Sun microsystems

SUN TECHNOLOGY

UNLEASHING 10 GIGABIT NETWORKS

White Paper
March 2007

Table of Contents

Chapter 1
Introduction

Today, every business relies on the corporate computing infrastructure to run the business. As more users access increasing numbers of faster systems to get work done — from running faster and more complex simulations and analyses, to logging retail and banking transactions and more — computing systems are generating, processing and sharing more information and data than ever before. To keep pace with these demands, computing system throughput is doubling approximately every two years, and networks must be able to keep pace in order for the entire computing infrastructure to meet business demand.

Faster Networks for Increasing Computing Demands

Over two decades ago, Ethernet emerged as the network technology of choice. Running at 10 Mbps, Ethernet cemented itself as the foundation for networking through its use of inexpensive wiring and easy integration into computing platforms. As system throughput and the adoption of client-server technology advanced, so did the need for faster networking capabilities. A direct extension of the 10Base-T Ethernet standard, Fast Ethernet increased network speeds to 100 Mbps, yet took advantage of the installed base of wiring already employed for traditional Ethernet networks, making the migration to faster networking technology easy.

System and networking capabilities continued to mature and advance, and businesses found they could use the computing infrastructure to get more work done. As a result, companies moved more tasks to the computing infrastructure. Systems were asked to perform a greater volume of work — and many different types of work — than ever before. With more sophisticated applications enabling global corporate collaboration, new and innovative network-intensive software placed further demands on the computing infrastructure. Just in time, Gigabit Ethernet technology emerged and raised networking speeds in support of these new demands.

Consolidation and Virtualization — Throttling Network Needs

Today, the nature of computing continues to change in fundamental ways. The explosive growth of corporate intranets and the Internet — including the need for more bandwidth, larger networks, and innovative digital devices — is creating new and challenging demands. As the number of users, systems and devices accessing services over the network grows, organizations need to find ways to provide more scalable and available compute power in the most efficient way possible, with users taking for granted that IT services will deliver the performance and predictability needed. However, as systems are replicated throughout the IT infrastructure to give them greater resiliency and throughput, the result is often a sprawling, complex network of systems that are costly and difficult to manage.

Today, *server consolidation* — bringing together applications, databases and servers onto fewer highly reliable hardware platforms — is taking center stage. Indeed, consolidation is essential if organizations are to streamline the datacenter and refocus IT infrastructures to support business priorities. Why is consolidation so important? It helps companies make more efficient use of resources with fewer business interruptions, thereby lowering total cost of ownership (TCO) and reducing costs.

Building on the foundation of consolidation technology, virtualization techniques are now becoming popular. *Virtualization* enables computing resources to be pooled and shared on an as-needed basis, further increasing resource utilization. By helping create a less complex IT environment based on a consistent architecture, consolidation and virtualization strategies help companies significantly improve efficiencies, increase resource utilization, simplify maintenance, and foster increased availability and manageability.

Parallelism — Driving Demand for Network Bandwidth

The consolidation of compute capacity, combined with processor and application parallel programming advancements, often results in a single system doing the work of tens or hundreds of machines. For example, chip multithreading technology (CMT) is an innovative processor design that provides massive amounts of thread-level parallelism and increased application throughput by supplying multiple cores per processor, and multiple threads per core. Combined with the software multithreading capabilities of many operating systems and applications, CMT is taking parallelism to a new level. Systems can run more applications simultaneously, operating systems and applications can send more work to more processors or cores at any given moment, and processors can execute more tasks in parallel — all leading to higher system and application performance and faster results.

With the use of hardware and software parallelism techniques on the rise, systems and networks are working harder. Network-centric applications are not only asking systems to make more computations faster, they are also making requests of local and remote resources, and moving data across local, regional and global enterprise networks. As a result, each system now has the network bandwidth requirements of many machines. Indeed, with more applications and users creating greater network traffic per system, network bandwidth requirements for each machine are increasing by a factor of 10 to 100. Unfortunately, supplying greater bandwidth by deploying large numbers of network interfaces in a single server is often impractical, or fails to fulfill demand.

10 Gigabit Ethernet Technology

While available as a technology for quite some time, 10 Gigabit Ethernet is finally gaining a foothold. Based on the 802.3ae and 802.3an standards, 10 Gigabit Ethernet enables organizations to scale existing networks and take advantage of 10 times the bandwidth of Gigabit Ethernet while leveraging the existing cabling infrastructure. This fast networking technology can be of the most value today when the networking requirements of a given system exceed the bandwidth that Gigabit Ethernet can provide, when hot spots develop in the network fabric, or traffic patterns become overly complex. Even though some aspects of the computing infrastructure have supported 10 Gigabit Ethernet for a few years, what has been missing is the technology to make it affordable and accessible to computing systems.

Sun's 10 Gigabit Ethernet Networking Technology

Sun's 10 Gigabit Ethernet Networking Technology is one of the industry's first network interfaces specifically designed to accelerate multithreaded application performance by optimizing I/O throughput in environments that utilize parallel threads. Providing a shared, multihomed, 10 Gigabit Ethernet interface, Sun's silicon innovation represents a revolutionary multithreaded approach to connecting systems to the network.

Figure 1-1. Sun's ASIC optimizes I/O throughput for threaded environments

The first realization of Sun's 10 Gigabit Ethernet Networking Technology is an ASIC that includes four Ethernet ports configured as two quad speed (10/100/1000/10000 Mbps) and two triple speed (10/100/1000 Mbps) Ethernet ports. With advanced processing functions optimized for throughput computing and network system architectures, including packet classification for load balancing, Sun's 10 Gigabit Ethernet Networking Technology provides a cost-effective, high-performance interface. Support for fine-grained virtualization enables network I/O to be virtualized through interface sharing and partitioning. Sun's 10 Gigabit Ethernet Networking Technology enables systems that employ it to:

- Take advantage of more network bandwidth
- Improve utilization of compute resources
- Minimize system I/O latency
- Leverage existing wiring infrastructure

With the introduction of its 10 Gigabit Ethernet networking technology, Sun is now the only vendor that can deliver a complete multicore, multithreaded computing environment optimized for network throughput — starting with the Solaris™ Operating System (OS), extending to the UltraSPARC® T1 processor, and running across the network interface. While Sun's new multithreaded 10 Gigabit Ethernet technology is optimized for CMT environments, it improves network performance for all processors. Now the tight integration between the Solaris OS, CMT processors, and the new 10 Gigabit Ethernet ASIC enables Sun to deliver an integrated, optimized network environment with unmatched throughput.

Sun and High-Speed Networking

For over 20 years, Sun has been instrumental in developing state-of-the-art networking technology and systems. Sun knows that networking is integral to computing and fundamental to keeping a business running at peak efficiency. Sun is constantly looking to, and shaping, the future of computing by investing in new technology. Significant improvements in high-performance network computing technologies, like CMT processors and Sun's 10 Gigabit Ethernet Networking Technology, and systems which incorporate them, ensure Sun customers will always have access to leading-edge technology that can help businesses maximize computing resources and create more effective enterprises.

Chapter 2

Sun Technology — Fueling the Network

Sun's 10 Gigabit Ethernet Networking Technology, and products which incorporate or take advantage of its capabilities, provide innovation that helps maximize network throughput.

Sun's 10 Gigabit Ethernet Networking Technology

Sun's 10 Gigabit Ethernet Networking Technology is an interface specifically designed to optimize I/O throughput and scale to support multicore and multithreaded environments. Providing a shared, multihomed, 10 Gigabit Ethernet interface, Sun's ASIC includes multiple ports and advanced processing and virtualization features that help accelerate application performance by optimizing I/O throughput in environments that utilize parallel threads. The ability of the ASIC to deliver high performance and enable I/O virtualization is made possible through an architecture that supports line rate packet classification and multiple DMA engines, and provides a wide variety of interrupt handling schemes and several other performance optimizations (Figure 2-1).

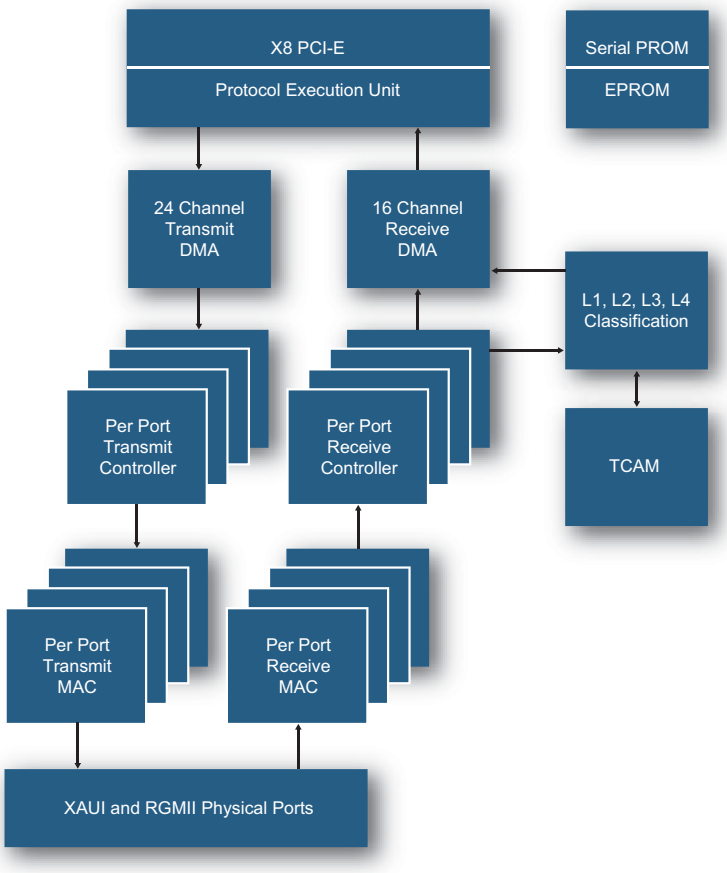

Figure 2-1. Functional block diagram

Line Rate Packet Classification

Until recently, the hardware to software interface was based on queuing packets into the system, and then distributing and processing those packets. To improve load balancing and performance, Sun's ASIC transforms the interface by moving to a distribute then queue model. Capable of handling over 30 million packets per second, the ASIC statically separates receive traffic into multiple logical queues that can be worked on in parallel. Packets can be classified based on packet classes, ternary matches, or hash functions.

Packet classification extends parallelism to I/O by enabling packets to be separated so that packets that do not depend on one another can be worked on simultaneously. Aggregate network workload is separated into many smaller increments — up to 16-way using Sun's technology — enabling multiple TCP streams to be allocated to different processor threads in parallel (Figure 2-2). Such a strategy spreads the network load among processing units, and frees CPUs and ports for other tasks.

Figure 2-2. Packet classification in Sun's 10 Gigabit Ethernet Networking Technology enables network packets to be worked on in parallel

Multiple DMA Channels and CPU and Thread Affinity

Sun's 10 Gigabit Ethernet Networking Technology employs multiple direct memory access (DMA) channels — 24 transmit and 16 receive channels — to help speed throughput. In addition, Sun's ASIC enables a one-to-one correlation of receive and transmit packets across the same TCP connection, and binding flexibility between DMA channels and ports. These capabilities keep caches warm, and can help avoid cross-calls and context switching to deliver greater performance while reducing the need for CPU resources to support I/O processing.

Virtualization Support

Support for fine-grained virtualization enables I/O to be virtualized through interface sharing, partitioning, and network virtualization. Resources can be partitioned logically into as many as eight groups. Virtualization can be based on the ingress port, virtual LAN (VLAN), MAC address, or service address. Up to 4,096 VLANs can be attached to a given port, up to 16 MAC addresses can be assigned to a 10 Gigabit Ethernet port, and up to 256 IP addresses or TCP/UDP ports can be associated with a service address. With these capabilities, Sun's ASIC can consolidate network attachments by aggregating attachment points to save switch ports and server I/O slots. In addition, fine-grained virtualization helps limit the compute resources assigned to a given connection. Such a scheme proves invaluable in the event of accidental or malicious high volume traffic, such as a Denial of Service attack, by ensuring the offending actions affect only the resources assigned to the virtual port or address and not the entire system.

Multiple Speed Ethernet Ports

Sun's ASIC includes up to four Ethernet ports configured as two quad speed (10/100/1000/10000 Mbps) and two triple speed (10/100/1000 Mbps) Ethernet ports. With four ports, the ASIC minimizes the number of hops through I/O bridges, thereby reducing the latency between system memory and the network. In addition, Sun's ASIC provides two XAUI 10 Gbps ports and four Reduced Gigabit Media Independent Interface (RGMII) 1000 Mbps ports, enabling the mixing and matching of ports and speeds based on system needs.

Interrupts

By handling various interrupts, the technology can accelerate performance for any platform and operating system capable of distributing its network load.

PCI-Express 1.1 Support

Designed for use in a wide variety of systems, Sun's 10 Gigabit Ethernet Networking Technology includes support for the PCI-Express 1.1 standard and is capable of operating at x1-, x4-, or x8-lane configurations. Up to 32 outstanding transactions can be handled at any given time. In addition Sun's ASIC supports relaxed memory access via on-chip transaction re-ordering to hide I/O and memory latencies, and improve throughput and efficiency.

Sun Multithreaded Networking Card

With parallel processing and chip multithreading technologies becoming more pervasive, high-performance applications are now more dependent on I/O throughput and memory latency for performance gains. Based on Sun's own ASIC and software for innovative Throughput Networking design, the new Dual 10 Gigabit Ethernet Sun Multithreaded Networking Card can provide the critical bandwidth, load balancing and granular resource allocation required for parallel applications and high-speed networks. Providing 10 Gbps wirespeed performance and significant cost savings compared to multiple Gigabit Ethernet network interface cards (NICs), the Sun Multithreaded Networking Card delivers the high bandwidth needed for application acceleration and I/O virtualization.

The Sun Multithreaded Networking Card is a dual 10 Gigabit Ethernet port x8 PCI Express 1.1 compliant, Fiber XFP MSA compliant, low profile, plug-in adapter (Figure 2-1). Compliant with the IEEE 802.3ae 2002 standard, the Sun Multithreaded Networking Card boosts performance by optimizing throughput in parallel-thread environments. Together, these capabilities enable the card to help deliver 10 Gbps Ethernet throughout the network stack on all single and multicore SPARC, x86, and x64 platforms.

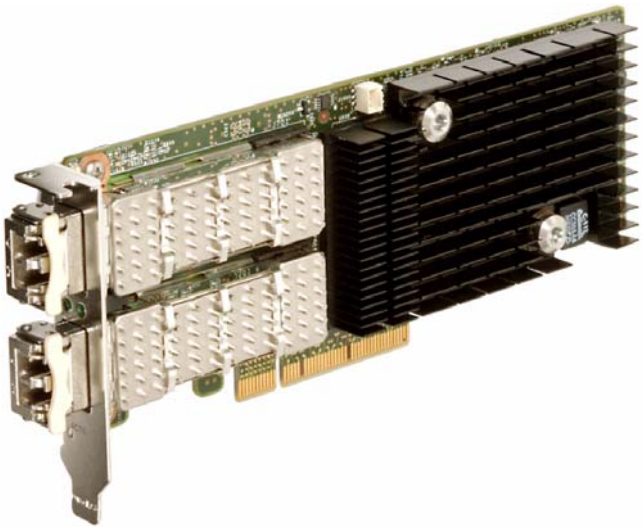

Figure 2-1. Sun's Multithreaded Networking Card brings 10 Gigabit Ethernet technology into the mainstream

- *High performance, low CPU utilization*
  The Sun Multithreaded Networking Card extends CPU and operating system parallelism to networking with its support for hardware-based flow classification and multiple DMAs. Using CPU thread affinity to bind a given flow to a specific CPU thread, it enables a one-to-one correlation of Rx and Tx packets across the same TCP connection. These capabilities can help avoid cross-calls and context switching to deliver greater performance while reducing the need for CPU resources to support I/O processing.

- *Increased speed in local area networks*
  The Sun Multithreaded Networking Card utilizes Sun's own innovative Network Interface Controller to map the 10 Gigabit Ethernet XAUI interface onto the PCI Express form factor. As a result, local area networks using the Sun Multithreaded Networking Card can immediately benefit from increased speed and efficiency.
- *Optimization of resources*
  The Sun Multithreaded Networking Card also makes it easier to optimize and customize resources for applications, services and users. Load balancing and granular resource control features enable incoming packets to be assigned to specific CPU threads or balanced across multiple compute elements based on L1 to L4 packet header information. As a result, the Sun Multithreaded Networking Card offers a high level of flexibility and granularity for assigning resources and can help avoid I/O bottlenecks.

Project Crossbow

One of many industry-leading endeavors undertaken via the OpenSolaris™ Project, Project Crossbow enables the Solaris OS to take full advantage of the new virtualization functionality provided by the Sun Multithreaded Networking Card. The next logical step in the evolution of the Solaris OS networking stack, Project Crossbow brings bandwidth resource control and virtualization into the operating system architecture, instead of adding on layered functionality with heavy overhead and undue complexity.

Project Crossbow provides the building blocks for network virtualization and resource control by virtualizing the network stack and network interface card around any service protocol, such as HTTP, HTTPS, FTP, and NFS, or any virtual machine, including Solaris Containers, Xen, and Logical Domains (LDoms). Each virtual stack can be assigned its own priority and bandwidth on a shared NIC without degrading performance. The architecture dynamically manages priority and bandwidth resources, and can provide a better defense against denial of service attacks directed at a particular service or virtual machine by isolating the impact to that entity. All virtual stacks are separated by a hardware classification engine that ensures traffic for one stack does not impact other virtual stacks.

Some of the features of Project Crossbow include:

- Virtual NIC capabilities that enable policy-based sharing of bandwidth and other resources by multiple Solaris Containers
- Parallelization of network workload across multiple threads and cores
- Capability to provide dedicated bandwidth and resources to separate services and protocols without a performance penalty

More information on Project Crossbow can be found on the OpenSolaris Project Web site located at *http://opensolaris.org*

Sun's Continuing Commitment

Sun understand the value of high-performance network computing, and continues to invest in research and development efforts to create products based on this promising technology. In fact, Sun plans to incorporate its unique 10 Gigabit Ethernet Networking Technology in future processors, motherboards, blades, and switches and across all product lines. Indeed, Sun's upcoming Niagara 2 processor will incorporate 10 Gigabit Ethernet Networking Technology directly on the processor.

Chapter 3

Changing the Face of Network Computing

Sun's 10 Gigabit Ethernet Networking Technology provides the missing link that enables parallelism throughout the entire computing architecture, supports virtualization techniques, and speeds network performance.

Completing the Parallelism Puzzle

As server consolidation and operating system virtualization technologies take hold, the role of the network takes on even greater importance. For years, developers designed applications to run as fast as possible, with any resultant network congestion handled by the network transport layer, if at all. Today, many applications utilize multithreaded programming techniques to speed performance, and rely on similar parallelism capabilities in the operating system to distribute work among processors and cores in the system. Unfortunately, even highly efficient applications and operating systems are limited at some point by the amount of information that can be sent and received across the network interface.

Assuming there are no significant performance bottlenecks in the system, network interface throughput is limited by the speed of the processor and its ability to handle software processing that occurs throughout the entire network stack. Typically, every Gigabit/second of network I/O performance requires 1 GHz of CPU performance. As a result, a 10 Gigabit Ethernet network interface could require a 10 GHz CPU in order to perform well. Systems using Sun's 10 Gigabit Ethernet Networking Technology could require even more — a 40 GHz CPU — in order to sustain line rate traffic in both directions on the network.

While many fast processors exist, none are available on the scale needed to keep data flowing over the network interface. Toward this end, many IT strategies utilize multiple CPUs and aggregate speed in order to handle network processing. Unfortunately, traditional network interface cards cannot scale to handle the throughput demands of multiple processors. As a result, many organizations use brute force parallelism, deploying multiple NICs and switch ports. While this strategy can work well for single threaded environments, it fails to keep pace when multicore systems and multithreaded computing techniques are employed. Indeed, it is impractical to plug 32 or 64 network cards into a system — many systems just do not provide sufficient slot capacity. Even if enough slots are available, deploying many cards can be prohibitively expensive or simply fail to manage bandwidth appropriately.

These challenges are exacerbated when viewed from the application perspective. Indeed, applications are inherently parallel in nature, either via statistical mutliplexing in time or by aggregation of parallel users. For example, Web servers direct hundreds or

thousands of small requests over the network to one or more systems for information, and telecommunications companies support millions of clients with very slow connection rates that are all handled by the same server. As a result, few applications require an extremely high network connection for an extended period of time — often many applications are needed to saturate a high-speed network interface on a sustained basis. Sun's 10 Gigabit Ethernet Networking Technology solves this problem and changes the nature of I/O parallelism by giving systems and applications the ability to share access to very high bandwidth through aggregated attachment points that save switch ports and server I/O slots.

Enabling Network Virtualization

Today, organizations are using server consolidation and virtualization techniques to enable more services and applications to run on a single system. At the same time, it is important to ensure every user and application can access the resources needed — from the operating system and CPU down to the network itself. Being successful requires finding ways to gain control and establish isolation mechanisms in order to improve resource utilization. With a finer granularity of control, enterprises can ensure all workloads have access to an appropriate amount of resources and that no workload consumes any or all resources.

Virtualization with Logical Domains

Virtualization is the technique that makes sharing computing resources possible. Resource management, hardware partitioning, operating system virtualization, and virtual machines comprise the four basic categories of virtualization technology available today. Supported in all Sun servers which utilize UltraSPARC T1 processors, a Logical Domain (LDom) is a full virtual machine that runs an independent operating system instance and contains virtualized CPU, memory, storage, console, and cryptographic devices.

Within the Logical Domains architecture, the hypervisor is a small firmware layer that provides a stable, virtualized machine architecture to which an operating system can be written. As such, each logical domain is completely isolated and the number of virtual machines that can be created relies upon the capabilities of the hypervisor as opposed to the number of physical hardware devices installed in the system. By taking advantage of LDoms, organizations gain the flexibility to deploy multiple, diverse operating systems simultaneously on a single platform.

The Logical Domains architecture includes underlying server hardware, hypervisor firmware, virtualized devices, and guest, control, and service domains. The hypervisor firmware provides an interface between each hosted operating system and the server hardware. An operating system instance controlled and supported by the hypervisor is called a *guest domain*. Communication to the hypervisor, hardware platform, and other domains for creation and control of guest domains is handled by the *control domain*.

Guest domains are granted virtual device access via a *service domain* which controls the system and hypervisor and assigns I/O. While the service domain is involved in configuring and setting up I/O partitioning, it does not interact with Logical Domains. In these roles, the control and service domains act as intermediaries, interfacing with the hypervisor on behalf of a guest domain to manage assignment and access of hardware resources such as the OpenBoot™ PROM, CPU, memory, network, I/O, console, and cryptographic units.

Bringing Virtualization to the Network

Unfortunately, current computing platforms are only able to virtualize processor cores and memory. Sun's 10 Gigabit Ethernet Networking Technology brings virtualization technology to the I/O channel by extending the parallel computing model with interface sharing, partitioning, and network virtualization. Toward this end, Sun's technology provides integrated policy, VIP addressing, and resource mapping from the protocol stack to the device hardware.

Sun's 10 Gigabit Ethernet Networking Technology and Logical Domains work together to make network virtualization a reality. While current I/O virtualization technologies require guest domains to communicate via the service domain, Sun's 10 Gigabit Ethernet Networking Technology gives guest domains a direct communication channel to the network (Figure 3-2). Each guest domain believes it owns the entire NIC and the bandwidth it provides, yet in practice only a portion of the total bandwidth is allotted to the domain. As a result, every NIC can be configured as demand dictates, with each domain receiving bandwidth on an as-needed basis.

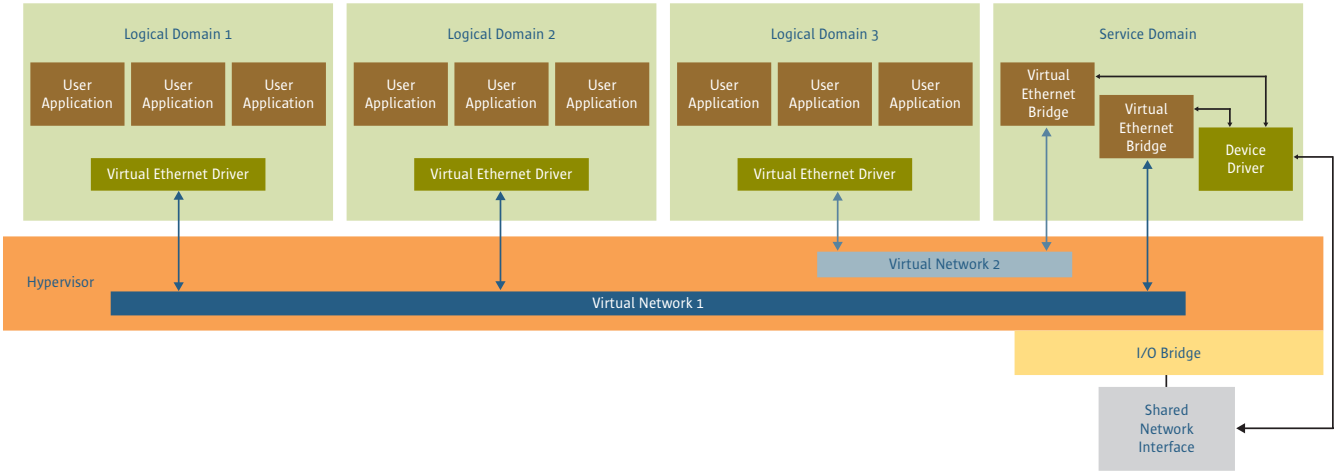

Figure 3-2. Sun's 10 Gigabit Ethernet Networking Technology enables data movement to occur directly between a Logical Domain and a virtualized device

Speeding Network Performance

Sun's 10 Gigabit Ethernet Networking Technology includes several innovations and optimizations that work together with parallelism and virtualization features to speed network performance.

- *Greater scalability*
  As organizations consolidate servers and more applications and services run on a single system, network demand rises. Single processor-based systems often fail to keep pace with the rate of network traffic and cannot process data quickly enough to keep the network interface busy. By exposing parallelism in the hardware, Sun's 10 Gigabit Ethernet Networking Technology lets organizations dedicate hardware resources per thread to avoid resource contention and utilize the full potential of available network bandwidth.

- *Reduced memory latency*
  Today, all network data is sourced from, and placed into, main memory. As a result, memory latency can become a critical factor in network performance. Sun's 10 Gigabit Ethernet Networking Technology reduces memory latency by moving networking closer to memory and utilizing heavy pipelining I/O transfers within a thread, and between threads. In addition, relaxed memory access and on-chip transaction re-ordering help maximize I/O to memory bandwidth.

- *Minimized translation lookaside buffer thrashing*
  Translation lookaside buffers (TLBs) are used in many processors and ASICs to cache virtual to physical memory address translations and associated page protection and usage information. When memory addresses are moved in and out of the TLB, thrashing can result and impact performance. Sun's 10 Gigabit Ethernet Networking Technology includes optimizations that improve TLB hit rates and reduce TLB thrashing in the I/O Memory Management Unit.

- *Reduced packet overhead*
  All network packets contain protocol overhead that can impact how quickly information is parsed and processed. Sun's 10 Gigabit Ethernet Networking Technology minimizes *per-packet overhead* — the common overhead that occurs with every packet — including programmed I/O (PIO) and direct memory access (DMA) overhead, and the impact of interrupt management. In addition, Sun's ASIC addresses the challenges posed by per-byte overhead associated with data by speeding the performance of data movement (copies) and checksum calculations.

- *Improved mutex lock contention optimizations*
  Software applications and operating systems use mutex locks to indicate a computing resource is allocated for exclusive access. Sun's 10 Gigabit Ethernet Networking Technology dedicates hardware resources to threads to avoid resource contention, mitigate the problems associated with deadlock conditions, and optimize network resource utilization.

Chapter 4
Summary

Companies rely on the network to collaborate and get work done. As datacenters consolidate systems and virtualize resources in order to host more applications and services, greater demand for network bandwidth is becoming commonplace. Sun's 10 Gigabit Ethernet Networking Technology helps accelerate multithreaded application performance by optimizing I/O throughput. By combining innovative chip technology with advanced NIC designs and sophisticated Solaris software, Sun delivers an integrated, optimized network environment with unmatched throughput.

For More Information

To learn more about innovative Sun products based on 10 Gigabit Ethernet technology, contact a Sun sales representative or visit the Web sites listed in Table 5-1 below.

Table 5-1. Related Web sites.

Description	URL
Logical Domains	sun.com/bigadmin/hubs/ldoms
OpenSolaris Project	opensolaris.org
Project Crosssbow	opensolaris.org/os/project/crossbow
Solaris Operating System	sun.com/solaris
Sun Multithreaded Networking Card	sun.com/products/networking/ethernet/10gigethernet
Sun Servers	sun.com/servers

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA **Phone** 1-650-960-1300 or 1-800-555-9SUN (9786) **Web** sun.com

Printed in USA 03/07