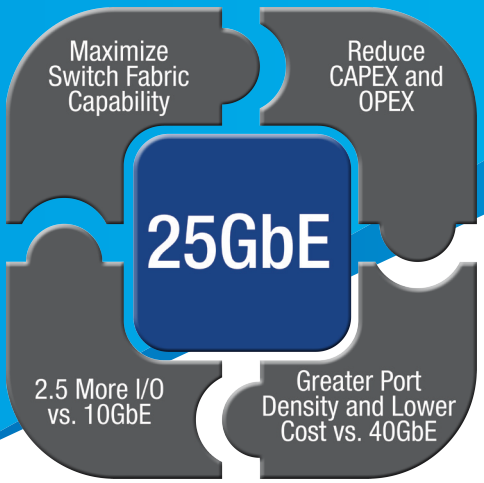
25Gb Ethernet

Accelerated Network Performance and Lower Costs for Enterprise Data Center and Cloud Environments



Driven by the bandwidth requirements of private and public cloud data centers and communication service providers (telco), 25Gbps Ethernet over a single lane will have a significant impact on server interconnect interfaces. It will help reduce capital expenditures (CAPEX) and operational expenditures (OPEX) while meeting the necessary I/O bandwidth requirements in data centers. With the introduction of Cavium™ FastLinQ® 41000/45000 Intelligent Ethernet Adapters, Cavium leads the market by delivering the industry's most comprehensive host and storage connectivity solution, offering support for more protocols and features than any other technology available in the market supporting the emerging 25GbE standard.

EXECUTIVE SUMMARY

Enterprise organizations continue to push the envelope as they deploy the latest technologies in order to keep up with exponential data growth and global operations. The availability of Intel® "Grantley" processors delivers a welcome boost to server performance. However, leading cloud providers are clamoring for even more network performance in order to meet the needs of their web-scale data centers and cloud-based services.

Many organizations currently deploy Top of Rack (ToR) architectures that utilize 10GbE. To keep up with the required network bandwidth, they would currently need to deploy twice as many 10GbE switches, along with additional cables, space, power, and cooling. To help address network performance needs, leading manufacturers explored the existing Institute of Electrical and Electronics Engineers (IEEE) 100GbE standard, which consists of four lanes of 25Gbps electrical signaling.

This resulted in the creation of the 25 Gigabit Ethernet Consortium, with goals that include "enabling the transmission of Ethernet frames at 25 or 50Gb per second (Gbps) and to promote the standardization and improvement of the interfaces for applicable products."

Given the growing demand for faster network performance and maintaining Ethernet economics, in July 2014, IEEE unanimously agreed to support the development of 25GbE standards for servers and switching. The upcoming IEEE 802.3by 25GbE standard is technically complete and expected to be ratified by June of 2016.

Utilizing 25GbE results in a single-lane connection similar to existing 10GbE technology—but it delivers 2.5 times more data. Compared to 40GbE solutions, 25GbE technology provides superior switch port density by requiring just one lane (vs. four with 40GbE), along with lower costs and power requirements. The 25GbE specification will enable network bandwidth to be cost-effectively scaled in support of next-generation server and storage solutions residing in cloud and web-scale data center environments. With the announcement of Cavium FastLinQ 41000/45000 Series adapters, Cavium leads the market by delivering the industry's most comprehensive host and storage connectivity solution, and offers support for more protocols and features than any other technology available in the market supporting the emerging 25GbE standard. Other products using 25GbE technology are also arriving in the marketthroughout 2016.

25GbE – AN EMERGING STANDARD

25GbE is a proposed standard for Ethernet connectivity that will benefit cloud and enterprise data center environments. In June 2014, the 25 Gigabit Ethernet Consortium was formed to promote the technology, and subsequently the IEEE P802.3by 25Gbps Ethernet Task Force was formed to develop the standard. In addition, the IEEE P802.3bq 40GBASE-T Task Force adopted objectives to also develop BASE-T support for 25GbE. Cavium is a member of the Consortium and engineers from Cavium have participated in the IEEE P802.3by 25Gbps Ethernet Task Force.

25GbE leverages technology defined for 100GbE implemented as four 25Gbps lanes (IEEE 802.3bj) running on four fiber or copper pairs. Quad small form-factor and 100Gb form-factor pluggable transceiver (QSFP28/CFP) modules have four lasers, each transmitting at 25Gbps. Each lane requires a Serializer/Deserializer (SerDes) chipset. The twisted pair standard was derived from 40GbE standards development. The following table provides a summary of key upcoming IEEE standard interfaces that specify 25GbE.

Table 1. IEEE 802.3 Standard Interfaces that Specify 25GbE

Physical Layer	Name	Error Correction
MMF Optics	25GBASE-SR	RS-FEC
Direct Attach Copper	25GBASE-CR	BASE-R FEC or RS-FEC
Direct Attach Copper	25GBASE-CR-S	BASE-R FEC or disabled
Electrical Backplane	25GBASE-KR	BASE-R FEC or RS-FEC
Electrical Backplane	25GBASE-KR-S	BASE-R FEC or disabled
Twisted Pair	25GBASE-T	N/A

The IEEE standard specifies two backplane and copper interfaces. These have different goals, hence the different interface. The –S short reach interfaces aim to support high-quality cables without Forward Error Correction (FEC) to minimize latency. Full reach interfaces aim to support the lowest possible cable or backplane cost and the longest possible reach, which do require the use of FEC. FEC options include BASE-R FEC (also referred to as Fire Code) and RS-FEC (also referred to as Reed-Solomon). RS-FEC has been used for a range of applications including data storage satellite transmissions. BASE-R FEC is a newer technology that is particularly well suited for correction of the burst errors typical in a backplane channel resulting from error propagation in the receive equalizer.

INTRODUCING THE CAVIUM FASTLINQ 41000/45000 SERIES CONTROLLERS AND ADAPTERS

The FastLinQ 41000/45000 Series adapters offer an extremely expansive set of features and protocols, delivering the following key benefits:

- Improved Server Utilization:
  - Hardware offload protocol processing, which significantly reduces the CPU burden and improves overall server efficiency. This includes:
    - Broad multi-protocol Remote Direct Memory Access (RDMA) interoperability, with CPU offload for RDMA over Converged Ethernet (RoCE), RoCEv2, and Internet wide area RDMA protocol (iWARP), delivering the first truly universal RDMA NIC (RNIC).
    - Hardened storage protocol interoperability, with CPU offload for iSCSI, Fibre Channel over Ethernet (FCoE), iSCSI Extensions for RDMA (iSER,) and Non-Volatile Memory Express (NVMe) over Fabrics, and provides a Converged Network Adapter (CNA) solution that supports not only traditional Storage Area Network (SAN) connectivity but also newer scale-out storage solutions that are based on object and file.
- Increased Scalability for Virtual Servers and Networks:
  - Advanced NIC virtualization capabilities with support for multi-queue, NIC partitioning (NPAR), and Single Root I/O Virtualization (SR-IOV). This enables faster performance in virtualized environments and tenant consolidation, which requires the use of fewer physical ports while maintaining quality of service (QoS) and service-level agreements (SLAs).
  - Support for Data Plane Development Kit (DPDK), which addresses network function virtualization (NFV) to enable and accelerate virtual network function (VNF) applications and provide flexible deployment of server nodes in data center carrier networks.
  - Extensive overlay networking (tunneling) capabilities with stateless offload support for network virtualization using Virtual Extensible Local Area Network (VXLAN), Network Virtualization using Generic Routing Encapsulation (NVGRE), Generic Routing Encapsulation (GRE), and Generic Network Virtualization Encapsulation (GENEVE). These capabilities optimize performance while reducing the cost of network encapsulation for virtualized workloads and hybrid cloud deployments.
- Efficient Administration:
  - Optimization for software-defined networking (SDN) and OpenStack deployments, which leverages NPAR, SR-IOV, tunneling offloads, DPDK support, and broad storage protocol integration.
  - Integration between the expansive set of features and protocols in FastLinQ 41000/45000 Series adapters and the Cavium QConvergeConsole® (QCC), which simplifies deployment, diagnostics and management from a single-pane-of-glass, streamlining storage and networking I/O connectivity orchestration.

KEY 25Gb ETHERNET BENEFITS

The 25GbE specification enables network bandwidth to be cost-effectively scaled in support of next-generation server and storage solutions residing in cloud and web-scale data center environments.

Key benefits include the following:

- Maximum switch I/O performance and fabric capability
  - 2.5 times more data vs. 10GbE
  - 4 times the switch port density vs. 40GbE (one lane vs. four lanes)
- Reduced CAPEX
  - Fewer ToR switches and fewer cables
  - Lower cost vs. 40GbE
- Reduced OPEX
  - Lower power, cooling, and smaller footprint requirements
- Quick maturation by leveraging the existing IEEE 100GbE standard

SIGNIFICANT COST BENEFITS

The proposed 25GbE standard delivers 2.5 times more data compared to existing 10GbE solutions. It also provides greater port density and a lower cost per unit of bandwidth by fully utilizing switch port capabilities when compared to 40GbE solutions.

The specification adopted by the consortium utilizes a single-lane 25GbE and dual-lane 50GbE link protocol. To help facilitate adoption, the new specification is available royalty-free by the consortium members to any data center ecosystem vendor or consumer who joins the consortium.

As shown in Figure 1, a recent five-year forecast by industry analyst Crehan Research projects that annual shipments of 25GbE ports will be 2.5 times greater than 40GbE ports by 2018.

TECHNICAL OVERVIEW – WHAT IS 25GbE?

Clock Rate

Within an Ethernet NIC or switch, all of the high-speed components are connected using a serial component called a SerDes, which takes data to be transferred and serializes it, and then the deserializer on the receiving side reconstructs the serial stream of bits into data for the ultimate receiver. SerDes technology over the years has progressed to the latest 25Ghz speed. The components that comprise current 10GbE switches run 12.5Ghz SerDes with a clock rate of 10.3125GHz. However, the current 40GbE NICs and switches use four parallel SerDes links with a clock rate of 10.3125GHz each. In contrast, the components that comprise the proposed 25GbE NICs and switches use a single lane SerDes with a clock rate of 25.78125GHz. (See Table 2.)

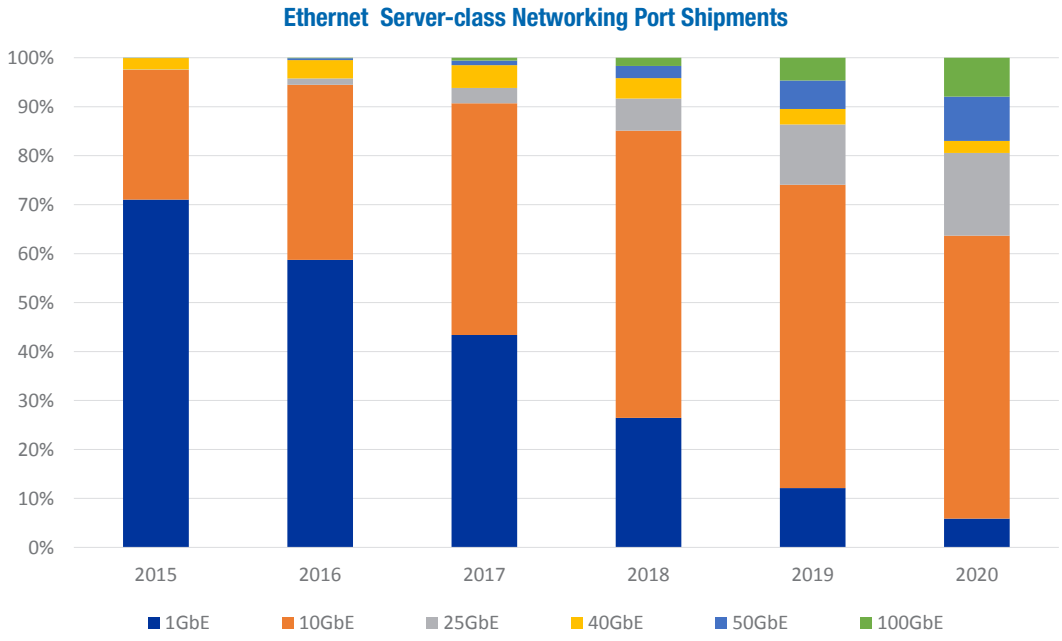


Figure 1. Server-class Adapter and LOM Controller Long Range Forecast (source: Crehan Research, January 2016)

Table 2. Clock Rates, Lanes, and Performance

Ethernet	Clock Rate	Lanes	Data Rate
1GbE	1.25Ghz	1	1Gbps
10GbE	10.31Ghz	1	10Gbps
25GbE	25.78Ghz	1	25Gbps
40GbE	10.31Ghz	4	40Gbps
100GbE	25.78Ghz	4	100Gbps

Number of Lanes

In a typical 10GbE/40GbE ToR Ethernet switch, the actual Ethernet ports are SerDes connections coming from the switching chip pins. These connections are then used to connect directly to the SFP+ and QSFP+ optics cages or other Ethernet or fabric chips (for blade servers). Communication between an SFP+/QSFP+ in the front of the switch and the switching chip runs on top of one of these SerDes connections. The number of SerDes connections required to drive a switch port are called "lanes."

Table 3. Optimize Switch ASIC I/O Bandwidth
Example: Ethernet switch with 3.2 Tbps capacity and 128 ports

Port Speed	Lane Speed (Gb/s)	Lanes Per Port	Usable Ports	Total BW (Gb/s)
10GbE	10	1	128	1280
25GbE	25	1	128	3200
40GbE	10	4	32	1280
100GbE	25	4	32	3200

The components used in today's switches all run SerDes with a clock rate around 10Ghz, providing a 10Gb transfer rate between each component (allowing for the encoding overhead). In the last few years, SerDes technology has advanced to the point that 25Ghz SerDes has become economically viable, and all of the various physics-related challenges in signal integrity have found reliable solutions.

The 40GbE interface is constructed from four parallel SerDes links between the Ethernet chip and the QSFP+ pluggable module. Extending QSFP+ onto fiber continues to require four parallel 10Gb streams to transport this to the receiving QSFP+ (Figure 2), i.e., parallel optics. Short-reach QSFP+ interfaces use four pairs of fiber between them. Long-reach QSFP+ interfaces utilize Wave Division Multiplexing (WDM) to transport the four 10Gb streams over a single pair of fiber. The requirement of four lanes significantly reduces switch port density per switching chip and increases the cost of cabling and optics.

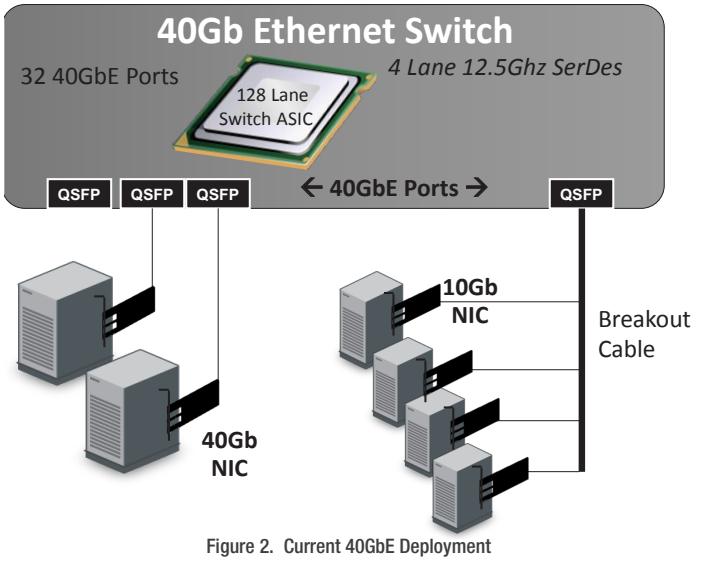


Figure 2. Current 40GbE Deployment

The proposed 25GbE standard leverages the availability of a 25Ghz SerDes and requires only a single SerDes lane, while delivering 2.5 times more throughput compared to current 10GbE solutions and significant CAPEX savings compared to 40GbE solutions (Figure 3). In addition, some blade server chassis solutions today are limited to only two SerDes lanes for their LAN on Motherboard (LOM) networking ports and therefore cannot implement a four-lane 40Gbps interface.

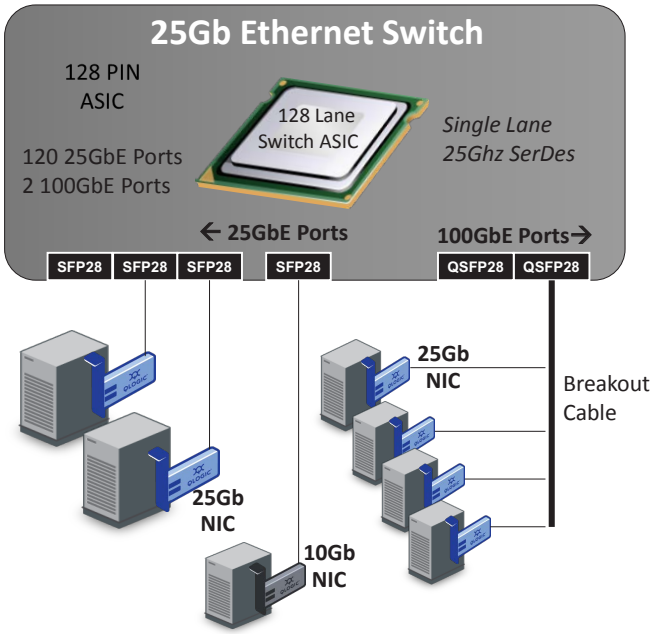


Figure 3. Proposed 25GbE Deployment Maximizes Bandwidth/Pin

Error Correction Code Options

Beginning with 10GbE and 10Gb Fibre Channel Inter-Switch Links (ISLs), the "64b/66b" encoding scheme is utilized to improve data transfer efficiency. The 64b/66b encoding results in a 3% overhead (66-64)/66 on the raw bit rate. To compensate, Clause 74 (Fire Code) FEC was introduced as an option with 10GbE to provide additional error protection. For 100Gb backplane, Clause 91 (Reed-Solomon) FEC was introduced to provide additional error protection. Clause 74 FEC and Clause 91 FEC are both supported by the 25GbE specification. Auto-negotiation can be used to determine whether Clause 74 FEC, Clause 91 FEC, or no FEC is employed on the link.

Auto-negotiation

The details of auto-negotiation capabilities are not fully settled or implemented. The proposed 25GbE and 50GbE solutions will be backward and forward compatible with 10GbE, 40GbE, 100GbE, and 400GbE products since they use the same IEEE 802.3 frame format. However, the capabilities of switch ports to automatically link at different speeds is a work in progress. For example, current switch support will allow Cavium 25GbE adapters to auto-negotiate link parameters at 25GbE. Cavium 100GbE adapters will auto-negotiate link parameters at 100GbE, 50GbE, 40GbE, and 25GbE.

Form Factors

The 25GbE physical interface specification supports a variety of form factors, as shown in the following table:

Table 4. 25GbE Interface Form Factors

Form Factor	Lanes and Speed
QSFP28	4 x 25 Gbps
SFP28	1 x 25 Gbps

Switches that are currently available do not support direct 25GbE connections using an SFP28 direct attach copper (DAC) cable. One of the recommended solutions is to use a breakout cable that allows four 25GbE ports to connect to a 100GbE QFSFP28 switch port. DAC cable lengths are limited to three meters for 25GbE. Active optic cable (AOC) solutions can also be used and support longer lengths.

PCI Express (PCIe) Interfaces

PCIe 3.0 is ubiquitous across shipping server platforms. The trend in cloud and web-scale server deployments is towards single-port Ethernet connectivity due to cost. These volume servers typically have PCIe 3.0 x4 slots. As outlined in Table 5, 25GbE is an easier upgrade path from 10GbE as it fits into the existing model and requires half the number of PCIe lanes compared to 40GbE, leading to better PCIe bandwidth utilization and lower power impact.

Table 5. PCIe 3.0 Lanes required for Ethernet Generations

Ethernet	Single Port	Dual Port
10GbE	2	4
25GbE	4	8
40GbE	8	16
100GbE	16	32

STANDARDIZATION ACTIVITY

Broad Industry Support

The new 25GbE specification is championed by the 25 Gigabit Ethernet Consortium, which consists of leading companies including Arista Networks™, Broadcom®, Brocade®, Cisco®, Google®, Cavium, and Microsoft®. The goal of the consortium is to support an industry-standard, interoperable Ethernet specification that boosts performance and slashes interconnect costs per Gb between the server NIC and ToR switch.

IEEE Standardization Status and Projected Completion

In July 2014, IEEE held a "call for interest" meeting and members unanimously agreed to support the development of a 25GbE standard for servers and switching. The IEEE P802.3by 25Gbps Ethernet Task Force was formed to develop the standard. In addition, the IEEE P802.3bq 40GBASE-T Task Force adopted objectives to also develop BASE-T support for 25GbE. Both of these task forces have completed the technical work, with projects in the Sponsor Ballot phase of the standards process. This fast progress was due to the high leverage of the existing standard for 100GbE as the base standard, and the 40Gbps twisted pair development for the 25GBASE-T specifications. Ratification of the IEEE 802.3by standard is expected to be completed by June 2016, with the IEEE 802.3bq standard on track for September 2016 ratification.

Interoperability Testing

Cavium joined with other members of the Ethernet Alliance at the University of New Hampshire's Interoperability Lab (UNH-IOL), an industry interoperability-testing event held in June of 2015. The UNH-IOL has a long history of providing interoperability and conformance testing of data networking and storage networking products. A wide range of pre-standard 25GbE equipment and cabling showed an unprecedented level of maturity.

Cable and Optics

According to the 25 Gigabit Ethernet Consortium, the 25Gb and 50Gb channels shall conform to all of the channel characteristics defined in IEEE standard 802.3bj, Clause 92 "Physical Medium Dependent (PMD) sublayer and baseband medium, type 100BASE-CR4," specifically those defined in 92.9. There are a number of connector and cable combinations that can meet these requirements. The 25Gbps and 50Gbps PMDs define support operation on low-cost, twin-axial copper cables, requiring only two twin-axial cable pairs for 25Gbps operation and only four twin-axial cable pairs for 50Gbps operation. Links based on copper twin-axial cables can be used to connect servers to ToR switches, and as intra-rack connections between switches and/or routers.

Unlike 10GbE, where the IEEE standard did not formally recognize the use of twin-axial cables, the cables are well specified for the 25GbE standard, including cable test specifications. This promises to improve interoperability of these popular, low cost cables.

VALUE PROPOSITION

Web-scale data centers and cloud-based services need greater than 10GbE connectivity from servers, while maintaining the cost-sensitive economics often associated with Ethernet deployments. Industry leaders and standards organizations recognize this need and have formed the 25GbE Consortium to target cloud-scale networks with the goal of standardizing 25GbE technology that uses a single 25Gbps lane, so that it can be widely developed and productized as quickly as possible. In contrast, there are no 40Gbps single-lane server-to-switch standardization efforts under way.

The IEEE has a new 50GbE Study Group and is initiating work on a serial (single-lane) server-to-switch standard. The P802.3cd effort leverages work underway in the IEEE P802.3bs 400Gbps Ethernet Task Force. The task force is just beginning the project and completion is projected for late 2018, with products likely to be introduced in a related time frame. Similar to the 25Gb Ethernet Consortium specifying a two-lane 50Gbps mode, this new IEEE project will also standardize a two-lane 100Gbps mode and a four-lane 200Gbps mode. While these standards and the resulting projects will provide a roadmap for the data center as we approach the close of the decade, 25GbE and 50GbE products will provide the most economical options for scaling beyond 10GbE.

The 25GbE specification extends the IEEE 802.3 standard to include operation at 25Gbps and 50Gbps over copper cables and backplanes. The capability to interconnect devices at these rates is important in next-generation data center networks that need to increase server network throughput beyond 10Gbps and 20Gbps without using more interconnect lanes. By providing a 25Gbps MAC rate, which leverages single-lane 25Gbps physical layer technology developed to support 100GbE, 25GbE maximizes server efficiency to access switch interconnects and provides an opportunity for optimum cost/performance benefits.

25GbE will provide up to 2.5 times faster performance than existing 10GbE connections while maximizing the Ethernet controller bandwidth/pin and switch fabric capability. This can deliver more than a 50% savings in rack interconnect cost per unit of bandwidth, and significantly improve an operator's bottom line. It will also increase network scale and accommodate higher server density within a rack than what is currently achievable with 40GbE ToR links. In short order, deploying 25GbE in a data center should quickly approach the cost of similar 10GbE solutions, something that solutions built on the more complex 40GbE standard can never achieve.

Table 6. Ethernet Comparison

Characteristics/ Requirement	10GbE	25GbE	40GbE	100GbE
PCIe 3.0 Lanes per Port	2	4	8	16
PCIe 3.0 Bandwidth Utilization	62.5%	78%	62.5%	78%
Clock Rate	10.31Ghz	25.78Ghz	10.31Ghz	25.78Ghz
SerDes Lanes	Single	Single	Quad	Quad
Servers/ToR with 3:1 Oversubscription	96	96	24	Future
Connector	SFP+	SFP28	QSFP	QSFP28
DAC Cabling	Thin 4-wire	Thin 4-wire	Bulkier 16-wire	Bulkier 16-wire
Cable Material Cost	Low	Low	High	High
Simpler Transition to 100GbE	-	Yes	No	-

TRUSTED SOLUTIONS

Cavium is a global leader and technology innovator in high-performance server and storage networking connectivity and application acceleration solutions. The company's leadership in product design and maturity of software stack make it the top choice of leading OEMs, including Cisco, Dell®, EMC®, Hitachi Data Systems, HPE®, IBM®, Lenovo®, NetApp®, and Oracle®, as well as channel partners worldwide for their virtualized, converged, and cloud environment solutions.

Cavium offers complete solutions to some of the most complex issues facing the data center. For more information, visit the Cavium Website at www.Cavium.com.

LEARN MORE

Cavium Overview of 2550100 – Fast and Smart

Cavium Technical Marketing presents an overview of new 25GbE, 50GbE and 100GbE technology, an industry go-to-market roadmap, and demonstrates the advantages of real 25GbE products in a live data center.

View Video Now

ABOUT CAVIUM

Cavium, Inc. (NASDAQ: CAVM), offers a broad portfolio of infrastructure solutions for compute, security, storage, switching, connectivity and baseband processing. Cavium's highly integrated multi-core SoC products deliver software compatible solutions across low to high performance points enabling secure and intelligent functionality in Enterprise, Data Center and Service Provider Equipment. Cavium processors and solutions are supported by an extensive ecosystem of operating systems, tools, application stacks, hardware reference designs and other products. Cavium is headquartered in San Jose, CA with design centers in California, Massachusetts, India, Israel, China and Taiwan.

CAVIUM™

Follow us:

Corporate Headquarters Cavium, Inc. 2315 N. First Street San Jose, CA 95131 408-943-7100

International Offices UK | Ireland | Germany | France | India | Japan | China | Hong Kong | Singapore | Taiwan | Israel