Storage Area Network Extension Design and Operation
BRKSAN-2704

Mark Allen
Manager Storage Technical Marketing
Agenda

- Business Drivers for SAN Extension
- Conflicting Challenges Faced Deploying SAN Extension
- Design Principles for Successful Deployments
- SAN Extension Network Solutions
- Using Enhanced Features to Improve SAN Extension
- Operation / Troubleshooting Reference
Business Drivers for SAN Extension
Why Do We Need SAN Extension?

Goal: Protecting Data

Disaster Recovery
Protecting Data Through Offsite Data Replication and Backup

Business Continuance
Restoration of Business After a Failure

Business Resilience
Continued Operation of Business During a Failure

Goal: Protecting Data
Regulations Driving Data Protection

- Currently there are more than 4000 U.S. state and federal laws and regulations dealing with electronic records
- Data retention, retrieval and protection guidelines are being specified

<table>
<thead>
<tr>
<th>Industry</th>
<th>Regulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial Services</td>
<td>Gramm-Leach-Bliley Act (GLBA)</td>
</tr>
<tr>
<td></td>
<td>Sarbanes-Oxley Act (SARBOX)</td>
</tr>
<tr>
<td></td>
<td>PCI Data Security Standard</td>
</tr>
<tr>
<td></td>
<td>Basel II Accord (EU)</td>
</tr>
<tr>
<td>Healthcare and Pharmaceuticals</td>
<td>HIPAA (Health Insurance Portability and Accountability Act of 1996)</td>
</tr>
<tr>
<td></td>
<td>FDA Title 21 Code of Federal Regulation (CFR) Part 11</td>
</tr>
<tr>
<td>Infrastructure and Energy</td>
<td>Guidelines for FERC and NERC Cybersecurity Standards</td>
</tr>
<tr>
<td></td>
<td>Chemical Sector Cyber Security Program</td>
</tr>
<tr>
<td></td>
<td>Customs-Trade Partnership Against Terrorism (C-TPAT)</td>
</tr>
</tbody>
</table>
Meeting Business Objectives: 
Recovery Point and Recovery Time

- Last Backup or Point Where Data in Usable State
- Disaster Strikes
- Systems Recovered and Operational

How Far Back? 
How Long to Recover?

Smaller RPO/RTO
- Higher $$$
- Replication
- Hot standby systems

Larger RPO/RTO
- Lower $$$
- Tape backup/restore
- Cold standby system
How Far is Enough?

- Disasters are characterized by their impact
  - Local, metropolitan, regional, global
  - Fire, flood, earthquake, attack

- Fundamental Question: Is the recovery / backup site within the threat radius you’re protecting against?

<table>
<thead>
<tr>
<th>Impact</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local</td>
<td>1–2 km</td>
</tr>
<tr>
<td>Metro</td>
<td>&lt; 50 km</td>
</tr>
<tr>
<td>Regional</td>
<td>&lt; 400 km</td>
</tr>
<tr>
<td>Global</td>
<td></td>
</tr>
</tbody>
</table>
Is Your Data Safe?

Data Protection Also Includes Security

- **Who has access to your data and your data network?**
  - Network Access Control (NAC)
  - Role Based Access Control (RBAC)

- **How is your data being protected on the network?**
  - Link Encryption on the network

- **How do you protect your data media?**
  - Media encryption

Other Challenges

- What type of Network and Bandwidth is available
  Native Fibre Channel or IP based

- What impact will latency have on your Applications? Replication? Backups?
  Performance impact due to latency can limit design options

- How much High Availability is built into the design
  Do you design around 1 point of failure? 2 points? Can you over design HA?
Adding It All Up

- Data is the most valuable commodity in the data center
- Businesses require comprehensive business plans for Business Continuity (BC) and Disaster Recovery (DR)
- SAN Extension is the foundation to a successful BC/DR plan
SAN Extension Design Principles
Typical SAN Design

- Servers with two Fibre Channel connections to storage arrays for high availability
  - Use of multipath software is required in dual fabric host design
- SAN extension fabrics typically separated from host access fabrics
  - Replication fabric requirements vary by replication method deployed
Basic HA SAN Extension Network

High-Availability Replication Design:

- Conventional approach is dual fabrics (e.g., yellow VSAN and blue VSAN) over distance
- “Client protection”—arrays provide protection against failures in either fabric
- May be augmented with additional “network protection” via portchannels and/or optical protection schemes
SAN Extension Design: Adding Link HA

Portchannels Increase Resilience for High-Availability with FC or FCIP Links

- Appears as a single logical link (up to sixteen member links)
  - Protecting the fabric from network failure
- Route portchannel member links over diverse geographic paths
- Load balancing on SRCID/DESTID or SRCID/DESTID/OXID basis (unidirectional per VSAN)
  - SCSI exchange is smallest atomic unit, so frame order kept intact

Site 1
Access VSANs
Replication VSANs

Site 2
Replication VSANs
“A” VSANs
“B” VSANs

Portchannels
Hybrid SAN Extension Implementations

Hub and Spoke

Central Site

IP Network

Secondary DC

Backup Site

Primary DC

Synchronous Replication

Asynchronous Replication/Backup

Multi-hop

Hub and Spoke

Remote Sites

Central Site
SAN Extension Network Solutions - Native Fibre Channel
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</td>
<td>Limited by Optics (Power Budget)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CWDM</td>
<td>Sync (1,2,4Gbps)</td>
<td>Limited by Optics (Power Budget)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DWDM</td>
<td>Sync (1,2,4,10Gbps per λ)</td>
<td>Limited by BB_Credits</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SONET/SDH</td>
<td>Sync (1,2Gbps + subrate)</td>
<td>Async</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IP</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDS9000 FCIP</td>
<td>Sync (Metro Eth)</td>
<td>Async (WAN,1Gbps)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dark Fiber

- Single 1/2/4/8/10 Gbps FC link per fiber pair
  - SW (850nm) over 62.5/125µm multimode
  - SW (850nm) over 50/125µm multimode (OM2, OM3 or OM4)
  - LW (1310nm) over 9/125µm single mode

- Client protection only; Upper Layer Protocol (ULP), either SAN or application, responsible for failover protection

Distance Based on Fiber Type, Optic Type, Link Speed

Portchannel 2-16
1/2/4/8/10 Gbps FC Over Two Diverse Paths

Diverse Paths—Multiple Fiber Pairs Each Path
Coarse Wavelength Division Multiplexing (CWDM)

- 8-channel xWDM technology using 20nm spacing (cf DWDM at <1nm spacing)
  - 1470, 1490, 1510, 1530, 1550, 1570, 1590, 1610nm
- “Colored” CWDM SFPs (or GBICs) used in FC switches (no transponder required)
- Optical multiplexing done in CWDM OADM (optical add/drop multiplexer)
  - Passive (unpowered) device; just mirrors and prisms
- Up to 30dB power budget (36dB typical) on Single Mode Fiber (SMF)
  - 1 / 2 Gb - ~100km point-to-point or ~40km ring
  - 4 Gb - ~40km point-to-point or ~20km ring
- 1/2/4 Gigabit Fibre Channel and 1 Gigabit Ethernet support
2-Site CWDM Storage Network

- HA resilience against fiber cut—“client” protection
  - 4-member portchannel - 2 x 2 diverse paths
  - Portchannel appears as single logical link
  - E_Port or TE_Port for carriage of VSANs
  - Load balance by src/dst (or src/dst/oxid)
  - Fiber cut will halve capacity from 32Gbps to 16Gbps but not alter fabric topology - no FSPF change

- MUX-8 would double capacity or leave spare wavelengths for GigE channels
CWDM Optics Without Multiplexor

- CWDM Optics do not require MUX
  - If dark fiber available, can be used like typical SFPs
- Can use different wavelengths or the same wavelengths on all interfaces
- Use of optical attenuators may be required for shorter distance fiber runs
  - Optical power meter used to measure signal strength

4 fibre paths between each switch

Portchannel 4 x 4 Gbps

- 4 Gbps CWDM SFPs - different wavelengths
- 4 Gbps CWDM SFPs - same wavelength
Dense Wavelength Division Multiplexing (DWDM)

- Higher density xWDM technology compared with CWDM
  
  32 lambdas or channels in the narrow band around 1550nm at 100GHz spacing (0.8nm)

- Erbium-Doped Fiber Amplifiers (EDFA) allow for longer distances than CWDM

- Carriage of 1, 2, 4 or 10 Gbps FC, FICON, GigE, 10GigE, ESCON, IBM GDPS

- Data center to data center

- Multiple Protection options: client, splitter, or line card
DWDM Protection Alternatives for Storage

Optical Splitter Protection

- Single transponder required
- Protects against fiber breaks
- Failover causes loss of light (and fabric change if only link)

Linecard or Y-Cable Protection

- Dual transponders required
  - More expensive than splitter-based protection
- Transmits over both circuits, but only one accepted
Traditional DWDM

- MDS
- Transponder
- ROADM

Support for 2G & 4G FC
No optical transponder required
DWDM sfp in MDS connects directly to OADM module of ONS 100 Ghz products
32 wavelengths mapping to ITU channels 21-24, 26-29, 31-34, 36-39, 41-44, 46-49, 51-54 and 56-59

FC and DWDM - MDS w/ Integrated DWDM Optics

- MDS
- ROADM
- DWDM SFP

32 wavelengths mapping to ITU channels 21-24, 26-29, 31-34, 36-39, 41-44, 46-49, 51-54 and 56-59
DWDM HA Storage Network Topology (1)

- **Client protection recommended**
  Fabric and application responsible for failover recovery

- **Portchannel provides resilience**
  Portchannel members follow diverse paths
  Single fiber cut will not affect fabric (no RSCNs, etc.)
  Loss of path reduces bandwidth of both “A” and “B” fabrics by 50%
  Use “Src/Dst” hash for load balancing (rather than “Src/Dst/Oxid” per exchange) for each extended VSAN
DWDM HA Storage Network Topology (2)

- **Client protection recommended**
  Fabric and application responsible for failover recovery

- **Portchannel provides resilience**
  Portchannel members follow diverse paths
  Single fiber cut will not affect fabric (no RSCNs, etc.)
  Loss of path reduces bandwidth of only one fabric by 50%
  Use “Src/Dst” hash for load balancing (rather than “Src/Dst/Oxid” per exchange) for each extended VSAN
FibreChannel over SONET/SDH

- FC over SONET/SDH (FCoS) follows same distance rules as other optical technologies
  - BB_Credits in Fibre Channel switch limits distance
- Outage in SONET/SDH network will not cause loss of light
  - Recovers in <50ms
  - May cause some loss BB_Credit loss from in flight traffic
  - MDS9000 has built in lost BB_Credit recovery mechanism
Optics Information

- For complete information on MDS optics, refer to the Cisco MDS 9000 Family Pluggable Transceivers Data Sheet at:

SAN Extension Network Solutions – Fibre Channel over IP (FCIP)
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
- Creates one logical FC fabric with single FSPF routing domain
### FCIP Frame Detail

- **Max FibreChannel frame is 2148 bytes plus optional extras**
- **FCIP will segment and reassemble FC frames if MTU too small (TCP payload only on second or subsequent packets)**
- **Jumbo frames may increase performance**
  - IP MTU of 2300 avoids splitting of TCP frames

#### 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

---

<table>
<thead>
<tr>
<th>Ethernet Header</th>
<th>IP Header</th>
<th>TCP Header</th>
<th>TCP Options</th>
<th>FCIP Header</th>
<th>EISL Header</th>
<th>Opt Header</th>
<th>FC Frame</th>
<th>Ethernet CRC32</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>20</td>
<td>20</td>
<td>12</td>
<td>28</td>
<td>4</td>
<td>8</td>
<td>0-16</td>
<td>4</td>
</tr>
</tbody>
</table>

Max 2148 (E_Port) + EISL and Opt Headers
Resolving Differences Between Storage and TCP/IP Networks

- **Storage traffic:**
  - Quite bursty
  - Latency sensitive (esp. sync replication and tape backup)
  - 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

- Designed to reduce probability of drops
  - Bursts are 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 mechanism
  - 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
Frame Buffering: FCIP and FC

- FCIP presents a lower bandwidth pipe (if WAN link)
  Drain rate (send rate) depends upon bandwidth and congestion
- Slow ramp up of traditional TCP can cause FC frame expiry in some conditions
  Mixture of slow link (e.g., <DS3/E3; retransmissions, many sources, big buffers)
FCIP TCP Packet Shaping: MDS9000

- Shaper sends at a rate consumable by the downstream path
  - Immediately sends at “minimum-bandwidth” rate (avoids early stages of traditional slow start)
  - Ramps up to “maximum-bandwidth” rate (using usual slow start and congestion avoidance methods)

- Requirements for shaper to engage:
  - Min-available-bandwidth > 1/20 max-bandwidth
  - SACK (Selective Ack) must be enabled
**MDS9000 FCIP TCP Behavior**

- **For example: a dedicated link**
  
  Entire link is always available, so…
  
  “min bandwidth” = “max bandwidth”
  
  FCIP will always send at 95% to 100% of max rate without ramp up
  
  Traffic is shaped at sending rate (max-bw)
  
  After retransmission (congestion), sender resumes at min (=max rate)

- **Behavior mimics UDP “blast” but with benefits of retransmission capability and shaping**
### Fibre Channel over IP (FCIP) Design

- **Same port channeling and VSAN trunking rules apply as with FC links**
- **Portchannel individual FCIP links to separate Ethernet switches/routers**
  - Each WAN link carries two FCIP tunnels
  - Load balancing on SRCID/DESTID or SRCID/DESTID/OXID basis (unidirectionally per VSAN)
  - Certain replication protocols require SRCID/DESTID load balancing
    - FICON, IBM PPRC, HP CA-EVA
Configuring FCIP

Define Local GE interface

Associate Local GE address with profile

Define maximum and minimum bandwidth and network latency

Associate profile to FCIP interface

Identify remote FCIP peer

feature fcip
interface GigabitEthernet 1/1
ip address 10.10.100.1 255.255.255.0
fcip profile 11
ip address 10.10.100.1
tcp max-bandwidth-mbps 1000 min-available-bw-mbps 950 round-trip-time-ms 10
interface fcip 11
switchport trunk mode off
use profile 11
peer-info ipaddr 10.10.100.2

feature fcip
interface GigabitEthernet 1/1
ip address 10.10.100.2 255.255.255.0
fcip profile 11
ip address 10.10.100.2
tcp max-bandwidth-mbps 1000 min-available-bw-mbps 950 round-trip-time-ms 10
interface fcip 11
switchport trunk mode off
use profile 11
peer-info ipaddr 10.10.100.1
SAN Extension Advanced Features
Extending Optical SAN Extension

FibreChannel Frame Buffering

- Buffer to buffer credits (BB_Credit) are negotiated between each device in a FC fabric; no concept of end to end buffering
- One buffer used per FC frame, irregardless of frame size; small FC frame uses same buffer as large FC frame
- FC frames buffered and queued in intermediate switches
- Hop-by-hop traffic flow paced by return of Receiver Ready (R_RDY) frames; can only transmit up to the number of BB_Credits before traffic is throttled
Extending Optical SAN Extension

**BB_Credits and Distance**

- **BB_Credits** are used to ensure enough FC frames in flight.
- A full (2112 byte) FC frame is approx 2 km long @ 1 Gbps, 1 km long @ 2 Gbps, ½ km long at 4 Gbps and ¼ km long @ 8 Gbps.
- As distance increases, the number of available BB_Credits need to increase as well.
- Insufficient BB_Credits will throttle performance - no data will be transmitted until R_RDY is returned.
Configuring Extended BB Credits

Configuring MDS to support 1000km at 4G over SONET/SDH Network

mds9513# configure terminal
Enter configuration commands, one per line. End with CNTL/Z.
mds9513(config)# feature fcrxbbcredit extended
mds9513(config)# interface fc 5/1
mds9513(config-if)# switchport fcrxbbcredit extended 2000
mds9513(config-if)# end
mds9513# show interface fc 5/1
fc1/18 is up
Hardware is Fibre Channel, SFP is short wave laser w/o OFC (SN)
Port WWN is 20:12:00:0d:ec:6a:2c:00
Admin port mode is auto, trunk mode is off
snmp link state traps are enabled
Port mode is E, FCID is 0x0a0000
Port vsan is 200
Speed is 4 Gbps
Rate mode is dedicated
Transmit B2B Credit is 128
Receive B2B Credit is 2000
Receive data field Size is 2112

Enable extended BB feature
Set BB credits (2000 required)
Verify BB credits set
# Extending Optical SAN Extension

## SAN Network Solutions for Increasing Distance

<table>
<thead>
<tr>
<th>Shared</th>
<th>Dedicated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FX-Port (Fixed)</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td><strong>32 Port 1/2/4/8/10 Gbps Advanced Switching Module (DS-X9232-256K)</strong></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>48 Port 1/2/4/8/10 Gbps Advanced Switching Module (DS-X9248-256K)</strong></td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) Assuming max frame size
# Extending Optical SAN Extension

SAN Network Solutions for Increasing Distance

<table>
<thead>
<tr>
<th>Shared</th>
<th>Dedicated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FX-Port</td>
</tr>
<tr>
<td></td>
<td>(Fixed)</td>
</tr>
<tr>
<td>24 Port Performance 1/2/4/8-Gbps (DS-X9224-96)</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>48 Port Performance 1/2/4/8-Gbps (DS-X9248-96)</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>48 Port Host 1/2/4/8-Gbps (DS-X9248-48)</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(1) Assuming max frame size
(2) Per port groups 1-4 and 5-8
## Extending Optical SAN Extension

### SAN Network Solutions for Increasing Distance

<table>
<thead>
<tr>
<th>Shared</th>
<th>Dedicated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FX-Port (Fixed)</td>
</tr>
<tr>
<td>16 Port 1/2-Gbps (DS-X9016)</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>32 Port 1/2-Gbps (DS-X9032)</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>14/2 Port 1/2-Gbps (DS-X9302-14K9)</td>
<td>N/A</td>
</tr>
<tr>
<td>9216i Fabric Switch</td>
<td></td>
</tr>
<tr>
<td>12 Port 1/2/4-Gbps (DS-X9112)</td>
<td>N/A</td>
</tr>
<tr>
<td>24 Port 1/2/4-Gbps (DS-X9124)</td>
<td>16</td>
</tr>
<tr>
<td>48 Port 1/2/4-Gbps (DS-X9148)</td>
<td>16</td>
</tr>
<tr>
<td>4 Port 10-Gbps (DS-X9704)</td>
<td>N/A</td>
</tr>
</tbody>
</table>

(1) Assuming max frame size
Extending Optical SAN Extension

Optical Solutions for Increasing Distance

ONS 15454 SL-Series Card

- Negotiates up to 255 BB_Credit with FC switch
- Spoofs R_RDYs to FC switch (Release 5.0)
- Has 1200 BB_Credits between SL cards
- Extends distances to 2300 km @ 1G FC or 1150 km @ 2G FC

Configuration Note: When using ONS 15454 R_RDY spoofing with MDS 9000, disable BB State Change Notification on connected MDS port:

```
interface fc 1/1
no switchport fcbbscn
```
Improving Optical Recovery

Port Track for Resilient SAN Extension Solutions

- Arrays recover from a link failure via I/O timeouts; however, this can take several seconds or longer
- MDS PortTrack addresses this by monitoring the WAN/MAN link and if it detects a failure, it will bring down the corresponding link connected to the array
- The array after detecting a link failure will redirect the I/O to another link without waiting for the I/O to timeout
Improving Optical Recovery

Port Tracking and ONS FLC or Squelching

- The MDS port tracking feature can be used with the ONS 15530 Forward Laser Control (FLC) or ONS 15454 squelching feature to further track failures in the network, improving the ability to detect failed paths.
- Forward Laser Control, squelching and port-tracking offer end to end path failure detection.
Comparison of Data Recovery
Port Tracking vs. No Port Tracking

![Graph showing data recovery comparison](image)

- **No Port Tracking Fabric A**
- **No Port Tracking Fabric B**
- **Port Tracking Fabric A**
- **Port Tracking Fabric B**

Fail ISL on Fabric “A”
Configuring Port Tracking

Configuring MDS to disable array port if both ISLs fail

mds9513-labD# configure terminal
Enter configuration commands, one per line.  End with CNTL/Z.
mds9513-labD(config)# feature port-track
mds9513-labD(config)# interface fc 5/9
mds9513-labD(config-if)# port-track interface fc5/1
mds9513-labD(config-if)# port-track interface fc5/2
mds9513-labD(config-if)#

mds9513-labD# show run interface fc5/9
version 4.1(3a)

interface fc5/9
port-track interface fc5/1
port-track interface fc5/2
switchport rate-mode dedicated
switchport mode auto
switchport trunk mode off
no shutdown

Enable port-track feature
Linked Array port
First Tracked ISL port
Second Tracked ISL port
FCIP Data Compression

- Compression used to increase link data capacity of FCIP interfaces
  - MDS 9000 18/4-port Multiservice Module, SSN-16 Storage Services Node, MDS 9222i (Current Generation)
    - Hardware and software-based compression, hardware based encryption and intelligent fabric based application services
  - MDS 9000 14/2-port Multiprotocol Services Module, MDS 9216i (Generation 2)
    - Hardware and software-based compression and hardware based encryption for FCIP

- Three compression algorithms used
  - Mode 1 – LZS based
  - Modes 2 & 3 – Deflate based
  - Auto – Optimizes Compression based on bandwidth and data rate

- Data 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 (2:1 on gen 2)

- Application throughput is the most important factor
Configuring FCIP Compression

Configure FCIP compression

```bash
mds9513-labD# conf t
Enter configuration commands, one per line. End with CNTL/Z.
mds9513-labD(config)# interface fcip 1
mds9513-labD(config-if)# ip-compression auto
```

Enable FCIP auto mode compression – must be same for both sides of FCIP tunnel

```bash
mds9513-labD# show interface fcip 1
fcip1 is up
    Hardware is GigabitEthernet
    IP Compression is enabled and set for auto

mds9513-labD# show interface fcip 1 counters fcip100
TCP Connection Information

IP compression statistics
  22958 rxbytes
  22958 rxbytes compressed, 0 rxbytes non-compressed
  1.28 rx compression ratio
  34660 txbytes
  20765 txbytes compressed, 4160 txbytes non-compressed
  1.39 tx compression ratio
  20641 txbytes compressed using auto
```
Link Layer Security

- Data Confidentiality requirements are part of business today
- Businesses need to ensure that data is not compromised while being transmitted between Data Centers
- FC TrustSec (FC) and IPSec (FCIP) used to secure data over ISLs between switches
Hardware-Based FC TrustSec Encryption

- Hardware-based 8G FC wire rate performance on Generation-3 and Generation-4 8G FC line cards
- Extension to standards-based FC-SP protocol to provide encryption of data
  - DH-CHAP used for peer authentication
  - Encryption: AES 128 bit key
Configuring FC TrustSec

Basic steps for configuring FC TrustSec

1. Enable FCSP feature (requires Enterprise License)

MDS9513# conf t

Enter configuration commands, one per line. End with CNTL/Z.

MDS9513(config)# feature fcsp
Configuring FC TrustSec

Basic Steps for Configuring FC TrustSec

2. Configure DHCHAP for switch MDS9513 with switch-wwn 20:00:00:0d:ec:07:e8:c0

   MDS9513(config)# fcsp dhchap password 7 fewhg123
   MDS9513(config)# fcsp dhchap devicename 20:00:00:0d:ec:07:e8:c0 password cisco123

3. Create one or more Security Associations (SA)

   MDS9513(config)# fcsp esp sa 400
   MDS9513(config-sa)# salt 0xDEADBEEF
   MDS9513(config-sa)# key 0x112233445566778899
   MDS9513(config-sa)# exit
Basic Steps for Configuring FC TrustSec

4. Shut down interface
   switch(config)# interface fc2/15
   switch(config-if)# shut

5. Configure fcsp feature on interface
   switch(config-if)# fcsp on

6. Configure Encapsulating Security Protocol (ESP) settings on interface
   switch(config-if)# fcsp esp manual
   switch(config-if-esp)# ingress-sa 400
   switch(config-if-esp)# egress-sa 400
   switch(config-if-esp)# exit

7. Bring up interface
   switch(config-if)# no shut

Full FC TrustSec configuration example using Fabric Manager and Device Manager at:
Hardware-Based IPSec Encryption

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

1. Enable IKE and IPSec features (requires Enterprise License)

MDS9222i# config t

Enter configuration commands, one per line. End with CNTL/Z.

MDS9222i(config)# feature crypto ike

MDS9222i(config)# feature crypto ipsec
Configuring IPSec

Basic Steps for Configuring FCIP IPSec

2. Configure IKE and ACLs

MDS9222i(config)# crypto ike domain ipsec
MDS9222i(config-ike-ipsec)# key LAB address 10.10.30.1
MDS9222i(config-ike-ipsec)# policy 1
MDS9222i(config-ike-ipsec-policy)# encryption 3des
MDS9222i(config-ike-ipsec-policy)# hash md5
MDS9222i(config-ike-ipsec-policy)# end

MDS9222i# conf t
Enter configuration commands, one per line. End with CNTL/Z.
MDS9222i(config)# ip access-list FCIP5 permit ip 10.10.30.2 0.0.0.0 10.10.30.1 0.0.0.0
MDS9222i(config)#
Configuring IPSec

Basic Steps for Configuring FCIP IPSec

3. Configure Transform Set and Crypto Map

MDS9222i(config)# crypto transform-set domain ipsec TFSET5 esp-aes 256 esp-sha1-hmac
MDS9222i(config)#

MDS9222i(config)# crypto map domain ipsec CMAP5 1
MDS9222i(config-crypto-map-ip)# match address FCIP5
MDS9222i(config-crypto-map-ip)# set peer 10.10.30.1
MDS9222i(config-crypto-map-ip)# set transform-set TFSET5
MDS9222i(config-crypto-map-ip)# set security-association lifetime seconds 3600
MDS9222i(config-crypto-map-ip)# set security-association lifetime megabytes 100
MDS9222i(config-crypto-map-ip)# set pfs group5
MDS9222i(config-crypto-map-ip)# end
MDS9222i#
Configuring IPSec

Basic Steps for Configuring FCIP IPSec

4. Bind the Crypto Map to the GE interface used by FCIP tunnel

MDS9222i# conf t
Enter configuration commands, one per line. End with CNTL/Z.
MDS9222i(config)# interface gigabitethernet 1/3
MDS9222i(config-if)# crypto map domain ipsec CMAP5
MDS9222i(config-if)# no shut

5. Enable the FCIP tunnel

MDS9222i(config)# interface fcip 5
MDS9222i(config-if)# no shut

Reference
Full FCIP IPSec configuration example at:
Application I/O Acceleration

- Distance between data centers impacts performance of disk replication and tape backups
- Latency introduced by distance is compounded by multiple round trips per command
- Different acceleration methods are available to accelerate data over distance
  - I/O Accelerator (IOA) for disk and tape over FC or FCIP
  - Write Acceleration for disk over FC (FC-WA)
  - Write Acceleration for disk over FCIP (FCIP-WA)
  - Tape Acceleration for tape over FCIP (FCIP-TA)
Acceleration Data Flow Concepts

Write Acceleration (WA)

- Synchronous Replication and Tape Backup are Similar – both have one outstanding I/O
- Tape drives further impacted by limited buffering and physical media
- Write Acceleration spoofs Transfer Ready only, Tape Acceleration spoofs Command Status

Tape Acceleration (TA)
Comparison of Acceleration Methods

FC-WA on SSM

- Ingress FC
- DPP Accelerate WA

Any FC Line Card

- Egress FC
- Ingress FC

FC-WA on SSM

- Egress FC
- DPP Accelerate WA

Any FC Line Card

FCIP-WA/TA on MSM

- Ingress FC
- FCIP Accelerate WA & TA
- Egress GE

Any FC Line Card

- FCIP Accelerate WA & TA
- Ingress GE

FCIP-WA/TA on MSM

- Egress FC
- FCIP Accelerate WA & TA

Any FC Line Card

FCIP on MSM or SSN-16

- Ingress FC
- Egress GE

FCIP on MSM or SSN-16

- FCIP
- Ingress GE

IOA on MSM or SSN-16

- IOA Accelerate WA & TA

IOA on MSM or SSN-16
### Choosing an Acceleration Method

<table>
<thead>
<tr>
<th></th>
<th>IOA</th>
<th>SSM-WA</th>
<th>FCIP-WA</th>
<th>FCIP-TA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Attached Devices</strong></td>
<td>1/2/4/8G</td>
<td>1/2G</td>
<td>1/2/4/8G</td>
<td>1/2/4/8G</td>
</tr>
<tr>
<td><strong>ISL Speed</strong></td>
<td>1/2/4/8/10G</td>
<td>1/2/4/8/10G</td>
<td>1GE</td>
<td>1GE</td>
</tr>
<tr>
<td><strong>Port Channels</strong></td>
<td>Yes, up to 16 ISLs</td>
<td>Yes, up to 16 ISLs</td>
<td>Yes, up to 16 ISLs</td>
<td>No</td>
</tr>
<tr>
<td><strong>Equal Cost Multi-Path</strong></td>
<td>Yes, up to 16</td>
<td>Yes, up to 16</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Disk Acceleration</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Tape Acceleration</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
IOA Overview

- Common Network Service for all acceleration
  Common Disk Acceleration and Tape Acceleration engine
- Transport & Speed Agnostic
  Independent of transport - FC, FCIP, FCOE
  Gigabit Ethernet, 1/2/4/8/10G FC
- Compression for FC and FCIP
  Reduces the WAN/MAN bandwidth requirements
- Highly Available
  Support for port channels and multiple paths for Data Center Interconnect
  Engine Clustering for redundancy and load balancing
- Highly Scalable
  Simple capacity and availability planning (no topological limitations)
- Transparent Fabric Service
  Deploy the IOA engine anywhere in the SAN using FC-Redirect
- Supported on 9222i, MSM18/4 and SSN-16
IOA Basics

Backup Servers
Primary DC
SAN
Primary Storage Array

Backup DC
SAN
Tapes/VTLs
Secondary DC
SAN
Secondary Storage Array

Backup/TA
FCIP
FC
Replication/WA
IOA Sites and Engines

- **IOA Site**
  - A local set of switches within the fabric (e.g. sjc-bldg6)

- **IOA Engine (interface)**
  - Represents a Service engine within the SSN-16 (4 engines) or MSM-18/4 (1 engine).
  - Ex. ioa2/1, ioa3/1-3/4

Note: a separate IOA license is required for each engine
IOA Clusters

- A set of IOA engines in a pair of IOA sites that operate as one group to provide IOA service
  - Ex. switches S1, S2, S3
- Automatic load balancing among engine pairs
- A site can be in multiple clusters (bunker sites)
  - A switch can be in multiple clusters

Note: an engine is bound to only one cluster
IOA Flows

- **IOA Flow**
  A flow accelerated within an IOA cluster.
  Each flow is identified by
  \{Initiator PWWN, Target PWWN, VSAN ID\}
  Eg. (SI₁, ST₁, V₁)

- **IOA Flow Group**
  A set of IOA flows classified for a given purpose.
  E.g. SRDF flow group and TSM flow group.

Backup Servers

Tape Library

IOA Cluster

Storage Array

Site : sjc-bldg6

Site : rtp-bldg8

Backup Servers

Tape Library

IOA Cluster

Storage Array

Site : sjc-bldg6

Site : rtp-bldg8

IOA Flows
How IOA Works

Flow is accelerated and sent towards normal routing path

Flow is re-directed to IOA engine

WAN or MAN links
## IOA Scalability Considerations

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of switches in a cluster</td>
<td>4</td>
</tr>
<tr>
<td>Number of Switches in the SAN for FC-Redirect</td>
<td>34</td>
</tr>
<tr>
<td>Number of IOA interfaces in a switch</td>
<td>44</td>
</tr>
<tr>
<td>Number of hosts per target</td>
<td>128</td>
</tr>
<tr>
<td>Number of flows in a cluster</td>
<td>1024</td>
</tr>
<tr>
<td>Number of flows across all clusters</td>
<td>1024</td>
</tr>
<tr>
<td>Number of flows per IOA service engine (hard limit)</td>
<td>128</td>
</tr>
<tr>
<td>Number of flows per IOA service engine (suggested soft limit)</td>
<td>64</td>
</tr>
</tbody>
</table>
# IOA Performance Considerations

<table>
<thead>
<tr>
<th></th>
<th>MSM-18/4</th>
<th>SSN-16</th>
</tr>
</thead>
<tbody>
<tr>
<td>WA</td>
<td>1182 MB</td>
<td>1196 MB</td>
</tr>
<tr>
<td>TA</td>
<td>1111 MB</td>
<td>1215 MB</td>
</tr>
</tbody>
</table>

1) The numbers are **application throughputs** measured with 64K I/O size. With 32K I/O size, the numbers observed were around 20-40 MBps lower.
2) Delay on the link was 30ms
3) Throughput per IOA engine was measured with 8 flows.
IOA Deployment Considerations

- Number of IOA Engines required
  Minimum of 1 engine per site
  2 engines per site gives better scalability and availability

  FCIP : 1-engine for every 3-FCIP engines
  Consolidate IOA and FCIP with SSN-16
  FC  : 1-engine for 1200 MB ISL bandwidth

- Licensing
  IOA license needed for each engine
IOA Deployment Considerations

- Limitations in FC-Redirect if Gen-1 modules are employed
  
  Limit of 32 targets per switch  -> Use Gen-2 or greater modules

- Zoning Considerations
  
  Non-Optimal zoning (all servers to all disks) is not scalable

- Cluster Name Restriction
  
  Same naming across all clustered applications (SME and IOA)
Write Acceleration Design Considerations

High Availability and Load Balancing:

- Can be used for native FC replication (SSM) or FCIP replication (IPS, MPS and MSM)
- If FCIP Write Acceleration HA is required, then Portchannels must be used
- Equal cost FSPF load balancing for FCIP Write Acceleration **not supported**
- Works with:
  - EMC SRDF, Mirrorview, SANCOPY
  - HDS TrueCopy
  - HP CA-XP, CA-MVA
  - IBM FlashCopy, FastT
Configuring FCIP Write Acceleration

Configure FCIP Write Acceleration

mds9513-labD# conf t
Enter configuration commands, one per line. End with CNTL/Z.
mds9513-labD(config)# interface fcip 1
mds9513-labD(config-if)# write-accelerator

Enable FCIP write acceleration – same for both sides of FCIP tunnel

mds9513-labD# show interface fcip 1 fcip100
TCP Connection Information

Write acceleration mode is configured on; operationally on
Additional FC-WA Information

- Fibre Channel Write Acceleration requires specific source and destination device configuration
  - FC-WA requires the initiator and target be connected to the SSM ports
    - Configure scsi-flow feature on specific SSM ports
    - Configure the specific initiator and target PWWNs and VSANs to use FC-WA
    - Configure the number of 2K buffers to be used by FC-WA
      \[(\text{Number of concurrent SCSI writes} \times \text{size of SCSI writes in bytes}) / \text{FCP data frame size in bytes}\]

- No restrictions on HA design; FSPF and Port Channeling may be used
Configuring FC Write Acceleration

Configure FC Write Acceleration

```bash
mds9513-labD# conf t
Enter configuration commands, one per line. End with CNTL/Z.
mds9513-labD(config)# boot ssi bootflash:m9000-ek9-ssi-mz.4.1.1i.bin module 2
mds9513-labD(config)# ssm enable feature scsi-flow module 2
mds9513-labD(config)# cfs distribute
mds9513-labD(config)# scsi-flow flow-id 1 initiator-vsan 100 initiator-pwn 40:00:00:e0:8b:0c:d6:25 target-vsan 100 target-pwn 40:00:00:e0:8b:0c:e5:f1
mds9513-labD(config)# scsi-flow flow-id 1 write-acceleration
Configure src/dest arrays (only required on MDS attached to initiator array)
Load SSI image on SSM module
Enable SCSI-flow feature on SSM module
Turn on CFS to distribute WA config

Configure the number of 2K buffers to be used by FC-WA
(Number of concurrent SCSI writes * size of SCSI writes in bytes)/FCP data frame size in bytes)
16 outstanding I/Os * 32K write block size / 2K FCP data frame size = 256
```
FCIP TA Design Requirements

- Tape Acceleration is not supported over Port-channels or equal cost FSPF paths
  Requirement due to state information kept in interface

- HA design requires costing parallel links so one path preferred over another
  HA through redundant links

- Can use separate VSAN and FCIP tunnel to allow Port-channels for Write Acceleration
Configuring FCIP Tape Acceleration

Configure FCIP Write Acceleration

```bash
mds9513-labD# conf t
Enter configuration commands, one per line. End with CNTL/Z.
mds9513-labD(config)# interface fcip 1
mds9513-labD(config-if)# write-accelerator tape-accelerator
```

Enable FCIP tape acceleration – same for both sides of FCIP tunnel

```bash
mds9513-labD# MDS9222i# sh int fcip 1
ccip1 is up.
Write acceleration mode is configured on; operationally on
Tape acceleration mode is configured on; operationally on
Tape read acceleration mode is operationally on
Tape Accelerator flow control buffer size is automatic
```
SAN Extension Fabric Stability

- Connecting existing SAN fabrics or extending a SAN fabric creates SAN design challenges
  - Minimize the impact of change in fabric services across geographically distributed sites
  - Limit fabric control traffic such as RSCNs and Build/Reconfigure Fabric (BF/RCF) to local VSANs
  - Connecting SAN fabrics with the same domain IDs

- Inter-VSAN Routing (IVR) can be used to address these challenges
  - IVR enables a layer 3 hierarchical layer design to SAN fabrics to isolate domains
  - IVR only sends selective RSCNs to edge switches, preventing disruption of fabric services
  - IVR with NAT allows two existing SAN fabrics with the same domain ID to be connected through a third transit VSAN
SAN Extension with IVR

- Any failure in transit VSAN_20 (network equipment, physical or logical failure) will not disrupt VSAN_10 or VSAN_30 fabric
- Works with any transport service (FC, SONET/SDH, DWDM/CWDM, FCIP)

VSAN_5 - Site 1 Host Fabric
VSAN_10 - Site 1 Replication Fabric
VSAN_20 - Inter-site SAN Extension Fabric
VSAN_30 - Site 2 Replication Fabric

But is it required? Host traffic is already protected via separate VSAN and if MAN/WAN goes down, replication traffic will still be impacted.
SAN Extension with IVR

In some case, YES. When host and replication traffic share the same physical port in the same VSAN, then IVR and transit VSANs are a must.

- Any failure in transit VSAN_20 (network equipment, physical or logical failure) will not disrupt VSAN_10 or VSAN_30 fabric.

VSAN_10 - Site 1 Host and Replication Fabric
VSAN_20 - Inter-site SAN Extension Fabric
VSAN_30 - Site 2 Host and Replication Fabric
QoS for FCIP SAN Extension

Most FCIP Implementations Use Dedicated Links, however:

- Acceptance of FCIP generating interest in converged IP network for FCIP SAN Extension

- No generally accepted DSCP values for FCIP traffic (unlike voice with DSCP EF)
  
  QoS—Define Marking and classification
  Mark DSCP according an agreed value
  Separate consideration of FCIP data and control packet

- Bandwidth Reservation
  FCIP has no support for reservation protocol
  Simulated using Min / Max B/W command
  (can be considered as a type of reservation)
FCIP QoS Mapping Proposal

- Synchronous Data Replication: Bursty, high bandwidth
  Can be mapped into Mission Critical (AF31 / DSCP 26)

- Asynchronous Data: Bursty, low to medium b/w
  Can be mapped into Transactional data (AF21 / DSCP 18)
  Can also mapped into Bulk data (AF11 / DSCP 10)

- Backup Data: 150 ~ 500 ms, constant (during backup), medium b/w
  Can be mapped into Bulk data (AF11 / DSCP 10)

- Control Packets
  Both Control and Data traffic can be assigned the same class
  If needed, Control Packets can be assigned CS6 or DSCP 48
## Classification and Marking Design

### Modified for SAN Traffic

<table>
<thead>
<tr>
<th>Application</th>
<th>L3 Classification</th>
<th>L2 CoS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routing / FCIP control</td>
<td>IPP=6, PHB=CS6, DSCP=48, CoS=6</td>
<td></td>
</tr>
<tr>
<td>Voice</td>
<td>IPP=5, PHB=EF, DSCP=46, CoS=5</td>
<td></td>
</tr>
<tr>
<td>Video Conferencing</td>
<td>IPP=4, PHB=AF41, DSCP=34, CoS=4</td>
<td></td>
</tr>
<tr>
<td>Streaming Video</td>
<td>IPP=4, PHB=CS4, DSCP=32, CoS=4</td>
<td></td>
</tr>
<tr>
<td>FCIP SYNC</td>
<td>IPP=3, PHB=AF31*, DSCP=26, CoS=3</td>
<td></td>
</tr>
<tr>
<td>Call Signaling</td>
<td>IPP=3, PHB=CS3*, DSCP=24, CoS=3</td>
<td></td>
</tr>
<tr>
<td>FCIP ASYNC</td>
<td>IPP=2, PHB=AF21, DSCP=18, CoS=2</td>
<td></td>
</tr>
<tr>
<td>Network Management</td>
<td>IPP=2, PHB=CS2, DSCP=16, CoS=2</td>
<td></td>
</tr>
<tr>
<td>FCIP backup / FCIP Async</td>
<td>IPP=1, PHB=AF11, DSCP=10, CoS=1</td>
<td></td>
</tr>
<tr>
<td>Scavenger</td>
<td>IPP=1, PHB=CS1, DSCP=8, CoS=1</td>
<td></td>
</tr>
<tr>
<td>Best Effort</td>
<td>IPP=0, PHB=0, DSCP=0, CoS=0</td>
<td></td>
</tr>
</tbody>
</table>

Value used when configuring MDS
Configuring FCIP QOS

Configure FCIP QOS for Synchronous Replication

mds9513-labD# conf t
Enter configuration commands, one per line.  End with CNTL/Z.
mds9513-labD(config)# interface fcip 1
mds9513-labD(config-if)# qos control 48 data 26

Enable FCIP QOS values on both sides

mds9513-labD# MDS9222i# sh int fcip 1
fcip1 is up
  .
  .
  .

QOS control code point is 48
QOS data code point is 26
TCP Connection Information
  2 Active TCP connections
FCIP Tuning
MDS FCIP Configuration Guidelines

- TCP Parameter settings
- Understanding FC Buffers and FCIP
- Monitoring for FC Frame expiry and why it is important
- MDS packet shaping methods so not to over run network
- Understanding MTU and the WAN network
MDS FCIP – TCP Parameters

- Select the proper bandwidth settings
tcp max-bandwidth and tcp min bandwidth
- Use ips measure-rtt or ping command determine RTT
  round-trip-time used to set initial TCP window size

```plaintext
fcip profile 1
  ip address 200.200.200.1
tcp max-bandwidth-mbps 800 min-available-bandwidth-mbps 500 round-trip-time-us 80
```

```plaintext
MDS9216-TOP# ips measure-rtt 200.200.200.2 int gig 2/4
Round trip time is 82 micro seconds (0.08 milli seconds)
or
MDS9216-TOP# ping 200.200.200.1
--- 200.200.200.1 ping statistics ---
  rtt min/avg/max/mdev = 0.068/0.079/0.091 ms
```
MDS FCIP – TCP Parameters

- Always use Selective ACK (SACK)
- If slow link (<T3/E3):
  
  Keep `tcp cwm burstsize` low (<20kB)
  
  Tune FC receive buffers—BB_Credits and performance buffers (`fcrxbbcredit`)
- If shared link (other traffic):
  
  Determine available bandwidth (link b/w – other traffic b/w) and set max-bandwidth = available bandwidth and min-available-bandwidth = 80-95% max-bandwidth
  
  Consider QoS policies; carving out bandwidth
- If shared link (other FCIP tunnel):
  
  Determine available bandwidth (link b/w – other traffic b/w) and set max-bandwidth = 50% available bandwidth and min-available-bandwidth = 80-95% max-bandwidth
- If a dedicated link:
  
  Set min=95%max and max=path bandwidth (FCIP will send at this rate)
  
  If GigE all the way, set `tcp cwm burstsize` to 100kB
Fibre Channel Frame Buffering

- FC frames buffered and queued in intermediate switches
- Buffer depth controlled by Rx BB_Credit parms
  
  switchport fcrxbbcredit <x>
  switchport fcrxbbcredit performance-buffers <y>

  Performance buffers added by default unless configured otherwise
- Hop-by-hop traffic flow paced by return of R_RDY frames
- Frames cannot sit in any switch for >500ms (or they will expire)
Frame Buffering: FCIP and FC

- FCIP presents a lower bandwidth pipe (If WAN link)
  Drain rate (send rate) depends upon bandwidth and congestion
- Slow ramp up of traditional TCP can cause FC frame expiry in some conditions
  Mixture of slow link (e.g. <DS3/E3; retransmissions, many sources, big buffers)

Increase TCP Send Buffer if:
- Slow WAN link is preventing FC traffic destined to other devices not across FCIP link to have issues

Backlog Here If Queue Can’t Drain Due to:
- Slow WAN link and long RTT
- Packet loss and retransmissions
- Many sources (only one shown)
- Buffer too big
Frame Expiration

- 500ms timer is fixed (not configurable)

- Behavior possible whenever FC Rx buffers cannot drain at a sufficient rate (at least 2x #Rx buffers/second)

  Caused by one or more of:
  
  too many sources (speed mismatch)
  slow downstream (slow receiver)
  FC Rx buffers too deep

- Possible causes are:

  Long optical links with low BB_Credits (i.e. can’t drain quickly) - misconfiguration.
  
  FCIP over long, slow links (e.g. T1 or T3 with 50+ms RTT) with occasional packet loss/retransmission
Monitor FCIP Frame Timeouts

MDS9216-TOP# show ips stats dma int gigabitethernet 2/4
Dma-bridge ASIC Statistics for port GigabitEthernet2/4

Hardware Egress Counters
1030008889 Good, 0 bad protocol, 0 bad header cksum, 0 bad FC CRC

Hardware Ingress Counters
1843269599 Good, 0 protocol error, 0 header checksum error
0 FC CRC error, 0 iSCSI CRC error, 0 parity error

Software Egress Counters
1030008581 good frames, 0 bad header cksum, 0 bad FIFO SOP
0 parity error, 0 FC CRC error, 0 RDL ok, 0 RDL drop (too big), 0 RDL ttl_1
1278210315 idle poll count, 116279 loopback
0 FCC PQ, 0 FCC EQ, 0 FCC generated
Flow Control: 0 [0], 0 [1], 0 [2], 0 [3]

Software Ingress Counters
1843269472 Good frames, 0 header cksum error, 0 FC CRC error
0 iSCSI CRC error, 0 descriptor SOP error, 0 parity error
141281 frames soft queued, 0 current Q, 732 max Q, 0 low memory
0 out of memory drop, 0 queue full drop
0 RDL ok, 0 RDL drop (too big)
Flow Control: 0 [0], 141281 [1], 0 [2], 0 [3]

Monitor for 500ms Timeout Issues
TCP Window and Buffer with FCIP

You cannot configure the TCP window size directly. This value is automatically calculated from the product of the maximum bandwidth x RTT.

In SAN-OS 2.0 and later, the RTT will dynamically adjust up to eight times the configured value in the FCIP profile. The TCP sender dynamically changes the maximum window size accordingly.

MDS9216-TOP# sh interface fcip 2 counters

TCP Parameters
Path MTU 1500 bytes
Current retransmission timeout is 200 ms
Round trip time: Smoothed 2 ms, Variance: 4
Advertized window: Current: 97 KB, Maximum: 97 KB, Scale: 5
Peer receive window: Current: 98 KB, Maximum: 98 KB, Scale: 5
Congestion window: Current: 52 KB, Slow start threshold: 103 KB
Current Send Buffer Size: 97 KB, Requested Send Buffer Size: 0 KB
CWM Burst Size: 50 KB

IP compression statistics
4059579130004 rxbytes, 273626035160 rxbytes compressed
2287961989360 txbytes 633365852164 txbytes compressed, 0 txbytes non-compressed
3.61 tx compression ratio

TCP send buffer can be increased to allow faster removal of frame off FC interface RX buffer
TCP send buffer would only require adjustment if other local exchanges were occurring off the same FC port that is being used by FCIP application; the minimum size is 14k and is dynamically calculated from bandwidth x rtt product.
FCIP Capacity Planning
FCIP – Capacity Planning
SAN Extension Tuner (SET)

SAN extension tuner - lightweight tool to assist in FCIP tuning by generating various SCSI traffic flows

- Creates a virtual N-port on an IPS, MPS, MSM or SSN-16 port that can act as both initiator and target to generate traffic
- Measures throughput and response time per I/O over the FCIP tunnels
- Model effect of storage data in a shared network
Using SAN Extension Tuner

- SAN Extension-Tuner allows you to configure a unused iSCSI interface as a FC Initiator and Target
- Used to create Virtual FC devices in the MDS to generate FC traffic
- FC traffic is sent across FCIP link (ISL) for WAN load test, help tune TCP parameters and exercise features; to achieve desired throughput across FCIP a number of parameters must be tuned to keep WAN pipe full
  - TCP Parameters (Window size, SACK…)
  - Outstanding SCSI I/Os allowed by application
  - Transfer size used by application
- Requires 2.0 SAN/OS or greater
  - Note: MSM 18+4/9222i require SAN-OS 3.3(1) or higher
SAN Extension Tuner

- Only read and write SCSI commands are used, entire SCSI suite in not supported
- Multiply Virtual N-ports can be created to:
  - Overload FCIP link to stress WAN Provisioning Compression Results
  - Troubleshoot issues without using FC real devices
- Use data provided with tool or use your own
- Simple Setup Wizard
- Wizard creates Zone for the virtual N-ports
SET Use from Fabric Manager

Click on FCIP Link

ERROR SAN Extension Tuner Wizard - /SAN/Networkers

Read I/O: Product of Transfer Size and Read I/O needs to be less than 128MB.
SET – Monitor Results, Make Adjustments

Monitor ISL Link Utilization Tool

Auto Created Zone
Q&A
Other Storage Related Sessions

- BRKSAN-1121  SAN Core Edge Design Best Practices
- BRKSAN-2047  FCoE - Design, Operations and Management Best Practices
- BRKSAN-2123  Storage Cloud Concept and Design for Disaster Recovery
- BRKSAN-3707: Advanced SAN Services
- BRKDCT-1044: FCoE for the IP Network Engineer
Additional Information

- Cisco Storage Networking

- Cisco Data Center Networking

- Storage Network Industry Association (SNIA)
  [http://www.snia.org](http://www.snia.org)

- Internet Engineering Task Force—IP Storage

- ANSI T11—Fibre Channel
  [http://www.t11.org/index.html](http://www.t11.org/index.html)
Complete Your Online Session Evaluation

- Receive 25 Cisco Preferred Access points for each session evaluation you complete.
- Give us your feedback and you could win fabulous prizes. Points are calculated on a daily basis. Winners will be notified by email after July 22nd.
- Complete your session evaluation online now (open a browser through our wireless network to access our portal) or visit one of the Internet stations throughout the Convention Center.
- Don’t forget to activate your Cisco Live and Networkers Virtual account for access to all session materials, communities, and on-demand and live activities throughout the year. Activate your account at any internet station or visit www.ciscolivevirtual.com.
Visit the Cisco Store for Related Titles
http://theciscostores.com
Thank you.
Advanced FCIP Designs
Advanced FCIP-TA Design

Scaling FCIP-TA Beyond 1 FCIP Tunnel/GE Interface

- IOA-TA supports ECMP over multiple FC or FCIP links, but FCIP Tape Acceleration is limited to 1 path. To work around this limitation:
  - Move the target tapes devices into separate VSANs
  - Make each FCIP interface only a member of one tape target VSAN and ensure trunking turned off
  - Use IVR to allow all backup servers to access all tape devices in all target VSANs
Using GE Sub-Interfaces, Multiple FCIP Tunnels and Port Channeling to Enable High b/w FCIP

- Use separate VSANs for data replication (100), backup servers (200) and tape backup (300, 400)
- Port channel FCIP tunnels for replication traffic
- Distribute tape drives over multiple VSANs and FCIP tunnels to permit Tape Acceleration, do not port channel or use equal cost paths
- Use IVR to allow access from backup servers (VSAN 200) to tape drives (VSANs 300, 400)
FCIP WA and TA—Multiple Tunnels

Port channel FCIP interfaces 11 & 21 for replication only

Do not port channel FCIP interfaces 12 & 22 for tape backup

**fcip profile 11**
- ip address 10.10.100.1
- tcp max-bandwidth-mbps 750 min-available-bw-mbps 700

**fcip profile 12**
- ip address 10.10.200.1
- tcp max-bandwidth-mbps 250 min-available-bw-mbps 200

**fcip profile 21**
- ip address 10.11.100.1
- tcp max-bandwidth-mbps 750 min-available-bw-mbps 700

**fcip profile 22**
- ip address 10.11.200.1
- tcp max-bandwidth-mbps 250 min-available-bw-mbps 200

**interface fcip 11**
- switchport trunk mode off
- channel-group 100 force
- use profile 11
- peer-info ipaddr 10.10.100.2
- write-accelerator

**interface fcip 12**
- switchport trunk mode off
- use profile 12
- peer-info ipaddr 10.10.200.2
- write-accelerator tape-accelerator

**interface fcip 21**
- switchport trunk mode off
- use profile 21
- peer-info ipaddr 10.11.100.2
- write-accelerator

**interface fcip 22**
- switchport trunk mode off
- use profile 22
- peer-info ipaddr 10.11.200.2
- write-accelerator tape-accelerator

**interface GigabitEthernet 1/1**
- no shut

**interface GigabitEthernet 1/1.100**
- ip address 10.10.100.1 255.255.255.0

**interface GigabitEthernet 1/1.200**
- ip address 10.10.200.1 255.255.255.0

**interface GigabitEthernet 1/2**
- no shut

**interface GigabitEthernet 1/2.100**
- ip address 10.11.100.1 255.255.255.0

**interface GigabitEthernet 1/2.200**
- ip address 10.11.200.1 255.255.255.0

**interface port-channel 100**
- no shut
- switchport trunk mode off
- switchport mode E
- channel mode active

**vsan database**
- vsan 100 interface fcip 11
- vsan 100 interface fcip 21
- vsan 100 interface port-channel 100
- vsan 300 interface fcip 12
- vsan 400 interface fcip 22
FCIP WA and TA—Multiple Tunnels

Port channel FCIP interfaces 11 & 21 for replication only

Do not port channel FCIP interfaces 12 & 22 for tape backup

fcip profile 11
ip address 10.10.100.2
tcp max-bandwidth-mbps 750 min-available-bw-mbps 700
cfc profile 12
ip address 10.10.200.2
tcp max-bandwidth-mbps 250 min-available-bw-mbps 200
cfc profile 21
ip address 10.11.100.2
tcp max-bandwidth-mbps 750 min-available-bw-mbps 700
fcip profile 22
ip address 10.11.200.2
tcp max-bandwidth-mbps 250 min-available-bw-mbps 200

interface fcip 11
switchport trunk mode off
channel-group 100 force
use profile 11
peer-info ipaddr 10.10.100.1
write-accelerator
tape-accelerator
interface fcip 12
switchport trunk mode off
use profile 12
peer-info ipaddr 10.10.200.1
write-accelerator
tape-accelerator
interface fcip 21
switchport trunk mode off
channel-group 100 force
use profile 21
peer-info ipaddr 10.11.100.1
write-accelerator

interface fcip 22
switchport trunk mode off
use profile 22
peer-info ipaddr 10.11.200.1
write-accelerator

tape-drives

Vsan database
vsan 100 interface fcip 11
vsan 100 interface fcip 21
vsan 100 interface port-channel 100
vsan 300 interface fcip 12
vsan 400 interface fcip 22
FCIP Troubleshooting
Troubleshooting Section

- Tools available is SAN-OS and NX-OS
- Practical use of SPAN, PAA & Cisco Analyzers in troubleshooting
FCIP CLI Commands

Useful Commands

- show interface fcip nnn
- show interface gigabitethernet n/n
- show ips stats all
- show ips stats buffer
- show ips stats dma-bridge
- show ips stats hw-comp
- show ips stats icmp
- show ips stats ip
- show ips stats mac
- show ips stats tcp

  Show fcip interface statistics
  Show Gigabit Ethernet interface statistics
  Show all ips stats
  Show ips buffer stats
  Show dma-bridge stats
  Show hw compression stats
  Show icmp stats
  Show ip stats
  Show ethernet mac stats
  Show tcp stats
Command Line Debugging

- Available debugs depend on features enabled in SAN-OS and NX-OS
- Many different options to select when turning on debugs
  - Interface and protocol level debugging available
- Where is it output going?
  - Logfile—data file in switch memory
    - Capture to direct to screen via console, telnet or ssh
- Requires admin privileges to run debugs
- Debugs can only be run from cli, no debug interface in Fabric Manager or Device Manager
FCIP Debugs

You Must Attach to the Module You Are Debugging

module-2# debug ips fcip ?
  all           Configure FCIP debugging
  ctlpkt        Configure FCIP Control Pkt debugging
  ctlpkt-detail Configure verbose FCIP Control Pkt debugging
  datapkt       Configure FCIP Data Pkt debugging
  datapkt-detail Configure verbose FCIP Data Pkt debugging
  err           Configure FCIP Error debugging
  fsm           Configure FCIP FSM debugging
  tape-acc      Configure FCIP Tape accelerator
  tape-acc-det  Configure FCIP Tape accelerator debug detail
  tape-acc-err  Configure FCIP Tape accelerator error
  write-acc     Configure FCIP Write accelerator
  write-acc-err Configure debug FCIP Write accelerator errors
  write-acc-event Configure debug FCIP Write accelerator events
  write-acc-pkt Configure debug FCIP Write accelerator packet
Gathering Protocol Traces for Analysis

- Using built-in FC Analyzer (CLI)
- Using Cisco Protocol and Traffic Analyzer on PC (local & remote)
- Using the MDS Port Analyzer Adapter (PAA)
- Using SPAN
- Using an external FC Analyzer

All non-disruptive to switch operations and traffic on the SAN
MDS FCAnalyzer
(SAN-OS and NX-OS Imbedded)

- Output is displayed to the console in readable sniffer like format
- Is only used to monitor Fibre Channel Traffic to and from supervisor on the MDS9000
  Traffic like Fabric Login’s, FSPF routing, Switch to switch control traffic
- Output can go direct to your console screen or to a workstation running a color Ethereal program

Note: SPAN is used not only for FC port to FC port monitoring, but also used to SPAN iSCSI and FCIP ports
FCanalyzer Options

- Local or remote - where to send the trace; can be to local devices or remote PAA attached to different MDS switch

- Brief or detailed - header information vs. full output of frame including hex; detail is default

- Limit-captured-frames - number of frames to capture; default is fcanalyzer will trace 100 frames; specifying zero is unlimited frame capture

- Limit-capture-size - allows to capture N number of bytes of frame; useful for not capturing frame data when it is not relevant to troubleshooting
Use of SPAN Feature

- Used for FC port to FC port analyzing
- Same type of tool as used on Cisco Catalyst® products (Catalyst uses Port Monitor)
- Can be left configured on switch
- Ingress and egress ports are sent to an FC-port setup as a SPAN Destination (SD-port type)
- No limits to where the ports are located on the MDS switch fabric
- Used to output to third-party test equipment or to Cisco Port Analyzer Adapter
Using the FCanalyzer with FCIP

MDS9216-TOP# conf t
Enter configuration commands, one per line. End with CNTL/Z.

MDS9216-TOP(config)# fcanalyzer local brief display-filter mdshdr.vsan==300

Warning: Couldn't obtain netmask info (eth2: no IPv4 address assigned).
Capturing on eth2

8.986146 ff.ff.fd -> ff.ff.fd 0x46a6 0xffffffff SW_ILS ELP
8.986233 ff.ff.fd -> ff.ff.fd 0x46a6 0x469a FC Link Ctrl, ACK1
8.987140 ff.ff.fd -> ff.ff.fd 0x469b 0xffffffff SW_ILS ELP
8.987539 ff.ff.fd -> ff.ff.fd 0x469b 0x46a7 FC Link Ctrl, ACK1
8.988273 ff.ff.fd -> ff.ff.fd 0x469b 0x46a7 SW_ILS SW_RJT (ELP)
8.988790 ff.ff.fd -> ff.ff.fd 0x46a6 0x469a SW_ILS SW_ACC (ELP)
8.989438 ff.ff.fd -> ff.ff.fd 0x469b 0x46a7 FC Link Ctrl, ACK1
8.991501 ff.ff.fd -> ff.ff.fd 0x46a6 0x469a FC Link Ctrl, ACK1
8.992965 ff.ff.fd -> ff.ff.fd 0x46a8 0xffffffff SW_ILS EFP
8.993037 ff.ff.fd -> ff.ff.fd 0x46a8 0x469c FC Link Ctrl, ACK1
9.001146 ff.ff.fd -> ff.ff.fd 0x469d 0xffffffff SW_ILS EFP
9.001603 ff.ff.fd -> ff.ff.fd 0x469d 0x46a9 FC Link Ctrl, ACK1
9.001719 ff.ff.fd -> ff.ff.fd 0x46a8 0x469c SW_ILS SW_ACC (EFP)

- Filter on VSAN 300
- No shut of FCIP interface
- Complete view of standard ISL initialization and switch fabric merge
- You do not capture FCIP protocol; FCIP protocol is viewed with debugs and/or SPAN (monitor) of WAN network
SPAN of FCIP Interface to PAA and Cisco Protocol Analyzer

[Diagram showing Fabric Manager interface and SPAN configuration]

PC w/Cisco Protocol Analyzer (ethereal)

DS-PAA
# PAA Capture Modes

## No Truncate Mode (NTM)

Fibre Channel frames are encapsulated into Ethernet frames without any modification to the payload. Ethernet devices must support jumbo frames for this mode to work.

<table>
<thead>
<tr>
<th>Field</th>
<th>Length</th>
<th>Description</th>
<th>Ethernet Header</th>
<th>SOF-T VSAN</th>
<th>FC Frame</th>
<th>EOF-T VSAN</th>
<th>Enet CRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Header</td>
<td>14 bytes</td>
<td>(EDA + SDA + Type)</td>
<td>14 bytes</td>
<td>2 bytes</td>
<td>0 - 2164 bytes</td>
<td>4 bytes</td>
<td>4 bytes</td>
</tr>
<tr>
<td>SOF-T VSAN</td>
<td>2 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EOF-T VSAN</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enet CRC</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Ethernet Truncate Mode (ETM)

The PAA truncates a Fibre Channel frame to a maximum payload of 1496 bytes. The Fibre Channel payload transmitted is 1472 bytes.

<table>
<thead>
<tr>
<th>Field</th>
<th>Length</th>
<th>Description</th>
<th>Ethernet Header</th>
<th>SOF-T VSAN</th>
<th>FC Frame</th>
<th>EOF-T VSAN</th>
<th>Enet CRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Header</td>
<td>14 bytes</td>
<td>(EDA + SDA + Type)</td>
<td>14 bytes</td>
<td>2 bytes</td>
<td>0 - 1490 bytes</td>
<td>4 bytes</td>
<td>4 bytes</td>
</tr>
<tr>
<td>SOF-T VSAN</td>
<td>2 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EOF-T VSAN</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enet CRC</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Shallow Truncate Mode (STM)

In STM, the PAA truncates the Fibre Channel frame to 256 bytes.

<table>
<thead>
<tr>
<th>Field</th>
<th>Length</th>
<th>Description</th>
<th>Ethernet Header</th>
<th>SOF-T VSAN</th>
<th>FC Frame</th>
<th>EOF-T VSAN</th>
<th>Enet CRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Header</td>
<td>14 bytes</td>
<td>(EDA + SDA + Type)</td>
<td>14 bytes</td>
<td>2 bytes</td>
<td>0 - 256 bytes</td>
<td>4 bytes</td>
<td>4 bytes</td>
</tr>
<tr>
<td>SOF-T VSAN</td>
<td>2 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EOF-T VSAN</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enet CRC</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Deep Truncate Mode (DTM)

In DTM, the PAA truncates the Fibre Channel frame to 64 bytes. By default, the PAA is configured in DTM mode.

<table>
<thead>
<tr>
<th>Field</th>
<th>Length</th>
<th>Description</th>
<th>Ethernet Header</th>
<th>SOF-T VSAN</th>
<th>FC Frame</th>
<th>EOF-T VSAN</th>
<th>Enet CRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Header</td>
<td>14 bytes</td>
<td>(EDA + SDA + Type)</td>
<td>14 bytes</td>
<td>2 bytes</td>
<td>0 - 64 bytes</td>
<td>4 bytes</td>
<td>4 bytes</td>
</tr>
<tr>
<td>SOF-T VSAN</td>
<td>2 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FC Frame</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EOF-T VSAN</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enet CRC</td>
<td>4 bytes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Display filters can be applied before or after capture.

Right click on field to filter.
GiGE Monitoring of FCIP

FCIP GigE Interface Must Be Up and Healthy

show ips stats ip interface gigabitethernet 2/2 detail
show ips stats tcp interface gigabitethernet 2/2 detail
show ips stats mac interface gigabitethernet 2/2

Device Manager Monitor Views Can Be Actively Kept Open and Graphed in Real Time