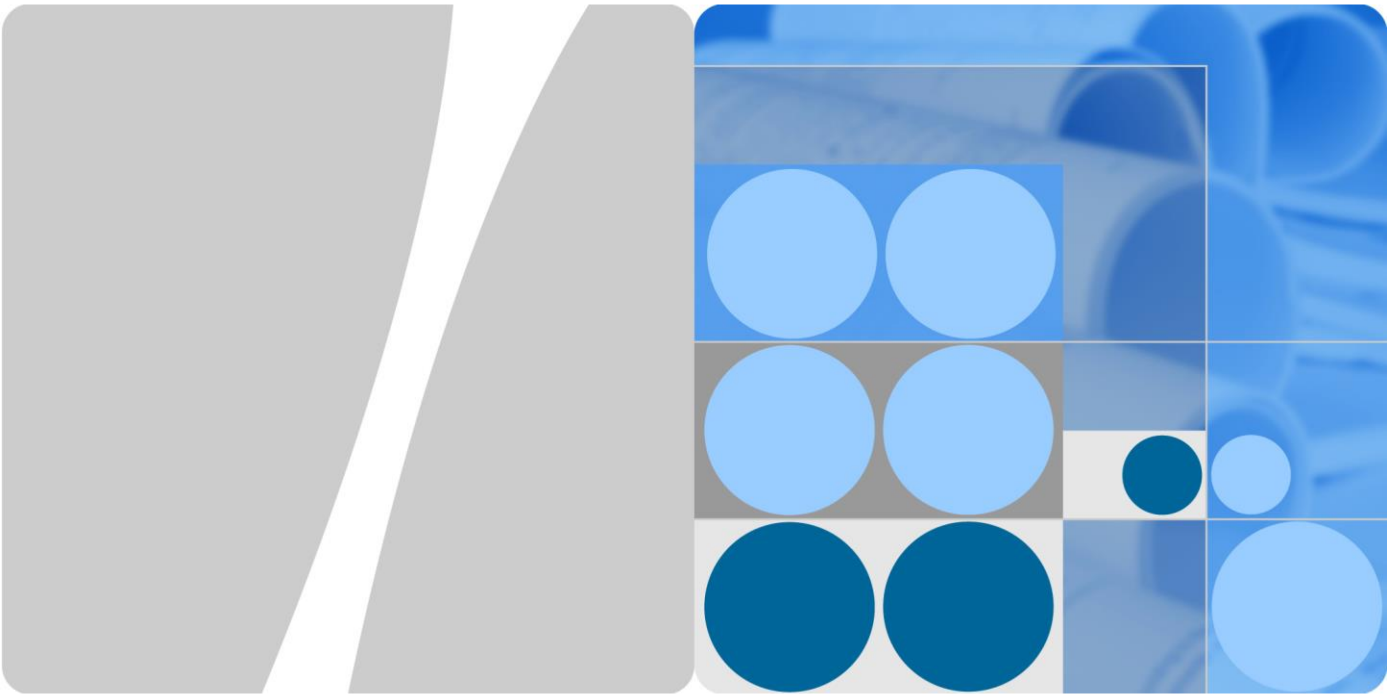


Segment Routing Technology White Paper

Issue	**01**
Date	**2018-08-24**

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

Notice

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

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

Huawei Technologies Co., Ltd.

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

Website: http://e.huawei.com

Contents

1 Segment Routing

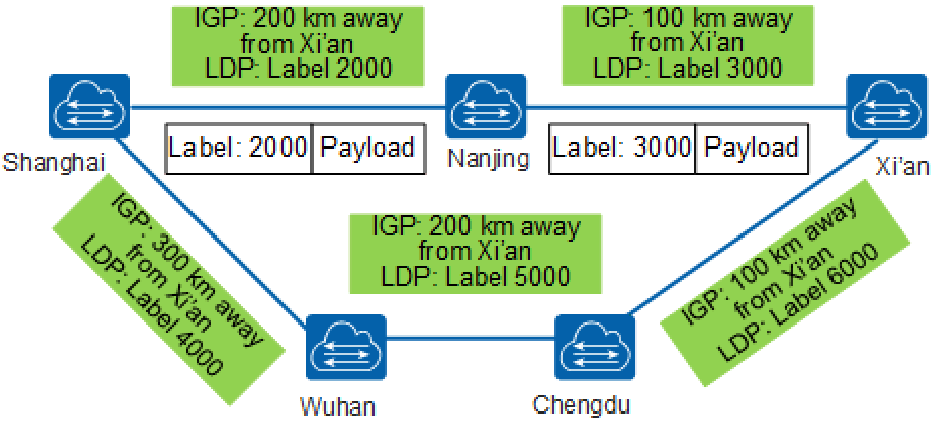
In the last issue of our Red Treasure Book, we introduced a very popular overlay technology — VXLAN, which is applied to data center networks. VXLAN technology enables Layer 2 interconnection on a data center network. However, there are many problems to be solved in terms of multi-DC interconnection.

In traditional data center interconnection solutions, enterprises focus on service isolation and traffic optimization. Therefore, MPLS technologies such as Resource Reservation Protocol-Traffic Engineering (RSVP-TE) or Label Distribution Protocol (LDP) are used no matter whether DCI networks are deployed by enterprise themselves or leased from carriers.

Segment Routing technology that we are going to introduce is an improvement to the preceding two types of main MPLS technologies. Before starting introduction, let's look at limitations of LDP and RSVP-TE.

2 Limitations of the Traditional MPLS Technologies

First, let's figure out how LDP works. In terms of forwarding behavior, LDP is the same as normal IP forwarding and it replaces IP addresses with labels as forwarding identifiers.


IGP: 200 km away from Xi'an LDP: Label 2000
IGP: 100 km away from Xi'an LDP: Label 3000
Shanghai
Label: 2000 Payload
Nanjing
Label: 3000 Payload
Xi'an
IGP: 300 km away from Xi'an LDP: Label 4000
IGP: 200 km away from Xi'an LDP: Label 5000
IGP: 100 km away from Xi'an LDP: Label 6000
Wuhan
Chengdu

As shown in the preceding figure, assume that a packet is being sent from Xi'an to Shanghai. First of all, all routers on the network must run the Interior Gateway Protocol (IGP) routing protocol. All nodes obtain corresponding routes based on cost values calculated using IGP. After obtaining routes, LDP starts distributing labels. For example, if we want to go to Xi'an from Shanghai, IGP will select the shortest path Shanghai — Nanjing — Xi'an for data forwarding. After the three cities have run the LDP protocol, Nanjing uses LDP signaling packets to inform Shanghai that there are 200 km between Nanjing and Xi'an, and packets labeled with 2000 can be sent to Nanjing directly. Similarly, Xi'an will inform Nanjing that packets labeled with 3000 are required. Therefore, Shanghai sends a packet labeled with 2000 to Nanjing, which receives the packet and figures out its destination, and replaces the label 2000 with 3000 before sending it to Xi'an.

We can see from it that LDP only distributes labels to forwarding packets based on their destination IP addresses (included in routing information), and then informs neighbors of mappings between destination IP addresses and labels. Neighbors will label packets according to their destination IP addresses. That is, after receiving a packet, a local device directly forwards it according to the label. Therefore, this forwarding behavior does not differ much from IGP forwarding, except that LDP uses labels instead of IP addresses.

Here, we can summarize characteristics of LDP as follows:

1. The forwarding mechanism of LDP depends on IGP completely and does not maintain any status. The forwarding behavior is similar to IGP.

2. The label is valid locally. For packets sent to Xi'an in the preceding example, the label is 2000 in Shanghai and 3000 in Nanjing.

The advantages of LDP are as follows:

- ECMP-supported: As described above, LDP relies on IGP to forward packets. Therefore, LDP supports ECMP naturally.
- Good scalability: The LDP protocol is simple and does not need to maintain any status, totally dependent on IGP. Therefore, the scale of an LDP-enabled network can be large. (Note that the scale of an SR-enabled network may be larger than that of an LDP-enabled network.)
- Simple configuration: The LDP configuration is simple. After IGP is configured, enable the LDP function on the interface so that the interface can send LDP signaling packets. No other configuration is required. There are only two core commands.

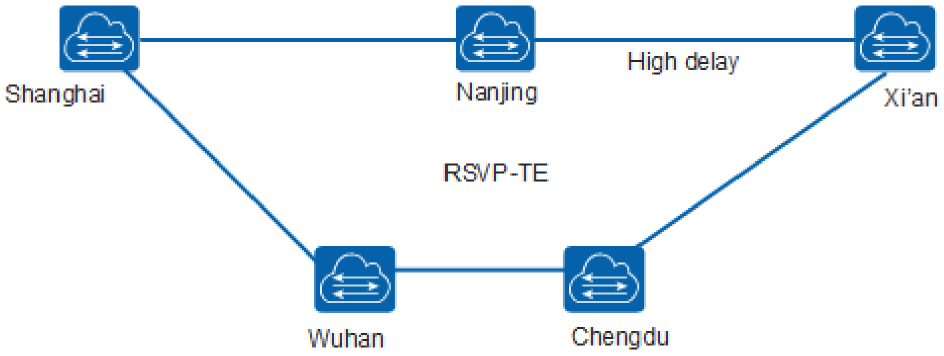
The disadvantages of LDP are as follows:

- Traffic blackhole: It is one of the enduring problems for which LDP has always been criticized. As shown in the preceding figure, in normal cases, after IGP calculates routes and LDP distributes labels, traffic is sent from Shanghai to Nanjing. However, in some cases, LDP connections are faulty. As a result, packets are not labeled, but IGP still sends them to Nanjing. Because no routes exist around the Nanjing node (like the BGP route blackhole, in which case no routes exist at both ends connected to a P node), packets can only be forwarded according to labels. In this case, packets will be discarded after they reach Nanjing.

  This problem can be resolved through association between LDP and IGP. That is, after LDP connections are faulty, IGP is informed of the fault. However, this method is a remedial measure and is not the fundamental way to solve the problem.
- No traffic optimization method: Since the forwarding mode of LDP is the same as IP forwarding, the path is selected only based on the cost value, but not based on other complex conditions such as bandwidth and latency.

Now, let's move on to RSVP-TE. Before the introduction, we need to know a concept called source route.

During common IP addressing or LDP label addressing, each device along a path calculates routes based on the destination IP address or the label. Devices only focus on the outbound interface of their next hop, but do not care about the subsequent traffic path. In the RSVP-TE mode, however, the forwarding model is based on the source route. That is, the path to the destination is determined when the traffic enters an RSVP-TE network.



Again, from Shanghai to Xi'an in the preceding example, LDP selects the relatively short route Shanghai — Nanjing — Xi'an based on the cost value calculated using IGP. If the link delay between Nanjing and Xi'an is high, service requirements cannot be met when traffic is transmitted through this path. According to the traffic engineering protocol, an end-to-end path that meets the current service requirements is determined when traffic is transmitted from the start node Shanghai.

In this case, we need some standards, such as latency and current bandwidth utilization, to measure whether a link meets the service requirements. According to these standards, the device designs a path on the source node. The path is not determined only based on the cost value calculated using IGP. Above is the basic principle of traffic engineering based on the source route mechanism.

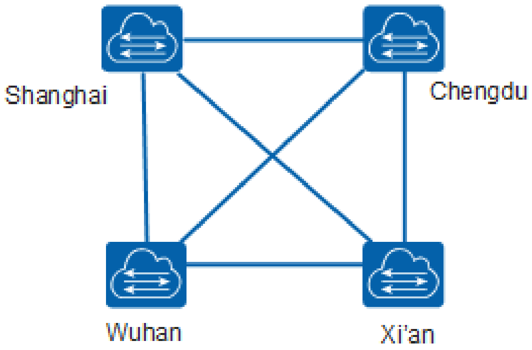
The following describes how RSVP-TE works.

1. Collect network-wide link information.

A proper link can be selected only when all link information is available. RSVP-TE extends the IGP protocol and carries the link information in IGP signaling packets. Details of IGP extension are not provided here. Intermediate system to intermediate system (IS-IS) extends some type-length-values (TLVs) and Open Shortest Path First (OSPF) uses several special link-state advertisements (LSAs) to carry link information. Then, the link information is synchronized on the entire network through the status synchronization mechanism of IGP. In this way, all devices share the network-wide link database. Note that the link database here is not the link state database (LSDB) that we often refer to. The LSDB functions only as a map, but the TE link information database includes both a map and real-time route information on the map.

2. Create tunnels at the start point and specify requirements of tunnels on routes.

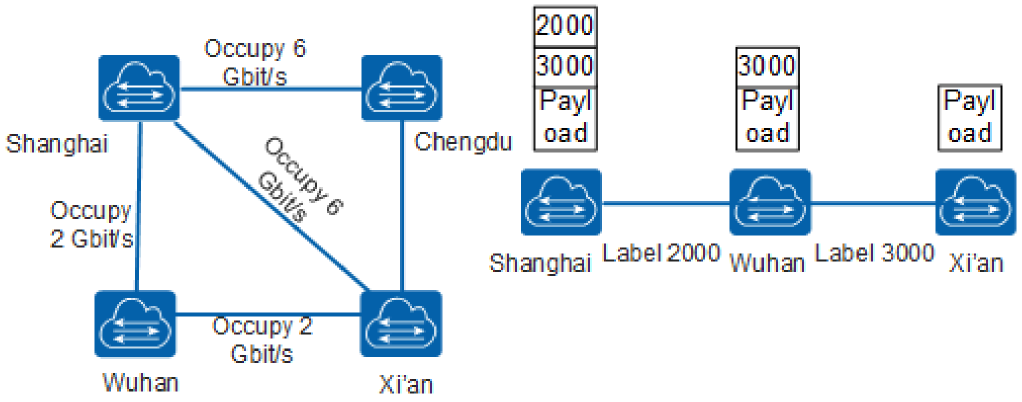
After the link information is collected, packets can be sent.



On the RSVP-TE network consisting of four nodes shown in the preceding figure, each node creates an RSVP-TE tunnel to another node. Requirements for each tunnel need to be specified, such as the bandwidth to be reserved and the delay. Because traffic is transmitted between a local node and all the other nodes, a total of 1/2*N* (N-1) tunnels need to be created if there are N nodes, and the routes for each tunnel must be configured. Therefore, the RSVP-TE configuration workload is heavy and RSVP-TE has not been widely used.

3. Route selection, label distribution, and forwarding

After tunnels are configured, route selection and packet forwarding are performed based on the current link condition.



As shown in the preceding figure, the bandwidth is 10 Gbit/s for all links between Xi'an and Shanghai. After tunnels are configured, at least 6 Gbit/s bandwidth is required. After checking link information, it is found that only the route Shanghai — Wuhan — Xi'an can meet the requirement.

Shanghai sends an RSVP request to Wuhan to ask for 6 Gbit/s bandwidth,

and Wuhan sends a similar RSVP request to Xi'an to obtain its permission.

Xi'an returns an RSVP signaling packet to Wuhan, indicating permission only if packets from Shanghai are labeled with 3000.

Similarly, Wuhan returns an RSVP signaling packet to Shanghai, indicating permission only if packets from Shanghai are labeled with 2000.

Then, Shanghai labels packets with 2000 and sends them to Wuhan. Wuhan replaces labels 2000 with 3000 for any packet destined for Xi'an.

The advantages of RSVP-TE are as follows:

- Flexible route selection RSVP-TE can flexibly adjust the traffic forwarding path according to the link status changes on the network. In contrast, IGP and LDP do not provide this function, but only select paths based on the cost values, which may cause congestion on a link that has a lower cost.

The disadvantages are as follows:

- Complex configuration: As mentioned above, RSVP-TE tunnels must be manually configured, and each tunnel has specific link requirements. The configuration is complex. The following presents how to configure RSVP-TE:

1. Configure basic MPLS capabilities and enable MPLS TE, RSVP-TE, and CSPF.

```
[~LSRA] mpls lsr-id 1.1.1.9
[*LSRA] mpls
[*LSRA-mpls] mpls te
[*LSRA-mpls] mpls rsvp-te
[*LSRA-mpls] mpls te cspf
[*LSRA-mpls] quit
[*LSRA] interface vlanif 100
[*LSRA-Vlanif100] mpls
[*LSRA-Vlanif100] mpls te
[*LSRA-Vlanif100] mpls rsvp-te
[*LSRA-Vlanif100] quit
[*LSRA] commit
```

2. Configure OSPF TE.

```
[~LSRA] ospf 1
[~LSRA-ospf-1] opaque-capability enable
[*LSRA-ospf-1] area 0
[*LSRA-ospf-1-area-0.0.0.0] mpls-te enable
[*LSRA-ospf-1-area-0.0.0.0] quit
[*LSRA-ospf-1] quit
[*LSRA] commit
```

3. Configure MPLS TE attributes on outbound interfaces of all nodes.

```
[~LSRA] interface vlanif 100
[~LSRA-Vlanif100] mpls te link administrative group 10001
[*LSRA-Vlanif100] quit
[*LSRA] commit
```

4. Create an MPLS TE tunnel on an ingress node.

```
[~LSRA] interface tunnel 1
[*LSRA-Tunnel1] ip address unnumbered interface loopback 1
[*LSRA-Tunnel1] tunnel-protocol mpls te
[*LSRA-Tunnel1] destination 4.4.4.9
[*LSRA-Tunnel1] mpls te tunnel-id 100
[*LSRA-Tunnel1] mpls te record-route label
[*LSRA-Tunnel1] mpls te affinity property 10101 mask 11011
[*LSRA-Tunnel1] quit
[*LSRA] commit
```

The above configuration is only for one tunnel on one device. If a small RSVP-TE network consists of four devices, you need to perform the configuration for five times. Even if you can use some configuration tools, it is difficult to maintain such a large quantity of tunnels, and the subsequent expansion is also very difficult.

- Complex protocol that is not suitable for large-scale deployment

Besides complex configuration, the RSVP-TE protocol mechanism is also complicated. All routers on the network need to maintain a large link information database. Therefore, the scale of an RSVP-TE network cannot be expanded, otherwise related work is very troublesome.

- ECMP not supported

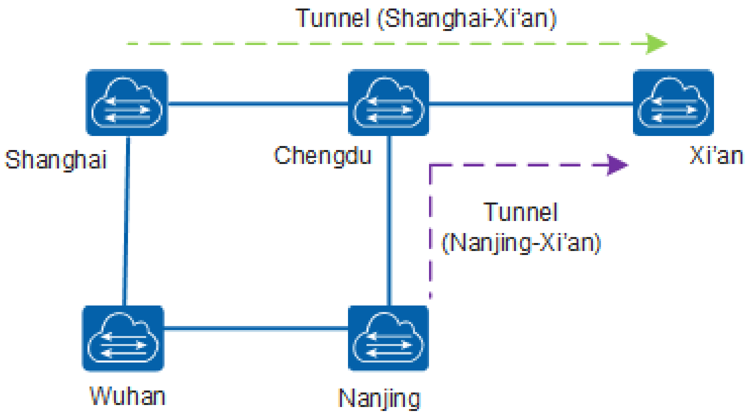
According to the RSVP-TE working principle, RSVP calculates paths based on link information, sets up tunnels, and distributes labels, not depending on IGP. RSVP-TE cannot perform equal-cost multi-path routing (ECMP) as IGP. If load balancing is required, another tunnel with the same source and destination IP addresses as the original one must be created. This is definitely complicated.

3 Take the Essence and Discard the Dregs

The preceding chapters elaborate largely on LDP and RSVP-TE, which are very necessary because Segment Routing combines the advantages of the two MPLS technologies. From a comprehensive perspective, LDP is like an energetic teenager who behaves casual and does not maintain any status. Forwarding is connectionless, and traffic can pass anywhere. In this manner, LDP is easy to deploy and supports load balancing. RSVP-TE, however, is like a rigorous middle-aged person, who plans the optimal route before each traveling and takes all situations into consideration. In this case, routing is reliable, but is inevitably inefficient. So, how can we be rigorous while improving efficiency? This is the goal of the Segment Routing protocol, which can be achieved through the following steps.

Step 1: Remove the RSVP-TE signaling mechanism.

The root cause of the complexity of RSVP-TE is that each device on a network needs to maintain a set of complex signaling. Can the signaling mechanism be removed? To answer this question, let's see why RSVP-TE requires signaling.



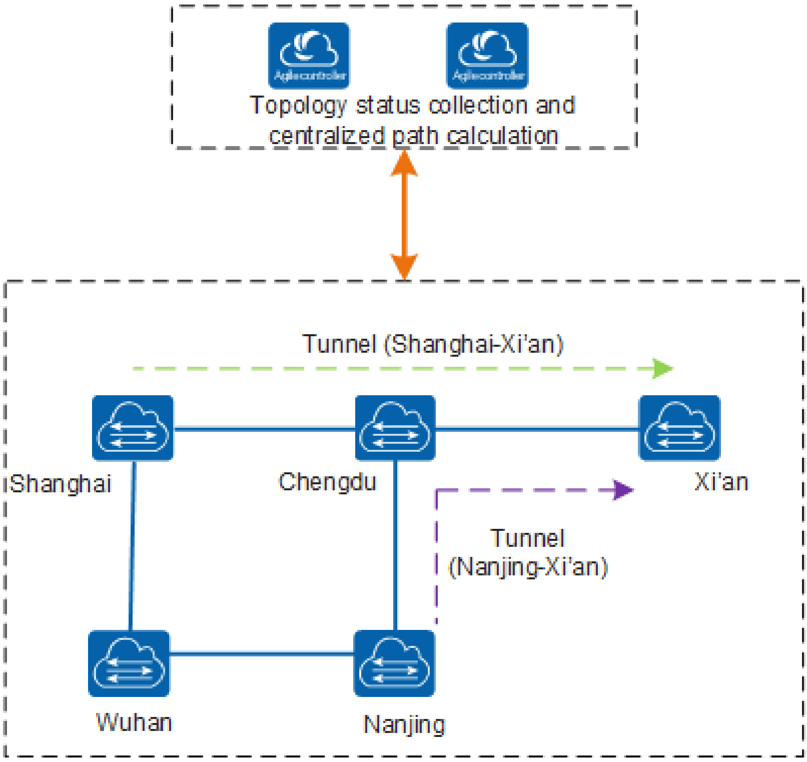
As mentioned above, after RSVP-TE obtains path information through the extended IGP, it calculates an optimal path to the destination, and then sends RSVP-TE signaling packets to establish tunnels. Take the route from Shanghai to Xi'an as an example. How can we remove the RSVP-TE signaling? That is, how can Shanghai send traffic to Chengdu without sending RSVP-TE signaling packets after calculating the path?

The answer is no. In the above figure, for example, Shanghai – Chengdu – Xi'an is an appropriate path for a packet to be sent from Shanghai to Xi'an. Can Shanghai directly encapsulate the packet and send it to Chengdu? No, because Shanghai is not 100% sure about whether the current link changes. At this moment, if a tunnel is generated between Nanjing and Xi'an and a certain amount of bandwidth is occupied, the route Shanghai – Chengdu –

Xi'an may not meet the service requirements. However, network convergence takes time. During path calculation, Shanghai does not receive related information.

Therefore, a mechanism is required to confirm the path and reserve bandwidth before packets are sent over each tunnel. This is the function of RSVP-TE signaling.

The reason for the above problem is that RSVP-TE uses a distributed architecture. Each device can only view its own status, and must use the signaling mechanism to know the status of other devices. If we change the distributed architecture into a centralized one by adding a centralized control node to calculate paths and distribute labels in a unified manner, the problem will be solved.

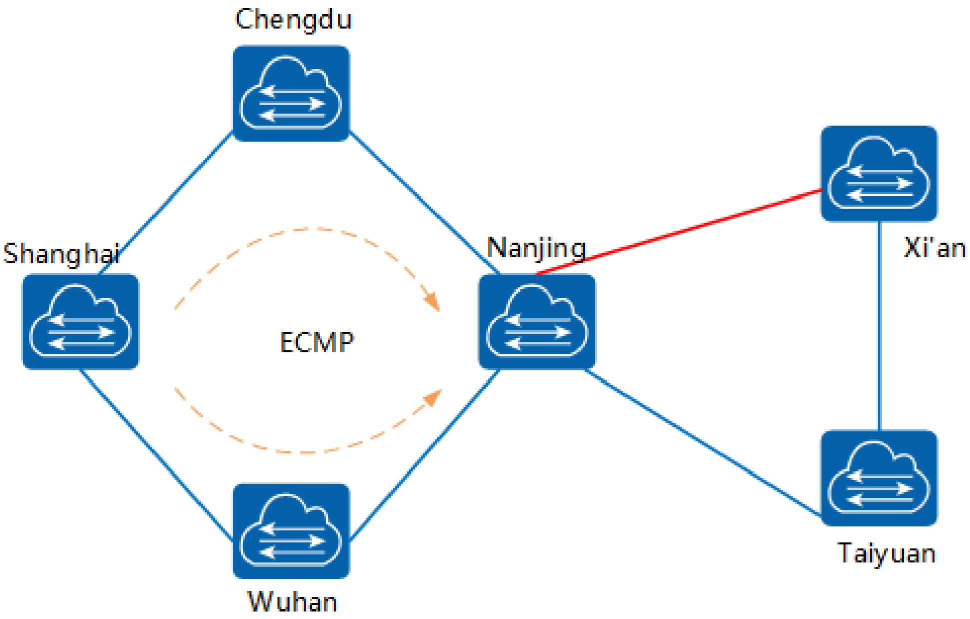


In this case, regardless of whether the route is from Shanghai to Xi'an or from Nanjing to Xi'an, all nodes request paths from the controller. The controller arranges routes in a unified manner, so the preceding problem will not occur.

This centralized architecture also fits the architecture of current SDN networks. Therefore, most of the centralized control nodes of Segment Routing are deployed on SDN controllers currently. It can be said that Segment Routing and SDN perfectly match each other. After a controller is deployed, the manual tunnel configuration can be omitted. The controller collects the link information configuration, which greatly reduces the configuration workload and deployment difficulty.

Step 2: Apply high efficiency and load balancing of LDP to RSVP-TE.

As mentioned above, RSVP-TE is like a rigorous man. It uses the source routing mechanism, which determines the path direction at the source end and specifies each hop along the path. This mechanism improves reliability, but is quite inefficient for traffic engineering. In addition, in cases of multiple paths meeting requirements, RSVP-TE does not support load balancing, which is a waste of bandwidth. Can high efficiency and load balancing of LDP be applied to RSVP-TE?



As shown in the preceding figure, according to path calculation, the link from Nanjing to Xi'an is congested and other links are available from Shanghai to Xi'an. If ECMP is used, only either of the following paths can be selected: Shanghai – Wuhan – Nanjing – Taiyuan – Xi'an, or Shanghai – Chengdu – Nanjing – Taiyuan – Xi'an. There is no doubt that a link is wasted. Can we adopt flexible load balancing like LDP between Shanghai and Nanjing and strictly follow the RSVP-TE mode to specify the path from Nanjing to Xi'an?

How does LDP implement load balancing? As mentioned above, LDP does not maintain any status information. In fact, LDP functions as a second transmitter of IGP to map each destination IP address to an MPLS label. Although label forwarding is used, it is essentially an IP forwarding model. A label is actually another manifestation of an IP address.

The LDP traffic blackhole has been criticized for a long time because LDP is just a second transmitter, whose status cannot be consistent with that of IGP. We wonder how about we remove the second-transmitter role from Segment Routing while retaining IP forwarding.

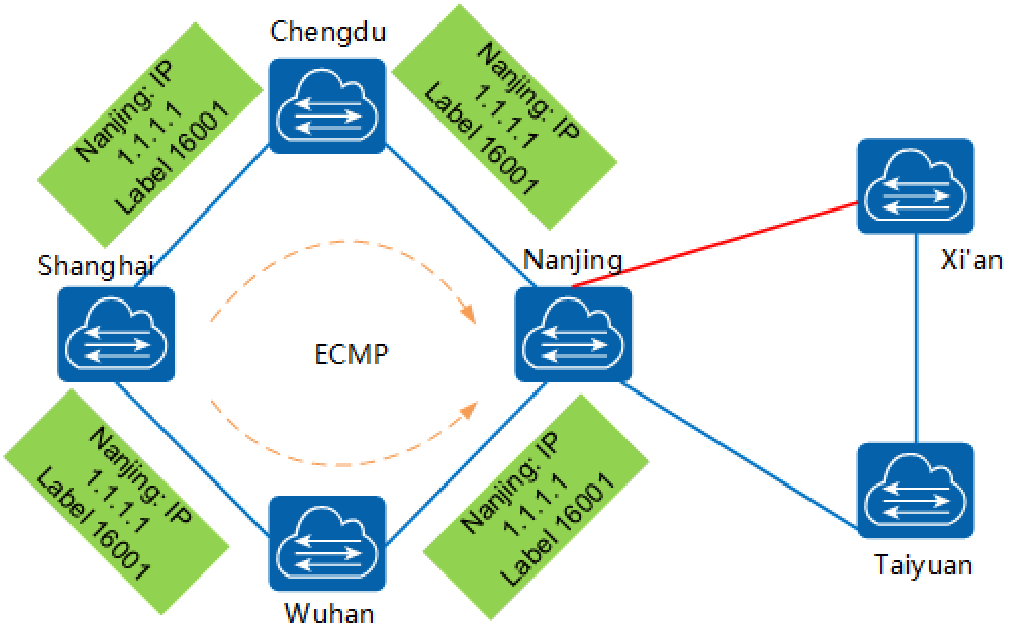
Several methods are used to optimize Segment Routing:

1. Use IGP to distribute labels.

Segment Routing directly extends IGP and carries label information through IGP signaling. In this way, the problem of traffic blackhole due to the second-transmitter role of LDP can be solved.

2. Set a node ID (global label).

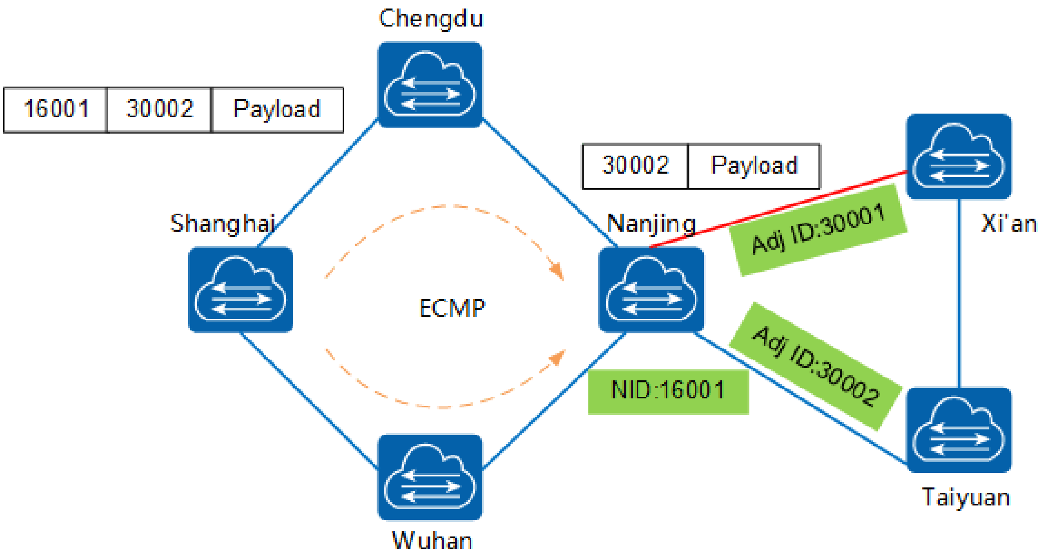
In LDP, labels are valid locally. In Segment Routing, a node ID is set to identify a device and is globally valid and unique. It can be seen as the loopback IP address of a device. Packets can be forwarded by looking up the route table based on a node ID, in the same way as IP forwarding through a global identifier. In this way, ECMP can be implemented.


Chengdu
Nanjing: IP 1.1.1.1 Label 16001
Nanjing: IP 1.1.1.1 Label 16001
Shanghai
Nanjing
Xi'an
ECMP
Nanjing: IP 1.1.1.1 Label 16001
Nanjing: IP 1.1.1.1 Label 16001
Taiyuan
Wuhan

For example, the IP address of Nanjing is 1.1.1.1, and the distributed global node ID is 16001. After receiving packets with the label 16001, a node such as Shanghai looks up the route table and forwards the packets to Nanjing through ECMP.

3. Set an adjacent ID.

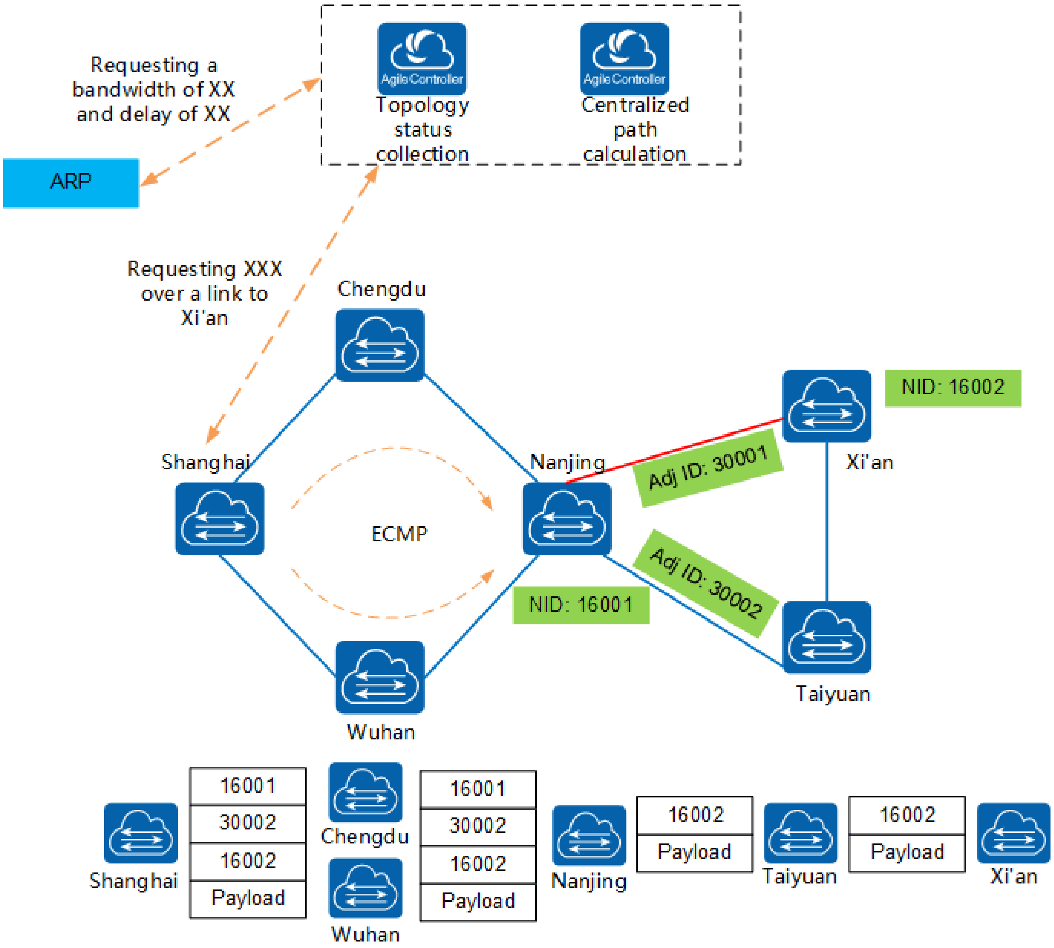
The ECMP function can be implemented with the node ID. However, if packets are forwarded only through the node ID, the optimal path cannot be selected in the same way as RSVP-TE. As shown in the following figure, if we only specify the next hop of Nanjing as Xi'an, Nanjing will select the route Nanjing – Xi'an according to the OSPF principle of IGP. However, this route is congested, represented by the red link in the figure. What should we do?



Segment Routing introduces the concept of adjacent ID. A node ID identifies a device globally and uniquely, and can be used for forwarding like a loopback IP address. An adjacent ID is valid locally and uniquely identifies a local link.

As shown in the preceding figure, Nanjing distributes the label 30001 to the red link and the label 30002 to the blue link. If you do not want traffic to be transmitted along the red link from Nanjing, add the label 16001 (NID of Nanjing) and label 30002 (adjacent ID of the blue link) to packets at the source end (Shanghai). After packets arrive in Nanjing, Nanjing removes the label 16001 and discovers the label 30002 indicating the blue link in the figure. Therefore, packets are sent out along the blue link.

4 Segment Routing Workflow



The Segment Routing end-to-end workflow is as follows:

1. IGP or BGP is running on a network to ensure that routes on the entire network are reachable.

2. A node ID is configured for each node on the network. IGP or BGP is used to advertise the node ID to ensure that the node ID is globally unique and reachable on the entire network.

3. Adjacent IDs are generated and can be configured manually or assigned by the controller.

4. The controller collects the network-wide topology and label information from devices. This information is reported to the controller through IGP or BGP.

5. An application sends a request to the controller, requesting the network to provide a link whose bandwidth is not lower than XX Mbit/s and delay is no longer than XX ms. After receiving the request, the controller informs Shanghai of a traffic flow from Shanghai to Xi'an, requesting a bandwidth of XX and delay of XX.

6. Shanghai requests path calculation to the controller. Based on the network-wide topology and label information, the controller calculates and concludes that the path from Nanjing to Xi'an (link in red) is congested and the optimal path should be Shanghai – Nanjing – Taiyuan – Xi'an.

7. The controller delivers a label stack (16001, 30002, 16002) to Shanghai.

8. After receiving the label stack, Shanghai sends the packet to Chengdu and Wuhan based on the first-hop label 16001. After receiving the packet, Chengdu and Wuhan send the packet to Nanjing based on the label 16001.

9. After Nanjing receives the packet, it finds that the outer label 16001 identifies itself and pops out the label. Nanjing then finds that the inner label 30002 identifies the adjacent ID of itself and pops out it too, and sends the packet along the link corresponding to 30002.

10. After Taiyuan receives the packet, it sends the packet to Xi'an according to the label 16002. The entire forwarding process is completed.

On a general basis, Segment Routing uses LDP/IGP for most nodes on a network, and forwards traffic based on node IDs. Load balancing can be implemented without specifying a path. For some special links, adjacent IDs are used to implement the RSVP-TE principle, which specifies a path. In this way, traffic engineering in complex situations can be achieved.

In addition, Segment Routing assigns the path calculation work on the control plane to the controller. This greatly alleviates the burden on devices, reduces the device configuration workload, and improves the scale and scalability of the network. The working principle of Segment Routing is actually the same as that of SDN.

5 Summary

This document describes the principles of Segment Routing using several MPLS technologies. Currently, Segment Routing is a new technology and needs to be further developed in terms of application scenarios and technology details. For example, end-to-end SR is available now. SR is used on underlay networks and EVPN runs on overlay networks. Even the Linux OS has announced to support SR. It can be said that SR is a very promising technology. We will further introduce application scenarios and advanced technologies of SR.