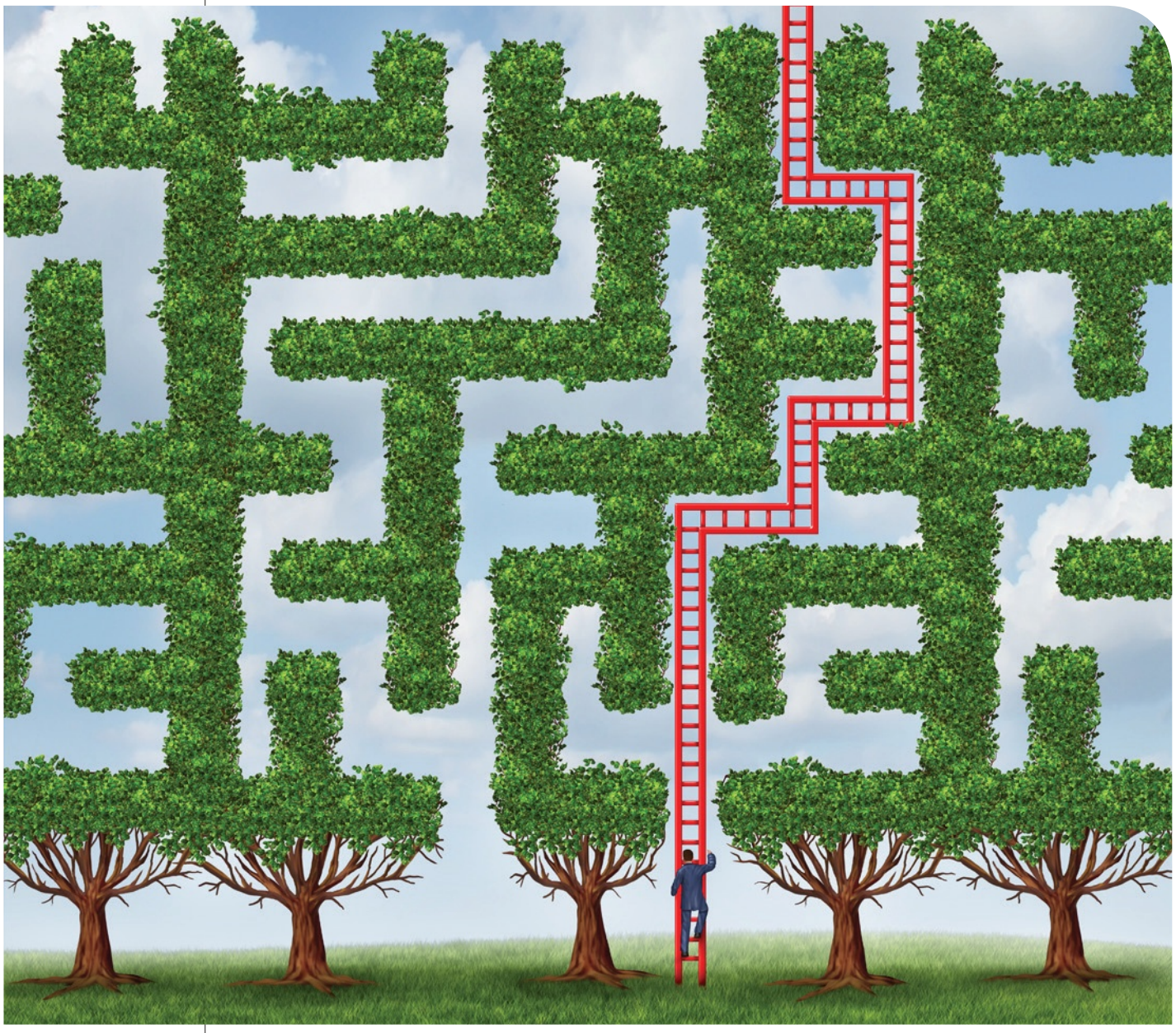
By Carrie Higbie, RCDD, NTS, CDCP, CDCS



NAVIGATING

the Pros and Cons of Fat-Tree Switch Fabric Architectures in the Data Center

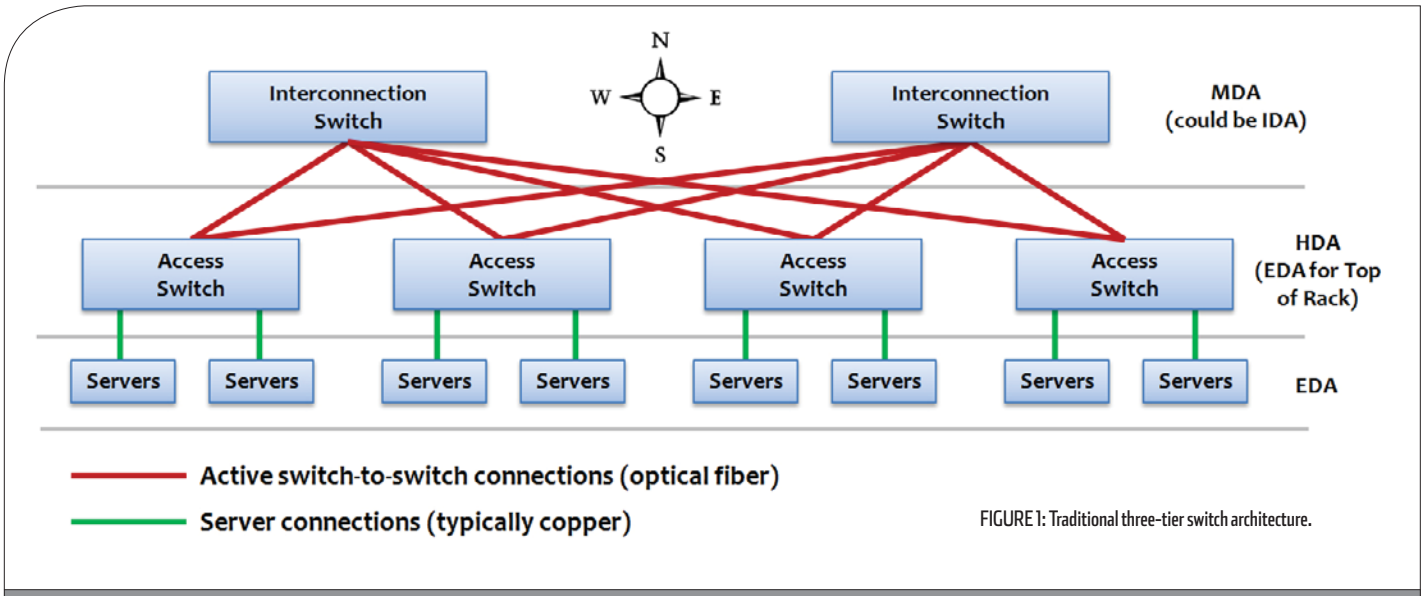


FIGURE 1: Traditional three-tier switch architecture.

Structured cabling allows for an "any-to-all" approach within the zone where any switch port can be connected to any server port within the zone.

Traditional three-tier Layer 3 switch architectures have been common practice in the data center environment for several years, but they are no longer ideal for large virtualized data centers. This has many data centers moving to fat-tree, sometimes Layer 2, switch fabric architectures that typically use only two tiers of switches to provide lower latency for improved traffic flow. With several configurations available for fat-tree switch fabrics, data center professionals and IT managers need to examine the pros and cons of each based on their specific needs and total cost of ownership (TCO).

Why Go Fat and Flat?

Traditional three-tier architectures consist of core switches located in the main distribution area (MDA); aggregation switches located in the MDA, intermediate distribution area (IDA) or horizontal distribution area (HDA); and access switches located in the HDA. Sometimes access switches are placed in the equipment distribution area (EDA) as seen with top of rack (ToR) architectures that place a smaller access switch in each server cabinet (see Figure 1).

While the traditional three-tier architecture has been well suited for data traffic between servers that reside on the same access switch, it does not adequately support the non-blocking, low-latency, high-bandwidth requirements of large virtualized data centers that divide single physical servers into multiple isolated virtual environments. Non-blocking refers to having sufficient bandwidth so that any port can communicate with any other port at the full bandwidth capacity of the port, while latency refers to the amount of time it takes for a data packet to travel from its source to its destination. With equipment now located anywhere in the data center, data traffic between two access switches in a three-tier architecture may have to traverse in a north-south traffic pattern through multiple aggregation and core switches, resulting in an increased number of switch hops and increased latency.

In a high-bandwidth, virtualized environment, the traditional north-south traffic pattern (i.e., switch to switch) causes the problem of links not having enough bandwidth to support the traffic. The need for inactive backup connections to additional aggregation switches for redundancy has also required enterprises to outfit their three-tier data centers with more switches than budgets often allow.

The limitations of the traditional three-tier architecture have many data centers moving to flattened switch fabric architectures with fewer switch-to-switch hops where devices connect to one another using network switches over multiple connection paths. Switch fabrics are typically limited to one or two tiers

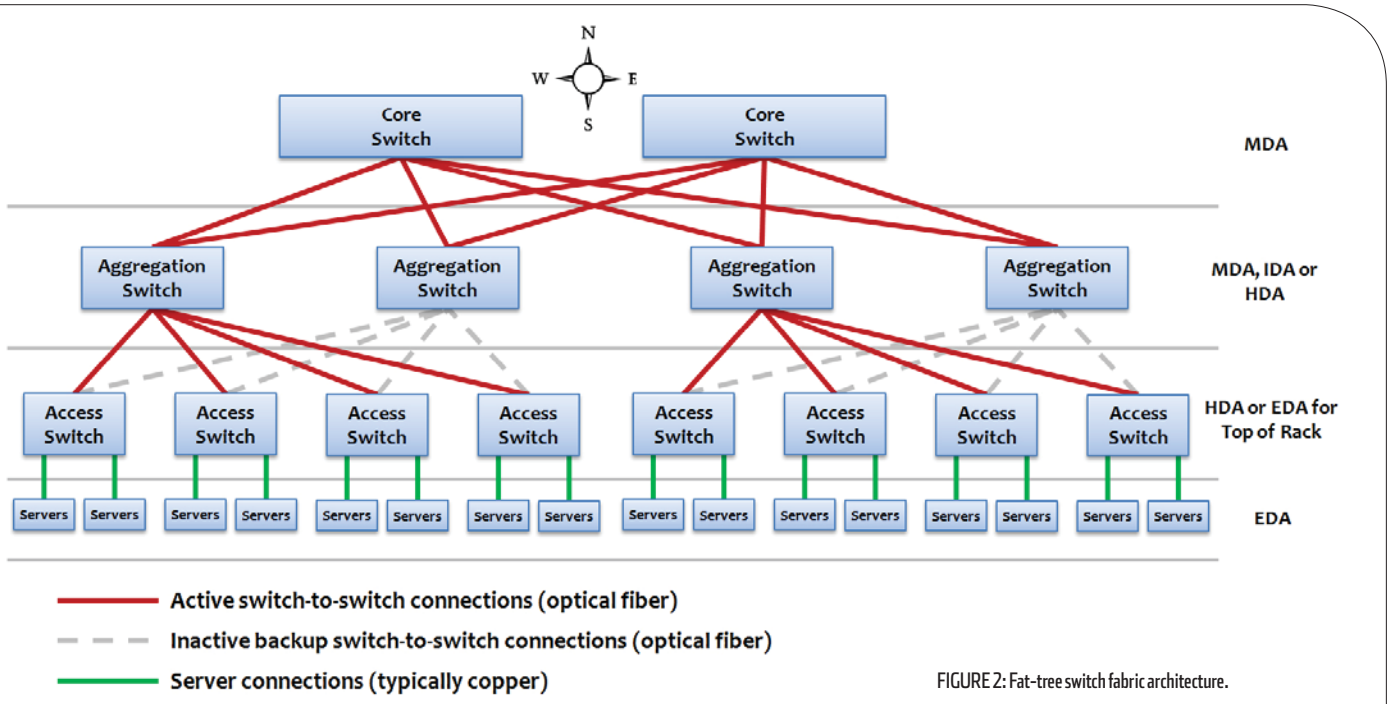


FIGURE 2: Fat-tree switch fabric architecture.

of switches to reduce the number of switch hops and provide higher bandwidth between any two points by taking advantage of wire-speed transmissions on the backplanes (i.e., port to port) of switches as opposed to uplinks (i.e., lower level switch to higher level switch). This enables dynamic east-west traffic (i.e., server to server) where it is needed, eliminating the need for communication between two servers to travel north-south through multiple switch layers.

Fat-tree switch fabrics, also referred to as a leaf and spine architectures, are one of the most common switch fabrics being deployed in today's virtualized data center. These flattened architectures leverage the reach of standards-based copper and optical fiber cabling to establish large numbers of active connections between fewer switches, providing multiple backplane routes to minimize latency. In addition to reducing the number of switches, fat-tree architectures lower TCO because networking and server resources can be deployed more rapidly. Lower latency also provides the performance needed to combine storage area networking (SAN) and network traffic onto the fabric, rather than having two distinct networks as is often required with traditional three-tier architectures.

The fat-tree architecture consists of interconnection (i.e., spine) switches placed in the MDA and access (i.e., leaf) switches placed in the HDA or EDA that each connect (i.e., uplink) to every interconnection switch in a mesh, typically via optical fiber (see Figure 2). With no more than one switch transmitting data between any two access switches that need to communicate with one another, the fat-tree architecture provides a non-blocking traffic environment with less hops that significantly reduces latency and enables more bandwidth from the access switches to the interconnection switches.

While fat-tree switch fabrics are rapidly becoming the norm, data center managers are faced with several configuration options and decisions regarding where to place access switches, whether to use structured cabling and how to handle longer cabling runs for active switch-to-switch connections. In April 2013, the Telecommunications Industry Association (TIA®) released ANSI/TIA-942-A-1, an addendum to the ANSI/TIA-942-A data center standard that provides cabling guidelines for switch fabrics. While the standard does an excellent job of explaining the fat-tree switch fabric architecture and its various configurations and functionality, real-world implementation warrants taking a closer look at the pros and cons of each option.

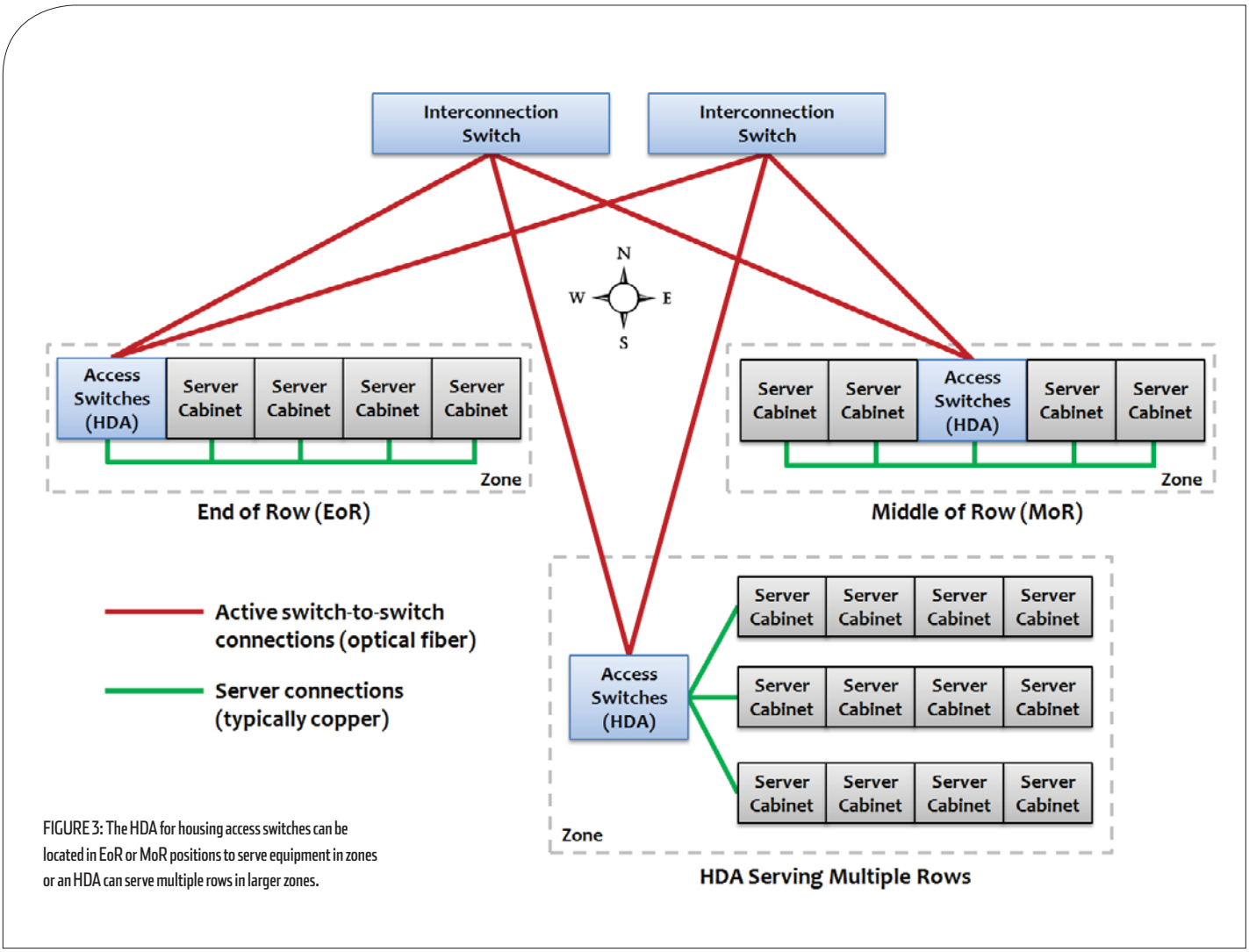


FIGURE 3: The HDA for housing access switches can be located in EoR or MoR positions to serve equipment in zones or an HDA can serve multiple rows in larger zones.

Getting into the Zone

Fat-tree switch fabrics support the use of HDAs for housing access switches that connect to servers and storage equipment placed in zones. In this scenario, the HDA can be located at the end of row (EoR) or middle of row (MoR) position to serve the equipment in that zone, or an HDA can serve multiple rows of cabinets in larger zones (see Figure 3).

By zoning switches in the aforementioned manner, unused ports are kept to a minimum because the switches are procured based on the number of servers that require a connection rather than on the number of cabinets as seen in ToR architectures. Another advantage of the EoR or MoR approach is that any two servers in a zone that need to "talk" to each other can be connected to the same switch for low-latency wire speed port-to-port communications rather than being connected to separate switches that result in higher latency switch-to-switch uplink communications.

A disadvantage with this zoned approach when using small form-factor pluggable (e.g., SFP+ and QSFP) twinaxial cable assemblies is the need to run point-to-point cabling in pathways from the access switch to each server in the zone, which can lead to "spaghetti" cabling over time. However, this can be easily alleviated by deploying 10GBASE-T over structured cabling via a cross-connect in the HDA that serves the entire zone. Structured cabling allows for an "any-to-all" approach within the zone where any switch port can be connected to any server port within the zone.

Another benefit to using structured cabling with a cross-connect is that the access switches

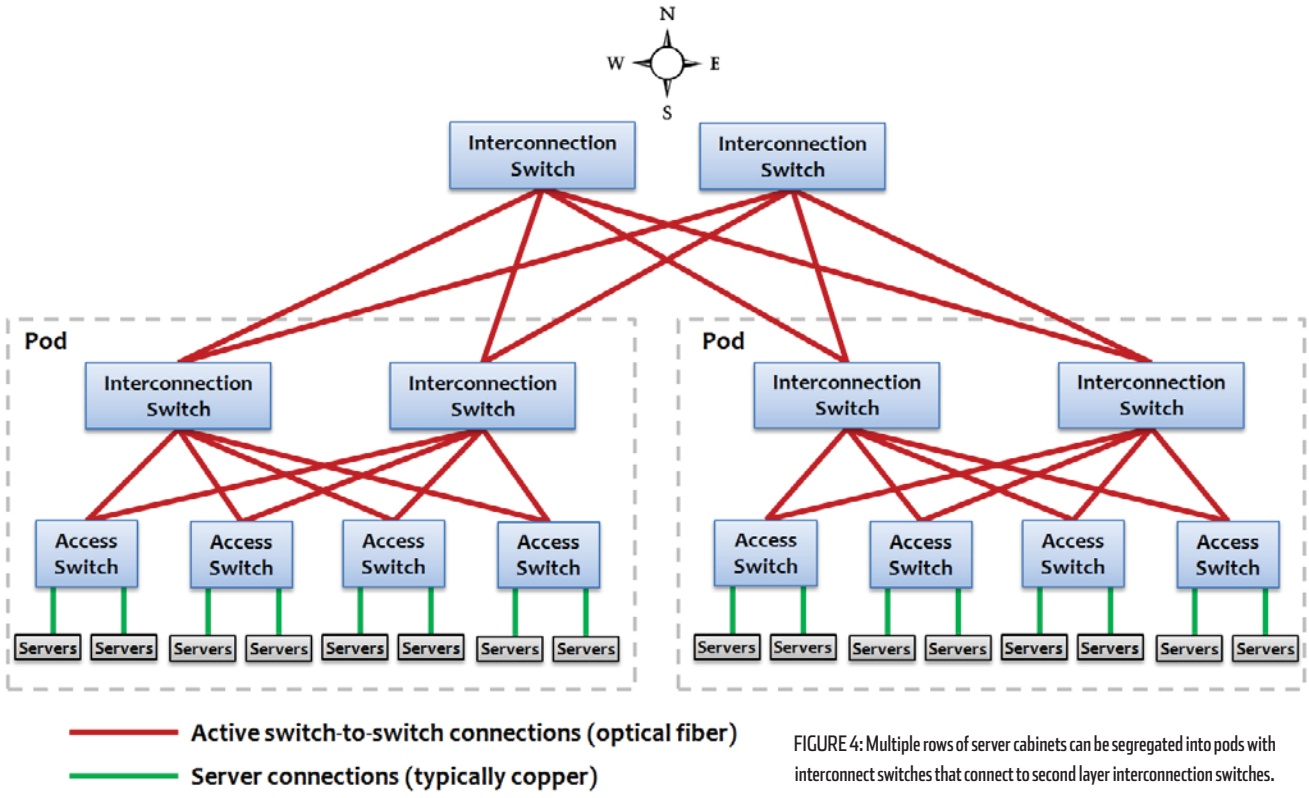


FIGURE 4: Multiple rows of server cabinets can be segregated into pods with interconnect switches that connect to second layer interconnection switches.

Maximum port utilization is easier to achieve when the access switches are placed in

in the HDA can remain separate and secure. Patch panels at the cross-connect can reside in a separate cabinet, so there is no need to access the switch cabinet. Structured cabling also lends itself well to software defined networking (SDN) where data traffic flow is controlled from a centralized control area—a practice that is especially advantageous for distributing carrier circuits in multi-tenant colocation environments. With structured cabling, SDN can be easily implemented without the need to reconfigure the entire cabling infrastructure. Structured cabling also works with all 10GBASE-T switches, regardless of vendor. This is not the case with proprietary SFP+ and QSFP twinaxial cable assemblies—these will likely need to be swapped out if another vendor's switch is deployed.

The use of zones provides scalability and flexibility, and eases deployment. Each zone can be configured in the same way so that they essentially become modules that operate as their own entity and are easily deployed using a repeatable, predictable process. With fat-tree switch fabrics, the idea of "zones" can be taken one step further. The switch fabrics themselves can be interconnected via interconnection switches located in IDAs that connect to a second layer of interconnection switches. In this scenario, multiple rows of server cabinets can be segregated into separate functional fabric areas called "pods," or modular subsets of the data center (see Figure 4). For example, one pod may be dedicated to finance applications, while another is dedicated to back-office functions. Each pod, existing as its own fat-tree switch fabric, supports east-west traffic within that pod. North-south traffic is only required when servers in one pod need to communicate with servers in another pod.

However, data center design should limit north-south traffic by placing servers that need to

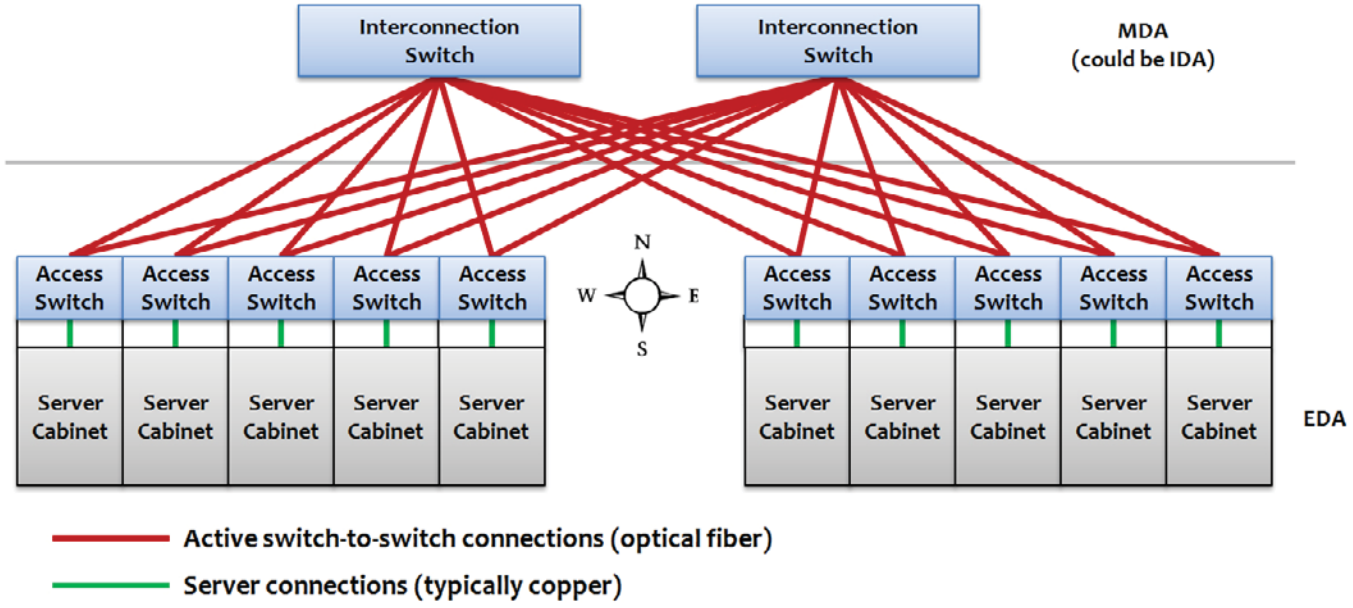


FIGURE 5: In a ToR configuration, small access switches placed in the top of each cabinet connect directly to the equipment in the cabinet via point-to-point cabling.

the HDA because the switch ports can be divided up to any equipment within a zone.

frequently communicate with one another within the same pod. This pod method is also ideal when switch fabrics need to coincide with traditional three-tier or other data center architectures—the fat-tree switch fabric remains as its own pod but connects to the other traditional architecture via core switches in the MDA.

Taking it to the Top

The alternative to placing access switches in the HDA at the EoR or MoR position is to place the access switches in the EDA in the ToR configuration. In this scenario, the HDA is eliminated and backbone cabling runs from each interconnection switch in the MDA to each cabinet. In the cabinet, the ToR access switch connects directly to the equipment in that cabinet via direct point-to-point cabling (see Figure 5). ANSI/TIA-942-A-1 specifies that this point-to-point cabling should be no greater than 10 m (33 ft).

The ToR configuration is geared toward dense 1 rack unit (1RU) server environments, enabling fast server-to-server connections within a rack versus with a zone. It is an ideal configuration for data centers that prefer a cabinet-at-a-time deployment versus a zone-at-a-time deployment. However, ToR eliminates the opportunity to use structured cabling with convenient cross-connects for making changes and connecting any switch port to any server port. All changes are made at the cabinet level and all communication from one cabinet to another requires uplink and downlink (i.e., switch-to-switch) transmission.

Another key consideration for the ToR configuration is that enough power and cooling must be supplied to the cabinet to support a full complement of servers and take advantage of all ports on the access switch. If a cabinet cannot support a full complement of servers, data center managers risk having unused switch ports in each cabinet. One

With no more than one switch transmitting data between any two access switches that need to communicate with one another, the fat-tree architecture provides a non-blocking traffic environment with less hops that significantly reduces latency and enables more bandwidth from the access switches to the interconnection switches.

solution is to provide one ToR switch for every two cabinets. However, point-to-point connections across cabinets can become difficult to manage over time. Maximum port utilization is easier to achieve when the access switches are placed in the HDA because the switch ports can be divided up to any equipment within a zone. With an access switch in each cabinet for ToR configurations (or two for primary and secondary networks), there is also a greater amount of active equipment and subsequent power supplies that need to be purchased, connected and maintained. That equipment also takes away from the number of servers that can be installed in each cabinet.

Instead of using ToR access switches, fat-tree architectures can also deploy port extenders in each cabinet that connect to the access switches in the HDA at the end or middle of a row. Also referred to as fabric extenders, port extenders are devices that provide additional ports to their controlling access switches (i.e., parent switch). They are essentially physical extensions of the parent access switches, and they can allow for connecting several lower-speed server ports to fewer high-speed ports on the access switch. Port extenders do not actually perform "switching" but are fully managed by their parent access switch. Port extenders cannot function without a parent switch and, in some cases, they are proprietary to their parent switch and cannot function with other vendors' switches. Port extenders rely on multiple uplink ports to access switches, which can also result in costly optical fiber cabling infrastructure that exceeds the costs of copper structured cabling.

The use of port extenders is a fairly new concept. Many see the port extender as combining the best of EoR and ToR configurations, but there are still some concerns surrounding this technology. While less expensive than ToR access switches, port extenders for each cabinet can ultimately cost more than a blade in a chassis access switch, if a blade slot is available. For the fat-tree switch fabric to remain non-blocking, the connection between the parent access switch and the port extender must also provide as much bandwidth as the total bandwidth of the server ports that connect to the port extender.

Cabling and Connectivity Considerations

To reap the benefits of the fat-tree switch fabric and ensure low latency, sufficient bandwidth must be provided between the access switches and the interconnection switch. In other words, the bandwidth of the server connections to each access switch must be less or equal to the sum of the bandwidth of all uplinks from the access switch to the interconnection switch. For example, a 32-port 10 gigabit access switch with a total bandwidth for server connections of 320 gigabit would need at least 32 10-gigabit uplinks or eight 40-gigabit uplinks to the interconnection switch. Consequently, the size of the switch fabric is dependent on the number of ports available on the interconnection switch. Where port extenders are used, non-blocking fabrics can be difficult or impossible to achieve for speeds of 10 gigabit or higher using current technology. These configurations typically rely on oversubscription of traffic, which is the assignment of more traffic to a link than the bandwidth capacity of the link. They are therefore often designed with lower uplink speeds based on the assumption that not all ports will communicate over the uplinks at the same time. However, that may not necessarily always be the case.

Another consideration in the fat-tree switch fabric is the length of the optical fiber backbone cabling from the interconnection switch in the MDA to the access switch in the HDA or to the ToR switches in the

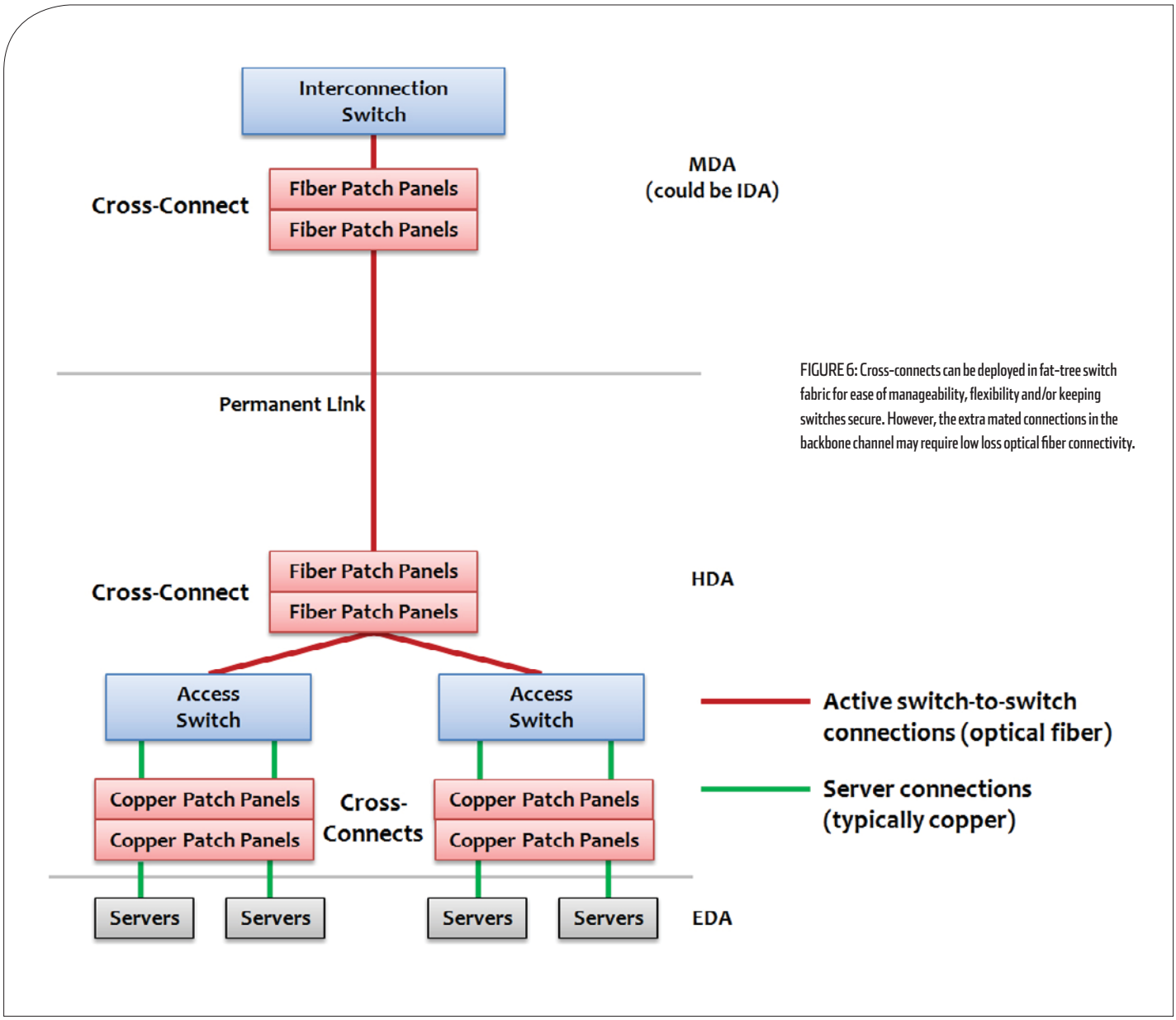


FIGURE 6: Cross-connects can be deployed in fat-tree switch fabric for ease of manageability, flexibility and/or keeping switches secure. However, the extra mated connections in the backbone channel may require low loss optical fiber connectivity.

EDA. As speeds increase, the distance limitations of multimode optical fiber systems may require a move to singlemode optical fiber and more expensive electronics.

Some data center managers also prefer to deploy traditional structured cabling for switch-to-switch connections with cross-connects at the interconnection switch and/or access switch for manageability, flexibility and/or keeping switches separate and secure (see Figure 6). Ideal for larger data centers or when optical fiber is distributed to multiple zones, the use of cross-connects can allow for one-time deployment of fixed high-fiber-count cabling from the cross-connect at the MDA to the cross-connect at the IDA and/or HDA. This allows optical fiber backbone cabling to be used for various purposes (e.g., SAN or uplink ports) without multiple moves, adds and changes.

A key consideration of deploying cross-connects between the interconnect switch and the access switch is optical insertion loss budget restraints. Each connection point in a channel adds additional loss, and data center managers need to carefully calculate the insertion loss of both cable and connectors to ensure they stay within the standards-based optical loss limitations that ensure maximum performance. Multi-

fiber push on (MPO) or mechanical transfer push on (MTP) connectors are recommended for switch-to-switch connections due to their preterminated plug-and-play benefits and ease of scalability from 10 gigabit to 40 and 100 gigabit speeds. However, typical MPO/MTP insertion loss may not allow for having more than two or three mated connections in the channel. If deploying cross connects in this scenario, low loss MPO/MTP connector solutions are required—their lower loss supports multiple mated connections for flexible patching options over a wide range of distances and configurations while remaining within the link loss budget.

Cabling for the switch-to-server connections is also a consideration. For EoR and MoR configurations, standards-based category 6A twisted-pair cabling will support 10GBASE-T up to 100 m (328 ft) distances and the use of a cross-connect. The 10GBASE-T standard includes a short reach (i.e., low power) mode that only works on category 6A and higher performing cabling up to 30 m (98 ft). Short reach mode can save up to 1.5W per port for improved energy efficiency.

While many data center managers choose SFP+ and QSFP twinaxial cable assemblies for the point-to-point connections in the rack when deploying a ToR configuration, standards-based, interoperable category 6A cabling assemblies offer significant advantages for this connection.

First of all, some ToR switch vendors require the use of proprietary SFP+ and QSFP twinaxial cable assemblies for server connections. Several ToR switches are even designed to check vendor security identification (ID) on the cables connected to each port and either display errors or prevent ports from functioning when connected to an unsupported vendor ID. While this helps ensure that vendor-approved cable assemblies are used with corresponding electronics, it can limit data center design options by locking data center managers into a proprietary solution. It also means that if the switch vendor changes, the cables will also need to change. Many of these proprietary SFP+ and QSFP twinaxial cable assemblies are only supported by a 90-day product warranty, while most category 6A vendors offer a 15- to 25-year system and application support warranty.

Another concern when using SFP+ and QSFP twinaxial cable assemblies is that these interfaces do not support autonegotiation—the ability for the switch to automatically and seamlessly switch between different speeds on individual ports depending on the connected equipment. Category 6A twisted-pair cabling supports autonegotiation, which enables partial switch or server upgrades on an as-needed basis. Without autonegotiation, a ToR switch upgrade requires all the servers connected to that switch to also be upgraded, incurring full upgrade costs all at once. SFP+ and QSFP cable assemblies are also typically more expensive than a structured twisted-pair cabling system, causing additional cost considerations.

Conclusion

While fat-tree fabric switch architectures provide the low latency and high bandwidth connections needed to support today's high-speed virtualized server environments, there is no single fat-tree configuration for every data center. Data center managers should carefully weigh the pros and cons of each option based on their specific needs. There are many factors to consider when choosing the configuration and supporting cabling infrastructure, including the benefits of using structured cabling and low loss optical fiber components with cross-connects in a fat-tree switch fabric when access switches are placed in EoR or MoR distribution area locations. ◄

AUTHOR BIOGRAPHY: Carrie Higbie, RCDD, NTS, CDCP, CDCS, is the global director of data center solutions and services for Siemon. She has been involved in the industry for more than 30 years as an executive running data centers and performing data center infrastructure design and audits in all tier levels for end user and hosted facilities. She can be reached at carrie_higbie@siemon.com.