Emerging Fabric Technologies

technology brief

Abstract

This paper describes the different fabric technologies on the horizon and their most likely implementations in the data center.

Introduction

Fabric interconnect technologies are emerging opportunities as data centers try to efficiently scale computing power and storage capacity. One example of the growing importance and focus on fabric technologies is the industry leading ProLiant BL family of blade server products. Up to 280 of these servers can be placed in a single industry-standard rack. At the same time, data centers are consolidating hundreds of servers into a much smaller number of symmetric multiprocessing servers to increase processing power and reduce management costs. Data centers are also building high-speed storage networks to increase storage capacity and improve resource utilization. These advances in computing and storage technologies are placing a considerable burden on the data center's network infrastructure.

A typical data center uses a variety of interconnects to link servers to servers and servers to storage resources. Front-end web servers are usually connected to local area networks (LANs) using Ethernet. Today, back-end database servers and other application servers are connected to storage resources using Fibre Channel. Inside servers, local I/O interconnects such as ATA and SCSI technologies are prevalent. The use of multiple system and peripheral bus interconnects decreases compatibility and management efficiency and drives up the cost of equipment and the personnel needed to operate and maintain it. To increase efficiency and lower costs, today's data center network infrastructure must be transformed into a unified high-speed system.

The concept of a unified high-speed data center infrastructure is relatively new. Such a system requires a high-bandwidth, low-latency fabric that can move data smoothly between servers, storage, and applications. This paper describes new fabric emerging fabric technologies that can unify and standardize the data center infrastructure. These technologies include RDMA/TCP (Remote DMA over TCP/IP), InfiniBand, and Internet Protocol (IP) storage protocols such as Internet SCSI and Fibre Channel over IP.

Data center requirements for fabric technologies

Data centers require reliable, high-performance networks with scalable bandwidth and the ability to handle all classes of traffic: computing, storage, and communications. These requirements are described in more detail below.

- high scalability – This refers to the ability to add additional performance for a given application as user demand increases. Scaling has traditionally been accomplished along two paths, increasing the processing capability within a given server (called "vertical scaling, or "scaling up") or by sharing the processing workload across servers connected over a fabric or network ("horizontal scaling " or "scaling out").
- · high reliability – As business applications become increasingly mission-critical, the need for hardware and software to maximize solution uptime becomes more important. New fabric architectures hold the promise of increased availability by allowing workloads to be distributed across the network.
- · remote management – IT personnel must be able to remotely manage servers, storage and networking configurations from anywhere in the world with the same level of control they have in the data center. Fabric interconnects extend remote management by enabling increased control

over computing and storage resources which can be provisioned dynamically without having to physically reconfigure network infrastructures.

- · adaptability– To handle all classes of traffic, networks must be able to satisfy the data transmission requirements of many types of servers, operating systems (OSs), and protocols as well as have the ability to easily be deployed and redeployed among various needs.
- · open industry standards – Data centers require fabric technologies based on cost-effective, open industry standards. Industry standards are important because they are broadly adopted and operable with other complementary technologies. Therefore, customers can be confident that an infrastructure based on such unifying industry standards will protect their capital investment in technology, will have the full support of independent software and hardware vendors, and will have best-in-class industry expertise available.

Network interconnect challenges

Transmission Control Protocol/Internet Protocol (TCP/IP) is the suite of protocols that drive the Internet. Every computer connected to the Internet must use the protocol to send and receive information. Information is transmitted in fixed data block (packet) sizes so that heterogeneous systems can communicate in a standardized format. Computers implement the TCP/IP protocol stack to process outgoing and incoming packets. Today, TCP/IP stack implementations are usually in operating system software and, therefore, must be processed by the CPU. As network speeds move beyond 1 Gb/s, CPUs become burdened by a large amount of TCP/IP protocol processing.

Why does TCP/IP protocol processing require so much CPU power? The TCP/IP stack of protocols was developed to be an internetworking language for all types of computers to transfer data across different physical media. TCP/IP protocols involve over 70,000 software instructions that provide all the necessary reliability mechanisms, error detection/correction, sequencing, recovery, and other communications features. As a result, protocol processing of incoming and outgoing network traffic consumes CPU cycles that could be used for software applications. This also negatively impacts an application's ability to scale across a large number of servers.

The burden of protocol stack processing is compounded by a finite amount of memory bus bandwidth. Incoming network data consumes the memory bus bandwidth because each data packet must cross the memory bus at least three times. The receiving device writes data to the device driver buffer, copies it to an operating system buffer, and then copies it into the application's memory space. These copy operations add latency. Considering that the original data often must be split into smaller chunks, data moving at 1 Gb/s could consume a considerable percentage of the memory bus bandwidth and force all CPUs to stall for memory.

Key enabling technologies for fabrics

TCP/IP protocol overhead and constrained memory bandwidth are obstacles to deployment of faster Ethernet networks in the data center. The use of TCP/IP offload engines (TOE) and remote direct memory access (RDMA) technology can diminish these obstacles. RDMA is often confused with TOE, but they perform different functions. Both technologies are described in more detail below.

TCP/IP Offload Engines

To reduce CPU utilization, high-speed IP networks need TCP/IP protocol stack processing to be offloaded (moved) to a dedicated adapter card (host bus adapter or network interface card) with TOE capability. TOE is logic embedded in a chip or firmware on the adapter card. By offloading protocol processing from the CPU, TOE can reduce CPU utilization to less than 20 percent. It is important to note that TOE adapters do not remove the number of times that data is copied to system memory buffers; they move the processing to the network interface card (NIC). As described below, the

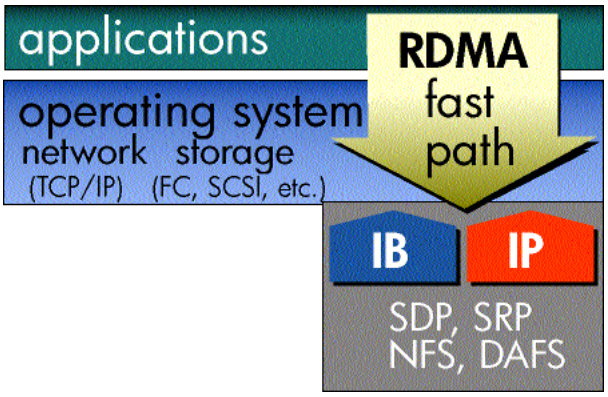
objective of RDMA technology is to both reduce the amount of network protocol processing that needs to take place and to reduce the number of memory data copy operations.

RDMA technology

RDMA technology was developed to move data from the memory of one computer directly into the memory of another computer with minimal involvement from their CPUs. Additional information included in the RDMA protocol allows a system to place the communicated data directly into its final memory destination without any additional or interim data copies. This "zero copy" or "direct data placement" (DDP) capability provides the most efficient network communication possible between systems.

As shown in **Figure 1**, RDMA provides a faster path for applications to transmit messages between servers. The RDMA protocol will be common for both InfiniBand™ (IB) and IP. Either interconnect (IP or IB) can support all of the emerging wire standards such as Sockets Direct Protocol (SDP), RDMA-enabled SCSI, Common Internet File System (CIFS), Network File System (NFS), and Direct Access File System (DAFS).

Figure 1. RDMA provides a faster path to the fabric (InfiniBand or IP).



HP advances RDMA and TOE technologies

HP is at the forefront of RDMA and TOE technology initiatives. HP has been a leader in the development of cluster interconnects such as ServerNet, the Virtual Interface (VI) Architecture, and InfiniBand™, which were the origins of RDMA technology. HP is a founding member of the RDMA Consortium, which is an independent group formed to develop the architectural specifications necessary to implement products that provide RDMA over TCP/IP networks.

Emerging fabric technologies

Multiple established and emerging fabric technologies are battling for space in the data center. The leading fabric technologies are RDMA over TCP (RDMA/TCP), InfiniBand, and IP storage protocols such as Internet SCSI (iSCSI) and Fibre Channel over IP (FCIP). These technologies are described below.

RDMA/TCP

Ethernet is the most prevalent network transport in use today. Customers have invested heavily in Ethernet technology and are unwilling to tear out their networks and replace them with new network

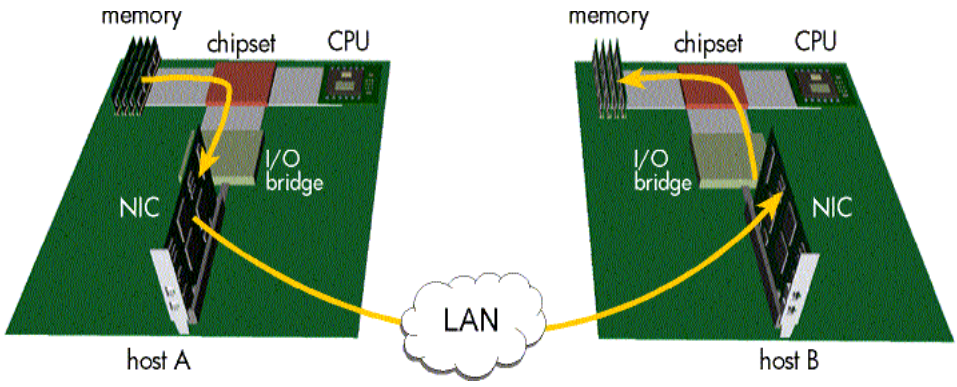
technologies. Their reliance on Ethernet is justified by its low cost, backward compatibility, and consistent bandwidth upgrades over time. Current Ethernet speeds in the data center are 100 Mb/s and 1 Gb/s. The next speed increase is 10 Gigabit Ethernet. Customer migration to 10-Gigabit Ethernet will be tempered by the I/O processing burden it places on servers. The use of TOE adapters will offload protocol processing from CPUs, but it will not remove the memory bandwidth barrier to 10 Gb/s speeds. The addition of RDMA capability to Ethernet will reduce CPU utilization and increase customer migration to 10-Gigabit Ethernet.

The addition of RDMA capability to fabric interconnects will allow data centers to converge their infrastructure over fewer types of interconnects. This simplifies server deployment and improves infrastructure flexibility. As a result, the data center infrastructure will become less complex and easier to manage. The data center will also have a more adaptive infrastructure. For example, bandwidth can be diverted from networking to iSCSI when provisioning a database server, and the bandwidth can be diverted back to networking if the server is reprovisioned as an application server.

How RDMA/TCP works

RDMA/TCP specifies a communication protocol layer that moves data directly between the memory of applications on the two nodes, with minimal work by operating system kernel, and without interim data copying into system buffers (**Figure 2**). This capability enables RDMA/TCP to work over standard TCP/IP-based networks that are commonly used in data centers today.

Figure 2. RDMA/TCP directly places the data from the memory of one host directly into the memory of another host without the need for multiple buffer copies or CPU intervention.



Benefits of RDMA/TCP

RDMA/TCP allows many classes of traffic (networking, I/O, storage, and interprocessor messaging) to share the same physical interconnect, enabling it to become the single unifying data center fabric. RDMA/TCP provides more efficient network communications, which can increase the scalability of CPU-bound applications. RDMA/TCP also leverages existing Ethernet infrastructures and the expertise of IT networking personnel.

InfiniBand

InfiniBand is a switched-fabric interconnect dedicated to and optimized for high-speed, low-latency, point-to-point communication. All computing, storage, and networking devices attach to the InfiniBand fabric using channel adapters that handle communications processing. Servers connect to the fabric using host channel adapters (HCAs) and target devices (storage and communication) connect to the fabric using target channel adapters (TCAs). HCAs and TCAs perform similar functions, but each TCA can be simplified based on the needs of the specific target device.

InfiniBand requires a central subnet manager service to provide topology discovery, interconnect management, performance analysis, and differentiated service configuration. One or more switches can connect any number of servers and target devices together to form large single-subnet solutions. Multiple subnets are expected to be connected through a router device, although a router specification has yet to be defined within the industry.

IP Storage protocols

IP storage protocols extend access to storage resources across local, metropolitan, and wide area networks. The functions of IP storage protocols may differ from each other, but they commonly provide block-level access to storage resources over IP networks. The following sections describe the iSCSI and FCIP protocols and their usage in more detail.

iSCSI protocol

The iSCSI protocol defines the rules and processes for transporting SCSI (block-level) data over a TCP/IP network. The iSCSI standard was developed by the Internet Engineering Task Force (IETF[1]).

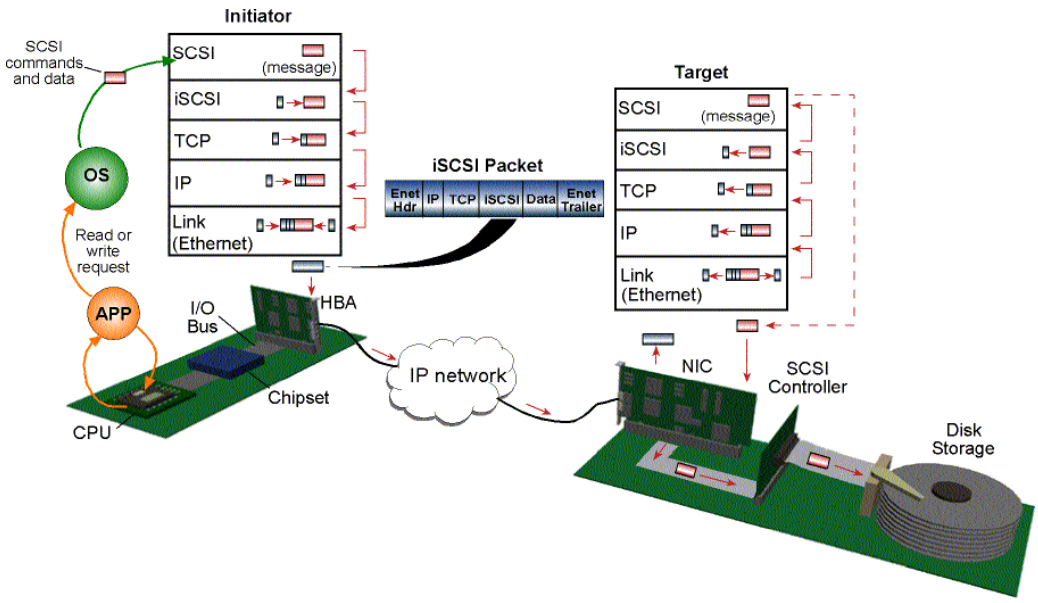
The intent of iSCSI is to run storage traffic over pervasive IP networks, which are primarily deployed over Ethernet infrastructures. Ethernet fabrics are typically organized into virtual LANs (VLANs). A VLAN is a logical workgroup of nodes (on the same or different physical LAN segments) that can communicate securely with each other as if they were all on the same physical LAN segment. VLANs allow IT personnel to partition or confine network traffic to a particular workgroup. This ensures secure, separate bandwidth for storage traffic over existing network infrastructures.

iSCSI follows the traditional SCSI architectural model, which is based on message exchange between an initiator and a target. In the SCSI model, initiators and targets are identified by a unique SCSI device name. Because iSCSI transport occurs over a network fabric instead of a direct cable connection, the initiator and target have IP addresses associated with their iSCSI names.

Figure 3 illustrates a message exchange between an initiator and a target. The process begins when an application sends a request to the OS to read or write data. The OS generates the appropriate SCSI commands and data request in the form of a message. Before the message can be sent over the network, it is processed through the iSCSI and TCP/IP protocol stack to split it into IP packets and to attach routing, error checking, and control information. This can be accomplished using software, or it can be offloaded to the host NIC (called the host bus adapter or HBA). The HBA transmits the packets over the IP network. When the packets reach the target device, they go through a reverse process to reassemble the message, which is then moved to the SCSI controller. The SCSI controller fulfills the request by writing data to or reading data from the target device. If it is a read transaction, the target returns data to the initiator using the iSCSI protocol.

[1] For more information on the IETF, go to http://www.ietf.org/.

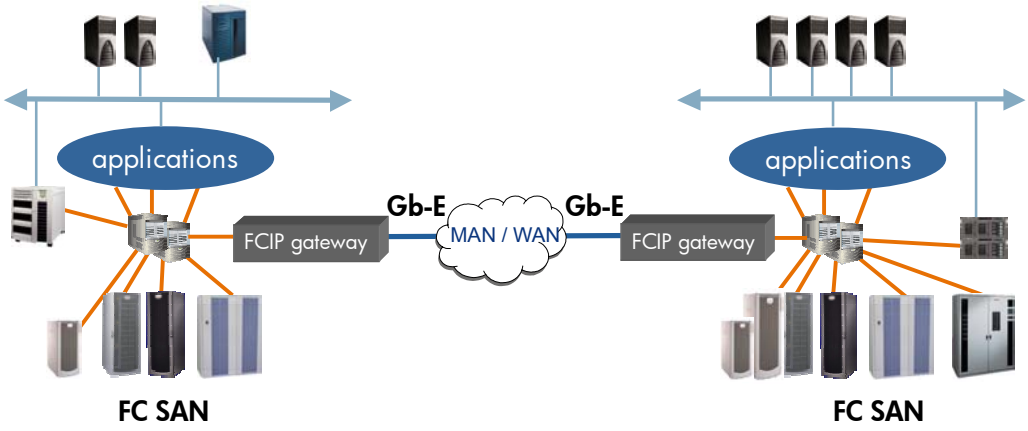
Figure 3. message exchange between an initiator and target using the iSCSI protocol model.



FCIP

Some enterprises use multiple FC SANs at different sites and these SANs are not connected as a single storage network. By linking these geographically separated SAN islands, customers can improve data assurance and accessibility across the enterprise. For customers needing to link existing FC SAN islands, FCIP technology is an option. FCIP encapsulates FC protocol data into IP packets and tunnels within TCP/IP. This capability allows native FC SANs to communicate with each other across Gigabit and 10-Gigabit Ethernet networks and allows all FC services to remain intact. FCIP requires gateway devices (bridges) that translate between FC and FCIP protocols (Figure 4).

Figure 4. FCIP gateways (bridges) tunnel through TCP/IP so that FC SANs can communicate.



Benefits of IP storage

IP storage protocols enable the following improvements in economics, operating distance, manageability, and security of NAS and SANs.

- · cost – For targeted applications, IP storage protocols have the potential to achieve a lower total cost of ownership than other network storage solutions. iSCSI and FCIP leverage existing network infrastructures and skilled network specialists, avoiding the need to train staff on another fabric technology.
- · distance – Because IP storage protocols can leverage widely available IP and Ethernet infrastructures, they offer the potential for globally secure connectivity over public and private IP networks.
- · manageability – Customers can leverage their existing IP network management tools and collectively manage NAS and SAN capacity over an IP network.
- · simplified host connectivity – By leveraging the Ethernet connections commonly used in servers today, iSCSI can reduce the need for specialized and dedicated storage networks for some applications.
- · security – IP networks have an existing security infrastructure (encryption and authentication) that can enhance the viability of using IP storage protocols for remote back up and disaster recovery applications.

Benefits of fabric technologies

Fabric technologies can support multiple classes of data traffic—computing, storage, and networking—on one interconnect standard. In other words, a single fabric port provides the same functionality as a ServerNet adapter, a Fibre Channel adapter, and an Ethernet NIC combined. In high-density computing environments, this will greatly reduce cabling complexity and simplify the deployment of blade servers. Fabric partitioning capability will allow data centers to create virtual systems across the network infrastructure. For example, data centers will be able to virtually partition and repartition hundreds or thousands of blade servers for dedicated tasks such as static web hosting, database applications, and high performance technical computing (Figure 5).

Figure 5. Fabrics will enable data centers to achieve higher CPU densities and better computing resource efficiency with blade servers.

By enabling the consolidation of computing, storage, and networking resources, fabric technologies will reduce the complexity of the data center. Other significant benefits of fabrics—scalable bandwidth, fault tolerance, and partitioning of resources—are described below.

Scalable bandwidth

Switched fabrics provide dedicated, collision-free communication between nodes with support for multiple simultaneous connections. This gives the switched fabric the ability to scale as more nodes are connected to it, with virtually no increase in latency.

Fault tolerance

Because point-to-point connections operate independently of each other, a failure in one connection doesn't affect other connections. In addition, tasks such as troubleshooting, upgrading, or replacing resources like storage devices will become "hot add" events. Some hot add events, such as adding storage, will require the application to be restarted or the server to be rebooted.

Partitioning

Fabric partitioning enables data centers to isolate the traffic of different applications or organizations from each other so they can share fabric resources securely. Partitioning of specific data center resources can control or limit access to multiple business processes (web servers, database servers, and application servers) within the organization. This type of resource partitioning prevents the data traffic for one business process from compromising that of another process in regards to performance and security.

Fabric positioning

Box-to-box interconnects include networking, interprocessor communication, and network storage technologies. InfiniBand and RDMA/TCP are the leading fabric technologies. InfiniBand is emerging as a specialized data center server-to-server fabric where extremely low latencies are required. Targeted fabric solutions must coexist with Ethernet-based solutions to provide a cost-effective and

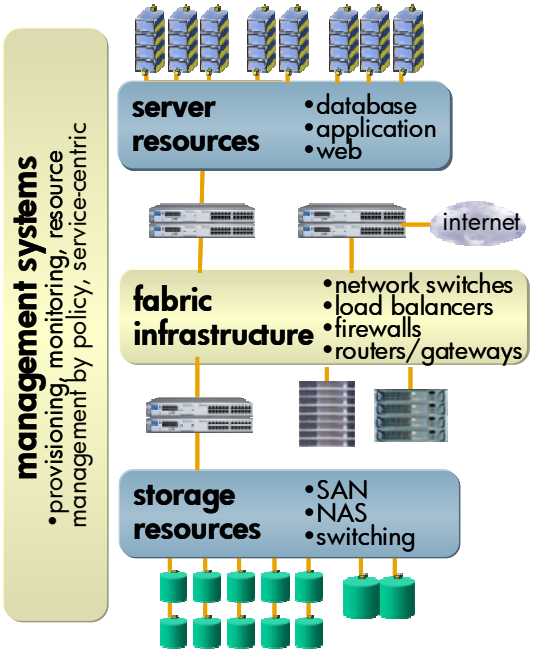
adaptive data center infrastructure. RDMA/TCP will be a unified fabric technology that can be used for networking, storage, remote management, and cluster communications. It will provide breakthrough economics and broad industry adoption to enable a larger number of fabric solutions.

The most widely adopted network storage interconnect is Fibre Channel. For the near future, latency-sensitive applications will require the high performance of Fibre Channel. The emergence of 10-Gigabit Ethernet and further development of offload adapters will increase the applicability of Ethernet (iSCSI) and FCIP as network storage alternatives.

Most likely data center implementations

With multiple existing and emerging fabric technologies for larger data centers, the most likely fabric implementations will be heterogeneous in nature. Today, Ethernet is the pervasive network interconnect and Fibre Channel is the predominant storage area network. Technologies such as RDMA/TCP will provide breakthrough fabric economics and broad industry adoption to enable a larger number of customer solutions. InfiniBand will be used where its unique capability is critical to a customer's success, such as specialized applications where performance and availability requirements justify a more costly, dedicated infrastructure (see Figure 6).

Figure 6. Dedicated infrastructure.



HP believes there will be two waves of adoption for fabric computing. The first wave will include early adopters who have a critical business need for a fabric solution. Because InfiniBand will be the first solution available, these customers will evaluate InfiniBand for their specialized needs. The second, much larger wave will occur soon after fabric computing becomes more accepted and migrates from the early adopters into the mainstream—as Ethernet-based solutions using RDMA/TCP become available. The coexistence of targeted InfiniBand solutions (applications requiring higher speeds and lower latency) and pervasive Ethernet-based solutions will provide a cost-effective and adaptive infrastructure for future data center solutions.

Conclusion

Many of the concepts and technologies used in fabric interconnects have been implemented for years in high-end servers such as HP NonStop and SuperDome servers. Because new fabric architectures are needed to meet the future needs of customers, HP is leading the development of future fabric interconnect technologies. HP is leading the development of RDMA/TCP technologies. The breakthrough economics and broad adoption of Ethernet, along with RDMA capabilities, open the possibility of truly pervasive fabrics in the future. By leveraging HP's experience, emerging fabric technologies will be more cost effective and able to compete in an industry-standard marketplace.

HP is also a founding member of the InfiniBand™ Trade Association. HP has contributed to the InfiniBand specification development and has helped to preserve the core fundamentals that customers value, such as investment protection and compatibility with existing technologies. The potential benefits of fabric computing are enormous, but fabric technologies are still evolving. HP continues to champion these technologies and apply our expertise in the high volume servers and Ethernet markets to take fabric computing into the mainstream.

Call to action

To help us better understand and meet your needs for ISS technology information, please evaluate this paper by completing the short survey at http://www.zoomerang.com/survey.zgi?XCCMYAX2QM746PKDNYQ3KD95.

Note: This URL will be active through 31 January 2004. Please send questions and further comments about this paper to: TechCom@HP.com.

TC031009TB, 10/2003