Redpaper

Jon Tate

An Introduction to Fibre Channel over Ethernet, and Fibre Channel over Convergence Enhanced Ethernet

Introduction

It is a matter of fact that Fibre Channel (FC) is the prevalent technology standard in the Storage Area Network (SAN) data center environment. This standard has spawned a multitude of FC-based solutions that have paved the way for high performance, high availability, and the highly efficient transport and management of data.

Without going into any great historical depth, the arrival of FC did solve many of the problems that existed in the data center. These problems included distance, performance, bandwidth, and overhead issues. The manner in which FC is implemented and its inherent functionality are designed to mitigate against the loss of data, against network congestion, and at the same time, provide a highly available and high performing network.

That is not to say that FC is the only technology choice available to us, nor does it belittle any other technologies that also have their place within the data center environment. We have iSCSI, InfiniBand (IB), Network Attached Storage (NAS), to name but three, which are almost guaranteed to spark the technological debate as to which should be the preferred data center choice, and they regularly bring up the possibility that FC is no longer economically viable, and that it is soon to be replaced. Estimates that there are over 10 million FC ports installed around the world indicate that FC is likely to be around for quite a while.

However, there is a new protocol emerging. This new protocol, Fibre Channel over Ethernet (FCoE), which is being developed within T11 as part of the Fibre Channel Backbone 5 (FC-BB-5) project, is not meant to displace or replace FC.

Technical Committee T11

This is the committee within the International Committee for Information Technology Standards (INCITS) responsible for Fibre Channel Interfaces. T11 (previously known as X3T9.3) has been producing interface standards for high-performance and mass storage applications since the 1970s.

Rather, FCoE is an enhancement that expands FC into the Ethernet by combining two leading-edge technologies (FC and the Ethernet).

Is it FCoE or FCoCEE?

FCoE or FCoCEE? Actually, it is both. FCoE is the standard that is driving convergence and the emergence of FC over the Ethernet. However, in and of itself it (FC over Ethernet) will not be enough to allow for fabric convergence. Without question, it is a move in the correct direction but without any enhancements (we will discuss these enhancements later) Ethernet itself does not meet the requirements for data center convergence. FCoCEE is Fibre Channel over Convergence Enhanced Ethernet (pronounced *eff-see-oh-see*), which is enabled by FCoE.

It is our aim in this IBM® Redpaper to provide a very brief introduction to the concepts and terminology surrounding the combination of FC and Ethernet.

In this paper, when we refer to Enhanced Ethernet, we are referring to an Ethernet that is full duplex and lossless when transporting Fibre Channel frames.

The FC-BB-5 proposal

Before we start to discuss FCoE in any great depth, the original proposal to the T11 technical committee that started the impetus for the creation of the FC-BB-5 standard is:

"This project proposal recommends the development of a set of additional and enhanced mechanisms, services, and protocols to connect Fibre Channel entities over selected non-Fibre Channel protocol infrastructures. Enhancements to Ethernet protocols, such as the Pause mechanism defined in IEEE 802.3-2005, make it possible to define a direct mapping of Fibre Channel over Ethernet (FCoE). This mapping provides several technological benefits over the currently defined Fibre Channel over IP (FCIP) mapping and gives a significant business advantage to Fibre Channel over competing technologies, such as iSCSI, because Fibre Channel provides seamless compatibility with existing storage, drivers, and management tools. The FCoE mapping allows Fibre Channel to be used in Ethernet-based I/O consolidated environments and will be especially useful in both the Data Center and Metro Ethernet environments.

Included within the scope of this project are functions, such as:

- *A direct mapping of Fibre Channel over selected full duplex IEEE 802.3 networks*
- *Any other item as deemed necessary during the development"*

This proposal has now been adopted within the International Committee for Information Technology Standards (INCITS) and is being actively worked on by the Storage technical committee for Fibre Channel Interfaces (T11). You can access the FC-BB-5 project, along with its current status, at:

http://tinyurl.com/3l7u5t

What is Convergence Enhanced Ethernet

As we have stated, the prevalence of FC storage in the world's data centers cannot be ignored when introducing a new technology. The current FC SAN investment must be protected, which is not negotiable. But, this still does not answer the question, "Why converge?" Well, the appeal of the convergence is that the existing FC infrastructure can be maintained, the management model is the same as the existing FC, fewer components will be required, and therefore, there is less power and cooling necessary, giving potential energy savings.

So before we fully answer our question, let us step back and look at a very simple example of the interfaces and networks that exist today. As we touched upon briefly, data centers and applications may all utilize different interfaces or adapters. For example:

- Ethernet (Ethernet network interface card (NIC))
- Fibre Channel (Fibre Channel host bus adapter (HBA))

Figure 1 shows a traditional server setup of today.

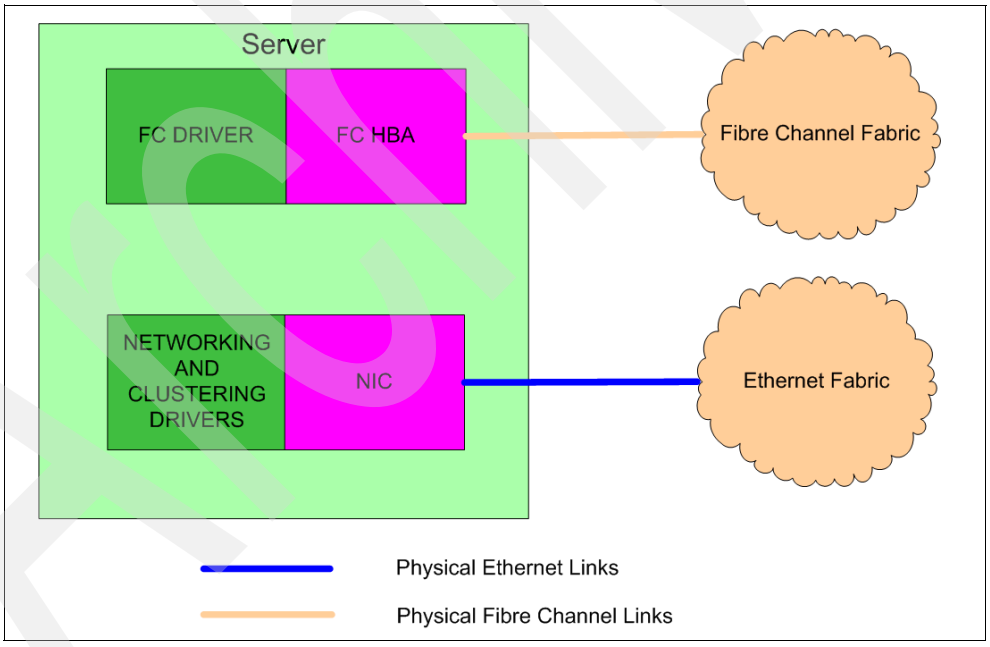

Figure 1 Traditional server of today

Using these examples, it is easy to see that we are presented with different networks. And each of these networks has its own adapters, fabrics, cables, tools, switches, management, and skills needed to maintain it. If somehow, all these of these components were combined, or converged, the potential for reducing cables, adapters, switches, and the skills required is obvious. Replacing multiple networks with one network is becoming closer to reality.

What are the components of a converged network

At a minimum, a converged network requires an adapter at the server that is capable of carrying FC and networking storage, and at the fabric/network level, FCoE capability will be required, which is sometimes referred to as the "*access layer*."

Using our "traditional server of today" diagram, now in Figure 2 we show how a converged network adapter in a server, connected to the Enhanced Ethernet, has the potential to reduce the components required.

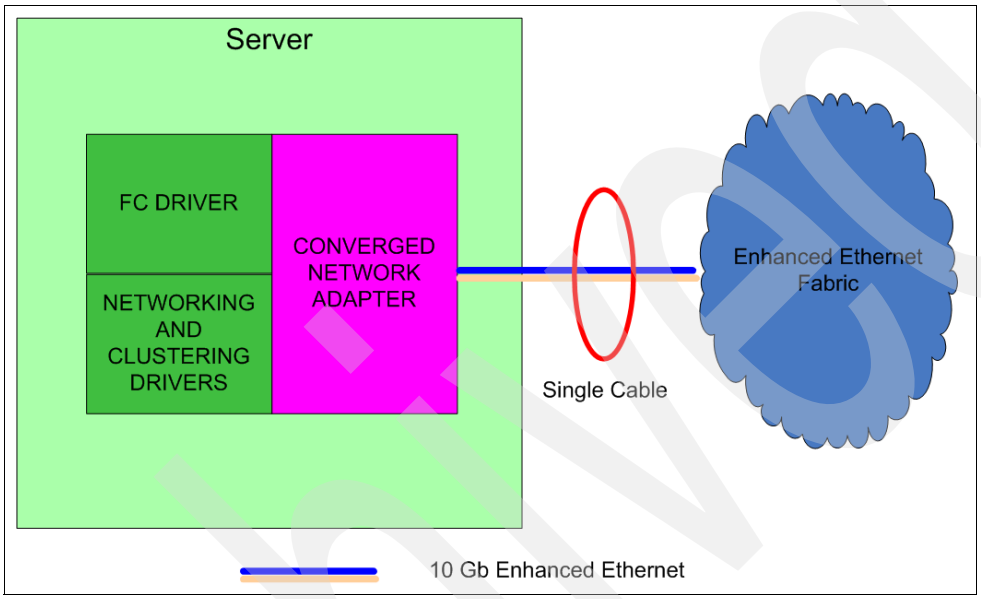

Figure 2 Converged Network Adapter

At the server level, we are already starting to see Converged Network Adapters (CNAs). Using a Fibre Channel driver, the CNA functionally represents a traditional Fibre Channel HBA to the server's operating system. Using NIC or clustering drivers, the CNA functionally represents a traditional networking or clustering device to the server's operating system. The Fibre Channel traffic is encapsulated into FCoE frames (as we describe in the topics that follow) and these FCoE frames are converged with networking or clustering traffic.

Within the fabric, we are beginning to see FCoE-capable switches that can pass Fibre Channel traffic to the attached SANs and Ethernet traffic to the attached Ethernet network. These switches need to be able to support the Enhanced Ethernet, and we discuss those demands in the topics that follow.

The 10 Gigabit Enhanced Ethernet

One of the inhibitors to using Ethernet as the base upon which an FCoE/FCoCEE network is built was the bandwidth limitations of the Ethernet. However, with the emergence of 10 Gbps Ethernet (10 GbE), we now have a base on which all FCoE/FCoCEE solutions will be built. Why? Well, with one large pipe, the value proposition means that storage is not transported to the exclusion of everything else. The large pipe creates a superhighway that allows for Voice over IP (VoIP), video, messaging, and storage to travel over a common Ethernet infrastructure. And it only gets better, and faster, with 40 GbE and 100 GbE in plan for the future.

The other important item to consider is that the FC network is lossless. The Ethernet is not lossless.

Enhanced Ethernet will include new extensions to the existing Ethernet standard that will eliminate the lossy nature of the Ethernet and make 10 GbE a viable storage networking transport. Other enhancements within the Enhanced Ethernet include:

- Congestion notification
- Priority-based flow control
- Enhanced transmission selection
- Data Center Bridging (DCB) Capability Exchange Protocol

We will cover these topics in “Data center bridging” on page 13 after we have introduced the terminology and concepts that make up the Convergence Enhanced Ethernet.

What is FCoE

As its name suggests, it is the transport, or mapping, of encapsulated FC frames over the Ethernet. Very simply, the Ethernet provides the physical interface, and FC provides the transport protocol, giving us an FC frame delivered in an Ethernet frame. Figure 3 shows an encapsulated FC frame within the Ethernet frame.

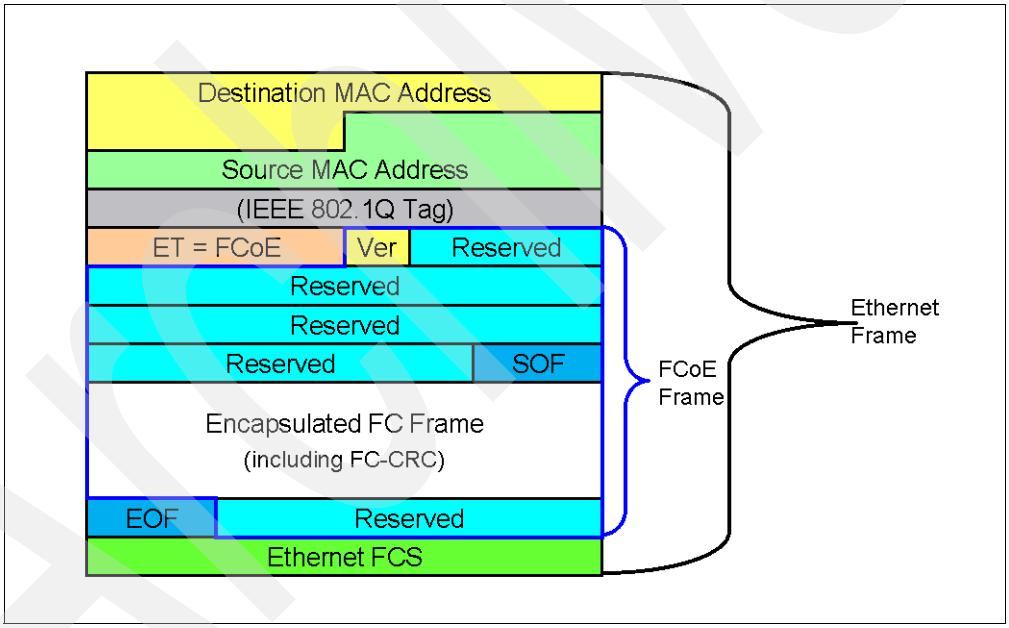

Figure 3 Encapsulated FC frame

Within the Ethernet frame, there is a Destination Media Access Control (MAC) Address and a Source MAC Address (as well as an IEEE 802.1Q Tag), but these components are *not* part of the encapsulation portion of the FCoE frame.

Of particular note is that the FC frame that is encapsulated contains the original 24 byte header and the payload. The reason that the original FC header is passed is to enable seamless processing of the frame without the requirement for a separate gateway.

Protocol stack changes

To send the packets over the network, changes had to be made to the protocol stack. Refer to Figure 4 for the changes that have been made to the stack to accommodate the Enhanced Ethernet (we assume a familiarity with the protocol stack).

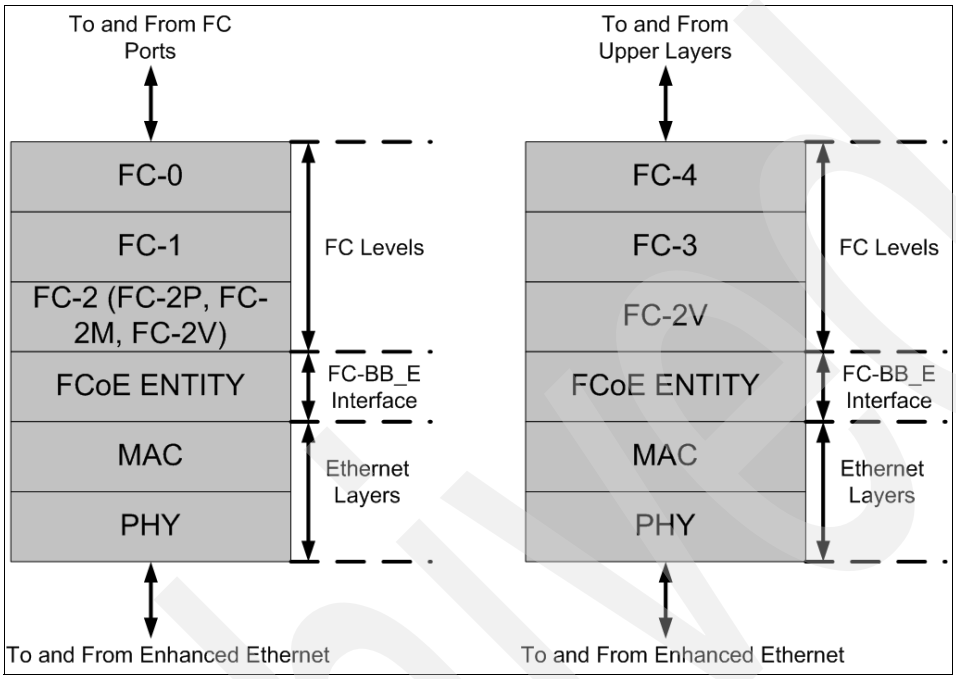

Figure 4 Protocol stack

The bottom two layers of the FC protocol stack have been replaced with their Ethernet protocol stack equivalents. The FC-BB-5 project group members are tasked with the responsibility for ensuring that the existing FC stack remains unaffected and that the work that has already gone into the existing FC stack is neither undermined or lost.

Note that within the FC-BB_E interface is a reference model within FC-BB-5 that defines the mappings for transporting Fibre Channel over Ethernet. Because an Ethernet network can lose frames, it is the extensions to the Ethernet that will allow FCoE to exhibit a lossless and full duplex behavior when carrying Fibre Channel frames.

Frame encapsulation

Figure 5 on page 7 shows the encapsulated frame in a little further detail. Note that the standard Ethernet frame is a maximum of 1518 bytes.

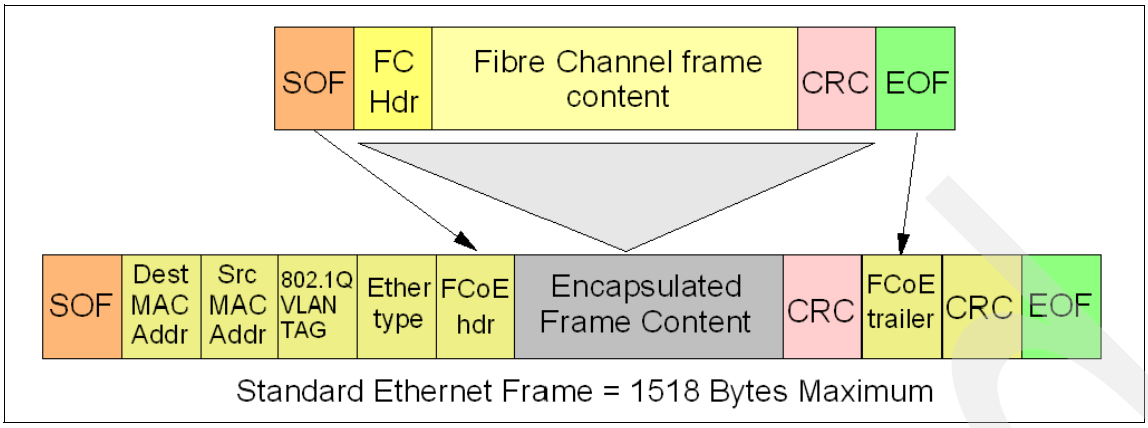

Figure 5 Ethernet frame with encapsulated FC frame (detail)

Contrast that size with the *maximum* FC frame size of 2148 bytes. As you can see, if we were to simply wrap the frames into an Ethernet frame, some form of fragmentation, or segmentation, must occur due to the difference in maximum sizes (1518 compared to 2148). Although this segmentation is possible, it is not desirable because it adds more processing overhead to the operation. After all, the aim of this convergence is to enhance and not add any performance degradation.

Minimum frame size

Additionally, Ethernet frames have a minimum allowable frame size of 64 bytes, and FC has a minimum allowable frame size of 28 bytes. So, some sort of framing standard or format was required to ensure that the most effective solution was found. The working group decided to add padding to all frames to avoid any need for a length field.

The answer to this difference in frame sizes is to increase the size of the Ethernet frame. Most of us are familiar with “jumbo frames.” Fortunately for us, jumbo frames are Ethernet frames with more than 1500 bytes of payload, and they meet our requirements. Conventionally, jumbo frames can carry up to 9000 bytes of payload, but variations exist and you must be careful when using the term jumbo frames. Many, but not all, Gigabit Ethernet switches and Gigabit Ethernet network interface cards support jumbo frames, but all Fast Ethernet switches and Fast Ethernet network interface cards support only standard-sized frames.

Although jumbo frames are not a standard, the quickest and simplest option was to require jumbo frame support for every device in the FCoE/FCoCEE network.

“Baby” jumbo frames of approximately 2500 bytes are desirable for the future.

From lossy to lossless

Storage requirements are very stringent and a given is that any new transport mechanism has to be lossless. If Ethernet is to be used, a “new” Ethernet must be built, an Enhanced Ethernet.

Additionally, congestion is to be avoided at all costs, and there must be no compromise to the availability that is inherent in the FC data center SAN that exists today. The FC-BB-5 standard will need to take into account that the “lossy” nature of the Ethernet needs to be

addressed if it is to gain a foothold in the transport of storage data. We describe several of the key characteristics and points that need addressing in the following sections.

Reliability

Just as in FC, FCoE needs to have the ability to guarantee frame delivery. The physical links have to be able to cope with low bit error rates (BER), and they have to ensure that if there is a buffer overflow or any form of congestion that no frames are dropped. Fortunately, both 1 GbE and 10 GbE have a BER requirement that matches that of FC — a 10-to-12 bit error rate (1 in 10^{12} bit error rate).

Flow control and avoiding packet loss

In the design of FC was the capability to maintain the speed and efficiency of the data center *channel* architecture that entailed the creation of a flow mechanism, which was done by using buffer to buffer "credits." Very briefly, a device cannot send any additional frames until the receiver says that it is OK to do so. FC handles this situation very well, and the same has to be true of the Enhanced Ethernet for FCoE to be adopted.

One of the problems with any Ethernet network is that without an adequate flow control mechanism, when a congestion condition arises, packets will be dropped (lost), which was not the answer that FCoE was looking for. Flow control that was similar to the buffer to buffer credit method was needed in the Ethernet network.

What we find is a flow control PAUSE mechanism that can be used to prevent packet loss. In a similar manner to buffer to buffer credit methods, the flow control PAUSE mechanism will ask a sender to hold off sending any more frames until the receiver's buffers are cleared. This mechanism is contained in the IEEE 802.3 Annex 31B flow control standard specification and goes part of the way to ensuring that storage traffic does not suffer frame loss and attempts to alleviate congestion.

However, one of the problems with the PAUSE mechanism is that it applies no intelligence to the PAUSE, and arbitrarily pauses all traffic. The IEEE is conscious of this problem and has a number of working groups looking at the issues of congestion management and quality of service (QoS) priority levels to ensure that the most important data gets to its destination first without suffering from any unwarranted congestion.

As the IEEE standard evolves, we can expect to see priority-based flow control that can be selectively applied to different classes of traffic. We will discuss flow control and other enhancements later in "Data center bridging" on page 13.

Addressing

In an FC network, the links are based on a "point-to-point" topology. The Ethernet network differs in this respect, because it does not create a point-to-point connection in the same way that FC does. FCoE will use the destination and source MAC addresses (as shown in Figure 3 on page 5) to forward a frame to its intended destination.

Addressing schemes

The two addressing schemes that will be encountered are:

- Server-Provided MAC Addresses (SPMA)
- Fabric-Provided MAC Addresses (FPMA)

Both of these addressing schemes have been accepted as valid addressing schemes, and it is up to the individual vendor as to which addressing scheme they choose to implement and support.

Server-Provided MAC Addresses

As its name suggests, an SPMA is a MAC address that is issued in accordance with Ethernet standards and set by the manufacturer at installation.

Fabric-Provided MAC Addresses

The FPMA is a fabric-unique address that is assigned by the fabric. The low-order 24 bits are equivalent to the N_Port ID (FC-ID) assigned during fabric login, and the high-order 24 bits are equal to the FCoE MAC address prefix (FC-MAP) associated with the fabric.

We discuss how these MAC addresses will be used to route frames after we describe the terminology that will be used in the FCoE/FCoCEE data center.

FCoE terminology

As we have stated, FCoE has the ability to transport FC, but unless FCoE is made lossless, it does not meet the behavior requirements that data centers demand of an FC port.

Before we introduce the basic FCoE concepts, it is important that we introduce the terminology that will be encountered and also the characteristics of FCoE. Where possible, we will draw parallels with FC to aid the understanding of FCoE. We will introduce the FCoE elements and concepts and then show how they all fit together.

Ports

To ensure that the FCoE ports meet the requirements (remember, our requirement is that FCoE ports behave like an FC port), they need to emulate FC ports and become “virtual” FC ports. So, using the FC terminology for nodes, ports, and inter-switch links (ISLs), FCoE will have a:

- Port in an Enhanced Ethernet node (ENode), which will be a Virtual N_Port (VN_Port)
- Port in an FCoE-capable Ethernet switch, which will be a Virtual F_Port (VF_Port)

The FCoE-capable Ethernet switch can also have an:

- ISL port, which will be a Virtual E_Port (VE_Port)

FCoE-capable Ethernet switch

This switch is capable of supporting the Enhanced Ethernet at a minimum, and one or more of these switches must be configured to support FCoE forwarding functions provided by an FCoE Forwarder (we discuss FCoE Forwarders later), as well as support for Fibre Channel fabric services.

From an addressing point of view, each FCoE Virtual Port will have its own MAC address associated with it, whether it has been assigned by SPMA or FPMA.

Links

In FC, every link between a node port and a switch port is a physical, point-to-point connection. In FCoE, the concept is different. An FCoE node port (VN_Port) has the ability to access more than one FCoE Switch port (VF_Port), based on its MAC address, giving a multitude of possibilities. Even stranger (conceptually speaking when compared to FC) is the ability of more than one FCoE node to be able to access the same FCoE Switch port.

This capability leads to the concept of virtual links, and there is an ENode MAC-to-FCoE Switch port-MAC relationship. This relationship is our FCoE Virtual Link, or just virtual link for simplicity.

ENode

An FCoE Node (ENode) is an FC node that is associated with one or more Enhanced Ethernet MACs, one or more FCoE Link End Points (FCoE_LEPs), and one or more VN_Ports. Note that each Enhanced Ethernet MAC is coupled with an FCoE Controller. We will discuss the function of the FCoE Controller later, but for now, the FCoE_Controller is responsible for the creation of VN_Ports, VF_Ports, VE_Ports, and FCoE_LEPs.

Furthermore, because each virtual port has its own MAC address, the FCoE Virtual Link is created by using the pair of MAC addresses of the two virtual link end points as source and destination MAC addresses.

But what is a virtual link end point? In its simplest form, a virtual link is the logical link created by a VN_Port communicating with a VF_Port, or a VE_Port communicating with a VE_Port. The component that facilitates this logical link is the FCoE Link End Point (FCoE_LEP). Each VN, VF, and VE_Port will have an FCoE_LEP associated with it, and this FCoE_LEP will also perform FC frame encapsulation and decapsulation. Because each virtual port also has a MAC address associated with it, it is easy to see the manner in which virtual link end points can be associated with each other.

Our FC/FCoE switching and interface element has an Ethernet port and the capability to handle, forward, or otherwise cope with FC frames. Within the switch is a component that is called an FCoE Forwarder (FCF). The FCF is an FC switching element and is associated with one or more Enhanced Ethernet MAC addresses. The FCF is the communication bridge between the Enhanced Ethernet and an FC fabric.

FCoE Forwarder

The FCoE Forwarder is a function that exists in a switch that has Ethernet ports and is responsible for translating the FCoE frames between the Enhanced Ethernet and an FC SAN. On the Enhanced Ethernet side, this function can be within a device, or it can be integrated into an Enhanced Ethernet switch. On the FC SAN side, the native FC ports connect to a Fibre Channel switch.

Both the ENode and the remote switch can have one or more FCoE_LEPs that are associated with one or more VN or VF and VE_Ports. FCoE_LEPs will reside at or in the ENode and also at or in the FCF.

What this means is that the virtual link connections allow for any VN_Port to connect to any VF_Port. This connection is in contrast to the FC point-to-point, physical relationship of N_Port to F_Port.

We show you an example of this virtual link connection in Figure 6 on page 11 where we show how ENode 1 and ENode 2 are connected to FCF Y and FCF Z. ENode 1 and ENode 2 each has a single physical Ethernet connection to the Enhanced Ethernet, as do FCF Y and FCF Z.

Multiple VN_Ports can be created at each ENode and associated with multiple VF_Ports (and VE_Ports can be associated with other VE_Ports), which can be created at each FCF, and these VN_Ports and VF_Ports can (and will) be connected to each other via virtual links.

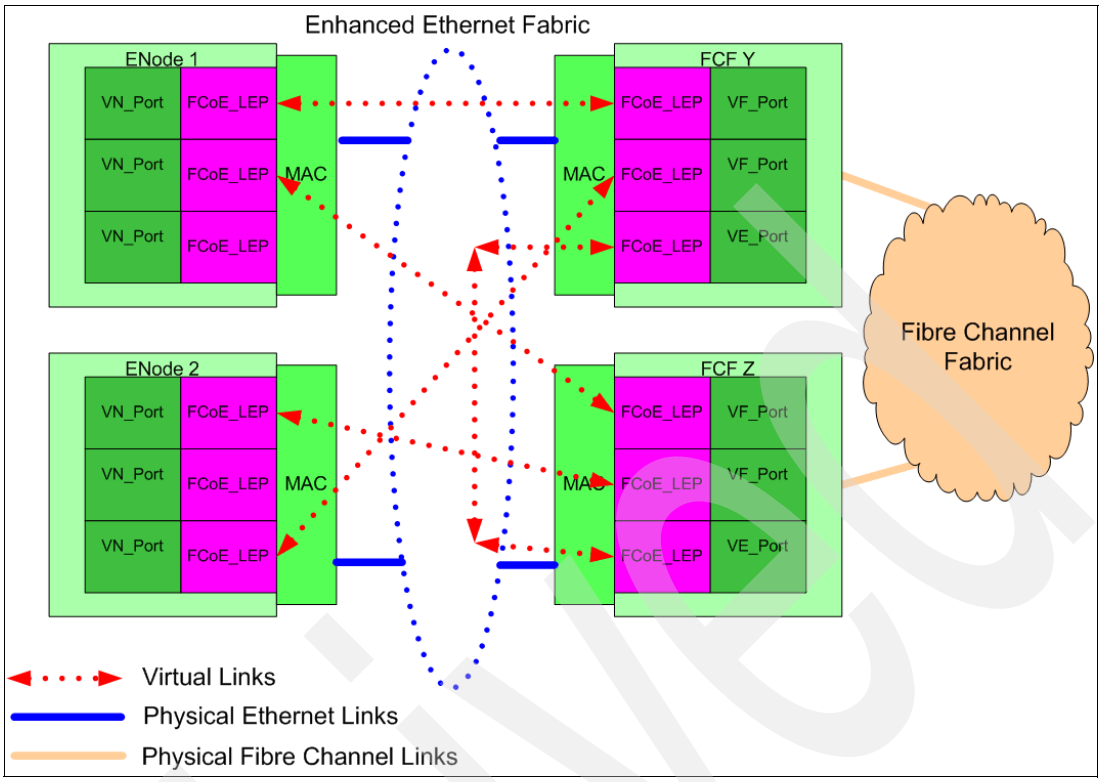

Figure 6 FCoE and Fibre Channel SAN high-level overview

When the FCoE_LEP decapsulates the FC frames from the FCoE frames, it is the FCoE_LEP that verifies that the destination address of the received FCoE frame is the same as the MAC address of the intended link end point. The FCoE_LEP will also verify that the source address of the received FCoE frame is the same as the MAC address of the remote link end point.

It is not a requirement that the FCF is connected to a Fibre Channel Fabric. The FCF can be a combined (“combo”) FC/FCoE switch provided as part of the Enhanced Ethernet Fabric.

Initialization, discovery, and port creation

In much the same way as FC already does, there needs to be a mechanism that discovers new ports, assigns and unassigns MAC addresses, and handles logins and logouts. This process is called the FCoE Initialization Protocol (FIP). Without going into great technical depth, which is not the intent of this paper, the FIP will be used by FCFs to discover other FCFs and to advertise their presence to nodes on the fabric. ENodes will use FIP to log in to the fabric. And it is through FIP that ENodes and FCFs will establish the virtual ports: VN_Ports, VF_Ports, and VE_Ports, and therefore create FCoE_LEPs.

Connectivity

FIP Link Keep Alive (FKA) is implemented in FCoE, because there needs to be a way of detecting if something has gone wrong in the path between VN_Ports and VF_Ports, and as well as between VE_Ports.

Primarily, FKA enables the VF_Port to discover that the VN_Port is unreachable, because the physical link is no longer a reliable indicator of this condition anymore (with FCoE, there might not be a direct connection to the FCF, because the path might be through intermediary switches). Timers associated with ENodes and FCFs are able to discover if a port is sending messages, and also if the port is still alive and not just inactive. To discover if a port is still alive, a periodic message needs to be sent, and these messages are sent as unsolicited advertisements.

These messages can be sent from FCF to ENode, and from FCF to FCF. However, because there is no unsolicited advertisement from an ENode to the FCF, a special FKA has been created to let the FCF know that the ENode and its VN_Ports are still alive. No response from the VF_Port is necessary, because the VN_Port discovers unreachable VF_Ports by the absence of the periodic advertisements that are multicast from the FCF. ENodes might send periodic FKAs for every MAC address to which it is capable of transmitting or from which it capable of receiving.

There is a functional entity in control of this process called an FCoE Controller, and it is responsible for executing the FIP. The FCoE Controller will be part of, or exist in, an FCF and an ENode.

Figure 7 shows how the FCoE Controller can exist in the ENode and the FCF.

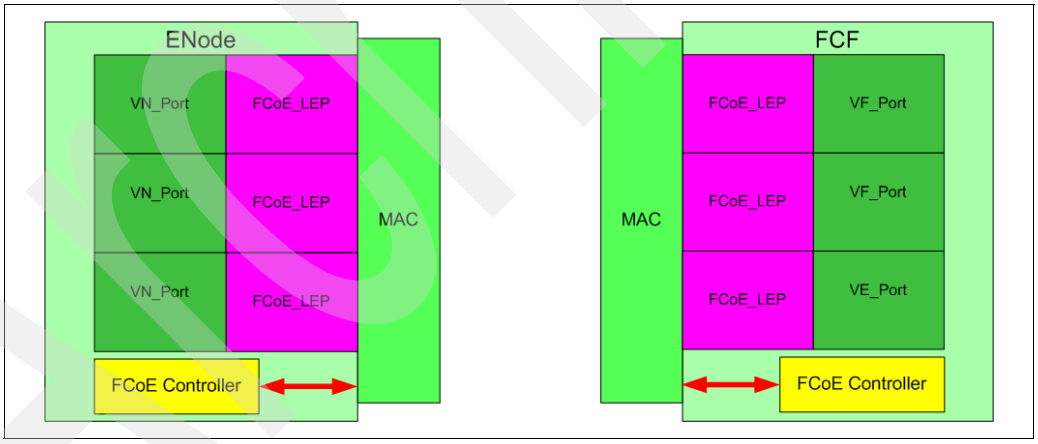

Figure 7 FCoE Controller

Pathing and routing mechanisms

It is essential that however the FCoE solution is implemented that the Ethernet and IP standards, alongside the FC standards, are supported for switching, path selection, and routing. Not only must the FCoE solution support the current standards, it must also be in a position to support any enhancements to the standards. Crucially, it has to have the ability to discern and adapt to FC and FCoE.

FC frames will still be handled by the Fabric Shortest Path Protocol (FSPF). The Spanning Tree Protocol (STP), Etherchannel, and its various versions and intended enhancements

(such as TRILL, which we mention next) will be used to move Ethernet frames. Because FC is layered on top of the Ethernet, there is no conflict between the two methods.

Under the auspices of the Internet Engineering Task Force (IETF), there is a working group that is looking at the Transparent Interconnection of Lots of Links (TRILL). The brief description of this group's goal is to design a solution for shortest-path frame routing in multi-hop IEEE 802.1-compliant Ethernet networks with arbitrary topologies, using an existing link-state routing protocol technology. It is expected that solutions in the future will be able to support this after TRILL is fully approved and ratified. What this means is that in the future, TRILL might be a protocol to watch for moving Ethernet frames around.

In essence, FCoE frames will be moved around by whatever method the Ethernet network uses.

Usage cases

We see three deployment usage cases for the initial generation of FCoCEE:

- **Usage case 1** is a rack upgrade scenario. In this case, an FCoCEE rack is deployed into an existing data center (DC), without changing the data center's Ethernet or FC infrastructure.
- **Usage case 2** is a new Dual-Fabric, data center scenario, where FCoCEE is used within each rack, but at the DC level, there are still two separate fabrics: Ethernet and FC.
- **Usage case 3** is a new Converged Fabric, data center scenario, where FC is used at the perimeter of the converged FCoCEE fabric to attach storage, but CEE and FCoCEE switches are used throughout the DC.

All of these usage cases will also enable a lower operational expense by integrating FC and Ethernet fabric management.

Data center bridging

As we have stated, there are issues to be overcome if we are to have a truly lossless Enhanced Ethernet. The IEEE will look at extensions to the Ethernet as part of its Data Center Bridging Task Group.

The Data Center Bridging (DCB) Task Group (TG) is a part of the IEEE 802.1 Working Group with the charter to provide enhancements to the existing 802.1 bridge specifications to satisfy the requirements of protocols and applications in the data center.

Data centers typically comprise multiple application-specific networks that run on different link layer technologies. The enhancements to the specifications will enable the 802.1 bridges and facilitate a converged network.

The projects that are being managed by the task group are:

- Priority-Based Flow Control (PFC) provides a link level flow control mechanism that can be independently controlled at a priority level and that can selectively pause different classes of traffic. The aim is to ensure zero loss due to congestion in data center bridging networks. The motivation behind this enhancement is to provide a no packet drop behavior, which is required by some data center applications (for example, FC and some IPC traffic). With priority-based flow control, separate flow control mechanisms can be used for different traffic classes.

Flow control comes into play when the network experiences congestion. When congestion happens, priority-based flow control can be engaged for the lower priority traffic classes where the no packet drop behavior is expected. By selectively pausing these lower priority traffic classes, other high priority traffic and delay sensitive traffic sharing the same link are not affected, which differs from the current IEEE 802.3x PAUSE?[1]?[3] mechanism where all traffic is affected when PAUSE is enabled. For example, with PFC, if storage traffic has a higher priority than LAN traffic and a large storage transfer causes congestion, PFC can be engaged to pause the storage transfer and let the LAN transfer proceed.

These enhancements are being addressed in the working group 802.1Qbb.

- Enhanced Transmission Selection (ETS) will provide a common management framework for assigning the appropriate and desired bandwidth to different traffic classes. For example, if a class does not use its available bandwidth, the bandwidth can be used by other traffic classes.

  Today, IEEE 802.1p defines a strict priority mechanism, where as long as there is higher priority traffic to be transferred, that traffic will take precedence over lower priority traffic. That is, it does not allow bandwidth to be allocated to different traffic priorities. The motivation behind ETS is the recognition that different traffic classes have different queuing requirements and need different resource allocation. ETS is the means to provide traffic differentiation, such that multiple traffic classes can share the same consolidated Ethernet link without impacting each other.

  Enhanced Transmission Selection is specified using two configuration tables. The first table maps the priority level as conveyed in IEEE 802.1p bits to Priority Groups, where each Priority Group represents a traffic class, such as LAN, SAN, IPC, and management.

  The first table defines whether a Priority Group is lossless or lossy and the type of priority it will use (strict as compared to non-strict). Strict priority scheduling is as defined in IEEE 802.1p, and there is no bandwidth check. For non-strict priority scheduling, the second table specifies the amount of link bandwidth allocated for each Priority Group. How bandwidth is allocated within a group is unspecified. When a group does not fully utilize its bandwidth allocation, the unused bandwidth is given to the other groups. Under an ETS-based scheduler, storage traffic can be managed as a group with configurable bandwidth guarantees to ensure that storage traffic will get its fair share of resources, and storage traffic will be lossless.

  These enhancements are being addressed in the working group 802.1Qaz.

- Data Center Bridging Capabilities Exchange Protocol (DCBCXP) is a discovery and capability exchange protocol that allows Enhanced Ethernet devices to convey and configure their Enhanced Ethernet capabilities with other attached Enhanced Ethernet devices. This protocol will ensure a consistent configuration across the network.

  The motivation behind this enhancement is to provide a way for discovering, initializing, and managing CEE compliant components. DCBCXP uses Link Layer Discovery Protocol (LLDP) as defined in IEEE 802.1AB to advertise connectivity and management information between two link peers. It uses DCBX Management Information Base (MIB) to configure and monitor CEE-compliant components.

  These enhancements are being addressed in the working group 802.1AB.

There are two additional mechanisms that further enhance convergence of FC with Ethernet, but they are not required for initial deployments. IBM views these additional CEE mechanisms as needed to support FC convergence with Ethernet at the data center level. The two additional mechanisms are:

- Congestion Notification (CN): Provides end to end congestion management for protocols that do not already have built-in congestion control mechanisms, which includes FCoE.

While mechanisms for congestion notification exist at the IP and TCP level, an enhancement at the Ethernet link level will provide the congestion management capabilities to applications that will exploit CEE but do not use TCP/IP. Link level congestion notification provides a mechanism for detecting congestion and notifying the source to back off the traffic flowing on the congested links. Link level congestion notification will allow a switch to send a signal that other ports need to stop or slow down their transmissions.

Because of the reactive nature of the mechanism, the life of the flow must be much greater than the network latency for congestion notification to be effective, which is useful in controlling unicast traffic in networks with long-lived data flows with respect to their bandwidth-delay product. However, while link level congestion notification might reduce the chance of deadlocks in the network and packet drops, it is not sufficient to guarantee a no packet drop behavior.

These enhancements are being addressed in the working group 802.1Qau.

- Link level shortest path first-based routing protocol: The routing schemes used in the current Ethernet link layer are inefficient due to the need to strictly avoid loops. This enhancement is intended to provide a mechanism that can provide shortest-path frame routing in multi-hop IEEE 802.1-compliant Ethernet fabrics with arbitrary topologies, using existing link-state routing protocol technology.

  This effort is currently being pursued in the IETF TRILL working group.

Summary

At the same time as providing cost reduction benefits, FCoE and FCoCEE will maintain all the services to which the Fibre Channel SAN is accustomed.

With the large install base of FC-based storage in the enterprise data center, a fabric convergence solution that aims to provide a consolidated network for IPC, LAN, and storage traffic needs to allow the users to protect their investment in FC storage. CEE enables fabric convergence by carrying FC traffic over a lossless transmission network, which allows a user to embrace fabric consolidation while retaining full use of the FC storage. Furthermore, features, such as traffic differentiation and priority-based flow control in CEE, provide value regardless of whether FC Channel storage is used or not.

Initially, CEE and FCoCEE are expected to play well in the High Performance Computing (HPC) and Analytics market as an alternative to InfiniBand®. As FCoCEE matures and meets the performance, reliability, and quality requirements of enterprise clients, we expect CEE will also play well in large enterprises wanting to pursue FC convergence with Ethernet. The usage cases described in this paper can provide a model for how FCoCEE deployment will progress over time.

This evolution with FCoE and FCoCEE makes network consolidation a reality by the combination of Fibre Channel and Ethernet. This network consolidation will still maintain the resiliency, efficiency, and seamlessness of the existing FC-based data center.

As you expect, IBM is investing heavily in the standards bodies, its technology, and its core competencies to ensure that FCoE preserves and enhances the SAN investment.

The writer

Jon Tate is a Project Manager for IBM System Storage™ SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 23 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist and an IBM SAN Certified Specialist. Jon also serves as the UK Chair of the Storage Networking Industry Association.

Errors, omissions, inaccuracies, and so on

Because this is an evolving protocol, it seems fair that there might be some unintended inaccuracies and omissions, and I respectfully ask that you contact me, *tatej@uk.ibm.com*, and although I do not guarantee to incorporate or respond to all suggestions, they will be considered.

My intent is to keep this paper as a living document and add items as the standards are ratified and as products begin to appear.

Thanks to

Renato Recio
IBM Distinguished Engineer, Chief Engineer eSystem Networks

Dan Eisenhauer
IBM Systems and Technology Group

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

This document REDP-4493-00 was created or updated on March 18, 2009.

Send us your comments in one of the following ways:
- Use the online **Contact us** review IBM Redbooks publication form found at:
  ibm.com/redbooks
- Send your comments in an e-mail to:
  redbooks@us.ibm.com
- Mail your comments to:
  IBM Corporation, International Technical Support Organization
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400 U.S.A.

Trademarks