White Paper

Performance Needs and Solutions in the Borderless Enterprise

Gidi Navon, Principal Architect
Switching BU, Networking Group, Marvell

November 2020

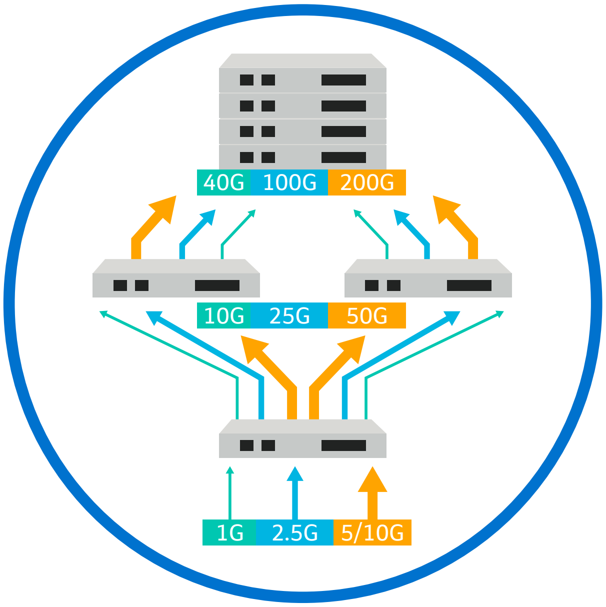
Abstract

Enterprise networks continue to evolve in many aspects, including the need for more performance. This need is caused by an increased number of wired and wireless network users, including IoT endpoints; increased demand for high-definition content; the adoption of new data-intensive applications and their transition to becoming borderless with workloads running in the cloud.

These demands are being catered by a new refresh cycle of the network infrastructure geared by new merchant and custom silicon devices. Devices built around new high-speed SerDes and new Ethernet port types, plus devices fabricated in new process nodes, are achieving higher performance and scale on a similar power envelope.

The solutions now available for the access, aggregation and core of the network include: new stackable switches with high-speed copper interfaces connected to Wi-Fi access points and other end points; new high-speed fiber transceivers connecting to the aggregation and core; and new chassis architectures based on high-performance switch silicon.

Marvell's newly announced Prestera® devices were developed to address these new performance needs and provide the building blocks for a new generation of enterprise switches with low TCO and investment protection for years to come.



Introduction

Network infrastructure generally evolves in two ways: Gradual change and growth aligned with minor organizational changes. During these gradual changes, IT managers will install and reconfigure systems from the same generation already installed. But after a few years, there comes a time for a new generation. A generation that can support a new level of performance based on newer technology.

In this white paper, we first look at the need for a change that starts in new clients such as new Wi-Fi access points, new bandwidth-hungry enterprise applications and new types of quickly emerging users, which are sometimes called IoT devices.

We next describe how different organizations choose to build their network topology in ways more suited to their needs, from access to the core.

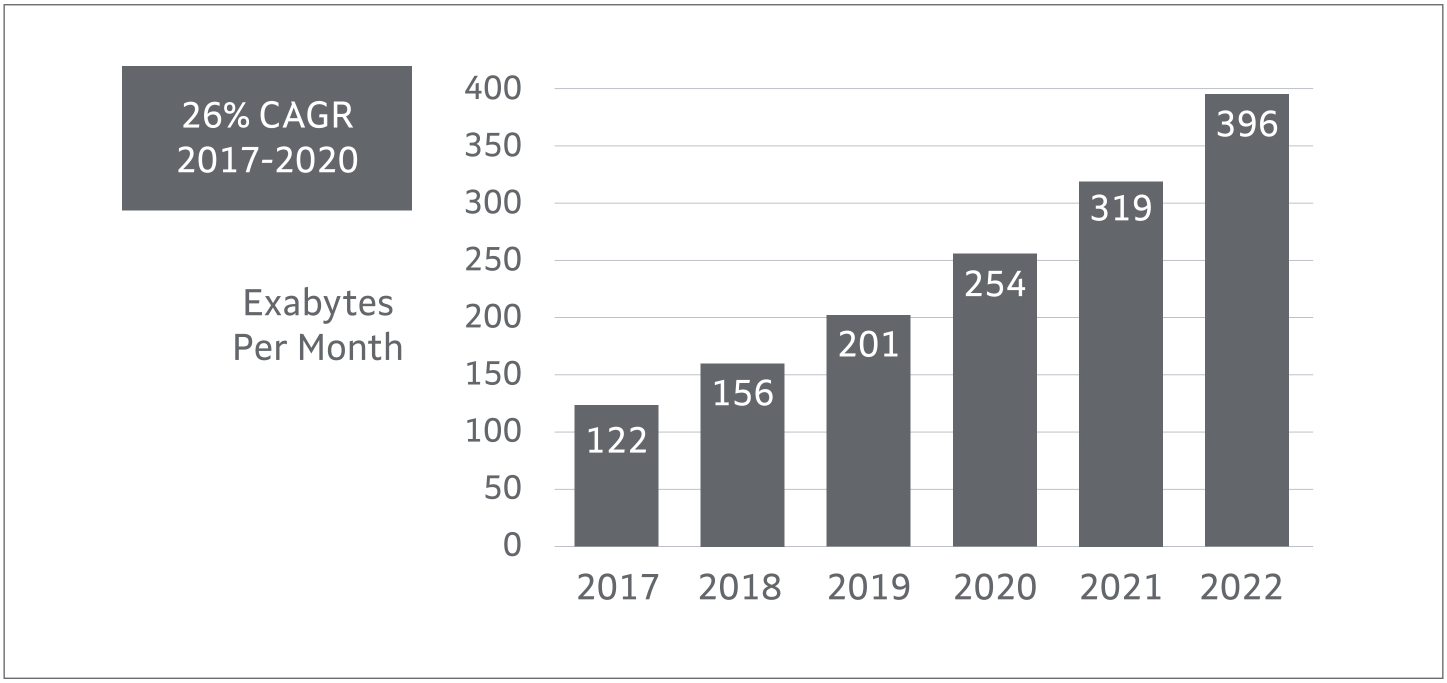
Lastly, we discuss a few migration paths that utilize higher speed interfaces, new stacking technology and paradigm changes in chassis design.

The Need for Change

Our digitized, always connected lives, are creating the need for change. Overall traffic growth continues to rise with the influx of new content, users and applications. In addition, the expectation for higher quality of experience continues to rise, pushing the envelope of current network infrastructure.

Traffic Growth

The general global growth in traffic usage continues to increase in Enterprise networks (as in other network segments) as outlined in Cisco's Visual Network Index (VNI) report. [1]The growth is mainly associated with the increase in video traffic including high-definition (HD) video streaming. Video content is consumed both in new conference rooms providing an immersive virtual reality (VR) type experience and also consumed by laptops and handheld devices across the borderless enterprise. New content includes the increased use of streaming video as training materials rich in content.

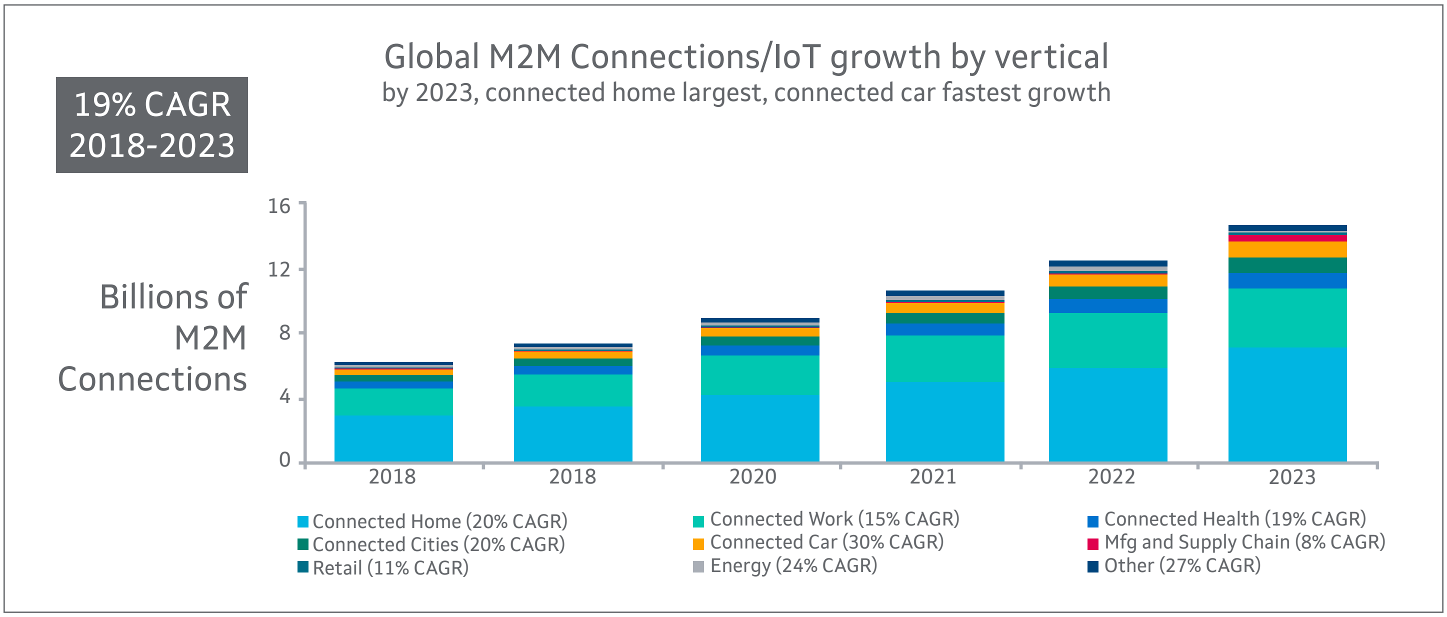


Source: Cisco VNI Global IP Traffic Forecast, 2017

Figure 1: Global Traffic Growth

A new user in town called IoT

The VNI report also shows the dramatic yearly growth of machine to machine (M2M) devices. While many of these devices are small sensors that affect scale much more than bandwidth, there are bandwidth-hungry devices such as HD IP video surveillance cameras that are now being installed in almost every corner of an organization. This of course also includes the entire new wave of automated retail stores that rely heavily on such devices for their business.



Source: Cisco Annual Internet Report, 2018–2023

Figure 2: Global IoT Growth

Quality of Experience

The need for more bandwidth is not only to address the consumption of the massive amount of information, but also to improve the quality of experience for users and applications. High-speed ports ensure latencies stay to a minimum at all times. This reduces serialization delay and avoids queuing delays, which are the largest contributors by far to latency.

Waiting for 1MB of congested packets stored in the packet buffer to be transmitted over a 1GbE link will take almost 10msec. This latency could be eliminated by replacing the 1GbE bottleneck link with a higher-speed port, avoiding any congestion.

Wi-Fi

The success of Wi-Fi in enterprise networks can be explained by our constant need for mobility. Mobility assists many organizations to become truly agile. Enterprise users expect the same high performance no matter if they are connected to their docking station or to a Wi-Fi connection in a meeting room. A high number of Wi-Fi access points are placed all around the campus buildings in traditional workplaces, such as meeting rooms, cafeterias, parking lots, or other informal locations, such as between buildings where employees can enjoy a nice spot in the shade.

The new Wi-Fi 6 technologies support multi-band and multi-user MIMO to achieve much higher air frequency utilization. The access points connected to the wired network via Ethernet cables must now run at higher speeds to fully enjoy these new capabilities. Examples are shown below:

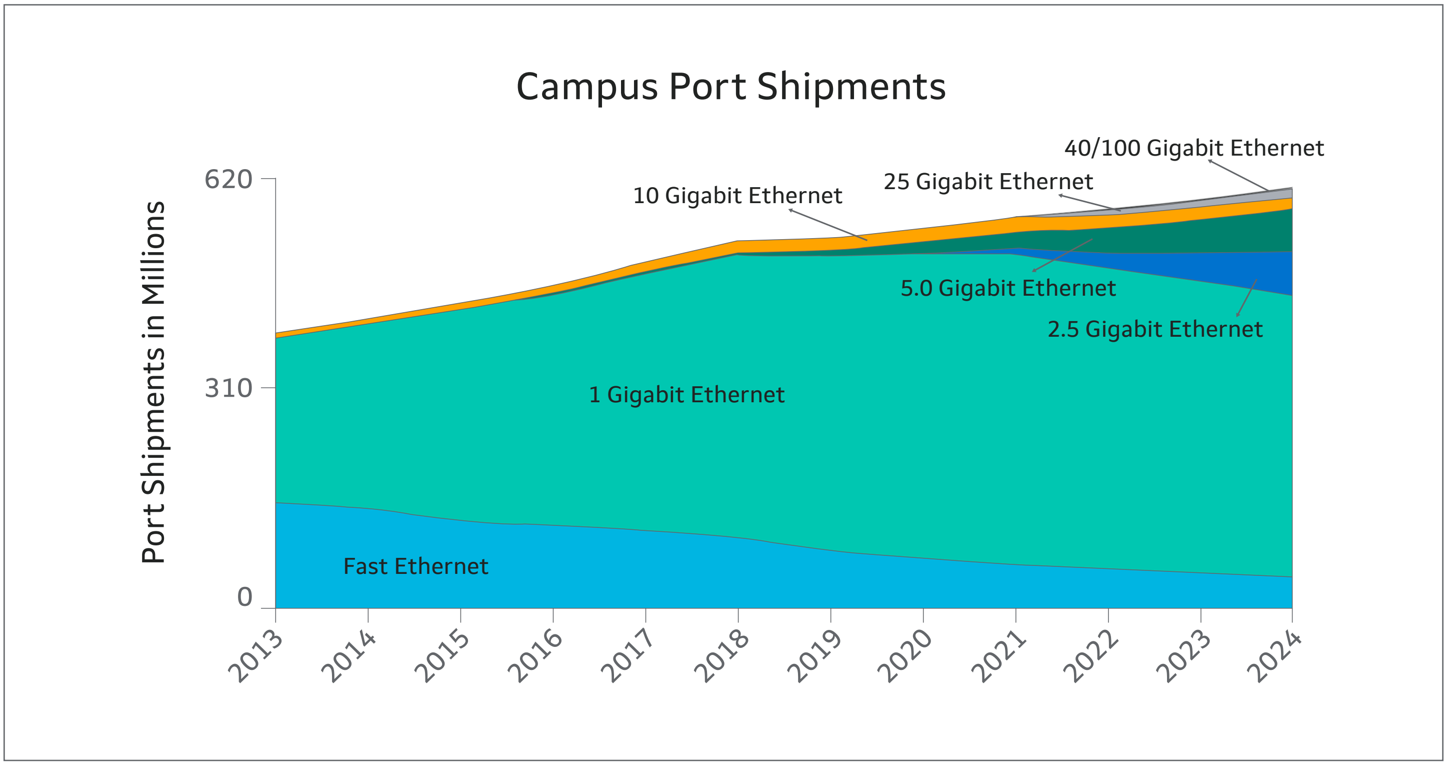
- IEEE 802.11ac (a.k.a. Wi-Fi 5) Access Points – Typically with 1GbE and 2.5GbE copper ports
- IEEE 802.11ax (a.k.a. Wi-Fi 6) Access Points – Typically with 1GbE, 2.5GbE and 5GbE copper ports

Some choose to even build their wireline infrastructure to support 10GbE Copper to accommodate future expected enhancements in this area.

2.5GbE to the desktop

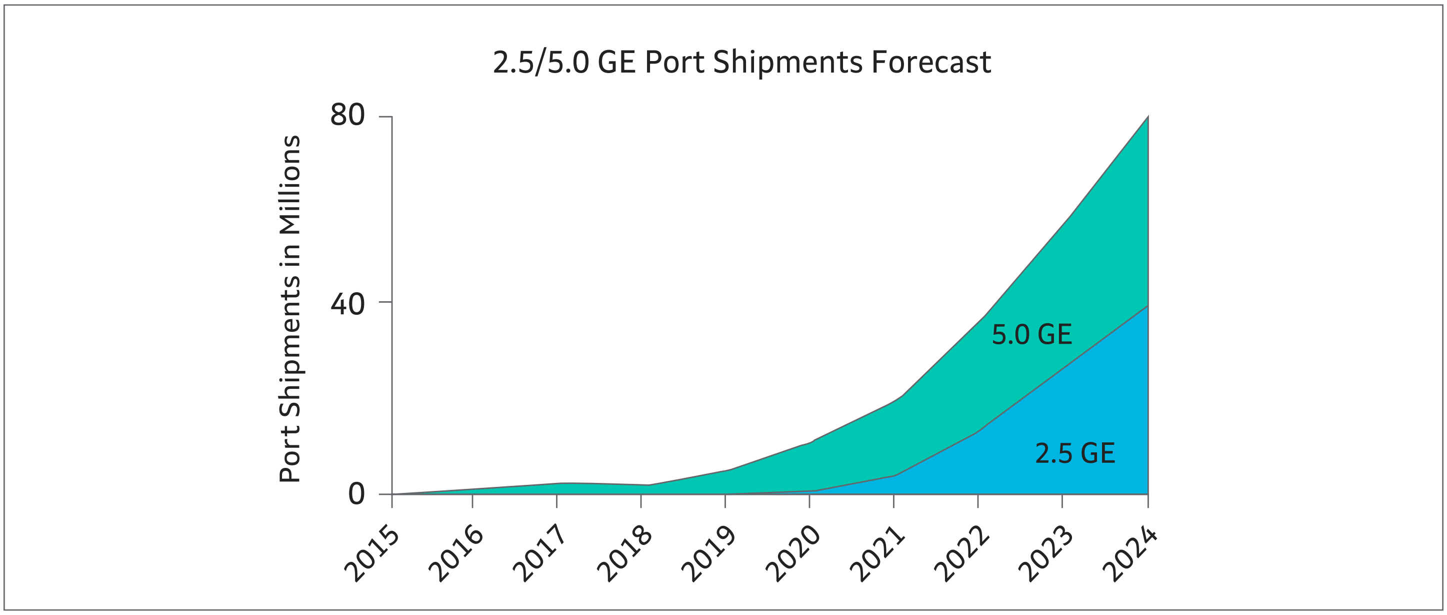
While access points that aggregate many users clearly are the first to request higher speeds than 1GbE, desktop and laptop users also want to gain from the high-speed copper technology. What started with desktop for gamers is now expected to reach PCs and laptops based on single-port low-power 2.5GbE PCIe Ethernet controllers.

Figure 3 and Figure 4 below show the expected volume of the different port types, with 2.5GbE and 5/10GbE gaining a similar share. (Note that when Dell'Oro Group says '5.0 Gigabit' they are referring to 5GbE and 10GbE Copper, and when they say 10 Gigabit, they mean 10GbE Fiber.)



Source: Dell'Oro Group

Figure 3: Campus Port Shipments



Source: Dell'Oro Group

Figure 4: 2.5/5G Port Shipments - Forecast

Borderless Enterprise

Enterprises are becoming borderless by moving compute and storage workloads to the cloud and operating in what is called a "hybrid cloud" model. In the past, you could have seen a department that had a server or two located at the department, and physically connected to the same access switch as the department members. This caused what is called 'East-West' traffic. But once resources are moved to the cloud, North-South traffic is more significant, and with that a lower oversubscription ratio between the total downlink bandwidth to the bandwidth of the uplinks.

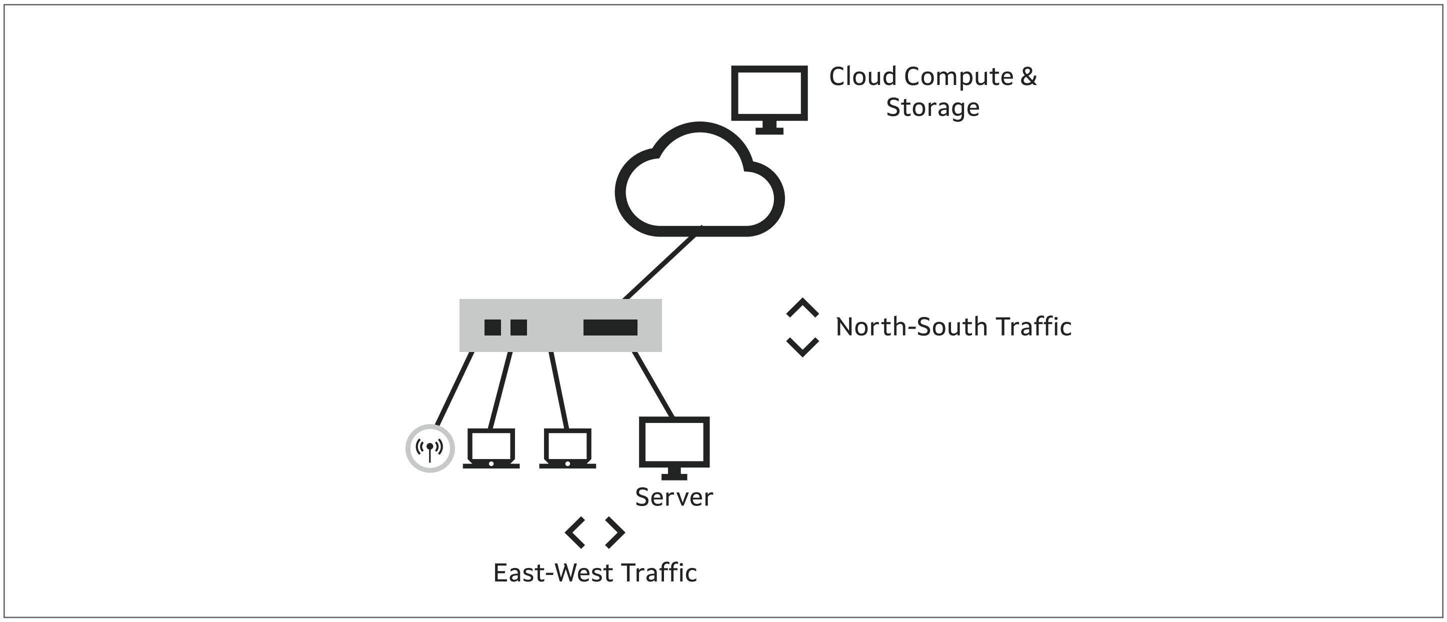


Figure 5: The move to the cloud

Technology is Here to Serve

New technologies have arrived to address these new needs providing solutions in a similar power envelope and at an even lower cost.

SerDes Technology

For many years the dominant SerDes (Serialization / Deserialization) technology has been 10Gbps (actually slightly higher). This did not mean that higher Ethernet speeds were not achievable. 40GbE was based on four 10G SerDes working together, and even a 100GbE standard was defined based on 10 such links. But for power- and cost-efficient solutions, transmission must be faster. Actually, 'faster' may not be the correct term, as propagation is limited by the speed of light. A better term would be 'shorter', as what happens is that the size of each bit becomes smaller and thus the duration of transmitting time is shorter.

10G SerDes was based on the well know NRZ (non-return-to-zero) modulation, and in recent years the same modulation technique was able to work at more than 25Gbps speeds (up to 28Gbps speeds) due to new CDR (Clock Data Recovery) capabilities. The 25G SerDes was the main enabler for the popularity of 100GbE, first in Data Center and then in Enterprise, which was based on four 25G links (a.k.a. 100G-R4).

But SerDes technology has not stopped here; a new modulation technique called PAM4 (Pulse Amplitude Modulation with four amplitude levels) allows for the transmission of 50G (up to 56Gbps) serially while at the same logic baud rate as a 28G NRZ-based signal. 50G PAM4 SerDes is driving new port speeds of 50GbE, 100G-R2 (i.e. 100GbE over two lanes of 50G SerDes), 200G-R4 and even 400G-R8.

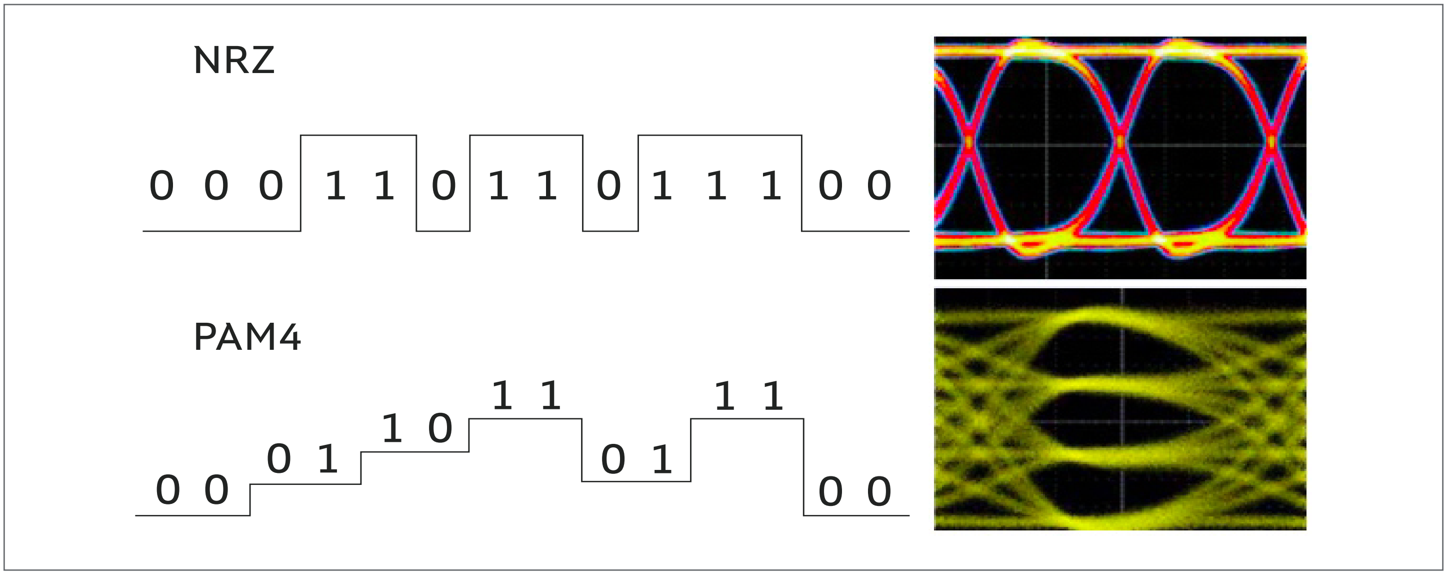


Image Credit: Tektronix

Figure 6: PAM4 vs. NRZ

The technology jump from 10G to 25G and even to 50G, usually does not require changing the existing fiber infrastructure, such as between building floors. The physical backward compatibility of the cage form factor, between SFP+, SFP28 and SFP56 connectors, allows for an easy and gradual transition. Moving from say 10GbE to 25GbE boosts the performance by 2.5x, with only a gradual increase in cost, yet a dramatic reduction in cost-per-bit.

The cost reduction is also achieved when IT managers can reduce the number of optical transceivers and reduce the inherent inefficiency associated with link aggregation. Deployments requiring 20Gbps of uplink traffic in the past required the installation of two 10GbE transceivers with link aggregation between them, which meant effectively less than 20G of traffic. A single 25GbE transceiver utilization of the same fiber will provide more for less.

Multi-Gig Technology

Breaking the barrier of 1Gbps speed over copper cabling was the main goal of the NBASE-T Alliance, formed in 2014 by Aquantia (now part of Marvell), Cisco and Xilinx. The goal of the Alliance was creating and promoting 2.5GbE and 5GbE technology for enterprise network infrastructure on existing enterprise cabling originally designed for 1GbE only.

The NBASE-T Alliance's efforts contributed to the standardization of 2.5GBase-T and 5GBase-T Ethernet transmission technologies under the auspices of the IEEE, which ratified the 802.3bz standard in October 2016. Subsequently, the NBASE-T Alliance merged with the Ethernet Alliance, which continues to promote the technology.

The 1000Base-T, 2.5GBase-T and 5GBase-T interfaces all use the very popular Cat5E cable that has a huge installed base around the world. Even when operating at 5GBase-T mode, it can reach distances of more than 100 meters. A standard for 10GbE over copper links was actually defined, earlier, in 2006, by the IEEE as 10GBase-T (or IEEE 802.3an), but the higher rate 10GBase-T requires higher quality cables, and to reach 100 meters, one needs to use Cat6A or Cat7 cabling.

This difference in cabling types suggest that 2.5G and 5G interfaces will be most popular, and 10GBaseT will be used when necessary and when the cabling installed in the organization allows it. 10GBaseT provides a future-proof technology that will be able to handle needs for years to come.

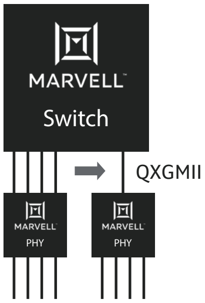
10GBase-T technology is also used to connect servers that are physically distant from the switch, thus coaxial DAC (Directly Attached Cable) cannot be used.

Deciding on the appropriate port speed does not only depend on the installed cabling infrastructure and the organizational needs, but also on the cost differences associated with each technology.

As can be imagined, cost is derived from multiple factors, including the switch device, copper PHY devices, PCB types, power supply and fans. Supporting multi-gig high-speed interfaces on copper is a big challenge that involves complex digital signal processing. Clearly the complexity for supporting 1Gbps speeds is much less than supporting 10Gbps speeds, and silicon vendors like Marvell have created dedicated PHY devices optimized for the different speeds.

Multiplexing Techniques

When looking at the switch device, one needs to look at the total switching capacity and the number of Ethernet ports and SerDes connections between the switch device and the PHY device. The number of SerDes can be lower than the number of ports by multiplexing several ports on a single SerDes. While in the 1G era, a common interface was QSGMII, which allowed multiplexing four 1G ports on a single SerDes. When using higher speed SerDes, one can multiplex more ports, and higher speed ports, and thereby simplify board design to create smaller packages and reduce the overall solution cost. The below combinations are worth noting:



OUSGMII, multiplexing up to 8x1G ports over a single 10G SerDes

10G-QXGMII, multiplexing up to 4x2.5G ports over a single 10G SerDes

20G-OXGMII, multiplexing up to 8x2.5G ports over a single 20G SerDes

20G-QXGMII, multiplexing up to 4x5G ports over a single 20G SerDes

20G-DXGMII, multiplexing up to 2x10G ports over a single 20G SerDes

Such multiplexing techniques are especially important on the switch side; one can imagine the burden on the switch to support 48 ports via 48 SerDes vs. only supporting 12 or even six SerDes for the same number of ports, depending on the multiplexing technique.

Different Organizations, Different Topologies

Organizations choose their network topology based on many factors: the geographical distribution of users, the expected traffic patterns, performance needs, budget constraints and more. They are also influenced by technological shifts and by trends in adjacent markets, such as recent changes in data centers.

The size of the organization (or more precisely the size of a specific location of the enterprise) is not the only factor that determines the number of layers the network should be built on. The number of ports in each layer is just as important, and in turn, depends on the form factor, connector types and more.

Different people name the layers differently, but usually a three-layer topology will be defined as: Access, Aggregation and Core. Terminology adopted from data center solutions is now also being used, i.e. Leaf, Spine and Core.

We will now discuss each of these layers.

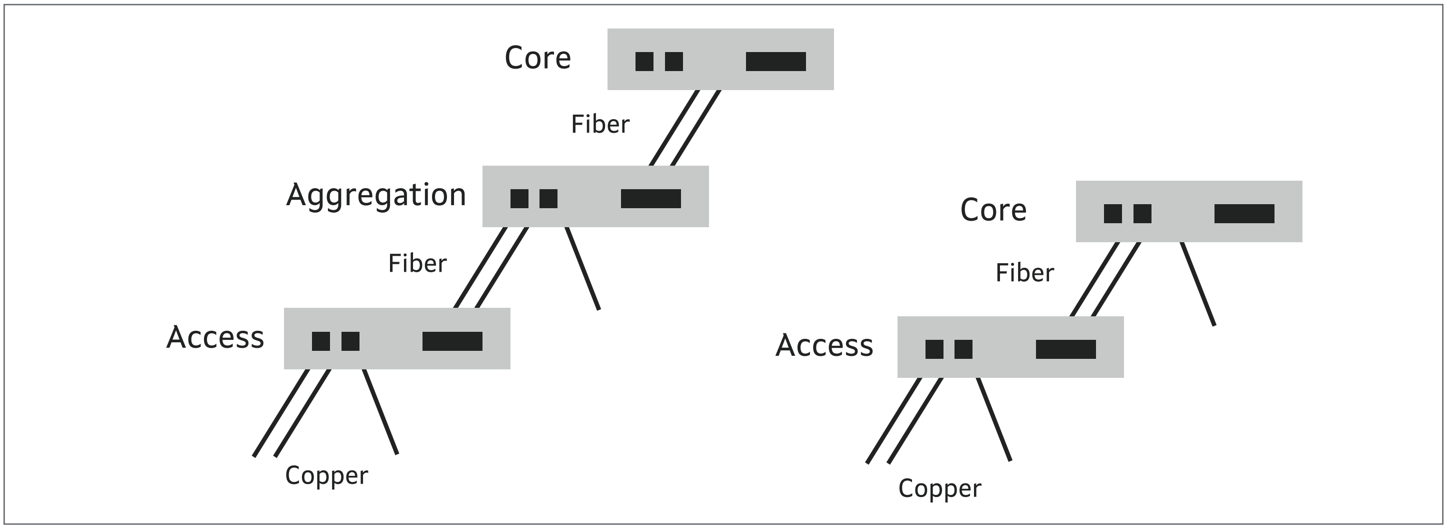


Figure 7: 2-Layer vs. 3-Layer Topology

Access Layer

Access switches are usually deployed in either stand-alone pizza boxes or in stackable switches. When stackable switches are concerned, the stacking can be done using some of the front panel uplinks or can be done via dedicated stacking ports, which are generally located in the back of the pizza box.

Stackable switches are a preferred choice by some for the following reasons:

Scalability – This means growing the size of the switch, while maintaining a single management interface for the entire stack. A similar capability can also be achieved in a chassis. This scalability also allows for the reduction in the number of layers in the network.

Reuse of Uplink Fibers – Instead of each pizza box having at least two uplink fibers for redundancy, which means more fibers that need to reach each floor, if the bandwidth allows, the uplinks can be shared for the entire stackable switches.

Pay As You Grow – This is unlike a chassis design that has higher initial costs of the mechanical passive components, the overhead of the fabric cards and the need for shared resources built for the max configuration, such as CPU, power supplies and the fans; in stackable switches, each box adds its own power, cooling and CPU resources.

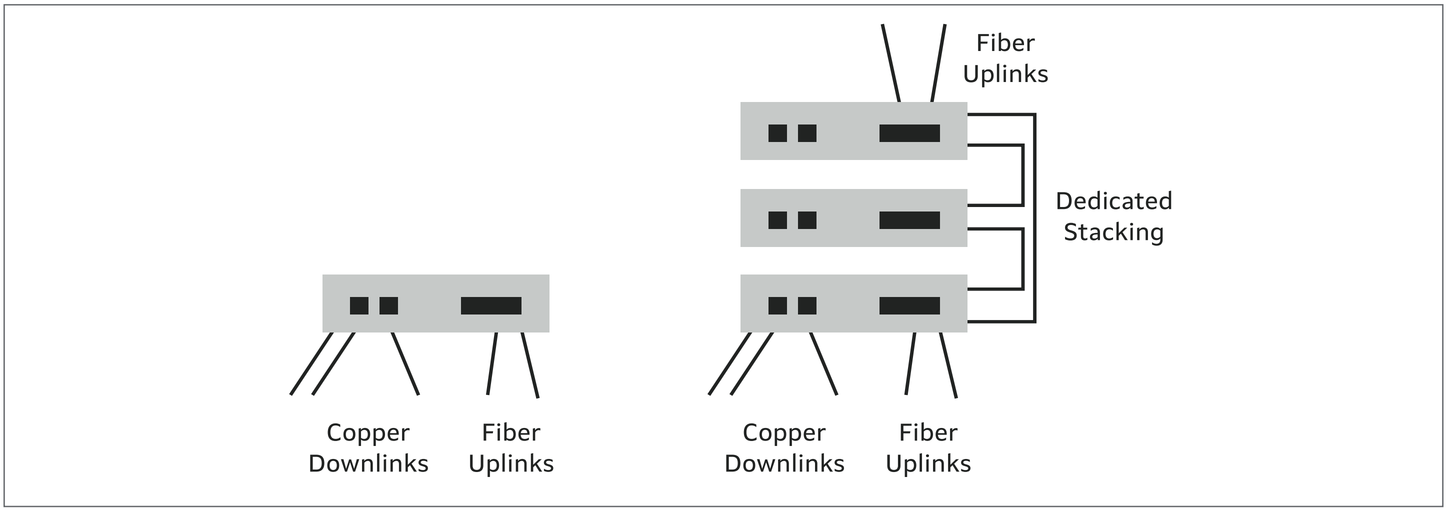


Figure 8: Standalone Switch vs. Stackable Switch

There is no one answer to what the bandwidth of the stacking port should be. The goal is for it to not get congested, while balancing the need to not over-extend resources for unrealistic scenarios. The bandwidth received from the copper downlink can go to a local uplink on the same switch (without using the stacking ports) or to uplinks on other switches in the stack. The bandwidth received on a stacking port can continue to flow in the ring or be sent to ports located on this switch (reducing the load on the stacking ports).

A typical rule-of-thumb is that the bandwidth of the two stacking ports is similar to the bandwidth of the downlink plus the uplink ports. For example, previous generation stacking systems used to have 48x1G downlinks + 4x10G uplinks + 2x40G stacking.

Stacking ports usually use DACs which allow for connecting of the switches over a few meters without the need for a PHY device or optical transceiver.

To mix or not to mix, that is the question

As discussed, different costs are achieved depending on if you build a 1G only solution or a multi-gig solution. At least initially, the demand for ports faster than 1GbE will be from Wi-Fi access points, which are lower in number than the Ethernet sockets spread in the campus for PCs, IP phones and laptops. Because of that, some system vendors have chosen to build switches that mix many 1G ports with a few multi-gig ports, for example, 32x1G + 16x2.5G ports reaching 48 RJ-45 connectors on the front panel.

While this has cost advantages, it puts a burden on IT personnel to make sure they connect the correct cable to the correct port, e.g. the cable that came from the Wi-Fi access point to the 2.5G-enabled port and the cables that come from the desktops to the 1G-enabled ports. This may lead to human error as all have the same RJ-45 mechanical connector.

Therefore, others prefer a unified configuration such as 48x2.5G copper downlink ports based on Marvell's new Prestera switches and PHY devices. This helps avoid installation mistakes. In actual deployments, not all ports will actually work at 2.5G, but once the end points are upgraded, the infrastructure is all ready for them.

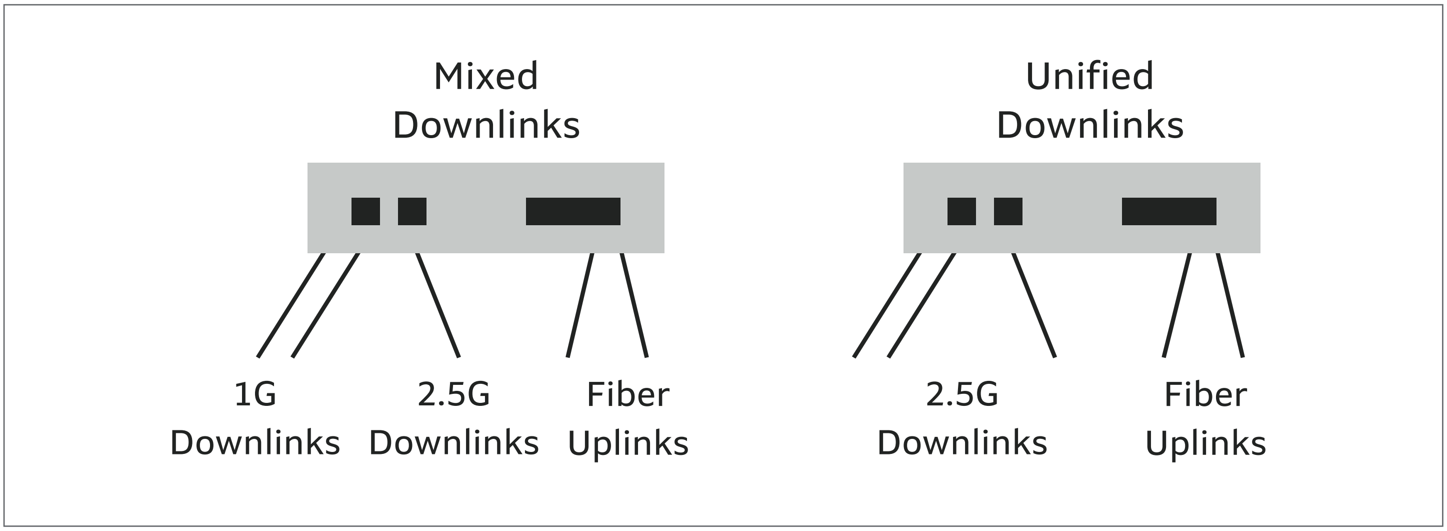


Figure 9: Mixed vs. Non-Mixed Copper downlinks

Aggregation/Core Layers

The number of ports in a fixed-configuration aggregation/core box depends on the type of connectors used. When QSFP connectors are used for 40G-R4 (40GbE based on four lanes of 10G SerDes), 100G-R4, 200G-R4 ports, it will most probably include 32 ports, in line with front panel constraints.

However, when SFP connectors are used (SFP+, SFP28, SFP56), one can easily build a 48-port aggregation pizza box. Similar to the stackable pizza boxes that increase the number of ports in the access layer and contribute to the flattening of the network, the move to SFP connectors in the aggregation layer also helps flatten the network.

Another solution that allows for building of a single layer with many ports in the aggregation and core, is the chassis architecture.

Chassis Design

Chassis design provides several benefits over pizza boxes that make it the design choice of many organizations, especially for the core of the enterprise network. These benefits include:

Scalability – Adding more line cards increases the number of supported users and reduces the number of layers in the network.

Flexibility – A variety of line card types address the different mixtures of interfaces.

High Capacity – Offers the ability to reach very high switch capacity by utilizing multiple devices working together.

Front Panel Space – The physical size of the chassis provides the area needed for many connectors.

Redundancy – The central card is duplicated, thus avoiding a single point of failure. Because users are connected to the line cards, each user must be connected to two line cards to ensure resiliency and to configure LAG/ECMP.

There are two main architectures to chassis design: distributed and centralized. In a distributed design, the packet processing functions are performed by the line cards, while in a centralized design, these functions are performed by the main card. These differences are substantial. But when looking at a chassis design from the outside, these differences cannot always be noticed. See Figure 10 below.

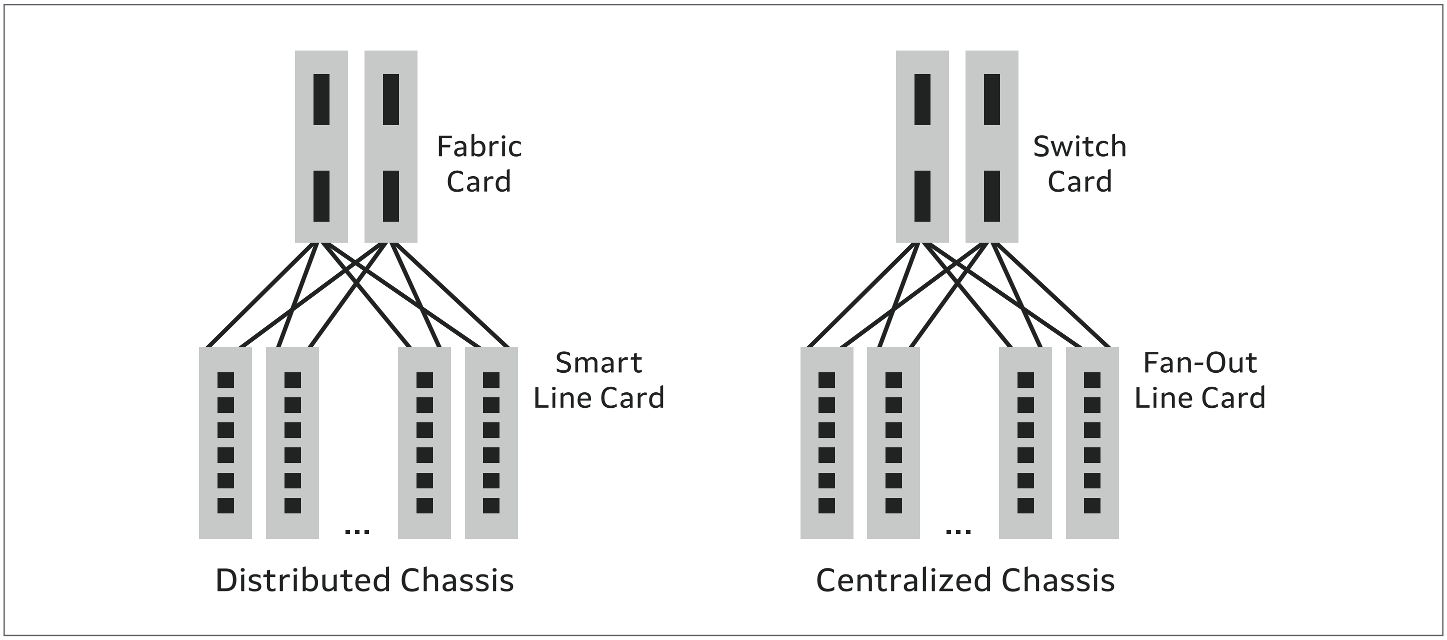


Figure 10: Chassis Architectures

When a packet arrives to the smart line card, it is fully processed and the destination of the packet is determined. This information is amended to the packet using a small tag so that once it arrives to the fabric card, it needs only to look at this tag, then is forwarded to the correct egress line card. The egress line may further process the packet and perform additional packet modification.

However, when a packet arrives to the line card in a centralized architecture, it basically only adds the source port from which it arrived and always sends it to the main card, which in turn is the one that performs the full packet processing, including packet modification. The egress line card basically performs port fan-out, then forwards the packet to the correct connector. In some cases, it can also do packet replication, usually when all copies of the packet look the same.

	DISTRIBUTED	CENTRALIZED
Local Switching	Possible	Always reaches the center
Software Complexity	High	Medium
Switching Capacity	Super high, via adding more fabric elements	High
Multicast	Efficient replication at egress line card	Packet may be sent multiple times on backplane

The main advantage of distributed chassis design is the ability to reach super high switching capacities. This has become less critical with the introduction of super strong single-chip devices reaching multi-terabit total bandwidth, and thus tipping the scale towards centralized chassis.

Migration Strategies

Organizations have different migration strategies depending on their existing topologies, performance needs, budget constraints and more.

Let's start by evaluating some migration strategies for the access layer, where volume is high.

Access Layer - Performance Jump

The vast majority of access switches have 24 or 48 1GbE copper downlink ports and 10GbE fiber uplinks. Organizations that choose a performance jump may decide to create a 2.5x jump in bandwidth as illustrated in Figure 11. Moving all 48 ports to be 2.5G ports is done for uniformity and simplicity, while in initial real-life installations, many of the downlink ports will actually be 1G. In such cases, fewer 25G uplinks will be used and the rest left for future upgrade.

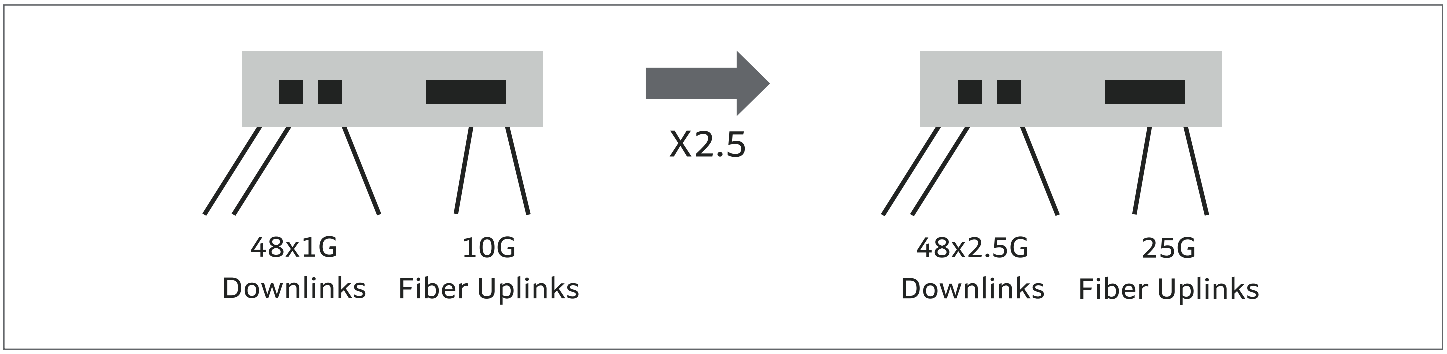


Figure 11: Performance Jump

Access Layer - More for Less

Organizations that choose to deploy stackable pizza boxes and want moderate bandwidth growth at the same or lower cost, may choose the solution illustrated in Figure 12. In this example, the downlink ports are a mixture of 1G and 2.5G copper ports to reduce cost. The use of 25GbE uplinks allows for fewer optical transceivers compared to 10GbE, which again reduces cost. Lastly, the cost of the stacking links over DAC cables is proportional to the number of lanes in the cable. 40GbE links are built from 4x10G links, while 50G stacking links are made from either 2x25G links or from a single 50G link. This reduces the cost of such connections.

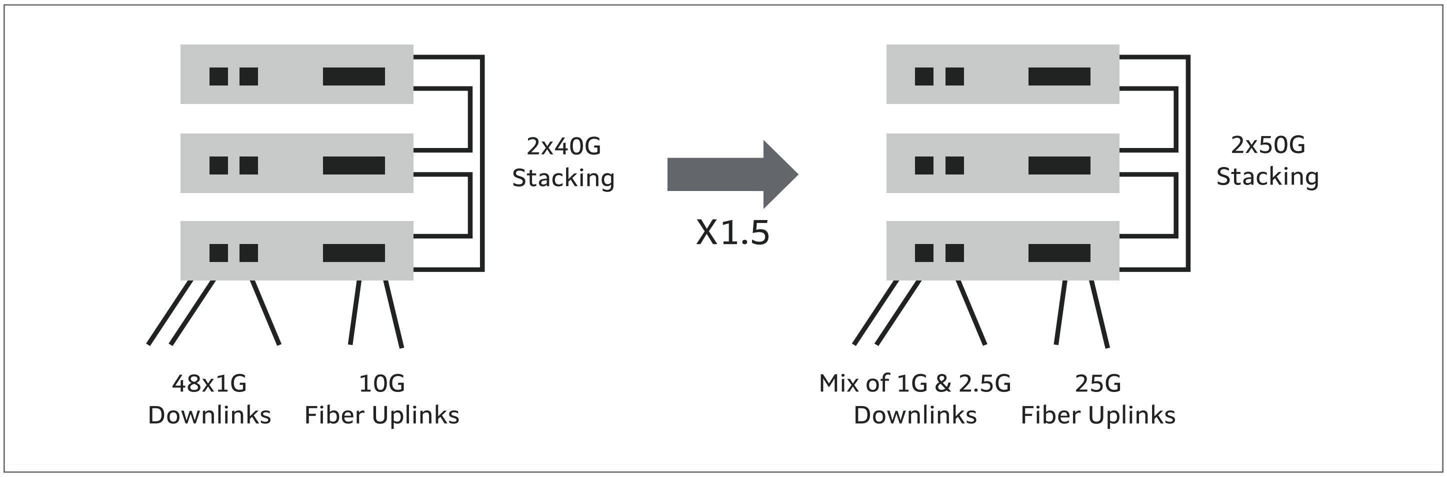


Figure 12: More for Less

Access Layer - Symmetric Growth

When moving from 1G copper interfaces to 5G interfaces, some organizations will choose to maintain the same ratio and grow the bandwidth of the uplinks and stacking accordingly. For example, they will move from a 40G-R4 DAC to a 200G-R4 DAC for the same 5x bandwidth growth.

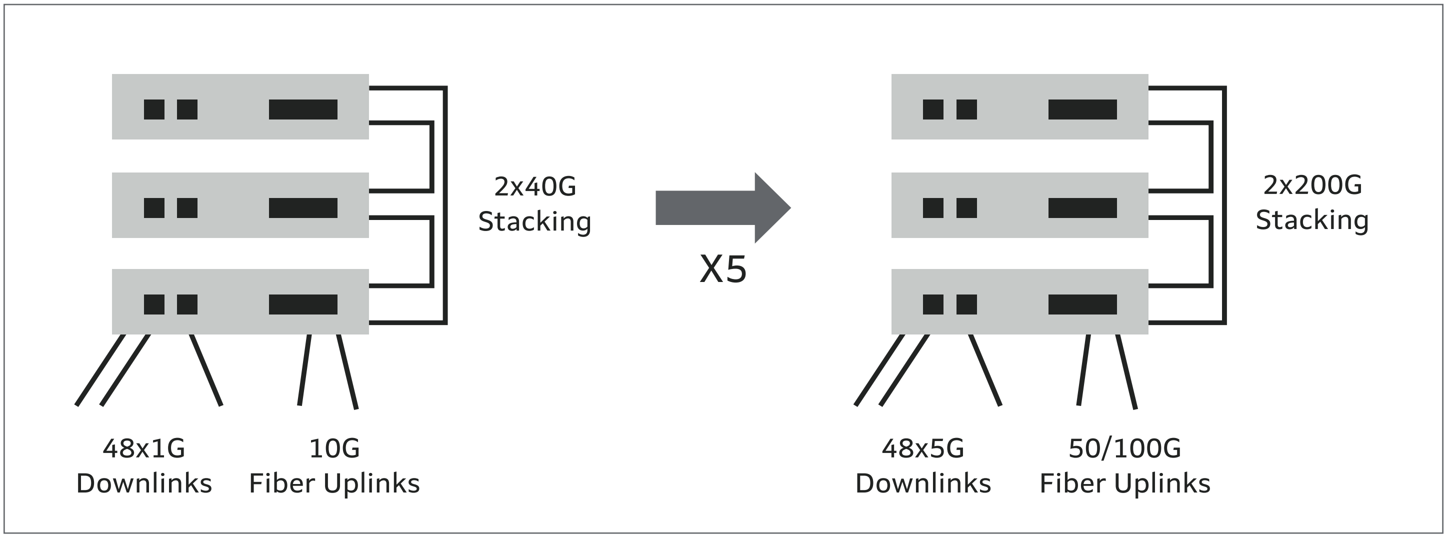


Figure 13: Symmetric Growth in Access

Aggregation and Core Layers

The aggregation layer needs to follow the same change happening in the access layer. When the access layer moves to 25G fiber uplinks, the aggregation layer needs to move to 25G fiber downlinks as well.

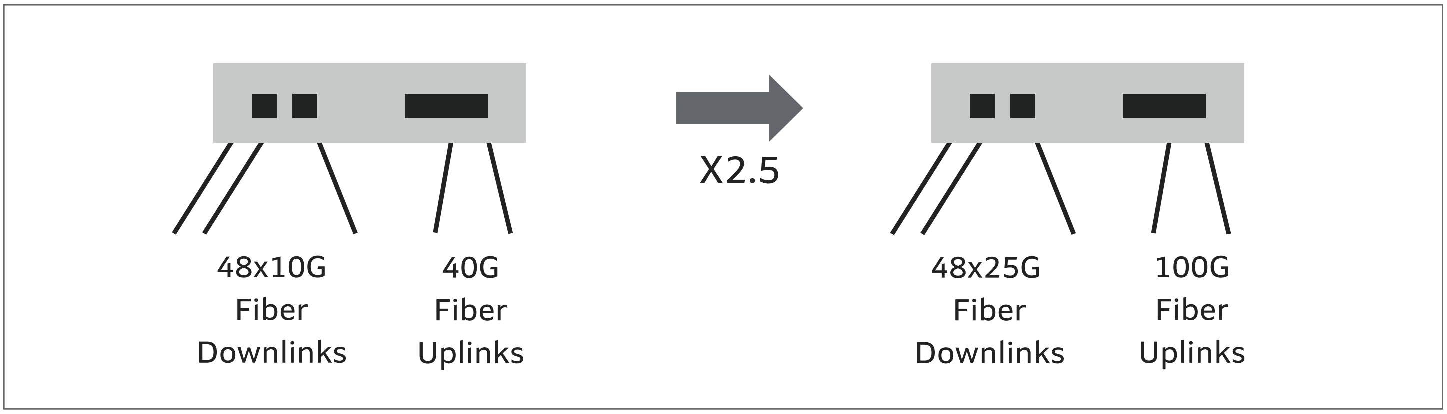


Figure 14: Aggregation Move to 25G

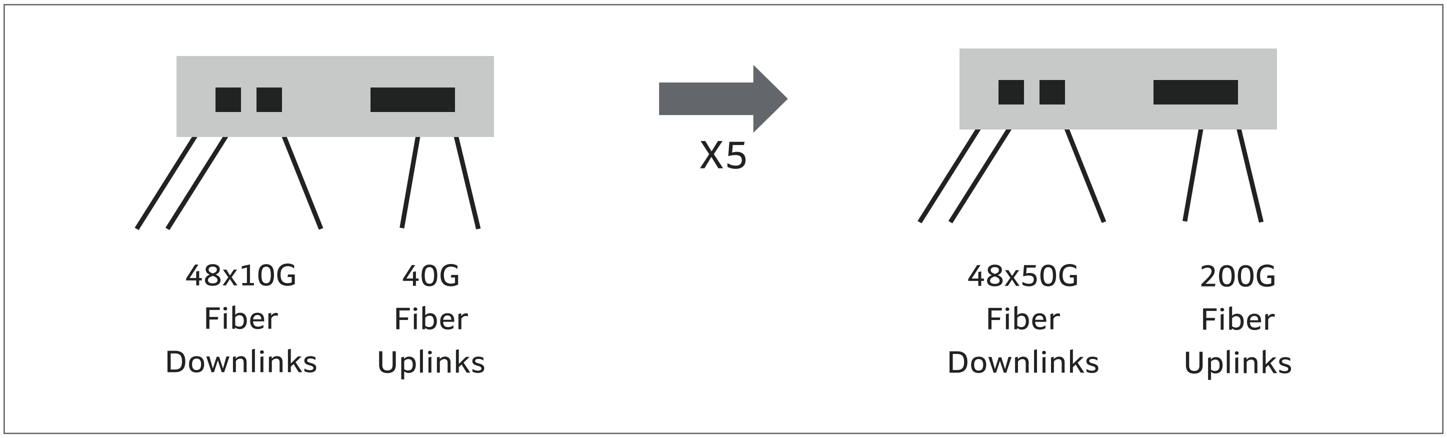


Figure 15: Aggregation Move to 50G

Similar matching is needed when an aggregation layer is connected to a core layer.

The above migration options are achieved using the new members of the Prestera 6K series connected to the Prestera 4K and 3K series.

Migrating in Harmony

Figure 16 below illustrates the need to align the changes in speed in each layer with the adjacent layer. It also underscores the importance of upgrading the entire network to be based on similar SerDes technologies. This ensures the entire network will work in harmony, without creating new bottlenecks in order to achieve the desired performance.

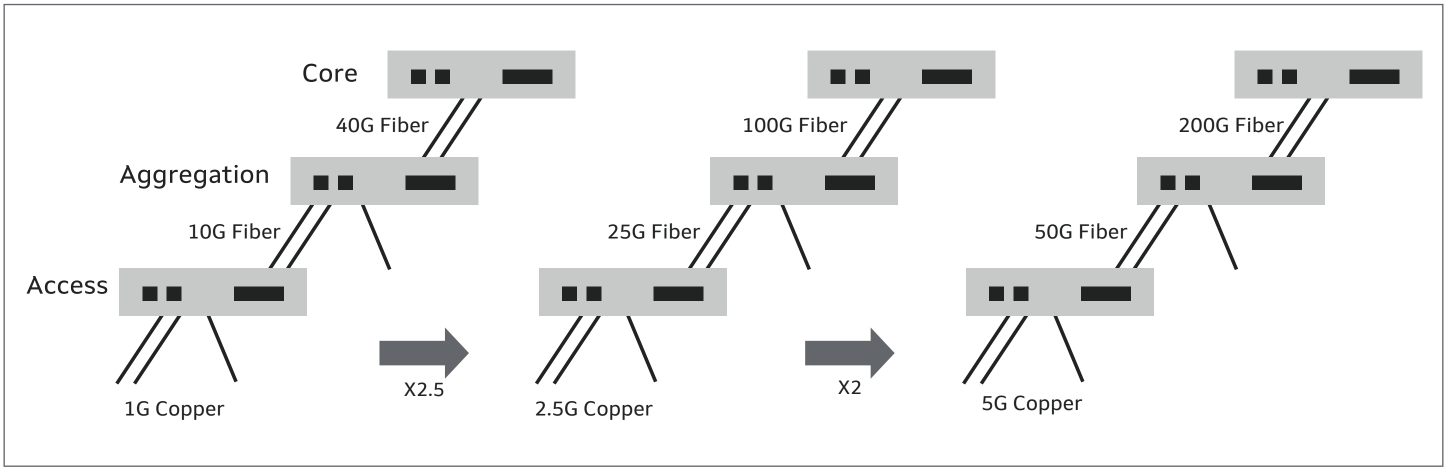


Figure 16: Migrating in Harmony from Access to Core

Conclusion

Marvell's newly announced Prestera devices provide the building blocks needed to develop a new generation of enterprise switches from access to core, addressing the new performance needs for years to come.

A refresh of the enterprise network built around new high-speed SerDes, new Ethernet port types, and new stand-alone, stackable and chassis designs will enable the adoption of higher bandwidth services and support for emerging real-time applications and an expansion to multi-cloud borderless environments.

About the Author

Gidi Navon

Principal System Architect

Gidi Navon is a member of the Switching CTO team at Marvell. In his role, Gidi defines new networking devices and software solutions, and drives the introduction of new technologies into Marvell's infrastructure products. Specifically, he is responsible for leading initiatives focused on defining new devices and network visibility solutions for the switching portfolio. Gidi joined Marvell eight years ago, after holding senior product and architectural positions at Nokia Siemens Networks for seven years, defining carrier packet platforms. Previous to that, he held various system architecture positions at leading silicon and system companies. Gidi received his Bachelor of Science degree in Electrical Engineering from the Technion Israel Institute of Technology and his MBA from Tel-Aviv University. He holds multiple patents in the field of networking and computer communication.

References

[1] Cisco Visual Networking Index (VNI) Complete Forecast Update, 2017–2022: APJC Cisco Knowledge Network (CKN) Presentation; https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/1213-business-services-ckn.pdf