STORAGE AREA NETWORK

Technology Overview: Why Fibre Channel over Ethernet?

Fibre Channel over Ethernet (FCoE) is not a replacement for conventional Fibre Channel (FC), but is an extension of Fibre Channel over a different link layer transport.

BROCADE

CONTENTS

INTRODUCTION

Over the past ten years, Fibre Channel (FC) has become the technology of choice for Storage Area Networks (SANs) worldwide. In the process, it has generated a wide range of new storage solutions, including higher-performance block transport, high-availability storage access, streamlined data center operations for backup and data protection, and higher-level storage services based on virtualization and advanced management utilities. Other technical innovations, such as InfiniBand, Network-Attached Storage (NAS), and iSCSI, however, periodically spark temporary debate about Fibre Channel's future and its viability as the next generation SAN transport. As of mid-2008, the Fibre Channel installed base is estimated to be well over 10 million ports.

Without compelling economic or functional advantages, new technologies rarely displace successful incumbents. Due to its economies of scale, for example, Ethernet displaced Token Ring for Local Area Network (LAN) transport despite Token Ring's higher performance (16 Mbit/sec versus 10 Mbit/sec at the time) and more robust operation. Asynchronous Transfer Mode (ATM), on the other hand, was unable to displace Ethernet to the desktop primarily due to its inability to co-opt Ethernet's large installed base. ATM's LAN Emulation (LANE) was simply too problematic. And although InfiniBand has demonstrated its end-user value for high-performance server clustering, it has been unable to compete as a transport for either local or storage area networking. Because only a few smaller vendors have introduced InfiniBand storage arrays, Fibre Channel continues to be the connectivity of choice for data center storage applications. In addition, InfiniBand's inability to demonstrate value as a LAN transport and its unique cabling scheme and limited distance support at high speeds have discouraged IT administrators from fork-lifting their existing networks.

THE SUCCESS OF FIBRE CHANNEL

Fibre Channel is a successful technology because it has solved many of the difficult problems associated with high-performance block data transport. It is, after all, *a channel architecture*, modeled after data center mainframe environments. A channel is characterized by high bandwidth and low protocol overhead to maximize efficient delivery of massive amounts of data within the circumference of the data center. To maintain consistent performance, Fibre Channel has internal mechanisms, such as buffer-to-buffer credits, to minimize the potential effects of network congestion. If a frame is lost, Fibre Channel does not stop to recover the individual frame as Transmission Control Protocol (TCP) does, but simply retransmits the entire sequence of frames at multi-gigabit speeds. For SANs, Fibre Channel pioneered storage-specific mechanisms, such as automated addressing, device discovery, fabric building, and state change notifications to facilitate transactions between initiators (servers) and targets (storage systems).

Fibre Channel has also introduced a set of higher-level services for scaling reliable and highly available fabrics. Fabric routing protocols, policy-based routing, hardware-based trunking, Virtual Fabrics, fabric security, and fault isolation have been built on top of a foundation of stable transport. New fabric-based application services for storage virtualization and data protection are further enhancing simplification and automation of storage administration. Collectively, Fibre Channel standards and standards-compliant products are optimized to deliver maximum performance and maximum availability of storage data. As a consequence, *Fibre Channel SANs are now powering every significant enterprise and institution worldwide.*

FCoE Standards Initiative

Recently, a new Fibre Channel standards initiative was created for running the Fibre Channel protocol over Ethernet (FCoE). Given the substantial investment in engineering resources, technical volunteers, and product development required to create a new standard and technology, and given the success Fibre Channel already has demonstrated for data center SANs, why is FCoE needed? Some industry observers have speculated that FCoE is an attempt by Fibre Channel vendors to compete with iSCSI, which, after all, also transports block storage data over Ethernet. When FCoE is compared to iSCSI, however, we see that the two protocols solve very different problems. iSCSI uses TCP/IP to move block storage data over potentially lossy and congested Local and Wide Area Networks (LANs and WANs) and is used primarily for low- and moderate-performance applications. The FCoE initiative, in contrast, intends to utilize new Ethernet extensions that replicate the reliability and efficiency that Fibre Channel has already demonstrated for data center applications. These new Ethernet enhancements are predicated on 10 Gbit/sec performance and are sometimes referred to as Converged Enhanced Ethernet (CEE).

FCoE is not a replacement for conventional Fibre Channel but is an extension of Fibre Channel over a different link layer transport. Enabling an enhanced Ethernet to carry both Fibre Channel storage data as well as other data types, for example, file data, Remote Direct Memory Access (RDMA), LAN traffic, and VoIP, will allow customers to simplify server connectivity and still retain the performance and reliability required for storage transactions. Instead of provisioning a server with dual-redundant Ethernet and Fibre Channel ports (a total of 4 ports), servers can be configured with 2 CEE-enabled 10 Gbit/sec Ethernet ports. For blade server installations, in particular, this reduction in the number of interfaces greatly simplifies deployment and ongoing management of the cable plant. *The main value proposition of FCoE is therefore the ability to streamline server connectivity using CEE-enabled Ethernet while retaining the channel characteristics of conventional Fibre Channel SANs*, as shown in Figure 1.

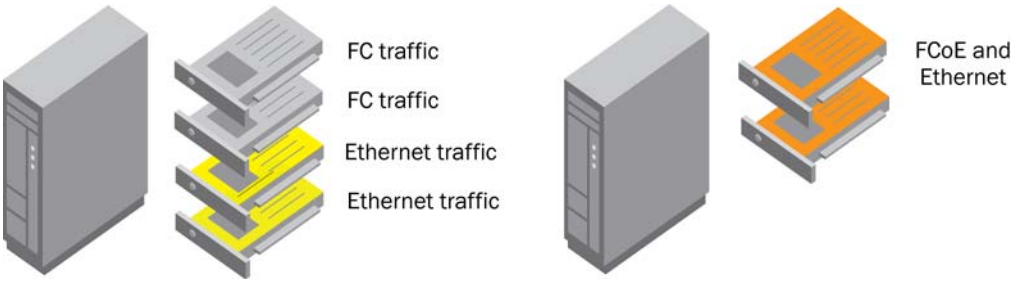

Figure 1. Consolidated server network interface using FCoE and CEE

Given the more rigorous requirements of storage transactions, FCoE is predicated on a new, hardened Ethernet transport that is both low loss and deterministic. Without the enhancements of CEE, standard Ethernet is too unreliable to support high-performance block storage transactions. Unlike conventional Ethernet, CEE provides much more robust congestion management and high-availability features characteristic of data center Fibre Channel.

The FCoE initiative is being developed in the ANSI T11 Technical Committee, which deals with FC-specific issues and is included in a new Fibre Channel Backbone Generation 5 (FC-BB-5) specification. Because FCoE takes advantage of further enhancements to Ethernet, close collaboration is required between ANSI T11 and the Institute of Electrical and Electronics Engineers (IEEE), which governs Ethernet and the new CEE standards.

One Big Network?

The ability to support both Fibre Channel and IP-based protocols over a common link layer transport may give the impression that customers will now be able to deploy a single large network to support all their applications and messaging traffic. The reality, though, is that the vast majority of data center administrators will continue to support one network for storage and another for messaging and data communications traffic. Why? Storage transactions are so mission critical and storage data integrity is so sacrosanct that few customers would risk merging their LANs and SANs, even if that were enabled by CEE-capable Ethernet switches.

Vendors who advocate a single homogeneous Ethernet network fail to understand the much more rigorous requirements of storage data transport. While temporary outages or disruptions are commonplace in the LAN and WAN world, any unplanned downtime or disruption is anathema in a storage networking environment. Over the past ten years, storage area networking has introduced new levels of resiliency and availability not found in conventional networking. Consequently, the admixture of LAN and SAN traffic on the same extended infrastructure would pose too great a risk to the stability required for storage data transactions. Although the unified network marketing message has some appeal in terms of streamlining deployment and management, it does not align with the practical realities of data storage.

FCoE enables a simplified connection on the server (initiator) side. Storage systems (targets), however, will for the foreseeable future remain native Fibre Channel. The storage network, therefore, can incorporate different link layer transports where appropriate, with Brocade® FCoE-enabled switches and directors providing the conversion between FCoE and Fibre Channel fabrics. Because all storage transactions from servers to storage are based on the Fibre Channel protocol, the inherent performance, stability and high availability features of Fibre Channel are maintained on both CEE and native FC links. It is therefore more accurate to describe this new architecture not as network unification but as *network convergence*, or the point at which separate SAN and LAN protocols converge on a common transport for simplified server connectivity. The new Data Center Fabric (DCF) will consist of traditional FC and FCoE supported on Brocade directors and the Brocade DCX Backbone platform, while the conventional outward-facing network for messaging and data communications will continue to be supported on LAN and WAN technologies.

Maintaining the Channel

Both Fibre Channel and Ethernet transports are link layer (Layer 2) protocols. In the Open Systems Interconnection (OSI) Reference Model, Layer 1 is the physical medium that supports network signaling. Layer 2 is the framing protocol that rides immediately on the medium, while upper layers handle higher-level services, such as network routing and session management. Because each additional upper layer imposes more protocol processing and overhead, Layer 2 is the most streamlined means to quickly transport data from one network node to another.

Fibre Channel was originally designed as a link layer transport protocol specifically to maintain the efficiencies of a data center channel. This has several implications. First, at gigabit and multi-gigabit speeds, a robust flow control mechanism is required to avoid frame loss due to congestion. Fibre Channel addresses the flow control issue with buffer-to-buffer credits. A device cannot send additional frames until the recipient's buffers are replenished and Receiver-Ready (R_RDY) signals are issued to the sender. Secondly, an FC fabric is essentially a single subnet with transactions between initiators and targets bounded by the data center. Although Fibre Channel now has auxiliary routing capability for SAN-to-SAN communication, Fibre Channel routing uses Network Address Translation (NAT) instead of Layer 3 routing overhead.

Over the years, Fibre Channel has evolved higher-level functions tailored to storage requirements:

- The Simple Name Server (SNS) hosted by every fabric switch, for example, provides device discovery for initiators seeking target resources.
- Zoning (based on port World Wide Name or Domain, Port identification) enables segregation of storage relationships and prevents unauthorized servers from communicating with designated storage assets.
- Registered State Change Notifications (RSCNs) provide a means to alert servers to the arrival or departure of storage systems on the fabric.
- The Fabric Shortest Path First (FSPF) protocol establishes optimum paths in a multi-switch fabric and allows multiple trunked links to increase bandwidth between switches.
- Fabric routing with fault isolation provides sharing of resources between autonomous SANs.
- Virtual Fabrics enable a common SAN infrastructure to be shared by separate departments or applications without infringing on each other.

To preserve the channel attributes and storage-centric services of conventional Fibre Channel, FCoE requires significant enhancements to conventional Ethernet networking and integrated controllers to provide device discovery, notifications, security, and other advanced storage services. Assuming that Ethernet can be hardened for data center use, FCoE would be a fairly straightforward means to wrap FC frames in Ethernet for frame translation between FCoE initiators and FC targets. To be viable for customer implementation, however, the rich set of Fibre Channel advanced fabric services must be preserved.

FCoE Switches

FCoE switches or Fibre Channel Forwarders (FCFs) provide the connectivity between FCoE initiators and conventional Fibre Channel fabrics. FCFs therefore offer both CEE ports and native FC ports for both device and switch-to-switch fabric connections. For CEE connectivity, FCoE device ports are VF_Ports (corresponding to Fibre Channel F_Ports), while switch-to-switch ports are VE_Ports (corresponding to conventional Fibre Channel E_Ports). Brocade FCF switches can be embedded as blade server FCoE switch modules, installed as rack-mount FCFs, or as FCoE blades inserted into the Brocade DCX Backbone. Customers will thus have the flexibility to deploy both native Fibre Channel and FCoE that aligns with their server strategy and platform mix over time.

Avoiding Packet Loss

One of the first challenges for FCoE development is to replicate the flow control provided by native Fibre Channel buffer-to-buffer credits. Ethernet switches do not have a comparable buffer-to-buffer mechanism, but Ethernet standards support a Media Access Control (MAC) frame to pace incoming traffic. The IEEE 802.3x flow control standard is based on a PAUSE frame, which causes a sender to hold off additional transmission until a specified time has elapsed. If a receiving device clears its buffers before that time has elapsed, it can reissue a PAUSE frame with pause time set to 0 (zero), which enables the sender to resume transmission until another PAUSE is received.

Because FCoE transactions must support reads and writes of storage data, all end devices and Ethernet switches in the network storage paths must support bidirectional 802.3x. While perhaps not as optimal as buffer-to-buffer credits, IEEE 802.3x PAUSE frames can provide comparable functionality in pacing storage traffic and helping minimize congestion and frame loss due to buffer overruns.

The conventional use of IEEE 802.3x PAUSE operates at the port level. This could prove problematic for FCoE, however, if other protocols on the same CEE port are the actual source of congestion. Consequently, successful implementation of PAUSE frames for storage environments must be more granular and operate on a per-application virtual channel basis. Because Ethernet does provide the ability to set priority bits in the Ethernet header, it is possible to combine 802.1Q prioritization with the 802.3x PAUSE mechanism. The priority-based flow control (PFC) CEE initiative is being worked in the 802.1Qbb workgroup and will provide the means to fine tune congestion control to the application layer.

In addition, 802.1Qbb can leverage prioritization to establish bandwidth allocation on a per-application basis. Time-sensitive applications such as inter-process communications (IPC) can be given a higher percentage of available bandwidth as needed while other applications are assured portions of the remaining available bandwidth. The enhanced transmission selection (ETS) algorithm will strengthen the ability of FCoE to reliably use Ethernet as a transport layer and minimize the chance of link congestion and frame loss.

Ethernet Congestion Management (ECM) uses another technique for implementing reliable flow control on an end-to-end basis. The IEEE 802.1Qau Congestion Notification Group is developing ways to mitigate congestion by reflecting frames back to their source when a congestion point occurs. When a host sees its own frames being reflected by a downstream switch, it will slow its own frame transmission until no more reflected frames are seen. Analogous to Backward Explicit Congestion Notification (BECN) in WAN technologies, ECM can slow the pace of issued frames until a congestion point is cleared.

Mapping Fibre Channel to Ethernet

FCoE must also resolve the disparity between Ethernet and FC frame sizes. A typical Ethernet maximum frame size is 1518 bytes. A typical FC maximum frame size is about 2112 bytes. Wrapping FC frames in Ethernet would therefore require segmentation of frames on the sending side and reassembly on the receiving side. This in turn would incur more processing overhead and undermine the FCoE effort to preserve streamlined channel performance end to end.

To align FC and Ethernet frame sizes, a larger Ethernet frame is needed. Although not an official IEEE standard, a de facto standard called “jumbo frames” allows for Ethernet frames up to about 9 kilobytes in length. The caveat for use of jumbo frames is that all Ethernet switches and end devices must support a common jumbo frame format.

Use of a maximum jumbo frame size of 9 kilobytes would allow four FC frames to be encapsulated in a single Ethernet frame. This would, however, complicate Fibre Channel link layer recovery as well as buffer flow control using 802.3x PAUSE commands. Instead, FCoE encapsulates a complete FC frame into one jumbo Ethernet frame, as shown in Figure 2. Because FC frames may include extended and optional headers or Virtual Fabric tagging information, the jumbo Ethernet frame size is not fixed and may vary depending on the requirements of the encapsulated FC frame.

Figure 2. Encapsulating a single FC frame in Ethernet

FCoE frames are native Layer 2 Ethernet frames with conventional six-byte destination and source MAC (media access control) addresses. The MAC addresses, however, are storage agnostic and are used only to switch frames from source to destination. The FCoE content of the frame retains Fibre Channel addressing required for storage transactions, and so some means is required to map Fibre Channel IDs (FCIDs) to Ethernet MAC addresses. The FCoE Initialization Protocol (FIP) allows for either fabric-provided MAC addresses or server-provided MAC addresses and FCoE switches (known as Fibre Channel Forwarders or FCF) maintain the mapping between Ethernet MAC addresses and the corresponding Fibre Channel addresses.

Redundant Pathing and Failover

The high availability characteristic of Fibre Channel SANs is typically based on flat or core-edge topologies, which provide redundant pathing from initiators to targets. The loss of a single Host Bus Adapter (HBA), link, switch port, switch, or storage port triggers a failover from primary to secondary paths. In some implementations, both paths are active, allowing for higher performance as well as availability. For FC fabrics, the Fabric Shortest Path First (FSPF) protocol is used to determine the optimum path between fabric switches based on bandwidth of each Inter-Switch Link (ISL) and traffic load.

For FCoE, the Ethernet infrastructure must provide comparable resiliency to ensure uninterrupted storage access. When multiple Ethernet switches are connected via ISLs (for example, in a full mesh topology), the IEEE 802.1D Rapid Spanning Tree Protocol (RSTP) is used to establish primary paths through the network and to avoid loops that would route frames in endless circles. Active bridging ports between switches are put into a forwarding state; inactive failover bridging ports are put into a blocking state. A blocked link, however, cannot be used for data transport and consequently every blocked link in the mesh network represents an unused and idle resource. Rapid Spanning Tree monitors the status of all bridging ports through control frames or Bridge Protocol Data Units (BPDUs). If a link, bridge port, or switch fails, the RSTP activates the requisite failover bridging ports to establish alternate pathing through the network.

An additional enhancement to alternate pathing in Ethernet networks has been defined by the IEEE 802.1s Multiple Spanning Tree Protocol (MSTP) and merged into the IEEE 802.1Q-2003 specification for Virtual LANs (VLANs). Analogous to hard zoning in Fibre Channel, VLAN tagging enables up to 4,096 segregated groups of nodes to coexist on a common Ethernet infrastructure. The MSTP enhancement to Spanning Tree allows for a separate Spanning Tree for each VLAN group. Consequently, a bridging port that is in blocking mode for one VLAN could be put into forwarding mode for another VLAN, allowing for fuller utilization of all network connectivity.

Even with MSTP enhancements, the RSTP reliance on simple forwarding and blocking states inevitably results in underutilized network links. More complex Layer 3 routing protocols, such as Open Shortest Path First (OSPF), calculate optimum paths between end nodes based on hop count, bandwidth, latency ,and other metrics and can enable load balancing along multiple paths. As a Layer 2 protocol, RSTP has not been able support this richer functionality and still maintain backward compatibility. Additional work is needed to find ways to bring Layer 3 routing functionality, such as load balancing, multiple points of attachment (for example, a node with two active links into the same Ethernet segment), broadcast, and multicast into Layer 2 Ethernet networks. Given the existing Fibre Channel support for trunking, load balancing, and multiple points of attachment, the continued evolution of RSTP or its next generation replacement would be required to make Ethernet more storage friendly and simplify FCoE deployment. *There is little benefit in using Ethernet as a storage transport if advanced services already provided by Fibre Channel are sacrificed.*

FCoE, iSCSI, AND FCIP

FCoE, iSCSI, and FC over IP (FCIP) are all storage protocols capable of transporting block storage data over Ethernet. However, each has been developed with very different goals and design criteria. FCoE is being developed as a streamlined data center storage protocol, which leverages the minimalist Layer 2 protocol efficiency of Fibre Channel and data center Ethernet. iSCSI was designed to reliably transport block storage data over any IP infrastructure, including LANs and WANs. As shown in Figure 3, iSCSI relies on the entire TCP/IP protocol stack at Layers 3 and above to support routing and packet recovery, and so can be used in potentially lossy networks. FCIP, in contrast, was designed as a simple tunneling protocol to link Fibre Channel SANs over distance on IP networks. Used primarily for remote storage access and disaster recovery, FCIP provides SAN-to-SAN connectivity over IP, but both end points are FC devices. Like iSCSI, FCIP carries the overhead of TCP/IP processing, which is essential for maintaining data integrity over long-distance storage applications. By linking FC SANs over distance, FCIP facilitates disaster recovery implementations that can span thousands of miles.

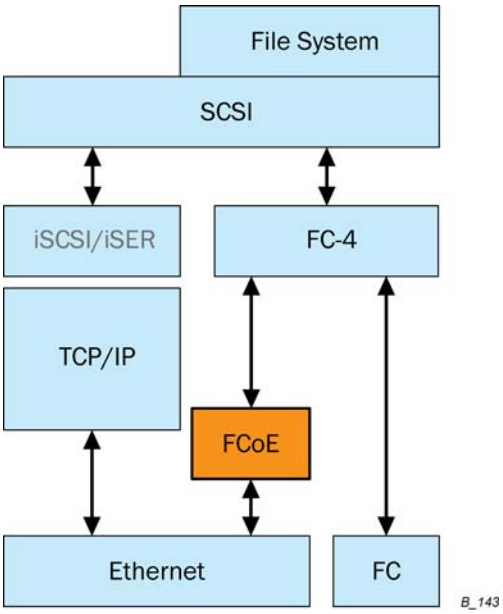

Figure 3. FCoE and iSCSI protocol stacks

A primary contribution of iSCSI is its ability to economically utilize free device drivers, commodity Ethernet Network Interface Cards (NICs), and commodity Ethernet switches and IP routers to transport SCSI block data between servers and storage. Although the per-server attachment and network infrastructure are inexpensive, iSCSI storage targets costs can vary depending on whether inexpensive disk drives are used and whether hardware-based or software-based controllers are used. Because there are no native iSCSI disk drives, iSCSI targets must rely on some form of protocol bridging (either iSCSI to SAS or SATA, or iSCSI to Fibre Channel) controller to store and retrieve block data. Thus there is no iSCSI equivalent to JBODs (just a bunch of disks) sometimes used in departmental Fibre Channel SANs.

At 1 Gbit/sec Ethernet, iSCSI is an affordable means to integrate lower performance, second-tier servers into existing FC data center SANs via gateways, or to provide shared storage for departmental use. At 10 Gbit/sec Ethernet, however, iSCSI loses its much publicized cost advantage. Use of 10 Gbit/sec Ethernet at the server attachment implies that the applications being hosted require high performance and reliability. Although standard NICs can be used at 1 Gbit/sec, server performance with iSCSI at 10 Gbit/sec is enhanced by auxiliary components, such as TCP off-load (TOE)-enabled adapters with iSCSI Extensions for RDMA (iSER) logic to avoid multiple memory copies of SCSI data

from the interface to application memory. Dedicated 10 Gbit/sec iSCSI adapters add significant cost per server attachment compared to 8 Gbit/sec FC HBAs, however, and undermine the value proposition of iSCSI at 1 Gbit/sec.

Although iSCSI-to-FC gateways enable iSCSI initiators to access FC storage targets, the requisite protocol conversion is considerably more complex than FCoE streamlined frame mapping to Fibre Channel. For iSCSI gateways, a complete address translation is required between iSCSI and FC address conventions. In addition, the gateway must proxy virtual FC initiators and virtual iSCSI targets and terminate sessions within the gateway between the two protocols. If the objective is to have Ethernet-attached servers access FC SAN targets, FCoE will require less protocol overhead and processing latency to span between Ethernet and Fibre Channel transports.

SUMMARY

The large installed base and maturity of Fibre Channel technology has generated a wide spectrum of storage-specific features and management tools to facilitate robust deployment of shared storage in the data center. The Converged Enhanced Ethernet (CEE) initiative will enable customers to combine storage, messaging traffic, VoIP, video, and other data on a common data center Ethernet infrastructure. FCoE is the component technology that enables highly efficient block storage over Ethernet for consolidating server network connectivity. By enabling customers to deploy a single server interface for multiple data types, FCoE and CEE will simplify both deployment and management of server network connectivity, while maintaining the high availability and robustness required for storage transactions. FCoE is thus not a replacement for, but an extension of, Fibre Channel and is intended to coexist with existing FC SANs.

Because FCoE takes advantage of further enhancements to CEE, its development will require close coordination of both Fibre Channel and Ethernet technologists and standards bodies. Although link layer issues, such as flow control and the limitations of Ethernet spanning tree protocols, present significant challenges, the more arduous task will be to create FCoE solutions that preserve the rich set of Fibre Channel advanced services that customers are deploying productively today. Even at 10 Gbit/sec, today's Ethernet technology will require substantial work to be suitable for data center storage applications. As a pioneer of Fibre Channel fabric technology, Brocade is bringing its expertise to the FCoE initiative to streamline the server network interface while preserving data center performance, reliability, and the proven benefits of advanced storage services.