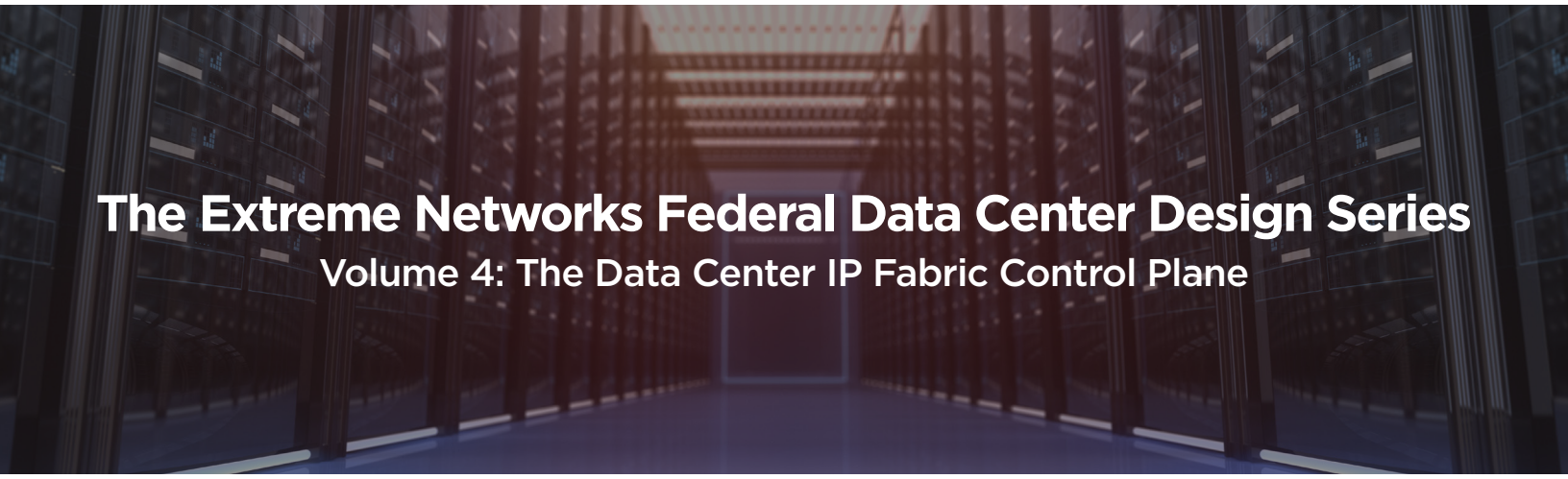


The Extreme Networks Federal Data Center Design Series

Volume 4: The Data Center IP Fabric Control Plane

Overview

Volume 3: Evaluation of Reliability and Availability of Network Services in the Data Center Infrastructure, discussed how agencies could be confident in their selection of either a single form factor or chassis based configuration in the underlay of the data center network. This is a result of the ultra high availability that may be achieved through proper design of the PoD. This volume discusses the IP Fabric Control plane in the data center, often referred to as the underlay. It will also mention the tools available by Extreme Networks that aid in establishing the control plane that reduce day zero risk and deliver easy add, moves and changes on Day 1 and beyond in the data center lifecycle.

IP Fabric Control Plane

How do we get from a Clos based architecture to operation like a full IP Fabric? One of the tenets for the Clos architecture is that all established paths between the compute/storage stack and the Top of Rack (ToR) Leaf, the Leaf to Spine connection, and the Spine to Border-Leaf or Super-spine connections provide any-to-any connectivity, remain on and pass traffic.

Here we look to describe some of the underpinnings in the Clos architecture in order to explain how these protocols combine to deliver the appropriate services.

From an underlay perspective, after the physical topology is created and arranged in a Leaf-spine PoD connected to a Border Leaf (3 stage Clos), or connected to a Super-Spine- which is connected to the border-leaf (5-Stage Clos)- we need a discovery process that identifies the physical reachability of the interfaces. To perform the initial detection and reachability, we enable LLDP protocols to find the extent of the connectivity. We also enable multi-chassis trunking (MCT) where appropriate or LACP (link aggregation control protocol). With these protocols we have enabled the discovery of physical connections. To ensure that these connections may be monitored, enabled and operating in stable manner, ports are enabled (Bidirectional Forward Detection) BFD functionality on the optics pairs to ensure that adjacencies, establishment or lost is handled immediately with minimal traffic disruption. BFD validates the forwarding plane between the two switches or routers. It will work with BGP (eBGP, iBGP), LDP or static routes. If a forwarding path fails, it can sever the adjacency between the iBGP instances. BGP has a keep-alive mechanism as well. A session between peers can be shut down when the neighbor becomes unreachable if the physical link goes down. With iBGP, when the link is down, there is still a use for forwarding as there can be more than one link between neighbors, BFD removes the specific adjacency from the forwarding database and reachability is maintained by the loopback interface.

The industry has predominantly settled upon a common set of practices to complete the underlay. From the control-plane perspective, there are two deployment options:

- Pervasive eBGP (all routers are BGP speakers)
- iBGP within a PoD

In this example (Figure 1), we show a notional design, Extreme recommends eBGP as the control plane in the fabric. IP Fabric is also referred to as routing to the top

of rack (ToR), in comparison with the traditional multitier access and aggregation networks where the L2/L3 boundary is on the aggregation devices. In IP Fabric, this boundary is moved to the edge or leaf. The leaf nodes directly advertise their server subnets into the BGP control plane. Reachability between the server subnets on various racks is established using control-plane learning.

Moving the intelligence to the leaf helps with scaling out the number of spines depending on the bandwidth or oversubscription issues in the network. For instance, the ratio between the bandwidth of the server ports and the uplinks ports (from leaf to spine layer). If the oversubscription is too high, additional spines may be added.

Pervasive eBGP

This deployment model refers to the usage of eBGP peering between the leaf and the spine in the fabric. This design using eBGP as a routing protocol within the data center is based on the IETF draft *Use of BGP for Routing in Large-Scale Data Centers*. In this model, each leaf node is assigned its own autonomous system (AS) number. The other nodes are grouped based on their role in the fabric, and each of these groups is assigned a separate AS number, as shown in Figure 1. Using eBGP in an IP fabric is simple and provides the ability to apply BGP policies for traffic engineering on a per-leaf or per-rack basis since each leaf or rack in a PoD is assigned a unique AS number. It is worth noting that private AS numbers are used in the IP fabric. One design consideration for the AS number assignment is that a 2-byte AS number provides a maximum of 1023 private AS numbers (ASN 64512 to ASN 65534); if the IP fabric is larger than 1023 devices, we recommend using 4-byte private AS numbers. As such, you will note that many vendors, including Extreme Networks typically begin their private As numbering between the following ASN range: ASN 4,200,000,000 to 4,294,967,294.

The following principals apply to the design:

- Each leaf in a PoD is assigned its own AS number
- Leafs advertise the server subnets directly into BGP
- All spines inside a PoD belong to one AS
- All super-spines are configured in one AS
- Edge or border leafs belong to a separate AS
- Each leaf peers with all spines using eBGP
- Each spine peers with all super-spines using eBGP
- Each border leaf peers with all super-spines using eBGP
- No BGP peering occurs between nodes in the same layer

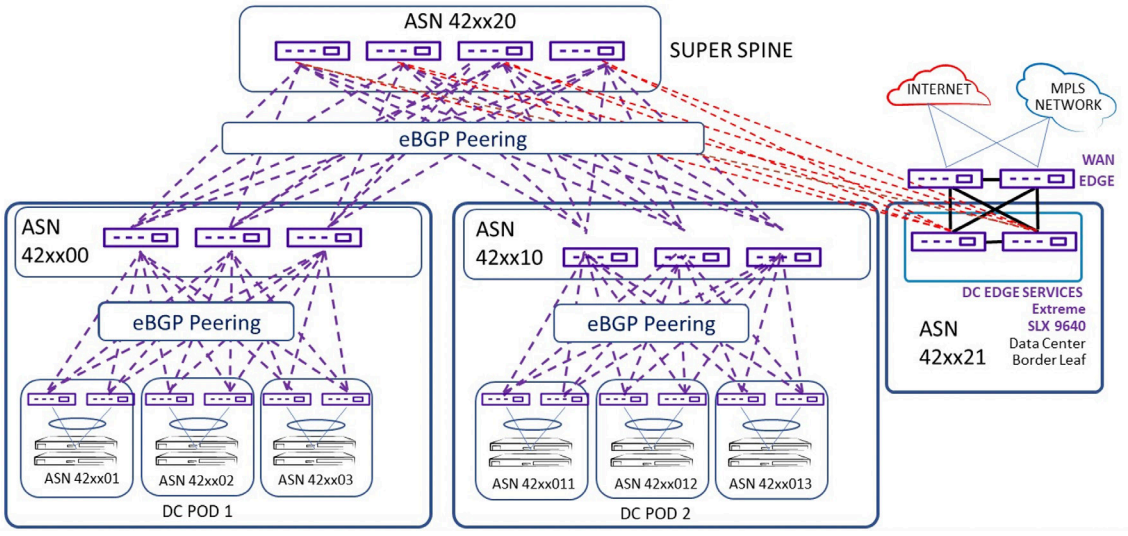
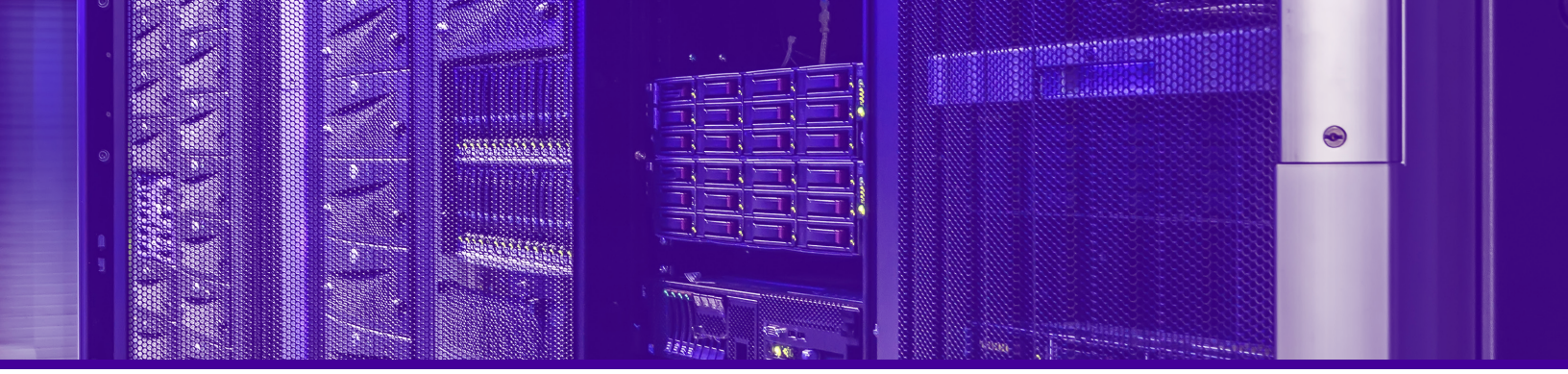


Figure 1: IP Fabric with eBGP as the Control Protocol



iBGP for Routing Inside a PoD

In this deployment model, each PoD and edge services PoD is configured with a unique AS number, as shown in Figure 2. The spines and leafs in a PoD are configured with the same AS number. The iBGP design is different than the eBGP design because iBGP must be fully meshed with all BGP-enabled devices in an IP fabric. To avoid the complexities of a full mesh, spines must act as route reflectors toward the leaf nodes inside the PoD. eBGP is used to peer between spines and super-spines. The super-spine layer is configured with a unique AS number; all super-spines use the same AS number.

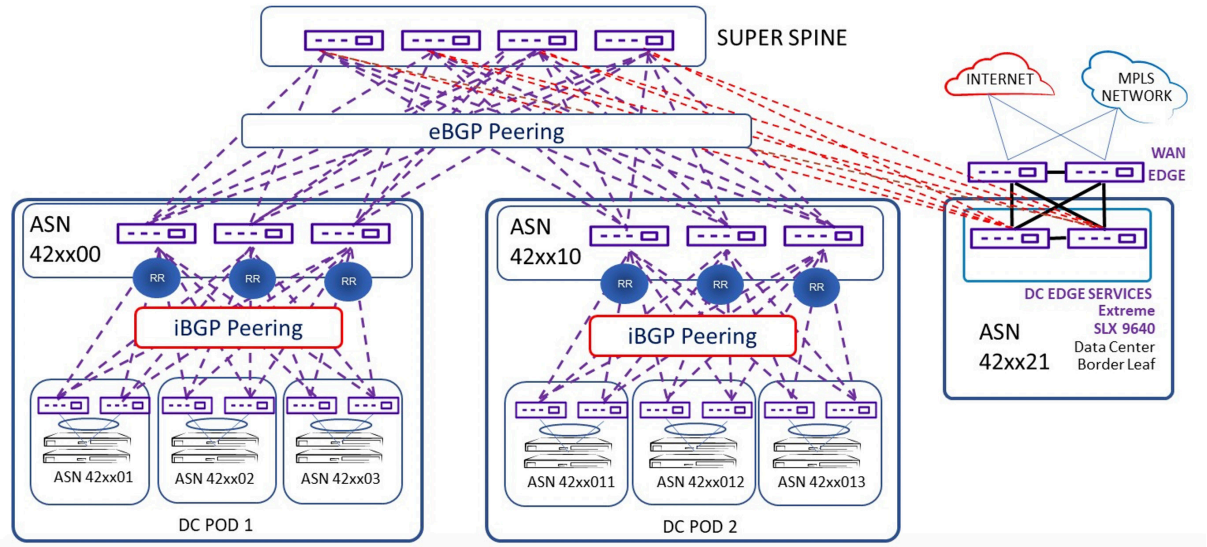


Figure 2: IP Fabric with iBGP as the Control Plane

When to use iBGP vs eBGP?

One of the common questions asked when discussing the implementation of an IP Fabric using eBGP, or iBGP is "When do I use iBGP vs eBGP?" In short, there are no hard/fast answers to this question, only preferences. For example, with eBGP, a private ASN can be used and a separate AS may be designated for each PoD. This may make troubleshooting easier. While it is not required, daily operations may represent over 95% or higher, as a percentage of the solution lifecycle. The use of private ASNs for each rack (2 byte (1023 devices) or 4 byte (almost 95 million)) may make management easier (understanding where traffic is heading (source AS to destination AS). While using eBGP within a Fabric (PoD) provides the ability to apply BGP policies on a per rack, unique AS basis (Recommended that private AS numbers are used).

With iBGP, having a full mesh of routes, with split horizon running to prevent loops under one large AS may be preferred. The bottom line, with using Extreme Networks Extreme Fabric Automation (EFA) is that by removing the provisioning load, extensive CLI usage (error-prone) is no longer a large consideration factor, the fabric is setup and validated automatically after a few short commands. For example, where a 12-node fabric may take 6 to 12 hours to configure (depending upon the OS/vendor, and the typist' speed) Extreme Networks' EFA would set the same fabric up within a couple of minutes, adding multiple PoDs with a Super-Spine (5-Stage Clos) is also supported. iBGP would run between the spine and leaf in a PoD. All spines act as BGP route reflectors to the Leafs for the underlay. EBGP is used to peer between the spines and super-spines. By automating the setup, Extreme has eliminated the potential for human error interfering with the configuration. If there

are design preferences, these may be edited in the EFA prior to running the fabric setup. To accomplish PoD customization and automated fabric or tenant creation.

Edge Services and Border Leafs

For two-tier and three-tier data center topologies, the role of the border leaf in the network is to provide external connectivity to the data center site. In addition, since all traffic enters and exits the data center through the border leaf switches, they present the ideal location in the network to connect network services like firewalls, load balancers, and edge VPN routers. The border leaf switches connect to the WAN edge devices in the network to provide external connectivity to the data center site. As a design principle, two border leaf switches are recommended for redundancy. The WAN edge devices provide the interfaces to the Internet and DCI (Data Center Interconnect) solutions. For DCI, these devices function as the Provider Edge (PE) routers, enabling connections to other data center sites through WAN technologies like Multiprotocol Label Switching (MPLS) VPN and Virtual Private LAN Services (VPLS). The Extreme validated design for DCI solutions is discussed in a separate Extreme Validated Design (EVD) document. There are several ways that the border leafs connect to the data center site. In three-tier (super-spine) architectures, the border leafs are typically connected to the super-spines. In two-tier topologies, the border leafs are connected to the spines. Certain topologies may use the spine as border leafs (known as a border spine), overloading two functions into one. This topology adds additional forwarding requirements to spines—the spines need to be aware of the tenants, VNIs, and VXLAN tunnel encapsulation and de-encapsulation functions.

From a customer or data center planners perspective the key takaway with understanding the underlay and overlay relationship is that the underlay is designed to enable fast set-up, maintain adjacencies, identify and communicate failure and restoral of links to preserve traffic forwarding and reachability in an any to any Clos based design. Extreme enables this behavior, quickly and dynamically in a validated fashion throughout the entire point of delivery. No other vendor can match the quick and easy deployment model offered by Extreme Networks offered with Extreme Fabric Automation. There is also a separate Solution Brief available that describes the benefits in more details..

Summary

In this document it was described how the underlay is the key component with IP Fabric networks serving the data center. The underlay protocols serve to ensure that traffic may be passed and that rules are maintained to ensure reachability across the fabric from the Leaf to the Spine, to the Super-spine and the Border Leaf. By utilizing standards-based protocols, interoperability is maintained, user familiarity with the expected behavior of the underlay, and assurance that operations of the fabric may be managed according to industry practices. Scalability is preserved by the protocols' inherent design to handle large-scale routing high capacity performance.

Extreme Fabric Automation (EFA) enables critical Day 0 fabric infrastructure CRUD (Create, Read/Monitor, Update and delete) operations of Extreme Networks SLX routing and switching portfolio. By leveraging this technology, provisioning Data Center IP Fabric based on BGP, EVPN and VXLAN are provisioned in seconds using simple commands from the application. There is no need to worry about configuring every switch manually with protocol specific commands, thus reducing the time it take to bring up a new IP Fabric and greatly reducing human errors.

Extreme networks

http://www.extremenetworks.com/contact