QSFP-DD MSA

Accelerating 400GbE Adoption with QSFP-DD

Using the QSFP form factor to effectively conquer the challenges facing the next generation of Ethernet equipment

3/10/2017

QSFP-DD is a new module and cage/connector system similar to current QSFP, but with an additional row of contacts providing for an eight lane electrical interface. It is being developed by the QSFP-DD MSA as a key part of the industry's effort to enable high-speed solutions.

www.QSFP-DD.com

The QSFP-DD Form Factor

The QSFP-DD module form factor is the industry's smallest 400GbE module providing the highest port bandwidth density. This form factor leverages the industry's manufacturing capability and cost structure that supports QSFP+ and QSFP28, the industry's de facto standards for 40GbE and 100GbE. QSFP-DD can support 36 ports of 400GbE in a single Rack Unit (RU) providing over 14Tb/s of bandwidth.
The QSFP-DD will support:

- 3m of passive copper cables
- 100m over parallel multimode fiber
- 500m over parallel single mode fiber
- 2 km and 10km over duplex single mode fiber
- Backward compatible with all QSFP based transceivers from 40G to 200G

This paper discusses:

- Traffic growth
- Evolution of module form factors
- Backward compatibility accelerating market adoption
- QSFP-DD design

Traffic growth is placing pressure on datacenters

Cisco's Visual Networking Index projected that annual global IP traffic should have exceeded 1 Zettabyte in 2016 and will more than triple by 2020. Traffic continues to grow in all dimensions: new Internet users, number of devices per user and usage per device —the last dimension boosted by the growth of video, augmented and virtual reality services as well as the transition to highly accessible mobile usage.

Figure 1 provides a high level comparison of the service adoption drivers between 2015 and 2020.[i]

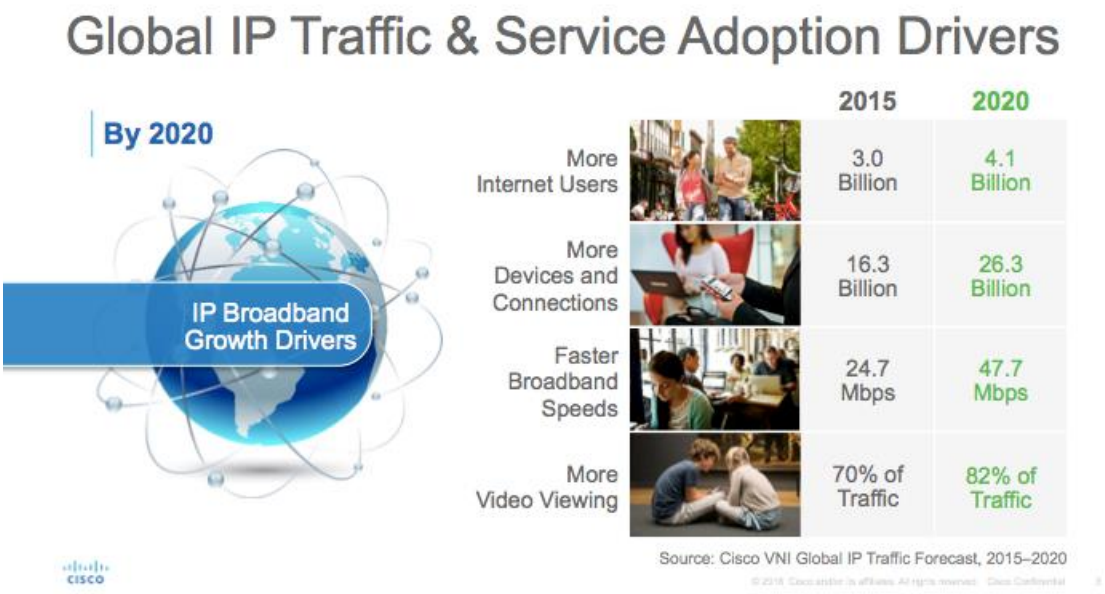

Figure 1: Global IP Traffic and Service Adoption Drivers

And where does this incredible amount of traffic come from? A virtuous cycle, as technology advances, costs are driven down, more services become viable, which drives more traffic to datacenters and in turn creates more demand. According to Cisco's Global Cloud Index, most Internet traffic since 2008 already originates or terminates in a data center. To manage this seemingly endless growth in traffic, higher capacity networks and data centers are being developed that can be scaled to economically provide these services. The market trend to manage the cost of services is to drive them to the cloud where resources can be quickly allocated as required. As a result, global data center IP traffic

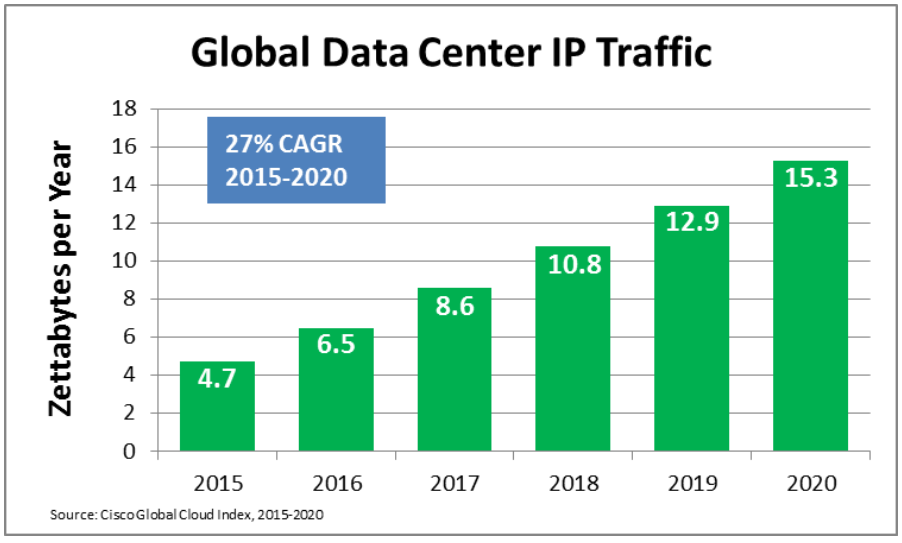

Figure 2: Global Data Center IP Traffic

is forecasted to outpace total global total IP traffic. Figure 2 shows the data center IP traffic forecast from Cisco's Global Cloud Index.

Intra-DC traffic dominates

Above, a figure of 1 ZB was cited for 2016 yet much larger numbers appear in Figure 2! Cloud services are often hosted by hyperscale data centers. The majority of the traffic in these data centers (east west traffic) stays within the data center itself, roughly 75%. This means that even though yearly global IP traffic today has broken the Zettabyte barrier in 2016, the amount of traffic managed within these data centers is higher yet.

These data centers use networking architectures that flatten network topologies to manage the services and allocate resources quickly. These flatter topologies (leaf and spine) require many high speed connections between switches (see Figure 3).

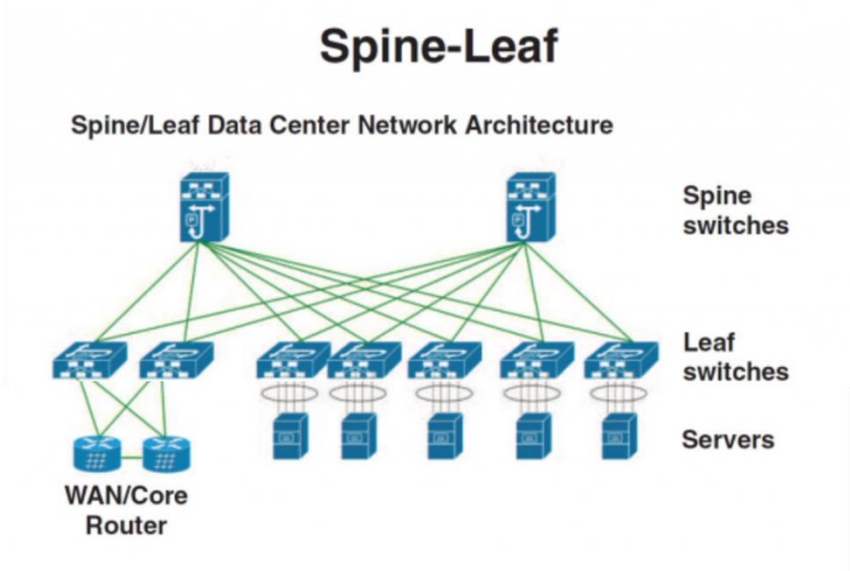

Figure 3: Spine/ Leaf Data Center Architecture

A single hyperscale data center may have hundreds of thousands of servers and thousands of Ethernet switches with as many as 36 ports per RU or even more. The number of connections at 10GbE and higher are staggering.

Furthermore, the growth of these high-density intra-datacenter links is compounded yet again by the disproportionate growth of the hyperscale portion of the total datacenter market. Roughly one quarter of all installed servers in 2016 were in hyperscale data centers but that share will rise to almost half by 2020.[ii]

Module evolution has chased ASIC development

At the core of the Ethernet switches are Applications Specific Integrated Circuits (ASICs) made by companies like Broadcom, Cisco, Mellanox and others. The usable bandwidth of ASICs in Ethernet switches can be limited by the physical dimensions of the pluggable ports themselves.

The most efficient switch design requires that the number of ports in a single RU match the capacity of the switch ASICs used. The physical size of the switch could be increased to accommodate larger form factor modules, but this is inefficient. Increased rack space, longer trace lengths would lead to higher trace loss on printed circuit boards, requiring more re-timers, etc. The size and bandwidth of module ports need to keep pace with the ASIC capability.

Back in 2010/2012 timeframe, ASICs that could provide over 1Tb/s bandwidth in a single RU were available. The SERDES from these ASICs operated at 10Gb/s. Up until this point, SFP+ form factor had been extensively used, but SFP+ would effectively limit the total port bandwidth to about half of the ASICs capability. For this next generation, QSFP+ was used. QSFP+ was able to provide 1.4Tb/s of bandwidth in a single RU. In fact in many applications, QSFP+ was used as a high density four by 10GbE port with breakout cables and optics.

But a few years later, history repeated itself. With the next CMOS node and continued design innovations, ASICs could support >3Tb/s of bandwidth from a single RU. The 40GbE QSFP+ provided no more 1.4Tb/s in a single RU with 36 ports, stranding more than half the capacity of the ASIC. So, the QSFP form factor had to transition to the QSFP28 which continued to use 4 lanes of electrical I/O, but increased the bandwidth by a factor of 2.5 to 25G SERDES per lane. This again aligned with the ASIC technology allowing QSFP28 to deliver 100GbE per port.

Figure 4 shows the relative bandwidth density of various optical modules, by dividing the module bandwidth by its width. The chart shows that QSFP-DD has the highest bandwidth density of all form factors.

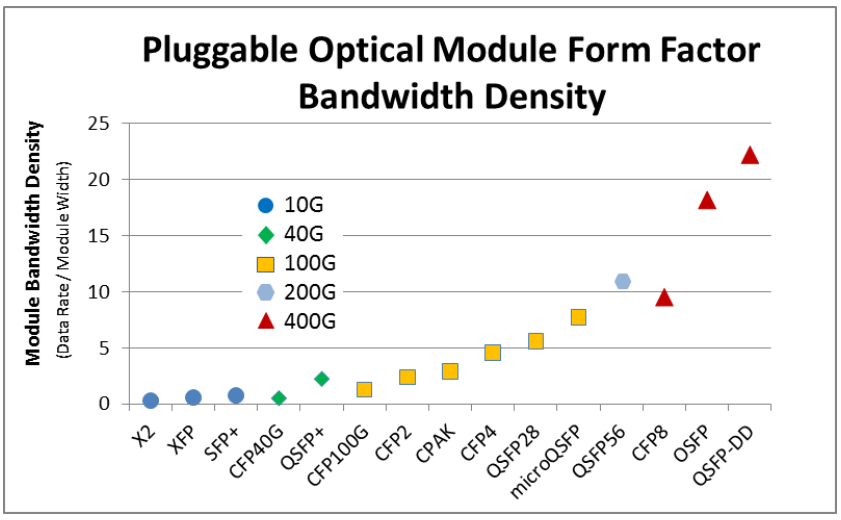

Figure 4: Relative Pluggable Optical Module Form Factor Bandwidth Density

Backward compatibility feature of QSFP+ and QSFP28 enabled networks to migrate quickly to the next generation

The importance of backward compatibility has been proven in multiple generations of pluggable optics. The benefit of backward compatibility is not only to accelerate adoption of the new module type but to accelerate overall network migration by providing a convenient bridge between existing infrastructure and the next generation. This was evident in the successful migration from SFP+ to QSFP+ as described above using breakouts.

For QSFP28 ports, the backward compatibility is enabled by the ASIC. Typically, QSFP28 ports on Ethernet switches and routers can also be configured for 100GbE or 40GbE for QSFP+ modules. Customers can deploy the latest equipment with the latest feature set and continue to use modules they have already invested in where necessary, along with some of the existing infrastructure at 40GbE until they are ready to transition the rest of their network to 100GbE ports. This has accelerated the deployment of QSFP28 equipment. Table 1 shows the evolution of the QSFP form factors I/O and form factor width.

Table 1: SFP+ and QSFP Module Features

Module Type	# of I/O lanes	Electrical I/O	I/O Baud Rate	Module BW	Module Width (mm)
SFP+	1	10Gb/s-NRZ	10G	10Gb/S	13
QSFP+	4	10Gb/s-NRZ	10G	40Gb/S	18
QSFP28	4	25Gb/s-NRZ	25G	100Gb/s	18
QSFP56	4	50Gb/s-PAM4	25G	200Gb/s	18
QSFP-DD	8	50Gb/s-PAM4	25G	400Gb/s	18

QSFP-DD continues the trajectory of QSFP+ and QSFP28

The twin success factors of density and backward compatibility have made the QSFP family of form factors extraordinarily successful. It is second only to SFP - the form factor of choice for 1GbE, 10GbE and 25GbE applications. By the end of 2016, the total number of QSFP modules that have been deployed at 40GbE and 100GbE since their introduction has topped 7M units according to market forecasts from Lightcounting. Lightcounting forecasts the total number of QSFP 40GbE and 100GbE modules to be deployed by 2020 is likely to exceed more than four times that number[iii].

The QSFP form factor has been widely adopted for cost sensitive, high performance applications in data centers, service provider, computing and enterprise market segments. The ecosystems for the QSFP form factors (connectors, cages, modules, cables, etc.) is well established, well down the cost track and continues to grow at double digit rates. QSFP provides a clear path for backward compatibility without the need for port adapters that waste power and increase thermal impedances, allowing users to leverage their investment in modules across at least two generation of equipment deployments.

Continuing forward with the next generation of optical modules using the QSFP form factor is the logical choice. For QSFP-DD (Double Density) designers have made two key changes to the module, while maintaining the key features that enabled the rapid acceptance of QSFP+ and QSFP28.

- The first is to increase the number of I/O lanes of the module from 4 to 8.
- The second is to double the data rate of each lane to 50Gb/s effectively quadrupling the overall bandwidth of the module compared to QSFP28. The IEEE made the choice to standardize on PAM4 signaling (designated as CEI-56G-VSR-PAM4) as opposed to NRZ for the ASIC to Module interface and the QSFP-DD MSA has followed the recommendation.

QSFP-DD extends backward compatibility feature of QSFP

In addition, to make the QSFP-DD port backward compatible to prior versions of QSFP, a second row of contacts was added to the electrical interface as shown in Figure 5. This approach allows QSFP+, QSFP28 and QSFP56 to contact the first row, making the port backward compatible. When a module is inserted in the QSFP-DD port, the port can recognize the type of module and configured to operate in NRZ mode at either 10Gb/s or 25Gb/s per lane for QSFP+ or QSFP28, or at 50G-PAM4 for QSFP56 or QSFP-DD. Since QSFP+, QSFP28 and QSFP56 have a shorter electrical connector, they will only connect to the first row of contacts as shown in Figure 5, engaging only 4 lanes of I/O. QSFP-DD module's electrical connector is longer and will connect to both rows of contacts engaging 8 lanes of I/O as shown in Figure 5.

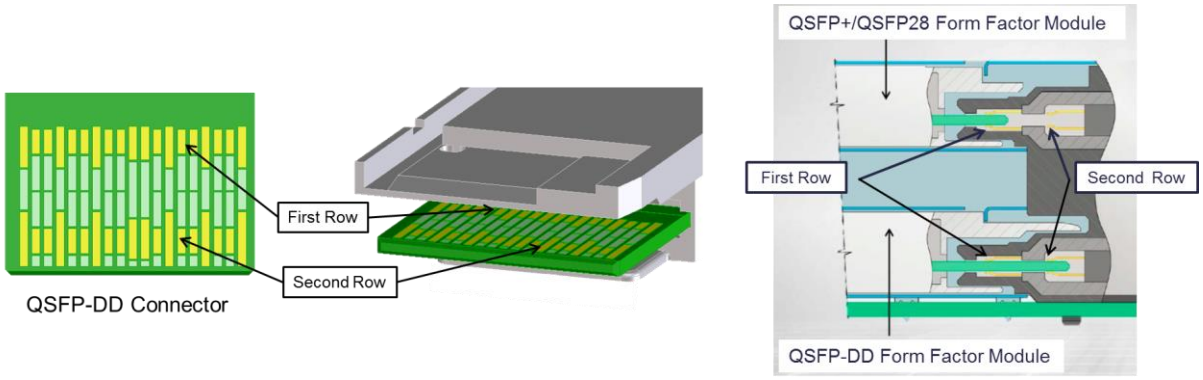

Figure 5: QSFP-DD Dual Row Electrical Connector and Engagement for QSFP and QSFP-DD

Wide variety of optical interfaces supported

One of the key factors for the adoption of QSFP for 40GbE and 100GbE other than the need for higher bandwidth pluggable optics, was the industry support of a wide variety of optical interfaces for a wide variety of applications and media. There have been a large number of optical interfaces deployed in the market. Some of these have been supported by the IEEE, some by MSA's and some are proprietary. Each of these has their own value proposition. As networks begin to transition to higher speeds, switches will likely need to interface at 40GbE and 100GbE as well as 400GbE. One way to accomplish this, is to provide switches with a mix of lower speed and higher speed ports or provide additional network elements to bridge between them. This approach might be appropriate in some cases, but it tends to add additional elements to a network, reducing efficiency.

Another way to provide connectivity to these optical interfaces is through breakouts. This was common for 40GbE initially as previously mentioned, but with the large variety of optical interfaces on the market, it may not always be feasible. This is where backward compatibility can really make a difference. For many interfaces, such as 40GBASE-LR4 and 100GBASE-LR4, CWDM4, do not have a direct path to optically aggregate to higher data rates as 10GBASE-SR can be aggregated to 40GBASE-SR4 through breakouts. Since the QSFP-DD ports can be configured to QSFP28 or even QSFP+, a customer has the option to reuse lower speed optics when necessary to connect to another piece of equipment, while using adjacent ports to connect to a 400GbE port. A switch with all QSFP-DD ports could be configured to accommodate nearly any module or cable available in QSFP+ or QSFP28. As requirements change, the switch can be reconfigured to meet the needs of the network.

It also means the QSFP-DD form factor on its own does not need to address the market with wide a portfolio of optical interfaces as a totally new form factor might without this multi-generational backward compatibility. Customers can leverage the prior investments in optics where it makes sense and focus new investments on upgrading the network and not patching portfolio gaps for a new form factor with another network element or port adaptors.

Advanced thermal design

Extensive modeling, integrated heatsink configurations and measurements have been performed by the MSA members to validate the specifications. Snapshots of those efforts can be seen in Figure 6. In one configuration, the QSFP-DD has been designed to take advantage of front to back airflow for cooling the module. Here the module can take advantage of cooler air from the aisle drawn across the cage's integrated heatsink. Air cooling in this manner means that the air is not preheated by other components on a line card before it is able to remove heat from the module heat sink. The QSFP-DD specification 2.0 has made provisions for power dissipation greater than 14W. This should be enough to manage the heat from any QSFP-DD or any other QSFP form factor module.

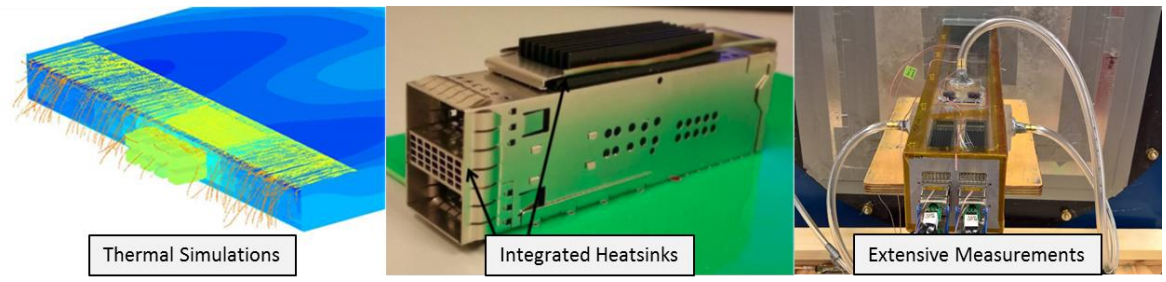

Figure 6: Extensive thermal modeling, design and testing has been performed on the QSFP-DD to validate thermal performance

Conclusion

The QSFP-DD MSA is currently supported by 52 member companies from cage, connector and cable manufactures to module and system vendors. Basing the design on the QSFP form factor, QSFP-DD will leverage the capability and cost structure of the de facto industry standard for 40GbE and 100GbE. The MSA has issued the initial release of the form factor specification in Sept '16 (available on the MSA website[iv]) and issued the 2.0 release in March '17. The backward compatibility is unique in the industry

and provides customers and designers unparalleled flexibility in meeting the demands of the market. The MSA to is open to all companies that expressed interest to participate. Over 200 individuals, reviewed and contributed to the specification making this a true industry collaboration.

For more information, please go to www.QSFP-DD.com

[i] Cisco Visual Networking Index Predicts Near-Tripling of IP Traffic by 2020, https://newsroom.cisco.com/press-release-content?type=webcontent&articleId=1771211

[ii] Cisco Global Cloud Index: Forecast and Methodology, 2015–2020, White Paper, http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/global-cloud-index-gci/white-paper-c11-738085.pdf

[iii] LightCounting Ethernet Transceivers Forecast, Sept 2016

[iv] www.QSFP-DD.com