FICON Channel Extension Technology and Customer Use

Mike Blair
Cisco Systems
mblair@cisco.com

Wednesday, February 6, 2013
Session 13062
Abstract

- This session will discuss FICON channel extension as a part of disaster recovery plan at many companies. When backup datacenters are across a campus from each other, there is nothing to worry about but with today's regulations and industry best practices, most companies have their datacenter tens, hundreds, even thousands of kilometers apart. This session will discuss in detail the technologies needed to assist the base FICON support of the mainframe for acceptable performance and usability.
Agenda

• Optical Solutions
• Issues for Extended Distance
• How does FCIP help this
• Protocol Acceleration
  • Tape
  • XRC
### SAN Extension Technology Options

#### Increasing Distance

<table>
<thead>
<tr>
<th>Data Center</th>
<th>Campus</th>
<th>Metro</th>
<th>Regional</th>
<th>National</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dark Fiber</td>
<td>Sync (Any Speed)</td>
<td>Limited by Optics (Power Budget)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CWDM</td>
<td>Sync (1,2,4 Gbps)</td>
<td>Limited by Optics (Power Budget)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DWDM</td>
<td>Sync (1,2,4,10 Gbps per λ)</td>
<td>Limited by BB_Credits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SONET/SDH</td>
<td>Sync (1,2 Gbps + Subrate)</td>
<td>Async</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDS9000 FCIP</td>
<td>Sync (Metro Eth)</td>
<td>Async (WAN,1 Gbps)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Coarse Wavelength Division Multiplexing (CWDM)

- 8-channel WDM at 20nm spacing (cf DWDM at <1nm spacing)
  - 1470, 1490, 1510, 1530, 1550, 1570, 1590, 1610nm
- “Colored” CWDM SFPs or XFPs used in FC/FICON directors
- Optical multiplexing done in CWDM OADM (optical add/drop multiplexer)
  - Passive (unpowered) device—just mirrors and prisms
- Power budget depends on speed
  - 2G support approx. 100km
  - 4G support approx. 40km
DWDM Schematic

- Routers
- ATM
- SAN
- Low Cost MM/SM 850nm/1310nm
- ITU-T grid 15xxnm Transponders
- DWDM Mux (Filter)
- Optical Amplifier
- Wavelength Multiplexed Signals
- Optically Amplified Wavelengths

OEO
Buffer Credit Concepts

- Flow Control Mechanism used in FC/FICON Networks
  - Based on idea of no packet loss / drops
- Each port has a configurable (or default) number of BCs
  - May be different based on port type / speed
- Cisco switches come with between 16 and 250 BCs / port
  - Extended BCs are available if licensed (up to 4K per port*)
- BC parameters are exchanged between peers at FLOGI
- Buffer credits are a hop-by-hop flow control – NOT End-to-End
- Each buffer credit is 2K
Buffer Credits in a Network

“I can receive 107 Buffers then you must stop until I acknowledge some or all of them”

“That’s fine – I can only receive 16 frames myself. This is ok since we are a few feet apart.”
Relationship of Speed to Buffer Credits

1 Gbps FC ~ Need 1 Buffer Credit per 2 Km

2 Gbps FC ~ Need 1 Buffer Credit per 1 Km

4 Gbps FC ~ Need 2 Buffer Credits per 1 Km

8 Gbps FC ~ Need 4 Buffer Credits per 1 Km

Theoretically - For 8 Km 1G needs 4, 2G needs 8, 4G needs 16, 8G needs 32
### Number of Buffer Credits - Reality

<table>
<thead>
<tr>
<th>Category</th>
<th>Value(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>76</td>
</tr>
<tr>
<td>Write</td>
<td>2084</td>
</tr>
<tr>
<td>Status Accept</td>
<td>68</td>
</tr>
</tbody>
</table>

- Simple 4K write
- Will not fit into 2 buffers because of headers for FC as well as SB3

Average = \( \frac{76+2084+2084+72+68}{5} = 877 \text{ Bytes} \)
FICON L4 Flow Control / IU Pacing

Mainframe    MDS1    MDS2    Device

Write 1

2
3
16

Write 1

16

Write 1

Complete your sessions evaluation online at SHARE.org/SFEval

2013
FICON L4 Flow Control / IU Pacing

Mainframe  MDS1  MDS2  Device

Write 1

Write 1

Write 1

Command Resp 1

Command Resp 8

Complete your sessions evaluation online at SHARE.org/SFEval
FICON L4 Flow Control / IU Pacing

Mainframe  MDS1  MDS2  Device

Write 1  2  16  Write 1

3

16

Command Resp 1

17

Command Resp 8

24

Command Resp 16

Write 17

24

Command Resp 1

Command Resp 8

Command Resp 16

Write 1
FICON L4 Flow Control / IU Pacing

Mainframe  MDS1  MDS2  Device

- Write 1
- Command Resp 1
- Command Resp 8
- Command Resp 16
- Command Resp 24
- Status

Device Idle
FCIP: Fibre Channel over IP

FCIP Is a Standard from the IETF IP Storage WG for Linking FibreChannel SANs over IP (RFCs 3821 and 3643)

- Point-to-point tunnel between FCIP link end-points
- Appears as one logical FC fabric with single FSPF routing domain
- FICON is just another upper layer protocol that can be transported over IP
- FICON over FCIP can provide cost-effective channel extension
FCIP Frame Detail

- Max FibreChannel frame is 2148 bytes plus optional extras
- FC frames are segmented and reassembled if MTU too small (TCP payload on second or subsequent packets)
- Jumbo frames may increase performance
  - IP MTU of 2300 avoids splitting of TCP frames

**Diagram:**
- Ethernet Header: 14 bytes
- IP Header: 20 bytes
- TCP Header: 20 bytes
- TCP Options: 12 bytes
- FCIP Header: 28 bytes
- EISL Header: 4 bytes
- Optional Header: 0-16 bytes
- FC Frame: Max 2148 (E_Port) + EISL and Opt Headers
- Ethernet CRC32: 4 bytes

**FCIP Overhead for Ethernet Frames:**
- 94 Byte Header + 4 Byte CRC = 98 Bytes

**EISL and Optional Headers:**
- If TE_Port, then 8 Bytes Added to FC Frame (after SOF) for VSAN Routing
Why Use FCIP?

• Network availability:
  • Lambdas or dark fiber not available or too expensive
  • IP network capacity already in place or only alternative

• Distance:
  • FCIP not limited by BB_Credits
  • Extension only limited by TCP max window (32MB for MDS9000 → 20,000km at 1Gbps)

• Application requirements:
  • Need Acceleration technologies built into FCIP
Storage Traffic and TCP

- Storage traffic:
  - Quite bursty
  - Latency sensitive (sync apps)
  - Requires high, instantaneous throughput
- Traditional TCP:
  - Tries to be network sociable
  - Tries to avoid congestion (overrunning downstream routers)
  - Backs off when congestion detected
  - Slow to ramp up over long links (slow start and congestion avoidance)
MDS FCIP TCP Behavior

• Reduce probability of drops
  • Bursts controlled through per flow shaping and congestion window control → less likely to overrun routers
• Increased resilience to drops
  • Uses SACK, fast retransmit and shaping
• Aggressive slow start q
  • Initial rate controlled by “min-available-bandwidth”
  • Max rate controlled by “max-bandwidth”

Differences with Normal TCP:
  ▪ When congestion occurs with other conventional TCP traffic, FCIP is more aggressive during recovery ("bullying" the other traffic)

  Aggression is proportional to the min-available-bandwidth configuration
Round Trip Time

Configuring the round-trip-time parameter:
• Not necessarily symmetric
• Use ping and measure-rtt commands to calculate
• Automatically calculated in Cisco MDS

12-ms
12-ms

End-to-end latency * 2 = 24 ms RTT.
TCP Maximum Window Size

Set the TCP MWS to **keep the pipe full.**

\[
\text{MWS} = \text{Maximum bandwidth} \times \text{RTT}
\]

Example: 5-ms latency = 10-ms RTT \times 155-\text{Mbps (OC-3)} = 192-\text{KB}

- Under dimensioning will throttle throughput
- Over dimensioning can cause congestion
- Use the bandwidth of the lowest speed link
TCP Maximum Bandwidth

Configure the TCP max-bandwidth value as follows:

- No larger than smallest pipe in the path
- If sharing the pipe, configure to be highest amount available to FCIP
- In a shared environment, configure QoS in the entire path
TCP Minimum Available Bandwidth

Configure TCP min-available-bandwidth value as follows:

- If dedicated path, min-available-bandwidth = max-bandwidth
- If shared path, use least amount that is always available to FCIP
- Lower if you see frequent retransmissions in a shared transport
- Must be at least 1/20 of max-bandwidth

<table>
<thead>
<tr>
<th>max_bw</th>
<th>min_bw</th>
</tr>
</thead>
<tbody>
<tr>
<td>45-Mbps (dedicated)</td>
<td>45-Mbps (shared)</td>
</tr>
</tbody>
</table>
Results of Packet Shaping

Traditional TCP Congestion Avoidance

Packet Shaping

- Shaper engaged during first RTT = min-available-bw
- Slow Start Threshold initialized to 95% MWS after one RTT
- Minimum threshold = min-available-bw
- Retransmission
- Congestion Avoidance (+2 cwnd per RTT)
Using GE Sub-Interfaces, Multiple FCIP Tunnels and Port Channeling to Enable High b/w FCIP

- Use separate VSANs for data replication (100) and tape access (200)
- Port channel FCIP tunnels for replication traffic for load balancing
- Tape access Links are not port channeled

Note – When multiple FCIP tunnels are on the same interface, they use a different TCP/IP port numbers
Now, Configure QOS based on business priorities of data

- VSAN 100 – high priority – disk mirroring
- VSAN 200 – med priority – Tape backups
- VSAN 300 (not shown) – low priority (open systems SAN stuff)

Making the assumption that this is a dedicated SAN WAN infrastructure – but within that, prioritization is needed.

Note: Routers and Switches **MUST** be QOS aware.
FCIP Data Compression

- Cisco uses RFC standard compression algorithms implemented in both hardware and software
- MDS 9000 18/4 MSM and SSN-16
  - Third Generation IP Services Modules
  - Hardware and software-based compression, hardware-based encryption, and intelligent fabric-based application services
- Three compression algorithms—modes 1–3 plus auto mode
- Compressibility is data stream dependent
  - All nulls or ones → high compression (>30:1)
  - Random data (e.g., encrypted) → low compression (~1:1)
- “Typical” rate is around 4:1, but may vary considerably
- Application throughput is the most important factor
IPSec Encryption for FCIP

FCIP Link Encryption Provides:

- Data confidentiality—sender can encrypt packets before transmitting them across a network
- Data integrity—receiver can authenticate packets sent by the IPSec sender to ensure that the data has not been altered during transmission
- Data origin authentication—receiver can authenticate the source of the IPSec packets sent; this service is dependent upon the data integrity service
- Anti-replay protection—receiver can detect and reject replayed packets
Hardware-Based IPSec Encryption

- Hardware-based GigE wire-rate performance with latency ~ 10 µs per packet
- Standards-based IPSec encryption—implements RFC 2402 to 2410, and 2412
  - IKE for protocol/algorithm negotiation and key generation
  - Encryption: AES (128 or 256 bit key), DES (56 bit), 3DES (168 bit)
Sun VSM to RTD Extension

VSM at local site writing both a local tape copy as well as a tape copy at the archived site.
FICON Tape Write Acceleration

- Accelerates Writes by means of local acknowledgement
  - Command Response
  - Status
- Data is never fully owned by the FTA
  - Sync command is not emulated – insures data integrity
- Flow control to limit amount of data read to WAN rate
- Tape control, label processing, etc are not accelerated
Backup protocol without acceleration

Mainframe → Cisco MDS ← FCIP → Cisco MDS → VSM/Tape Library

Write Chain

ScratchVol mount, Write VolHdrs etc.
Backup protocol without acceleration ...

Mainframe → Cisco MDS (FCIP) → Cisco MDS → VSM / Tape Library

- ScratchVol mount, Write VolHdrs etc.
- Write Chain
- Status

TAPE IDLE
Backup protocol without acceleration ...

Mainframe → Cisco MDS [FCIP] → Cisco MDS → VSM / Tape Library

- ScratchVol mount, Write VolHdrs etc.
- Write Chain
- SYNC
- TAPE IDLE
- TAPE IDLE
- All data on Media
- Rewind Unload

Status
Backup protocol with acceleration

Mainframe → Host-side FCIP → Tape-Side FCIP → VSM / Tape Library

- ScratchVol mount, Write VolHdrs etc.
- Write Chain 1
- Write Chain 1
- New OXID
Backup protocol with acceleration ...

Mainframe → Host-side FCIP → Tape-Side FCIP → VSM / Tape Library

ScratchVol mount, Write VolHdrs etc.

Write Chain 1
New OXID
Write Chain 2
Write Chain 1
Backup protocol with acceleration ...

Mainframe

Host-side FCIP

ScratchVol mount, Write VolHdrs etc.

Write Chain 1

SYNC

Rewind Unload

Tape-Side FCIP

Write Chain 1

Write Chain 2

New OXID

VSM / Tape Library

All data on Media

Complete your sessions evaluation online at SHARE.org/SFEval
FICON Tape Read Acceleration

- Accelerates Reads by
  - Learning Read Data Pattern
  - Entering Acceleration mode
  - Flowing off the host / Pre-reading data on CU
  - Stage data at host side – continue reading at the tape side
  - Start up the host reading the staged data
- If too much data is pre-read, FTA will reposition the tape
- Tape control, label processing, etc are not accelerated
FICON Tape Read Acceleration
Sample Performance for FTA Write

Empirix Distance Simulator
Results: Migration Time (write)

FTA Enabled (Seconds)
- 132
- 134
- 136
- 140
- 146
- 149
- 153
- 159
- 165

FTA Disabled (seconds)
- 131
- 391
- 708
- 1019
- 1504
- 1637
- 1955
- 2256
- 2583
- 2876
- 3186

1000 Km – 10ms
5000 Km – 50ms

Complete your sessions evaluation online at SHARE.org/SFEval
Results: Throughput (write)

- **Throughput – MB/s**
  - **FTA Enabled (MB/sec)**
  - **FTA Disabled (MB/sec)**

- **Site-Site Delay**
  - 1000 Km – 10ms
  - 5000 Km – 50ms

Throughput and site-site delay relationship graph.
FTA Configuration Information

• Only one active FTA-enabled FCIP link is allowed between two domains. Port Channels of FTA-enabled links are not supported.
  • This is due to the fact that state is kept on a per-port basis.
• Multiple/all ports on a IPS card can run FTA simultaneously
  • Each of the links must be trunking a different VSAN.
• 8 FICON VSANs are allowed per MDS Chassis – each of these with its own CUP for management.
• The number of write chains buffered is automatically adjusted based on the tape speed and the RTT of the FCIP connection
FTA – More Details

• There is support for both 3590 and 3490 real FICON tape drives. There is support for 3490 Virtual Tapes
  • IBM and STK have both only implemented 3490 Virtual Tapes in their VTS and VSM platforms respectively.
  • 3490 versus 3590 selection is dynamic and no configuration is needed for device selection. There can be 3490 and 3590 on the same FCIP link at the same time.
• Multipath is supported from the host to the tape.
  • These multiple paths must still transverse the same FCIP link but this gives higher host-side redundancy.
VSM - RTD Customer Example - EMEA

600 Miles short path
What is XRC?

- XRC = eXtended Remote Copy
  - Now officially “z/OS Global Mirror”
- Mainframe-based replication SW
- XRC clients include:
  - Over 300 installations worldwide (source: IBM)
  - Major Banks in Germany, Scotland, Italy, Turkey, Greece
  - Major US Banks / Brokerages / Insurance Co’s
  - Major Banks in Taiwan, Japan, China, Thailand, Korea
- Remote “System Data Mover” (z)
  - Reads data from remote primary DASD
  - Writes it to local secondary DASD
What is good and bad about XRC?

- **Disk Vendor Independent**
  - No lock into vendor unique implementations
  - Can copy from one vendor disk to another
  - Can be used for migrations from one vendor to another
- **Management and control from the mainframe**
  - No reliance on disk-to-disk replication changes
  - Performance management from Z
XRC Acceleration – How It Works

- Acceleration of RRS channel programs
- Read Record Sets (ie. Updates)
- In DSO (first command in RRS set) we know how to pre-read for the whole chain of data

- We pre-read the data and send across .. thereby filling the pipe.
- This works around the IU limitations in the FICON architecture.
How Fast is it?

• Some performance testing results:
  • Vs. no Acceleration:
    • Almost 5x faster at 1600 km
    • Almost 9x faster at 3200 km
XRC Acceleration – Other Facts

- Works with Cisco port channels
  - Allows for less disruption when loss of WAN occurs
- Works with all models of Z system
  - Backwards compatible with all older Z systems
  - Fully compatible with the new z10 Extended distance FICON
- Can utilize all compression/encryption on FCIP hardware
- Supports all 3 major vendor’s disk arrays
- Supports multi-reader, PAVs, and Hyper PAVs
- This is a separately licensed feature through IBM
XRC Customer Example – North America