ethernet alliance

Data Center Bridging

Version 1.0 November 2008

Authors:

Steve Garrison, Force10 Networks

Val Oliva, Foundry Networks

Gary Lee, Fulcrum Microsystems

Robert Hays, Intel

Table of Contents

1.0 Executive Summary

Networks are the essential part of any modern data center and they must deliver reliability, availability and high performance. Enterprises rely on their data centers to run business operations, service providers rely on their data centers to generate revenues by delivering network services, and content providers rely on their data centers to distribute revenue-producing content. Ethernet is the most widely deployed networking technology today. Currently, it fulfills increasingly demanding requirements for a variety of business needs. But can Ethernet technology evolve to help data centers improve cost-effectiveness and meet the demands for next-generation services?

2.0   Data Center Networks

Data center network designs can vary significantly depending on the vertical market, user profile, and size of enterprise. However, most data center networks have adopted a tiered design that is reflected in the logical and physical topology. Figure 1 shows a network topology of a simple data center. The tiers of the network are focused on the various requirements of data center functions:

1. Customer Network Access tier. Including, local access network (LAN), remote access using wide area network (WAN) services or virtual private networks (VPNs). The access tier can include numerous devices, such as web servers, WAN application accelerators, firewalls, and security devices.

2. Customer-Facing Services tier. Including, servers that provide network services for dynamic host configuration protocol (DHCP), domain name system (DNS), and enterprise or hosted applications.

3. Backend Services tier. Including, database applications, administration/ management, and access to high performance computing resources.
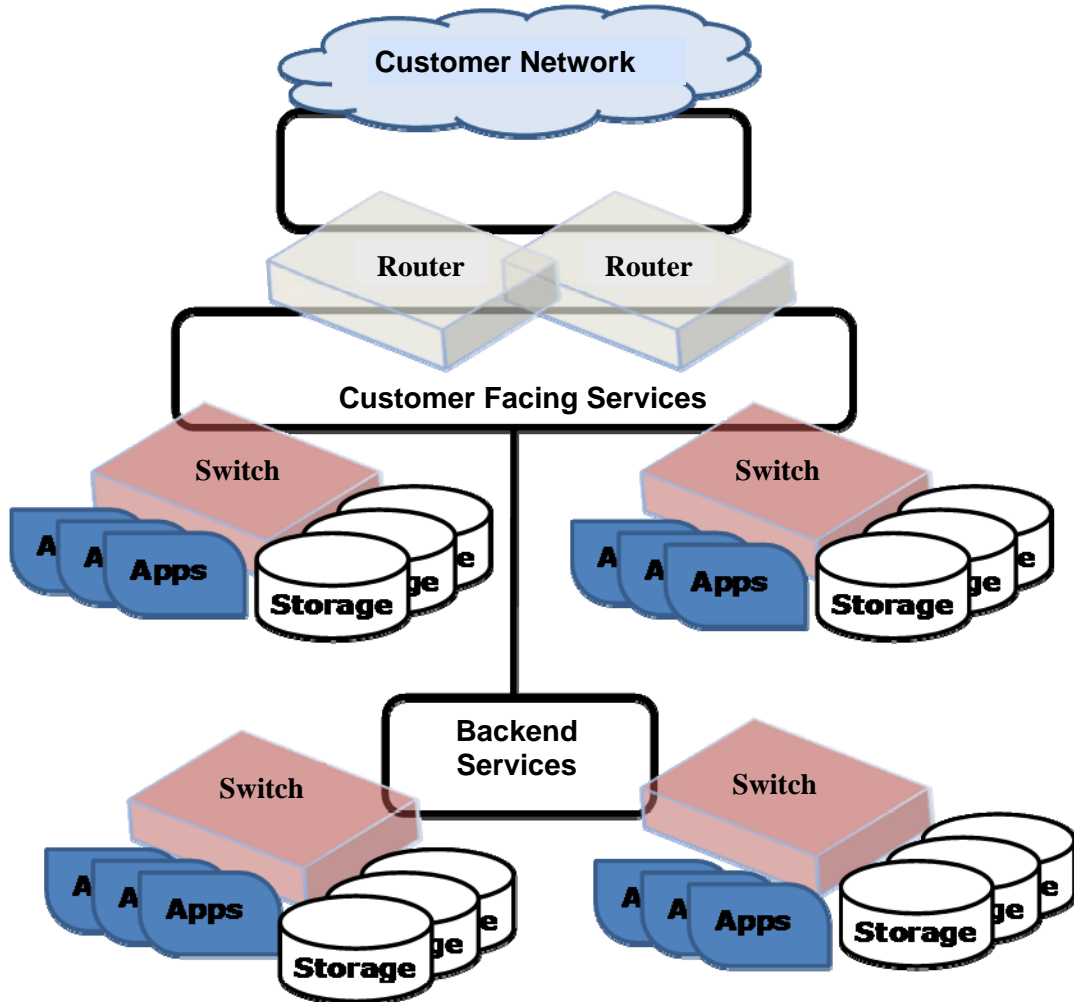
Figure 1—A Simple Data Center

Today, data centers must be highly secure, reliable, and provide scalable performance while remaining cost-effective. Balancing performance and other required attributes with cost-effectiveness can be challenging, especially in the face of increasing complexity as more application and user demands are placed on data center resources.

Ethernet technology has a long tradition of delivering the most cost-effective solution for high performance networking. Some of the traditional advantages of Ethernet networks include:

- **Lower Cost Interconnect:** Ethernet's large production volumes and a highly competitive market environment ensure that it will continue to offer the lowest cost switches and host adapters.

- **Ubiquitous Connectivity:** Virtually every computer system shipped today includes at least one Ethernet port as a standard feature on the motherboard. Moving forward, 10 Gigabit Ethernet will provide a cost-effective unified fabric for high performance computing (HPC), data and storage traffic throughout the data center.

- **Proven Interoperability:** The IEEE, specifically the 802.3 Ethernet Working Group, has been the guardian of the Ethernet standard and helped nurture its development. This assures the interoperability between the products of different vendors of network interface cards (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

- **Ease of Management**: Enhancements to Ethernet for general purpose networks or 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 fabrics and protocols.

As data center requirements evolve, Ethernet must continue to adapt to address issues within a data center such as:

- Increasing demand for higher bandwidth and performance
- Power conservation and energy efficiency
- Virtualization, including universal server access to networked storage resources
- Transition to service oriented application architectures (SOA)
- Low latency interconnect for high performance cluster computing and databases

3.0   Technologies for Data Centers

Ethernet technology is standard and deployed in virtually every data center throughout the world; cost-effectiveness and availability from many vendors made this possible. Despite Ethernet's traditional advantages, special purpose switching fabric technologies needed to be developed to address special requirements, such as storage area networking, to deliver a much lower latency.

However, upcoming enhancements to the Ethernet standards are making progress in meeting data center requirements for a 'unified switching fabric', removing the need for special purpose switching fabric. These enhancements also address major networking issues that data centers will face in the future, for example:

- 10GBASE-T (IEEE Standard 802.3an)
- Intelligent Ethernet NICs with TCP/IP Offload Engines (TOE), Internet Wide Area RDMA Protocol (iWARP), and internet small computer system interface (iSCSI) support
- Low Latency Ethernet Switching (cut-through 10 GbE switching at Layer 2)
- 40Gbps and 100Gbps Ethernet (IEEE P802.3ba™)
- Energy Efficient Ethernet (IEEE P802.3az™)
- Class-base pause frames (IEEE P802.3x™)
- Congestion Notification (IEEE P802.1Qau™)
- Enhanced Transmission Selection (IEEE P802.1Qaz™)
- Shortest Path Bridging (TRILL and IEEE P802.1aq™)
- Fibre Channel over Ethernet (T11 FCoE)

The following sections describe these upcoming enhancements to Ethernet.

3.1   10GBASE-T

Released in 2006, the 10GBASE-T or IEEE 802.3an™ standard provides for 10 Gbps connections over unshielded or shielded twisted pair copper cables. With a reach of up to 55 meters using existing Category 6 cabling or 100 meters on new enhanced Category 6A cabling, 10GBASE-T allows for gradual transition from existing 1000BASE-T connections with the support of autonegotiation (NWay).

With the introduction of 10GBASE-T products by several silicon, NIC, and switch system manufacturers, it is promising to be a valuable technology for data centers due to its greater range compared with existing copper cable solutions (e.g., CX4 and Direct Attach Twin-Ax) and at a significantly lower cost than fiber optic solutions.

3.2   Intelligent Ethernet NICs

Over the last few years, Ethernet Controller/NIC vendors along with industry associations, such as the Remote Direct Memory Access (RDMA) Consortium, Open Fabrics Alliance, and Internet Engineering Task Force (IETF) have been working on specifications to enable hardware acceleration for network protocol processing. Traditionally, all network protocol stack processing been done completely in software. Hardware acceleration for transmission control protocol/internet protocol (TCP/IP), iSCSI, or iWARP protocol processing delivers higher performance networking for applications such as compute cluster message passing and networked storage. These efforts have focused on the technologies shown in Figure 2, which provides a simplified overview of hardware-assisted end system protocol stacks.

The IETF standard for RDMA over TCP/IP and Ethernet is iWARP. iWARP's operating system (OS) kernel bypass allows applications running in the user space to post read/write commands that are transferred directly to the intelligent iWARP NIC or Remote Direct Memory Access NICs (RNIC). This process or function eliminates delay and overhead associated with copy

operations among multiple buffer locations, kernel transitions, and application context switches. iWARP NICs can reduce central processing unit (CPU) utilization for 10 Gbps transfers to less than 10%, and can reduce the host component of end-to-end latency to as little as 5-10 microseconds.
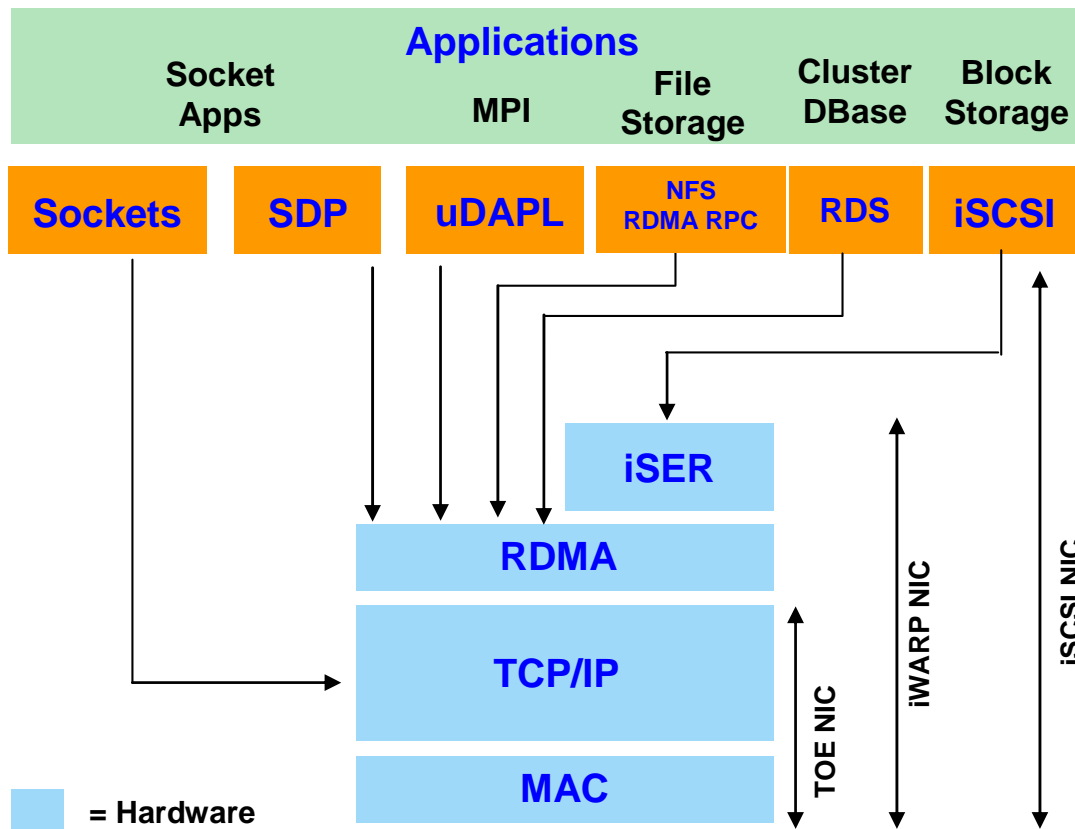
Figure 2 - Intelligent Ethernet NICs

Ethernet NICs that provide hardware acceleration for protocol processing are also known as TOE, RNIC, and iSCSI or Fibre Channel Protocol over Ethernet (FCoE) Host-Bus Adapters (HBA) depending on the type of hardware support they provide.

Ethernet NICs are also increasingly capable of supporting I/O Virtualization (IOV), which allows a single physical NIC to support multiple virtual NIC (VNIC) instantiations for the server's various cores and virtual machines. In addition, intelligent NICs can switch to host-based processing when legacy applications are running.

3.3   Low Latency Ethernet Switching

Some low latency 10 GbE switches use cut-through switching to reduce port-to-port latency, reducing the overall end-to-end network latency. Cut-through switches delay the packet long enough to read the Layer 2 packet header (e.g., virtual LAN tag and 802.1p® priority field) to make a forwarding decision. Switching latency is reduced because packet processing is restricted to the header itself rather than the entire packet, therefore, packet forwarding can begin.

Cut-through switches also reduce serialization time in multi-hop networks by serializing only once rather than once per hop as in a network with store-and-forward switches. Cut-through switches are applicable in Layer 2 networks where systems and switch ports operate at 10 Gbps because it removes speed changes required to store packets. As a result, there is a natural fit between cut-through switches and intra-tier Layer 2 data center connectivity among servers and networked storage resources.

3.4   40Gbps and 100Gbps Ethernet

The IEEE 802.3ba™ task force is developing the next generation of higher speed Ethernet. Today's agreed upon speeds are 40 and 100 Gbps.

Ethernet speeds beyond 10 Gbps are currently a requirement in many data centers, especially for content and service providers. The growing demand for online video content is driving the need for higher bandwidths to connect switches and routers. For example, YouTube currently delivers 2.9 billion online videos per year in the U.S. alone.

Figure 3 shows the recent annual growth in Internet traffic flowing through the London Internet Exchange (LINX)[1]. In the enterprise data center, demand for higher speed Ethernet will be driven by accelerated adoption of 10 GbE

---

1 LINX is an Internet Exchange Provider (IXP) that connects many network service providers using an Ethernet-built network fabric.

server connectivity enabled by the maturation of 10GBASE-T. 40 GbE and 100 GbE will be able to offer cost-effective data center connectivity with 100 meters of reach over multi-stranded multimode optical fiber.
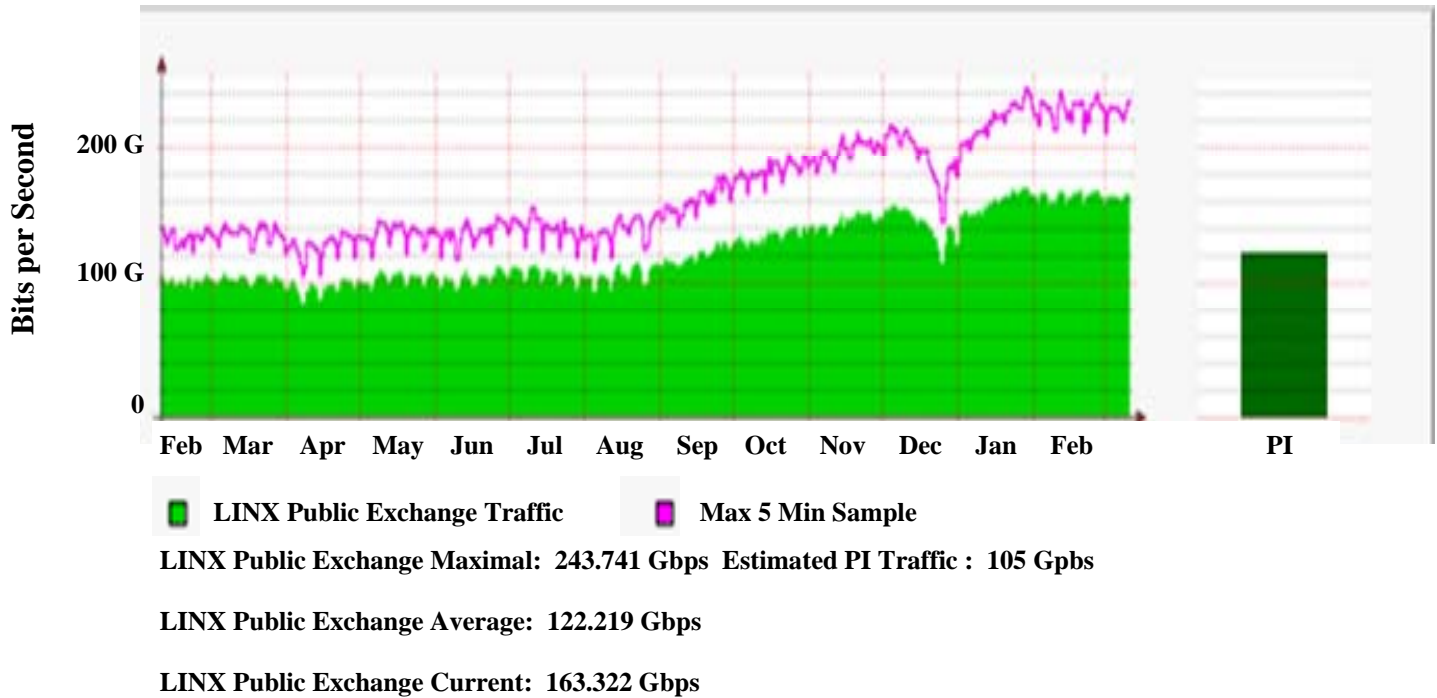

LINX Public Exchange Maximal: **243.741 Gbps** Estimated PI Traffic : **105 Gpbs**

LINX Public Exchange Average: **122.219 Gbps**

LINX Public Exchange Current: **163.322 Gbps**

Figure 3 - Yearly traffic view from the LINX

3.5   Energy Efficient Ethernet

Currently, Ethernet interfaces typically consume the same amount of power regardless of traffic load or network utilization. The Energy Efficient Ethernet (IEEE P802.3az™) task force is writing a specification to reduce the power of links while idle or operating under a light load.

The solution selected by the task force, introduces a new Low-Power Idle state that enables systems to reduce power consumption between transmission periods without inhibiting high-speed communications. This solution also provides a method to optionally negotiate the wake or resume time from low -power idle, allowing systems to safely enter deeper sleep states to save ad-

ditional power. With Energy Efficient Ethernet, energy consumption will be more directly proportional to bandwidth utilization, in contrast with current Ethernet devices which consume similar power regardless of the utilization rate or traffic load.

The IEEE 802.3az standard of Energy Efficient Ethernet will complement other Green Networking efforts, such as the U.S. EPA's Energy Star program, improving the energy-efficiency of network devices and reducing data center operational costs. According to a Gartner Group study, the annual costs of data center power and cooling can exceed the capital costs of the server hardware.

3.6 Class-based Pause Frames

Class-based pause is being developed by the IEEE as an extension of the IEEE P802.3x™ pause frame definition. The new class-based pause frame can differentiate up to 8 classes of service in order to provide lossless operation for HPC and storage traffic in the datacenter. This is accomplished by instructing the upstream link partner to only pause traffic classes which are experiencing congestion. In this way, uncongested classes can continue transmission across the link.

3.7 Congestion Notification

The IEEE P802.1Qau task force is investigating enhanced congestion management capabilities for Ethernet that are intended to minimize congestion hotspots in multi-stage Ethernet fabric configurations. This methodology is based on bridges that use Quantized Congestion Notification (QCN) to cause traffic sources to rate limit or pause transmission.

QCN congestion management will be aware of different IEEE 802.1p QoS™ traffic classes and will send a QCN frame back to the traffic source in order to control the traffic rate by using 'pause' notifications across the range of traffic classes. This methodology is also intended to be applicable across multiple tiers of cascaded Layer 2 switches, such as those typically found in larger data centers for cluster interconnect and SAN fabrics.

3.8   Enhanced Transmission Selection

Another task force, IEEE P802.1Qaz, is focused on Enhanced Transmission Se-lection that will allocate unused bandwidth among the traffic classes includ-ing the priority classes specified by 802.1Qau. Together P802.1Qau and 802.1Qaz are part of the IEEE standards to support 'Data Center Ethernet'.

3.9   Shortest Path Bridging

The IETF is working on a project to develop an Ethernet link-layer routing pro-tocol that uses a shortest-path, adaptive routing protocol, for Ethernet for-warding in switch topologies. Transparent Interconnection of Lots of Links (TRILL) features:

- Routing loop mitigation
- 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

A similar effort is being pursued by the IEEE P802.1aq task force to define a standard for shortest path bridging of unicast and multicast frames (based on the intermediate system-intermediate systems protocol) supporting multiple active topologies.

3.10   Fibre Channel over Ethernet (FCoE)

The T11 Working Group is currently developing a specification to support Fibre Channel over Ethernet (FCoE). FCoE is intended to enhance the Ethernet's ability to serve as a unified data center switching fabric; no longer requiring servers to separate network adapters for data and storage networking. FCoE

will allow IT professionals to control equipment cost, power consumption, and complexity by using a single set of adapters, cables, and switches.

FCoE will leverage the efforts of the IEEE as related to lossless Ethernet congestion management and enhanced transmission selection as described above. FCoE will also allow end users to protect their previous investments in Fibre Channel as they move toward adopting Ethernet as a unified switching fabric. With FCoE, the unified Ethernet fabric will be capable of supporting NAS, iSCSI, and Fibre Channel storage resources, as shown in Figure 4.
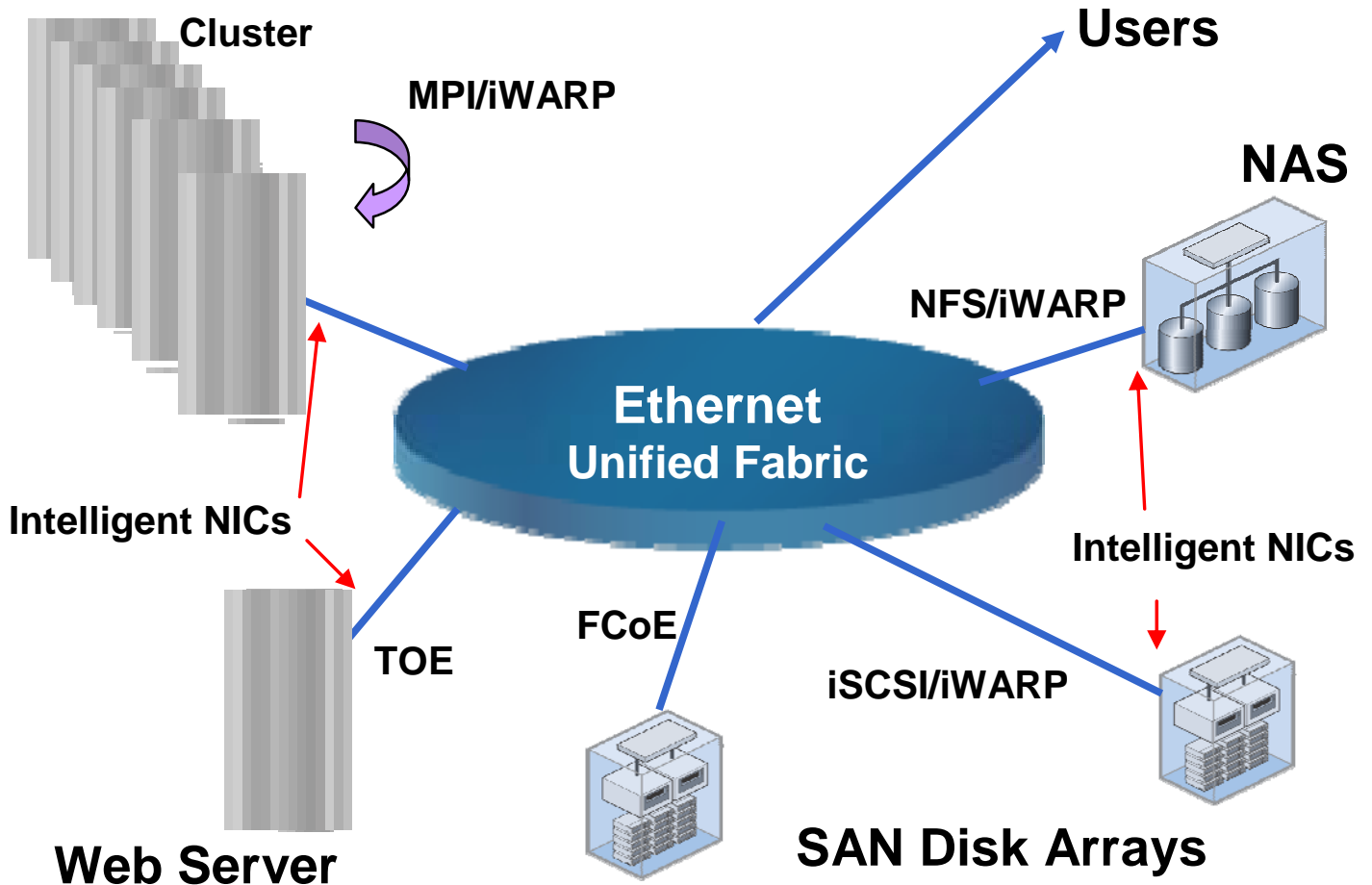
Cluster

Users

MPI/iWARP

NAS

NFS/iWARP

Ethernet
Unified Fabric

Intelligent NICs

Intelligent NICs

TOE

FCoE

iSCSI/iWARP

Web Server

SAN Disk Arrays

Figure 4 - Ethernet Unified Fabric with FCoE

4.0   Summary

Emerging Ethernet technologies will focus on enhancing the Ethernet's ability to meet the networking requirements of current and future data centers. These technologies will create an opportunity for data center managers to implement unified switching fabric strategies that leverage the strengths of Ethernet and Internet Protocol (IP) as pervasive networking standards supported by ecosystems of semiconductor, system, and software vendors.

In order to help coordinate the numerous development efforts, the Ethernet Alliance has recently chartered the Data Center Ethernet subcommittee to work with users, vendors, and standards bodies to ensure a cohesive Ethernet solution set for data center networking.

Chartered as a resource for IT professionals for data center-focused Ethernet technologies and standards, the Data Center Ethernet subcommittee was formed to address the demands on Ethernet networking in the data center, as the role of Ethernet expands from transporting TCP/IP packets to technologies such as low latency data center application networking, unified fabric over Ethernet, and energy-efficient Ethernet.

In addition to assisting users and developers of Ethernet, the Data Center subcommittee will identify areas where technical work may be needed to provide enhancements; enabling Ethernet to remain the best choice for IT professionals to meet their growing data center networking requirements.