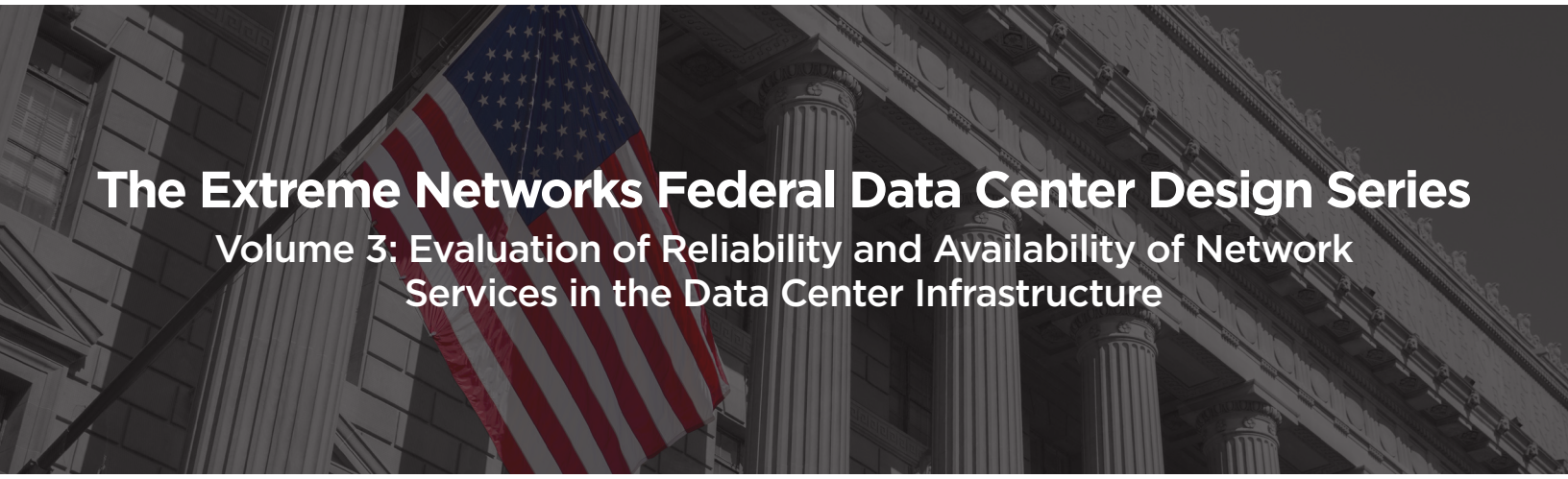

The Extreme Networks Federal Data Center Design Series

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

Overview

Volume 2: Achieving Any-to-Any Connectivity with Time-Tested Design Approaches and Proven Methodologies described 3-Stage and 5-Stage Clos architectures and how they support any-to-any connectivity. This document describes the arrangement of the elements in the architecture and how it relates to reliability, leveraging the availability calculation method that would typically be used and how various equipment configurations may be compared.

Measuring Reliability for the Data Center

As mentioned previously, one of the top priorities of the Office of Management Budget (OMB) is to maintain high availability of services; which requires a reliable network infrastructure as its foundation. It may seem obvious, but reliability in the data center is of greater importance today for a few reasons.

- The amount of data traffic being processed in the data center, I.e. server traffic, is doubling every 12 to 15 months
- The amount of traffic flowing to and from the data center compute, storage, security, networking and application based in federal architectures constitutes nearly all mission critical network traffic

The importance of reliability and availability become more critical with each passing monthly traffic report. Another reason that the actual hardware being deployed must be assembled into a resilient design, is that the protocols being used to deploy the data center use policy, not metrics, and peering, not link-state protocols to preserve multiport connections. Therefore, the age-old consideration of 1RU/2RU form factors versus chassis-based equipment presents itself.

One of the benefits of the IP Fabric deployed in a Clos based architecture is that it provides efficiency in establishing any-to-any connections. As a networking connection system, versus the switching element within a product design (original evaluation by Clos 1953) provides the higher network system availability; this a direct result from the overlay and underlay working together to keep services running for the subscriber community of the various applications hosted in the Data Center PoD (Point of Delivery). Because of utilizing a single form-factor based infrastructure, deployed in a framework that reflects a crossbar switching platform (like modern chassis-based packet switching and routing systems), the IP Fabric underlay provides the basis for a system availability that, model dependent, exceeds several nines availability (>5).

One standard that assists planners in selecting high availability elements for their design is the manufacturer calculated MTBF (mean-time between failure). To determine the element availability and the system availability the MTBF of the individual devices is used to and calculate the network availability. A risk assumption of 8 hours MTTR (mean time to restoral) was added. To determine system availability, we utilized the element availability rate of the

device and used it to calculate system availability using the parallel availability method using the model of a 4 leaf (A), 4 spine system to a 4 Leaf system (B) as the baseline. This is a calculation based upon the element MTBF, the organizations MTTR (mean-time to restoral) targets, to arrive at the availability (A).

Element A = MTBF/(MTBF+MTTR)

Calculating the Availability of a Single Form-factor based Fabric

While Clos architecture principles were originally applied to the transistor and switching chipset technology of the time, the model works well for a system architecture, if the individual element MTBF/Availability numbers are known.

Expressed below is the availability calculation using the typical leaf switch element of a single form factor 1RU (rack unit) system using the parallel availability calculation.

Parallel System A = (L1) x (L2) = #9's

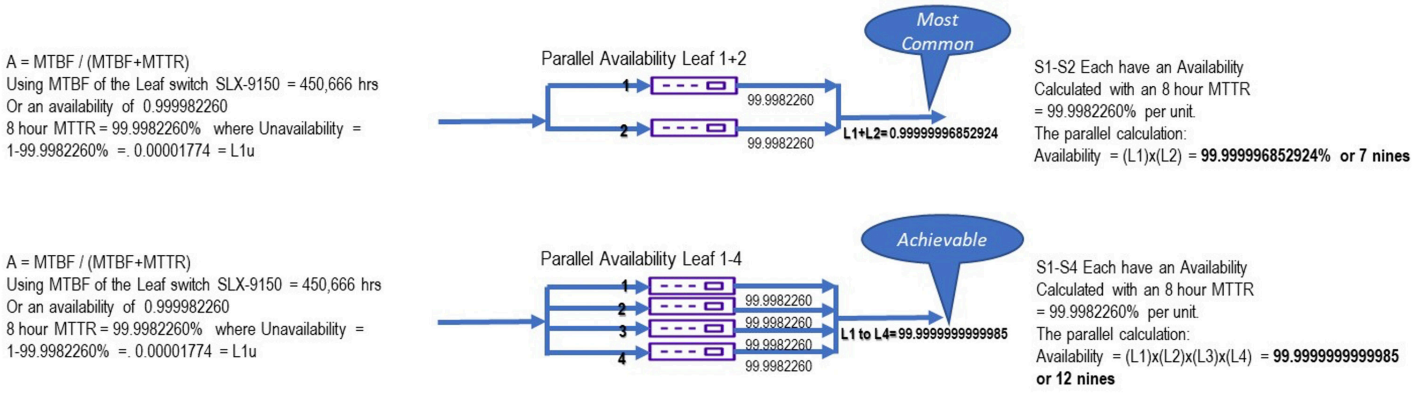

Figure 1: Availability Calculation of a parallel leaf-based system* with 2 access ports to the Leaf, and 4 access ports to the leaf.

Note While many restoral models or SLA's regarding mean-time to restoral may be calculated to include the lowest numbers, such as 15-minute hot-swappable cards, or 4-hour on-site engineering support, for this example, increased (longer) restoral or outage time up to 8 hours were used in the calculation (versus 0:15hrs or 4:00hrs). additionally, the availability model for the IP Fabric has also been calculated the availability based upon a 24-hour outage and the resulting availability calculation was not lowered by a single nine.*

Thus, the Leaf (Access to the fabric) availability, (network availability) to the server is 99.9999999999985% or 13 nines. The Spine calculation would be performed in the same manner as a Leaf system. As shown below, the availability of the Leaf-Spine-Leaf fabric architecture utilizes a serial calculation where the unavailability is added L1-4u (Leaf 1 through 4 unavailability) to S1-4u (Spine elements 1-4 unavailability) to L5-L8u (3rd stage of the traffic flow in a server to server model Leaf 5-8 unavailability) to arrive at the system availability from server to server for East West traffic. In a Typical Data Center, the traffic flowing East-West accounts for up to 80% of the traffic across the fabric.

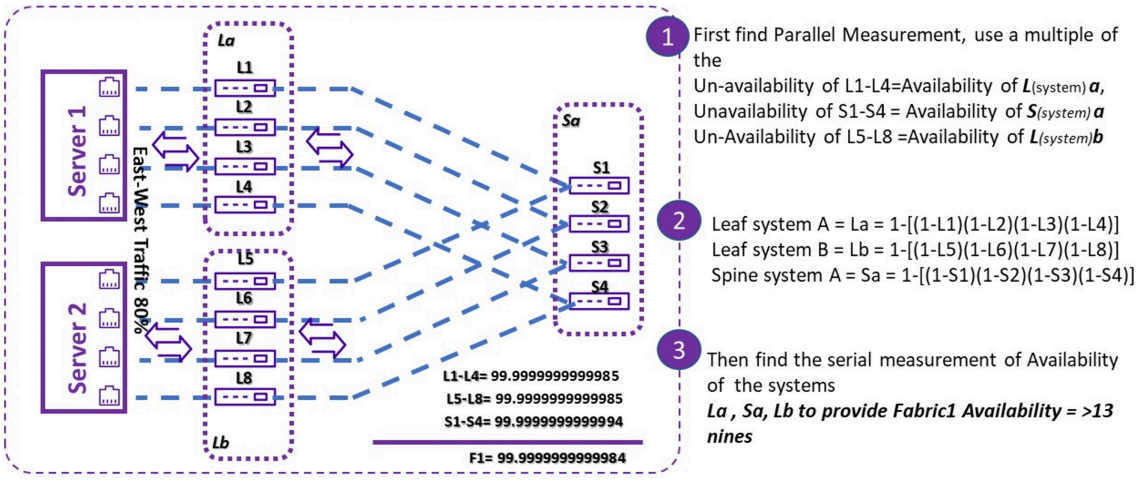

Figure 2: Availability Calculation of 3-stage Clos PoD Architecture based upon Extreme Networks SLX-series switches

The resulting availability between the elements of the 3 stage Clos based underlay of the IP fabric F1 (Fabric 1) or PoD is 99.9999999999984 or 13 nines. This amounts to an expected 5.07673902916395000000000e-07 seconds unavailability per year, or could be also expressed as 5.0767395 nanoseconds/yr. This equates to slightly > one half of one billionth of a single second per year.

A three stage Clos configuration yields very high availability as a direct result of the architecture used. In addition, we maintain the benefit of the 3-stage Clos maintaining any-to-any capability, while delivering ultra-high availability (>5 nines, in this case 13 nines). Systems with lower MTBF values may yield lower system availability overall, but what may be the question to ask is: Does this design create an unacceptable level of unavailability?

Calculating the Availability of a Chassis-Based Fabric

A chassis-based infrastructure yields similar results, as we view network availability to a set of servers to a fully redundant SLX 9850 chassis-based system. The chassis-based hardware, when configured with the appropriate level of N+1 redundancy componentry. When calculating the individual common equipment within the system, we find that these elemental blocks such as Switch Fabric (S1,S4,S3,S4) Power Supplies (P1, P2, P3, P4,), Fan assemblies (F1-F6) can be calculated as mathematically 100% availability of the system. Due to such high levels of redundancy, the chassis common equipment number for unavailability can be lower that that <1/googol of a second (128 places). This is achieved because of extensive hardware redundancy that consists of subcomponents working in N:1 or N+1 redundant configuration. For example, the switching fabric modules measures 8 nines availability. To arrive at the load sharing availability of all installed switch fabrics (Sa) supporting parallel paths from the line cards, we multiply the unavailability of switch module S1-S6. The difference between single form factor and the chassis-based solution, when implemented in a PoD yields similar results.

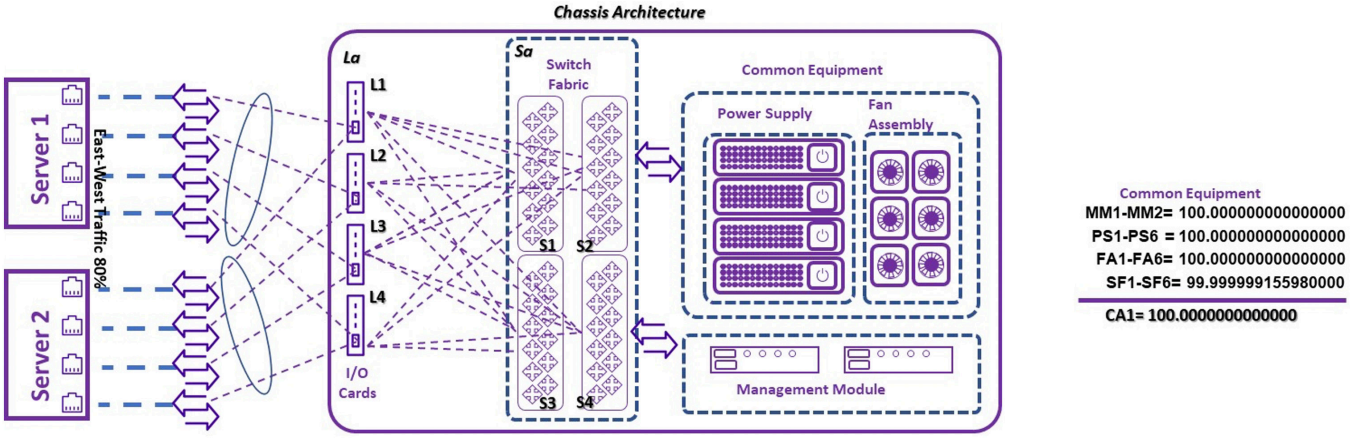

Figure 3: Chassis Architecture

The general conclusions can be made that the chassis-based infrastructure buys an additional ½ of a nanosecond availability when compared to the single form factor implementation. The exercise also demonstrates that the cost of embedding more than 3 access links (LAG) or Leaf to Spine connections achieve connectivity assurance, but do not yield significantly higher availability. In this model the 3 stages at server to Line card (L1-L4) with parallel connections to switch fabrics 1-6 (S1-S6) (stage 2) back to the egress line card to the server port (3rd stage).

To conclude, when costs and ease of expansion are factored into the decision, a single form factor implementation may be less expensive to acquire, deploy and operate and better facilitate a 'pay-as-you-go' cost modeling. While traditionally in the past, the importance of the traffic, the amount of growth and proven higher availability of chassis systems dictated their usage, the overall system configuration (3/5 stage Clos based PoDs with 1RU/2RU elements) negates the added benefit vs. cost consideration typically incurred by the selection of the chassis solutions.

There is, however, one inescapable benefit of utilizing the chassis-based system as an element within an IP fabric data center. Which is that it enables higher scale at the access layer while reducing the number of Spines. This should be weighed against the impact of a failure of one of the Spine switches, versus a single form factor-based Spine switch (Risk, Mean Time to Recovery). In a similar configuration with a 1RU based single form factor spine consisting of 8 units with 8 x 100G links from Leaf to Spine, if one unit fails, only 12.5% of the overall bandwidth is lost. When a chassis based system fails as a Leaf, 100% of the access bandwidth is gone, or if deployed as a spine (with 8 x 100G links from the Leaf layer to 2 units of 4 slot SLX 9850's) 50% of the bandwidth is lost.

Therefore, the general question to ultimately ask the technical director or program manager is:

- Does the Data Center planner want to risk 50% of the Data Center PoD bandwidth to a system that has a < ½ of a nanosecond of unavailability per year risk?
- Or does the Data Center planner want to risk 12.5% of the data center PoD bandwidth to a system with a slightly > ½ nanosecond of calculated unavailability per year.

Where the decision may have been clear in previous decades that the risk involved with choosing a chassis system or not choosing a chassis may have appeared clear, today's choice involves splitting mere nanoseconds.

How do Extreme Networks Data Center elements compare using this availability model?

The resulting network availability calculations for each model of SLX Switches in the Federal portfolio are provided in the chart below. In this chart we have provided the availability/unavailability for a 1,2 and 4 port access port channel deployment. We have indicated the number of 'nines' that the availability level achieves with an 8 hour mean time to restoral (MTTR). The availability is calculated and expressed as a percentage. As you will note from the chart below, The unavailability is expressed in seconds (1 unit) milliseconds (2 ports, 2 units), micro-seconds (not shown- 3 ports with 3 separate units) and nanoseconds (unit of measurement needed for 4 ports over 4 access/leaf switches). As you will note that the availability calculation for 4 ports connected to 4 separate leaf switches is 100% (<1 Nano-second to 10's or hundredths of pico-seconds.)

Time measurement calculation definitions in the chart:

Millisecond	Millisecond is a unit of time. It is equal to 0.001 second.
Microsecond	Microsecond (μs) is a multiple of the second, a unit of time, prefixed by the standard-base multiplier micro (μs), equal to 0.000001 second
Nanosecond	Nanosecond (ns) is a multiple of the second, a unit of time, prefixed by the standard-base multiplier Nano (n), equal to 0.000000001 second
Picosecond	Picosecond (ps) is a multiple of the second, a unit of time, prefixed by the standard-base multiplier Pico (p), equal to 1e-12 second.

The chart determines the calculated availability based upon the MTBF (Mean Time Between Failure) and the MTTI (Mean-Time-To-Innocence). For example, a single Leaf access port for a server connected to an SLX-9150 would be provided 4 nines or 99.99823% availability; however, with a 4-port access configuration where the access links connect to 4 separate Leaf switches. Over 9 nines of availability are calculated. This represents the access tier.

When the same SLX-Switch models are utilized for the Spine, with multiple connections, the same calculation for that tier of the PoD is performed and the Spine availability then known. After these calculations are done, a series calculation is performed to provide the availability of any to any connections within the data center.

Leaf to Spine Series Unavailability = the unavailability of 2 port access (L1/L2 parallel unavailability) + Unavailability of 2 spine connections (S1/S2 parallel unavailability).

To arrive at the L1/L2 unavailability and the S1/S2 unavailability for a 2-port access/Leaf connection traversing a 2-spine configuration in a 3 stage Clos architecture the unavailability of each tier in the model is additive, and then subtracted from the availability.

Thus, the access tier (2 switches L1 and L2) unavailability of an SLX-9150 is 0.000000000314724749117937 (year). And the Spine tier (If also constructed with SLX 9150, S1 and S2) would be 0.000000000314724749117937 (year). Expressed a formula:

(Leaf Layer Unavailability) + (Spine Layer Unavailability) = Fabric Unavailability (year)

Or as follows:

0.00000000031472474911793 7 + 0.00000000031472474911793 7

(L1/L2 Leaf layer unavailability) (S1/S2 Spine layer unavailability)

= 0.00000000062944949823587 4

(total Fabric unavailability)

When expressed as the number 'nines' of availability the formula is as follows:
#"nines" = 100% (time) minus total (Serial Leaf-Spine unavailability) = 9 nines

100% – total unavailability or 0.00000000062944949823587 4) = 99.9999999370551%.

This method of calculating the Leaf to Spine Layer availability is performed by taking the parallel calculation of the Access/ Leaf and Spine tiers separately and plugging their values into the serial calculation to identify the unavailability. We then take the unavailability per year calculation and subtract it from the 100%, expressed as 1, of the time in one year.

Product	"MTBF Calculated Value (Hrs)"	"(MTBF(MTBF+ MTTR) (8hrs)"	Unit 1 Unavailability	Time (Seconds)	Unit 1 Availability %	Nines (Availability)	2 Unit Unavailability %	Unavailability Time (Milliseconds)	2 Unit Availability %	Nines (Availability)
		1 Unit Standalone					2 Units Access to Leaf Port			
BR-SLX-9140-48V-DC-F or R	303683	99.99737%	0.000026342565	1,385	99.9973657434695%	4+	0.0000000069393075%	36.47	99.9999999306069%	9+
BR-SLX-9140-48V-AC-F or R	324414	99.99753%	0.000024659240	1,296	99.9975340759874%	4+	0.0000000060807812%	31.96	99.9999999391922%	9+
SLX9150-48XT-6C-AC-F	363534	99.99780%	0.000022005710	1,157	99.9977994289518%	4+	0.0000000048425129%	25.45	99.9999999515749%	9+
SLX9150-48XT-6C-AC-B	380460	99.99790%	0.000021026735	1,105	99.9978973264506%	4+	0.0000000044212361%	23.24	99.9999999557876%	9+
SLX9150-48Y-8C-AC-F	385325	99.99792%	0.000020761264	1,091	99.9979238736366%	4+	0.0000000043103007%	22.65	99.9999999568970%	9+
SLX9150-48Y-8C-AC-B	450938	99.99823%	0.000017740483	932	99.9982259516661%	4+	0.0000000031472475%	16.54	99.9999999685275%	9+
BR-SLX-9240-32C-AC-F	327539	99.99756%	0.000024423976	1,284	99.9975576024204%	4+	0.0000000059653059%	31.35	99.9999999403469%	9+
BR-SLX-9240-32C-AC-R	327539	99.99756%	0.000024423976	1,284	99.9975576024204%	4+	0.0000000059653059%	31.35	99.9999999403469%	9+
BR-SLX-9240-32C-DC-R	306419	99.99739%	0.000026107360	1,372	99.9973892640009%	4+	0.0000000068159425%	35.82	99.9999999318406%	9+
BR-SLX-9240-32C-DC-F	306419	99.99739%	0.000026107360	1,372	99.9973892640009%	4+	0.0000000068159425%	35.82	99.9999999318406%	9+
SLX9250-32C-AC-F	377974	99.99788%	0.000021165029	1,112	99.9978834970978%	4+	0.0000000044795845%	23.54	99.9999999552042%	9+
SLX9250-32C-AC-R	450666	99.99467%	0.0000532516808	2799	99.9946748319244%	4+	0.0000000283574150%	149.05	99.9999997164258%	9+
BR-SLX-9540-48S-AC-R and -F	306419	99.99739%	0.0000261073600	1372	99.9973892640009%	4+	0.0000000068159425%	35.82	99.9999999318406%	9+
BR-SLX-9540-24S-DC-R and -F	327539	99.99756%	0.000024423976	1,284	99.9975576024204%	4+	0.0000000059653059%	31.35	99.9999999403469%	9+
EN-SLX-9640-24S-12C- DC	306419	99.99739%	0.000026107360	1,372	99.9973892640009%	4+	0.0000000068159425%	35.82	99.9999999318406%	9+
SLX-9640-24S-12C-AC-F	327539	99.99756%	0.000024423976	1,284	99.9975576024204%	4+	0.0000000059653059%	31.35	99.9999999403469%	9+
SLX9740-40C-AC-F	189747	99.99578%	0.000042159627	2,216	99.9957840373113%	4+	0.0000000177743414%	93.42	99.9999998222566%	8+
SLX9740-40C-AC-F	106232	99.99247%	0.000075301205	3,958	99.9924698795181%	4+	0.0000000567027145%	298.03	99.9999994329729%	8+
SLX9740-80C-AC-F	131836	99.99393%	0.000060677771	3,189	99.9939322229301%	4+	0.0000000368179186%	193.52	99.9999996318208%	8+
SLX9740-80C-AC-F	64300	99.98756%	0.000124401319	6,539	99.9875598681346%	3+	0.0000001547568808%	813.40	99.9999984524312%	7+

Table 1: Calculated Parallel Availability for 1,2 and 4 port Server port groups when accessing an IP Fabric system by Product.

Product	4 Unit Unavailability %	"Time (Nanoseconds)"	4 Unit Availability (Network)	3 Unit Unavailability %	Time (Microseconds)	3 Unit Availability %	Nines (Availability)
	4 Units Access to Leaf Port			3 Units Access to Leaf Port			
BR-SLX-9140-48V-DC-F or R	0.00000000000000000482%	0.0253	100.0%	0.0000000000001828%	0.96	99.99999999999982%	13+
BR-SLX-9140-48V-AC-F or R	0.00000000000000000370%	0.0194	100.0%	0.0000000000001499%	0.79	99.99999999999985%	13+
SLX9150-48XT-6C-AC-F	0.00000000000000000234%	0.0123	100.0%	0.0000000000001066%	0.56	99.99999999999989%	13+
SLX9150-48XT-6C-AC-B	0.00000000000000000195%	0.0103	100.0%	0.0000000000000930%	0.49	99.99999999999991%	14+
SLX9150-48Y-8C-AC-F	0.00000000000000000186%	0.0098	100.0%	0.0000000000000895%	0.47	99.99999999999991%	14+
SLX9150-48Y-8C-AC-B	0.00000000000000000099%	0.0052	100.0%	0.0000000000000558%	0.29	99.99999999999994%	14+
BR-SLX-9240-32C-AC-F	0.00000000000000000356%	0.0187	100.0%	0.0000000000001457%	0.77	99.99999999999985%	13+
BR-SLX-9240-32C-AC-R	0.00000000000000000356%	0.0187	100.0%	0.0000000000001457%	0.77	99.99999999999985%	13+
BR-SLX-9240-32C-DC-R	0.00000000000000000465%	0.0244	100.0%	0.0000000000001779%	0.94	99.99999999999982%	13+
BR-SLX-9240-32C-DC-F	0.00000000000000000465%	0.0244	100.0%	0.0000000000001779%	0.94	99.99999999999982%	13+
SLX9250-32C-AC-F	0.00000000000000000201%	0.0105	100.0%	0.0000000000000948%	0.50	99.99999999999991%	14+
SLX9250-32C-AC-R	0.00000000000000000841%	0.4227	100.0%	0.0000000000015101%	7.94	99.99999999999849%	14+
BR-SLX-9540-48S-AC-R and -F	0.00000000000000000465%	0.0244	100.0%	0.0000000000001779%	0.94	99.99999999999982%	13+
BR-SLX-9540-24S-DC-R and -F	0.00000000000000000356%	0.0187	100.0%	0.0000000000001457%	0.77	99.99999999999985%	13+
EN-SLX-9640-24S-12C- DC	0.00000000000000000465%	0.0244	100.0%	0.0000000000001779%	0.94	99.99999999999982%	13+
SLX-9640-24S-12C-AC-F	0.00000000000000000356%	0.0187	100.0%	0.0000000000001457%	0.77	99.99999999999985%	13+
SLX9740-40C-AC-F	0.00000000000000003159%	0.1661	100.0%	0.0000000000007494%	3.94	99.99999999999925%	15+
SLX9740-40C-AC-F	0.00000000000000032152%	1.6899	100.0%	0.0000000000042698%	22.44	99.99999999999573%	14+
SLX9740-80C-AC-F	0.00000000000000013556%	0.7125	100.0%	0.0000000000022340%	11.74	99.99999999999777%	14+
SLX9740-80C-AC-F	0.00000000000000239497%	12.5880	100.0%	0.0000000000192520%	101.19	99.99999999998075%	13+

Table 1 (cont.): Calculated Parallel Availability for 1,2 and 4 port Server port groups when accessing an IP Fabric system by Product.

In table 1, industry standard methods of device/network availability calculations have been utilized to provide a comparison of various elements. To deliver on the OMB tenet of high availability to the federal data center, planners can use this chart in concert with their requirements increase the level of availability of their data center network fabric in the PoD.

In the scenario where it is provided that >1 port channel connection from each server is connected to >1 Leaf switch, each with parallel connections to all available spine switches, a high availability network access is accomplished. With respect to availability, it is arguable that using the single-form factor vs. chassis choice is largely reduced to personal preference because the measurable difference is miniscule. Each choice provides greater availability when deployed in parallel configuration as an underlay. Either choice enables designers to achieve the OMB priority of High Availability for data centers. Costs should be evaluated and compared to provide a better picture between the solutions.

Summary

In this document, industry standard definitions for element availability were used for comparison purposes. A uniform standard means to define the network level availability by tiers in the data center was identified. The use of single form factor switching elements was compared with chassis-based systems. Various elements provided a means of comparison when deployed in series or parallel and the resulting industry standard availability measurement when deployed in a Point of Delivery (PoD), or data center fabric system.

Upon reviewing the total availability levels of the fabric, the underlay can now be programed onto the hardware. From this point, the element software maintains reachability for the rock-solid foundation built from these elements that are configured to deliver the any to any connectivity discussed in the previous document.

Continue to Volume 4: The Data Center IP Fabric Control Plane

This document discusses the protocols utilized to handle the IP control plane of the data center fabric. It also discusses the ability to provide deployment simplicity and uniformity with tools that create an automated underlay.