

Ethernet: The High Bandwidth Low-Latency Data Center Switching Fabric

The transition to 10 GbE is under way in the data center, and the ubiquity and reliability of Ethernet make it a desirable solution for delivering a high performance data center network, storage and compute fabric. Ethernet technology has reached a maturity and is a compelling proposition. The mass-market availability of 10 GbE adapters and switches enables Ethernet based networks to deliver high bandwidth, high throughput and low latency solutions for the rigorous demands of data center applications.

Major Trends in the Data Center

There are a number of emerging trends that have the potential to significantly increase the complexity of data center network infrastructures. These trends are being driven by the following technology transitions:

Accessibility of High Performance Computing (HPC):

Clusters of commodity servers have rapidly evolved into a highly cost-effective form of supercomputer. As the technology has matured and costs have declined, enterprises across a wide range of industries have begun leveraging HPC for product design and simulation, data analysis and other highly compute-intensive applications that were previously beyond the reach of IT budgets. Off-the-shelf clusters frequently use Gigabit Ethernet as the cluster interconnect technology, but a number of cluster vendors are exploiting more specialized cluster interconnect fabrics that feature very low message-passing latency.

Multi-core Processing and Cluster Computing for

Enterprise Applications: With current semiconductor technology, the escalation of on-chip power consumption is preventing major increases in clock speed. As a result, future Moore's Law microprocessor performance improvements will come primarily from increased aggregate performance of multiple-cores per chip, coupled with only modest increases in clock rate. This shift in microprocessor technology means that compute-intensive enterprise applications will be transitioning to a multi-threaded programming model that supports efficient, parallelized program execution on multi-core servers and clusters of multi-core servers. As is the case for HPC, performance for multi-threaded enterprise applications will be highly dependent on low end-to-end latency and high bandwidth between sub-processes running on different processor cores or servers. For example, the Oracle database application (including versions 9i RAC, 10g and 11g) has been adapted to cluster computing in order to improve performance on commodity server platforms. On-line transaction processing (OLTP) environments running on Oracle clusters have found that the transactions per second (tps)

throughput are inversely proportional to the end-to-end latency of the cluster interconnect. As more enterprise applications transition to multi-threaded programming models, low latency cluster interconnect can be expected to become a mainstream requirement for enterprise data centers.

Virtualization: Many data centers are evolving toward a virtual data center model where enterprise applications have the flexibility to draw on pools of shared computing, storage and networking resources rather than being rigidly constrained to dedicated physical resources. Virtualization not only greatly increases the flexibility of the data center infrastructure to accommodate changing workload requirements, but also improves resource utilization and power efficiency, thereby reducing total cost of ownership (TCO). Maximizing the effectiveness and flexibility of resource virtualization implies that entire pools of computing resources can share resources at very high speed. This degree of resource sharing requires not only a high speed local area network (LAN) for server interconnect but also a high speed, low-latency storage area network (SAN) for shared access to virtual machine images and data. For more information on network architectures and designs optimized for virtualized servers, please see the Force10 Networks white papers: ***The High Performance Data Center: The Role of Ethernet in Consolidation and Virtualization or Design Guide: 10 Gigabit Ethernet Virtual Data Centers.***

As the technology trends described above continue to evolve, data center managers may find themselves facing the prospect of deploying three distinct switching fabrics to support the LAN, the SAN and low latency cluster interconnect. However, for most enterprise data centers, multiple switching fabrics would prove to be excessively expensive in terms of both CAPEX (with multiple switch fabrics and host adapters per server) and OPEX (with a much broader range of required operational and management expertise). Furthermore, additional interconnect complexity can prove to be a hindrance in achieving a

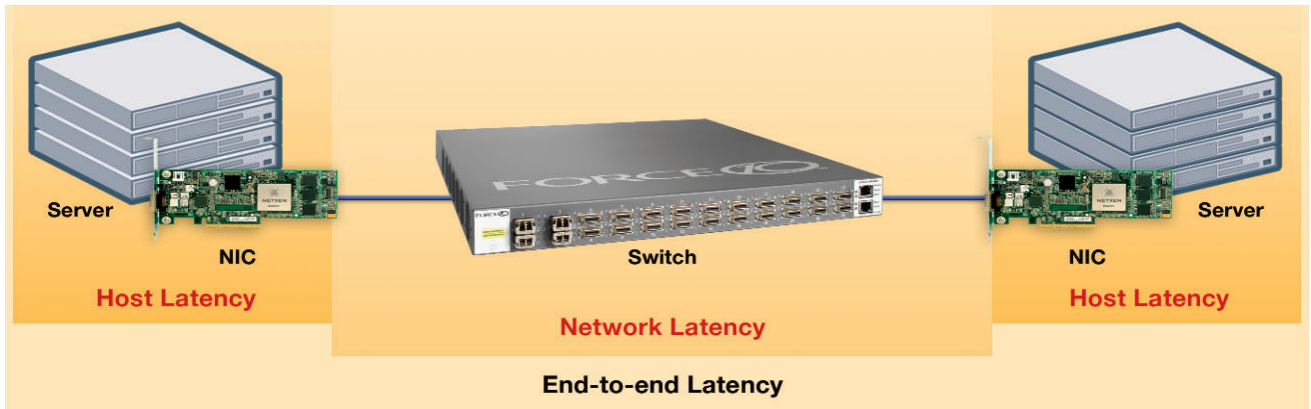


Figure 1. End-to-end latency

fully virtualized data center infrastructure and in developing the level of automated management of virtual resources required to implement service oriented architectures (SOA) for enterprise applications.

The alternative to multiple switching fabrics is to standardize on a single (unified) data center switching fabric that can simultaneously provide LAN interconnect, SAN interconnect and low-latency cluster interconnect. The remainder of this paper focuses on Ethernet as the best option for data center switch fabric unification.

Ethernet as the Unified Data Center Switch Fabric

Although Ethernet is the de facto technology for the general purpose LAN, Gigabit Ethernet has been considered as a sub-optimal switching fabric for very high performance cluster interconnect and storage networking. This is due primarily to performance issues stemming from the fact that GbE has lower bandwidth than InfiniBand and Fibre Channel, and typically exhibits significantly higher end-to-end latency and CPU utilization. However, this situation has changed dramatically due to recent developments in low-latency 10 GbE switching and intelligent Ethernet NICs that offload protocol processing from the host processor. These enhancements allow server end systems to fully exploit 10 GbE line rate, while reducing one-hop end-to-end latency to less than 10 microseconds and CPU utilization for line-rate transfers to less than 10 percent. As a result, 10 GbE end-to-end performance now compares very favorably with that of more specialized data center interconnects, eliminating performance as a drawback to the adoption of an Ethernet unified data center fabric. Off-loading protocol processing from the central CPU to the intelligent NIC can also improve the power efficiency of end stations because off-load ASIC processors are generally considerably more power efficient in executing protocol workloads.

Low Latency Switching

In the absence of congestion, end-to-end latency in the LAN includes two basic components, as shown in Figure 1:

- **Network latency** is the delay involved in serializing the message and switching it through network nodes to its destination.
- **Host latency** is the delay incurred in the end systems and NICs for transferring data between the application buffers and the network. Host latency is incurred in both the sending and receiving end system.

Low latency Ethernet switches employ cut-through switching to reduce the network latency component of end-to-end latency. With cut-through switching, the switch delays the packet only long enough to read the Layer 2 packet header and make a forwarding decision based on the destination address and other header fields (e.g., VLAN tag and 802.1p priority field). Switching latency is reduced because packet processing is restricted to the header itself rather than the entire packet, and the packet does not have to be fully received and stored before forwarding can begin. Low latency switching also reduces serialization time in multi-hop networks because the packet is serialized only one time rather than once per hop as in a network of store-and-forward switches.

Table 1 shows the network latencies vs packet size that are achievable with a typical low latency switch as a function of Layer 2 network diameters (number of switch hops). With store-and-forward 10 GbE switching, the network latencies would be typically 10-40x higher due to higher switching latency and the additional serialization delay at every hop across the network.

Intelligent Ethernet NICs

Traditionally, TCP/IP protocol processing has been performed in software by the end system's CPU. The load on the CPU increases linearly as a function of packets

Network Latency		
One Hop	64B	351 ns
	1500B	1.5 μs
Three Hop	64B	951 ns
	1500B	2.1 μs
Five Hop	64B	155 μs
	1500B	2.7 μs

Table 1. Network latencies with cut-through 10 GbE switching

processed, with the usual rule of thumb being that each bit per second of bandwidth consumes about a Hz of CPU clock (e.g., 1 Gbps of network traffic consumes about 1 GHz of CPU). As more of the host CPU is consumed by the network load, both CPU utilization and host send/receive latency become significant issues.

Intelligent Ethernet NICs offload protocol processing from the application CPU thereby eliminating software performance bottlenecks, minimizing CPU utilization, and greatly reducing the host component of end-to-end latency.

Over the last few years, vendors of intelligent Ethernet NICs, together with the RDMA Consortium and the IETF, have been working on specifications for hardware-accelerated TCP/IP protocol stacks that can support the ever-increasing performance demands of general purpose networking, cluster IPC and storage interconnect over GbE and 10 GbE. The efforts have focused on the technologies shown in Figure 2, which provides a simplified overview of hardware-assisted end system protocol stacks.

Intelligent NICs are frequently classified as TOE, RDMA or iSCSI NICs depending on the level of hardware support they provide for protocol processing.

A **TOE NIC** includes a dedicated hardware-based TCP offload engine (TOE). The TOE can offload essentially all the TCP processing from the host CPU. This reduces CPU utilization and also reduces latency because the protocols are executed in hardware rather than software. Applications generally access the TOE-supported stack via the sockets interface. Tests have shown that TOE NICs can improve web server performance as much as 10X vs. conventional 10 GbE NICs. A server with a TOE NIC is fully interoperable with end systems that are running conventional host-based TCP/IP protocol stacks.

An **RDMA NIC** incorporates a TOE for transport processing and also provides hardware support for a remote direct memory access (RDMA) mechanism. RDMA allows a host to read/write data directly between its user memory space and the user memory space of another RDMA host on the network, without any involvement of the host operating systems. The IETF has developed a set of standards for RDMA over TCP/IP and Ethernet called iWARP (Internet wide area RDMA protocol). iWARP's OS kernel bypass allows applications running in user space to post read/write commands that are transferred directly to the intelligent **iWARP NIC** or **R-NIC**. This eliminates the delay and overhead associated with copy operations among multiple buffer locations, kernel transitions and application context switches. iWARP NICs can reduce CPU utilization for 10 Gbps transfers to less than 10 percent and can reduce the host component of end-to-end latency to as little as 5–10 microseconds. As shown in Figure 2, a wide variety of network applications can gain the benefits of iWARP

RDMA via appropriate APIs and upper level interfaces.

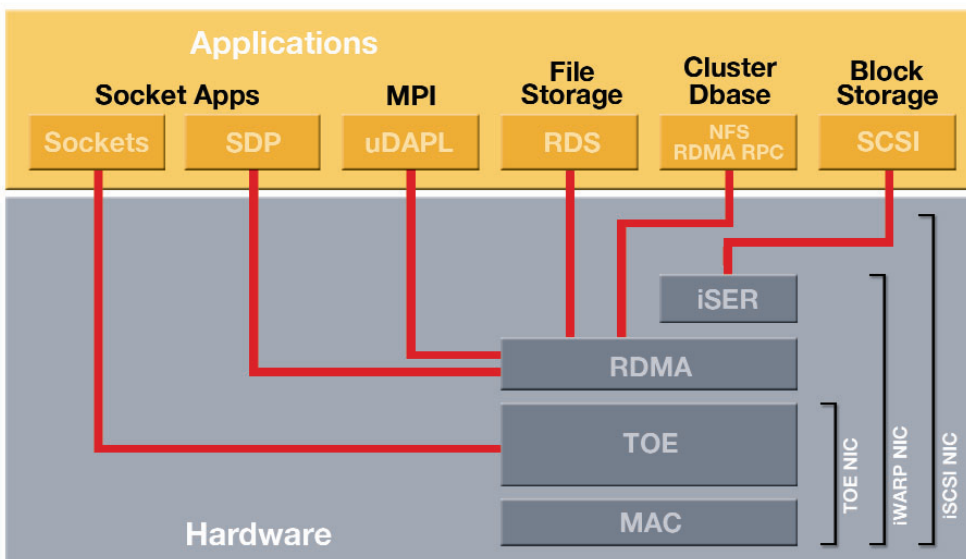
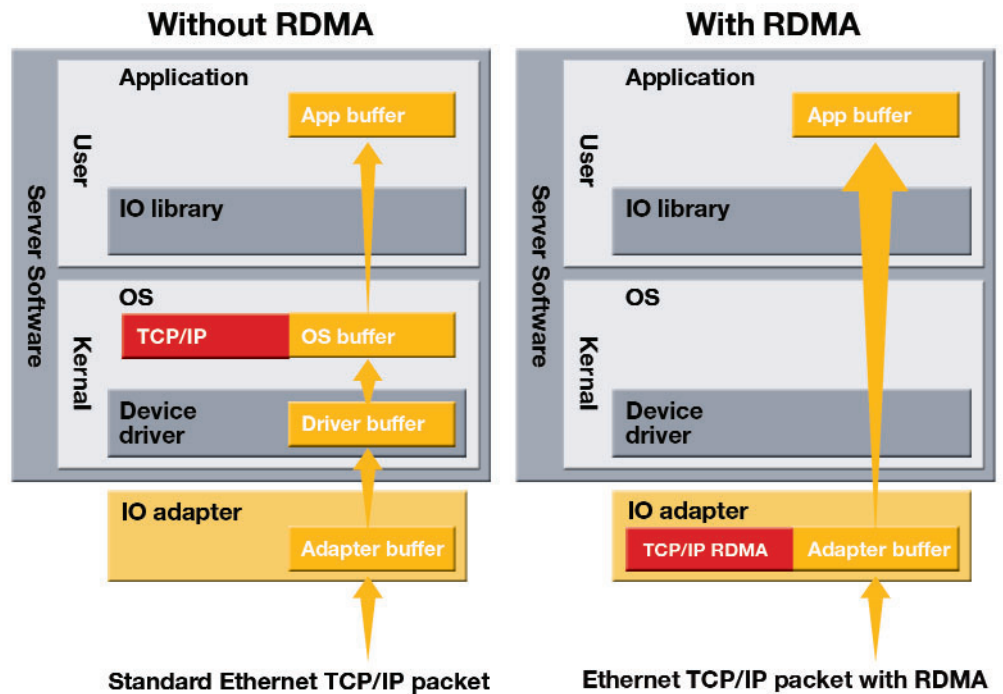


Figure 2. Intelligent Ethernet NIC protocol stacks

- Sockets direct protocol (SDP) allows unmodified socket applications to gain direct access to iWARP/RDMA-optimized data transfers
- User-level direct access transport APIs (uDAPL) serves as a lightweight API interface for cluster applications based on message-passing mechanisms, such as various versions of the message passing interface (MPI).
- NFS/RDMA RPC is a version of the ONC RPC that is being developed to access RDMA stacks in order to accelerate NAS file access speeds
- Reliable datagram sockets (RDS) is an open source IPC mechanism optimized for clustered databases by the Oracle Corporation. RDS allows messages to be sent reliably to multiple destinations from a single socket and simplifies migration of legacy-clustered databases to RDMA-enabled switch fabrics.
- iSCSI extensions for RDMA (iSER) is a component of iWARP that provides a datamover architecture (DA) extension to offload iSCSI SAN data movement and placement operations to the RDMA hardware, with the control aspects remaining in iSCSI host software. An *iSCSI NIC* is an iWARP NIC that supports iSER, as well as possibly offloading additional aspects of the host iSCSI processing to the NIC.

Figure 3 shows how the iWARP combination of TOE, RDMA and OS bypass essentially eliminate the CPU overhead related to networking by reducing host utilization for line rate 10 GbE to approximately 10 percent and end-to-end latency to less than 10 microseconds.

A "converged" Intelligent NIC is one that supports the full offload suite shown in Figure 2 (TOE, iWARP and iSCSI). The converged NIC delivers the maximum benefit from a unified Ethernet fabric because it can provide the highest levels of reduced host overhead and protocol offload to the full suite of data center applications and services, including cluster computing, web and enterprise applications, and storage networking, as shown in Figure 4. Also some intelligent NICs are capable of supporting I/O virtualization (IOV), which allows a single physical NIC to support multiple virtual NIC (vNIC) instantiations. For example, I/O virtualization can allow a single physical intelligent NIC to simultaneously provide multiple iWARP vNICs, multiple iSCSI vNICs and multiple TOE vNICs for the server's various cores and virtual machines. In



Source	Percent CPU Overhead Related to etworking	iWARP Techniques
Transport (TCP/IP) processing	40	Transport offload
Intermediate buffer copies	20	RDMA
Application context switching	40	OS bypass

Figure 3. Reduction of CPU latency and utilization with iWARP NICs

iWARP and the Open Fabrics Enterprise Edition (OFED) Stack

The iWARP protocol layers are shown in Figure 4 and consist of the following specifications:

- **iWARP verbs** are user level APIs that defines the behavior of iWARP R-NIC functionality as viewed from the host's upper layer protocols.
- **RDMA Protocol (RDMAP)** includes information in each packet that associates it with a specific application memory buffer location.
- **Direct Data Placement (DDP)** allows data segments to be placed directly in the destination application's buffer, even when packets arrive out of order.
- **Marker PDU Aligned framing for TCP (MPA)** defines how packets with RDMA and DDP headers are framed with TCP.

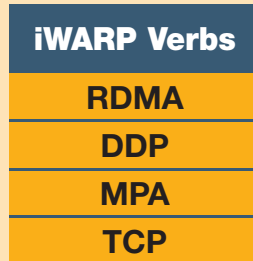


Figure 4. The iWARP RDMA stack

Recognizing the high degree of commonality between the two RDMA switching fabrics InfiniBand and iWARP/Ethernet, the Open Fabrics Alliance (OFA) has developed the Open Fabrics Enterprise Edition (OFED) RDMA stack of APIs and upper level interfaces that maximizes the common ground between IB and iWARP, while also recognizing specific components of each technology. OFED offers significant advantages to both software developers and end users. From the developer's perspective, application software based on OFED can readily be ported from one underlying RDMA technology to the other. This reduces development costs and increases the size of the available market. The end user benefits include the ability to readily migrate custom and off-the-shelf enterprise applications from one switching fabric to the other without major rewrites or other costly disruption of operations.

The OFED distribution includes an open source version of the Open Fabrics protocol stack supporting both iWARP and InfiniBand for the Linux operating system. Release 1.0 was made available by OFA in June 2006. The current release is 1.2.5. The OFED stack has been picked up by both the RedHat and the Novell/SUSE Linux distributions. Microsoft supports deployment of RDMA technology today using Windows Server 2003 with Winsock Direct Protocol (a precursor to SDP). Microsoft has also announced a roadmap for adding full iWARP support in the future RDMP Chimney release of their protocol stack. The current TCP Chimney stack supports TOE. A version of Windows OpenFabrics (WinOF) that supports Infiniband, but not iWARP, was released in July 2007. The current release is Release 1.0.1 (October 5, 2007).

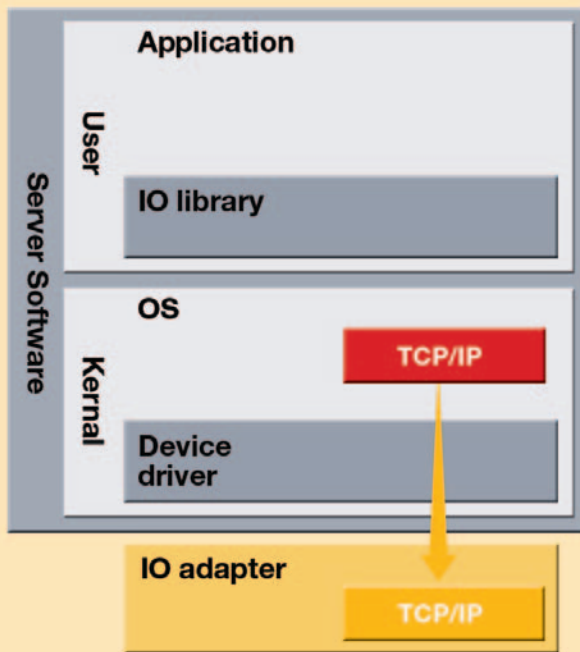


Figure 5. The Open Fabrics Enterprise Edition RDMA stack

addition, if there is a legacy application on the server that requires interaction with a host-based software protocol stack, the intelligent NIC with IOV can support a vNIC that emulates a traditional "dumb" Ethernet NIC that bypasses the RDMA/TOE protocol engine.

In the past, virtualized servers were rarely used for applications that require high levels of client/server I/O because of the overhead of multiple virtual machines sharing conventional GbE NICs. This limitation has been eliminated by the emergence of intelligent 10 GbE NICs that provide hardware support for I/O virtualization, offloading many of the traditional hypervisor tasks from the host software to the adapter. These IOV NICs allow virtualized applications to achieve performance levels previously experienced only on dedicated physical servers and dedicated physical NICs.

Traditional Advantages of Ethernet

Considering the significant performance improvements discussed earlier in this section, data center managers can give full consideration to the traditional advantages that Ethernet has over more specialized switching fabrics:

- **Lower Cost Interconnect:** Very high production volumes and a highly competitive market environment ensure that the Ethernet will continue to offer the lowest cost switches and host adapters.
- **Ubiquitous Connectivity:** Virtually every computer system shipped today comes with Ethernet built in. An increasingly popular option for servers is to incorporate the LAN on the motherboard, with higher performance servers beginning to support on-board 10 GbE intelligent NICs.
- **Proven Interoperability:** From the outset, Ethernet networks have relied on interoperability between the

products of multiple vendors of NICs, hubs, switches and routers. Proven multi-vendor interoperability across a very broad ecosystem has continued to be a major strength of Ethernet through successive generations of higher link speeds, including 10 Gbps today and 40 Gbps/100 Gbps within the next 3+ years.

- **Ease of Management:** Ethernet-based cluster and storage interconnect can be readily assimilated in the existing Ethernet network management environment without requiring the additional management tools or training needed for use of special purpose switch fabric and protocols.

The Future of Ethernet in the Data Center

There are a number of on-going standards bodies efforts to further enhance Ethernet's effectiveness as a unified data center switching fabric:

1. The Higher Speed Study Group under IEEE 802.3 is developing the next generation of higher speed Ethernet at 40 Gbps and 100 Gbps.
2. The IEEE 802.1Qau group is investigating enhanced congestion management capabilities for Ethernet that are intended to eliminate packet loss due to buffer overflow on congested ports. The methodology is based on bridges that use backward congestion notification (BCN) to cause end stations to rate limit or pause transmission. BCN congestion management would be aware of different .1p QoS traffic classes and would use a priority scheme to control rate limiting and "pause" notifications across the range of traffic classes. The methodology is also intended to be applicable across multiple tiers of cascaded Layer 2 switches, such as those typically found in large cluster interconnect and SAN fabrics.

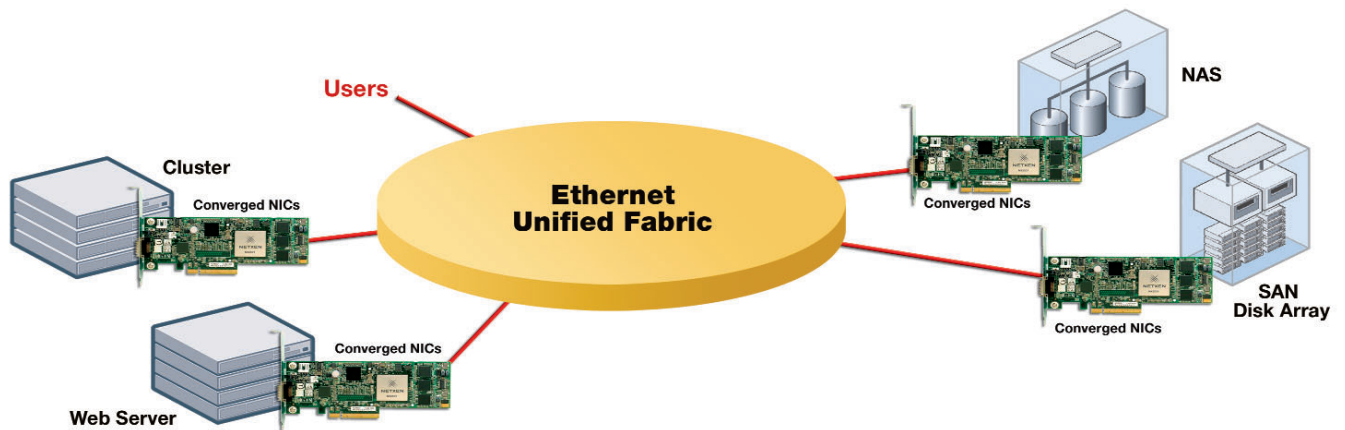


Figure 6. Ethernet unified fabric with converged NICs

3. Another working group that IEEE 802.1Qav is focused on enhanced transmission selection that will allocate unused bandwidth among the traffic classes including the priority class specified by 802.1Qau. Together 802.1Qau and 802.1Qav are part of the IEEE data center Ethernet features.
4. The Internet Engineering Task Force (IETF) is working on a project to develop an Ethernet link-layer routing protocol that will provide a shortest-path, adaptive routing protocol for Ethernet forwarding in arbitrary switch topologies. TRILL (Transparent Interconnection of Lots of Links) will feature:
 - Routing loop mitigation (TRILL is an alternative to STP that will not need to disable links to deal with loops)
 - Load splitting among multiple inter-switch links and multiple paths through a multi-hop network
 - Support for broadcast and multicast
 - Auto-configuration or minimal configuration
5. A T11 Working Group is currently developing a specification to support the Fibre Channel protocol directly over Ethernet (FCoE). FCoE, is designed to make converged networks possible for data centers. Servers will no longer require separate network adapters for data and storage networking. IT professionals can increase functionality while reducing equipment cost, power consumption and

complexity by using a single set of adapters, cables and switches. FCoE-converged network adapters (CNAs) appear to the operating system as Fibre Channel HBAs. The FCoE effort will also leverage the IEEE work on data center Ethernet for congestion management and "lossless Ethernet" described above. FCoE will allow end users to protect their previous investments in Fibre Channel as they move toward adopting Ethernet as a unified switching fabric. FCoE will also present an opportunity for intelligent NIC vendors to provide hardware support for the Fibre Channel upper layer protocols in their RDMA NICs.

Conclusion

The combination of low latency 10 GbE switching and intelligent iWARP RDMA NICs gives data centers the opportunity to gain the full benefits of 10 GbE interconnect without the drawbacks of high CPU utilization and high end-to-end latency that were issues with traditional GbE. This opens the door for data center managers to implement unified switching fabric strategies that leverage the many strengths of TCP/IP and Ethernet as pervasive networking standards supported by huge ecosystems of semiconductor, system and software vendors.