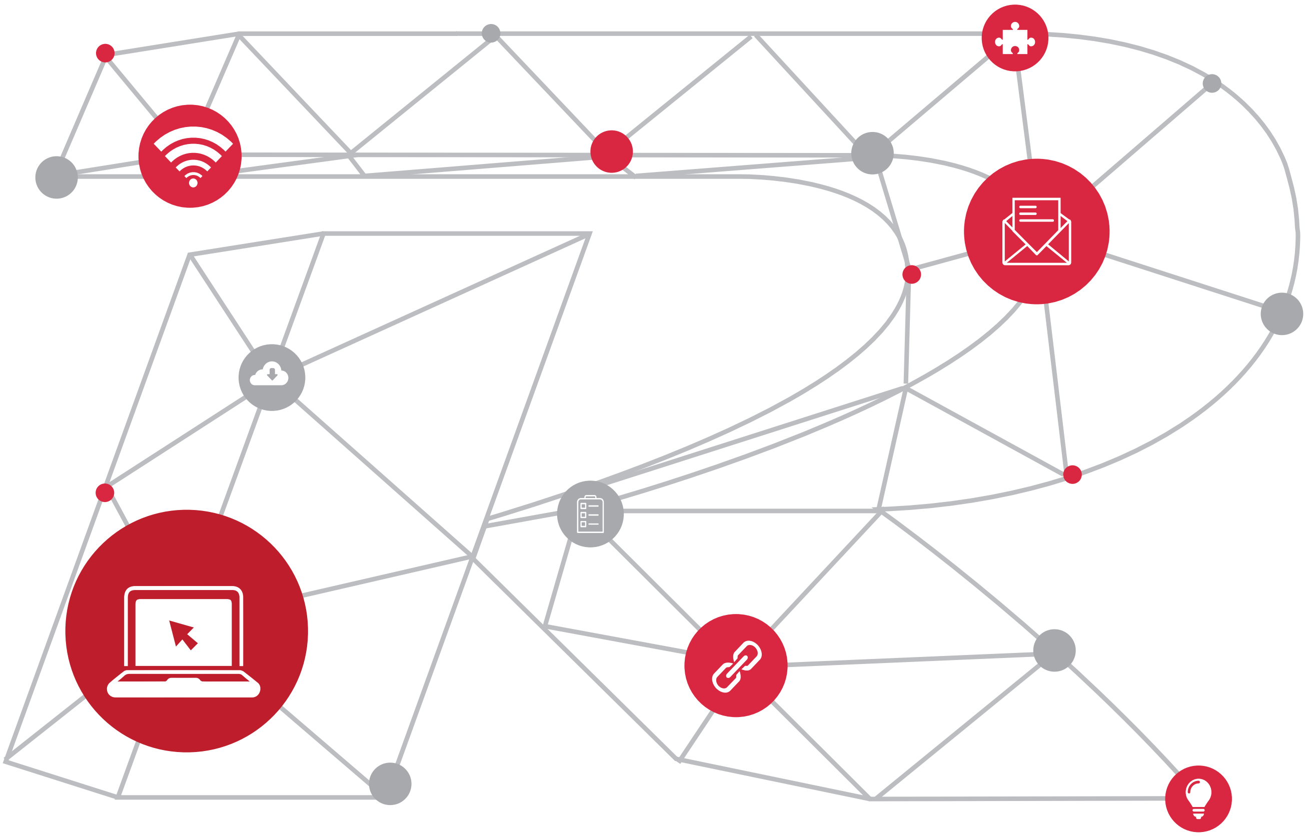
White Paper

A Feast of Technologies
Planning BGP Networks for Large IDCs



Contents

Overview

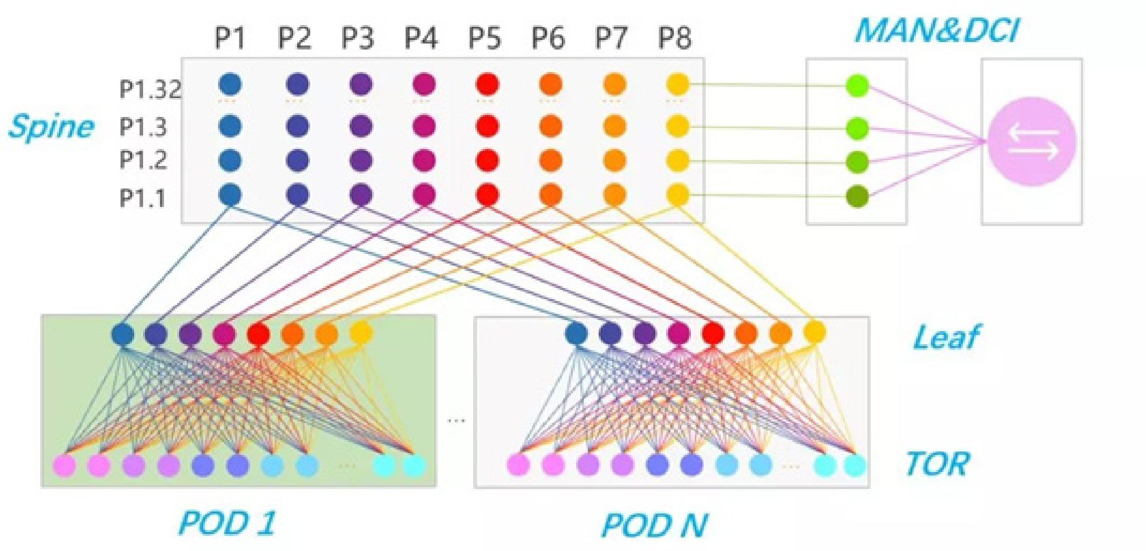
This artical draws on practical experience from large Internet companies in and out of China and summarizes some methods for planning and running Border Gateway Protocol (BGP) networks.

Preface

As described in the previous article (Selecting a Routing Protocol for Large Data Centers), BGP has become the preferred routing protocol for large Internet data centers (IDCs). It is widely known that BGP was initially designed for interconnection between different autonomous systems, not for IDCs. When BGP was introduced in IDC scenarios, various errors occurred due to incompatibility. In face of these problems, what optimizations have Internet engineers made on BGP? What problems must they consider when planning BGP networks for IDCs? This article draws on practical experience from large Internet companies in and out of China and analyzes some cases for your reference.

Network Architecture of Large IDCs

Figure 1: Internal Leaf-Spine Network Architecture of Large IDCs



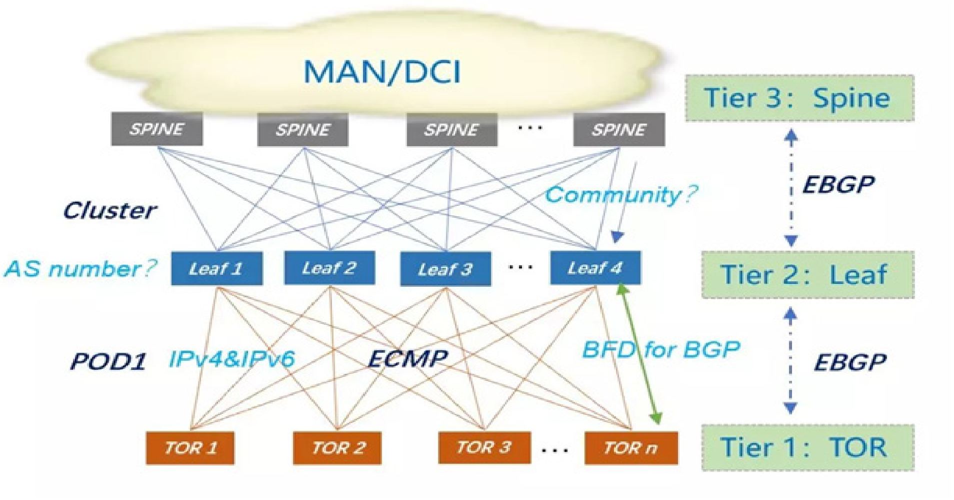
Services for IDCs require high reliability. In modern times, when you design an IDC network, an important method is to assume that all network devices and links are unreliable. In this way, when one of these unreliable devices or links becomes faulty, adverse impact on services can be eliminated through self-recovery. Therefore, the leaf-spine network architecture has become the mainstream for IDCs. As shown in Figure 1, this type of Clos multi-tier switch network brings a large number of equivalent devices and paths for IDCs. This eliminates single point of failure, enabling the network architecture to scale out with high reliability and performance.

In this network architecture, BGP is usually deployed at all tiers of the Clos network to build a simple and uniform ultra-large network for the IDC. In Figure 1, BGP is deployed on devices such as top-of-rack (ToR) switches, leaf switches, and spine switches. When you deploy BGP, you must ensure basic IPv4 and IPv6 route transfer capabilities and quick convergence, flexible control, and convenient maintenance of BGP.

Key Points of BGP Deployment

This article aims to provide some methods for deploying BGP on IDCs for your reference. The underlay routing in IDCs is used as an example.

Figure 2: Key Points of Deploying BGP on IDCs



As shown in Figure 2, BGP deployment is divided into the following parts in a typical three-tier Clos IDC network:

1. Plan basic BGP capabilities, including:

* Plan autonomous system (AS) numbers for devices at tiers 1 to 3.

* Configure basic BGP parameters and establish BGP neighbors between devices.

* Generate equal-cost multi-path (ECMP) routing for the Clos network.

* Control different types of BGP routes by configuring routing properties.

* Design route transfer rules.

* Enable IPv4/IPv6 dual stack.

2.Plan BGP operation & maintenance (O&M) capabilities, including:

* Plan autonomous system (AS) numbers for devices at tiers 1 to 3.

* Configure basic BGP parameters and establish BGP neighbors between devices.

• Planning Basic BGP Capabilities

Planning of AS Numbers

BGP AS numbers include public AS numbers and private AS numbers. Inside an IDC, although AS numbers are not advertised to external networks, we recommend that you use private AS numbers to ensure security.

In earlier BGP versions (described in RFC 1771), an AS number occupies two bytes, and only 1023 private AS numbers are available (64512–65534), which are insufficient for numerous network elements in large IDCs. Two solutions are available to address this issue:

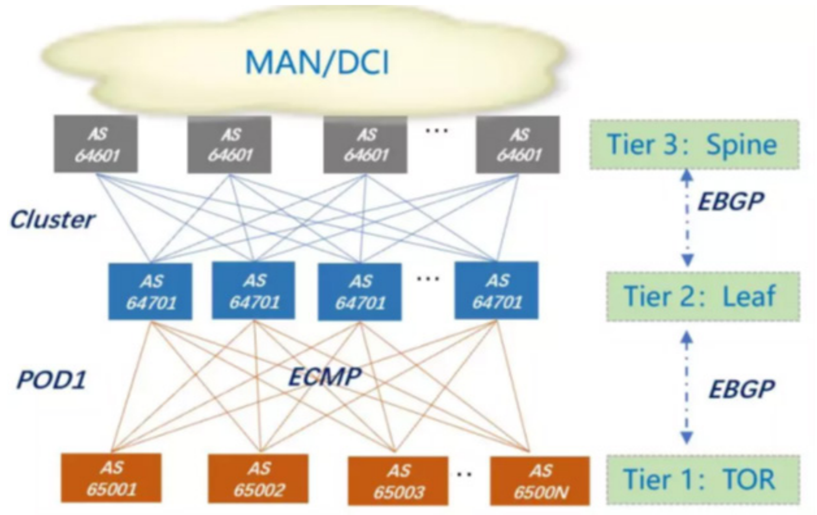
* Use four-byte BGP AS numbers defined in RFC 4893, BGP Support for Four-octet AS Number Space. In this protocol, AS numbers are increased to the same range as IPv4 addresses. Up to 90 million private AS numbers in the range from 4200000000 to 4294967294 are available. As a result, each network element and even each host in IDCs can be assigned one separate AS number.

* Use private AS numbers in the range from 64512 to 65534 and globally plan the AS numbers. This ensures that the use of AS

numbers is simple and AS numbers are supported on all devices. Global planning also allows the same AS number to be reused by multiple devices at the same time.

The following example shows recommended assignment of AS numbers.

Figure 3: Assignment Example of AS Numbers in an IDC



Device Role	AS Number Planning Rule	Assignment Mode	Example
ToR	AS numbers for ToR switches must be unique within a Pod, but can be the same across Pods.	Set the AS number to the sum of 65000 and the SN of the ToR switch.	Assign 65001 to the first ToR switch. Assign 65002 to the second ToR switch.
Leaf	AS numbers for leaf switches must be the same within a Pod.	Set the AS number to the sum of 64700 and the SN of the Pod.	Assign 64701 to leaf switches in the first Pod. Assign 64702 to leaf switches in the second Pod.
Spine	AS numbers for spine switches must be unique within a metropolitan area network (MAN).unique within a metropolitan area network (MAN).	Globally plan the AS numbers.	Assign 64601 to the spine switch in the XX cluster or district.
MAN	AS numbers for MAN switches must be unique within the internal network.	Globally plan the AS numbers.	Assign 64513 to the XX MAN.

• Basic BGP Parameters

Configuring basic BGP parameters is the basis to implement BGP interconnection in IDCs. We recommend that you configure as follows:

BGP Neighbor

BGP connections are established based on Transmission Control Protocol (TCP). Therefore, you must specify an IP address for BGP to set up BGP sessions. Within an IDC, we recommend that you use the IP address of a direct-connected port on a device to set up BGP sessions.

BGP Router ID

You can set the identifier to the IP address of the management port of the switch or the loopback address.

BGP Timers

BGP uses keepalive (KA) messages to check whether a session is alive. Based on this, BGP can determine whether the next hop is reachable. As described above, BGP was initially designed for interconnection between different autonomous systems of service providers. Route stability between ASs is more important than fast convergence. To prevent route flapping, BGP sets the default timer to a great value. The keepalive timer is set to 60s and the hold timer is set to 180s. For routers within IDCs, fast convergence after failures occur is more important. We recommend that you set the keepalive timer to 1s and the hold timer to 3s to accelerate convergence. BGP also provides an advertisement interval, an important timer that specifies the interval between the sending of routing updates. Within this period, BGP events are cached. After the timer expires, the BGP events are sent together. BGP sets the default advertisement interval to 30s. In IDCs, where instant advertisement of routing updates is required, we recommend that you set the advertisement interval to 0s.

For example, in Ruijie RGOS switches, you must configure the timer in BGP processes. The following table shows the configurations.

Configuration Command	Description
timers bgp 1 3	Sets the keepalive timer to 1s and the hold timer to 3s.
neighbor XX advertisement-interval 0	Sets the advertisement interval to 0s.

Other Recommended Configurations

bgp log-neighbor-changes: records BGP status changes when the debug mode is disabled.

• BGP ECMP Routing Paths

For Clos networks, ECMP routing is the cornerstone of network reliability and stability.

Before you implement ECMP routing, you must enable the multi-path feature. For example, the following table shows the configurations you must specify for Ruijie RGOS switches.

Configuration Command	Description
maximum-paths ebgp 32	BGP supports a maximum of 32 ECMP routing paths. We recommend that you set the maximum number of ECMP routing paths to 32 for ToR switches and 64 for leaf switches.

The preceding configuration enables the BGP multi-path feature. Next, you must add the next-hop IP addresses of the multiple links to the routing table based on BGP route selection rules. In this way, ECMP routing paths can be established. Among the 13 BGP route selection rules, the criterion for determining whether two routes are of equal cost and can load balance traffic is that the first eight conditions are the same. Among the first eight conditions, you only need to check the AS_PATH parameter when you plan BGP for IDCs, because other conditions are the same or can be ignored in IDCs.

By default, you must precisely compare the AS_PATH parameter with the AS number. ECMP routing paths can be established only when their lengths are the same. As described in AS number planning, each ToR switch has one different AS number. As a result, southbound routes from a leaf switch to two ToR switches in the same group cannot balance traffic. A solution to this issue is to enable loose comparison of AS-PATH on the leaf switch. For example, the following table shows the configurations you must specify for Ruijie RGOS switches.

Configuration Command	Description
bgp bestpath as-path multipath-relax	Only the length, and not the value, of the AS-PATH parameter is compared.

As described in AS number planning, AS numbers for all leaf switches must be the same within a Pod. Therefore, the AS-PATH parameter shown to ToR switches are always the same, regardless of which leaf switch sends routing updates. In this case, you do not need to enable loose comparison on leaf switches.

In addition, a large number of ECMP neighbors exist between leaf switches and ToR switches. Configuration policies for these ECMP neighbors are the same. During deployment, we recommend that you use the BGP peer-group feature to simplify configuration.

For example, the following table shows how to configure this feature on Ruijie RGOS switches.

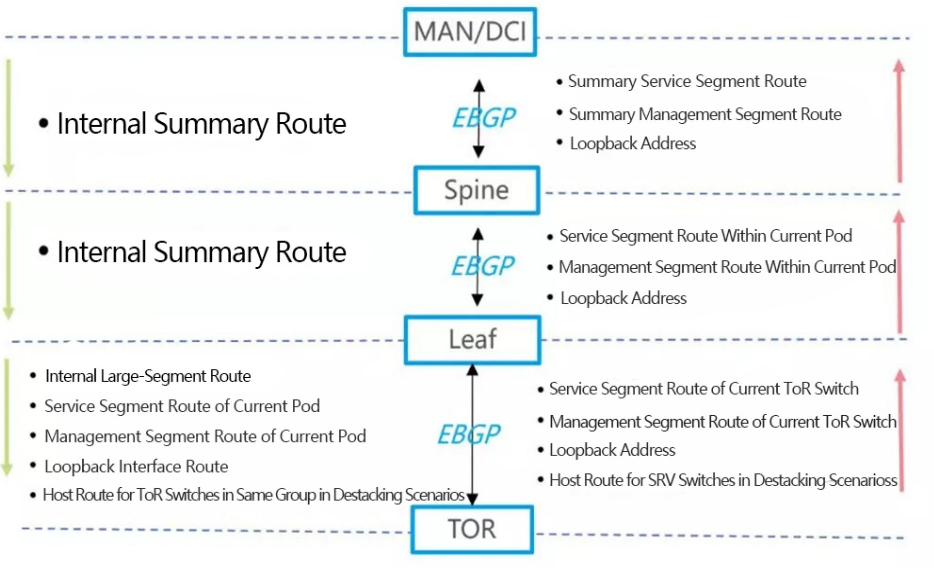
Configuration Command	Description
neighbor abc peer-group	Creates a peer group named abc.
neighbor [Neighbor IP] peer-group abc	Adds the neighbor to peer group abc.

• BGP Routing Attributes

BGP provides abundant extension attributes to implement powerful routing control. The BGP community attribute is the most commonly used in IDCs to simplify routing policies. In IDCs, we usually use the private community attribute to add a management flag to a prefix. The private community attribute is in the AS:number format. AS indicates the local AS number or the peer AS number and number indicates a group to which you can apply the policy. The number is locally assigned. You can also use a simplified community flag. For example, you can add the 1:1 attribute to the service network segment and the 2:2 attribute to the internal summary route. Then, you can exercise precise control over route transfer based on this.

• Route Transfer Rules

Figure 4: Advertisement of Routing Updates in the IDC BGP Network


MAN/DCI
• Internal Summary Route
EBGP
• Summary Service Segment Route
• Summary Management Segment Route
• Loopback Address
Spine
• Internal Summary Route
EBGP
• Service Segment Route Within Current Pod
• Management Segment Route Within Current Pod
• Loopback Address
Leaf
• Internal Large-Segment Route
• Service Segment Route of Current Pod
• Management Segment Route of Current Pod
• Loopback Interface Route
• Host Route for ToR Switches in Same Group in Destacking Scenarios
EBGP
• Service Segment Route of Current ToR Switch
• Management Segment Route of Current ToR Switch
• Loopback Address
• Host Route for SRV Switches in Destacking Scenarioss
TOR

As shown in Figure 4, a Pod includes multiple groups of ToR switches and leaf switches. A Pod is the minimum delivery unit and serves as the basic physical design unit in IDCs. A spine switch horizontally connects multiple Pods. MAN and Data Center Interconnect (DCI) switches connect switches across regions. The following content describes our suggestions on BGP routing planning in IDCs:

* North-bound Route Transfer

Service segment routes, management segment routes, and loopback addresses are advertised from ToR switches to leaf switches, spine switches, and MAN and DCI switches tier by tier. In destacking scenarios, ToR switches must advertise host routes to spine switches.

* South-bound Route Transfer

Summary routes of the entire internal network are transferred from MAN and DCI switches to spine switches and leaf switches tier by tier. For example, 10.0.0.0/8;172.16.0.0/12;192.168.0.0/16 is transferred. In addition to internal summary routes, service segment routes, management segment routes, and loopback addresses of the current Pod also need to be advertised from leaf switches to ToR switches. If the uplink of a leaf switch is faulty, traffic of the same Pod can match detailed routing and be forwarded by spine switches as usual.

Note

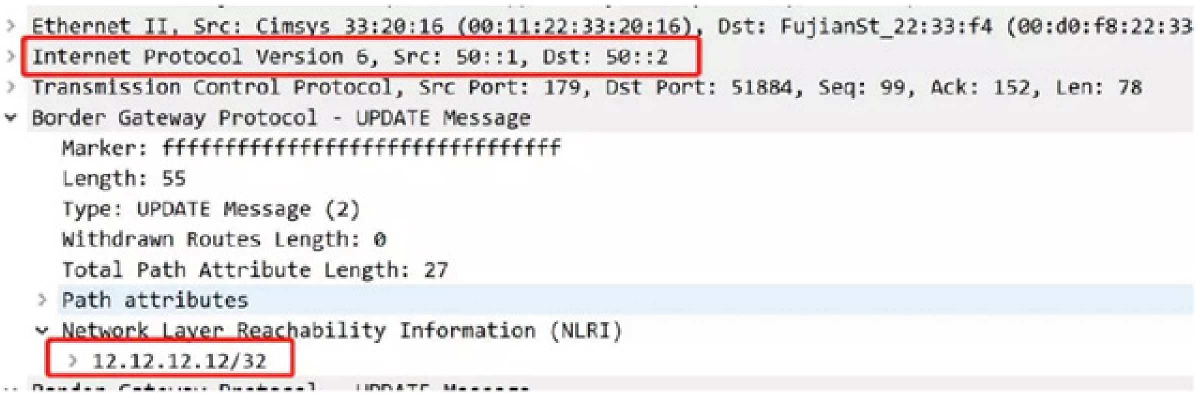
The destacking technology is used more and more widely at the ToR tier to implement dual homing. For more information, see another article of the "A Feast of Technologies" series: How to Destack the Network Architecture of IDCs. In destacking scenarios, a leaf switch receives a large number of host routes from ToR switches. Tens of thousands of host routes may exist, depending on the number of hosts within the Pod. As a result, when the leaf switch transfers host routes between ToR switches, the maximum number of routes for ToR switches may be reached. To address this issue, you must specify a route receiving policy for ToR switches to filter out host routes from other ToR switches.

• BGP Dual Stack

China has been promoting the development of IPv6 addresses over recent years. Private IP addresses for large IDCs are running out. Therefore, deployment of IPv4/IPv6 dual stack within IDCs is urgently needed.

BGP supports multiple protocols and therefore supports IPv4/IPv6 dual stack in the same BGP process. Generally, BGP sessions are separately set up for BGP IPv4 and IPv6 neighbors. However, this doubles configuration and maintenance work. Actually, BGP IPv4 update messages can be sent over TCP connections that are established by using IPv6 addresses, and vice versa. In other words, a single connection can be used to advertise update messages of multiple protocol families.

Figure 5: Advertising IPv4 Update Messages over an IPv6 Session



```
> Ethernet II, Src: Cimsys 33:20:16 (00:11:22:33:20:16), Dst: FujianSt_22:33:f4 (00:d0:f8:22:33
> Internet Protocol Version 6, Src: 50::1, Dst: 50::2
> Transmission Control Protocol, Src Port: 179, Dst Port: 51884, Seq: 99, Ack: 152, Len: 78
v Border Gateway Protocol - UPDATE Message
    Marker: ffffffffffffffffffffffffffffffff
    Length: 55
    Type: UPDATE Message (2)
    Withdrawn Routes Length: 0
    Total Path Attribute Length: 27
  > Path attributes
  v Network Layer Reachability Information (NLRI)
    > 12.12.12.12/32
> Border Gateway Protocol - UPDATE Message
```

As shown in Figure 5, Ruijie Network provides an optimized solution: Set up a single session to carry the routes of dual stack. This solution simplifies configuration, saves IP addresses, and reduces a half of performance costs of deployment of BFD for BGP and other protocols.

Planning BGP O&M Capabilities

In addition to basic BGP capabilities, high BGP O&M capabilities are also required in IDCs. Common BGP O&M capabilities include the following:

Accelerating BGP Network Convergence by Using BFD

Although IDC networks are built based on high redundancy, their network reliability is restricted by the capabilities of detecting failure by using network devices and rerouting traffic to other paths. This restriction is evident especially when a one-way problem occurs on the optical modules or optical fiber. IDCs require extremely low convergence time. The convergence time for cloud services must be within subseconds. As described above, you can modify the BGP timers to accelerate convergence. However, this slow hello mechanism can only reduce the convergence time to seconds, which fails to meet the requirements.

BFD provides the detection precision of milliseconds. You can combine BFD with BGP to enable quick convergence of BGP routes, ensuring service continuity. We recommend that you enable the BFD for BGP feature in IDCs. To ensure high performance of the devices, when all ports are enabled, we recommend that you set the parameters to 300 ms and 3.

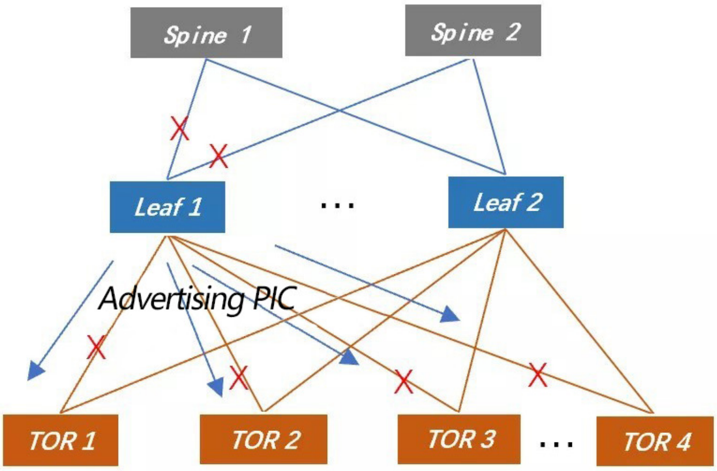
For example, the following table shows the BFD configurations on Ruijie RGOS switches.

Configuration Command	Description
neighbor XX fall-over bfd	Enables BFD.
bfd interval 300 min_rx 300 multiplier 3	Sets the detection interval to 300 ms and the timeout threshold to 3, indicating that detection times out when advertisement fails three times.

Ensuring Non-stop Service — BGP PIC

To accelerate BGP route convergence, we must delete failed routes and add routes on the routing table and correspondingly add and delete routes on the forwarding table on the chip. However, when a large number of routes exist, deleting routes one by one and refreshing the routing table take a long time and the convergence time may reach seconds or event tens of seconds. Ruijie RGOS switches optimize route convergence by supporting prefix-independent convergence (PIC). As shown in Figure 6, when all external BGP (eBGP) neighbors from leaf 1 switch to spine switches fail, leaf 1 switch notifies all ToR switches that the AS to spine switches is unreachable. When ToR switches receive this message, the ToR switches search the ID-based index and perform next-hop switching by advertising the forwarding table. The ID-based index is pre-assigned based on the AS numbers of spine switches and the router IDs of leaf switches. The switching accelerates convergence for services. Convergence is no longer restricted by the number of route entries. According to a large Internet company, the convergence time of 12,000 routes is only 0.7s.

Figure 6: BGP PIC



Ensuring Non-stop Service — BGP NSR

To ensure high reliability, most leaf and spine switches in IDCs are configured with dual management boards. In stacking scenarios, ToR switches also implement the effects similar to those of dual management boards. When active and standby management boards are switched over, inconsistency between status information causes protocol oscillation.

Non-stop routing (NSR) is used to ensure that routes are not interrupted when the protocol restarts during a switchover between the active and standby management boards. When the BGP NSR feature is enabled for a neighbor, the TCP non-stop service of the neighbor is enabled. In addition, the related neighbors and route information are backed up to the standby management board. During a switchover between the active and standby management boards, the NSR feature ensures stable network topology, maintains neighbor status and forwarding tables, and guarantees non-stop key services.

Ensuring Non-stop Service — Smooth Exiting and Delayed Publishing of BGP

Smooth exiting: In Clos networks of IDCs, when you upgrade a device in the isolated mode, the smooth exiting feature of BGP ensures that no or few services are interrupted.

Smooth exiting is performed as follows:

* Advertises the route of the lowest priority together with the gshut community attribute to a neighbor. In particular, if the value of local-preference is 0 or the value of med is 4294967295, the route is of the lowest priority. The gshut community attribute enables the neighbor to update routes and switch traffic to a backup link or other equivalent links in advance.

* **Shuts down the BGP connection with the neighbor after ensuring that route learning is complete.**

Delayed publishing: When a device restarts, routing information is advertised to a neighbor, but the routing table may have not been delivered to local hardware entries. As a result, an error occurs when the neighbor switches traffic in advance. To prevent this problem, you can adjust the priority of the route that you want to publish to the lowest upon device restart.

We recommend that you preconfigure this capability on the device. For example, the following table lists the configurations on Ruijie RGOS switches.

Configuration Command	Description
bgp advertise lowest-priority on-startup 120	The value 120 is the duration of publishing a route of the lowest priority. You can change the value as needed.

Afterword

Planning, building, and operating a BGP network for IDCs are not easy. A lot of experience is required. Fortunately, BGP is maturing in IDCs. We can reference many cases and practices from large Internet companies and operators. It is a great honor for Ruijie Network to participate in this great project. We have delivered several large IDC BGP networks for Tencent, Alibaba, ByteDance, and other companies.

For BGP performance optimization and more BGP O&M features, stay tuned for subsequent articles of the "A Feast of Technologies" series.