Lippis Consulting

Industry Paper

A Simpler Data Center Fabric Emerges

For The Age of Massively Scalable Data Centers

by

Nicholas John Lippis III
President, Lippis Consulting

June 2010

Table of Contents

Abstract

A number of independent trends are driving a new age of massively scalable data centers. One of these trends include a new IT delivery model based upon cloud computing, where large hosting facilities provide a range of IT services to corporations and governments. Further, high performance computing (HPC) facilities built via server clusters on the order of thousands to tens of thousands of servers and more has ushered in new favorable economics thanks to its use of x86 commodity hardware. The growth of public hosting and HPC facilities will only continue as efficient data center economics point to a fewer number of highly dense sites. It is this data center market segment, where the number of servers per facility is greater than 5,000, that we focus this white paper from a perspective of fabric, connecting servers and storage to internet/intranet via high performance Ethernet networking. For IT architects and designers of high-end data centers, this is the most important network design paper you will read this year.

The Problem of Data Center Scale

For IT business leaders that manage overall IT, it is no secret that approximately a third of all IT spending is consumed in the data center, according to Sanford Bernstein. With such a large share of IT Total Cost of Ownership (TCO) concentrated in the data center, changes in architecture can materially impact IT spend and corporate competitiveness. This is especially true in today's IT market, where virtualization and cloud computing is redefining the boundaries between compute, storage and networking to enable economic and energy efficiency, plus scale and elastic IT service delivery. This concentration trend of compute power in data centers is only beginning, as multi-core semiconductors are packaged into increasingly dense blade systems by computer manufactures with increasingly favorable economics. Add the trend and associated benefits of virtualization to pack more operating systems and applications into one physical server and the IT industry have served up a receipt for massively scalable data centers. Note that when we discuss data centers in this report, we refer to those facilities with 5,000 or more servers.

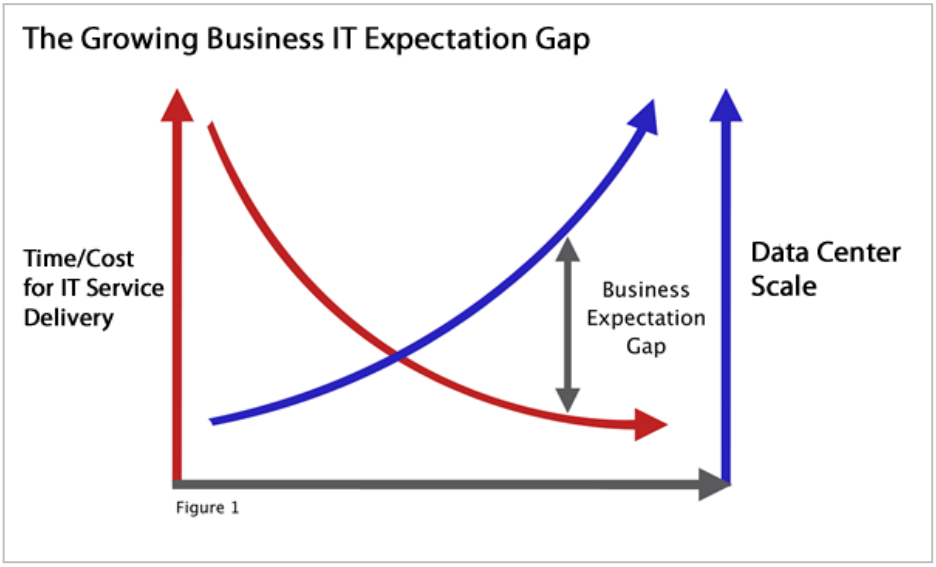


Figure 1

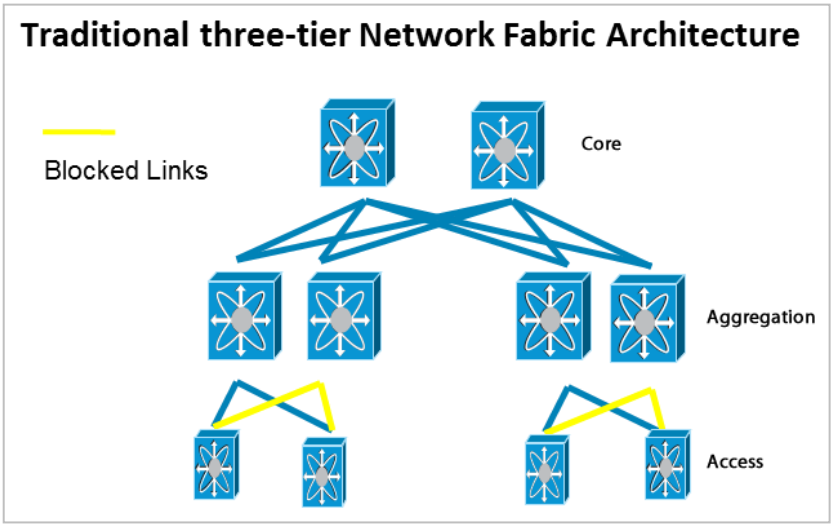
While the trends of virtualization and cloud computing offer data center architecture opportunities, there are also challenges. High-end data center design is challenged with increasing complexity, the need for greater workload mobility and reduced energy consumption. Traffic patterns have also shifted significantly, from primarily client-server or as commonly referred to as north-to-south flows, to a combination of client-server and server-server or east-to-west plus north-to-south streams. These shifts have wreaked havoc on application response time and end user experience, since the network is not designed for these Brownian motion type flows.

In short, most IT business leaders are exploring ways to make their data centers more energy efficient and dynamic. To address this challenge, these leaders are searching for data center architectures that accommodate increased application level flexibility, which translates into a more scalable, flatter, dynamic, highly available network infrastructure at the hardware layer, so that resources can be pooled and virtual machines (VM) freely moved. Key attributes of this new architecture are simplicity of design and operations, scale, high performance, resiliency, and flexibility. For high scale and VM mobility, a larger and common Ethernet or layer two domain that contains aspects of layer three or IP routing is being proposed within the industry.

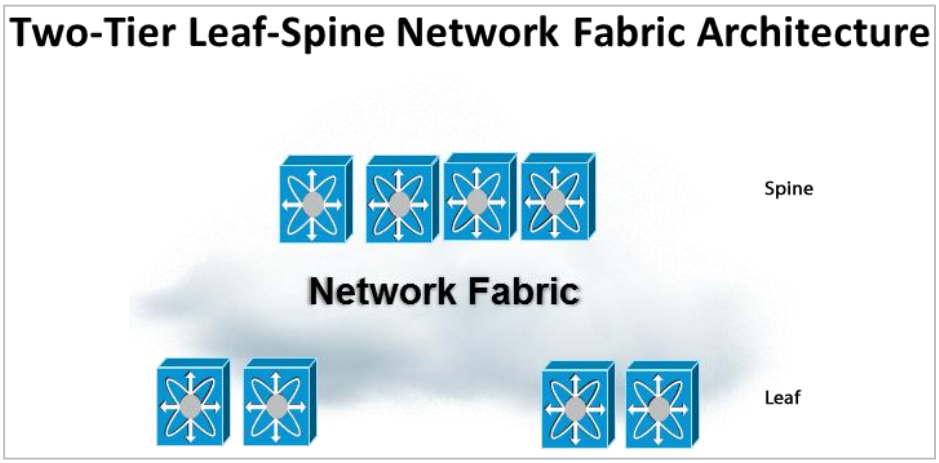
Traditionally, network architects in search of high scalability and reliability have utilized layer three or IP addressing to segment and scale networks, and this will continue to be the design of choice in most data centers. However, expanding use cases such as VM mobility or distributed applications, will require large data center scale without IP segmentation. To meet these needs, new innovations are being introduced that can dramatically scale, layer two or Ethernet networks. This is the cornerstone of a new data center network architecture discussed below.

Scalable Data Center Fabrics

In the section below, we discuss a new architecture for the data center fabric. But before we do, a word about the real world. A three-tier network architecture is the dominant structure in data centers today, and will likely continue as the standard design for many networks. By three tiers, we mean access switches/Top-of-Rack (ToR) switches, or modular/End-of-Row (EoR) switches that connect to servers and IP based storage. These access switches are connected via Ethernet to aggregation switches. The aggregation switches are connected into a set of core switches or routers that forward traffic flows from servers to an intranet and internet, and between the aggregation switches. It is common in this structure to over-subscribe bandwidth in the access tier, and to a lesser degree, in the aggregation tier, which can increase latency and reduce performance, but more on this later. Inherent in this structure is the placement of layer 2 versus layer 3 forwarding that is Virtual Local Area Networking or VLANs and IP routing. It is common that VLANs are constructed within access and aggregation switches, while layer 3 capabilities in the aggregation or core switches, route between them. Within the high-end data center market, where the number of servers is in the thousands to tens of thousands and east-west bandwidth is significant, and also where applications need a single layer 2 domain, the existing Ethernet or Layer 2 capabilities within this tiered architecture do not meet emerging demands.



One way to design a scalable data center fabric is often called a “fat-tree” and has two kinds of switches; one that connects servers and the second that connect switches creating a non-blocking, low latency fabric. We use the terms ‘leaf’ switch to denote server connecting switches and ‘spine’ to denote switches that connect leaf switches. Together, a leaf and spin architecture create a scalable data center fabric. Another design is to connect every switch together in a full mesh, with every server being one hop away from each other.



Many IT leaders in Global 2000 firms will have a need for both a typical tiered network structure supported by layer 3 routing and a scalable layer 2 fabric, as different deployment models are used for different applications. These leaders need a network equipment supplier that possess product architecture flexibility and can support both approaches in a single platform. A robust network Operating System (OS) that can be configured to support multiple use cases is also important as IT operations gain efficiency to manage the fabrics as configuration and management are consistent. In addition, a common network OS offers rapid absorption of innovation to IT operations as new OS features are available at the same time to all fabrics. Also, the benefit of using a common product set to build different designs offers value around operational efficiency, training, sparing and ease of evolution between different deployments. In short, the network switch needs to be flexible and general purpose versus narrowly defined for a special purpose use. It is this type of flexibility that will enable IT leaders to address the challenges of scale and efficiency outlined above. The alternative is introducing different suppliers for different use cases, which could provide incremental gains in certain use cases but often drives greater complexity, cost, and risk into the network overall.

In addition to product flexibility, some networking suppliers take a systems approach to their fabric design meaning that a solution is built and pre-tested before it arrives on site. This ensures that IT does not have to perform system integration. With the increased concentration of computing and IT dollars into data centers, it is only obvious that data centers are long-term corporate commitments. Therefore, it is only appropriate that the networking supplier of choice also has a proven long-term commitment to their product architecture. Perhaps the best example of this is Cisco's Catalyst 6000 switching architecture and its two-year-old Nexus product line. The Catalyst investment protection is well documented as it has been in operation for over a decade, which Cisco customers enjoy continued innovation and value added to this platform. The Nexus product line has a similar investment protection philosophy with a fifteen-year plus lifespan expectation. Common to both Catalyst and Nexus is the fact that these products are built on silicon and developed at Cisco, affording investment protection from one generation of the hardware to the next.

A Unified Fabric

The concept of a unified fabric is to virtualize data center resources and connect them through a high bandwidth network that is very scalable, high performance and enables the convergence of multiple protocols

onto a single physical network. These resources are compute, storage and applications, which are connected via a network fabric. In short, the network is the unified fabric and the network is Ethernet.

The industry tends to focus on storage transport over Ethernet as the main concept behind a unified fabric with technologies such as Fiber Channel over Ethernet or FCoE, iSCSI over Ethernet, iWARP over Ethernet and even Infiniband over Ethernet. However, this is a narrow view of a unified fabric, which is being expanded thanks to continual innovation of Ethernet by the vendor community and standards organizations such as the IEEE and IETF. Ethernet innovations such as FCoE, Data Center Bridging or DCB, Cisco's VN-Link, FEX-Link and virtual PortChannel or vPC have enhanced Ethernet networking to support a wide range of new data center fabric design options. In addition to these protocol enhancements, the IEEE is scheduled to complete its work on defining 40Gb and 100Gb Ethernet in June 2010, significantly increasing Ethernet's ability to scale bandwidth. To demonstrate how Ethernet is evolving to be the unified fabric for high-end data centers, we explore Cisco's new FabricPath innovation.

FabricPath: Multipath Ethernet Scaling Inter-Switch Bandwidth

FabricPath provides a new level of bandwidth scale to connect Nexus switches and delivers a new fabric design option with unique attributes for IT architects and designers. FabricPath is a NX-OS innovation, meaning that its' capabilities are embedded within the NX-OS network OS for the data center. FabricPath essentially is multipath Ethernet; a scheme that provides high-throughput, reduced and more deterministic latency, and greater resiliency compared to traditional Ethernet.

FabricPath combines today's layer 2 or Ethernet networking attributes and enhances it with layer 3 capabilities. In short, FabricPath brings some of the capabilities available in routing into a traditional switching context. For example, FabricPath offers the benefits of layer 2 switching such as low cost, easy configuration and workload flexibility. What this means is that when IT needs to move VMs and/or applications around the data center to different physical locations, it can do so in a simple and straightforward manner without requiring VLAN, IP address and other network reconfiguration. In essence, FabricPath delivers plug and play capability, which has been an early design attribute of Ethernet. Further, large broadcast domains and storms inherent in layer 2 networks that occurred during the mid 1990s have been mitigated with technologies such as VLAN pruning, Reverse Path Forwarding, Time-to-Live, etc.

The layer 3 capabilities added to FabricPath deliver scalable bandwidth allowing IT architects to build much larger layer 2 networks with very high cross-sectional bandwidth eliminating the need for oversubscription. In addition, FabricPath affords high availability as it eliminates the Spanning Tree Protocol (STP), which only allows one path and blocks all others, and replaces it with multiple paths between endpoints within the data center. This offers increased redundancy as traffic has multiple paths in which to reach its final destination.

FabricPath employs routing techniques such as building a route table of different nodes in a network. It possesses a routing protocol, which calculates paths that packets can traverse through the network. What is being added to FabricPath is the ability for the control plane or the routing protocols to know the topology of the network and choose different routes for traffic to flow. Not only can FabricPath choose different routes, it can use multiple routes simultaneously so traffic can span across multiple routes at once. These layer 3 features enable FabricPath to use all links between switches to pass traffic as STP is no longer used and would shut down redundant links to eliminate loops. Therefore, this would yield incremental levels of resiliency and bandwidth capacity, which is paramount as compute and virtualization density continue to raise driving scale requirements up.

New Network Design Enabled by FabricPath

In essence, FabricPath is link aggregation on steroids. As mentioned before, when using STP between two upstream switch chassis, one path is active and one is standby. To get around this limitation, companies such as Cisco offered vPC, which allows link aggregation between two chassis with both links active. What FabricPath's multipathing allows is to scale link aggregations up to 16 chassis. This is significant as network design completely changes when link aggregation scales from 1-to-2 then to 16 links.

When network design was limited to link aggregate to one or two, chassis planners were forced to build a multi-tiered hierarchical architecture; that is the three-tier structure of access switch, aggregation and core. In all of these different tiers, planners had to build in a certain level of oversubscription based on north-to-south and/or east-to-west traffic patterns.

With multipathing to 16 ways, IT architects now have an opportunity to build very large, broad, scalable topologies, without having to build multiple tiers. What this enables in HPC, cloud computing or hosting environments is that the network is transformed from a hierarchical architecture to a flat topology where only one hop separates any node in the data center! This completely changes the economics and dynamics with which IT designers build data center fabrics by affording a simpler and very scalable network. In addition, there is no single point of failure as in a three-tiered architecture since losing a tier or switch in such a design would effectively reduce bandwidth in half. With multipathing there are 16 ways between chassis and even if one fails the system loses only a 16th of its fabric interconnect bandwidth; even less depending upon design.

FabricPath Switching System or FSS

The value of FabricPath can be significantly extended when combined with Nexus hardware to build a very large data center switch fabric by interlinking a number of Nexus switches via FabricPath. This structure is called the FabricPath Switching System or FSS. The scale of this fabric is an unprecedented first for the IT industry. With Nexus switches and FabricPath, an IT architect can build a 160 Terabit/sec of switching capacity

fabric in a single layer 2 domain. That is nearly two orders of magnitude greater than other industry design. This 160 Tb fabric can connect more than 8,000 10 GigE ports of server facing, non-blocking switching capacity supporting full mesh traffic flows. The fabric itself supports over 24,000 10GbE ports for both server and fabric facing connections. This level of switching scalability is beyond previous technology. IT architects will view FSS as a system since it is multiple chassis interlinked via FabricPath and managed as a single system with Cisco's Data Center Network Manager (DCNM) management software. It is also pre-tested by Cisco. To obtain such a high level of 10GbE port density, Cisco had to develop a new high performance 32-port 10GbE module supporting Unified Fabrics called the F-Series module. The new module offers 32 ports of auto-sensing 1/10GbE and is essentially for server access and aggregation. Some of the module specs are: 320 Gbps switching capacity, 230 Gbps slot bandwidth, 512 ports per Nexus 7018 system, 5 microseconds of port-to-port latency and approximately 10W per 10GbE port consumption. With typical mesh traffic patterns, IT planners can expect 32 Gbs of line rate forwarding on a single card. It supports both the emerging TRILL standards and Data Center Bridging (DCB) with software upgrade capability to support FCoE implementations. Not only is this hardware industry leadership in terms of energy efficiency and latency, but also when combined with FabricPath, applications will have much fewer hops to traverse, increasing application response time. On a system level, FSS in combination with the new F-Series module will yield much lower power consumption and lower latency.

One Thousand Plus Servers Connected Into a Unified Fabric

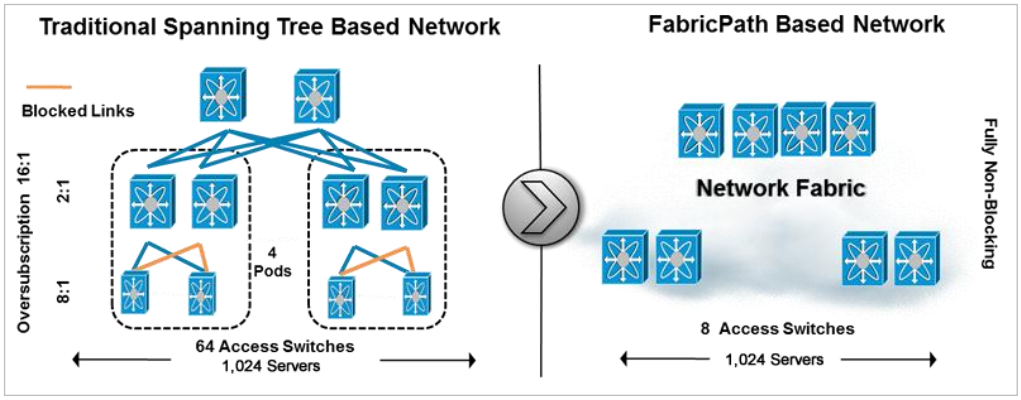
Consider a 1,024-server data center where all servers are dual home connected into the fabric via 10GbE. This example could be a Global 2000 company data center, but many Global 2000 companies and service provider hosting companies have larger scale requirements in the multiple tens of thousands of servers. In this example, approximately 2,048 10GbE connections are needed. Let us consider this requirement using traditional approaches versus FSS.

Traditional Spanning Tree Approach

If designing this data center fabric with traditional spanning tree-based networking, there would be blocked links between access and distribution. The IT architect would rely upon a three tier structure that forces an oversubscription of nearly 8:1 between access and aggregation and 2:1 between aggregation and core or a total of 16:1 oversubscription. There would be 64 access switches, 8 aggregation switches and two core switches required and four pods to house access and aggregation switches.

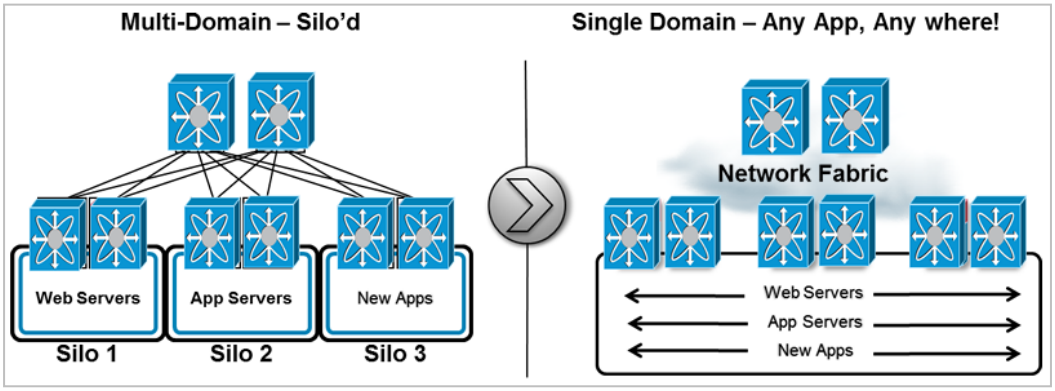
FSS with F-Series Module

With FabricPath, the IT architect would dual home all 1,024 servers into the fabric with a total of 12 Nexus switches reducing the number of devices managed from 74 to 12, increasing bandwidth performance by a factor of 16 and significantly reducing energy consumption. This fabric would be a non-blocking switching capacity too. Network availability would increase by more than a factor of two, thanks to greater redundancy in the FSS spine vs. the aggregation or core layer and every server would be one routing hop away from each other, lowering latency and increasing application performance. This fabric is flat too, meaning that servers and VMs can be moved freely without network configuration changes required. Also, there is only one network fabric tier in this design.



Workload Mobility

To highlight the ability to move data center resources freely, we focus on how a flatter network fabric changes application silos, which have traditionally been built through segmentation to isolate various applications. For scale considerations, many IT architects segment applications such as web servers, CRM, email, etc., by placing them in different subnets or different IP address domains; in effect siloing applications via subnet domains. This approach increases cross-sectional bandwidth thanks to layer 3 demarcations between silos but is very inflexible. For example, each subnet is usually a silo made up of a physical rack with multiple VLANs. Consider a common scenario when incremental server capacity is needed in subnet one, but there is no more physical space. There may be space within subnet two, but unfortunately the IT architect cannot pool physical space across subnets, thanks to segmentation.



Consider how this approach changes with FabricPath. First, a single domain layer 2 fabric connects all servers, offering a much more scalable and flexible network fabric. Since all servers reside within one common layer two network, IT operations can move servers around as needs require, meaning that there is no movement constraints. The level of flexibility and scale is just significantly enhanced. In addition, for VM mobility this is an essential requirement. FabricPath translates into simpler and lower operational costs and increased flexibility in terms of deployment. When FabricPath is introduced, IT architects scan scale more broadly with subnet barriers eliminated, allowing a VLAN to span across the entire physical infrastructure. Even though the physical infrastructure is still isolated via FabricPath, IT architects no longer have to separate application groups in to logical silos.

Designing A 160 Tbps Data Center Fabric

The following details the design of a 160 Tbps switching fabric with FabricPath and the F-Series module for high performance data centers using Cisco's Nexus 7000 switches. This architecture can support over 8,000 severs connected at 10GbE or 4,000 servers dual homed at 10GbE with attributes of being non-blocking, low latency (5 microseconds), high bandwidth, reliability, plus simplicity of workload movement.

To build a 160 Tbps single tier fabric, thirty-two Nexus 7018 switches populated with F-Series 10GbE modules would connect servers. These thirty switches are leaf switches. Each leaf chassis provides 256 10GbE ports to connect servers and another 256 10GbE ports to connect into spine switches. Therefore, each leaf is directly connected to each spine with sixteen FabricPath ports at 10GbE equaling a total of 256 10GbE ports for each leaf switch. There are sixteen spine switches each accepting 512 10GbE FabricPath ports. A single leaf chassis connects 256 10GbE ports into a spine equaling approximately 2.5Tbs. Multiplying each thirty-two leaf's contribution into the fabric yields 80Tbs. As Ethernet is full-duplex, the total fabric switching capacity is 160 Tbps. Therefore, 160Tbps of switching fabric is available across all thirty-two leaf chassis. As 256 10GbE equals 2.5 Tbs, which also equals 16 FabricPath links to each one of sixteen spine switches, yields 2.5 Tbs, the fabric is non-blocking.

As for layer 2 and layer 3 forwarding, the job of the spine is to forward packets from leaf switches at layer 2, creating a single tier fabric. A key attribute of this architecture is that each 16-way FabricPath links are Equal Cost Multipathing or ECMP. What 16-way FabricPath ECMP provides are two benefits: 1) It delivers more paths for traffic to flow, which increases available bandwidth in the fabric and 2) as they're distributed across all switches, diversity of routes is enabled to distribute packet forwarding. In essence what 16-way FabricPath ECMP provides is a very low latency, high bandwidth approach to supporting both north-to-south and east-to-west traffic flows simultaneously.

Getting Started

The above discussion is useful to demonstrate scale of a layer 2 based single tier fabric. To put FSS and the F-Series to work, we recommend that IT architects start by considering scope and size of the data center. How many servers need to be connected? Are servers single or dual homed? Will they connect via 1 and/or 10GbE? What are the mobility requirements etc? Once the scope and scale is defined, we recommend starting with a pilot to test FSS and gain insight, understanding and quantification of the attributes above. These are new product features and need to thoroughly be tested before placed into operations. The pilot phase is a wonderful time for IT to develop the skills needed to scale up the implementation assuming all goes well. Once IT is comfortable with the pilot, then a production test of a FSS implementation with F-Series modules is suggested to gain experience with a flat fabric and the attributes detailed above. The pre-validation process conducted by Cisco prior to release of FSS should also build confidence that implementations will go smoothly.

Industry Recommendations

The following recommendations are focused upon IT architects and planners who are responsible for building massively scalable data centers.

1. Start pilots now of FSS to gain IT skills, understand limitations and knowledge to plan deployment at scale.

2. Consider eliminating Spanning Tree Protocol in high-end data centers with multi-path technologies that scale up to the maximum number of spine switches in your implementation.

3. Consider a scalable layer 2 10Gb Ethernet based fabric to increased cross-sectional bandwidth, reliability, performance, flexibility and mobility of high-end data centers.

4. Consider a scalable layer 2 10Gb Ethernet fabric as a general-purpose platform to deliver storage integration options such as FCoE, iSCSI over Ethernet, etc.

5. Look for suppliers that support both rich Layer 3 routing services and scalable Layer 2 Ethernet capabilities to ensure choice and flexibility of three tier and scalable fabric implementations. Such suppliers thus offer products that can be configured in multiple uses, cases and topologies where modules are inter-changeable, skills transferable and operations common between both fabric approaches.

About Nick Lippis

Nicholas J. Lippis III is a world-renowned authority on advanced IP networks, communications and their benefits to business objectives. He is the publisher of the Lippis Report, a resource for network and IT business decision leaders to which over 35,000 business and IT executive leaders subscribe. Its Lippis Report podcasts have been downloaded over 80,000 times; i-Tunes reports that listeners also download the Wall Street Journal's Money Matters, Business Week's Climbing the Ladder, The Economist and The Harvard Business Review's IdeaCast. Mr. Lippis is currently working with clients to transform their converged networks into a business platform.

He has advised numerous Global 2000 firms on network architecture, design, implementation, vendor selection and budgeting, with clients including Barclays Bank, Microsoft, Kaiser Permanente, Sprint, Worldcom, Cigitel, Cisco Systems, Nortel Networks, Lucent Technologies, 3Com, Avaya, Eastman Kodak Company, Federal Deposit Insurance Corporation (FDIC), Hughes Aerospace, Liberty Mutual, Schering-Plough, Camp Dresser McKee and many others. He works exclusively with CIOs and their direct reports. Mr. Lippis possesses a unique perspective of market forces and trends occurring within the computer networking industry derived from his experience with both supply and demand side clients.