SOLUTION BRIEF

PCIe 4.0 has Arrived to Feed the Appetite of Data-Intensive Applications

Executive Summary

Artificial Intelligence, Virtual Machines, containerization, and 5G mobile wireless networks are key drivers for next-generation high-performance systems. However, current servers with PCIe Express (PCIe) 3.0 require wide busses to keep up with the latest Ethernet or InfiniBand speeds or with the performance demands of new NVMe solid-state drives (SSD). For example, the bandwidth of an 8-lane PCIe 3.0 interface supports a single 40 Gigabit Ethernet connection but creates a bottleneck with dual ports and at greater speeds. Likewise, PCIe 3.0 is already seen as a speed limitation for SSDs. A faster solution is required since increasing lane width with 3.0 is not efficient in terms of cost, complexity, higher power, complex circuit board layout, and component fanout.

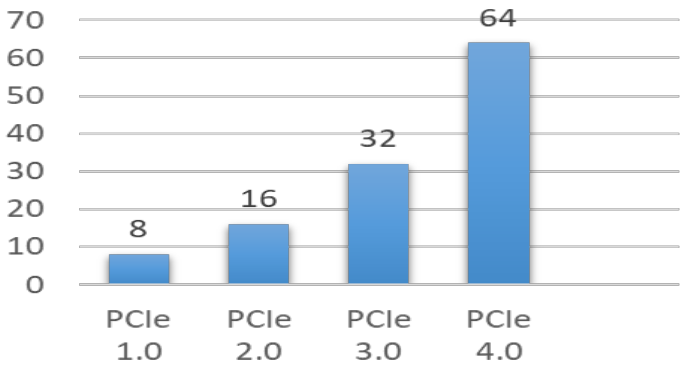
To meet the requirements of data-intensive applications, PCIe 4.0 doubles the bandwidth of servers, creating a superhighway to increase a server's data handling capacity to 64 GBytes/s. This will dramatically enhance data access capabilities so that servers will be able to analyze more data, more effectively for real-time insights, and access shared data quicker to fully utilize NVMe drives. Other advantages include a smooth transition for 4K and 8K video enabling four times sharper video footage and a platform that can manage the traffic that will be created by new 5G mobile networks.

Key Takeaways:

- With PCIe 4.0, performance is dramatically increased to 16 GT/s per lane and total bandwidth increased by up to 64 Gbytes
- High-speed interfaces such as 100G/200G Ethernet & InfiniBand and NVMe solid-state drives are require greater bandwidth
- SmartNICs and FPGAs are expected to be the first to be deployed on PCIe Gen 4 starting in 2019
- Lenovo ThinkSystem servers with AMD EPYC 7002 processors are among the first servers available with support for PCIe Gen 4
- Intelligent and high performance Mellanox interconnects exploit the advantages of PCIe Gen 4

PCI Express Generation 4.0

It's been almost seven years since PCIe 3.0 was released, and the industry is finally moving on to the next high-speed interconnect version, PCIe 4.0. PCIe 4.0 doubles the PCIe 3.0 I/O performance and is 6 times faster than the original PCIe 1.0. It enables connection speeds of up to 251 GBytes/s while using 16 lanes concurrently, providing flexibility, scalability, and lower power consumption improvements due to the simultaneous enhancements of advanced integrated circuit design and manufacturing technology. This new specification maintains the same 128b / 130b encoding scheme of PCIe 3.0 and remains backward compatible with PCIe 1.0. With no interface changes, PCIe 4.0 enables an easy migration path and maintains existing behaviors of PCI 3.0 and lower for a seamless integration.



PCIe Total Bandwidth in GB per Sec.

In term of performance, throughput is 16 gigatransfers per second (GT/s) per lane. This equates to 251 GBytes/s of unidirectional bandwidth for a PCIe 4.0 x16 slot. This sort of speed will help ensure future network demand will efficiently keep pace with the ever-increasing quantity of data.

Theoretical vs Actual Bandwidth

With all this said, it still leaves some confusion on what the actual bandwidth of a PCIe 4.0 interconnect can achieve and why it's needed. To explain this, first, let's discuss the overhead associated with the PCIe specification. The overhead is related to the encoding process within the transmission procedure. Due to this overhead, some of the theoretical maximum throughput, listed as gigatransfers per second, will be lost (see *PCI Express Standards and Associated Speeds* chart below). PCIe 1 and 2 utilize an 8b/10b encoding, meaning that for every 10 bits transmitted, 2 bits (or 20 percent) of the theoretical bandwidth is lost. With PCIe 3.0 (and above) the specification shifted to a more efficient encoding scheme of 128b/130b, so the overhead is much less - only 1.54 percent.

The maximum possible PCIe bandwidth can be calculated by multiplying the PCIe width and speed. From that number, we reduce ~1 GBytes/s for error correction protocols and the PCIe header overhead. The overhead is determined by both the PCIe encoding, and the PCIe MTU. The formula is listed below:

Maximum PCIe Bandwidth = SPEED * WIDTH * (1 - ENCODING) – 1 GBytes/s.

After overhead, the maximum per-lane data rate of PCIe 1.0 is eighty percent of 2.5 GT/s or 2 GBytes/s or 250 MBytes/s per lane.

PCI Express Standards and Associated Speeds

Standards Version	Year Introduced	Raw Bit Rate	Throughput per lane each direction	Actual Bandwidth (x16)
PCIe 1.0	2003	2.5 GT/s	250 MBytes/s	2 GBytes/s (4 GBytes/s)
PCIe 2.0	2007	5.0 GT/s	500 MBytes/s	64 GBytes/s (8 GBytes/s)
PCIe 3.0	2010	8.0 GT/s	985 MBytes/s	126 GBytes/s (15.75 GBytes/s)
PCIe 4.0	2017	16.0 GT/s	1.96 GBytes/s	251 GBytes/s (31.51 GBytes/s)

All bandwidth are unidirectional. Bi-directional bandwidth would be doubled.

Let’s take a look at a x16 PCIe Gen 4 device:

Maximum PCIe Bandwidth = 16G * 16 * (1 - 2/130) - 1G = 256G * 0.985 - 1G = ~251GBytes/s.

Why is Gen 4.0 Needed?

CPUs and GPUs are manipulating ever-increasing data sets. Flash storage, and NVMe drives are far faster than spinning media from the past therefore, media and entertainment companies must support higher-definition content and interconnect speeds are quickly moving towards 200 GBytes/s. All this means an improved local I/O mechanism is needed to keep system latency low and to prevent bandwidth bottlenecks. PCIe 4.0 provides faster speeds to handle the greater bandwidth that new and more powerful component and applications demand.

To explain this further, let's look at a 40 GBytes/s Ethernet adapter to understand why Gen 4 is needed. The industry standardizes on PCIe 3.0 x8 for a standard 40 Gb adapter. The max bandwidth of a single PCIe 3.0 lane is 985 MBytes/s (or 7.88 Gb/s). Multiply that by 8 to get the max bandwidth of a x8 slot and a PCIe 3.0 x8 slot can provide 63.04 Gb/s, plenty to handle a single port adapter. However, there is not enough bandwidth for a dual-port adapter. Another example is the huge growth in M.2 NVMe SSDs that utilize PCI Express connectivity. The x4 M.2 NVMe SSDs using PCIe 3.0 peak at 3.94 GBytes/s, a number which will double to 7.88 GBytes/s with PCIe 4.0. It's easy to see that most data center will see a real benefit from PCIe 4.0. With Ethernet moving from 100 Gb/s speeds, which are currently available, to 200 Gb/s at the end of this year the additional bandwidth of PCIe 4.0 arrives just in time to stay ahead of the curve as a x16 slot can provide enough bandwidth for a 200 Gb/s adapter.

Adoption of Gen 4

Expansion of the Internet, ubiquitous smartphone usage, increasing acceleration of AI solutions and the Internet of Things (IoT) will increase the need for fast and more efficient data center environments. As well as expanding and emerging markets like streaming 4k/8K video and 5G mobile networks, higher speed interfaces such as 100 Gb Ethernet, 200 Gb InfiniBand, and NVMe solid-state drives are providing larger pipes. For servers to keep pace, they must increase their data handling capacity. PCIe 4.0 does just this, more throughput to prevent bottlenecks in the rise of these applications and interfaces.

Several PCIe 4.0 servers are available now while they will begin flooding the market in 2020. Data intensive application will require high- performance computing products such as SmartNICs, FPGAs and NVMe SSDs, which are expected to be the first to be deployed on PCIe 4.0 systems and will occur this year. 4K and 8K video and gaming will continue to drive adoption in 2020. And in the meantime, the ecosystem for Gen 4 support will continue to evolve to support a wide range of vertical markets.

Lenovo Among the First With Support

Lenovo ThinkSystem’s are among the first generally available x86 servers to enable the use of PCIe Gen 4 and are equipped with an AMD EPYC 7002 single-socket processor. Designed to capture the improved I/O capabilities of PCIe Gen 4, the ThinkSystem SR635 is offered in a 1U form factor while the SR655 is 2U. Both servers maximize availability by providing eight PCIe Gen 4 slots that enable high throughput to handle the data driven from the 128 PCIe lanes enabled by the AMD EPYC 7002 CPU. You can use the 128 lanes of PCIe Gen 4 connectivity to tie together HPC clusters or satisfy insatiable needs for east-west bandwidth of virtualized traffic.

Connecting at higher speeds allows greater acceleration from GPU clusters or to expand NVMe drive connectivity. The ThinkSystem servers can take full advantage of the bandwidth by loading the 1U SR635 system with up to sixteen 2.5” NVMe drives without oversubscription or 32 NVMe drives in the SR655 2U system (up to 28 without over-subscription). Lenovo customers can take full advantage of PCIe Gen 4 to boost performance and server efficiency. Enabling data centers to scale in a cost-effective way and adapt to constantly changing requirements.

Maximize I/O

Delivering these specialized high-end systems often requires partnership with technology experts who can provide the underlying high-performance building blocks. Mellanox delivers these building blocks through an end-to-end interconnect networking portfolio of high-performance 25G, 50G, and 100G Ethernet and InfiniBand adapters, cables and switches. By the end of 2019, Mellanox will be shipping each with up to 200 Gb speeds. These offerings are well proven within the largest hyperscale and telecom data centers who have adopted Mellanox due to our intelligent, high-performance and efficient components which help them achieve total infrastructure efficiency.

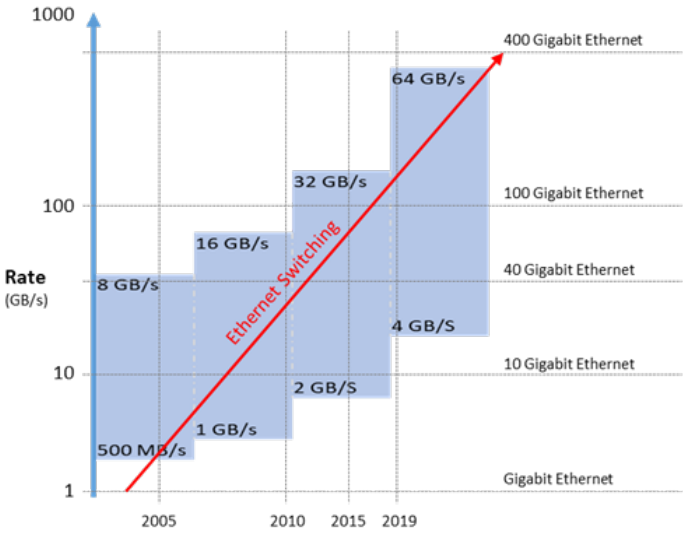
Data-intensive applications require the movement of large volumes of data causing the CPU to devote much of their processing time to I/O and the manipulation of data. Mellanox ConnectX family of network adapters exploit the advantages of PCIe Gen 4 to allow for high-bandwidth. The ConnectX also combines networking offloads and accelerators to remove the tasks of moving the huge volumes of data without burdening the CPU. With multiple times the performance packed into the same infrastructure footprint, end-users deploying Mellanox on Gen 4 systems can quickly reap the benefits of the new specification to satisfy the most demanding applications and remove fears of becoming I/O-bound.

Furthermore, to maximize system performance Mellanox LinkX cables have been tested and show a very low Bit Error Rate (BER) of less than 1E-15, significantly better than other competitors. Cables with poor signal quality can impact high-speed data transmissions therefore adversely affect data. throughput and slowing down network performance, minimizing the benefits of PCIe Gen 4 servers.

Mellanox tests every cable to ensure they can support the speed and data-carrying capabilities required for high bandwidth networks.

Aligning Bus and Network Bandwidth

An important element for scaling PCIe 4.0 must include network and with Ethernet switching speeds currently at 100 Gb and approaching 200 Gb, the speed of the PCIe bus much match or surpass that of the network. The chart below shows Ethernet switching slowing passing that of the internal PCIe bus by the end of PCIe 3.0's life. A shift to PCIe 4.0 puts the two back in sync, removing any potential bottlenecks.



PCIe 4.0 Aligns Compute Bus and Network Bandwidth

Equally important, switches must deliver high bandwidth and consistent performance. Mellanox Spectrum Ethernet switches do this in several ways. First, by utilizing a fully-shared buffer to provide packet burst absorption capabilities to avoid dropping packets. Also, by employing cut-through switching, Spectrum switches forward packets faster than store-and-forward switches, delivering lower latency and improved consistency in performance.

Additionally, to regulate traffic at a granular flow, Spectrum utilizes an intelligent congestion management feature to ensure data flows freely at the highest speeds and with unprecedented levels of performance.

The Network is a critical part of the infrastructure that determines the overall system performance. The interconnecting fabric is the glue that holds the

entire system together to reliably transport data packets. Standardizing on Mellanox ensures that PCIe Gen 4 systems achieve their maximum potential and paves the way for faster workloads

Conclusion

In today's era of data-centric computing, to truly deliver superior performance, data center servers must be designed differently. They need advanced I/O buses with enhanced bandwidth and latency capabilities to deliver higher capacity for data-intensive workloads. PCIe Gen 4 prepares a data center to handle the new demands that specialized applications need and provides a foundation to support high-performance and efficient network connectivity to support a wide range of link speeds and evolving network acceleration engines such as SmartNICs. In turn, SmartNICs will complement the performance boost by offering acceleration engines for specific functions like networking, security, and storage to alleviate the CPU from the burden of I/O tasks.

Want To Learn More?

Link to Lenovo servers that are mentioned:
https://www.lenovo.com/us/products/new severs

For detailed information on Mellanox OEM-specific products for Lenovo visit:
http://www.mellanox.com/oem/lenovo

Mellanox end-to-end Ethernet connectivity:
http://www.mellanox.com/ethernet-storage-fabric/

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403

Version 7.25.19