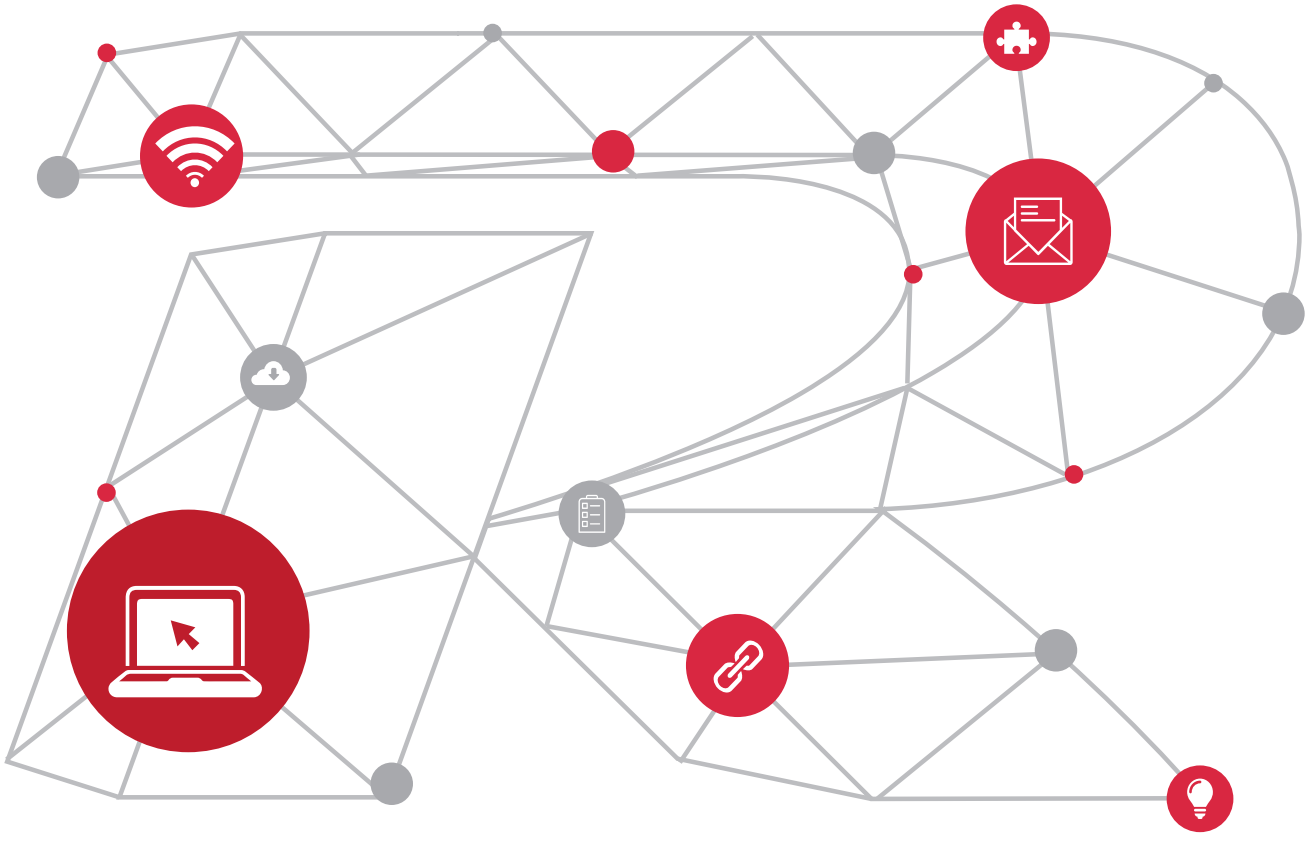
White Paper

A Feast of Technologies Optimized BGP Features in DCN Scenarios



Contents

Preface

As the scale advantages of ultra-large Internet data centers (IDCs) become more prominent, especially in the IPv4/IPv6 dual-stack mode, network engineers are facing growing pressure on construction and maintenance. In the article **Planning BGP Networks for Large IDCs**, we discussed the large-scale deployment of Border Gateway Protocol (BGP) in IDCs, which can greatly improve routing performance and simplify network planning. However, IDC networks are different from conventional Wide Area Networks (WANs). Requirements for BGP deployment and maintenance are also different from those on WANs. In this case, optimizing BGP can further improve routing performance and simplify maintenance. This article shares optimized BGP features in IDC scenarios based on the data center network (DCN) construction experience of Xiao Li, an engineer at an Internet company.

I am Xiao Li, a handsome guy.

When I was in college, I fought wits with Ruijie's authentication and accounting system deployed on the campus network for several years in order to save network fees. Then, I fell in love with networks even though I majored in law, and finally obtained the Ruijie Certified Internetwork Expert (RCIE) certificate. After graduation, I entered an Internet company, dealing with construction, planning, configuration, and changes of all kinds of networks every day. And now I become a really experienced old hand.

The following tells my story. Please listen to me carefully.

Network Construction

One morning, as soon as Xiao Li entered the office, he received an important task from his boss: to build an IDC that can serve more than 50,000 servers and run the servers in IPv4/IPv6 dual-stack mode. Xiao Li was asked to give a detailed network design and plan first. During network planning, in addition to physical networking, routing and address planning are relatively complex. Having read **Planning BGP Networks for Large IDCs** published by Ruijie, Xiao Li had no doubts about routing protocol selection and planning, but felt agitated when it came to the large number of interface addresses and management addresses in the dual-stack mode.

Xiao Li was facing the following urgent tasks:

* Run servers in the dual-stack mode. This means that the dual-stack mode must be enabled for the network.

* Plan IPv4 and IPv6 interface addresses and management addresses for a large number of devices.

* Configure IPv4 and IPv6 neighbors based on BGP.

Certainly, these tasks can be done by using the conventional configuration method. However, as an innovative engineer of a leading Internet company, Xiao Li did not rush to make plans based on experience. He wondered whether a simpler plan was available. After communicating with manufacturers, Xiao Li adopted the plan offered by Ruijie Networks:

1. Set up sessions based on link-local addresses: This simplifies assignment of IPv6 addresses.

A link-local address is a new type of address introduced by the IPv6 stack. When IPv6 is enabled on an interface, a link-local address can be automatically configured using the link-local prefix FE80::/10. The address is valid for local links. Devices can establish multiple BGP neighbors based on the link-local address, freeing users from assignment of independent IPv6 addresses.

2. Activate dual-stack routing for each BGP session: This reduces BGP neighbors.

Users can set up sessions between BGP neighbors based on only IPv4 addresses or IPv6 addresses to transfer both IPv4/IPv6 dual-stack routes. This reduces neighbor entries on devices.

Xiao Li found that these two features were perfect in this scenario: Users can establish BGP neighbors based on IPv6 link-local addresses by simply specifying a neighbor interface, and then can activate IPv4/IPv6 dual-stack routing based on each neighbor. In this way, users can use a single IPv6 session to transfer IPv4/IPv6 dual-stack routes.

Xiao Li's plan was approved right away after being submitted. Then, he immediately got down to construction. While bringing the servers online in batches, Xiao Li received a new task: Plan a point of delivery (POD) in the IDC, where servers of the POD must run Docker containers, and hosts must exchange routes with top-of-rack (ToR) switches based on BGP. At that time, network segments for hosts had been planned, but IP addresses would be available not until the business team brought the hosts online.

Xiao Li was shocked because this meant that he had to configure connections to BGP neighbors every time the business team brought a host online. The business team usually brings hosts online in off-peak hours. Xiao Li wondered if he had to work overtime every day only to spend less than one minute to establish one BGP neighbor.

It is fine working overtime, but not when the work has little value. Therefore, Xiao Li started thinking of establishing BGP neighbors based on the planned network segments. If this method was feasible, there would be no need to work overtime.

Xiao Li looked up the device manual and found the following feature:

Passive Session Setup Based on Network Segments

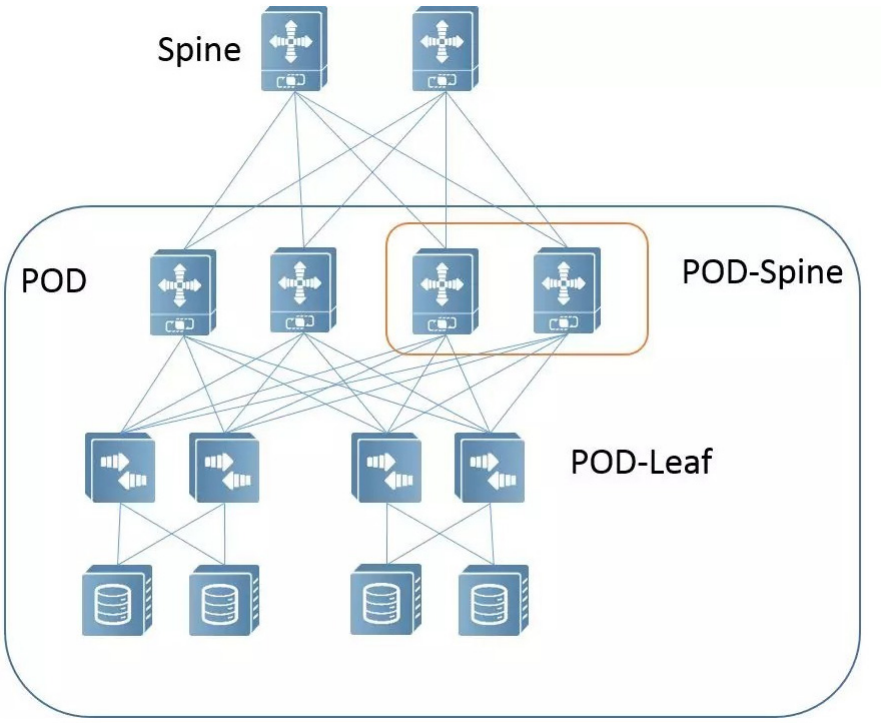
A network device can configure a BGP neighbor based on the network segment. After configuring this mode, the network device does not actively initiate BGP neighbor establishment requests. Instead, it passively receives the establishment requests initiated by the peer neighbor. Then, it generates a real neighbor based on the neighbor's address and sets up a session.

It was eight o'clock on the evening, and Xiao Li had completed the configuration. Xiao Li did not need to work overtime and just happily clocked off. There was even enough time for an exercise. He was satisfied.

Network Expansion

One day, Xiao Li received a new task from his boss: The company wanted to launch a batch of artificial intelligence (AI) services recently. Though service scale was small, the convergence of the original network was relatively high and may fail to meet the requirements for high-performance computing. Therefore, Xiao Li was required to expand one POD to reduce the convergence ratio without affecting existing online services. Figure 1 shows the expanded network architecture.

Figure 1: Network Architecture of the Expanded POD



After receiving the task, Xiao Li was delighted, because the spine-leaf architecture facilitates scale-out. He migrated and brought online the first POD-spine device as planned. However, the business team soon reported that a few of packets were lost. Although the packet loss is minor, Xiao Li pondered the cause for the fault.

He checked and found that the configurations were correct and route learning was normal. So what caused the packet loss? After careful analysis, Xiao Li found that the root cause was the difference in the installation of routing entries: After the new POD-spine device was brought online, the POD-spine device advertised its routes to POD-leaf devices and spine devices and learned the routes of the network. For a POD-leaf device, the number of equal-cost multi-path (ECMP) routing paths changes from two to three at once, and the POD-leaf device sent data traffic to the new POD-spine device. At this time, although the new POD-spine device had learned of network routes, installation of these routing entries took certain time and had not completed. As a result, a time difference existed between traffic sending and receiving, resulting in the lost packets. How to solve the problem? Xiao Li wondered whether this time difference could be eliminated if the device can learn routes and install the routing entries before advertising its complete routes to its neighbors. Thinking of this, Xiao Li looked up the device manual again and found the following feature:

Delayed Advertisement for BGP Route

BGP allows a device to install learned routes in hardware routing entries first and then advertise these routes to its neighbors.

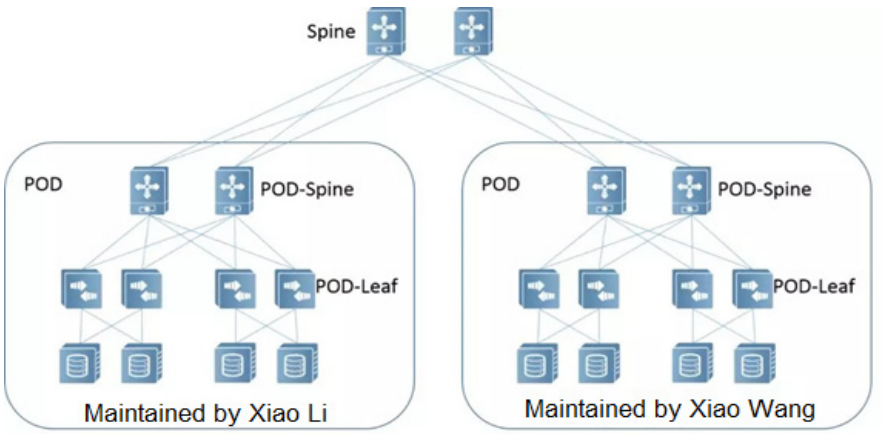
This feature could be the solution to the problem! With this idea in mind, Xiao Li immediately brought the second POD-spine device online and enabled this feature. He also asked colleagues from the business team to monitor the packet loss status of the server in real time. The second device did not cause any packet loss. Nail it!

Troubleshooting

Human fortunes are unpredictable as the weather, and bad things will happen anyway even if the probability is low.

Figure 2 shows the maintenance regions.

Figure 2: Maintenance regions



One day, engineer Xiao Wang told Xiao Li that he was complained by the business team for a packet loss of more than 10 seconds due to the failure and system breakdown of spine devices. Xiao Wang considered such packet loss as usual because the network included more than 10,000 routes and convergence was time-consuming. After all, the services were not interrupted. But then, Xiao Wang found that the maintenance region handled by Xiao Li was not affected by the fault. He was curious about the reason.

Xiao Li told Xiao Wang about his secret code in coping with the black swan event:

BGP PIC

Prefix independent convergence (PIC) provided in BGP allows the network to achieve convergence independent of the route scale. In this way, a large scale of routes can be quickly switched.

The PIC feature is implemented based on autonomous system (AS) numbers and enabled between External Border Gateway Protocol (eBGP) neighbors. After the PIC feature is enabled, BGP adds the PIC-extended community attribute to a route and then advertises the route. The switch that receives the BGP route allocates a unique index ID based on the AS number of the advertiser and the router ID. Then, the switch adds the index ID to the route and delivers the route to the forwarding plane after optimal computing. If all the uplinks of the advertiser are interrupted and cannot receive the routing information of this AS, the switch searches for the index ID and instructs the forwarding plane to switch the route associated with this ID. As such, quick convergence is achieved without the need to delete routes one by one for convergence. Common route convergence requires one-by-one deletion of failed routes and therefore is strongly affected by the route scale.

Simply put, for the routes advertised by a failed spine device, a POD-spine device classifies them by the private attribute of BGP, adds the private attribute to the routes, and then advertises the routes to POD-leaf devices. If the spine device fails, the POD-spine device can quickly switch to other routes and use the private attribute to notify all POD-leaf devices that all its routes fail and they can switch to other routes. This is why quick convergence can be implemented within the POD handled by Xiao Li. This implementation is independent of prefixes and is highly applicable to switching of a large scale of routes. The following shows the results of convergence tests performed based on 12,000 routes:

* When PIC is disabled, the convergence time is 13s.

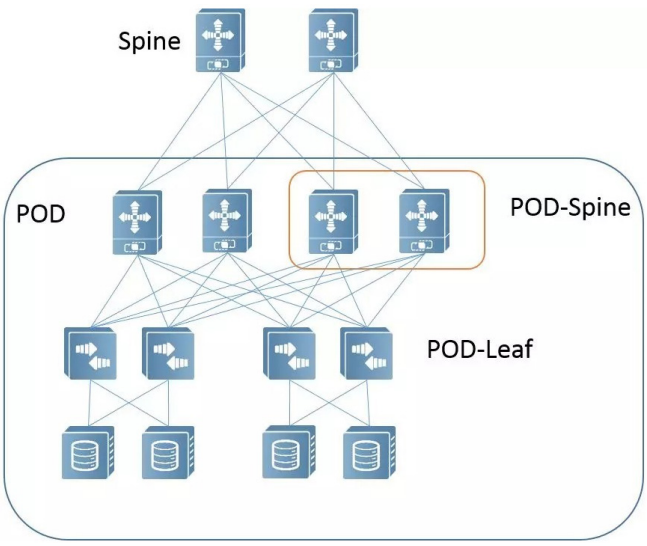
* When PIC is enabled, the convergence time is 0.7s. This convergence time does not change with the route scale, and therefore is applicable to scenarios with a large number of routes.

Xiao Wang was confused. If the attribute is private, it can be identified only by the specified device. Doesn't this affect route learning? The answer is no. Devices that do not support this attribute automatically filter out this attribute. Therefore, route learning of other devices is not affected. After expressing his admiration and gratitude to Xiao Li, Xiao Wang immediately went back to his workstation and applied this method to his maintenance region, which successfully reduced the loss caused by the black swan event.

Core Migration

Someone may ask how serious this cold winter of the Internet industry will be. Xiao Li may say that no matter how serious it is, there will always be new tasks. He was right. He received another task of building an IDC. Xiao Li immediately checked the network plan and the hardware, and found that two more POD-spine devices were needed. Should these devices be migrated from the old IDC? The answer was yes. The services of the old IDC were not deployed as expected. The traffic was not high. Therefore, Xiao Li needed to increase the convergence ratio and remove two POD-spine devices, without affecting the services.

Figure 3: Network Architecture of the Old IDC



Before removing a device, Xiao Li needed to migrate the traffic of the device first. To ensure that services are not affected, Xiao Li communicated with the device's manufacturer and learned of several methods for migrating BGP traffic.

1. Neighbor shutdown

This method sends a notification message to a neighbor to inform that the neighbor relationship has been shut down. It is usually used to change the isolation of a single line card in a frame-type device.

2. Graceful shutdown

This method sends an UPDATE message to a neighbor to advertise the route of the lowest priority together with the **gshut community** attribute to the neighbor. If the value of **local-preference** is **0** or the **MED** value is **4294967295**, the route has the lowest priority. The **gshut community** attribute enables the neighbor to update routes and switch traffic to a backup link or other ECMP paths in advance.

3. BGP advertise-map

This method sends an UPDATE message that carries the **withdraw routes** field to a neighbor, to inform the neighbor that the route is invalid. After receiving the UPDATE message, the neighbor updates the local routing table to delete all related routes. In this way, the traffic of the routes is switched to a backup link or other ECMP paths.

After comparative analysis on the principles of the traffic migration methods, Xiao Li summarized several differences, as described in Table 1.

Table 1: Summary of BGP Traffic Migration Methods

Isolation Method	Neighbor Shutdown	Graceful Shutdown	BGP advertise-map
TCP sessions torn down or not	Yes	Yes	No
Routes canceled or not	No	Yes	Yes
Flexibility	Low Depending on neighbors or the global network	Low Depending on neighbors or the global network	High Supporting selection of to-be-advertised routes based on Access Control Lists (ACLs)
Efficiency	Effective immediately	Low	Effective immediately
Application scenario	Isolation of a single line card in a frame-type device	Rarely used	Isolation of the whole device
Packet loss occurring in isolation or not	Yes, within 1s	No	No

After comparative analysis, Xiao Li found that: Neighbor shutdown is violent and causes packet loss. Graceful shutdown does not cause packet loss but leads to a long route convergence time. The BGP advertise-map method directly advertises invalid routes, which is convenient and causes no packet loss. In addition, this method allows users to control route advertisement based on ACLs in special scenarios. The third method won out.

Conclusion

There is a long road to technical knowledge. One can move forward only through continuous learning, accumulation, and practice. Optimizing BGP in DCN scenarios focuses on simplifying BGP deployment in the dual-stack mode, improving the convergence performance of BGP, and ensuring smooth traffic migration. Xiao Li has shared his experience to you. Hope that you can learn something from him. Putting on his favorite plaid shirt, Xiao Li continues his journey of technologies.