The Extreme Networks Federal Data Center Design Series

Volume 2: Achieving Any-to-Any Connectivity with Time-Tested Design Approaches and Proven Methodologies

Overview

Volume 1: Industry Trend towards a de facto standard: IP Fabric Overview, discussed how agencies were being tasked to deliver on the priorities of the OMB in their data centers, and listed the priorities in brief fashion. This volume discusses how Extreme Networks delivers against these priorities. Data center network planners will be confident that by selecting Extreme Networks solutions they are meeting or exceeding their agency's requirements, but also adhering to the priorities of the OMB.

Have you had a conversation with a colleague or peer about the latest technology and found that you each worked on some ancestral solution that bore a canny resemblance to the latest technology? It may be a bit unfair to characterize a new technology as an old one with a new spin. There must be reason for the thread of familiarity between the old and the new technologies. It is often the case that the new technology must preserve previous functionality and expand the functionality. In some cases, market drivers force deployments of an inferior solution that may implemented in a fashion that meets the needs of the enterprise, but perhaps not as efficiently as the market would desire. For example, when spanning tree was initially introduced, its' purpose was to provide loop free networks at layer 2. With the arrival of SPB (shortest path bridging) and TRILL (transparent interconnect with lots of links), the main deficiencies of the spanning tree protocol were finally overcome. Perhaps no networking challenge has been more persistent than delivering any-to-any connectivity in the most efficient manner. In the data center, delivering any-to-any connectivity is why IP Fabric technologies have been developed.

IP Fabric Overview

As noted previously, the industry is rapidly adopting IP Fabric. IP Fabric provide a Layer 3 Clos deployment architecture for data centers. With Extreme Networks IP Fabric, all links in the Clos topology are Layer 3 links leveraging an overlay network to extend layer 2 and layer 3. In comparison with the traditional 2-tier access-aggregation topologies and Layer 2 fabrics where the L2/L3 demarcation happens on a device typically more than a hop away from the access port, the L2/L3 boundary in IP fabric is pushed to the Top of Rack (ToR) or the leaf node itself (a.k.a.: routing to the ToR). In an IP fabric model, Leafs advertise the server subnets attached to them directly into the routing control-plane protocol. Modern data centers zeroed in on BGP as the preferred control-plane protocol. Because the infrastructure is built on IP, many advantages are leveraged including loop-free communication using industry-standard routing protocols, Equal Cost Multi-Path (ECMP), superior scalability (compared to traditional networks design and Layer 2 based fabrics), and standards-based interoperability. Other advantages of the IP Fabric include high-bandwidth scaling, predictable low-latency, and nonblocking server-to-server connectivity.

Some of the historical issues resolved with this architecture would include a familiar list of problems (now in the past): unpredictable latency, the need for virtual machine mobility, and server port scale-out requirement for data centers. In a traditional network with a 3 tier architecture with distinct Access/Aggregation/Core layers, latency was inconsistent, and ports were blocked due to spanning tree loop

detection. Some organizations tried to utilize static unused resources by load balancing services across disparate VLANs from unused ports. Hence, extensive resources were utilized to design and implement "programmed complications" into the network design, simply to allow data traffic to 'occupy' what would be unused resources in a relatively inefficient manner. Once live migration of storage and virtual machines became commonplace, any-to-any server port connectivity was a needed. The 1st generation fabrics delivered against this requirement; however, many vendors took a proprietary approach. These first-generation layer 2 based fabrics were limited in interoperability and scale. The IP fabric provides the any-to-any connectivity, predictable low latency, embedded security, and uniform availability.

IP Fabric offer unmatched scalability in comparison with Layer 2 fabrics, where the number of ToR switches, VLANs, MAC entries, IP subnets, ARP/ND entries, route scale, etc. are limited by the L2/L3 boundary, typically found at the spine of a Layer 2 fabric. The implication is that the VLANs or broadcast domains must be pruned properly per membership or interest to avoid large broadcast domains. In IP Fabric implementations, the L2/L3 boundary resides at the ToR switch.

This type of topology has the predictable latency and provides the ECMP (Equal Cost Multi-Path) forwarding in the underlay network. The number of hops between two leaf devices is always limited to two hops within the fabric. This topology also enables easier scale-out in the horizontal direction as the data center expands and is limited only by the port density and bandwidth supported by the spine devices.

As a design recommendation, to provide for better performance predictability, hardware mixing at the spine layer is not recommended. For example, at the ToR layer, it is fine to utilize a mix of switch types, for example SLX 9150 or SLX 9250 to accommodate various link speeds at the access layer. However, at the spine layer, where uniform scaling is desired, it is recommended to use "like" devices. Validated Spine and Leaf devices are depicted in the "place in the network" (PIN) chart below (see Figure 1).

			Places in the Network (PIN) Civilian/Enterprise/IC										MILDEP/JITC/IC				
Extreme Networks Platforms	Buffer	Application	Leaf	Spine	Super Spine	Border Leaf	Non-Clos (IP Fabric)	DCI	Border Router	L2/L3 MPLS (RSP)	Ring / G.8032 (RSP)	MPLS P & PE (IXP)	ADRP (DISA)	Core (MCN/CN-CEB-CDN/IR)	Aggregation (ADN/DN/DR)	ASLAN (Access/Agg/ Core EUB/CAN/AN)	SF Server Farm
SLX 9150	Shallow	Access	✓				✓										✓
SLX 9250			✓	✓			✓									*See Extreme*	✓
SLX 9540	Deep	BGP-EVPN VXLAN, VLL/VPLS/ MPLS	✓			✓		✓	✓	✓	✓	✓	✓	✓	✓	*Campus JITC/APLITS Listings*	✓
SLX 9640						✓		✓	✓	✓	✓	✓	✓	✓	✓		✓
SLX 9740				✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
SLX 9850				✓	✓								✓	✓	✓		✓
Data Center IP Fabric								Service Provider -Transport Backbone					Military Bases / Guard Units / Nato Partners				

Figure 1: Extreme Networks Guide to 'Places in the Network" (PIN). Referred to as a PIN chart.

The IP Fabric consists of an Underlay, and an Overlay. The underlay includes the physical hardware infrastructure which is made up of elements and the physical topology that provides the IP connectivity. The overlay is layered over the hardware to implement the logical construction. Some engineers include the control protocol in their definition of underlay, while defining the overlay as including protocols such as VXLAN.

Underlay Achieving Any-to-Any Connectivity for the Agile Data Center

One of the primary goals in designing an architecture where any to any (application, service or resource) connectivity is to deliver a design where access is scaled, but the number of connections are controlled. It has been proven throughout the history of communications networking that achieving an any-to-any design can easily become cost prohibitive if not executed carefully. In the 1940's (telephony) access lines were shared, and the

supporting circuitry were eventually developed to have central processing point of all connections performed by the local central office. By the 1970's local in-building post branch exchange (PBX) switches were implemented with a numbering plan and a hierarchy to make possible connections from any desk, to any other desk with minimal wiring and circuitry. With the advent of bandwidth managers in the 1980's and early 1990's paths were shared and trunked to central nodes where better decisions could be made about call placement or data switching. By the time packet switched services became prevalent (example ATM), it became apparent to designers that achieving N x N connectivity was needed for nodes offering transport to deliver the services. The resulting technology that would resolve this issue was ATM switched virtual circuits, where intelligent nodes routed the traffic between other intelligent nodes. However, the process became unwieldy because the cost of achieving any to any connectivity became design intensive, it was typically manually implemented and the advent of competing IP systems were offering methods to groom the traffic, preserve the bandwidth needed, and offered prioritization using methods such as ToS (Type of Service) and differentiated services had become widespread. The next problem introduced was the delay in the decision-making capability of the routers involved to perform all these functions at scale, while the adoption of disparate features and the network buildout were still underway.

With virtualization and advanced computing methods, the intelligence again shifted. We witnessed the shift from an intelligent core design model to one where the network elements make decisions regarding services. The approach proved to be fragmented, time-consuming and complex to implement. Modern virtualization methods demanded multi-channel access capabilities; thus, all the ports needed to become available to all services without further complication being added. By the 2010's vendors offered multi-pathing, with controlled latency providing any to any connectivity in the data center. In the data center became the central point where services had become consolidated due to many factors, such as virtualization and efficiency-based practices such as green IT (consumption). Because of consolidation, facilities were designed specifically for the efficient use of network, environmental, compute and storage resources. Disciplines arose (arguably resurrected) to measure and compare the efficiency of the facility where the data center was built, for example PUE (Power Utilization Effectiveness), a measurement ratio of the amount of power required to deliver a watt from energy source to the equipment. Just as the service provider delivered services in the 1990's and previously that had the power required for delivery supplied by the central office, the data center supplies power to the equipment providing the service. Employing practices that deliver efficiencies in resource utilization become the norm.

The industry coalesced around the premise that costs needed to be controlled, age-old problems such as spanning tree needed to be resolved, and traffic flow behavior needed to be assured, and predictable. In fact, the typical data center traffic flows ran east and west from application server to other application servers within the data center walls. It grew to represent 80% of the total traffic within the data center network system. In some data centers, the Northbound Traffic shrank to 3-5% of the entire data center traffic.

Some systems, such as the Extreme Networks Virtual Clustering System (VCS), addressed these problems and the speed to implementation issues by automating the underlay. The VCS was designed to operate efficiently, but also created with an efficient architecture that reduced the tiers in the data center network from 3 to 2. The architecture eliminated the aggregation tier. But what underlay was to be automated? Some of these early fabric implementations had their virtues. Would the benefits of TRIILL (Brocade/Extreme), SPB (Avaya/Extreme) and other networking protocols carry over to the next generation system? How could it be made to function more efficiently, accommodate increased speeds (40G/100G or 400G) that typically carried higher adoption costs than 1G or 10Gb systems? With so many choices of early competing fabrics, many of which had fallen short on scale, had also become obsolete because they did not address ease of implementation. In some cases, the fabric functionality, (due to the complicated implementation programming required), didn't even get turned on. Vendors went back to the drawing board, looking back at early switching architecture

methods that would address these issues. Problems that were resolved in the first generation of fabrics (i.e. TRILL, SPB), such as spanning tree, MAC tables, constrained forwarding tables, fine grained labeling and automation are resolved with new approaches with IP Fabric.

Where Did Data Center Equipment Design Engineers Look?

While it cannot be attributed directly to every designer, solving the any-to-any problem, throughout history, can be traced back to a common point. While problem has been referred to in later years as Metcalfe's Law, or - described as the effect of a telecommunications network is proportional to the number of users of the system (n^2). While Metcalfe's law deals primarily with the internet today as a system, the same problem was studied much earlier at a circuit level by research engineer who worked for Bell Labs. In 1953 Charles Clos author of "The Bell System Technical Journal (Volume: 32, Issue: 2, March 1953)" A Study of Non-Blocking Switching Networks, proposed a methodology to implement switching in a manner that controlled the path of a circuit and the scale at which it could be implemented to the number of nodes that enable non-blocking designs with providing absolute dedicated paths between all services within the architecture. One of the key problems solved was that whenever a new connection was needed or implemented, only some of the resources required any type of rearrangement. The problem to be solved in his research was to identify a means to reduce the number of cross points required for any 2 Input/Outputs of a circuit at scale. This has also by referred to as the n^2 problem.

One of the outcomes of the paper explained how a crosspoint system could resolve switching arrays in stages. The chart below illustrates that the Value of N (64 in the example), number of connections, in our case nodes/routers/switches in a square array required 4,096 crosspoints. In a 3-stage array, that number was reduced to 2880. In a 5-stage array, this value increased value of 3248 crosspoints. From the chart below, it would seem that 5-stage arrays require more crosspoints than a 3-stage array. However, as the number N scales up, the number of crosspoints in a 5-stage array increases at a far lower rate than a square array. Using this design approach worked in latter-day technologies such as crossbar architectures in routing switches, but would this approach work for the data center as a system? It would be far more efficient to address the connection of up to 10,000 ToR switches in an any-to-any fashion, without having to make a direct connection from every switch.

Table II - Crosspoints for Several Values of N

N	Square Array	3-Stage Array	5-Stage Array
64	4,096	2,880	3,248
729	531,441	115,911	95,013
1000	1,000,000	186,737	146,300
10,000	100,000,000	5,970,000	3,434,488

Figure 2: Extract from "Bell195303 C Clos- a Study of Non-Blocking Switching Networks"

Extreme Networks has adopted this approach for implementation, being mindful of prioritization, security, any-to-any connectivity, east-west utilization, and preserve the automation features that thousands of the company's customers embraced, implemented and continue to run in their data centers today.

At Extreme Networks, the design and architecture solution based upon this methodology became known as IP Fabric. The underlay network equipment would be running services such as BGP with ECMP for multi-pathing. The physical topology supporting the platform is primarily based upon 3-stage and 5-stage Clos architectures. Not every customer needs a data center at scale, so Extreme also offers a small data center option referred to as the Collapsed Core (Non-Clos) data center. One of the goals was to preserve the number links between systems, and still accommodate rapid deployment, easy expansion and scale.

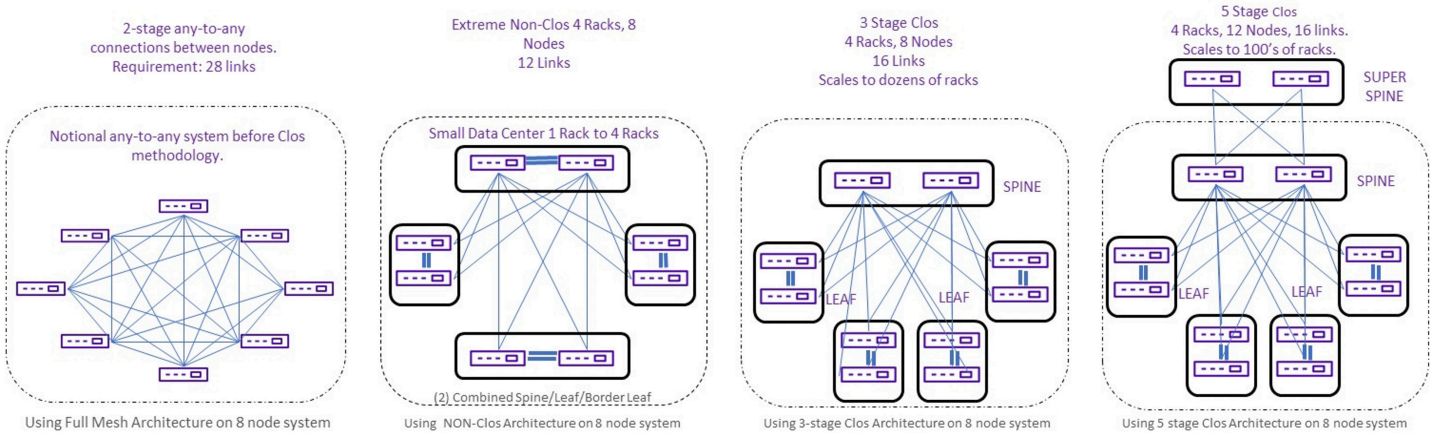

Figure 3: Basic example of 2-Stage N^2, small Non-Clos, 3-Stage and 5-Stage Clos Implementations. Like the chart in figure 2. A Clos based architecture may require more initial system elements for any to any connectivity, but scales beyond traditional systems and methods and gains greater efficiency as the data center fabric grows.

3-Stage Clos Architecture

The 3-Stage Clos Architecture is based on a spine-leaf 3-stage Clos network, is designed to deliver the same performance and expansion capability and provides a graceful means to grow the solution to the size of the mission. It has two tiers to the architecture, a Leaf layer to connect all the server and compute elements to the network, and a Spine layer that ensures that all access interfaces from the leaf layer may be reachable within the PoD (Point of Delivery). The Spine layer also connects to the Border Leaf for access to the ingress and egress to/ from the network backbone. The Border Leaf is also the point where non-IP Fabric may connect to the data center, such as VDX based Clusters (VCS fabrics), Campus Fabrics or traditional IP network architectures. So why is the 3-stage Clos architecture considered better than traditional Access-Aggregation-Core (3 tier legacy data center) architectures? In traditional networks, the path from server port to server is not always uniform, therefore latency is unpredictable. In a 3-stage Clos architecture, the spines are of like devices form factors (example SLX 9250 or SLX 9740), latency is predictable and uniform across the fabric from server to server port.

Data Center Fabric Architecture– 3-Stage Clos

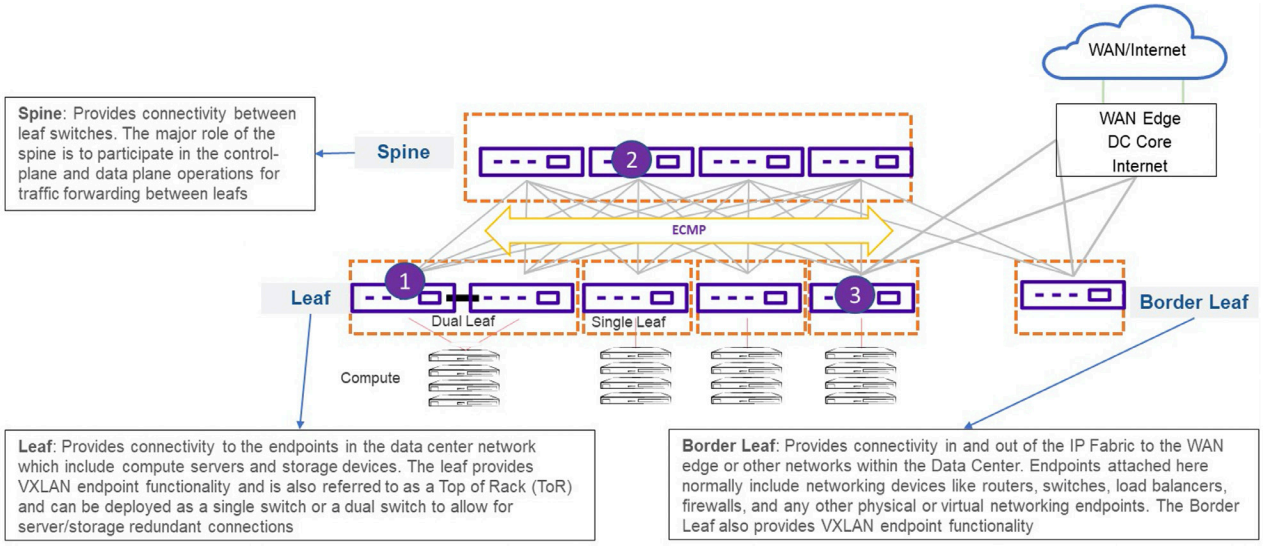

Figure 4: The 3-Stage Clos Architecture as the foundation of a Data Center Point of Delivery (PoD).

As a design principle, the following requirements apply to the leaf-spine topology:

- Each leaf connects to all spines in the network through 40G or 100G Ethernet link(s)
- Spines are not interconnected with each other
- Leafs are not interconnected with each other for data-plane purposes. (Two leafs may be interconnected for control-plane operations, such as forming a server-facing vLAG. This is referred to as vLAG pair leaf.)
- Multi-chassis trunking (MCT) at the leaf pair for dual homing end devices, (i.e. servers)
- LLDP is used to discover the connections between leaf and spines
- The network endpoints connect to leaves, no connection to the spines
- A Single instance of eBGP is used (alternatively iBGP may be used, but no IGP is required, next hop reachability via LLDP)
- Single IP per switch (loopback/router ID) for BGP peering
- Private use Autonomous System number assignments are used (16 and 32 bit AS numbers reserved by IANA)**
- Bidirectional Forwarding Detection (BFD) for faster convergence
- MD5 Hash for BGP peer authentication

*** IANA "Autonomous System (AS) Numbers" registry*
http://www.iana.org/assignments/as-numbers/

The Clos topology has the predictable latency, fast convergence, high availability and provides the ECMP forwarding in the underlay network. The number of hops between two leaf devices is always two within the fabric. This topology also enables easier scale-out in the horizontal direction as the data center expands and is limited by the port density and bandwidth supported by the spine devices.

Once the 3-stage Clos based IP fabric is constructed and connected to the compute, storage and application components, a 3-stage Clos IP Fabric would be considered an element of a PoD, or point of delivery for service. The idea of a PoD is to provide a repeatable design with common components; which helps with manageability, scalability, and common operation. What if there is more than one Point of Delivery? How do you get an application in one PoD reachable to another application instance residing in another PoD?

5-Stage Clos Architecture

Multiple PoDs based on leaf-spine topologies can be connected for higher scale in an optimized 5-stage folded Clos (three-tier) topology. Where the 3-stage Clos consisted of a Leaf and Spine tier, the 5-stage Clos topology adds a new tier to the network, known as a super-spine. Super-spines function like spines: they utilize a BGP control and data-plane. Traffic forwarding between the PoDs crosses the Super-spine; while traffic from the PoDs to destinations outside the fabric is passed from the super-spine via the border leafs. Also note that no endpoints are connected to the super-spines. Figure 5 shows four super-spine switches connecting the spine switches across multiple data center PoDs.

Data Center Fabric Architecture – 5-Stage Clos

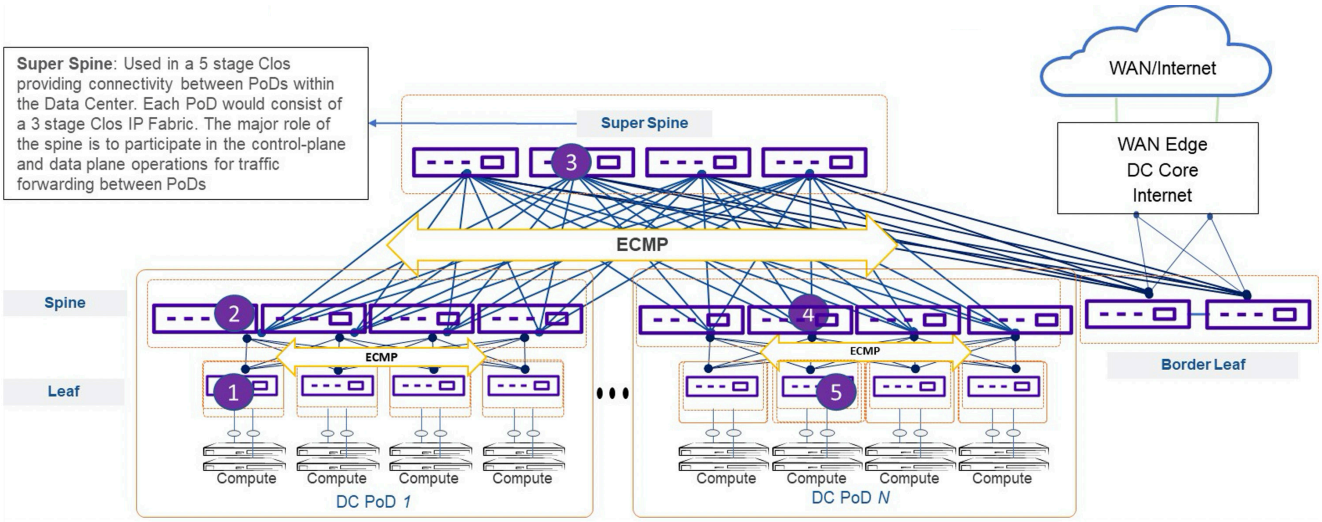

Figure 5: The 5-Stage Clos Fabric Architecture address the need to scale by the utilization of a Super-Spine, which connects PoDs that contain 3-Stage Clos fabrics.

The connection between the spines and the super-spines follows the Clos principles:

- Each spine connects to all super-spines in the network.
- Neither spines nor super-spines are interconnected with each other.
- Switching between PoDs occurs via the Super-Spine
- Any-to Any connectivity is maintained.

The 5-Stage Clos enables the fabric to maintain reachability between PoDs and Border Leaf while simultaneously increasing scale without resorting to the N squared connection points. As a result, relative economy is achieved in that the number of connections required for any application in PoD 1 could access resources in the other PoDs (N).

Summary

Extreme Networks delivers Data Center IP Fabric that scale from 2 to 8 switches in a collapsed core configuration, or to larger PoDs based upon 3-stage Clos architecture designs. The PoDs are part of a larger data center fabric based upon 5-Stage Clos architectures that maintain any to any connectivity with large buffers at the access layer (Leaf layer) and shallow buffers at the Spine and Super-Spine Layers to ensure high performance and efficient ingress and switching of data center traffic from any point in the data center to either a local Leaf, to another Leaf in different PoD or to the egress of the data center at the border leaf. The use of Clos based architectures ensures a methodology of expanding and scaling out the data center IP fabric, deployed with a methodology that ensures an efficient deployment of the underlay. The Clos based design uses a proven methodology of ensuring efficient delivery of any to any connectivity. Just as using new disciplines like designing to lower PUE for power management, the use of 3- and 5-stage Clos architectures ensure efficient use of hardware resources to deliver any to any connectivity. The IP fabric based data center is the product of furtherance in the design evolution in the data center with a nod to the past. It provides a disciplined resource-conscious, solid foundation for greater efficiencies to deliver enterprise services. More product information can be found in the Extreme Networks SLX Switching and Routing Portfolio.

Continue to Volume 3: Evaluation of Reliability and Availability of Network Services in the Data Center Infrastructure

This document discusses the industry standard methodologies of calculating, for purposes of comparison, various underlay element types. It compares single form factor switching elements at the Leaf, Spine and Super-spine Layers with their chassis-based counterparts, and identifies the mean-time-between-failures (MTBFs), Mean-time-between-repairs (MTTR), and their relation to element and network availability. It also provides availability measurements and the unavailability calculations based upon the elements placement in the network, and the number of port connections between compute and application services platforms.

©2021 Extreme Networks, Inc. All rights reserved. Extreme Networks and the Extreme Networks logo are trademarks or registered trademarks of Extreme Networks, Inc. in the United States and/or other countries. All other names are the property of their respective owners. For additional information on Extreme Networks Trademarks please see http://www.extremenetworks.com/company/legal/trademarks. Specifications and product availability are subject to change without notice. 37641-0421-09