Solutions for Disaster recovery: SAN Extension Design

Rajinder Singh

Cisco Networkers Solutions Forum 2007
Agenda

- Basic SAN Extension Principles
  Dual path, HA options

- Alternate Designs
  Hub-spoke, multi-hop

- SAN Extension Choices and Enhancements
  Optical: Extended B2B credits, B2B credit spoofing, Port-tracking
  FCIP: Compression, Encryption, Application Acceleration

- MDS 9000 Product Family
Design Criteria

Factors to consider for SAN Extension

- Applications using SAN Extension
  - Synchronous/Asynchronous Replication, Data Backup/Restore
- Application latency requirements
  - Applications that use synchronous replication may be impacted
- Application throughput requirements
  - Determines bandwidth requirements
- Transport options available
  - What choices are available for SAN Extension
Typical SAN Design

- Servers with two FibreChannel connections to storage arrays for high availability
  - Use of multipath software is required in dual fabric host design
- SAN extension fabrics typically separate from host access fabrics
  - Replication fabric requirements generally specified by array vendor
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
Fabric Consolidation with VSANs

- Separate physical fabrics
- Over-provisioning ports on each island
- High number of switches to manage

Collapsed Fabric with VSANs
- Common redundant physical infrastructure
- Less over-provisioning required – lower $$
- Fewer switches to manage
- Move unused ports non-disruptively
Virtual SANs and fabric zoning are very complimentary

- Hierarchical relationship – provision VSANs then assign independent zones per VSAN
- VSANs divide up the physical infrastructure
- Zones provide added security and allow sharing of device ports
- VSANs provide traffic statistics
- VSANs only changed when ports needed per virtual fabric
- Zones can change frequently (backup)
- Ports are added/removed non-disruptively to VSANs
Inter-VSAN Routing (IVR): Sharing Resources Across VSANs

- Allows sharing of centralized storage services such as tape libraries and disks across VSANs – without merging separate fabrics (VSANs)
- Provides high fabric resiliency and VSAN-based manageability
  - Distributed, scaleable, and highly resilient architecture
  - Transparent to third-party switches
- Enables blade-per-VSAN architecture for blade servers
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
SAN Extension Design: Adding Fabric HA

Distribute Green and Blue VSANs Between All Switches

- All portchannels trunk both VSANs…TE_Port (EISL)
- Each VSAN protected by:
  - Portchannels over diverse paths
  - Multiple switches/directors
- Same per VSAN load balancing:
  - SRCID/DESTID, or
  - SRCID/DESTID/OXID
Other SAN Extension Implementations

Hub and Spoke

Multi-hop
## SAN Extension Technology Options

The image illustrates the options for extending SANs with increasing distance across different data centers: Center, Campus, Metro, Regional, and National. The options are categorized under Optical and IP technologies.

### Optical Technologies
- **Dark Fiber**
  - Sync
  - Limited by Optics (Power Budget)
- **CWDM**
  - Sync (1,2Gbps)
  - Limited by Optics (Power Budget)
- **DWDM**
  - Sync (1,2,10Gbps per λ)
  - Limited by BB_Credits
- **SONET/SDH**
  - Sync (1,2Gbps + subrate)
  - Async
- **SONET/SDH**
  - Sync (1,2Gbps + subrate)
  - Async

### IP Technologies
- **MDS9000 FCIP**
  - Sync (Metro Eth)
  - Async (1Gbps)
Dark Fiber

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

- Client protection only; Upper Layer Protocol (ULP), either SAN or application, responsible for failover protection
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 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 SM fiber
  ~100km point-to-point or ~40km ring
- 1/2 Gigabit Fibre Channel and 1 Gigabit Ethernet currently
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 8 Gbps to 4 Gbps 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
Dense Wavelength Division Multiplexing (DWDM)

- Higher density than CWDM
  
  32 lambdas or channels in narrow band around 1550nm at 100GHz spacing (0.8nm)

- Erbium-Doped Fiber Amplifier (EDFA) amplifiable allows for longer distances than CWDM

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

- Data center to data center

- Protection options: client, splitter, or linecard
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
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.)

Use “Src/Dst” hash for load balancing (rather than “Src/Dst/Oxid” per exchange) for each extended VSAN
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 will recover lost BB_Credits
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 and ½ km long at 4 Gbps
- As distance increases, the BB_Credits need to increase as well
- Insufficient BB_Credits will throttle performance—no data will be transmitted until R_RDY is returned
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
SAN Network Solutions for Increasing Distance

- **MDS 9000 16-port FibreChannel Switching Module**
  
  All ports have up to 255 BB_Credits
  Extends distances to 510 km @ 1G FC or 255 km @ 2G FC

- **MDS 9000 14/2-port Multiprotocol Services Module**
  
  Extended BB_Credits available on FC ports 1-12
  Up to 3500 BB_Credits can be configured on any one port
  Extends distances up to 7000 km @ 1G FC or 3500 km @ 2G FC

- **MDS 9000 12/24/48-port FibreChannel Switching Modules**
  
  Extended BB_Credits available on any FC port
  Up to 4095 BB_Credits can be configured on any one port
  Extends distance up to 8190 km @ 1G, 4095 km @ 2G or 2048km @ 4G
Extending Optical SAN Extension
Optical Solutions for Increasing Distance

ONS 15454 SL-Series Card

- Negotiates up 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
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

Port Tracking vs. No Port Tracking

- No PT, No FLC No PT, FLC Fabric A
- No PT, No FLC No PT, FLC Fabric B
- PT and FLC PT & FLC Fabric A
- PT and FLC PT & FLC Fabric B

Fail ISL on Fabric “A”
MDS FCIP SAN Extension Design

- Same port channeling and VSAN trunking rules apply as with FC links
- Portchannel individual FCIP links to alternate 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
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 on second or subsequent packets)
- Jumbo frames may increase performance
  
  IP MTU of 2300 avoids splitting of TCP frames
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
  Initial rate controlled by “min-avail-bandwidth”
  Max rate controlled by “maximum-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-avail-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
FCIP Data Compression

- Cisco uses RFC standard compression algorithms implemented in both hardware and software
- MDS9000 IP Storage Services Module
  - Software-based compression for FCIP
- MDS9000 Multiprotocol Services Module
  - Hardware and software-based compression and hardware based encryption for FCIP
- 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 2:1, but may vary considerably
- Application throughput is the most important factor
FCIP Data Compression and TCP Windowing

- Compression has the effect of a variable bandwidth path
- TCP window applies to data stream before compression
  - If window size not increased, throughput will not increase
  - Need to compensate with larger TCP max window size
- MDS9000 incorporates moving average feedback to dynamically adjust TCP window according to compression rate
- Feedback mechanism is not available when using IP network-based compression solutions—manual adjustment of TCP window size required
IP vs. SAN Network-Based Compression

### Network-Based Compression
- 7200 with VAM, VAM2, VAM2+
- **Advantages**
  - Compression for some or all IP traffic (as required) in shared network
- **Disadvantages**
  - No feedback mechanism to TCP, need to manually adjust TCP window size
  - Compression modules required for each network endpoint/path

### SAN-Based Compression
- MDS9000 with IPS or MPS Module
- **Advantages**
  - Compression feedback for TCP window adjustments
  - Compression on multiple FCIP tunnels through different network paths
- **Disadvantages**
  - Compression of storage data only in a shared network
Protecting Data through Encryption

Encryption in a SAN Extension Network - Secure Data for Business and Regulatory Requirements

- 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 ~ 10µs per packet
- Standards-based IPSec Encryption - implements RFC 2402 to 2410, & 2412
  IKE for protocol/algorithim negotiation and key generation
  Encryption: AES (128 or 256 bit key), DES (56 bit), 3DES (168 bit)
IP vs. SAN Network-Based Encryption

Network-Based Encryption
- 7200 with VAM, VAM2, VAM2+
- Cisco Catalyst® 6500 with VPNSM
- Advantages
  Encryption for some or all IP traffic (as required) in shared network
- Disadvantages
  Data not encrypted over part of IP network (SAN to switch or router)
  Encryption modules required for each network endpoint/path

SAN-Based Encryption
- MDS9000 with MPS Module
- Advantages
  Encryption on multiple FCIP tunnels through different network paths
- Disadvantages
  Encryption of storage data only in a shared network
Write Acceleration

- Enables extended distance capabilities for remote replication technologies
- Better performance using FC/FCIP-WA – Up to 2X the performance over given distance
  - Reduces effective I/O latency within SAN extension solutions
- Built into Services modules (IPS, MPS, SSM) - transparent to disk arrays
- Highly resilient solution – no data stored in MDS 9000 switch

### FCIP Write Acceleration (WA)

- **WRITE** line for data transfer
- **XFER_RDY** line for status
- **STATUS** line for error handling

**Write Acceleration Ratio (at various link speeds and write sizes)**

- **RTT (ms)**: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
- **Ratio**:
  - 1Gbps
  - 622Mbps
  - 155Mbps
  - 45Mbps

**Legend:**
- 32kB 45M
- 32kB 155M
- 32kB 622M
- 32kB 1G
Write Acceleration Design Considerations
SANOS 2.0

High Availability and Load Balancing:

- Can be used for native FC replication (SSM) or FCIP replication (IPS / MPS)
- Portchannels may be used for HA
- 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
SAN Extension Fabric Stability

- Connecting existing SAN fabrics or extending a SAN fabrics creates SAN design challenges
  
  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 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)

Host to Local Array Fabric is VSAN_5
Site A Replication Fabric is VSAN_10
Site B Replication Fabric is VSAN_30
SAN Extension Fabric is VSAN_20
MDS 9000 Fabric Switch Positioning

Cisco positioned to extend reach all market segments

Industry-Leading Investment Protection Across a Comprehensive Product Line

Cisco MDS 9000 Family SAN-OS
Agenda

- Basic SAN Extension Principles
  - Dual path, HA options

- Alternate Designs
  - Hub-spoke, multi-hop

- SAN Extension Choices and Enhancements
  - Optical: Extended B2B credits, B2B credit spoofing, Port-tracking
  - FCIP: Compression, Encryption, Application Acceleration

- MDS 9000 Product Family
Q and A