Innovations in Storage Networking: Next-Gen Storage Networks for Next-Gen Data Centers

Dennis Martin
President, Demartek
Agenda

Next Generation Data Center Storage Networking

- About Demartek
- What drives the need for increased bandwidth?
- Ethernet – 10 Gigabit and futures
- Fibre Channel – 16 Gigabit and futures
- Converged networks (DCB, FCoE, iSCSI, etc.)
- Network (I/O) virtualization (NPIV, SR-IOV, etc.)
- Cabling considerations and recommendations
- Performance results
About Demartek

● Industry analysis with on-site test lab
● Lab includes servers, networking and storage infrastructure
  - Fibre Channel – 4, 8 & 16 Gbps
  - Ethernet – 1 & 10 Gbps: NFS, SMB (CIFS), iSCSI & FCoE
  - Servers – 8+ cores, large RAM
  - Virtualization – ESX, Hyper-V, Xen
● We prefer to run real-world applications to test servers and storage solutions
  - Currently testing various SSD, 16GFC and other technologies
● Website: www.demartek.com
The Need for More Bandwidth

What is Driving This?

- **Server virtualization**
  - How many VMs per physical box do you deploy?
  - Compare the number of VMs today vs. one & two years ago
- **Application growth**
  - Applications processing more data today
- **SSD**
  - How many are deploying enterprise SSDs today?
- **New generation of servers (1H 2012)**
  - New servers support 10GigE on the motherboard
  - PCI-Express 3.0 (up to 40 PCIe lanes per processor)
10 Gigabit specification was ratified in 2002
- 10GBASE-T specification was ratified in 2006
Early adopters – Switch-to-switch trunking
Adoption of 10GigE increasing over the last 2-3 years
Blade server chassis have two or more 10GigE ports
Connector types:
- SFP+ – Fiber-optic cables and direct attach copper
- RJ45 – Twisted-pair “traditional Ethernet” (10GBASE-T)
  - 10GBASE-T a few years behind SFP+ but expected to gain acceptance relatively quickly
Some 10GigE switches support both SFP+ and RJ45
Ethernet – 1GigE vs. 10GigE

● 1GigE
  - Not unusual to see 4, 6 or 8 NIC ports in a server
    ● Requires 4, 6, or 8 Ethernet cables
  - Used for management and network/storage traffic
  - 1GigE NICs can be quad-port, dual-port or single-port
  - Can consume two, four or more I/O slots in a server

● 10GigE
  - A dual-port 10GigE NIC provides bandwidth and failover
  - May be a good choice for 1U servers that have few I/O slots
  - Slot requirements
    ● Dual-port 10GigE NIC – PCIe 2.0 x8
    ● Single-port 10GigE NIC – PCIe 2.0 x4 or PCIe 1.0 x8
Ethernet – 40GigE and 100GigE

- IEEE 802.3ba (40GigE and 100GigE) ratified June 2010
- The fastest Ethernet cables and connectors today are 10 Gbps per lane or channel
- Higher speeds today are achieved by bundling
  - 40GigE today = 4 x 10 Gbps together
  - 100GigE today = 10 x 10 Gbps together
- 25 Gbps connectors expected in late 2012 or 2013
  - These connectors support up to 28 Gbps (“25/28G”)
  - 100GigE (future) = 4 x 25 Gbps together
  - 250GigE (future) = 10 x 25 Gbps together
  - End-user products with 25 Gbps expected in 2013 or 2014
Fibre Channel – 16 Gigabit

- 16 GFC is backward compatible with 4 GFC and 8 GFC
- Uses 14 Gbps single-lane connectors
  - Doubles speed of 8 GFC due to newer 64b/66b encoding
- The first 16 GFC switches shipped in 2011
- The first 16 GFC HBAs shipped in 2011
  - Others announced in 2012
  - Some of these HBAs can also function as 10 GB NICs
- 16 GFC storage targets expected late 2012 or in 2013
- Fibre Channel speeds and server slots
  - 4 Gb: PCI-X 2.0, PCIe 1.0
  - 8 Gb: PCIe 2.0 x4 or PCIe 1.0 x8
  - 16 Gb: PCIe 2.0 x8
Fibre Channel – History

- **FC SAN interface** – Doubles in speed every 3-4 years*
  *Source: Fibre Channel Industry Association (FCIA)*
  - 1997 – 1 Gb/s FC products
  - 2001 – 2 Gb/s FC products
  - 2005 – 4 Gb/s FC products
  - 2008 – 8 Gb/s FC products
  - 2009 – 10 Gb/s FCoE products
  - 2011 – 16 Gb/s FC products

- **FC Disk Drive interface**
  - 4 Gb/s was the highest speed for disk drive FC interface
  - HDD/SSD vendors have moved to 6 Gb/s SAS for enterprise drives
    - First 12 Gb/s SAS drives announced in Spring 2012
Fibre Channel – 32 Gigabit and 64 Gigabit

● Formal statement of direction:

“The INCITS Technical Committee T11 is currently working on the 32 GFC Fibre Channel specifications. The 32 GFC specifications are expected to be stabilized in mid-2013. Work has not yet begun in T11 for developing the 64 GFC specifications, but 64 GFC is on the FCIA Speed roadmap.”

Steve Wilson, Director of Technology and Standards, Brocade and INCITS Technical Committee T11 Chairman

● 32 GFC will use 28 Gbps connectors (25/28G), and will double the speed of 16 GFC
Converged Networks

- Combined LAN and SAN networks
  - Lossless features of Fibre Channel with ubiquity of Ethernet
  - Within a rack (short-term)
  - Entire infrastructure (long-term)
- DCB – Data Center Bridging
  - Enhanced Ethernet to support FC storage traffic and more
- FCoE – Fibre Channel over Ethernet
  - First major application for DCB
- CNA – Converged Network Adapter
  - Supports 10 Gb Ethernet and 10 Gb FCoE at the same time on the same wire
Converged Networks Standards

● DCB
  - A collection of architectural Ethernet extensions designed to improve Ethernet networking and management in the data center
    ● Identifies and manages different traffic types on the same connection using Enhanced Transmission Selection (ETS)

● Ethernet
  - IEEE 802.1: 802.1Qau, 802.1Qaz, 802.1Qbb

● Fibre Channel
  - INCITS T11: FC-BB-5 (“FCoE”) approved June 2009
  - INCITS T11: FC-BB-6 enhancements to FCoE underway
    ● Possible ratification by end of 2012
Converged Networks – Fibre Channel over Ethernet

- FCoE places the FC protocol on a new physical link
  - Uses Lossless Ethernet (DCB) physical links
  - Protocol and behavior is the same as traditional FC

- FCoE fabrics must be built with FCoE/DCB switches
  - Interoperate with traditional FC fabrics
  - Support all FC advanced features
  - Operate identically on FCoE and FC fabrics
Converged Networks – Switch and Adapter Technology

● DCB/FCoE Switches
  - Blades for blade-architecture switches
  - Top-of-rack switches
  - Includes 10 Gb Enhanced Ethernet (DCB)
  - Supports FCoE and iSCSI
  - Optional – 4 or 8 Gb native FC ports
  - Some switches offer “universal” ports

● Adapter offload characteristics
  - CNAs – FC & FCoE supported in hardware
  - 10GbE NICs – FC & FCoE supported by software
  - Ethernet – Similar to good server-class NIC
Converged Networks – Organizational Issues

- In typical large shops today, networking and storage are separate departments
  - Networking – Dynamic (more changes)
  - Storage – Stable (fewer changes)

- Other areas of convergence
  - Consider voicemail & email

- Those that learn networking and storage will be in the best position
Network (I/O) Virtualization

- Virtualizing the *I/O path* between a server and an external device
- De-couple the logical from the physical
  - Hardware can be split into smaller logical units
  - Hardware can be represented as multiple units
  - Hardware can be combined into larger units
- Can apply to anything that performs I/O or works with an I/O adapter in a server, such as:
  - Ethernet Network Interface Cards (NICs) and switches
  - Disk Controllers (including RAID controllers)
  - Fibre Channel Host Bus Adapters (HBAs) and switches
  - SSDs mounted on internal cards
Network Virtualization – Existing Forms

- **NIC Teaming**
  - A virtual NIC composed of two or more physical NICs

- **Virtual LAN**
  - Multiple, smaller logical LANs within a physical LAN infrastructure

- **Fibre Channel NPIV**
  - Multiple logical N_Port IDs sharing one physical N_Port

- **Virtual SAN Fabrics**
  - Multiple, smaller logical SANs within a physical SAN infrastructure
Network Virtualization – Single-root I/O Virtualization (SR-IOV)

- Multiple VMs sharing one I/O adapter
- Bandwidth of the I/O adapter is shared among the VMs
- Virtual adapters created and managed by SR-IOV adapter (not hypervisor)
- Improved performance for VMs and their apps (near-native) by offloading I/O management and mapping functions to the adapter
Network Virtualization – Multi-root I/O Virtualization (MR-IOV)

- Multiple servers & VMs sharing one I/O adapter
- Bandwidth of the I/O adapter is shared among the servers
- The I/O adapter is placed into a separate chassis
- Bus extender cards are placed into the servers
Network Virtualization – SR-IOV Virtual Functions

- Virtual functions are the way that the adapter makes multiple versions of itself visible to the Hypervisor.
- The Hypervisor assigns a VF to a guest:
  - Many VFs can be created per physical port
  - Hypervisor has no visibility into the VF
- Guest sees a new adapter that it can use for anything that adapter can do.
Network Virtualization – SR-IOV Availability Today

- Available today for Ethernet only
- SR-IOV Support in Operating System & Hypervisor:
  - Citrix XenServer
  - RHEL 6 KVM
  - VMware 5.1
    - vMotion not supported with SR-IOV
  - Windows Server 2012 (Hyper-V and guest)
    - Live Migration is supported with SR-IOV
Network Virtualization – SR-IOV Dependencies

● Hardware
  - Many, but not all PCIe 2.0 and 3.0 servers meet these criteria
  - Processor (CPU) support (specific support for SR-IOV)
  - Motherboard support (chipset, etc.)
  - BIOS support
  - SR-IOV capable NIC

● Software
  - Hypervisor IOV-enabled virtual switch
  - VF driver for the guest O.S.
  - Windows registry: IovEnableOverride in parent partition
Encoding Schemes

Data Transfer “On the Wire”

- **8b/10b**
  - For every 8 bits, adds 2 bits for command and control
  - 20% overhead = \( \frac{10-8}{10} \)

- **64b/66b**
  - Used by 10 GigE and 16 GFC
  - For every 64 bits, adds 2 bits for command and control
  - 3% overhead = \( \frac{66-64}{66} \)

- **128b/130b**
  - Used by PCIe 3.0
  - For every 128 bits, adds 2 bits for command and control
  - 1.5% overhead = \( \frac{130-128}{130} \)
PCI-Express

- Bus used in modern computers for I/O adapters
- Measured in gigatransfers/second (GT/s)
  - Bandwidth specified by indicating number of lanes such as “x1”, “x2”, etc., and generally spoken as “by 1”, “by 2”, etc.

<table>
<thead>
<tr>
<th></th>
<th>GT/s</th>
<th>Encoding</th>
<th>x1</th>
<th>x2</th>
<th>x4</th>
<th>x8</th>
<th>x16</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCIe 1.x</td>
<td>2.5</td>
<td>8b/10b</td>
<td>250 MB/s</td>
<td>500 MB/s</td>
<td>1 GB/s</td>
<td>2 GB/s</td>
<td>4 GB/s</td>
</tr>
<tr>
<td>PCIe 2.x</td>
<td>5</td>
<td>8b/10b</td>
<td>500 MB/s</td>
<td>1 GB/s</td>
<td>2 GB/s</td>
<td>4 GB/s</td>
<td>8 GB/s</td>
</tr>
<tr>
<td>PCIe 3.x</td>
<td>8</td>
<td>128b/130b</td>
<td>1 GB/s</td>
<td>2 GB/s</td>
<td>4 GB/s</td>
<td>8 GB/s</td>
<td>16 GB/s</td>
</tr>
</tbody>
</table>

- **PCIe 4.0** – In November 2011, the PCI-SIG announced the approval of 16 gigatransfers per second as the bit rate for the next generation of PCIe architecture, known as PCIe 4.0. Final specifications are expected in 2014-2015
Fibre Channel Adapter Specifications Overview

- Fibre Channel can run in full-duplex mode, but storage protocols generally operate in half-duplex mode
  - Throughput numbers below are half-duplex (one-way)
- Host Adapter Requirements are for dual-port cards

<table>
<thead>
<tr>
<th>Fibre Channel</th>
<th>Throughput (MBps)</th>
<th>Encoding</th>
<th>Line Rate (Gbaud)</th>
<th>Host Adapter Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1GFC</td>
<td>100</td>
<td>8b/10b</td>
<td>1.0625</td>
<td>PCI-X</td>
</tr>
<tr>
<td>2GFC</td>
<td>200</td>
<td>8b/10b</td>
<td>2.125</td>
<td>PCI-X</td>
</tr>
<tr>
<td>4GFC</td>
<td>400</td>
<td>8b/10b</td>
<td>4.25</td>
<td>PCI-X 2.0 or PCIe 1.0 (x4)</td>
</tr>
<tr>
<td>8GFC</td>
<td>800</td>
<td>8b/10b</td>
<td>8.5</td>
<td>PCIe 1.0 (x8) or PCIe 2.0 (x4)</td>
</tr>
<tr>
<td>16GFC</td>
<td>1600</td>
<td>64b/66b</td>
<td>14.025</td>
<td>PCIe 2.0 (x8) or PCIe 3.0 (x4)</td>
</tr>
</tbody>
</table>
Cabling Considerations and Recommendations – Fiber-optic Cables

- Fiber optic cabling service life – 15 to 20 years
  - The choices made today need to support legacy, current and emerging standards

- **Recommendation** – OM4 cables for current and future
  - OM4 will support 40/100 GigE and higher speeds of FC

<table>
<thead>
<tr>
<th>Jacket color</th>
<th>OM1</th>
<th>OM2</th>
<th>OM3</th>
<th>OM4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orange</td>
<td>300m</td>
<td>500m</td>
<td>860m</td>
<td>–</td>
</tr>
<tr>
<td>1 Gb/s</td>
<td>150m</td>
<td>300m</td>
<td>500m</td>
<td>–</td>
</tr>
<tr>
<td>2 Gb/s</td>
<td>70m</td>
<td>150m</td>
<td>380m</td>
<td>400m</td>
</tr>
<tr>
<td>4 Gb/s</td>
<td>21m</td>
<td>50m</td>
<td>150m</td>
<td>190m</td>
</tr>
<tr>
<td>8 Gb/s</td>
<td>33m</td>
<td>82m</td>
<td>Up to 300m</td>
<td>Up to 400m</td>
</tr>
<tr>
<td>10 Gb/s</td>
<td>15m</td>
<td>35m</td>
<td>100m</td>
<td>125m</td>
</tr>
<tr>
<td>16 Gb/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cabling Considerations and Recommendations – Copper Cables

● 10 GigE – SFP+ Copper
  - SFP+ copper cables are known as Direct Attach Copper (DAC)
  - SFP+ “transceiver” is directly attached to the cable
  - Common lengths of 10 GigE DAC are 3 and 5 meters

● 10 GigE – RJ45 / 10GBASE-T
  - Cables must be certified to at least 500MHz to ensure 10GBASE-T compliance
  - **Recommendation** – Cat6a & Cat7 up to 100 meters
  - Cat6 can be used up to 55 meters, but should be tested first
  - Cat5e is not recommended for 10 GigE
As interface speeds increase, expect increased usage of fiber-optic cables and connectors for most interfaces

- At higher Gigabit speeds, passive copper cables and interconnects become too “noisy” except for short distances (within a rack or to adjacent racks)

- Expect to see “active copper” for some higher-speed connection types
## Connectors

**Single-lane – SFP, SFP+**

**Four-lane – QSFP, QSFP+**

<table>
<thead>
<tr>
<th></th>
<th>SFP</th>
<th>SFP+</th>
<th>QSFP+</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ethernet</strong></td>
<td>1GbE</td>
<td>10GbE</td>
<td>40GbE</td>
</tr>
<tr>
<td><strong>Fibre Channel</strong></td>
<td>1GFC, 2GFC, 4GFC</td>
<td>8GFC, 16GFC</td>
<td>–</td>
</tr>
<tr>
<td><strong>Infiniband</strong></td>
<td>–</td>
<td>–</td>
<td>QDR, FDR</td>
</tr>
</tbody>
</table>
Demartek Test Lab Performance Results

- Unified Storage – 10GbE with FCoE, NFS, iSCSI
- 16G Fibre Channel
Performance Results –
10 GigE, NFS, iSCSI and FCoE

Test Configuration
Performance Results – 10 GigE, NFS, iSCSI and FCoE

ETS – 50% FCoE, 50% IP

Performance Results – 10 GigE, NFS, iSCSI and FCoE

ETS – 40% FCoE, 60% IP

Performance Results – 16 GFC – IOPS Test

Configuration

MySQL #1
MySQL #2
MySQL #3
MySQL #4

2Px6 server running RHEL
DB1 DB2 DB3 DB4
Volume Management
Dual-port 16 Gbps FC HBA
16 Gbps FC switches

All flash storage A
DB1A DB1A DB1A
DB1A

All flash Storage B
DB2A DB2A DB2A
DB2A

Rack Layout:

Redundant 10 GbE Switches
20 physical servers (2Px6 cores)
Four MySQL nodes/Server
Each server mounts 2 TB mirrored from all-flash Appliances

Redundant 16 Gbps FC switches
10-80 TB all flash storage appliances
Performance Results – 16 GFC – IOPS Test

Test Results – 1 Million IOPS

- MySQL Clusters
  - 10 servers, 4 instances of MySQL on each server
  - Databases – 256 GB each instance
  - 8 mirrored storage volumes, all flash

Performance Results – 16 GFC – Bandwidth Test

Configuration

Rack Layout:

- Server 1
- Server 2
- Server 3

16Gb Fibre Channel x5

16Gb FC Switch

8Gb Fibre Channel x10

10 TB All-flash Array

Specs:
16Gb dual-port FC HBA (in each server)

Specs:
16Gb FC Switch
Performance Results – 16 GFC – Bandwidth Test

Test Results – 7200 MB/sec sustained

- Oracle RAC single large database cluster
  - “Select count(*)” all 10.2 billion rows completed in < 8 minutes

Demartek Free Resources

- Demartek FC Zone
  - www.demartek.com/FC
- Demartek FCoE Zone
  - www.demartek.com/FCoE
- Demartek iSCSI Zone
  - www.demartek.com/iSCSI
- Demartek SSD Zone
  - www.demartek.com/SSD
- Demartek SSD Deployment Guide
Demartek Storage Interface Comparison

Contents
- Acronyms
- Storage Networking Interface Comparison Table
- Transfer Rate, Bits vs. Bytes, and Encoding Schemes
- History
- Roadmaps
- Cables: Fiber Optics and Copper
- Connector Types
- PCI Express (PCIe)

- www.demartek.com/Demartek_Interface_Comparison.html

- Or search for “storage interface comparison”
Demartek YouTube Channel
Free Monthly Newsletter

Demartek publishes a free monthly newsletter, *Demartek Lab Notes*, highlighting recent reports, articles and commentary.

Look for the newsletter sign-up at: [www.demartek.com/Newsletter](http://www.demartek.com/Newsletter)
Thank You!

Contact Information

(303) 940-7575
www.demartek.com
http://twitter.com/Demartek
YouTube: www.youtube.com/Demartek
Skype: Demartek

Dennis Martin, President
dennis@demartek.com
www.linkedin.com/in/dennismartin

#SDSF2012