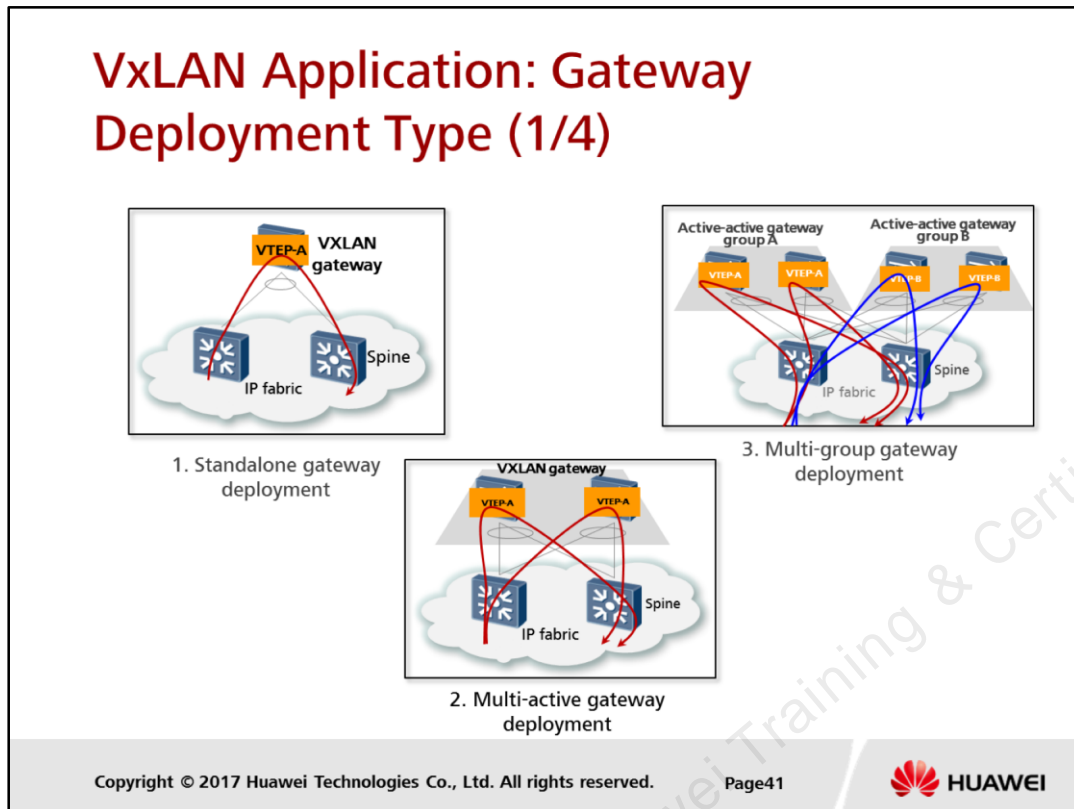
- For the VxLAN deployment in Cloud DCN scenario, the gateway deployment can be divided into 2 physical deployment type:-

### 1. Gateway and spine convergence

- The gateway and VxLAN termination point is deployed on the same spine switch; which means the exit gateway connecting to internet function also is realized on this spine switch. This realizes 2 layer architecture in the physical underlay deployment
- The convergence deployment reduces the number of network devices and lowers the network deployment cost.
- The gateway nodes are closely coupled with the spine nodes, making network expansion difficult. This deployment is applicable to a data center that does not need to be expanded in the near future.
- Gateways cannot be deployed in multi-group mode.

### 2. Gateway and spine decoupling

- Exit gateway and VxLAN gateway is realized on different equipments; this leads to 3 layers architecture in the underlay deployment.
- The decoupling deployment facilitates network expansion. Expansion of the spine, leaf, or gateway nodes will not greatly affect the other nodes.
- Multiple groups of gateways can be deployed on a large-sized network.
- Gateways can be deployed in multi-group, multi-active mode.



- There are 3 different types of gateway deployments, which are listed below:-

### 1. Standalone gateway deployment

- A standalone switch or stack system act as the VXLAN gateway. It supports 4K tenants (2K tenants in VPN access scenario), 4K VRFs, 4K subnets, 4K VNIs, 125K VMs (or 25K VMs + physical servers), 32K-1M RIB entries, and 5K ACLs.

### 2. Multi-active gateway deployment

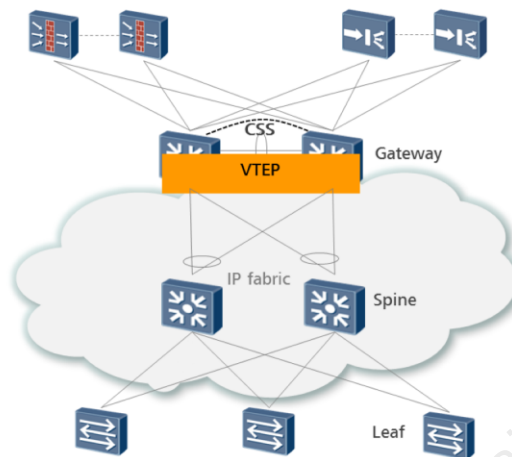
- Two or four gateway devices set up a DFS group and are configured with the same gateway address and VTEP IP address to act as one logical gateway for VMs.

### 3. Multi-group gateway deployment

- More subnets can be supported by deploying more gateway groups, but the forwarding capability and reliability of each gateway group remain unchanged.

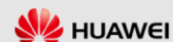
## VxLAN Application: Gateway Deployment Type (2/4)

- Gateway standalone Deployment



Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

Page42



### Requirements

- A CSS system has been deployed in a DC.
- The VxLAN gateway needs to be deployed on the CSS system.

### Solution design

- The CSS system acting as the gateway has a vBDIF IP address, virtual MAC address, and VTEP IP address configured.
- Value-added service devices (FW/LB) are dual-homed to the CSS system in bypass mode.
- Value-added service devices are expanded together with the gateway devices.

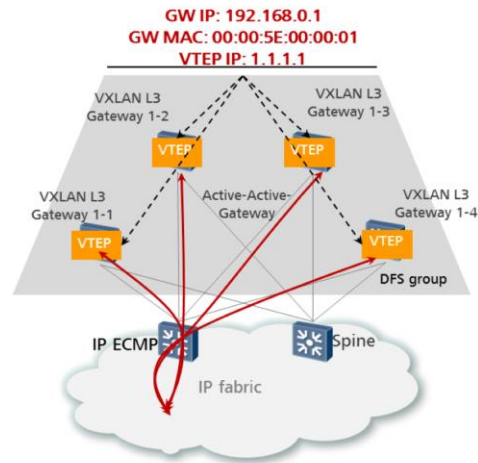
### Characteristics

- This solution can be used for VLAN-to-VxLAN evolution.
- The CSS system is managed and configured as an independent logical device, which simplifies device management and facilitates network O&M.
- The CSS gateway is more reliable than a standalone gateway device.



## VxLAN Application: Gateway Deployment Type (3/4)

- Multi-active gateway deployment



Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

Page43

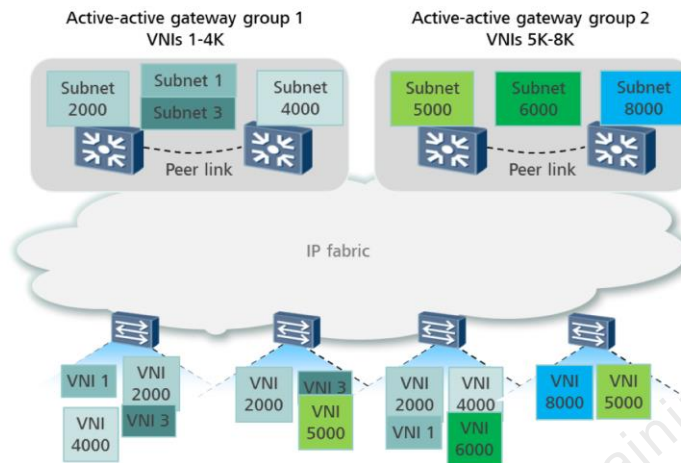


- Multiple gateway devices are configured with the **same gateway address and VTEP IP address**. VMs are unaware of locations and number of gateway devices.
- Multi-active gateway devices **set up a DFS group and use the peer link between them to synchronize ARP and MAC entries**, so that they save the same traffic forwarding information.
- Each gateway device in a multi-active gateway group has all forwarding information and can work independently, providing high gateway reliability.
- The underlay network uses IP ECMP to implement load balancing, which enables traffic to be evenly distributed to gateways and improves forwarding performance.
- FYI:** In multi-active gateway deployment, ping to a virtual or physical server from the gateway may fail because of inconsistent forward and reverse paths. This is a normal situation.
- This solution cannot increase the number of VRF/Subnet/RIB/FIB/ARP/MAC entries supported on gateway devices.
- This deployment is applicable to private cloud data centers requiring high reliability.



## VxLAN Application: Gateway Deployment Type (4/4)

- Multi-group gateway deployment



Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

Page44



### Solution description

- Deploy new gateway groups to increase the number of resources by multiple times. The new gateway groups run independently and do not affect the original gateway group. VNIs on the same ToR switch can belong to different gateway groups.
- The network scale in a POD expands by multiple times, but the capacity of a single gateway group remains unchanged.

### Solution design

- Deploy multiple gateway groups in PODs with a large number subnets.
- A POD supports a maximum of four gateway groups. Each gateway group supports 4K routing domains, 4K subnets, and 25K ARP entries. (This is the specification in a scenario with both virtual and physical servers. 125K ARP entries are supported if there are only VMs.)
- **A gateway group can contain a single gateway, active-active gateways, or quad-active gateways.**
- A tenant can select a gateway group when creating the first VRF. Subsequent VRFs created by the tenant are automatically assigned to the selected gateway group. (Subnets of a tenant must be deployed on the same gateway group.)
- The Agile Controller can assign gateway groups to tenants based on loads of gateway groups.
- Traffic sent from a spine node to a multi-active gateway group is load balanced among IP ECMP paths.

### Characteristics

- This solution is applicable to large-scale private cloud DCs.



## Contents

1. VxLAN Overlay Overview
2. VxLAN Basic Concepts
3. VxLAN Applications in SDN AC-DCN Cloud Fabric Network
4. **VxLAN Configuration Examples in SDN AC-DCN Cloud Fabric Network**

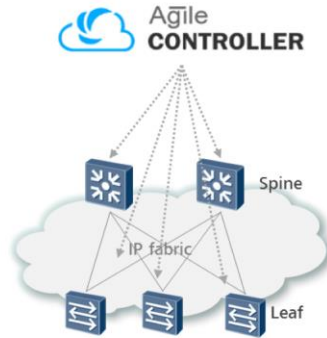
## Contents

### 4. VxLAN Configuration Example in AC-DCN

#### 4.1 Configuration between AC and Switches

#### 4.2 Configuring VxLAN Overlay Network

## Configuration between AC and Switches



Protocol	Function
SNMP	It is used for AC to be able to add forwarders into topology and manage device status and alarms remotely.
Netconf	For AC to deploy and obtain forwarders' configurations
Openflow	Through Openflow protocol, AC can send and receive VxLAN information and ARP mapping table. After SNMP and Netconf connection is established between AC and forwarders, Openflow configuration on forwarders can be performed through Netconf; no manual configuration on forwarder is needed.

Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

Page47



- There are 3 types of protocol configured between AC and forwarders; all protocols are configured serving for different function.
- SNMP and Netconf configuration must be configured manually on switches while Openflow configuration can be deployed through AC to switches after SNMP and Netconf connection is established.

## Netconf Configuration Example on Switches

### 1. Configure SSH users

```
<Gateway-CE12808-1>system-view
[~Gateway-CE12808-1]user-interface vty 0 4
[~Gateway-CE12808-1-agent-ui-vty0-4]authentication-mode aaa
[~Gateway-CE12808-1-ui-vty0-4]protocol inbound ssh
[~Gateway-CE12808-1-ui-vty0-4]commit
[~Gateway-CE12808-1-ui-vty0-4]quit
[~Gateway-CE12808-1]aaa
[~Gateway-CE12808-1-aaa]local-user client@huawei.com password irreversible-cipher Huawei@123
[~Gateway-CE12808-1-aaa]local-user client@huawei.com service-type ssh
[~Gateway-CE12808-1-aaa]local-user client@huawei.com level 3
[~Gateway-CE12808-1-aaa]commit
[~Gateway-CE12808-1-aaa]quit
```

### 2. Create local RSA key

```
[~Gateway-CE12808-1] rsa local-key-pair create
The key name will be: netconf-agent_Host
The range of public key size is (512 ~ 2048).
NOTE: If the key modulus is greater than 512,
      It will take a few minutes.
Input the bits in the modulus [default = 512] :
[~Gateway-CE12808-1] commit
```

- Steps above shows an example of the manual configuration needed to be done on forwarders.

## Netconf Configuration Example on Switches

### 3. Configure SSH user authentication type and service type

```
[~Gateway-CE12808-1] ssh user client@huawei.com authentication-type password  
[~Gateway-CE12808-1] commit  
[~Gateway-CE12808-1] ssh user client@huawei.com service-type snetconf  
[~Gateway-CE12808-1] commit
```

### 4. Enable Netconf function in global

```
[~Gateway-CE12808-1] snetconf server enable  
[~Gateway-CE12808-1] commit
```

## SNMPv3 Configuration Example on Switches

### 1. Configure SNMPv3 user group, user name, authentication mode and privacy mode and passwords.

```
[*Gateway-CE12808-1] snmp-agent usm-user v3 admin group dc-admin
[*Gateway-CE12808-1] snmp-agent usm-user v3 admin authentication-mode sha
Please configure the authentication password (8-255)
Enter Password:          //Enter Password; password used here is Huawei@123
Confirm Password:       //Reconfirm Password; password used here is Huawei@123
[*Gateway-CE12808-1] snmp-agent usm-user v3 admin privacy-mode aes128
Please configure the privacy password (8-255)
Enter Password:          //Enter Password; password used here is Huawei@123
Confirm Password:       //Reconfirm Password; password used here is Huawei@123
```

### 2. Configure SNMPv3 trap function

```
[*Gateway-CE12808-1] snmp-agent trap enable feature-name trunk
[*Gateway-CE12808-1] snmp-agent trap enable
[*Gateway-CE12808-1] snmp-agent trap source loopback0
[*Gateway-CE12808-1] commitrd used here is Huawei@123
```

### 3. Configure SNMPv3 MIB view

```
[*Gateway-CE12808-1] snmp-agent mib-view included iso-view iso
[*Gateway-CE12808-1] snmp-agent mib-view included nt iso
[*Gateway-CE12808-1] snmp-agent mib-view included rd iso
[*Gateway-CE12808-1] snmp-agent mib-view included wt iso
[*Gateway-CE12808-1] snmp-agent group v3 dc-admin privacy read-view rd write-view wt notify-view nt
[*Gateway-CE12808-1] commit
```

- The AC-DCN obtains LLDP link information from the MIB view specified by SNMP. In this case, the specified MIB view must be iso-view, and the MIB sub-tree of the specified OID must be iso.

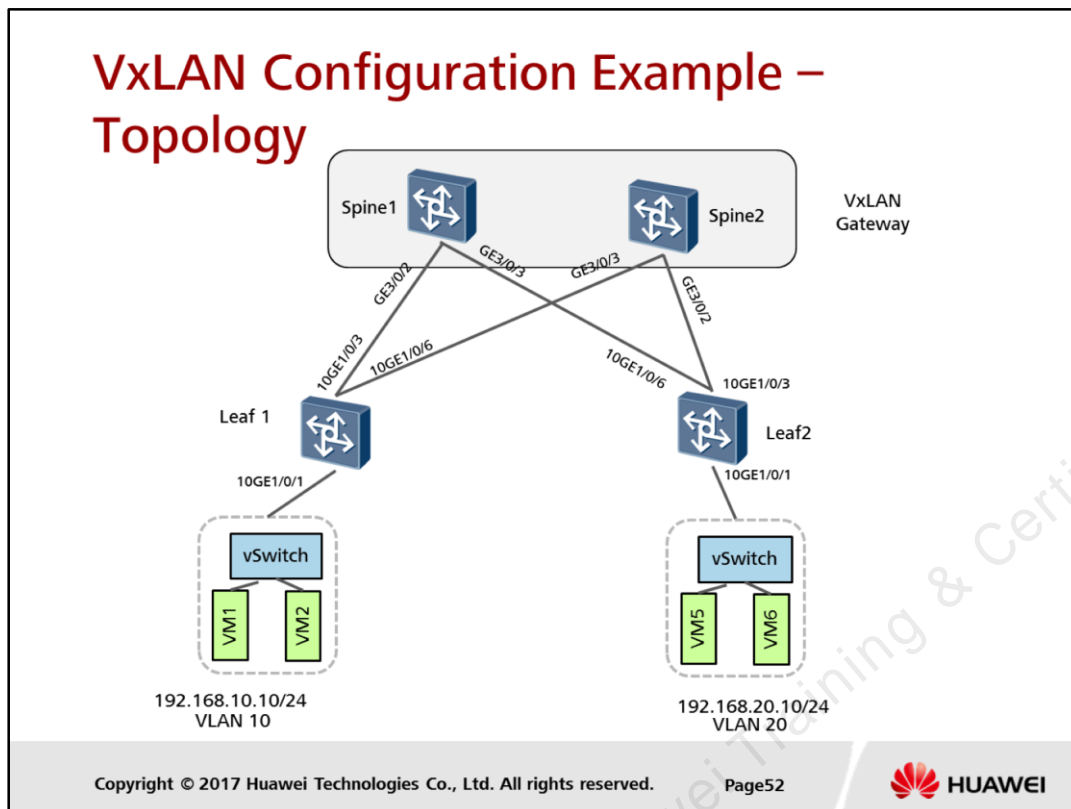
## Contents

### **4. VxLAN Configuration Example in AC-DCN**

4.1 Configuration between AC and Switches

**4.2 Configuring VxLAN Overlay Network**





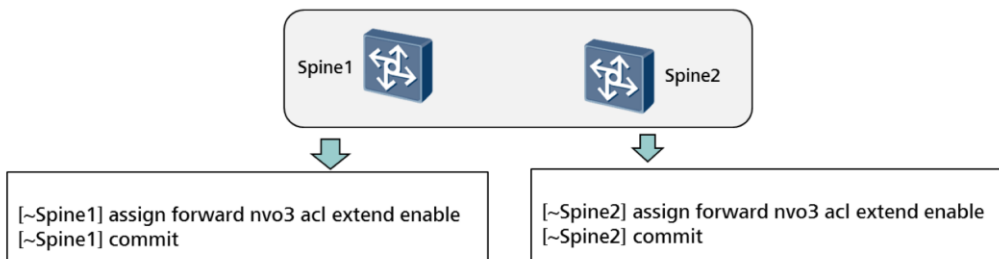
- As shown in the topology above, the DCN shown in the diagram is deploying gateway and spine converged 2 layer DC architecture and centralized multi-active gateway-group. Spine 1 and Spine 2 is located in the core layer while Leaf1, Leaf2, and Leaf 3 are located in the access layer. Full meshed connection is established between spines and leaves, performing ECMP for higher redundancy. No connection between Spines and Leafs.
- VMs are belonged to VLAN 10, 20 and 30 respectively; Bridge domain to be configured are BD10, BD20, and BD30; VxLAN VNI ID are VNI 5000, VNI 5001 and VNI 5002 respectively.

## VxLAN Configuration Example – Configuration Roadmap

Step	Description
Pre-requisite	Configure OSPF on Leaf1 and Leaf2 as well as Spine1 and Spine2 to ensure Layer 3 network connectivity.
1	Enable the NVO3 ACL extension function on Spine
2	Configure multi-active gateway on Spine1 and Spine 2 by configuring DFS group.
3	Configure VXLAN on Leaf1 and Leaf2 as well as Spine1 and Spine2 to construct a large Layer 2 VXLAN network over the basic Layer 3 network.
4	Configure service access points on Leaf1 to Leaf2 to distinguish traffic from servers and forward the traffic to the VXLAN network.
5	Configure VXLAN Layer 3 gateways on Spine1 and Spine2 to implement communication between VXLAN networks on different network segments and between VXLAN and non-VXLAN networks.

- Step 1 OSPF configuration is omitted.

## Step 1: Enable the NVO3 ACL extension function



- NOTE: After modifying the tunnel mode or enabling the NVO3 ACL extension function, you need to save the configuration and restart the device to make the configuration take effect. You can restart the device immediately or after completing all the configurations.

- NOTE: After modifying the tunnel mode or enabling the NVO3 ACL extension function, you need to save the configuration and restart the device to make the configuration take effect. You can restart the device immediately or after completing all the configurations.

## Step 2: Configure multi-active gateway on Spine1 & 2

Loopback 0:1.1.1.1/32      Loopback0:1.1.1.1/32  
 Loopback1: 100.1.1.1/32      Loopback1: 100.1.1.2/32

```
[~Spine1]dfs-group 1
[~Spine1-dfs-group-1]source ip 100.1.1.1
[~Spine1-dfs-group-1]active-active-gateway
[~Spine1-dfs-group-1-active-active-gateway]peer 100.1.1.2
[~Spine1]commit
```

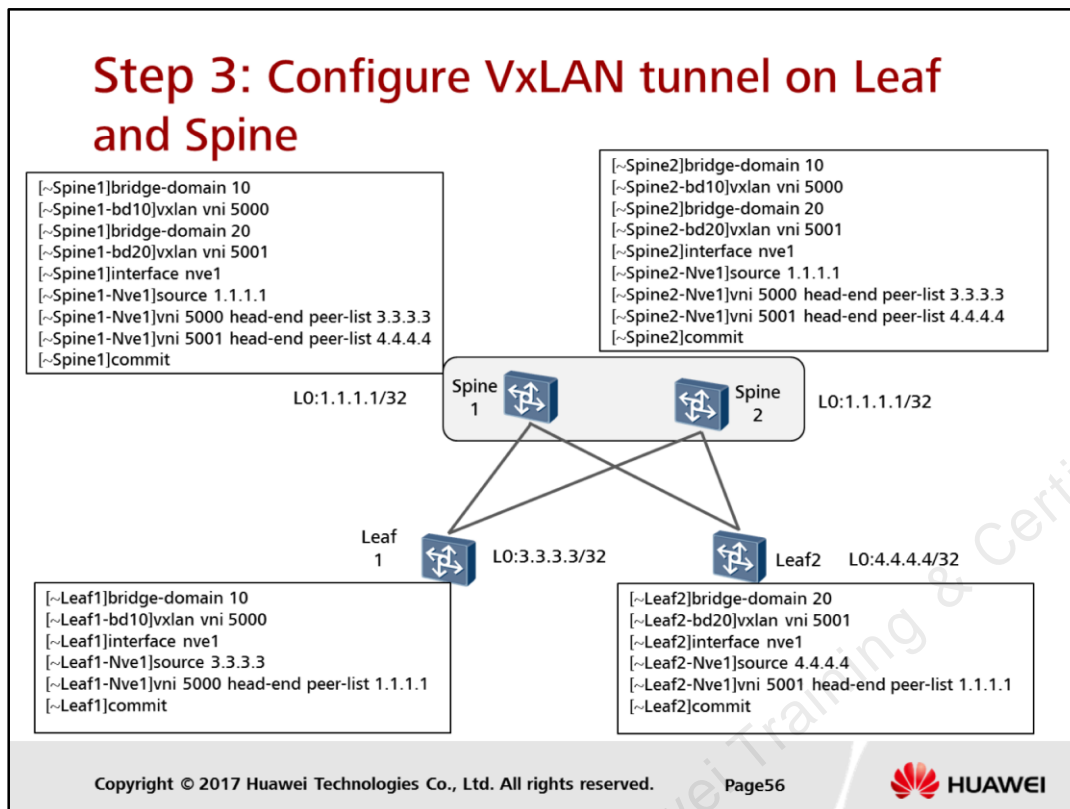
```
[~Spine2]dfs-group 1
[~Spine2-dfs-group-1]source ip 100.1.1.2
[~Spine2-dfs-group-1]active-active-gateway
[~Spine2-dfs-group-1-active-active-gateway]peer 100.1.1.1
[~Spine2]commit
```

```
[Spine1]display dfs-group 1 active-active-gateway
A:Active I:Inactive
-----
Peer      System name  State  Duration
100.1.1.2 Spine2       A      1:38:37
```

Copyright © 2017 Huawei Technologies Co., Ltd. All rights reserved.

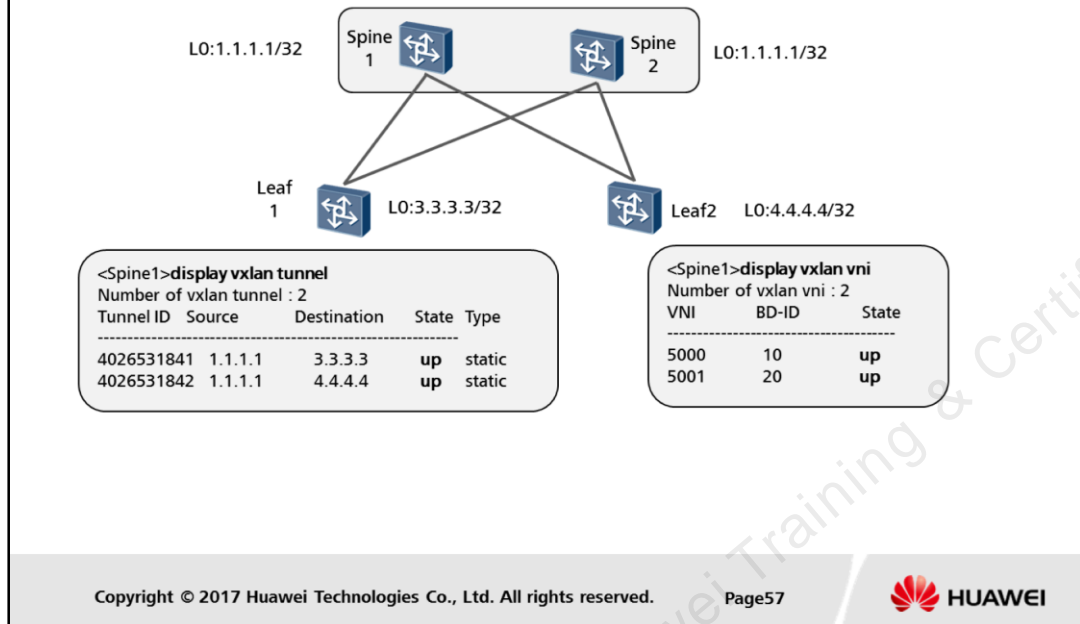
Page55

- After the configuration is complete, run the **display dfs-group 1 active-active-gateway** command on Spine1 and Spine2. If the state is shown "A:active", it means that the multi-active gateway connection is established.



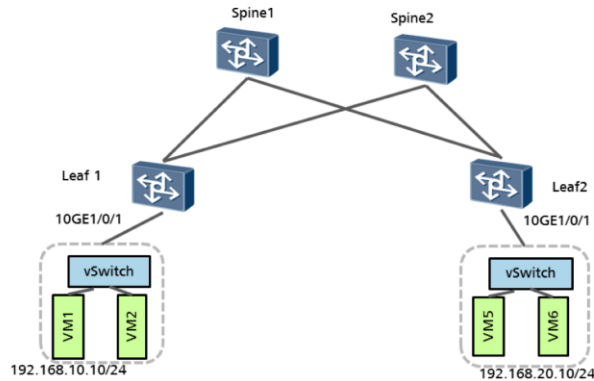
- As Spine 1 and spine2 is working in multi-active DFS group, the loopback 0 configured on both spines must be the same IP. Leaf will see them as 1 device.

## Step 3: Configure VxLAN tunnel on Leaf and Spine - Verification



- Example above shows the verification done on Spine1.
- After VxLAN tunnels are established, run the **display vxlan tunnel** command to check tunnel information
- After a VxLAN is configured, to check the VNI status and BD to which the VNI is mapped, run the **display vxlan vni** command. The command output helps you determine whether the VxLAN is correctly configured.

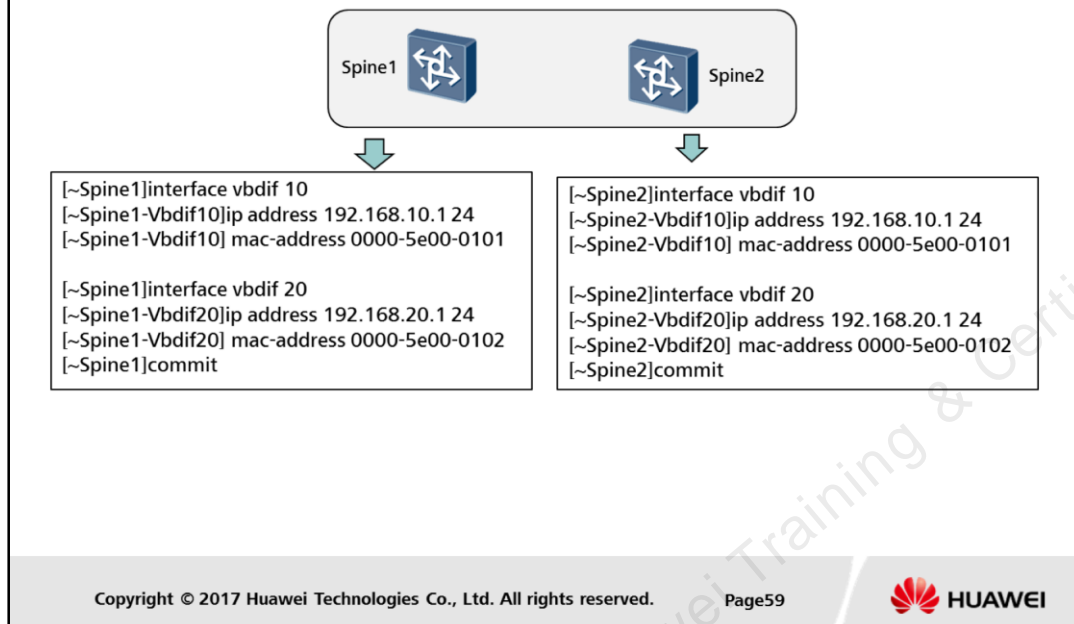
## Step 4: Configure Service Access Points at Leafs



```
[~Leaf1]vlan 10
[~Leaf1]interface 10ge1/0/1
[~Leaf1-10GE1/0/1]port link-type trunk
[~Leaf1-10GE1/0/1]undo port trunk allow-pass vlan 1
[~Leaf1-10GE1/0/1]port trunk allow-pass vlan 10
[~Leaf1]bridge-domain 10
[~Leaf1-bd10]l2 binding vlan 10
[~Leaf1]commit
```

```
[~Leaf2]vlan 10
[~Leaf2]interface 10ge1/0/1
[~Leaf2-10GE1/0/1]port link-type trunk
[~Leaf2-10GE1/0/1]undo port trunk allow-pass vlan 1
[~Leaf2-10GE1/0/1]port trunk allow-pass vlan 20
[~Leaf2]bridge-domain 20
[~Leaf2-bd20]l2 binding vlan 20
[~Leaf2]commit
```

## Step 5: Configure VxLAN L3 Gateway on Spines



- The configuration on Spine 1 and 2 must be same because they are working in active-active gateway group.



## Configuration Verification

- Once all the configuration completed corrected, VM1 and VM5 in different network segment can ping to each other, meaning that the inter-network segment communication is successful.
- You can also check the MAC address learnt in VXLAN tunnel using the command "display mac-address bridge-domain xx"; example is shown below

```
[~Spine1]dis mac-address bridge-domain 20
Flags: * - Backup
BD   : bridge-domain
-----
MAC Address  VLAN/VSI/BD      Learned-From  Type
-----
5451-1b84-0318  -/-/20          4.4.4.4      dynamic
-----
```

- The **display mac-address bridge-domain** command displays MAC address entries in a specified bridge domain (BD).

## Summary

- As a summary for this topic, we have covered:
  1. VxLAN overview including how VxLAN solves issues in traditional DCN networking
  2. VxLAN basic concepts, terms and forwarding models.
  3. VxLAN application in SDN AC-DCN network including VM communication and fabric network
  4. VxLAN configuration examples in SDN AC-DCN Cloud Fabric Network

**Thank you**

[www.huawei.com](http://www.huawei.com)

Huawei Training & Certification Huawei Training & Certification



## Recommendations

- Huawei Learning Website
  - <http://learning.huawei.com/en>
- Huawei e-Learning
  - <https://ilearningx.huawei.com/portal/#/portal/ebg/51>
- Huawei Certification
  - [http://support.huawei.com/learning/NavigationAction!createNavi?navId=\\_31&lang=en](http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=en)
- Find Training
  - [http://support.huawei.com/learning/NavigationAction!createNavi?navId=\\_trainingsearch&lang=en](http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=en)



## More Information

- Huawei learning APP

