Considerations for choosing top-of-rack in today's fat-tree switch fabric configurations

Pros and cons exist for each configuration, and depend largely on an individual data center's needs.

BY ROBERT CARLSON, COMMUNICATIONS CABLE AND CONNECTIVITY ASSOCIATION

Three-tier switch architectures have been common practice in the data center environment for several years. However, this architecture does not adequately support the low-latency, high-bandwidth requirements of large virtualized data centers. With equipment now located anywhere in the data center, data traffic between two servers in a three-tier architecture may have to traverse in a north-south traffic (i.e., switch-to-switch) pattern through multiple switch layers, resulting in increased latency and network complexity. This has many data centers moving to switch fabric architectures that are limited to just one or two tiers of switches. With fewer tiers of switches, server-to-server communication is improved by eliminating the need for communication to travel through multiple switch layers.

Fat-tree switch fabrics, also referred to as leaf and spine, are one of the most common switch fabrics being deployed in today's data center. In a fat-tree switch fabric, data center managers are faced with multiple configuration options that require decisions regarding application, cabling and where to place access switches that connect to servers. In a fat-tree switch fabric, access switches can reside in traditional centralized network distribution areas, middle of row (MoR) positions or end of row (EoR) positions—all of which use structured cabling to connect to the servers. Alternatively, they can be placed in a top of rack (ToR) position using point-to-point cabling within the cabinet for connecting to the servers.

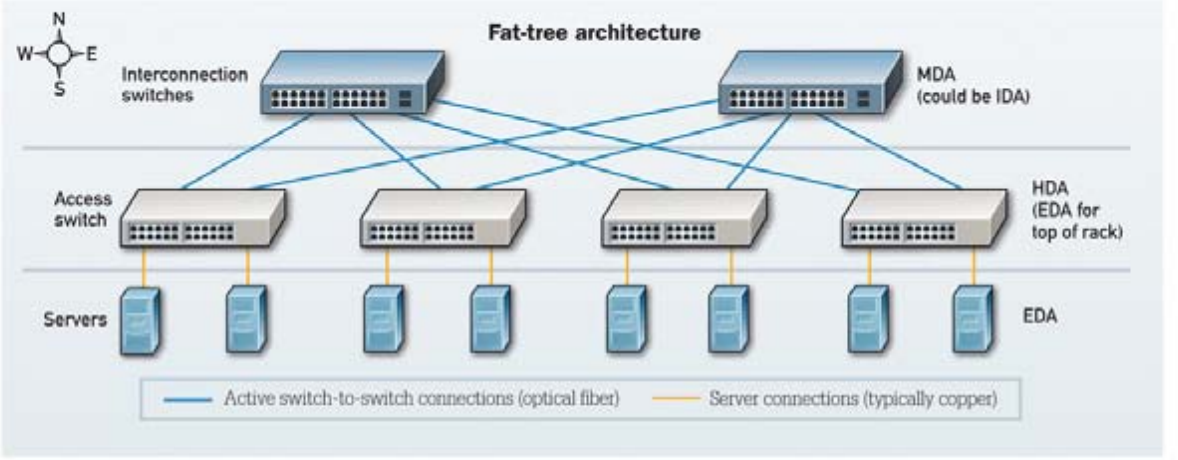
There is no single ideal configuration for every data center, and real-world implementation of newer fat-tree switch fabric architectures warrants CIOs, data center professionals and IT managers taking a closer look at the pros and cons of each option based on their specific needs.

A closer look at options

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; 942-A-1 provides cabling guidelines for switch fabrics. The fat-tree switch fabric outlined in the addendum consists of interconnection (spine) switches placed in the main distribution area (MDA), and access (leaf) switches placed in the horizontal distribution area (HDA) and/or the equipment distribution area (EDA). Each access switch connects to every interconnection switch in a mesh topology, typically via optical fiber.

In a fat-tree switch fabric, access switches that connect to servers and storage equipment in rows can be located at the MoR or EoR position to serve the equipment in that row, or they can be located in a separate dedicated area to serve multiple rows of cabinets. MoR and EoR configurations, which function in the same manner, are popular for data center environments where each row of cabinets is dedicated to a specific purpose, and growth is accomplished on a row-by-row basis. For the purposes of this article, we will concentrate on the more-popular EoR configuration.

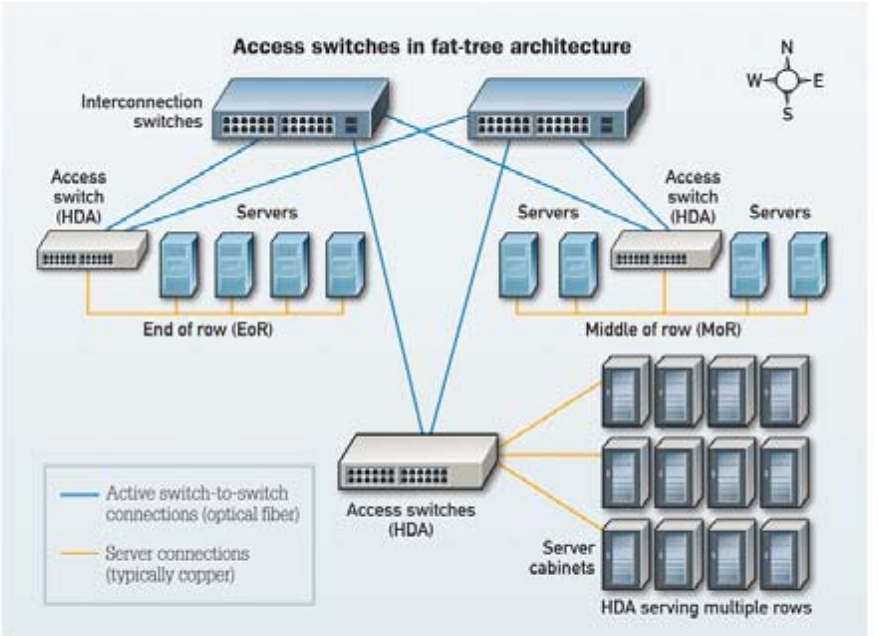
EoR configurations that place access switches at the end of each row use structured cabling with passive patch



Fat-tree architecture (Source: ANSI/TIA-942-A-1)

panels to serve as the connection point between the access switches and servers. Patch panels that mirror the switch and server ports (crossconnect) at the EoR location connect to corresponding patch panels at the access switch and in server cabinets using permanent links. The connection between switch and server ports are made at the cross-connect via patch cords.

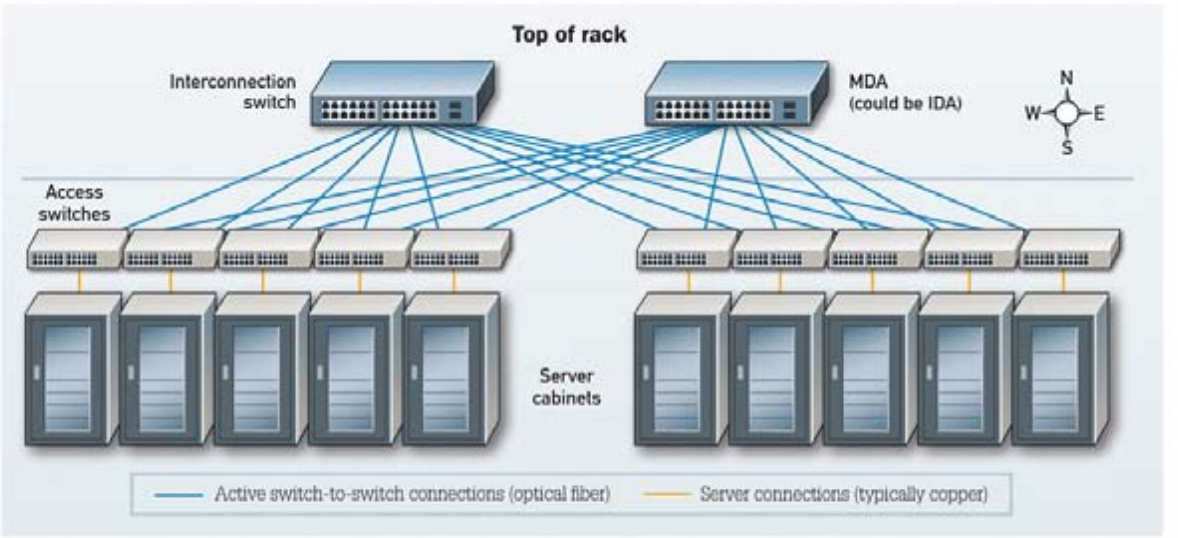
The alternative to placing access switches in the EoR position is to place them in the ToR position. In this scenario, fiber cabling runs from each interconnection switch in the MDA to smaller (1RU to 2RU) access switches placed in each cabinet. Instead of access switches, active port extenders can be deployed in each cabinet. Port extenders, sometimes referred to as fabric extenders, are essentially physical extensions of their parent access switches. For the purposes of this article, we will refer to ToR switches in general to represent both access switches and port extenders placed in the ToR position.



In a fat-tree architecture, the access switches (HDA) can be located in MoR or EoR positions to serve equipment in rows, or they can be located in separate dedicated areas to serve multiple rows.

Within each cabinet, the ToR switches connect directly to the servers in that cabinet using point-to-point copper cabling often via short preterminated small-form-factor pluggable (e.g. SFP+, QSFP) twinaxial cable assemblies, active optical cable assemblies or RJ45 modular patch cords.

ToR configurations are geared toward dense 1-rack-unit (1RU) server environments, enabling fast server-to-server



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. (Source: ANSI/TIA-942-A-1)

connections within a rack versus within a row. ToR is ideal for data centers that require a cabinet-at-a-time deployment and cabinet-level management.

The use of a ToR configuration places the access switch in the EDA, eliminating the HDA and patching area for making connections between switches and servers. In fact, ToR is often positioned as a replacement for and reduction of structured cabling. However, structured cabling offers several benefits, including improved manageability and scalability, and overall reduced TCO. These factors should be considered when evaluating ToR and structured cabling configurations in today's fat-tree switch fabric environments.

Manageability considerations

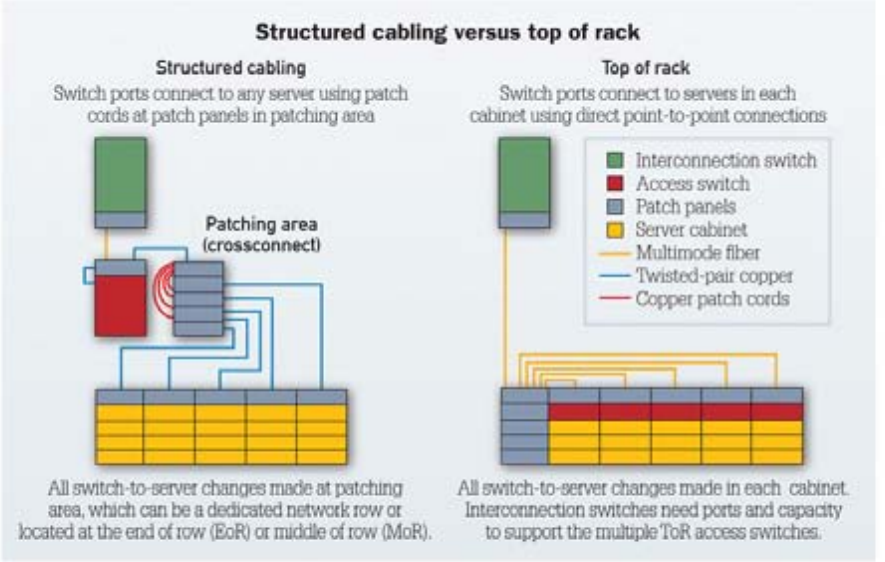
With structured cabling, where connections between active equipment are made at patch panels that mirror the equipment ports, all moves, adds and changes (MACs) are accomplished at the patching area. Any equipment port can be connected to any other equipment port by simply repositioning patch cord connections, creating an "any-to-all" configuration.

Because ToR switches connect directly to the servers in the same cabinet, all changes must be made within each individual cabinet rather than at a convenient patching area. Depending on the size of the data center, making changes in each cabinet can become complicated and time-consuming. Imagine having to make changes in hundreds of server cabinets versus being able to make all your changes at the patching area in each EoR location.

With structured cabling, the patch panels that mirror active equipment ports connect to the corresponding panels in patching areas using permanent, or fixed, links. With all MACs made at the patching area, the permanent portion of the channel remains unchanged, which allows the active equipment to be left untouched and secure. The patching area can reside in a completely separate cabinet so there is no need to access the switch cabinet. This scenario can be ideal for when switches and servers need to be managed by separate resources or departments.

ToR configurations do not allow for physically segregating switches and servers into separate cabinets, and MACs require touching critical switch ports. The ToR configuration can be ideal when there is a need to manage a group of servers and their corresponding switch by application.

Another manageability consideration is the ability for servers across multiple cabinets to "talk" to each other. While ToR enables fast server-to-server connections within a rack, communication from one cabinet to another requires switch-to-switch transmission. One advantage of the EoR approach is that any two servers in a row, rather than in a cabinet, can experience low-latency communication because they are connected to the same switch.



Structured cabling versus ToR topology. The ToR topology eliminates the convenient patching area for making changes.

Cabling distance limitations can also impact manageability. For ToR configuration, ANSI/TIA-942-A-1 specifies that the point-to-point cabling should be no greater than 10 meters (33 feet). Moreover, the SFP+ twinaxial cable assemblies often used with ToR switches limit the distance between the switches and the servers to a length of 7 meters in passive mode. The cabling channel lengths with structured cabling can be up to 100 meters, which allows for more-flexible equipment placement throughout the life of the data center.

Cooling considerations

ToR configurations can also land-lock equipment placement due to the short cabling lengths of SFP+ cable assemblies and data center policies that do not allow patching from cabinet to cabinet. This can prevent placing equipment where it makes the most sense for power and cooling within a row or set of rows.

For example, if the networking budget does not allow for outfitting another cabinet with a ToR switch to accommodate new servers, placement of the new servers may be limited to where network ports are available. This can lead to hot spots, which can adversely impact neighboring equipment within the same cooling zone and, in some cases, require supplemental cooling. Structured cabling configurations avoid these problems.

A ToR switch can be placed in the middle or bottom of a cabinet, but they are most often placed at the top for easier accessibility and manageability. According to the Uptime Institute, the failure rate for equipment placed in the top third of the cabinet is three times greater than that of equipment located in the lower two-thirds. In a structured cabling configuration, patch panels are generally placed in the upper position, leaving cooler space for the equipment.

Scalability considerations

ToR configurations allow for cabinet-at-a-time scalability, which can be a preference for some data centers, depending on the budget and business model in place. Once several cabinets are deployed, a widespread switch upgrade in a ToR configuration obviously will impact many more switches than with structured cabling. An upgrade to a single ToR switch also improves connection speed to only the servers in that cabinet. With an EoR structured cabling configuration, a single switch upgrade can increase connection speeds to multiple servers across several cabinets in a row.

Application and cabling for the switch-to-server connections is also a consideration when it comes to scalability. For EoR configurations with structured cabling, standards-based Category 6A twisted-pair cabling is typically the cable media of choice. Category 6A supports 10GBase-T up to 100-meter distances. The 10GBase-T standard includes a short-reach (i.e., low-power) mode that requires Category 6A and higher-performing cabling up to 30 meters (98 feet). Recent advancements in technology also have enabled 10GBase-T switches to rapidly drop in price and power consumption, putting them on par with ToR switches.

For direct switch-to-server connections in a ToR configuration, many data center managers choose SFP+ twinaxial cable assemblies rather than

Category 6A modular patch cords. While these assemblies support low power and low latency, which can be ideal for supercomputing environments with high port counts, there are some disadvantages to consider.

Standards-based Category 6A cabling supports autonegotiation, but SFP+ cable assemblies do not. Autonegotiation is the ability for a switch to automatically and seamlessly switch between different speeds on individual ports depending on the connected equipment, enabling partial switch or server upgrades on an as-needed basis. Without autonegotiation, a switch upgrade requires all the servers connected to that switch to be upgraded also, incurring full upgrade costs all at once.

For decades, data center managers have valued standards-based interoperability to leverage their existing cabling investment during upgrades regardless of which vendors' equipment is selected. Unlike Category 6A cabling that works with all Base-T switches, regardless of speed or vendor, higher-cost proprietary SFP+ cable assemblies may be required by some equipment vendors for use with their ToR switches. While these requirements help ensure that vendor-approved cable assemblies are used with corresponding electronics, proprietary cabling assemblies are not interoperable and can require cable upgrades to happen simultaneously with equipment upgrades. In other words, the SFP+ assemblies will likely need to be swapped out if another vendor's switch is deployed.

Some ToR switches are even designed to check vendor-security IDs on the cables connected to each port and either display errors or prevent ports from functioning when connected to an unsupported vendor ID. SFP+ cable assemblies also typically are more expensive than Category 6A patch cords, causing additional expense during upgrades. In addition, many of the proprietary cable assemblies required by switch vendors come with an average 90-day warranty. Depending on the cable vendor, Category 6A structured cabling typically carries a 15- to 25-year warranty.

Equipment, maintenance, energy costs

In a ToR configuration with one switch in each cabinet, the total number of switch ports depends on the total number of cabinets, rather than on the actual number of switch ports needed to support the servers. For example, if you have 144 server cabinets, you will need 144 ToR switches (or 288 if using dual primary and secondary networks for redundancy). ToR configurations therefore can significantly increase the amount of switches required, compared to the use of structured cabling configurations that use patch panels to connect access switches to servers in multiple cabinets.

Having more switches also equates to increased annual maintenance fees and energy costs, which impacts TCO. This is especially a concern as power consumption is one of the top concerns among today's data center managers. As data centers consume more energy and energy costs continue to rise, green initiatives are taking center stage. Reducing the number of switches helps reduce energy costs while contributing to green initiatives like LEED, BREEAM or STEP.

Based on a low-density 144-cabinet data center using a fat-tree architecture, Table 1 compares the cost for installation, maintenance and annual power consumption for a ToR configuration using SFP+ cable

Low-density, 144 server cabinets, 14 servers per cabinet

Material, Power, Maintenance	ToR (SPF+)	EoR (10GBase-T)
Material cost	$11,786,200	$8,638,300
Annual maintenance cost	$1,655,200	$1,283,100
Annual energy cost	$101,400	$44,400
Total cabling cost (included in material cost)	$1,222,300	$70,300
Total cost of ownership	**$13,542,800**	**$9,965,800**

High-density, 144 server cabinets, 40 servers per cabinet

Material, Power, Maintenance	ToR (SPF+)	EoR (10GBase-T)
Material cost	$26,394,000	$21,596,100
Annual maintenance cost	$3,371,900	$2,737,900
Annual energy cost	$177,600	$106,700
Total cabling cost (included in material cost)	$5,123,900	$2,078,200
Total cost of ownership	**$29,943,500**	**$24,440,700**

assemblies to an EoR configuration using Category 6A 10GBase-T structured cabling. The ToR configuration ultimately costs 30 percent more than using an EoR configuration.

The example assumes an average 5 to 6 kW of power per cabinet, which supports approximately 14 servers per cabinet. It also assumes primary and secondary switches for redundancy. Installation costs include all switches, uplinks, fiber line cards, fiber backbone cabling and copper switch-to-server cabling. Annual maintenance costs are based on an average 15 percent of active equipment costs. Annual power costs are based on the maximum power rating of each switch for 24x7 operation. The example excludes the cost of software, servers, cabinets, and pathways.

The next table compares a ToR configuration to an EoR configuration with the same assumptions but in a high-density environment that assumes an average of 15 to 20 kW of power per cabinet to support 40 servers per cabinet. In this scenario, the total cost of ownership for ToR is still 20 percent more than that for EoR.

Low-density, 144 server cabinets, 14 servers per cabinet

	ToR (SPF+)	EoR (10GBase-T)
Total unused ports	5,184	576

High-density, 144 server cabinets, 40 servers per cabinet

	ToR (SPF+)	EoR (10GBase-T)
Total unused ports	6,912	224

Switch port utilization

Low utilization of switch ports also equates to higher total cost of ownership. In a low-density environment of 5 to 6 kW that can accommodate just 14 servers in a cabinet, server switch port demand will be lower than the 32 switch ports available on a ToR switch. Table 3 shows the same 144-cabinet example used in Table 1 equates to 5,184 unused ports with ToR, versus just 576 unused ports with EoR. That equates to more than 162 unnecessary switch purchases and related maintenance and power. Using an EoR configuration with structured cabling allows virtually all active switch ports to be fully utilized because they are not confined to single cabinets. Via the patching area, the switch ports can be divided up, on demand, to any of the server ports across several cabinets in a row.

Even when enough power and cooling can be supplied to a cabinet to support a full complement of servers, the number of unused ports can remain significantly higher with ToR than with EoR and structured cabling. As shown in Table 4, the same high-density 144-cabinet example used in Table 2 with 40 servers per cabinet equates to 6,912 unused ports with ToR versus only 224 unused ports with EoR. The reason is that two 32-port ToR switches are required in each cabinet to support the 40 servers, or four for a dual network using primary and secondary switches. That equates to 24 unused ports per cabinet, or 48 in a dual network. In a 144-cabinet data center, unused ports quickly add up.

In reality, the only way to truly improve switch-port utilization with ToR is to limit the number of servers to no more than the number of switch ports per cabinet. However, limiting the number of servers to the number of ToR switch ports is not always the most efficient use of power and space. For example, in a high-density environment that supports 40 servers per cabinet, limiting the number of servers per cabinet to 32 (to match the number of switch ports) results in 8 unused rack units per cabinet, or 1,152 unused rack units across the 144-cabinet data center. Once the number of servers surpasses the number of available switch ports, the only option is to add another ToR switch (two for dual networks). This significantly increases the number of unused ports.

Regardless of the configuration being considered, it's always best to consider port utilization when designing the data center and ensure that either empty rack spaces or unused ports can be managed.

There is no single cabling configuration for every data center. However, many data center environments can benefit from the manageability, cooling, scalability, lower cost and better port utilization provided by Category 6A structured cabling and 10GBase-T used in end of row, middle of row or centralized configurations.

Robert Carlson, vice president of global marketing for Siemon, authored this article on behalf of the CCCA (www.cccassoc.org).