Comparing Server I/O Consolidation Solutions: iSCSI, InfiniBand and FCoE

Dr. Walter Dey, Cisco Systems
The material contained in this tutorial is copyrighted by the SNIA.

Member companies and individuals may use this material in presentations and literature under the following conditions:

- Any slide or slides used must be reproduced without modification
- The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.

This presentation is a project of the SNIA Education Committee.
Abstract

Comparing Server I/O Consolidation: iSCSI, Infiniband and FCoE

This tutorial gives an introduction to Server I/O consolidation, having one network interface technology (Standard Ethernet, Data Center Ethernet, InfiniBand), to support IP applications and block level storage (iSCSI, FCoE and SRP/iSER) applications. The benefits for the end user are discussed: less cabling, power and cooling. For these 3 solutions, iSCSI, Infiniband and FCoE, we compare features like Infrastructure / Cabling, Protocol Stack, Performance, Operating System drivers and support, Management Tools, Security and best design practices.
Agenda

➤ Definition of Server I/O Consolidation

➤ Why Server I/O Consolidation

➤ Introducing the 3 solutions
  iSCSI
  InfiniBand
  FCoE

➤ Differentiators

➤ Conclusion
Definition of Server I/O Consolidation
What is Server I/O Consolidation

- IT Organizations operate multiple parallel networks
  - IP Applications (including NFS, NAS,…) over a Ethernet network *)
  - SAN over a Fibre Channel network
  - HPC/IPC over an InfiniBand network **)

- Server I/O consolidation combines the various traffic types onto a single interface and single cable

- Server I/O consolidation is the first phase for a Unified Fabric (single network)

*) In this presentation we cover only Block Level Storage solutions, not File Level (NAS, NFS,..)

**) For the remaining part, we don’t cover HPC; for lowest latency requirements, InfiniBand is the best and most appropriate technology.
I/O Consolidation Benefits

- Adaptor: NIC for Ethernet/IP, HCA for InfiniBand, Converged Network Adaptor (CNA) for FCoE
- Customer Benefit: Fewer NIC’s, HBA’s and cables, lower CapEx, OpEx
Why Server I/O Consolidation?
The drive for I/O Consolidation

- Multicore – Multisocket CPUs
- Server Virtualization software (Hypervisor)
- High demand for I/O bandwidth
- Reductions in cables, power and cooling, therefore reducing OpEx/CapEx
- Limited number of interfaces for Blade Servers
- Consolidated Input into Unified Fabric
Virtual networks growing faster and larger than physical
- Network admins are getting involved in virtual interface deployments
- Network access layer needs to evolve to support consolidation and mobility

Multi-core Computing driving Virtualization & new networking needs
- Driving SAN attach rates higher (10% → 40% → Growing)
- Driving users to plan now for 10GE server interfaces

Virtualization enables the promise of blades
- 10GE and FC are highest growth technologies within blades
- Virtualization and Consolidated I/O removes blade limitation

Network Virtualization enables CPU & I/O Intensive Workloads to be Virtualized
- Enable broader adoption of x86 class servers
Multi-Core CPU architectures allowing bigger and multiple workloads on the same machine

Server virtualization driving the need for more bandwidth per server due to server consolidation

Growing need for network storage driving the demand for higher network bandwidth to the server

Multi-Core CPUs and Server Virtualization driving the demand for higher bandwidth network connections
### Evolution of Ethernet Physical Media

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cable</th>
<th>Distance</th>
<th>Power (each side)</th>
<th>Transceiver Latency (link)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SFP+ CU Copper</td>
<td>Twinax</td>
<td>10m</td>
<td>~0.1W</td>
<td>~0.1μs</td>
</tr>
<tr>
<td>SFP+ USR ultra short reach</td>
<td>MM OM2</td>
<td>10m</td>
<td>1W</td>
<td>~0</td>
</tr>
<tr>
<td></td>
<td>MM OM3</td>
<td>100m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SFP+ SR short reach</td>
<td>MM 62.5μm</td>
<td>82m</td>
<td>1W</td>
<td>~0</td>
</tr>
<tr>
<td></td>
<td>MM 50μm</td>
<td>300m</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10GBASE-T</td>
<td>Cat6</td>
<td>55m</td>
<td>~8W</td>
<td>2.5μs</td>
</tr>
<tr>
<td></td>
<td>Cat6a/7</td>
<td>100m</td>
<td>~8W</td>
<td>2.5μs</td>
</tr>
<tr>
<td></td>
<td>Cat6a/7</td>
<td>30m</td>
<td>~4W</td>
<td>1.5μs</td>
</tr>
</tbody>
</table>
Introducing the three solutions
Server I/O Consolidation Solutions

- **iSCSI**
  - LAN: Based on Ethernet and TCP/IP
  - SAN: Encapsulates SCSI in TCP/IP

- **InfiniBand**
  - LAN: Transports IP over InfiniBand (IPoIB); Socket Direct Protocol (SDP) between IB attached servers
  - SAN: Transports SCSI over Remote DMA protocol (SRP) or iSCSI Extensions for RDMA (iSER)
  - HPC/IPC: Message Passing Interface (MPI) over InfiniBand network

- **FCoE**
  - LAN: Based on Ethernet (Data Center Ethernet) and TCP/IP
  - SAN: Maps and transports Fibre Channel over Data Center Ethernet (lossless Ethernet) *)

* Data Center Ethernet is an architectural collection of Ethernet extensions designed to improve Ethernet networking and management in the Data Center
Encapsulation technologies

- **iSCSI**
  - FCP*
  - iSCSI
  - FCIP
  - TCP
  - IP
  - Fibrechannel
    - 1, 2, 4, (8), 10 Gbps

- **FCoE**
  - FCP*
  - iFCP
  - TCP
  - IP
  - Ethernet
    - 1, 10 Gbps

- **InfiniBand**
  - SRP
  - iSCSI
  - iSER
  - InfiniBand
    - 10, 20 Gbps

**Operating System / Applications**

- **SCSI Layer**
  - FCP*
  - FCIP
  - iFCP
  - TCP
  - IP
  - FCoE
  - Data Center Ethernet
    - 1, 10 Gbps
  - (1), 10 Gbps

* Includes FC Layer

Comparing Server I/O Consolidation: iSCSI, InfiniBand and FCoE

© 2008 Storage Networking Industry Association. All Rights Reserved.
A SCSI transport protocol that operates over TCP

- Encapsulates SCSI CDBs (operational commands: e.g. read or write) and data into TCP/IP byte-streams (defined by SAM-2—SCSI Architecture Model 2)
- Allows iSCSI Initiators to access IP-based iSCSI targets (either natively or via iSCSI-to-FC gateway)

Standards status

- RFC 3720 on iSCSI
- Collection of RFCs describing iSCSI
  - RFC 3347—iSCSI Requirements
  - RFC 3721—iSCSI Naming and Discover
  - RFC 3723—iSCSI Security

Broad industry support

- Operating System vendors support their iSCSI drivers
- Gateway (Routers, Bridges) and Native iSCSI storage arrays
iSCSI Messages

Contains routing information so that the message can find its way through the network

Provides information necessary to guarantee delivery

Explains how to extract SCSI commands and data

Ethernet Header

IP Header

TCP Header

iSCSI Header

iSCSI Data

22 Bytes

20 Bytes

20 Bytes

48 Bytes

nn

4 Bytes

PDU

TCP Segment

IP Packet

Ethernet Frame

Contains routing information so that the message can find its way through the network

Provides information necessary to guarantee delivery

Explains how to extract SCSI commands and data

Ethernet Header

IP Header

TCP Header

iSCSI Header

iSCSI Data

22 Bytes

20 Bytes

20 Bytes

48 Bytes

nn

4 Bytes

PDU

TCP Segment

IP Packet

Ethernet Frame
iSCSI Topology

Allows I/O consolidation

- iSCSI is proposed today as an I/O consolidation option
- Native (iSCSI Storage Array) and Gateway solutions
View from Operating System

- Operating System sees:
  - 1 Gigabit Ethernet adapter
  - iSCSI Initiator
iSCSI based I/O Consolidation

- Overhead of TCP/IP Protocol
- It’s SCSI not FC
- LAN/Metro/WAN ( Routable )
- Security of IP protocols ( IPsec )
- Stateful gateway ( iSCSI <-> FCP )
- Mainly 1G Initiator ( Server )
- 10G for iSCSI Target recommended
- Can use existing Ethernet switching infrastructure
- Offload Engine ( TOE ) suggested ( virtualized environment support ?)
- QoS or separate VLAN for storage traffic suggested
- New Management Tools
- Might require different Multipath Software

Overhead of TCP/IP Protocol
- It’s SCSI not FC
- LAN/Metro/WAN ( Routable )
- Security of IP protocols ( IPsec )
- Stateful gateway ( iSCSI <-> FCP )
- Mainly 1G Initiator ( Server )
- 10G for iSCSI Target recommended
- Can use existing Ethernet switching infrastructure
- Offload Engine ( TOE ) suggested ( virtualized environment support ?)
- QoS or separate VLAN for storage traffic suggested
- New Management Tools
- Might require different Multipath Software
InfiniBand

- Standards-based interconnect
  - http://www.infinibandta.org
- Channelized, connection-based interconnect optimized for high performance computing
- Supports server and storage attachments
- Bandwidth Capabilities (SDR/DDR)
  - 4x—10/20 Gbps: 8/16 Gbps actual data rate
  - 12x—30/60 Gbps: 24/48 Gbps actual data rate
- Built-in RDMA as core capability for inter-CPU communication

OS / Applications
SCSI Layer
  - SRP
  - iSCSI
  - iSER

IB

10, 20 Gbps (4X SDR/DDR)
InfiniBand:
SCSI RDMA Protocol (SRP)

- SCSI Semantics over RDMA fabric
- Provides High Performance block-level storage access
- Not IB specific - Standard specified by T10
  http://www.t10.org
- Host drivers tie into standard SCSI I/F in kernel
- Storage appears as normal SCSI/FC disks to local host
- Can be used for end-to-end IB storage (No FC)
- Can be used for SAN Boot over IB
InfiniBand: iSCSI Extensions for RDMA (iSER)

- IETF Standard
- Enables iSCSI to take advantage of RDMA.
- Mainly offloads the data path
- Leverages iSCSI management and discovery architecture
- Simplifies iSCSI protocol details such as data integrity management and error recovery
- Not IB Specific
- Needs a iSER Target to work end to end
InfiniBand Gateway Topology: Gateways for Network and Storage

- IP Application Traffic
- Block Level Storage

Single InfiniBand link for:
- Storage
- Network

Fibrechannel Fabric

Infiniband Switch with stateful Gateways

Ethernet LAN

Fibre Channel to InfiniBand gateway for storage access

Ethernet to InfiniBand gateway for LAN access

Fibrechannel

Ethernet

InfiniBand
Physical vs. Logical view

Physical View
- Servers connected via IB
- SAN attached via public AL
- Ethernet attached via Gig Etherchannel

Logical View
- Hosts present WWNN on SAN
- Hosts present IP address on VLAN
View from Operating System
InfiniBand based I/O Consolidation

Requirements:

- Requires new Eco system (HCA, cabling, switches)
- Mostly copper cabling, limited distance
- Datacenter protocol
- New driver (SRP)
- Stateful Gateway from SRP to FCP (unless native IB attached disk array)
- RDMA capability of HCA used
- Low CPU overhead
- Payload is SCSI not FC
- Concept of Virtual links and QoS in InfiniBand

OS / Applications

SCSI Layer

SRP
iSCSI
iSER

IB
IB

10, 20 Gbps (4X SDR/DDR)
- From a Fibre Channel standpoint it’s Fibrechannel encapsulated in Ethernet
- From an Ethernet standpoint it’s just another ULP (Upper Layer Protocol)
- FCoE is an extension of Fibre Channel onto a Lossless (Data Center) Ethernet fabric
- FCoE is managed like FC at initiator, target, and switch level, completely based on the FC model
  - Same host-to-switch and switch-to-switch behavior of FC
  - in order frame delivery or FSPF load balancing
  - WWNs, FC-IDs, hard/soft zoning, DNS, RSCN
- Standards Work in T11, IEEE and IETF not yet final

* Includes FC Layer
FCoE Enablers

- 10Gbps Ethernet
- Lossless Ethernet (Data Center Ethernet)
  - Matches the B2B credits used in Fibrechannel to provide a lossless service
- Ethernet jumbo frames (2180 Bytes)
  - Max FC frame payload = 2112 bytes

Ethernet V2 Frame, Ethertype = FCoE

Same as a physical FC frame

Control information: version, ordered sets (SOF, EOF)
Data Center Ethernet

्Enhanced Ethernet for Data Center Applications

- Priority Flow Control (Priority Pause) *
- Link Scheduling
- Congestion Management
- Layer 2 Multipathing
- Configuration Management

Transport of FCoE

Enabling Technology for I/O Consolidation and Unified Fabric

*) T11 BB 5 group has only required that Ethernet switches have standard Pause, and baby Jumbo frame capability; which means no I/O consolidation support.
FCoE Enabler: Priority Flow Control

- Enables lossless Fabrics for each class of service
- PAUSE sent per virtual lane when buffers limit exceeded
- Network resources are partitioned between VL’s (E.g. input buffer and output queue)
- The switch behavior is negotiable per VL
- InfiniBand uses a similar mechanism for multiplexing multiple data streams over a single physical link
FCoE Enabler:
Consolidated Network Adapter (CNA)
View from Operating System

- Standard drivers
- Same management
- Operating System sees:
  - Dual port 10 Gigabit Ethernet adapter
  - Dual Port 4 Gbps Fibre Channel HBAs
FCoE I/O Consolidation Topology

**FCoE Target:**
- Dramatic reduction in adapters, switch ports and cabling
  - 4 cables to 2 cables per server
- Seamless connection to the installed base of existing SANs and LANs
- High performance frame mappers vs. gateway bottlenecks
- Effective sharing of high bandwidth links
- Consolidated network infrastructure
  - Faster infrastructure provisioning
- Lower TCO
FCoE based I/O Consolidation

- FCP layer untouched
- Requires Baby Jumbo Frames (2180 Bytes)
- Nonroutable Datacenter protocol
- Datacenter wide VLAN’s
- Same management tools as for Fibre Channel
- Same drivers as for Fibre Channel HBA’s
- Same Multipathing software
- Simplified certifications with storage subsystem vendors
- Requires lossless (10G) Ethernet switching fabric
- May require new host adaptors (unless FCoE software stack)

* Includes FC Layer
Differentiators
<table>
<thead>
<tr>
<th></th>
<th>iSCSI</th>
<th>FCoE</th>
<th>IB-SRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payload</td>
<td>SCSI</td>
<td>Fibre Channel</td>
<td>SCSI</td>
</tr>
<tr>
<td>Transport</td>
<td>TCP/IP</td>
<td>Data Center Ethernet</td>
<td>InfiniBand</td>
</tr>
<tr>
<td>Scope</td>
<td>LAN/MAN/WAN</td>
<td>Datacenter</td>
<td>Datacenter</td>
</tr>
<tr>
<td>Bandwidth/Performance</td>
<td>Low/Medium</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>CPU Overhead</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Gateway Overhead</td>
<td>High</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>FC Security Model</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>FC Software on Host</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>FC Management Model</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Initiator Implementation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Target Implementation</td>
<td>Yes</td>
<td>Yes/Future</td>
<td>Yes</td>
</tr>
<tr>
<td>IP Routable</td>
<td>Yes</td>
<td>No</td>
<td>N/A</td>
</tr>
</tbody>
</table>
## Storage Part of I/O Consolidation

<table>
<thead>
<tr>
<th>Feature</th>
<th>iSCSI</th>
<th>FCoE</th>
<th>IB-SRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virtual Lanes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Congestion Control</td>
<td>TCP</td>
<td>Priority Flow Control</td>
<td>Credit based</td>
</tr>
<tr>
<td>Gateway Functionality</td>
<td>stateful</td>
<td>stateless</td>
<td>stateful</td>
</tr>
<tr>
<td>Connection Oriented</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Access Control</td>
<td>IP/VLAN</td>
<td>VLAN / VSAN</td>
<td>Partitions</td>
</tr>
<tr>
<td>RDMA primitives</td>
<td>defined</td>
<td>defined</td>
<td>defined</td>
</tr>
<tr>
<td>Latency</td>
<td>100s of us</td>
<td>10s of us</td>
<td>us</td>
</tr>
<tr>
<td>Adapter</td>
<td>NIC</td>
<td>CNA</td>
<td>HCA</td>
</tr>
</tbody>
</table>
Conclusion
Encapsulation Technologies

Comparing Server I/O Consolidation: iSCSI, Infiniband and FCoE

Encapsulation Technologies

iSCSI

FCoE

InfiniBand

Operating System / Applications

SCSI Layer

Ethernet

FCP*

iSCSI

IP

TCP

Fibrechannel

1, 2, 4, (8), 10 Gbps

1, 10 Gbps

10, 20 Gbps

FCP*

FCIP

iSCSI

TCP

IP

Fibrechannel

1, 2, 4, (8), 10 Gbps

IP

TCP

Ethernet

Data Center Ethernet

(1), 10 Gbps

SRP

iSCSI

iSER

InfiniBand

* Includes FC Layer

Comparing Server I/O Consolidation: iSCSI, Infiniband and FCoE
© 2008 Storage Networking Industry Association. All Rights Reserved.
Conclusion

- Server I/O Consolidation is driven by high I/O bandwidth demand
- I/O Bandwidth demand is driven by Multicore / Socket Server and Virtualization
- TCP/IP (iSCSI), Data Center Ethernet (FCoE) and InfiniBand (SRP, iSER) are generic transport protocols allowing Server I/O Consolidation
- Server I/O Consolidation is the first phase, consolidating input into a Unified Fabric
Thank You !
Check out SNIA Tutorial:

- Fibre Channel Technologies: Current and Future
- IP Storage Protocols - iSCSI
- InfiniBand Technology Overview
- FCoE: Fibre Channel over Ethernet
Q&A / Feedback

Please send any questions or comments on this presentation to SNIA: tracknetworking@snia.org

Many thanks to the following individuals for their contributions to this tutorial.
- SNIA Education Committee

Gilles Chekroun  
Errol Roberts  
Walter Dey
Howard Goldstein  
Bill Lulofs  
Dror Goldenberg
Marco Di Benedetto  
Carlos Pereira  
James Long
References

http://www.fcoe.com/

http://www.t11.org/fcoe

http://www.fibrechannel.org/OVERVIEW/FCIA_SNW_FCoE_WP_Final.pdf

http://www.fibrechannel.org/OVERVIEW/FCIA_SNW_FCoE_flyer_Final.pdf

http://www.fibrechannel.org/FCoE.html