BGP in the datacenter explained

Miguel Moraga Muñoz

Part 1: A new landscape

BGP (Border Gateway Protocol) is the most common routing protocol among the EGPs. Since it replaced the EGP standard in 1989, its different versions have been established as the Internet routing protocol. In spite of this, it has always been considered a protocol for long distances, while for IGPs, other protocols such as OSPF, RIP or IS-IS are more frequent. However, at these days, it starts to make sense to use BGP as the only protocol in our data center.

BGP is called a path vector routing protocol, in other words, it is a distance vector routing protocol with several additional attributes. Thesel attributes allow users to acquire more flexibility in order to manipulate routing decisions. That means they also have higher flexibility in controlling network traffic.

But first, let's see how the landscape has changed in recent years. Traditional data centers are far from current needs, having changed the main problems they face. Of the two types of traffic flows that exist in a data center, the traditional data centers assume a greater traffic of the North / South type, while in the current data centers, it is the East / West type that monopolizes most of them. Server to server communication has been increasing significantly.

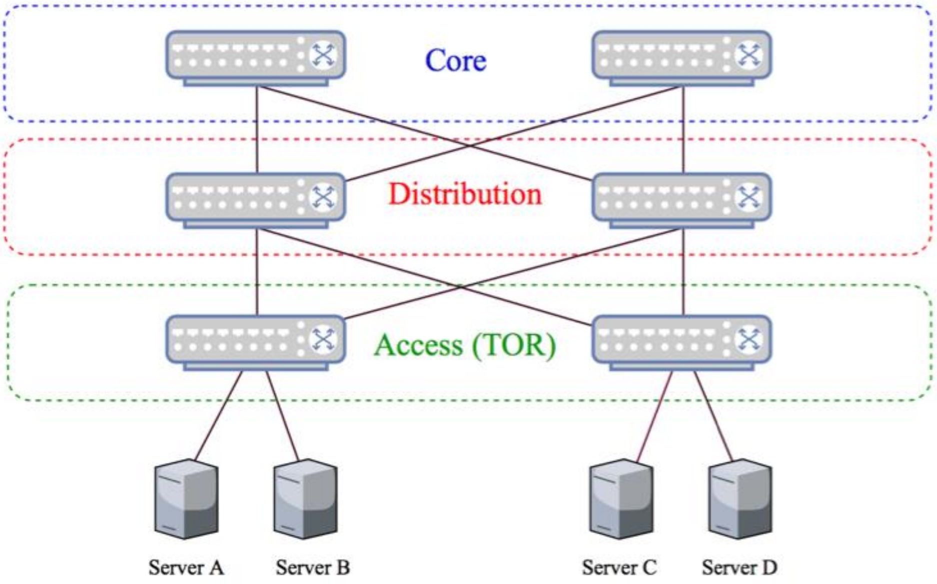

Figure 1: Traditional Data Center Architecture

When the datacenter grows, this architecture may not be able to scale due to various types of limitations. Adding new devices to the distribution layer will result in adding new devices to the core at some point, because the core layer has to be adjusted based on the lower layers' increased bandwidth requirements. This means the data center has to scale vertically, since it was developed based on North / South traffic considerations.

Part 2: A new topology

To solve these problems, we need a simple topology providing significant amount of bisection badwith. Here's how to overcome these challenges, let us take a closer look at CLOS Architecture in the data center.

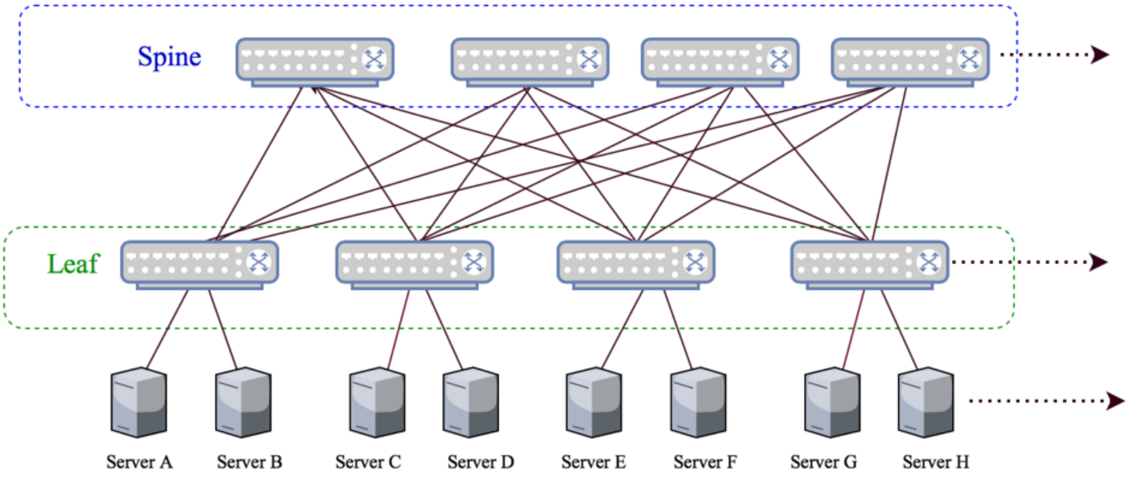

Figure 2. CLOS Data Center Topology

A CLOS topology is comprised of spine and leaf layers. Servers are connected to leaf switches (Top of Rack) and each leaf is connected to all spines. There is no direct leaf-to-leaf and spine-to-spine connection.

There are various advantages in this architecture. First of all, regardless of being in the same vlan condition, each server is three hops away from the others. That's why this is called 3-stage CLOS topology. It can be expanded to 5-stage CLOS by dividing the topology into clusters and adding another top-spine layer (also known as a super-spine layer). No matter how many stages there are, total hop count will be the same between any two servers. Therefore, consistent latency can be maintained throughout the data center.

Multi-Chassis Link Aggregation Group (MLAG or MCLAG) is available on the server side. Servers can be connected to two different leaf or TOR switches in order to have redundancy and load balancing capability. On the other hand, as the connectivity matrix is quite complex in this topology, failures can be handled gracefully. Even if two spine switches go down at the same time, the connectivity between servers will remain.

Finally, the CLOS topology scales horizontally, which is very cost effective. The bandwidth capacity between servers can be increased by adding more spine-leaf links as well as adding more spine switches. As newly added spine switches will be connected to each leaf, server to server bandwidth/throughput will increase significantly.

Part 3: BGP in the datacenter

At first look, the choice of a protocol for our architecture may seem to be highly dependent on the size of the data center. IGP could easily scale if there are only a few thousand prefixes in total. This isn't wrong, but there are more considerations than sizing:

- From a configuration perspective, IGP is easier to deploy compared to BGP, especially considering the number of peers to be configured in BGP. Therefore, that is not a problem with automated configuration generation.
- IGP is more likely to be supported by TOR switches. In BGP deployments, this limitation could result in aggregating the TOR and leaf layers.
- BGP is better when dealing with a high number of prefixes, and it has a limited event propagation scope. On the other hand, in such a complex connectivity matrix (depends on the number of switches), IGP will propagate link state changes throughout the network, even to the irrelevant nodes which are not impacted. SPF calculations will take place on each node. Some mechanisms such as Incremental SPF or Partial SPF could avoid this but may add more complexity.
- If IGP is used, BGP will still remain in the data center, most likely at the edge. This means there will be two routing protocols in play and redistribution will be necessary. That being said, BGP can be the only routing protocol if chosen.
- With its attributes and filters, BGP provides much better flexibility on controlling the traffic, and it provides per-hop traffic engineering capability. BGP AS path visibility also helps operators troubleshoot problems more easily.
- By default, IGP is more compatible with ECMP, where BGP configuration should be adjusted to meet the load balancing requirements.

When looking to enable BGP in data center networks, whether to use iBGP or eBGP is a good question. There are similar obstacles in both. However, eBGP-based design is seen more commonly, because iBGP can be tricky to deploy, especially in large-scale data centers.

One of the issues with iBGP is route reflection between nodes. Spine switches can be declared as route reflectors, but this causes another issue: Route reflectors reflect only the best paths. Therefore, nodes won't be able to use ECMP paths. The BGP add-path feature has to be enabled to push all routes to the leaves. On the other hand, "AS_Path Multipath Relax (Multi-AS Pathing)" needs to be enabled in eBGP. Even so, eBGP does not require maintaining route reflectors in data centers.

As stated before, BGP has extensive capabilities on per-hop traffic engineering. With iBGP, it is possible to use some part of this capability but eBGP's attributes provide better visibility, such as directly comparing BGP-Local-RIB to Adj-RIB-In and Adj-RIB-Out. In terms of traffic engineering and troubleshooting, it is more advantageous than iBGP.

As seen in the figure below, spine switches share one common AS. Each leaf and each TOR switch has its own AS. Why are spines located in one AS while each TOR switch has its own? This is to avoid path hunting issues in BGP.

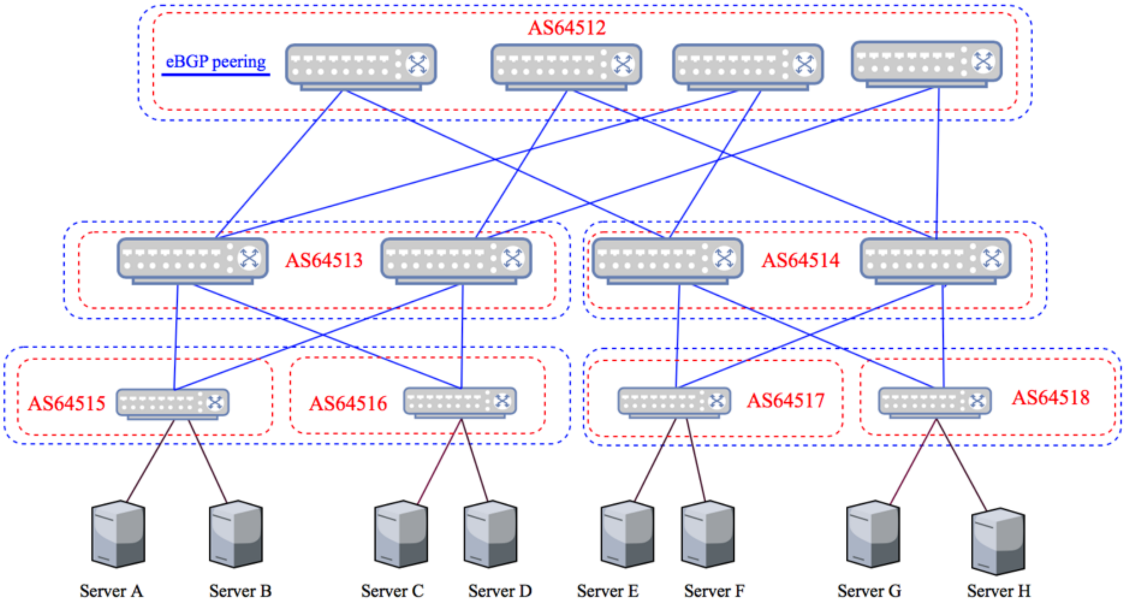

Figure 3. CLOS Topology using eBGP for Spine, Leaf and TOR Switches

Here are some concerns about this architecture:

- ECMP (equal-cost multi-path routing):

ECMP is one of the most critical points, because otherwise it wouldn't be possible to use all links. IGP can easily deal with this requirement, whereas BGP needs some adjustments.

The topology above is multiple pod/instance design, which reduces the number of links between leaf and spine layer. It also allows operators to put more than one leaf node in each AS. This type of setup is commonly seen, especially in large data center networks.

It seems quite straightforward in this topology, but what if each leaf switch has its own AS or servers are connected to more than one TOR switch for the sake of redundancy? Then, the length of each AS Path attribute will be the same, which does not suffice for using multi-paths in BGP. Even if the length is identical, in order for BGP to route traffic over multiple paths, all attributes have to be the same, including the content of the AS Path attribute. Fortunately, there is a way to meet this requirement: The "AS Path multi-path relax" feature needs to be enabled to let BGP ignore the content of the AS Path attribute, while installing multiple best routes as long as the length is identical.

- Convergence:

Due to its nature and area of use, convergence time initially was not one of the first concerns in BGP. Stability was prioritized over fast convergence. However, some fast convergence enhancements have been introduced to BGP subsequently:

- BGP neighbor fall-over (iBGP) and/or BGP fast external fall-over (eBGP) should be enabled.
- BFD can also be used, but until recently, there was a limitation that BFD did not take any action in case of a single link failure on a LAG. Still, it needs to be checked with vendors.
- The Events Advertisement Interval in eBGP peering should be set to zero, which is the default value in iBGP peering.
- The Keepalive timer should be set to five seconds at most and hold time should be set to 15 seconds.

- Number of ASes to be used:

Since each TOR switch is located in its own AS, how are operators able to scale the number of ASes, especially considering the number of TOR switches in a large data center?

- There are 1,023 private ASNs. If this is not sufficient, one of the options is using 4-byte ASNs, which enables millions of ASNs.
- TOR ASNs can be used more than once. In this case, the BGP Allow-AS-In feature needs to be enabled on TOR switches. This will turn off one of BGP's main loop avoidance mechanisms, and TOR switches will accept the routes even though they see their own ASN in received updates.
- If instances/pods are used in the topology where leaf switches share ASes, summarization might create black holes when a specific prefix is withdrawn on the TOR side. To avoid these kinds of issues, specific prefixes should not be hidden.

BGP is a protocol that meets several requirements and architectural needs in various segments of the network. I personally think the CLOS with eBGP-based architecture is going to be seen more commonly in data centers.

Sources:

https://www.nanog.org/meetings/abstract?id=1942

https://www.nanog.org/meetings/abstract?id=2137

https://www.packetdesign.com/blog/bgp-in-the-data-center-part-1/?cn-reloaded=1

https://www.youtube.com/watch?v=yJbqnOdD3cg&ab_channel=TeamNANOG