

Fiber Channel Over Ethernet (FCoE)

Using Intel[®] Ethernet Switch Family

White Paper

November, 2008



Legal

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life saving, life sustaining, critical control or safety systems, or in nuclear facility applications.

Intel may make changes to specifications and product descriptions at any time, without notice.

Intel Corporation may have patents or pending patent applications, trademarks, copyrights, or other intellectual property rights that relate to the presented subject matter. The furnishing of documents and other materials and information does not provide any license, express or implied, by estoppel or otherwise, to any such patents, trademarks, copyrights, or other intellectual property rights.

The Controller may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Intel and Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2011. Intel Corporation. All Rights Reserved.

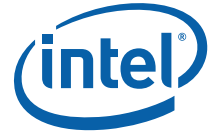


Table of Contents

Introduction	4
Fibre Channel over Ethernet (FCoE).....	4
Intel® Ethernet Switch Family QoS and Congestion Management.....	4
Traffic Isolation	5
Traffic Classes	6
Watermarks and Flow Control.....	6
Baby Jumbo Frame Support	6
Scheduling	7
Ingress Policing and Rate Limiting.....	7
Egress Scheduling and Shaping	7
FCoE Frame Forwarding.....	8
Zoning and Security.....	9
Using VLANs	9
Using ACLs	9
Fabric Management	10
FIP Snooping	10
Converged Datacenter Example	11
Conclusion	12



Introduction

Fibre Channel over Ethernet (FCoE) is being defined as part of the initiative to converge storage fabrics and data fabrics within the datacenter. This white paper will describe how the advanced features in the Intel[®] Ethernet Switch Family can be used to support the FCoE standard. For further information on these features, see the FM4000 data sheet and the Telecom Congestion Management Application Note.

Fibre Channel over Ethernet (FCoE)

Ethernet is the dominant networking protocol in the enterprise today, but due to previous limitations in the Ethernet standard, Fibre Channel (FC) has emerged as a widely used fabric for storage area networking. Because of this, servers in the data center require connections to both fabrics, increasing costs due to separate adapter cards, cables and switches.

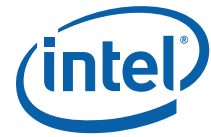
Fibre Channel over Ethernet (FCoE) attempts to converge the Fibre Channel storage fabric into the Ethernet data fabric in order to reduce both the up-front costs as well as total cost of ownership. It does this by encapsulating FC frames within a FCoE header and then adds further encapsulation within an Ethernet frame. For this to work, the Ethernet fabrics that forward these frames must support the following features.

- 10G fabric ports in order to maintain 4G and 8G FC performance
- Lossless operation
- Bounded latency through the fabric
- Storage zoning
- Data security

This paper will discuss how the Intel[®] Ethernet Switch Family contains the features to support these requirements. This paper will not discuss Fibre Channel Forwarding (FCF), which is implemented in end-point devices such as Converged Network Adapters (CNAs) or specialized edge switches.

Intel[®] Ethernet Switch Family QoS and Congestion Management

Quality of Service (QoS) and congestion management are key features that are required to support FCoE. A block diagram representing the Intel[®] Ethernet Switch Family QoS and congestion management features is shown in [Figure 1](#). As frames arrive at the ingress, token buckets can be used to police and rate limit incoming traffic based on



ACL rules. ACL rules can also be used to assign frames to traffic classes. The traffic class determines the memory partition as well as the egress scheduler queue that a frame will be associated with.

Class-Based Pause (CBP) frames can be sent to ingress devices such as FCoE CNAs based on traffic class to memory partition mapping. For example, if a memory partition crosses a programmed watermark, all traffic classes assigned to that memory partition will be paused. Token buckets can also be used to provide ingress rate control per traffic class using CBP frames. At the switch egress, CBP frames generated by a line card can be used to flow control the 8 scheduler queues at each switch egress port. VCN frames, which report the status of these egress queues, can also be multicast to ingress traffic managers to eliminate head-of-line blocking if needed.

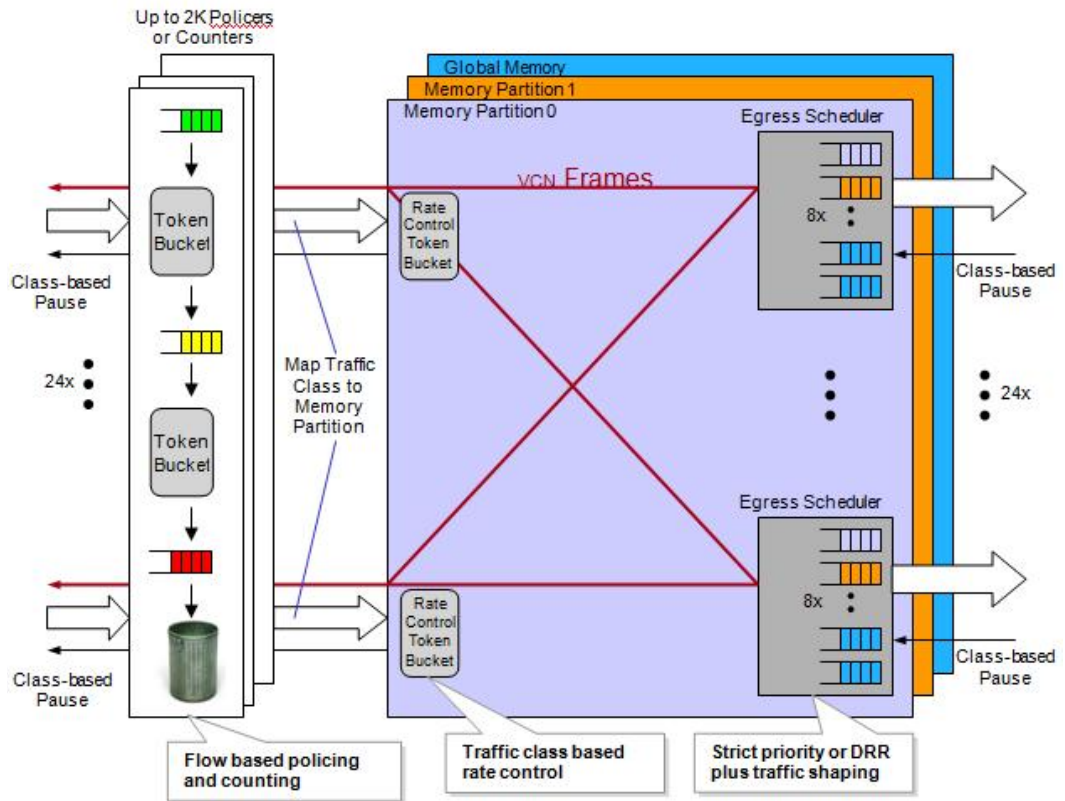


Figure 1. Intel® Ethernet Switch Family QoS and Congestion Management

Traffic Isolation

Fibre Channel was developed as a transport for iSCSI, which cannot tolerate dropped frames or unbounded latency due to its excessive time-out recovery delay. Ethernet switches were originally designed to drop frames during periods of high congestion. To support FCoE, the



switch needs to isolate storage traffic from other sources of congestion such as bursty data traffic. This, in effect, creates a separate virtual fabric for storage. The Intel[®] Ethernet Switch Family does this by assigning frames to traffic classes, which are then assigned to separate memory partitions for flow control.

Traffic Classes

Traffic classes are used for two purposes within a Intel[®] Ethernet Switch Family switch. At the ingress, traffic classes are assigned to memory partitions, which are flow controlled separately. At the egress, traffic classes are used for scheduling. The Intel[®] Ethernet Switch Family devices can map various header fields such as FCoE ethertype to one of 8 traffic classes using ACL rules. The Intel[®] Ethernet Switch Family assigns a higher priority to larger traffic class numbers when it is relevant. This traffic class mapping is global for all egress ports.

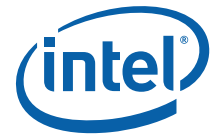
Watermarks and Flow Control

The Intel[®] Ethernet Switch Family uses a shared memory structure to store frames. An ingress crossbar forwards frames from the receive ports into shared memory. An egress crossbar forwards frames from shared memory to the transmit ports. Watermarks are used to define regions where frames can be stored in shared memory and can also trigger certain actions when these memory fill levels are exceeded.

The Intel[®] Ethernet Switch Family ACL rules can be used to match on the FCoE Ethertype field and assign these frames to a give switch priority and traffic class. By assigning only FCoE frames to one of the memory partitions, and all other frames to the other memory partition, FCoE traffic will be effectively isolated from data traffic. Class-Based Pause frames are generated when these memory partitions cross pre-programmed watermarks. Assuming the upstream CNA can react to these CBP frames, lossless operation can be guaranteed for storage traffic.

Baby Jumbo Frame Support

The FCoE standard requires that the Ethernet infrastructure supports frame sizes as large as 2.5KB (baby jumbo frames). The FM4000 contains 2MB of shared packet memory. For a 24-port switch, this allows for over 32 baby jumbo frames per egress port. Or for mixed data and storage traffic, each egress port can support over 16 baby jumbo FCoE frames along with 4 10K data jumbo frames. This is enough buffer memory margin to accommodate class-based pause flow control latency without any frame drops.



Scheduling

So far in this paper, we have shown how the Intel® Ethernet Switch Family can provide traffic isolation and lossless operation for storage traffic. This section describes how shaping and scheduling can be used to guarantee a maximum latency for FCoE traffic. The key to maximum latency is to make sure that FCoE traffic has a guaranteed minimum bandwidth and that other traffic types do not exceed bandwidth limits.

Ingress Policing and Rate Limiting

At the switch ingress, data traffic can be policed or rate limited to make sure there is enough bandwidth allocated at this port for FCoE traffic. Policers use ACL rules to look at header information in order to provide flow based policing. For example, all traffic that does not contain an FCoE Ethertype can be policed in order to limit its ingress bandwidth using token buckets.

An alternative method is to use the ingress rate limiters that are based on memory partition fill levels. By the proper setting of data memory partition watermarks, data traffic can be rate limited using CBP frames. By assigning the FCoE traffic class to a separate memory partition as described above, the rate limiting of data traffic can provide a minimum bandwidth allocation for FCoE traffic.

Egress Scheduling and Shaping

At the switch egress, FCoE and data traffic can be assigned different traffic classes and therefore be scheduled differently. FCoE traffic can be given strict high priority, although this will have a tendency to starve data traffic during bursts of storage activity. A better way is to use Deficit Round Robin (DRR) scheduling for storage and data traffic which can provide a minimum bandwidth guarantee for FCoE while also providing minimum bandwidth guarantees for certain types of data traffic such as video distribution. For example, a 10G egress port can be assigned a minimum bandwidth of 4G for FCoE traffic and a minimum bandwidth of 2G for video traffic, leaving 4G for all other data traffic.

Egress traffic shaping can be used to create an upper bound on the bandwidth for a traffic class and can be used to reduce latency jitter. If DRR is used, it is expected the maximum shaping bandwidth will be set higher than the minimum DRR bandwidth. Consecutive traffic class numbers can be in the same shaping group such that the aggregate bandwidth from that group does not exceed a maximum value. For

example, traffic shaping can be used for a group of data traffic classes to make sure they do not impact downstream FCoE bandwidth allocations.

FCoE Frame Forwarding

FCoE frames can be forwarded using the Intel[®] Ethernet Switch Family parser, TCAM and ARP table. The FCoE frame format is shown in Figure 2.

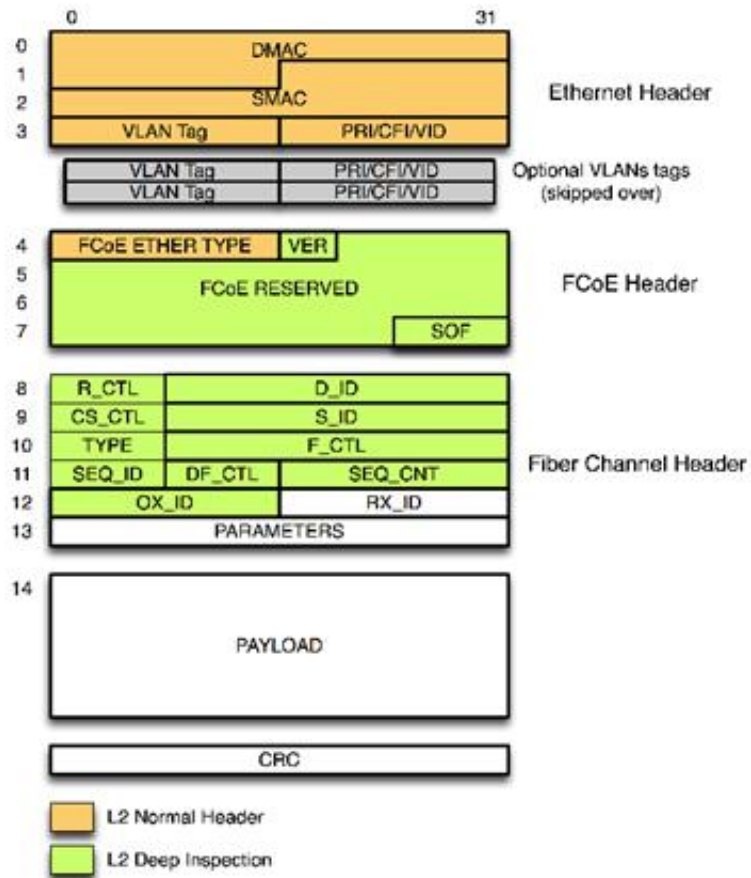
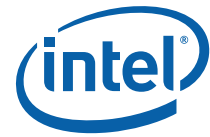


Figure 2. FCoE Frame Format

The Intel[®] Ethernet Switch Family parser and TCAM can identify FCoE packets using the ethertype field. Using deep packet inspection, forwarding decisions can be made using the FC header information. To do this, the TCAM is used to match the FC destination ID field. When a match occurs, it produces an index into the ARP table. The ARP table can contain the next hop MAC address in the same way that it does for



IP routing. For FCoE frames, the destination MAC address is replaced by the next hop MAC address and the source MAC address is replaced by the switch MAC address.

Zoning and Security

Zoning is used within a FC network for storage domain isolation and to provide a layer of data security. For example, zoning can be used to provide isolated storage for sensitive information that should not be accessed by certain groups within an organization.

Using VLANs

A simple way to support zoning within an Ethernet switch is to use VLAN IDs. This allows up to 4096 zones to be supported within the Ethernet fabric. Frames will be denied to ports that are not members of the correct VLAN. Unfortunately, this is not very secure since VLAN IDs are fairly easy to read and spoof.

Using ACLs

The Intel[®] Ethernet Switch Family contains an advanced Frame Filtering and Forwarding Unit (FFU) that can be used to implement ACL rules. The frame header is presented to the FFU, which associates one or more ACL actions with the frame. This unit contains 32 consecutive slices containing both TCAM and SRAM. The egress ACL unit follows the last slice and can optionally be used for egress ACLs that apply in parallel to multiple egress ports.

Each slice has the following elements:

- Configuration mask to select which frame header fields are selected for the TCAM comparison which is known as a key
- 512-entry x 36-bit TCAM block used for key comparison
- Hit detection circuitry which can cascade across consecutive slices to create keys larger than 36 bits
- Priority of hit detection to determine the hit in each slice with the highest priority which will take precedence
- 512-entry x 40-bit SRAM block to store ingress ACL actions associated with the highest priority hit

Note: The last slice can optionally feed its final 512 hits to the egress ACL unit to determine if any egress ACL actions are to be taken.

The Intel[®] Ethernet Switch Family provides a comprehensive set of ACL rules that can be used for FCoE zoning and data security. Up to 78-bytes within the first 128-bytes of the frame header can be mapped



into a TCAM. This includes the Ethernet header, the FCoE header and the FC header. If a match occurs in the TCAM, a set of actions can be applied to the frame including the following:

- Route the frame using the ARP table
- Deny forwarding of the frame and it is dropped at the egress
- Permit a frame to be forwarded through the switch
- Log the frame by sending a copy to the CPU port
- Trap the frame to the CPU port and do not forward
- Mirror the frame to a specified port
- Count the frame
- Police the frame (see the previous section)
- Change the frame DSCP field to the specified value
- Change the switch priority to the specified value
- Change the user bits in the ISL tag to the specified value
- Change the VLAN to the specified value
- Change the VLAN priority to the specified value

The FFU produces several different “action fields” which are subsequently used to modify the frame and/or determine its destination as described above. There may be multiple action entries, which attempt to modify the “action fields” for a given frame. As long as each action entry attempts to modify different action fields, then there is no conflict. However, if one action field is modified by more than one action entry that has hit, then the conflict is resolved using a precedence field assigned to each action entry.

Fabric Management

The storage fabric can be managed through the CPU port on each switch chip, or by using Fulcrum In-band Management (FIBM) frames. Fulcrum In-Band Management is the management of one or more Intel® Ethernet Switch Family chips through management commands encapsulated within Ethernet frames. This means that all Intel® Ethernet Switch Family chips do not require an attached CPU, and they can be managed by a CPU somewhere else in the network. One CPU can manage multiple switch chips, and there is no fixed limit to the number of switch chips that one CPU can manage using FIBM.

FIP Snooping

FC Initiation Protocol (FIP) Snooping is a control plane mechanism for endpoint discovery in FCoE networks. It uses a special Ethertype for discovery and login. The Intel® Ethernet Switch Family can support this by redirecting these frames to an attached control plane processor. This processor could be connected through a switch port, or connected



directly to the switch CPU interface. The switch can identify (snoop) FCoE frames using ACL rules based on Ethertype and mirror or redirect these frames to the control plane processor port.

The control processor must identify switch ports that are connected to Fibre Channel Forwarders (FCFs) and apply FCoE filters to non-FCF ports using ACL rules. FCF addresses are then learned by using FIP snooping and the control processor updates the switch ARP table based on this information. As new addresses are discovered, FCoE filters can be removed from designated ports by updating the switch ACL rules.

Converged Datacenter Example

Figure 3 shows an example of a converged datacenter network. Here, a high port count fat tree using Intel[®] Ethernet Switch Family switches forms the heart of the datacenter fabric, transporting both LAN and storage traffic. Legacy Fibre Channel systems connect to the Intel[®] Ethernet Switch Family fabric using FC bridges (FCFs) that contain specialized silicon for FCoE encapsulation. It is not cost effective to use these bridge switches in the heart of the datacenter fabric where no FC bridging is required. As the datacenter evolves, Converged Network Adapters (CNAs) will be used with servers, providing both NIC and HBA functionality while providing FCoE encapsulation. This, combined with storage arrays containing FCoE controller cards, will relegate FCoE bridges to the edge of the network in support of legacy Fibre Channel systems.

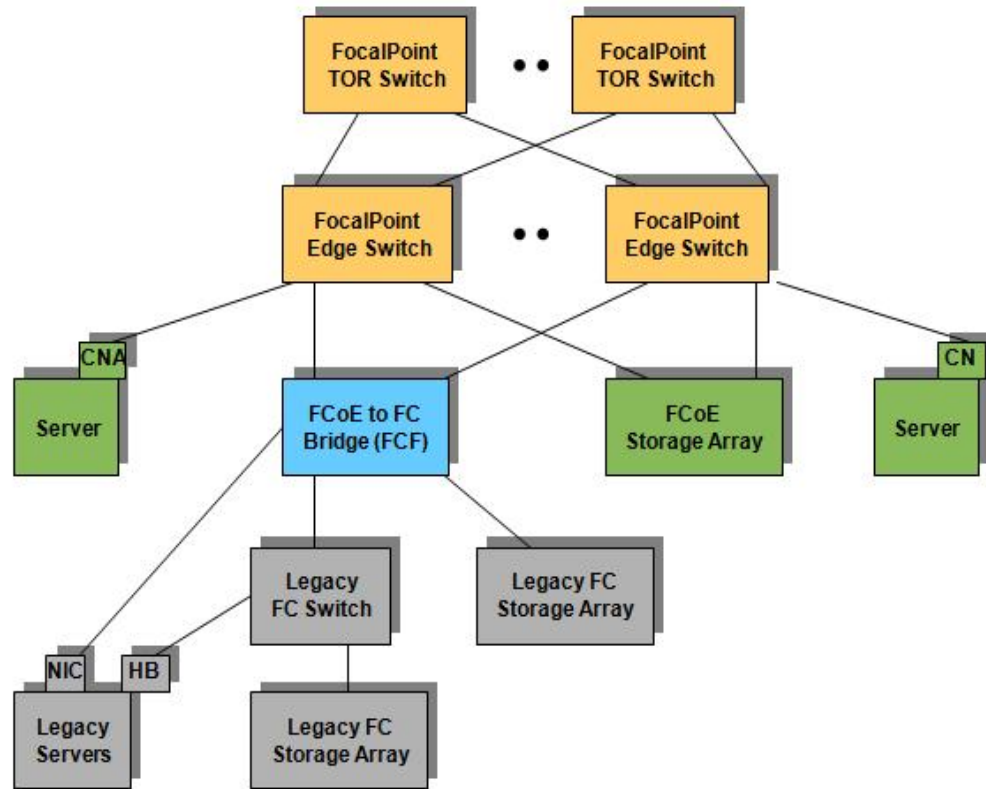


Figure 3. Example Converged Datacenter Network

Conclusion

FCoE is an emerging standard that will reduce data center costs by converging the storage and data switch fabrics. To support this convergence, Ethernet switches must support the advanced features required for storage traffic. The Intel[®] Ethernet Switch Family provides the required QoS, congestion management, zoning and data security required for FCoE.

§ §