I/O Consolidation in the Data Center

A complete guide to Data Center Ethernet and Fibre Channel over Ethernet

Silvano Gai
Claudio DeSanti

ciscopress.com

I/O Consolidation in the Data Center

A Complete Guide to Data Center Ethernet and Fibre Channel over Ethernet

Silvano Gai, Claudio DeSanti

Published by:

Cisco Press

800 East 96th Street

Indianapolis, IN 46240 USA

Printed in the United States of America 1 2 3 4 5 6 7 8 9 0

First Printing September 2009

Library of Congress Cataloging-in-Publication Data is available upon request.

ISBN-13: 978-1-58705-888-2

ISBN-10: 1-58705-888-X

Warning and Disclaimer

This book is designed to provide information about Data Center Ethernet and Fibre Channel over Ethernet. Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied.

The information is provided on an "as is" basis. The authors, Cisco Press, and Cisco Systems, Inc., shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book or from the use of the discs or programs that may accompany it.

The opinions expressed in this book belong to the author and are not necessarily those of Cisco Systems, Inc.

Feedback Information

At Cisco Press, our goal is to create in-depth technical books of the highest quality and value. Each book is crafted with care and precision, undergoing rigorous development that involves the unique expertise of members from the professional technical community.

Readers' feedback is a natural continuation of this process. If you have any comments regarding how we could improve the quality of this book, or otherwise alter it to better suit your needs, you can contact us through e-mail at feedback@ciscopress.com. Please make sure to include the book title and ISBN in your message.

We greatly appreciate your assistance.

Preface

This book describes the work done by Nuova Systems and Cisco on the evolution of Ethernet as a Data Center Network in 2006 and 2007. The technologies described herein have been accepted by industry and, starting in 2008, made their way into both products and standards.

In particular, the FC-BB-5 standard, which defines Fibre Channel over Ethernet (FCoE), has been approved by the International Committee for Information Technology Standards (INCITS) T11 Fibre Channel committee on June 4, 2009, and was forwarded to INCITS for publication as an American National Standards Institute (ANSI) standard. This book reflects the FCoE standard.

This book describes new Data Center technologies with an educational view. The reader will find here updated material compliant with current standards and material part of proposals for future standards.

Standards are expected to evolve; therefore this book should not be used as a basis for designing standards-compliant products. Designers should refer always to the most recent standards when designing products.

This book probably contains errors. The authors would appreciate if you email any corrections to the following address:

dc_book@ip6.com

Chapter 1

I/O Consolidation

Introduction

Today Ethernet is by far the dominant interconnection network in the Data Center. Born as a shared media technology, Ethernet has evolved over the years to become a network based on point-to-point full-duplex links. In today's Data Centers, it is deployed at speeds of 100 Mbit/s and 1 Gbit/s, which are a reasonable match for the current I/O performance of PCI, based servers.

Storage traffic is a notable exception, because it is typically carried over a separate network built according to the Fibre Channel (FC) suite of standards. Most large Data Centers have an installed base of Fibre Channel. These FC networks (also called fabrics) are typically not large, and many separate fabrics are deployed for different groups of servers. Most Data Centers duplicate FC fabrics for high availability reasons.

In the High Performance Computing (HPC) sector and for applications that require cluster infrastructures, dedicated and proprietary networks like Myrinet and Quadrix have been deployed. A certain penetration has been achieved by Infiniband (IB), both in the HPC sector and, for specific applications, in the Data Center. Infiniband provides a good support for clusters requiring low latency and high throughput from user memory to user memory.

Figure 1-1 illustrates a common Data Center configuration with one Ethernet core and two independent SAN fabrics for availability reasons (labeled SAN A and SAN B).

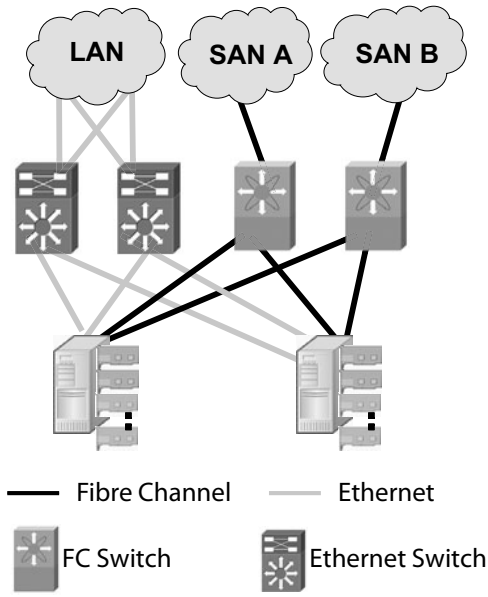


Figure 1-1 Current Data Center Architecture

What Is I/O Consolidation

I/O consolidation is the capability of a switch or a host adapter to use the same physical infrastructure to carry multiple types of traffic, each typically having peculiar characteristics and specific handling requirements.

From the network side, this equates in having to install and operate a single network instead of three (see Figure 1-2). From the hosts and storage arrays side, this equates in having to purchase fewer Converged Network Adapters (CNA) instead of Ethernet NICs, FC HBAs, and IB HCAs. This requires a lower number of PCI slots on the servers, and it is particularly beneficial in the case of Blade Servers.

The benefits for the customers are

- Great reduction, simplification, and standardization of cabling
- Absence of gateways that are always a bottleneck and a source of incompatibilities
- Less need for power and cooling
- Reduced cost

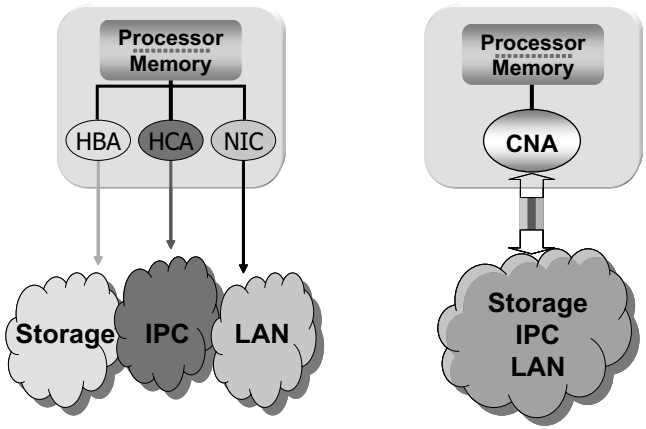


Figure 1-2 I/O Consolidation in the Network

To be viable, I/O consolidation should maintain the same management paradigm that currently applies to each traffic type.

Figure 1-3 shows an example in which 2 FC HBAs, 2 Ethernet NICs, and 2 IB HCAs are replaced by 2 CNAs.

Merging the Requirements

The biggest challenge of I/O consolidation is to satisfy the requirements of different traffic classes with a single network.

The classical LAN traffic that nowadays consists mainly of IPv4 and IPv6 traffic must run on native Ethernet [4]. Too much investment has been done in this area and too many applications assume that Ethernet is the underlying network for this to change. This traffic is characterized by a large number of flows. Typically these flows were not sensitive to latency, but this is changing rapidly, and latency now must be taken into serious consideration. Streaming Traffic is also sensitive to latency jitter.

Storage traffic must follow the Fibre Channel (FC) model. Again, large customers have massive investments in FC infrastructure and management. Storage provisioning often relies on FC services like naming, zoning, and so on. Because SCSI is extremely sensitive to packet drops, in FC losing frames is not an option. FC traffic is characterized by large frame sizes, to carry the typical 2KB SCSI payload.

Inter Processor Communication (IPC) traffic is characterized by a mix of large and small messages. It is typically latency, sensitive (especially the short messages). IPC traffic is used in

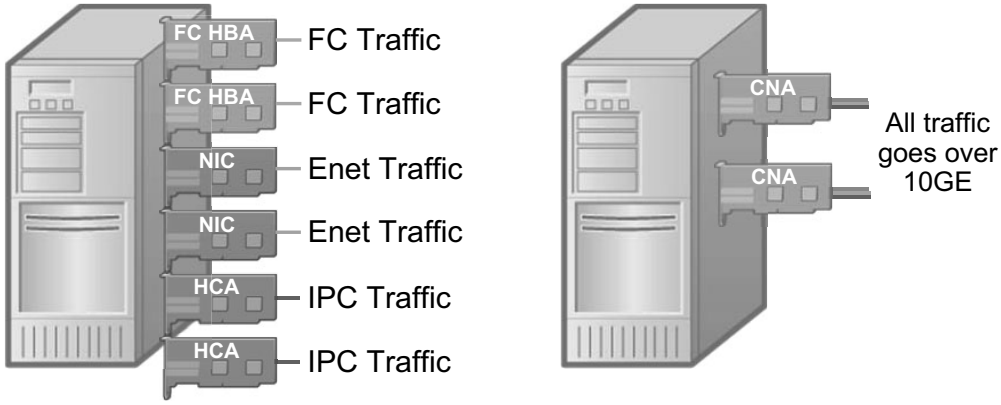


Figure 1-3 I/O Consolidation in the Servers

Clusters (i.e., interconnections of two or more computers). Examples of server clustering in the data center include

- Availability clusters (e.g., Symantec/Veritas VCS, MSCS)
- Clustered file systems
- Clustered databases (e.g., Oracle RAC)
- VMware virtual infrastructure services (e.g., VMware VMotion, VMware HA)

Clusters do not care too much about the underlying network if it is cheap, it is high bandwidth, it is low latency, and the adapters provide zero-copy mechanisms.

Why I/O Consolidation Has Not Yet Been Successful

There have been previous attempts to implement I/O consolidation. Fibre Channel itself was proposed as an I/O consolidation network, but its poor support for multicast/broadcast traffic never made it credible.

Infiniband has also attempted I/O consolidation with some success in the HPC world. It has not penetrated a larger market due to its lack of compatibility with Ethernet (again, no good multicast/broadcast support) and with FC (it uses a storage protocol that is different from FC) and to the need of gateways that are bottlenecks and incompatibility points.

iSCSI has been probably the most significant attempt at I/O consolidation. Up to now it has been limited to the low performance servers, mainly because Ethernet had a maximum speed

of 1 Gbit/s. This limitation has been removed by 10 Gigabit Ethernet (10GE), but there are concerns that the TCP termination required by iSCSI is onerous at the 10Gbit/s speed. The real downside is that iSCSI is "SCSI over TCP," it is not "FC over TCP," and therefore it does not preserve the management and deployment model of FC. It still requires gateways, and it has a different naming scheme (perhaps a better one, but anyhow different), a different way of doing zoning, and so on.

Fundamental Technologies

The two technologies that will play a big role in enabling I/O consolidation are PCI-Express and 10 Gigabit Ethernet (10GE).

PCI-Express

Peripheral Component Interconnect (PCI) is an old standard to interconnect peripheral devices to computer that has been around for many years [1].

PCI-Express (PCI-E or PCIe) [2] is a computer expansion card interface format designed to replace PCI, PCI-X, and AGP. It removes one of the limitations that have plagued all these I/O consolidation attempts (i.e., the lack of I/O bandwidth in the server buses), and it is compatible with current operating systems.

PCIe uses point-to-point full duplex serial links called lanes. Each lane contains two pairs of wires: one to send and one to receive. Multiple lanes can be deployed in parallel: 1x means a single lane; 4x means 4 lanes.

In PCIe 1.1, the lanes run at 2.5 Gbps (2 Gbit/s at the datalink), and 16 lanes can be deployed in parallel. This supports speeds from 2 Gbit/s (1x) to 32 Gbit/s (16x). Due to protocol overhead 8x is required to support a 10GE interface.

PCIe 2.0 (i.e., PCIe Gen 2) doubled the bandwidth per lane from 2 Gbit/s to 4 Gbit/s and extended the maximum number of lanes to 32x. It is shipping now.

PCIe 3.0 will approximatively double the bandwidth again: "The final PCIe 3.0 specifications, including form factor specification updates, may be available by late 2009, and could be seen in products starting in 2010 and beyond." [3].

10 Gigabit Ethernet

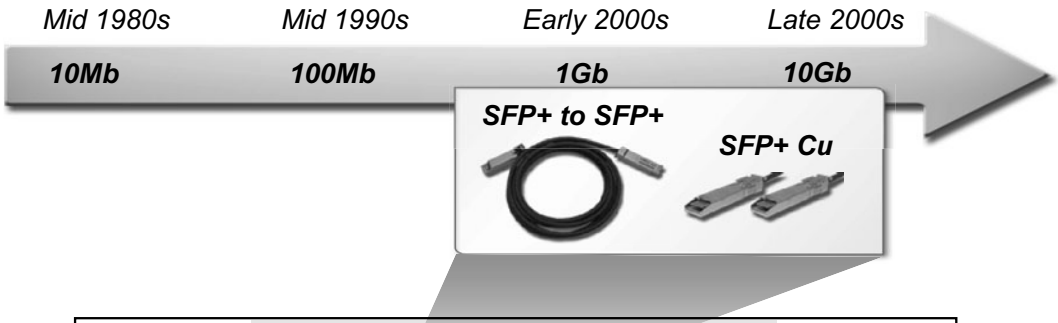
10GE is a practical interconnection technology since 2008. The standard has reached the maturity status and cheap cabling solutions are available. Fiber continues to be used for longer distances, but copper is deployed in the Data Center for its lower cost.

Switches and CNAs have standardized their connectivity using the Small Form-factor Pluggable (SFP) transceiver. SFPs are used to interface a network device motherboard (i.e., switches, routers, or CNAs) to a fiber optic or copper cable. SFP is a popular industry format supported by several component vendors. It has expanded to become SFP+, which supports data rates up to 10 Gbit/s [9]. Applications of SFP+ include 8GFC and 10GE.

The key benefits of SFP+ are:

- A comparable panel density as SFP
- A lower module power than XENPAK, X2, and XFP
- A Nominal 1W power consumption (optional 1.5W high power module)
- Backward compatibility with SFP optical modules

The IEEE standard for twisted pair cabling (10GBASE-T) is not yet a practical interconnection technology, because it requires an enormous number of transistors, especially when the distance grows toward 100 meters (328 feet). This translates to significant power requirements and into additional delay (see Figure 1-4). Imagine trying to cool a switch linecard that has 48 10GBASE-T ports on the front-panel, each consuming 4 watts!



Technology	Cable	Distance	Power (each side)	Transceiver Latency (link)
SFP+ Cu Copper	Twinax	10m	0.1W	0.1μs
SFP+ USR ultra short reach	MM OM2 MM OM3	10m 100m	1W	0
SFP+ SR short reach	MM 62.5μm MM 50μm	82m 300m	1W	0
10GBASE-T	Cat6 Cat6a/7 Cat6a/7	55m 100m 30m	8W 8W 4W	2.5μs 2.5μs 1.5μs

Figure 1-4 Evolution of Ethernet Physical Media

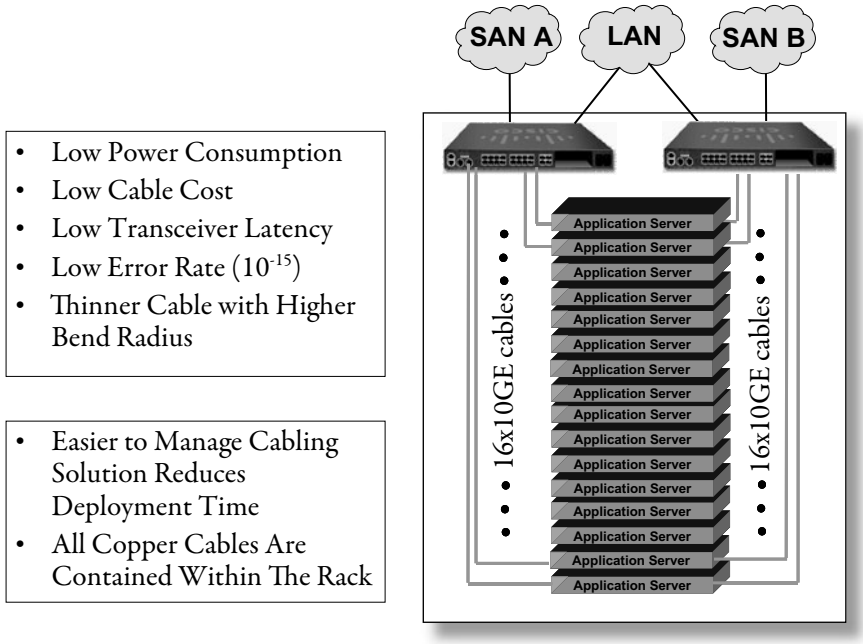


Figure 1-5 Twinax Copper Cable

A more practical solution in the Data Center, at the rack level, is to use SFP+ with copper Twinax cable (defined in SFF-8431, see [9]). The cable is flexible, approximately 6 mm (1/4 of an inch) in diameter, and it uses the SFP+ themselves as the connectors. Cost is limited; power consumption and delay are negligible. It is limited to 10 meters (33 feet) that are sufficient to connect a few racks of servers to a common top of the rack switch.

These cables are available from Cisco, Amphenol, Molex, Panduit, and others.

Figure 1-5 illustrates the advantages of using Twinax cable inside a rack or few racks.

The cost of the transmission media is only one of the factors that need to be addressed to manufacture 10GE ports that are cost competitive. Other factors are the size of the switch buffers, and Layer 2 versus Layer 3/4 functionality.

Additional Requirements

Buffering Requirements

Buffering is a complex topic, related to propagation delays, higher level protocols, congestion control schemes, and so on. For the purpose of this discussion, it is possible to divide the networks into two classes: lossless networks and lossy networks.

This classification does not consider losses due to transmission errors that, in a controlled environment with limited distances like the Data Center, are rare in comparison to losses due to congestion.

Fibre Channel and Infiniband are examples of lossless networks (i.e., they have a link level signaling mechanism to keep track of buffer availability at the other end of the link). This mechanism allows the sender to send a frame only if a buffer is available in the receiver, and therefore the receiver never needs to drop frames. Although this seems attractive at a first glance, a word of caution is in order: Lossless networks require to be engineered in simple and limited topologies. In fact, congestion at a switch can propagate upstream throughout the network, ultimately affecting flows that are not responsible for the congestion. If circular dependencies exist, the network may experience severe deadlock and/or livelock conditions that can significantly reduce the performance of the network or destroy its functionality. These two phenomena are well known in literature and easy to reproduce in real networks. This should not discourage the potential user, since Data Center networks have simple and well-defined topologies.

Historically Ethernet has been a lossy network, since Ethernet switches do not use any mechanism to signal to the sender that they are out of buffers. A few years ago, IEEE 802.3 added a PAUSE mechanism to Ethernet. This mechanism can be used to stop the sender for a period of time, but pragmatically this feature has not been successfully deployed. Today it is common practice to drop frames when an Ethernet switch is congested. Several clever ways of dropping frames and managing queues have been proposed under the general umbrella of Active Queue Management (AQM), but they do not eliminate frame drops and require large buffers to work effectively. The most used AQM scheme is probably Random Early Detection (RED).

Avoiding frame drops is mandatory for carrying native storage traffic over Ethernet, since storage traffic does not tolerate frame drops. SCSI was designed with the assumption of running over a reliable transport in which failures are so rare that it is acceptable to recover slowly from them.

Fibre Channel is the primary protocol used to carry storage traffic, and it avoids frame drops through a link flow control mechanism based on credits called buffer-to-buffer flow control (also known as buffer-to-buffer credit or B2B credit). iSCSI is an alternative to Fibre Channel that solves the same problem by requiring TCP to recover from frame drops; however iSCSI has not been widely deployed in the Data Center.

In general, it is possible to say that lossless networks require fewer buffers in the switches than lossy networks and that these buffers may be accommodated on-chip (cheaper and faster), although large buffers require off-chip memory (expensive and slower).

Both behaviors have advantages and disadvantages. Ethernet needs to be extended to support the capability to partition the physical link into multiple logical links (by extending the IEEE 802.1Q Priority concept) and to allow lossless/lossy behavior on a per Priority basis.

Finally, it should be noted that when buffers are used they increase latency (see page 10).

Layer 2 Only

A significant part of the cost of a 10GE inter-switch port is related to functionalities above Layer 2, namely IPv4/IPv6 routing, multicast forwarding, various tunneling protocols, Multi-Protocol Label Switching (MPLS), Access Control Lists (ACLs), and deep packet inspection (Layer 4 and above). These features require external components like RAMs, CAMs, or TCAMs that significantly increase the port cost.

Virtualization, Cluster, and HPC often require extremely good Layer 2 connectivity. Virtual Machines are typically moved inside the same IP subnet (Layer 2 domain), often using a Layer 2 mechanism like gratuitous ARP. Cluster members exchange large volumes of data among themselves and often use protocols that are not IP-based for membership, ping, keep-alive, and so on.

A 10GE solution that is wire-speed, low-latency, and completely Ethernet compliant is therefore a good match for the Data Center, even if it does not scale outside the Data Center itself. Layer 2 domains of 64,000 to 256,000 members are able to satisfy the Data Center requirement for the next few years.

To support multiple independent traffic types on the same network, it is crucial to maintain the concept of Virtual LANs and to expand the concept of Priorities (see page 20).

Switch Architecture

This section deals with the historical debate of store-and-forward versus cut-through switching. Many readers may correctly complain of having heard this debate repeatedly, with some of the players switching sides over the course of the years, and they are right!

When the speed of Ethernet was low (e.g., 10 or 100 Mbit/s), this debate was easy to win for the store-and-forward camp, since the serialization delay was the dominating one. Today, with 10GE available and 40GE and 100GE in our close future, the serialization delay is low enough to justify looking at this topic again. For example, a 1-KB frame requires approximately 1 microsecond to be serialized at the speed of 10 Gbit/s.

Today many Ethernet switches are designed with a store-and-forward architecture, since this is a simpler design. Store-and-forward adds several serialization delays inside the switch and therefore the overall latency is negatively impacted [10].

Cut-through switches have a lower latency at the cost of a more complex design, required to save the intermediate store-and-forward. This is possible to achieve on fixed configuration switches like the Nexus 5000, but much more problematic on modular switches with a high port count like the Nexus 7000.

In fixed configuration switches a single speed (for example 10 Gbit/s) is used in the design of all the components, a limit to the number of ports is selected (typically less than 128), and these simplified assumptions make cut-through possible.

In modular switches, backplane switching fabrics are multiple (also to improve high availability, modularity, and serviceability) and run dedicated links toward the linecards at a speed as high as possible. Modular switches may have thousands of ports because they may have a high number of linecards and a high number of ports per linecard. The linecards are heterogeneous (1GE, 10GE, 40GE, etc.), and the speed of the front panel ports is lower than the speed of the backplane (fabric) ports. Therefore, a store and forward between the ingress linecard and the fabric and a second one between the fabric and the egress linecard are almost impossible to avoid.

Cut-through switching is not possible if there are frames already queued for a given destination and if the speed of the egress link is higher than the speed of the ingress link (data underrun). Cut-through is typically not performed for multicast/broadcast frames.

Finally, cut-through switches cannot discard corrupted frames, since when they detect that a frame is currupted, by examining the Frame Control Sequence (FCS), they have already started transmitting that frame.

Low Latency

The latency parameter that cluster users care about is the latency incurred in transferring a buffer from the user memory space of one computer to the user memory space of another computer. The main factors that contribute to the latency are

1. The time elapsed between the moment in which the application posts the data and the moment in which the first bit starts to flow on the wire. This is determined by the zero-copy mechanism and by the capability of the NIC to access the data directly in host memory, even if this is scattered in physical memory. To keep this time low most NICs today use DMA scatter/gather operations to efficiently move frames between the memory and the NIC. This in turn is influenced by the type of protocol offload used (i.e., stateless versus TOE [TCP Offload Engine]).
2. Serialization delay: This depends only on the link speed. For example, at 10 Gbit/s the serialization of one Kbyte requires 0.8 microseconds.

3. Propagation delay: This is similar in copper and fiber; it is typically 2/3 of the speed of light and can be rounded to 200 meters/microsecond one way or to 100 meters/microsecond round-trip delay. Some people prefer to express it as 5 nanoseconds/meter, and this is correct as well. In published latency data, the propagation delay is always assumed to be zero. The size of Data Center networks must be limited to a few hundreds meters to keep low this delay, otherwise it becomes dominant and low latency cannot be achieved.
4. Switch latency varies in the presence or absence of congestion. Under congestion the switch latency is mainly due to the buffering occurring inside the switch and low latency cannot be achieved. In a noncongested situation the latency depends mainly on the switch architecture, as explained on page 9.
5. Same as in point 1, but on the receiving side.

Native Support for Storage Traffic

The term native support for storage traffic indicates the capability of a network to act as a transport for the SCSI protocol. Figure 1-6 illustrates possible alternative SCSI transports.

SCSI was designed assuming the underlying physical layer was a short parallel cable, internal to the computer, and therefore extremely reliable. Based on this assumption, SCSI is not efficient in recovering from transmission errors. A frame loss may cause SCSI to time-out and recover in up to one minute.

For this reason, when the need arose to move the storage out of the servers in the storage arrays, the Fibre Channel protocol was chosen as a transport for SCSI. Fibre Channel, through its buffer-to-buffer (B2B) credit-based flow control scheme, guarantees the same frame delivery reliability of the SCSI parallel bus and therefore is a good match for SCSI.

Ethernet does not have a credit-based flow control scheme, but it does have a PAUSE mechanism. A proper implementation of the PAUSE mechanism achieves results identical to a credit-based flow control scheme, in a distance-limited environment like the Data Center.

To support I/O consolidation (i.e., to avoid interference between different classes of traffic) PAUSE needs to be extended per Priority (see page 20).

RDMA Support

Cluster applications require two message types:

- Short synchronization messages among cluster nodes with minimum latency.
- Large messages to transfer buffers from one node to another without CPU intervention. This is also referred to as Remote Direct Memory Access (RDMA).

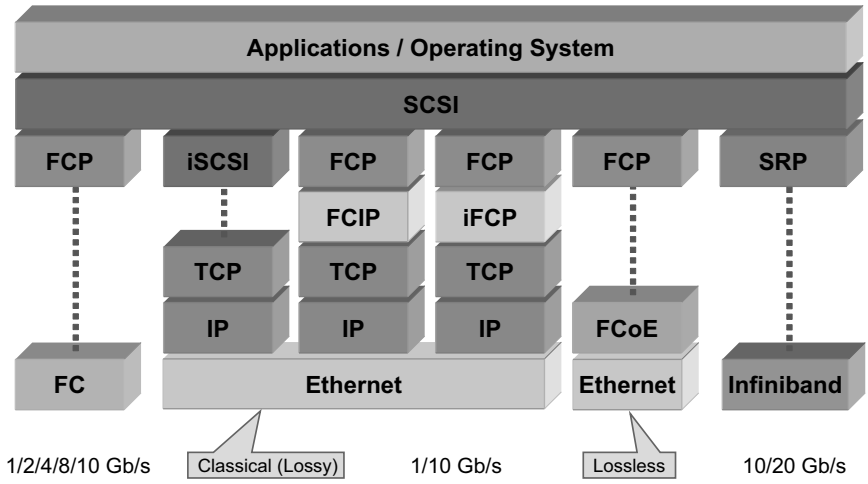


Figure 1-6 SCSI Transports

In the latter case the buffer resides in the user memory (rather than in the kernel) of a process. The buffer must be transferred to the user memory of another process. User memory is virtual memory, and it is therefore scattered in physical memory.

The RDMA operation must happen without CPU intervention, and therefore the NIC must be able to accept a command to transfer a user buffer, gather it from physical memory, implement a reliable transport protocol, and transfer it to the other NIC. The receiving NIC must verify the integrity of the data, signal the successful transfer or the presence of errors, and scatter the data in the destination host physical memory without CPU intervention.

RDMA requires in-order reliable delivery of its messages by the underlying transport.

In the IP world, there is no assumption on the reliability of the underlying network. iWARP (Internet Wide Area RDMA Protocol) is an Internet Engineering Task Force (IETF) update of the RDMA Consortium's RDMA over TCP standard. iWARP is layered above TCP, which guarantees in-order delivery. Packets dropped by the underlying network are recovered by TCP through retransmission.

Over networks with limited scope, such as Data Center networks, in-order frame delivery can be achieved without using a heavy protocol such as TCP. As an example, in-order frame delivery is successfully achieved by Fibre Channel fabrics and Ethernet networks.

As discussed in Chapter 2, Ethernet can be extended to become lossless. In Lossless Ethernet dropping happens only because of catastrophic events, like transmission errors or topology reconfigurations. The RDMA protocol may therefore be designed with the assumption

that frames are normally delivered in order without any frame being lost. Protocols like LLC2, HDLC, LAPB, and so on work well if the frames are delivered in order and if the probability of frame drop is low.

Lossless Ethernet can also be integrated with a congestion control mechanism at Layer 2.

Another important requirement for RDMA is the support of standard APIs. Among the many proposed, RDS, IB verbs, SDP, and MPI seem the most interesting. RDS is used in the database community and MPI is widely adopted in the HPC market.

Open Fabrics Alliance (OFED) is currently developing a unified, open-source software stack for the major RDMA fabrics.

Index

Symbols

A

B

C

D

E

F

H

I

L

M

N

O

P

Q

R

S

T

U

V

W

Z