White Paper

Intel® Server Adapters

New Trends Make 10 Gigabit Ethernet the Data-Center Performance Choice

Technologies are rapidly evolving to meet data-center needs and enterprise demands for efficiently handling and managing increasingly bandwidth-hungry applications. These technologies include the advent of multi-core servers, server consolidation and virtualization, high-density computing, and networked storage over Ethernet. This white paper discusses those evolving technologies, their impact on input/output (I/O) performance needs, and how 10 Gigabit Ethernet is evolving to support those technologies.

Table of Contents

Why 10 Gigabit Ethernet? Why Now?

Since ratification of the first 10 Gigabit Ethernet (10GbE) standard in 2002, 10GbE usage grew primarily in niche markets demanding the highest bandwidth available. This established 10GbE as a stable, standards-based connectivity technology and the next evolutionary stage above Gigabit Ethernet (GbE). However, connectivity needs today extend beyond just providing a higher-bandwidth pipe.

Certainly, bandwidth and I/O throughput are important aspects of data-center connectivity needs today. Beyond that, though, connectivity products today — whether they are GbE or 10GbE — must support the performance features of the evolving data-center technologies. This includes:

- Scaling on multi-core processor-based servers
- Supporting the I/O arbitration needs of multiple virtual machines (VMs) in sever consolidation and virtualization
- Providing economical, energy-efficient, high-performance connectivity for server blades and high-density computing
- Internet Small Computer System Interface (iSCSI) support for storage-over-Ethernet applications

Second-generation Intel® 10 Gigabit Ethernet products for PCI Express* (PCIe) introduce a variety of new technologies to support evolving data-center technologies. For the data center, this means more energy-efficient, cost-effective delivery of high-bandwidth capacity for growing application needs. The attractiveness of 10GbE performance is further enhanced by the trend toward lower prices in 10GbE switches, server adapters, and related infrastructure items. The 2006 ratification of the 10GBase-T standard for 10GbE on Category-6 or better twisted-pair copper wire should provide further price reductions for 10GbE capability.

Multi-Core Platforms Drive I/O Capacity Needs

Multi-core processors, starting with the Dual-Core Intel® Xeon® processor, are the answer to Moore's Law running into the brick wall of physical reality. Specifically, traditional approaches of increasing processor performance by moving to higher clock rates eventually ran into power consumption and expensive cooling issues. Multi-processor systems provided an interim solution, and they are still viable for some applications. However, multi-core processors have the advantages of a smaller footprint with greater performance per watt. This makes multi-core processors ideal for traditional server architectures and particularly well suited for blades and other high-density computing needs where space and cooling are at a premium.

Servers based on Intel multi-core processors allow data centers to grow compute power without growing space or cooling requirements. Replacing older single-core servers with multi-core servers can deliver three to five times the compute power within the same hardware footprint. Even greater efficiencies can be achieved by consolidating applications onto fewer, more powerful servers.[1]

The multiplied capabilities and efficiency of multi-core processor-based systems also raises the demand for I/O capacity. While the multi-core server does provide ample headroom for consolidating multiple applications onto the server, the aggregation of application I/O traffic can easily require the additional bandwidth of 10GbE connectivity for optimum network performance.

However, additional connectivity bandwidth alone is not the complete answer to improved throughput. Potentially significant bottlenecks exist throughout the various server I/O processes. Intel® I/O Acceleration Technology (Intel® I/OAT) is designed specifically to address system-wide bottlenecks.

Intel I/OAT is a suite of features that enables efficient data movement across the platform – network adapters, chipset and processors – thereby improving overall system performance by improving CPU utilization and lowering latency. The different features include Intel® QuickData Technology, Direct Cache Access (DCA), Message Signaled Interrupts-Extended (MSI-X), low latency interrupts and Receive Side Coalescing (RSC).

Intel QuickData Technology enables data copy by the chipset instead of the CPU, and DCA enables the CPU to pre-fetch data, thereby avoiding cache misses and improving application response times. MSI-X helps in load-balancing interrupts across multiple MSI vectors, and low latency interrupts automatically tune the interrupt interval times depending on the latency sensitivity of the data. RSC provides light-weight coalescing of receive packets that increases the efficiency of the host network stack.

Intel I/OAT is a standard feature on all Intel network connections for PCIe and on Dual-Core and Quad-Core Intel® Xeon® processor-based servers. It accelerates TCP/IP processes, delivers data-movement efficiencies across the entire server platform, and minimizes system overhead.

In addition to supporting Intel I/OAT, all Intel 10 Gigabit Server Adapters for PCIe are tuned to optimize throughput with multi-core processor platforms. These new networking features increase performance by distributing Ethernet workloads across the available CPU cores in the system. These server adapter features include:

- MSI-X – distributes network controller interrupts to multiple CPUs and cores. By spreading out interrupts, the system responds to networking interrupts more efficiently, resulting in better CPU utilization and application performance.
- Multiple Tx/Rx queues – is a hardware feature that segments network traffic into multiple streams that are then assigned to different CPUs and cores in the system. This allows the system to process the traffic in parallel for improved overall system throughput and utilization.
- Receive Side Scaling – called Scalable I/O in Linux,* distributes network packets to multiple CPUs and cores. This improves system performance by using software to direct packets to the appropriate CPU based on IP, TCP, and port address.
- Low Latency – allows the server adapter to run a variety of protocols while meeting the needs of the vast majority of applications in high-performance computing (HPC) clusters and grid computing. Intel has lowered Ethernet latency with adaptive and flexible interrupt moderation and by streamlining different operating system (OS) stacks.

Consolidation and Virtualization Need Still More I/O

Data centers are migrating in greater numbers to server consolidation and virtualization in order to reduce server proliferation and provide greater management efficiencies and quality of service. Typically, the first step in the process consists of consolidating different instances of like applications onto a single server or fewer servers than used in the old single-application-per-server paradigm.

More typical today, however, is the trend of including virtualization. Virtualization allows mixed applications and OSs to be supported on a single sever by defining multiple virtual machines (VMs) on the server. Each VM on a server operates in essence like a standalone, physical machine, but because the VMs are under the auspices of a single server, IT gains the advantages of a reduced server inventory, better server utilization, data-center consolidation, and more-efficient centralized management of resources.

Figure 1 illustrates the basic concept of a virtualized server. The VMM software defines and manages each VM and can define additional VMs as necessary to handle application load increases. As can be imagined, the overhead of running a VMM and multiple VMs requires a high-performance server and the better the performance, the more that can be virtualized. Multi-core Intel Xeon processor-based servers are uniquely suited to these needs because their performance far exceeds that of previous server generations. More than that, however, they also include Intel® Virtualization Technology (Intel® VT), which reduces the need for compute-intensive software translations between the guest and host OSs.[2] This allows consolidation of more applications on fewer physical servers.

Multiple VMs mean multiple I/O streams, the aggregate of which increases the I/O bandwidth and throughput needs for each physical machine. Use of a 10GbE network interface card (NIC) or a dual-port 10GbE server adapter provides maximum available connectivity bandwidth for virtualized environments.

Additional assists from Intel I/OAT and optimization for multi-core servers provide further performance gains from Intel 10 Gigabit Server Adapters. However, there is still another special assist for virtualized environments. This assist is Virtual Machine Data Queues, or VMDq.

VMDq is a networking hardware feature on Intel Server Adapters that provides acceleration by assigning packets to various virtual machines (VMs) in a virtualized server. Received packets are sorted into queues for the appropriate VM and are then handed up to the virtual machine monitor (VMM) switch, thereby reducing the number of memory copies the system needs to make to get packets to VMs. VMDq also handles transmission of packets from the various VMs on the host server to ensure timely and fair delivery to the network. This reduces the significant I/O penalty created by overhead associated with the added layer of VMM software for sharing ports and switching data between multiple VMs. In brief, VMDq provides acceleration with multiple queues that sort and group packets for greater server and VM I/O efficiency.

Figure 2 illustrates the key elements of VMDq. The virtualized server is at the top of the diagram and has several VMs with an intervening virtualization software layer that includes a Layer 2 software switch. The server adapter with VMDq provides the connection between the virtualized server and the LAN.

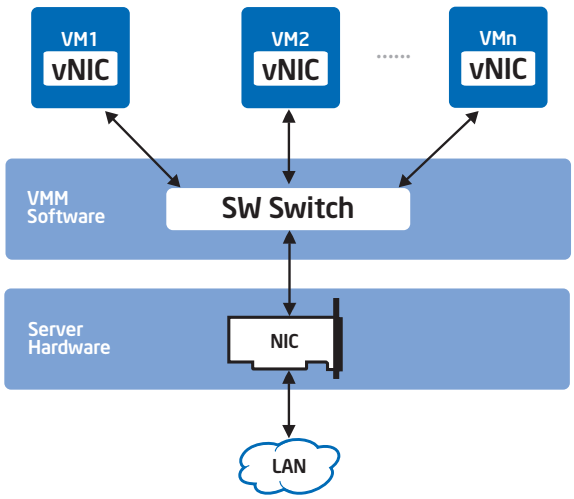

Figure 1. Virtualized server. Multiple VMs can be defined on a single server, with each VM running different applications under different operating systems.

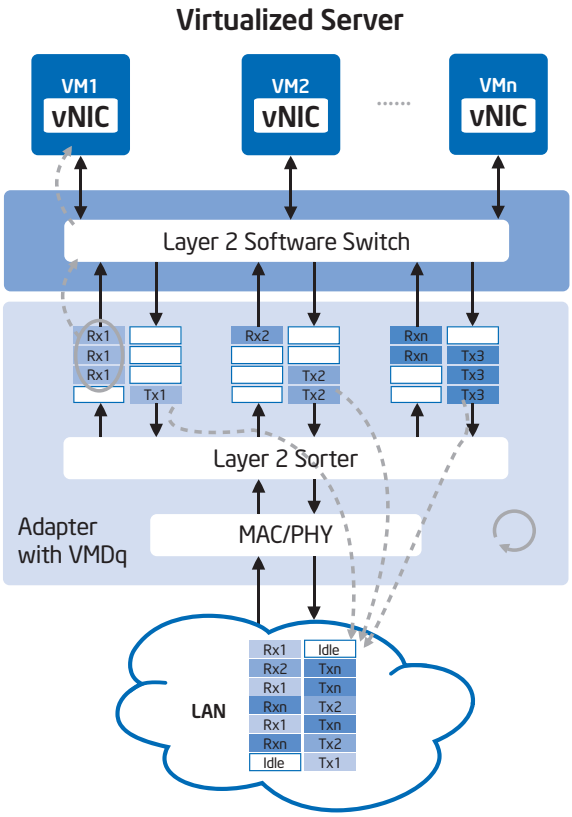

Figure 2. VMDq and Intel® 10 Gigabit Ethernet Server Adapters. VMDq provides more efficient I/O in virtualized servers.

On the receive side, VMDq sorts I/O packets into queues (Rx1, Rx2,... Rxn) for the destination VMs and sends them in groups to the Layer 2 software switch. This reduces the number of decisions and data copies required of the software switch.

On the transmit side (Tx1, Tx2,... Txn), VMDq provides round-robin servicing of the transmit queue. This ensures transmit fairness and prevents head-of-line blocking. The overall result for both the receive-side and transmit-side enhancements is a reduction in CPU usage and better I/O performance in a virtualized environment.

High-Density Computing Environments

The performance multiples offered by multi-core processors — along with their lower energy and cooling requirements — make them an increasingly popular choice for use in high-density computing environments. Such environments include high-performance computing (HPC), grid computing, and blade computer systems. There is also a newly emerging breed referred to as dense computing, which consists of a rack-mounted platform with multiple motherboards sharing resources.

While their architectures differ, high-density computing environments have many commonalities. They are almost always power, cooling, and space critical, which is why multi-core processors are taking over in this environment. Additionally, their I/O requirements are quite high today. This is typified by the blade system shown in Figure 3.

Typical blade systems in the past used Fibre Channel* for storage connectivity and GbE for network connectivity. However, blade systems today are moving to Quad-Core Intel Xeon processor-based blades with 10GbE connectivity, as shown in Figure 3. This is supported by the emergence of universal or virtual fabrics and

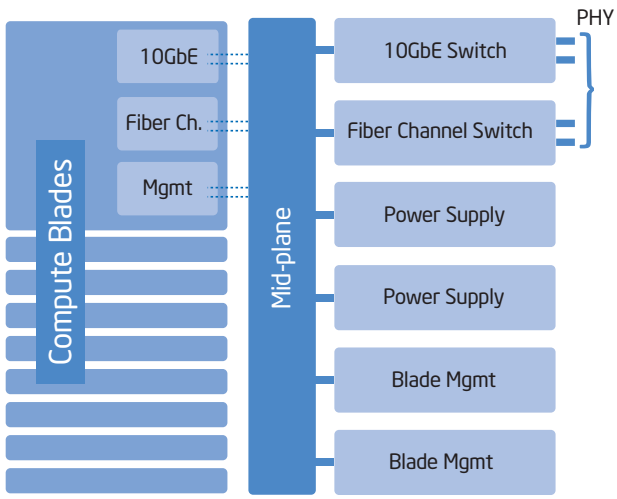

Figure 3. Typical blade system architecture. Recent advances have moved 10GbE connectivity to the blade level and reduced per-port costs by consolidating expensive PHY-level devices at the switch port.

the migration of 10GbE from the blade chassis to the blade architecture. Moreover, instead of costing thousands of dollars per port, 10GbE in blades is now on the order of a few hundred dollars per port.

A key reason for the cost reduction of 10GbE in blades is the fully integrated 10GbE Media Access Control (MAC) and XAUI ports[3] on the new Intel® 82598 10 Gigabit Ethernet Controller. When used as LAN on motherboard (LOM) for blades, the integrated MAC and XAUI ports allow direct 10GbE connectivity to the blade system mid-plane without use of an expensive Physical (PHY) layer device. As a result, PHY devices can be pushed out of the blades and consolidated at the switch ports, as shown in Figure 3. Since PHY devices, especially for fiber connectivity, constitute as much or more than half the cost of NICs, the switch-level PHY consolidation and sharing indicated in Figure 3 results in significant reductions in 10GbE cost per port for blade systems. Such significant performance increases and cost reductions, especially with the advent of 10GbE in twisted pair, will promote 10GbE connectivity throughout the data center.

The Emergence of Storage over Ethernet

So far, discussion has focused primarily on compute platform and I/O performance as driving the need for 10GbE connectivity. Storage is another related area that can benefit from the bandwidth benefits and falling prices of 10GbE.

Within the network and data center, there are three traditional types of storage. These are direct-attached storage (DAS), network-attached storage (NAS), and the storage-area network (SAN). Each has its distinct differences and advantages; however, SAN is emerging as being the most advantageous interms of scalability and flexibility for use in data centers and high-density computing applications.

The main drawback to SAN implementation in the past has been equipment expense and the specially trained staff necessary for installing and maintaining the Fibre Channel (FC) fabric used for SANs. Nonetheless, there has been sufficient demand for the storage benefits of SANs for Fibre Channel to become well established in that niche by virtue of its high bandwidth.

Although 10GbE has been around for several years, it is now poised to take a position as an alternative fabric for SAN applications. This was made possible by the Internet Small Computer System Interface (iSCSI) standard. The iSCSI standard

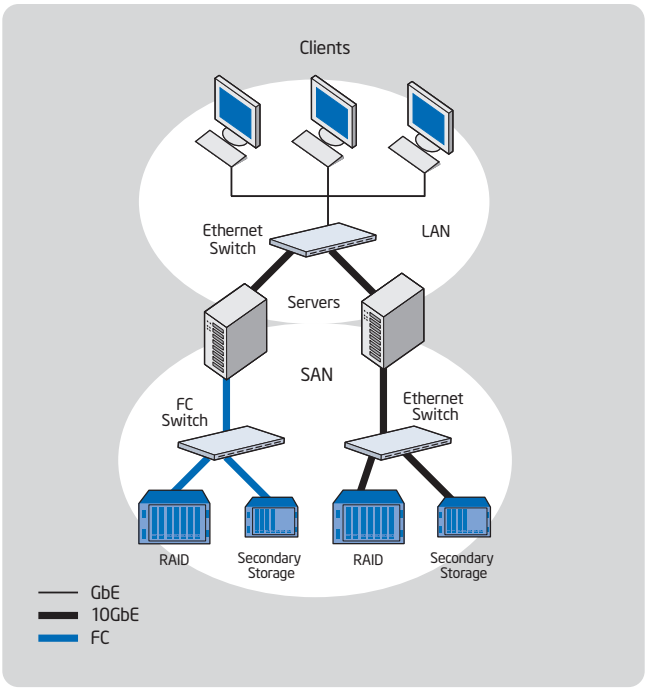

Figure 4. 10GbE in the SAN environment. 10GbE can coexist with Fibre Channel (FC) and is a flexible means of expanding an existing SAN.

is an extension of the SCSI protocol for block transfers used in most storage devices and also used by Fibre Channel. The Internet extension defines protocols for extending block transfer protocol over IP, allowing standard Ethernet infrastructure elements to be used as a SAN fabric.

Basic iSCSI capability is implemented through native iSCSI initiators provided in most OSs today. This allows any Ethernet NIC to be used as a SAN interface device. However, lack of a remote-boot capability left such implementations lacking in the full capabilities provided by Fibre Channel fabrics. Initially, iSCSI host bus adapters (HBAs) offered a solution, but they were expensive specialty items much like Fibre Channel adapters.

To resolve the remote-boot issue, Intel provides iSCSI Remote Boot support with all Intel PCIe server adapters, including the new generation of Intel 10 Gigabit Ethernet Server Adapters. This allows use of the greater bandwidth of 10GbE in new SAN implementations. Additionally, Ethernet and Fibre Channel can coexist on the same network. This enables use of the greater performance and economy of 10GbE in expanding legacy Fibre Channel SANs as shown in Figure 4.

In addition to the bandwidth advantage over Fibre Channel, 10GbE adapters with iSCSI Remote Boot offer a variety of other advantages in SAN applications. These include:

- **Reduced Equipment and Management Costs –** GbE and 10GbE networking components are less expensive than highly specialized Fibre Channel components and, because they are Ethernet, they do not require the specialized Fibre Channel skill set for installation and management.
- **Enhanced Server Management –** Remote boot eliminates booting each server from its own direct-attached disk. Instead, servers can boot from an OS image on the SAN. This is particularly advantageous for using diskless servers in rack-mount or blade server applications as well as for provisioning servers and VMs in server consolidation and virtualization. Additionally, booting servers from a single OS image on the SAN ensures that each server has the same OS with the same patches and upgrades.
- **Improved Disaster Recovery –** All information on a local SAN – including boot information, OS images, applications, and data – can be duplicated on a remote SAN for quick and complete disaster recovery. Remote boot and an iSCSI SAN provides even greater assurance of disaster protection and recovery. This is because iSCSI SANs can be located anywhere Internet connectivity is available, allowing greater geographic separation and protection from local or regional disasters such as earthquakes and hurricanes.

Figure 4, which shows the coexistence of FC and iSCSI SANS, hints at a pattern of emerging convergence into a unified fabric. The April 2007 announcement by Intel and other industry leading vendors of support for a Fiber Channel over Ethernet (FCoE) standard carries this pattern of convergence further toward a unified fabric. FCoE provides a second option for connecting servers to existing FC SANs as shown in Figure 5. The advantage would be convergence of storage and LAN traffic onto a single, low-cost multifunctional NIC. Additionally, volume rack servers and blades could be more readily connected to existing FC SANs through a common Ethernet fabric.

Future Directions: Convergence on Ethernet

Ethernet is the ubiquitous network fabric. However, in the past its bandwidth limitations kept it from being the fabric of choice for some application areas, especially storage applications and interprocess communication (IPC). Consequently, various other fabrics emerged to meet high-bandwidth, low-latency, no-drop needs. Fibre Channel and InfiniBand are the most notable of these high-performance fabrics.

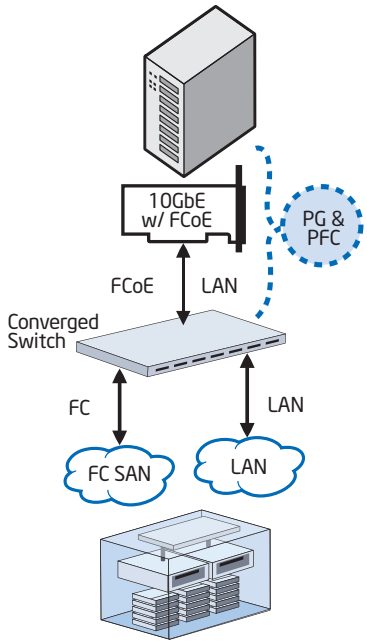

Figure 5. Converged LAN and SAN traffic. FCoE would allow migration toward a unified data-center fabric with lower cost and easier manageability.

The inclusion of iSCSI initiators in popular operating systems opened the way for Ethernet to take a role as a storage fabric. Development of iSCSI Remote Boot opened the path further, allowing Ethernet server adapters to provide price, distance, and simplicity advantages as a SAN fabric. Now, lowering prices and increased performance features in 10GbE server adapters are providing bandwidth and cost/performance advantages over Fibre Channel.

This extension of Ethernet, especially 10 Gigabit Ethernet, into the storage domain is the beginning of a trend toward I/O convergence on Ethernet. Ethernet is a logical choice for this convergence since it is already so prevalent and well known. I/O convergence on Ethernet would provide numerous benefits including lowering the cost of infrastructure and management, standardization of interconnects, and a single highly flexible fabric from server backplanes to the outer reaches of the network.

For I/O convergence to occur, Ethernet still needs some end-to-end quality-of-service enhancements. These include:

- No dropped packets for storage (would eliminate need for iSCSI)
- Lower latency on par with what InfiniBand now provides
- Virtualized network links
- SAN remote boot

Additionally, server adapters and switches will need enhancements that include:

- Priority groups
- End-to-end congestion management
- Flow control
- Bandwidth reservation

Intel is already developing technologies and products that are moving toward the concept of I/O convergence and is driving IEEE* standards to ensure multi-vendor support. When I/O convergence on Ethernet becomes a reality, multiple traffic types (LAN, storage, and IPC) can be consolidated onto one easy to use and truly ubiquitous network fabric.

Conclusion

The significant cost-per-port reduction for 10GbE – projected to decline as much as 41 percent from 2005 to 2007 – along with the significant performance gains and energy efficiencies of multi-core processors in blades and other platforms portends a significant rise in 10GbE connectivity throughout data centers and networks. This is especially true for server virtualization and high-density computing environments where I/O bandwidth and throughput are critical. Additionally, with the emergence of iSCSI initiators in OSs and iSCSI Remote Boot support in Intel® Server Adapters for PCIe, 10GbE is well poised for an expanding role in SAN applications. The future role of 10GbE and Ethernet in I/O convergence promises to provide a lower-cost network infrastructure that is both more agile and responsive in meeting business needs.

In short, 10GbE deployment in data centers is on a growth path driven by needs for higher performance and supported by falling per-port prices for 10GbE capability. 10GbE is the here-and-now upgrade and migration path to the high-performance networks of the future. To ensure a smooth and cost-effective migration path, Intel has created a new generation of 10GbE server adapters with a balanced set of features specially optimized to take advantage of new and emerging technologies.

For More Information

To find out more about,

Intel® 10 Gigabit Server Adapters, visit: www.intel.com/go/10GbE

Intel® I/OAT, visit: www.intel.com/go/ioat

Intel® VT, visit: www.intel.com/go/virtualization

Intel® iSCSI support, visit: www.intel.com/network/connectivity/products/iscsiboot.htm

[1] For more information on server performance, visit: www.intel.com/business/technologies/energy-efficiency.htm and www.intel.com/performance/server/xeon/intthru.htm

[2] Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, Virtual Machine Monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

[3] In XAUI, the X is Roman numeral X for 10, AUI stands for attachment unit interface. XAUI is a standard for connecting 10GbE ports to each other and to other electronic devices.

*Other names and brands may be claimed as the property of others.

Printed in USA 0907/BY/OCG/HOP/300 Please Recycle 318466-001US