



# FICON Over IP – Technology and Customer Use

Tony Almeida  
Cisco Systems  
[talmeida@cisco.com](mailto:talmeida@cisco.com)

Thursday, August 11, 2011  
Session 9853

## Abstract

- Many customers need to send FICON traffic between two datacenter sites - or between their datacenter and their tape archive. In the past this meant dedicated SONET circuits but today FICON over IP allows customers to send this storage data over IP networks that are much more cost effective. This IP network can be a dedicated circuit or even be a virtual tunnel within collapsed corporate networks. First in this session we will discuss the technology to do this and how it insures the RAS required for storage and then in the second half of the session, we will discuss a customer network where this was implemented and is running in production today.

# Agenda











- What is FCIP
  - Where/Why is it mostly used
- Basic FCIP – What do I need to worry about
- Advanced FCIP
- Customer Examples
  - Customer 1
  - Customer 2

# SAN Extension Technology Options

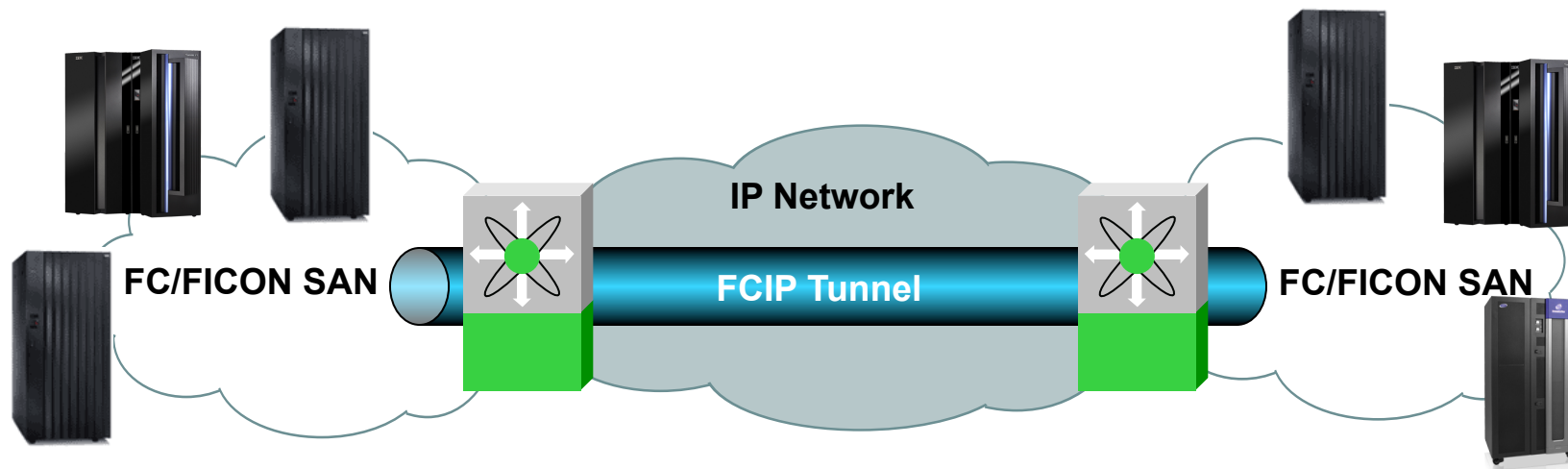
Increasing Distance 

Data

Center Campus Metro Regional National Global

		Increasing Distance 	
		Data	
		Center Campus Metro Regional National Global	
Optical	Dark Fiber	Sync (Any Speed) 	Limited by Optics (Power Budget)
	CWDM	Sync (1,2,4 Gbps) 	Limited by Optics (Power Budget)
	DWDM	Sync (1,2,4,10 Gbps per $\lambda$ ) 	Limited by BB_Credits
	SONET/SDH	Sync (1,2 Gbps + Subrate) 	Async 
IP	MDS9000 FCIP	Sync (Metro Eth) 	Async (WAN, 1 Gbps) 

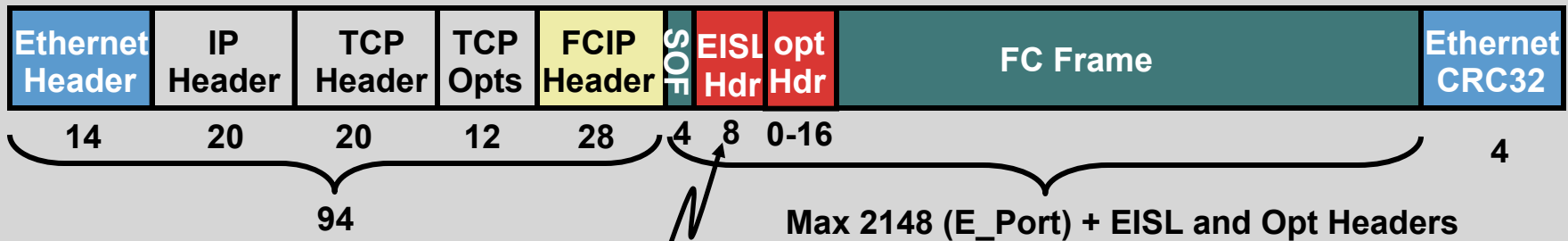
# FCIP: Fibre Channel over IP



## 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



**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

- 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

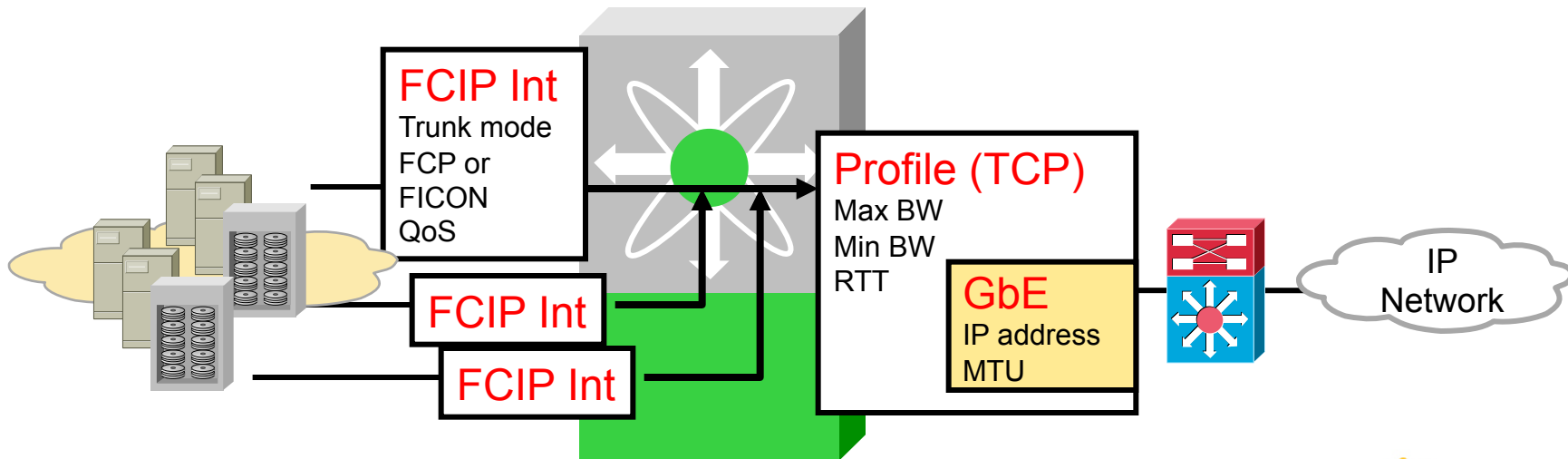
# Why Use FCIP?

- 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

# FCIP Configuration Overview

To create a basic FCIP configuration, follow these steps on each peer switch:

1. Configure the Gigabit Ethernet interface
2. Configure static IP routes
3. Enable the FCIP feature
4. Create an FCIP profile
5. Create an FCIP interface (up to 4 per profile)





# Basic FCIP Configuration Example

## FCIP configuration on MDS A

```
interface GigabitEthernet2/1
 ip address 10.1.21.21 255.255.255.0
 no shutdown

ip route 10.2.21.0 255.255.255.0
 10.1.21.254 interface gig2/1

fcip enable
fcip profile 1
 ip address 10.1.21.21

interface fcip1
 use-profile 1
 peer-info ipaddr 10.2.21.11
 no shutdown
```

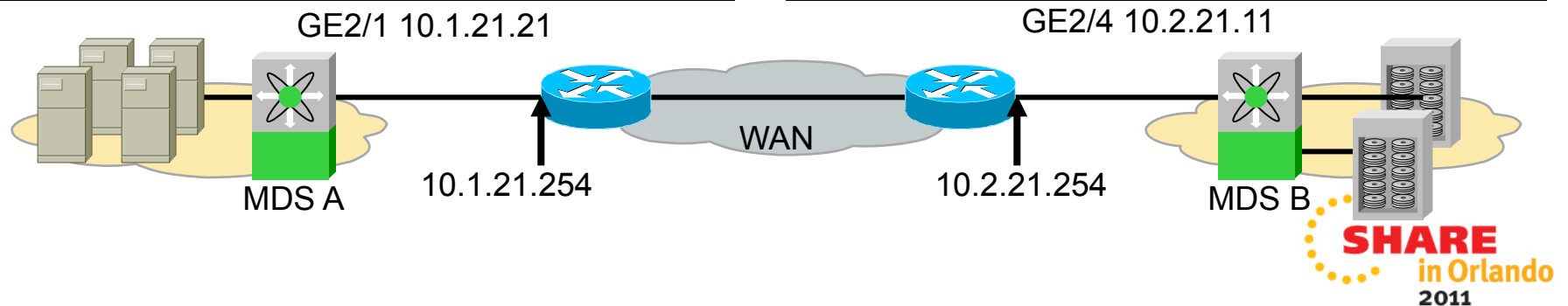
## FCIP configuration on MDS B

```
interface GigabitEthernet2/4
 ip address 10.2.21.11 255.255.255.0
 no shutdown

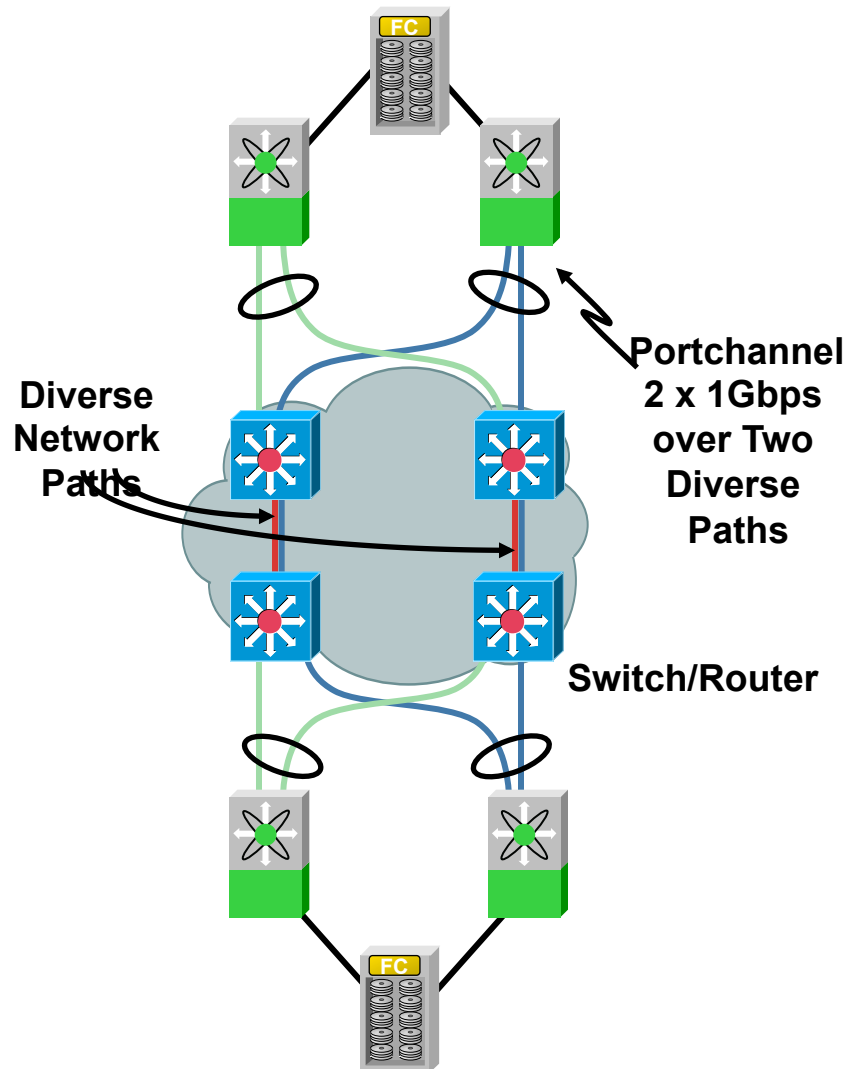
ip route 10.1.21.0 255.255.255.0
 10.2.21.254 interface gig2/4

fcip enable
fcip profile 1
 ip address 10.2.21.11

interface fcip1
 use-profile 1
 peer-info ipaddr 10.1.21.21
 no shutdown
```



# MDS FCIP SAN Extension Design



- Same port channeling and VSAN trunking rules apply as with FC links
- Port channel individual FCIP links to alternate Ethernet switches/routers
  - Each WAN link carries two FCIP tunnels

# 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”

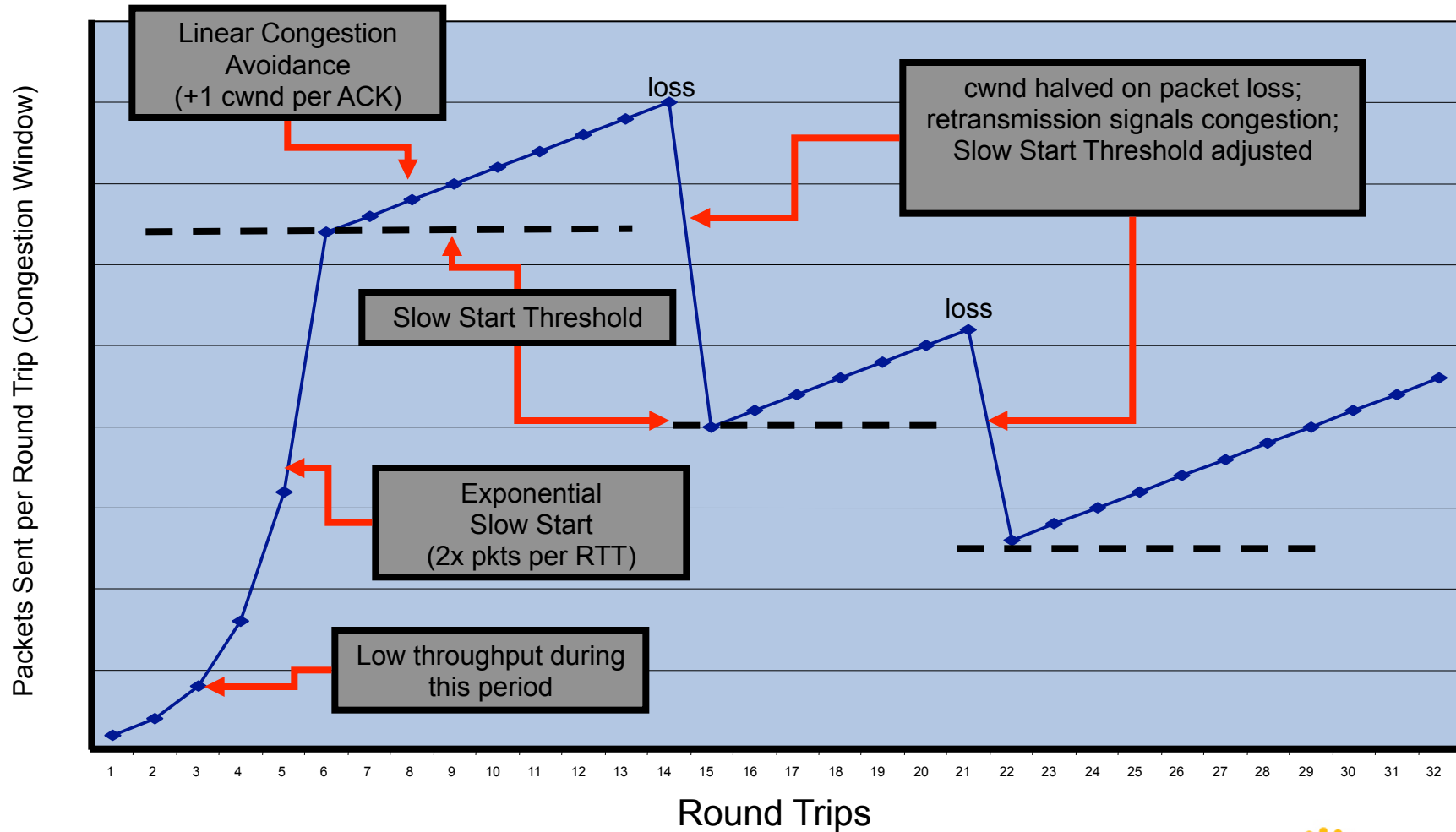
## Differences with Normal TCP:

- 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

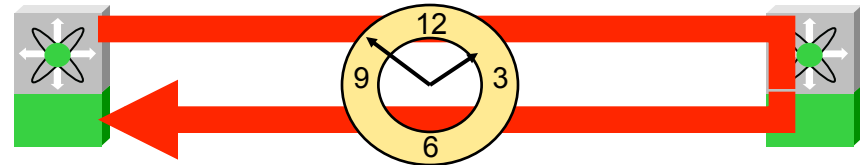
# The Impact of TCP Congestion Avoidance

## Simplified TCP Congestion Avoidance

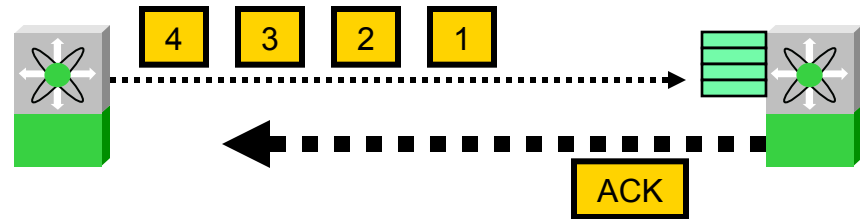


# FCIP Flow Control Design Factors

Round Trip Time:

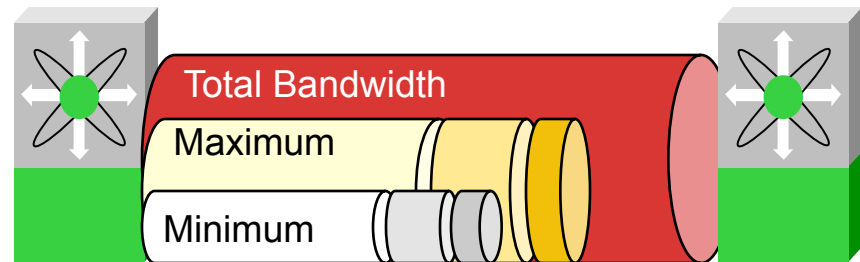


Maximum Window Size:



Packet Shaping:

- Maximum bandwidth
- Minimum available bandwidth



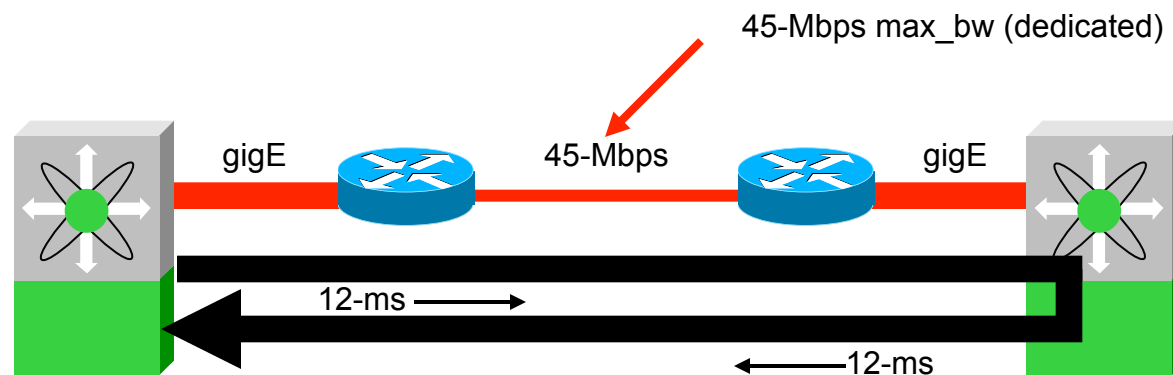
Congestion window monitoring



# Round Trip Time

Configuring the **round-trip-time** parameter:

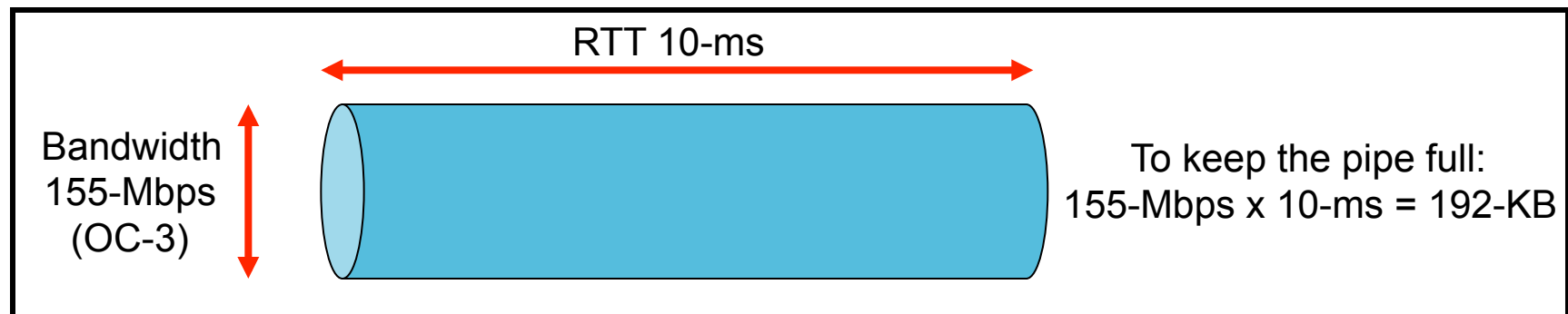
- Not necessarily symmetric
- Use **ping** and **measure-rtt** commands to calculate
- Automatically calculated in Cisco MDS



End-to-end latency \* 2 = 24 ms RTT.

# TCP Maximum Window Size

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



$MWS = \text{Maximum bandwidth} \times \text{RTT}$

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

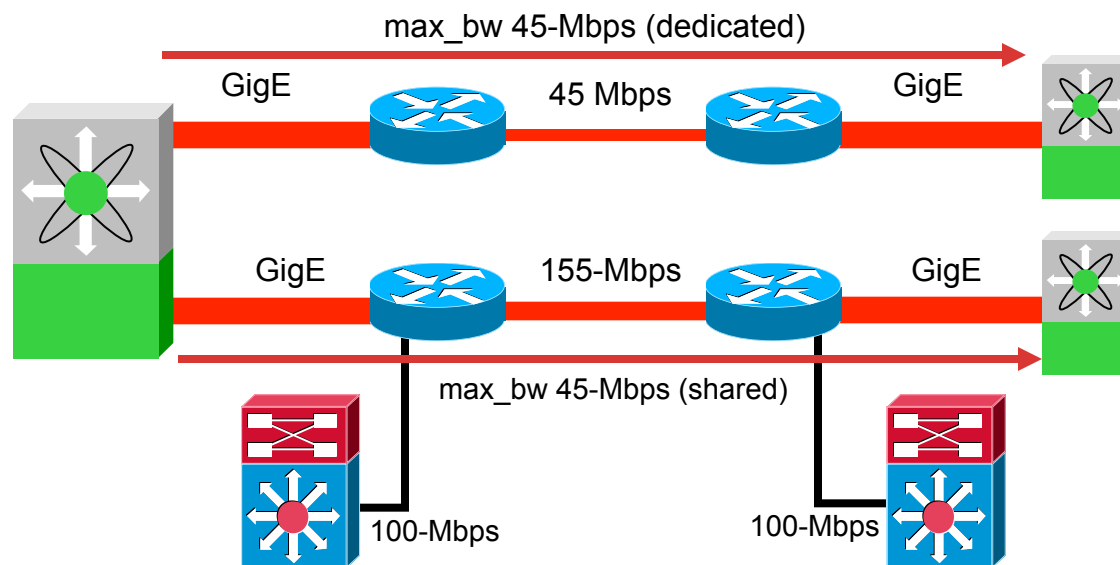
- Under dimensioning will throttle throughput
- Over dimensioning can cause congestion
- Use the bandwidth of the lowest speed link



# TCP Maximum Bandwidth

Configure the TCP max-bandwidth value as follows:

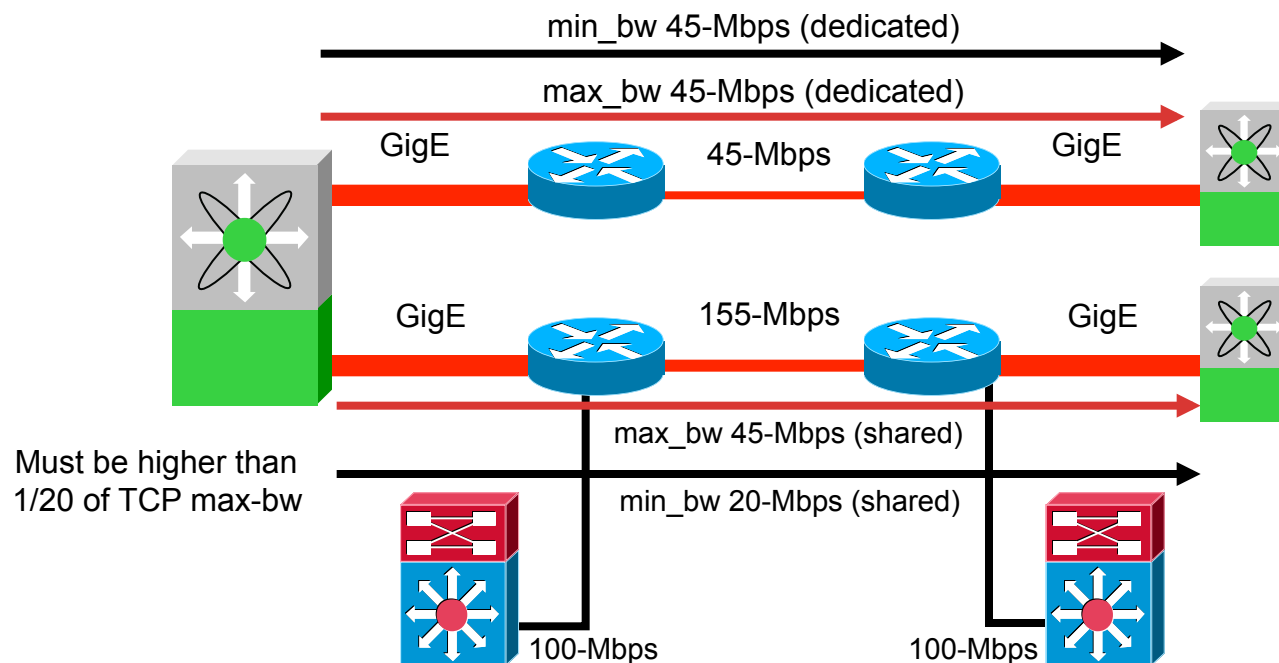
- No larger than smallest pipe in the path
- If sharing the pipe, configure to be highest amount available to FCIP
- In a shared environment, configure QoS in the entire path



# TCP Minimum Available Bandwidth

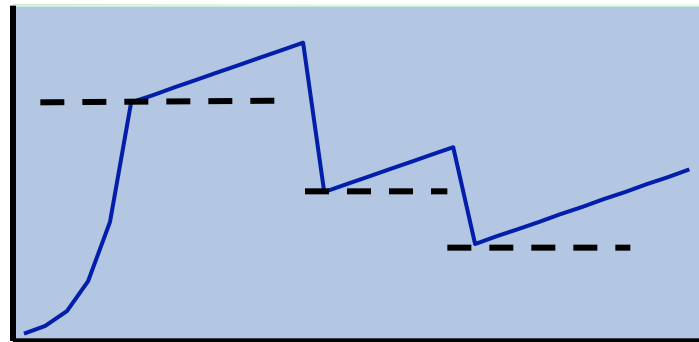
Configure TCP min-available-bandwidth value as follows:

- 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

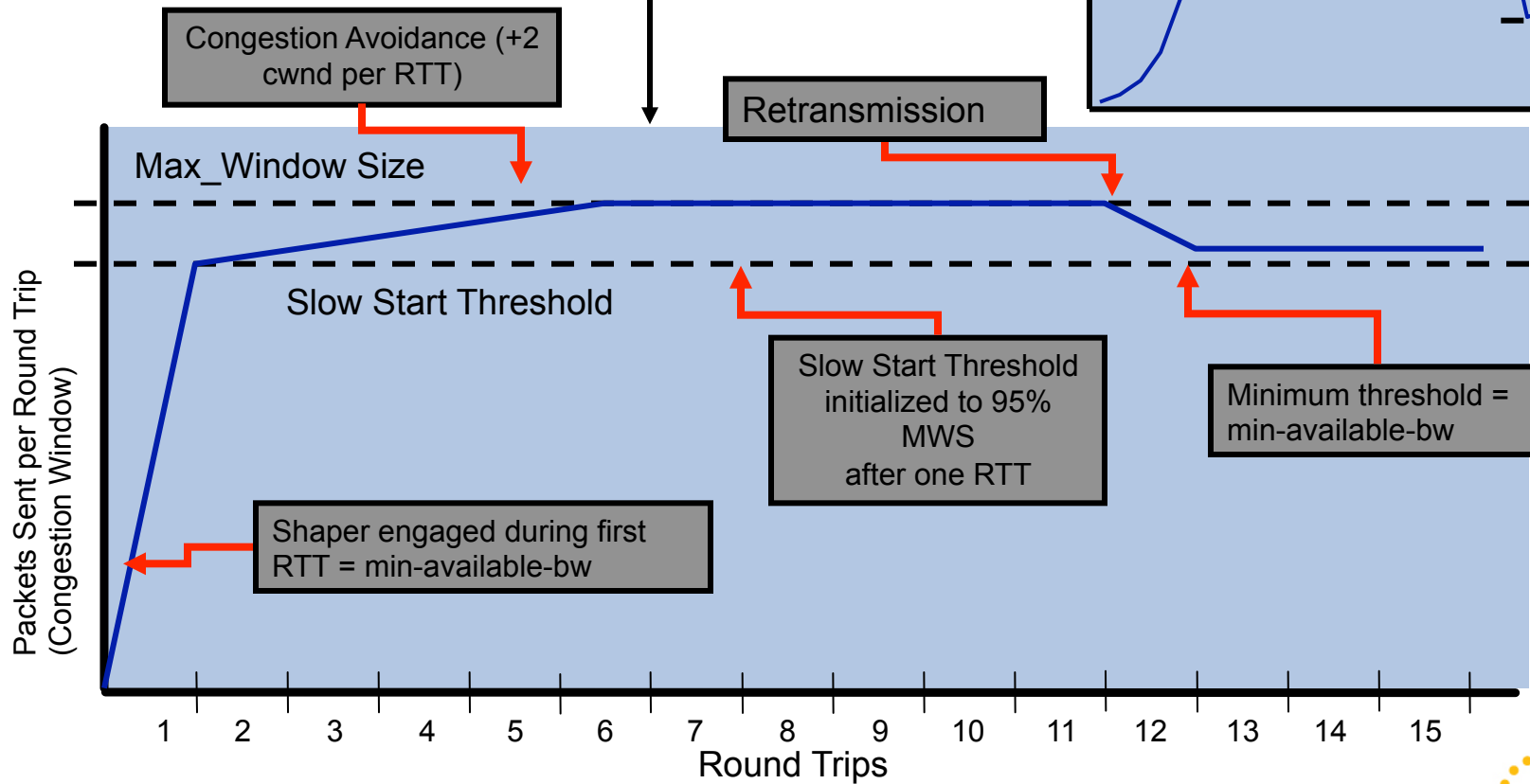


# Results of Packet Shaping

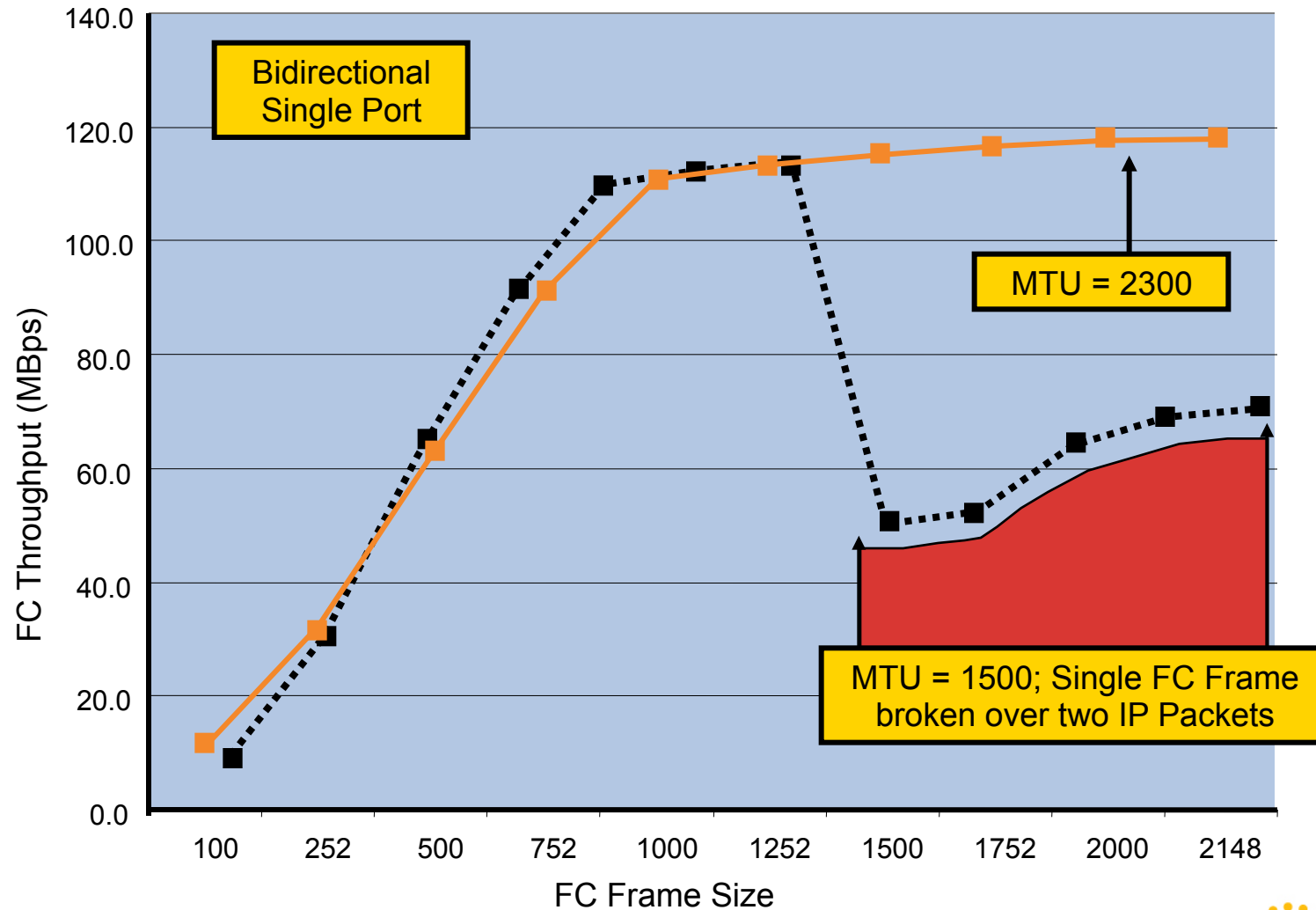
Traditional TCP Congestion Avoidance



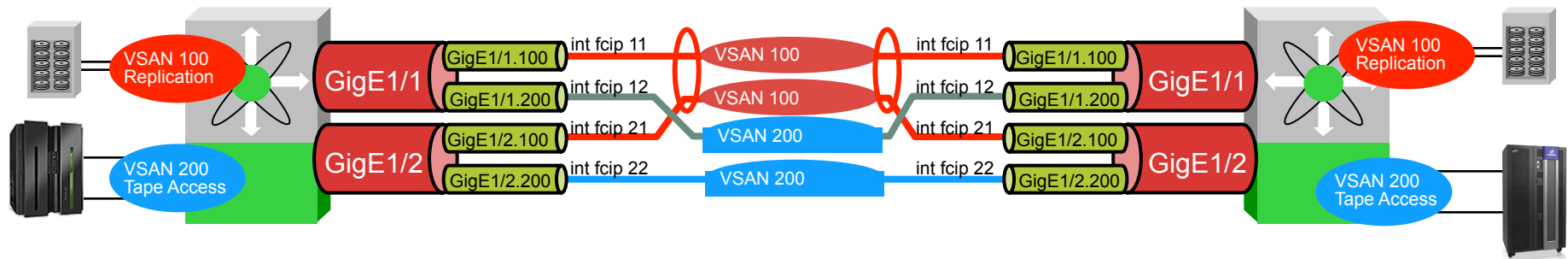
## Packet Shaping



# MTU Size: Performance Comparison



# FCIP – Multiple FCIP Tunnels



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 channelled

Note – When multiple FCIP tunnels are on the same interface, they use a different TCPIP port numbers

# 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 standard 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 can assign CS6 or DSCP 48

# FCIP QoS Markings

Customer networks can have several types of business-critical traffic, including voice over IP (VoIP), video, FCIP, business applications, etc...

Traffic is normally classified as it enters the network, where it is marked for appropriate treatment.

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

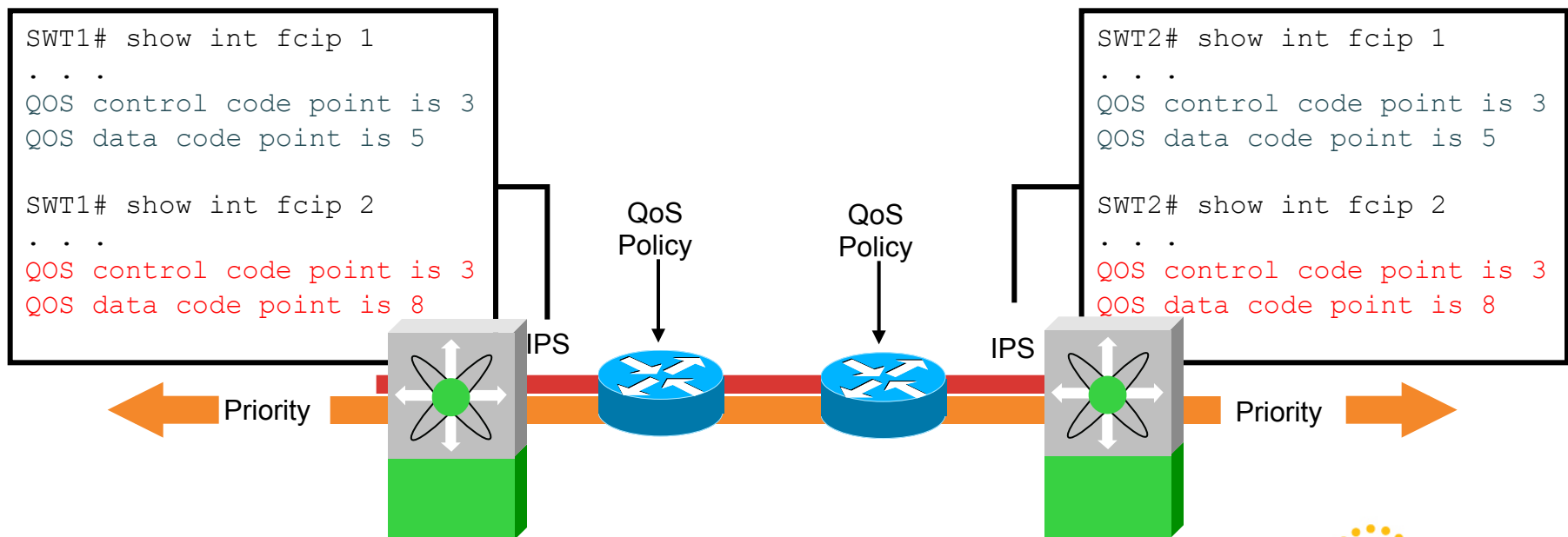
High → (points to top of table)  
 ← Med (points to right of table)  
 Low → (points to bottom of table)



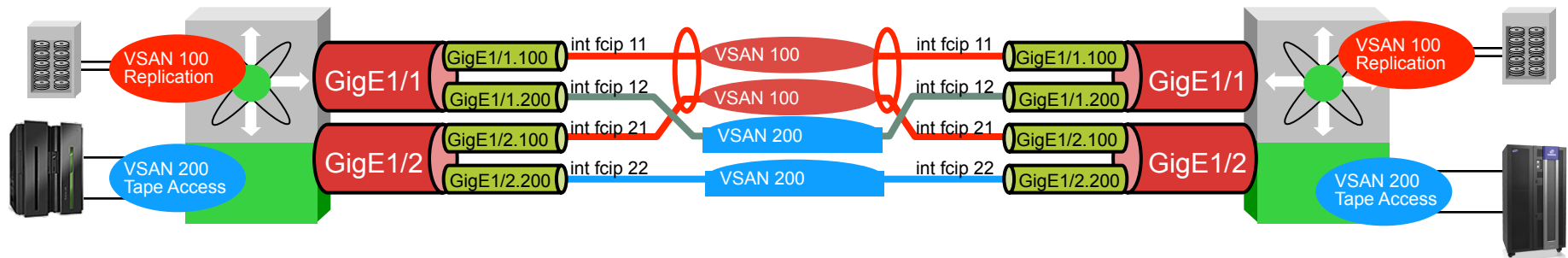
# IP Quality Of Service

The Cisco MDS 9000 can tag control and data traffic for every FCIP link with a DSCP value between 0 and 63:

- QoS-aware WANs can then recognize and treat the tagged traffic according to enterprise QoS policies
- FCIP traffic tagged at a higher DiffServ priority is treated more favorably if congestion occurs on the WAN



# FCIP – Multiple FCIP Tunnels



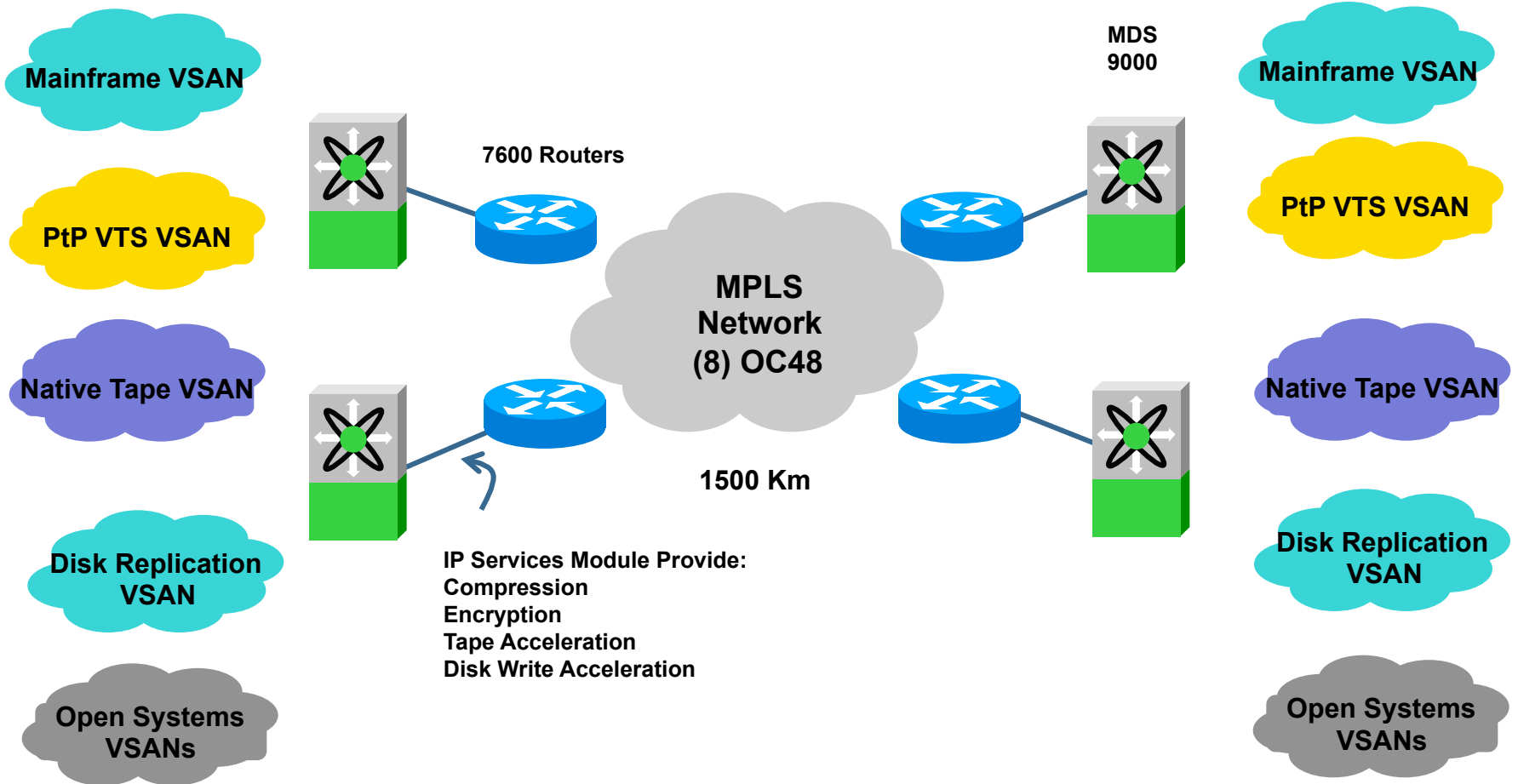
Now, Configure QOS based on business priorities of data

- 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.

# Large Provider of Business Outsourcing Services - FCIP



# FCIP Data Compression

- Cisco uses RFC standard compression algorithms implemented in both hardware and software
- MDS 9000 18/4-port Multiservice Module
  - Third Generation IP Services Module
  - 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

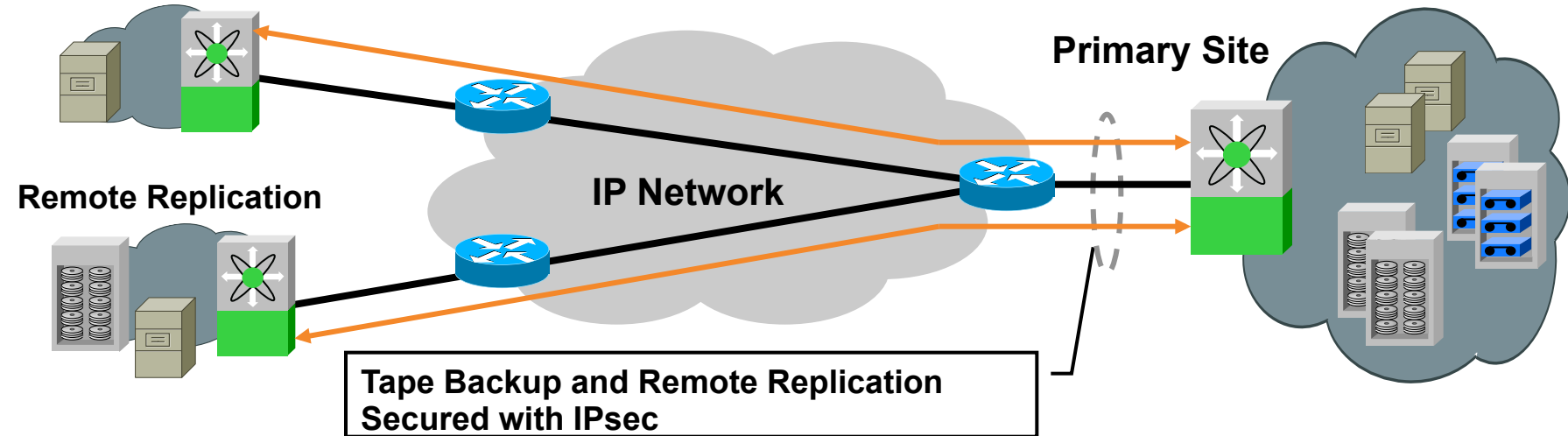
# IPSec Encryption for FCIP

## 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

## Remote Tape Backup



- Hardware-based GigE wire-rate performance with latency ~ 10  $\mu$ s 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)

# 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
- Tape control, label processing, etc are not accelerated
- Has been shipping for multiple years and numerous customers in production

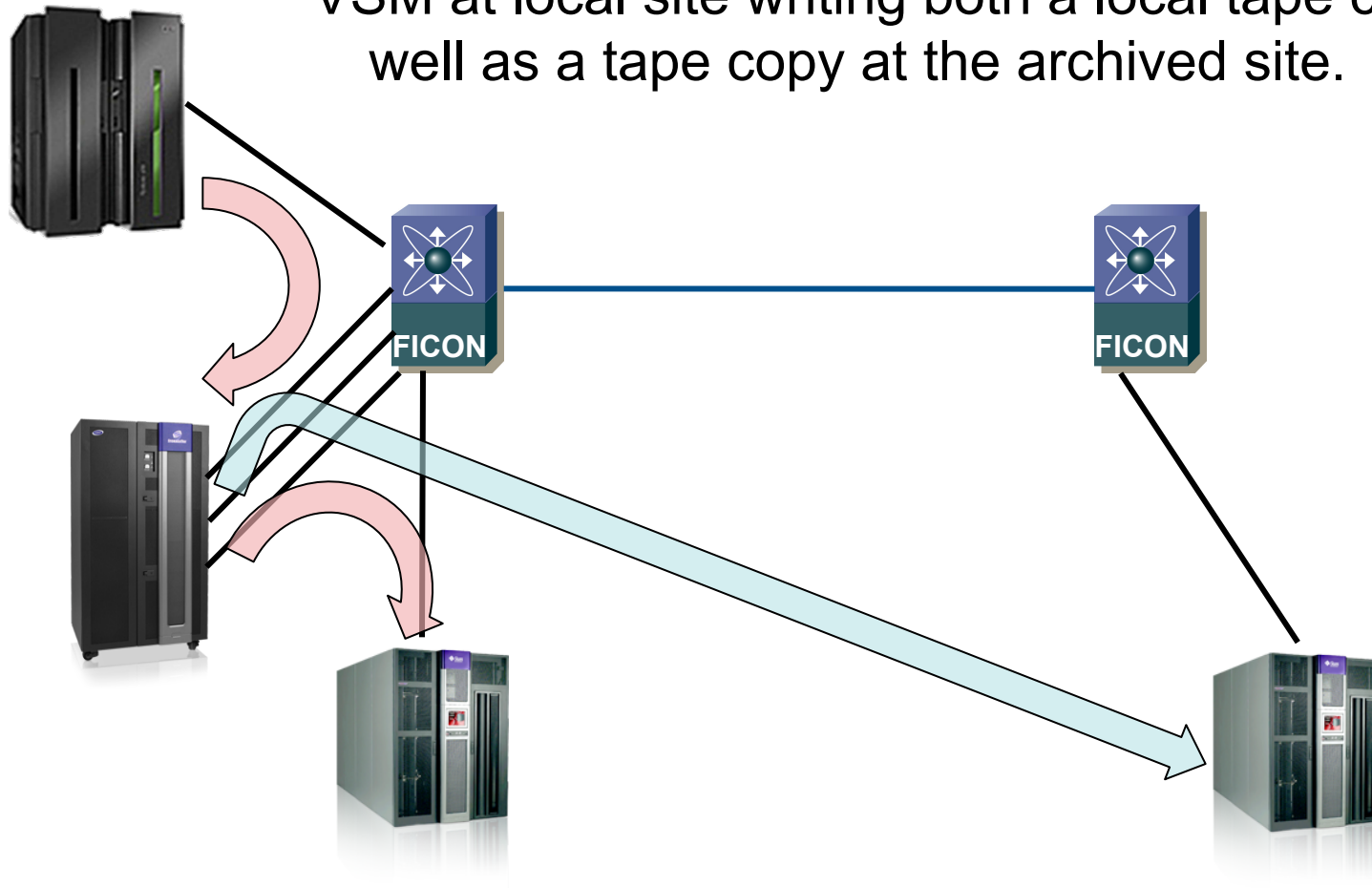
# FICON Tape Read Acceleration

- Accelerates Reads by
  - Flowing off the host until data ready
  - 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

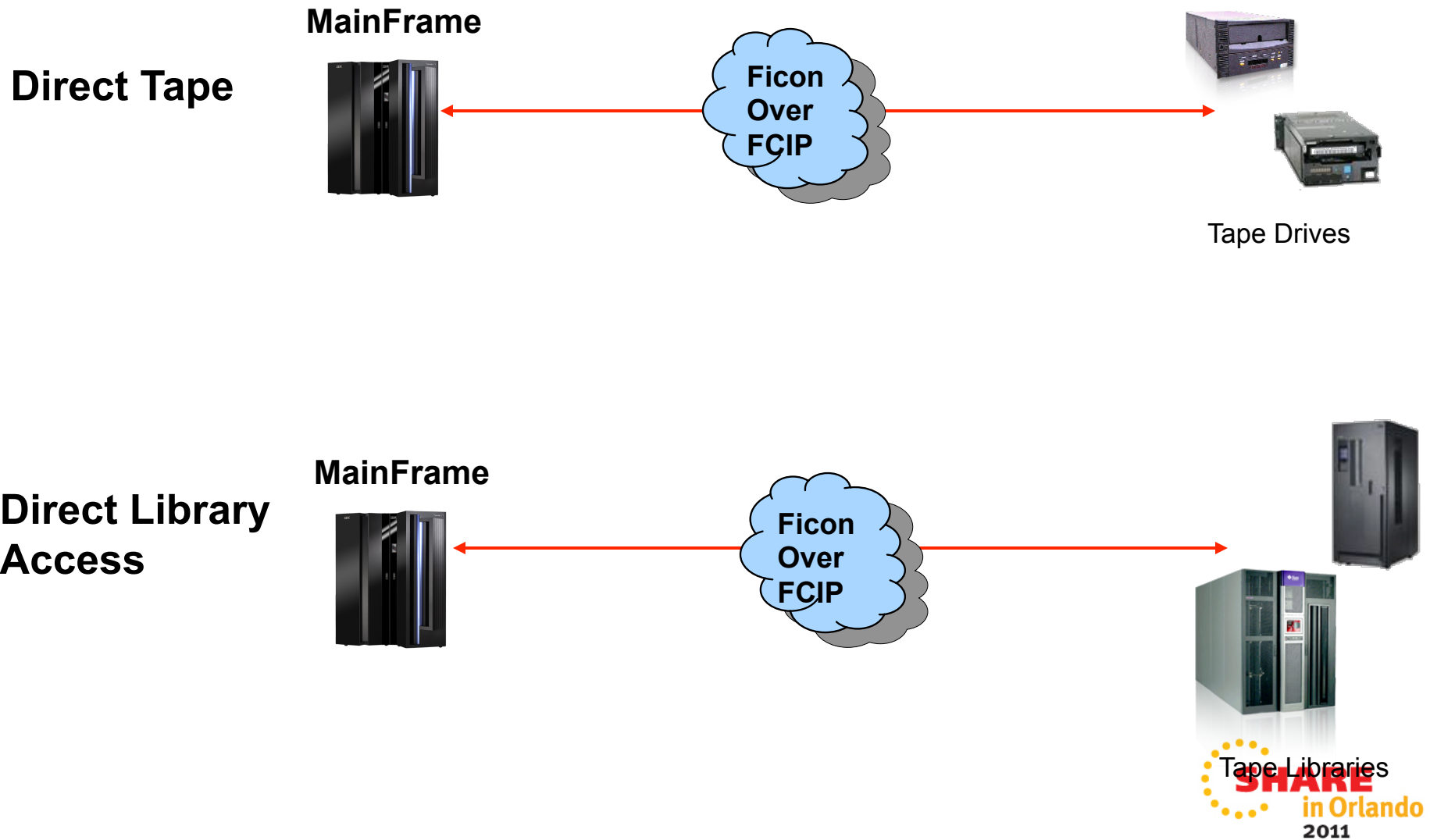


# Sun VSM to RTD Extension

VSM at local site writing both a local tape copy as well as a tape copy at the archived site.



# FICON Tape Acceleration



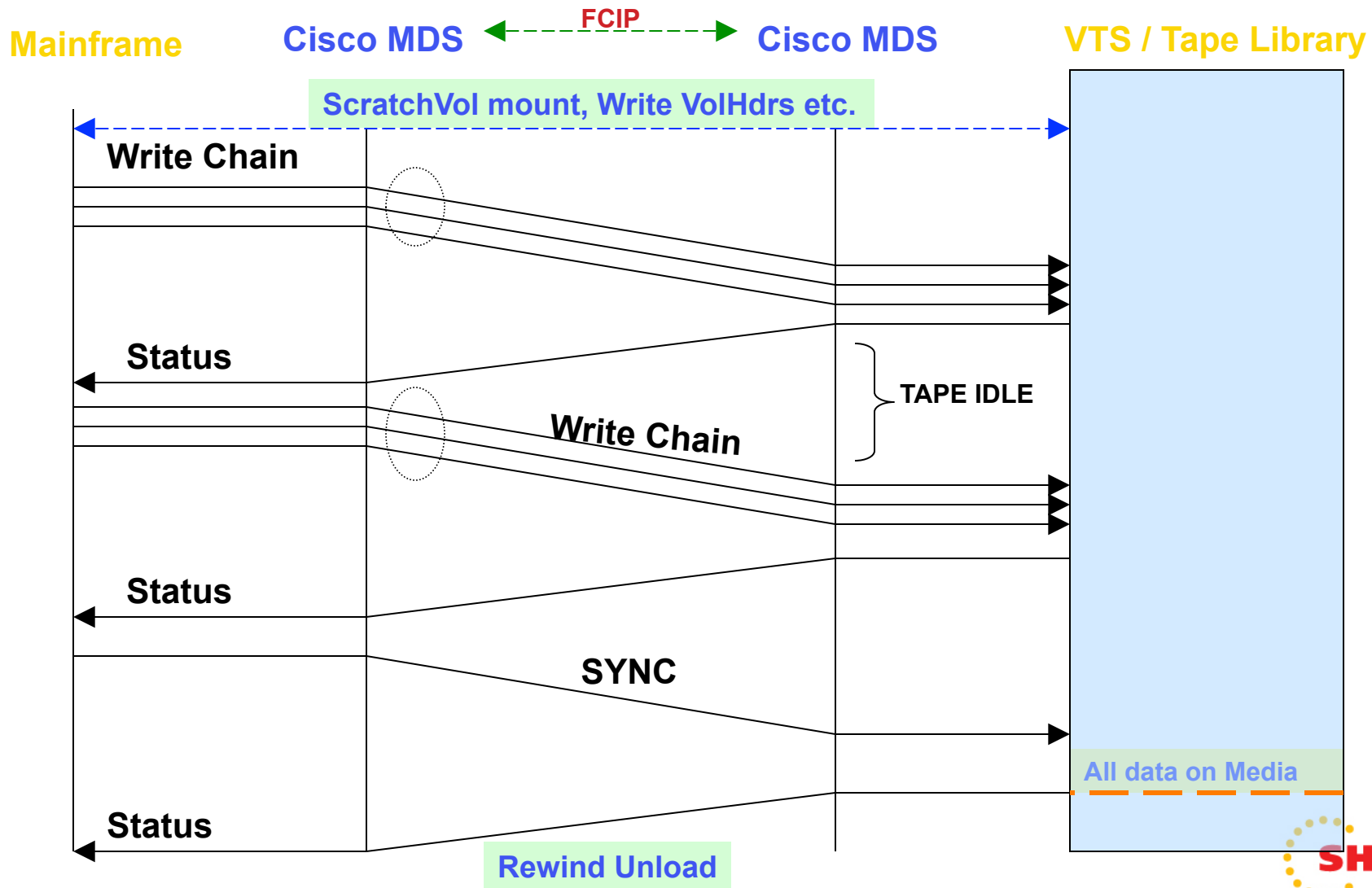
# FTA Configuration Information

- 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

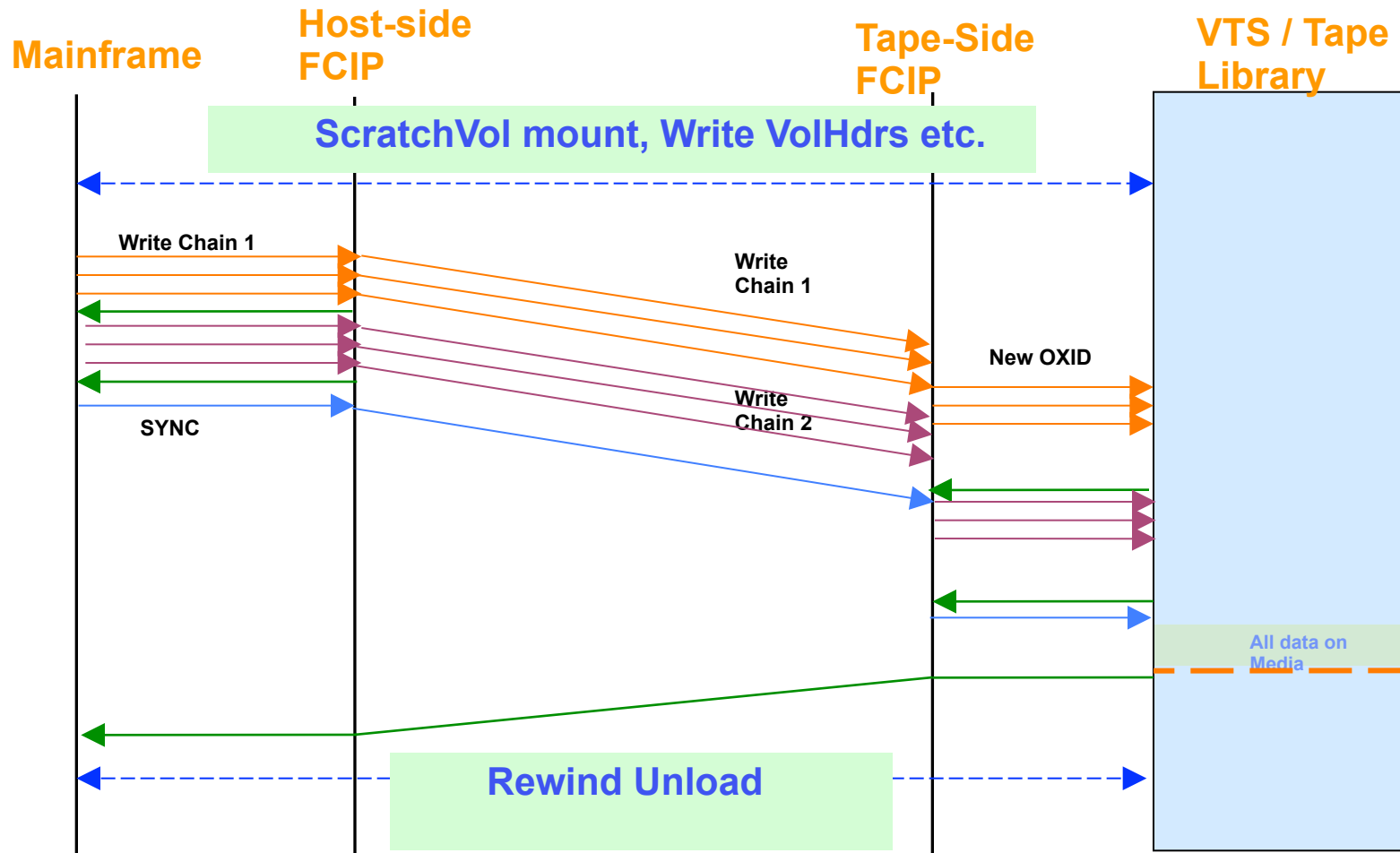
## FTA – More Details

- 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.

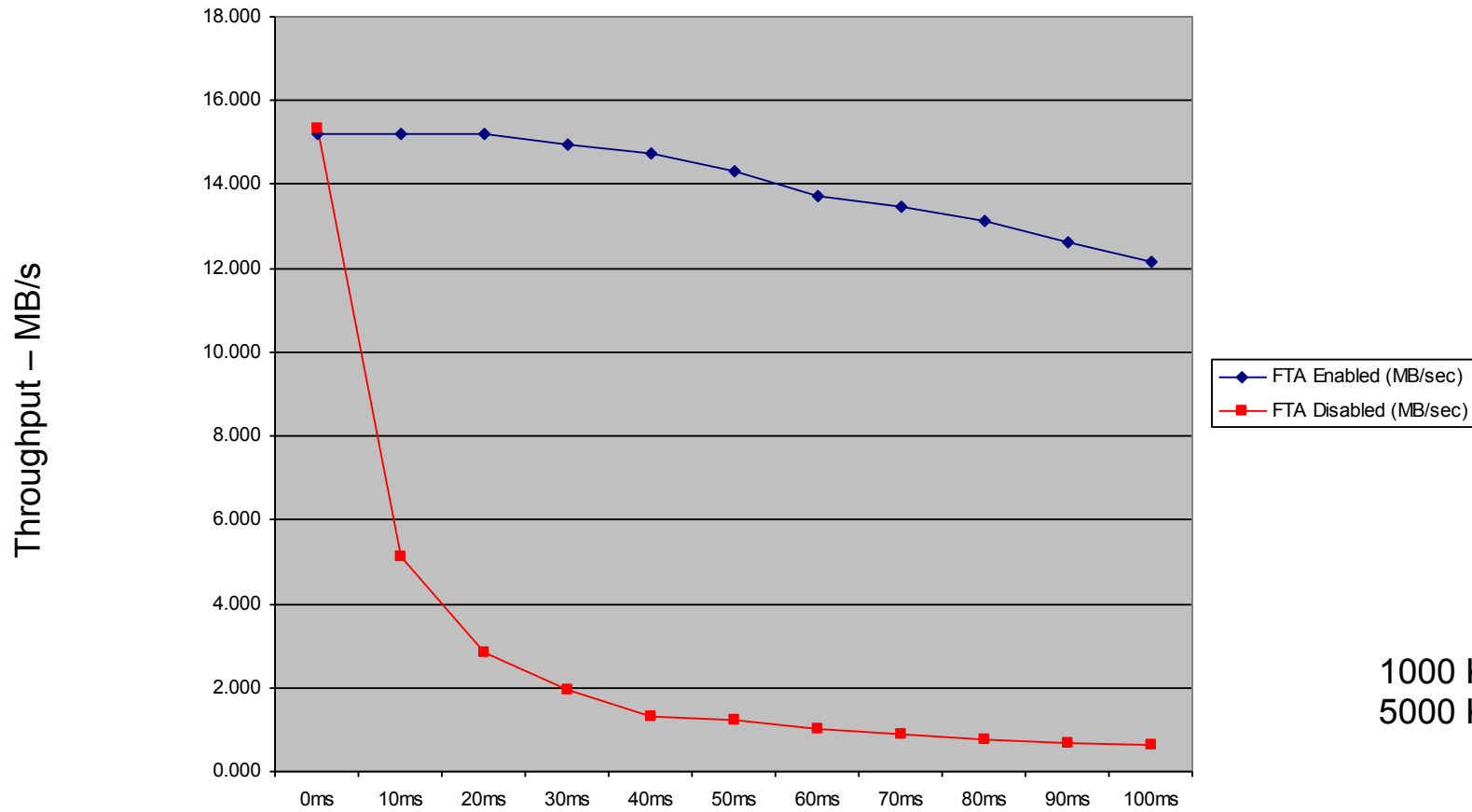
# Backup protocol without acceleration ...



# Backup protocol with acceleration ...



# Results: Throughput

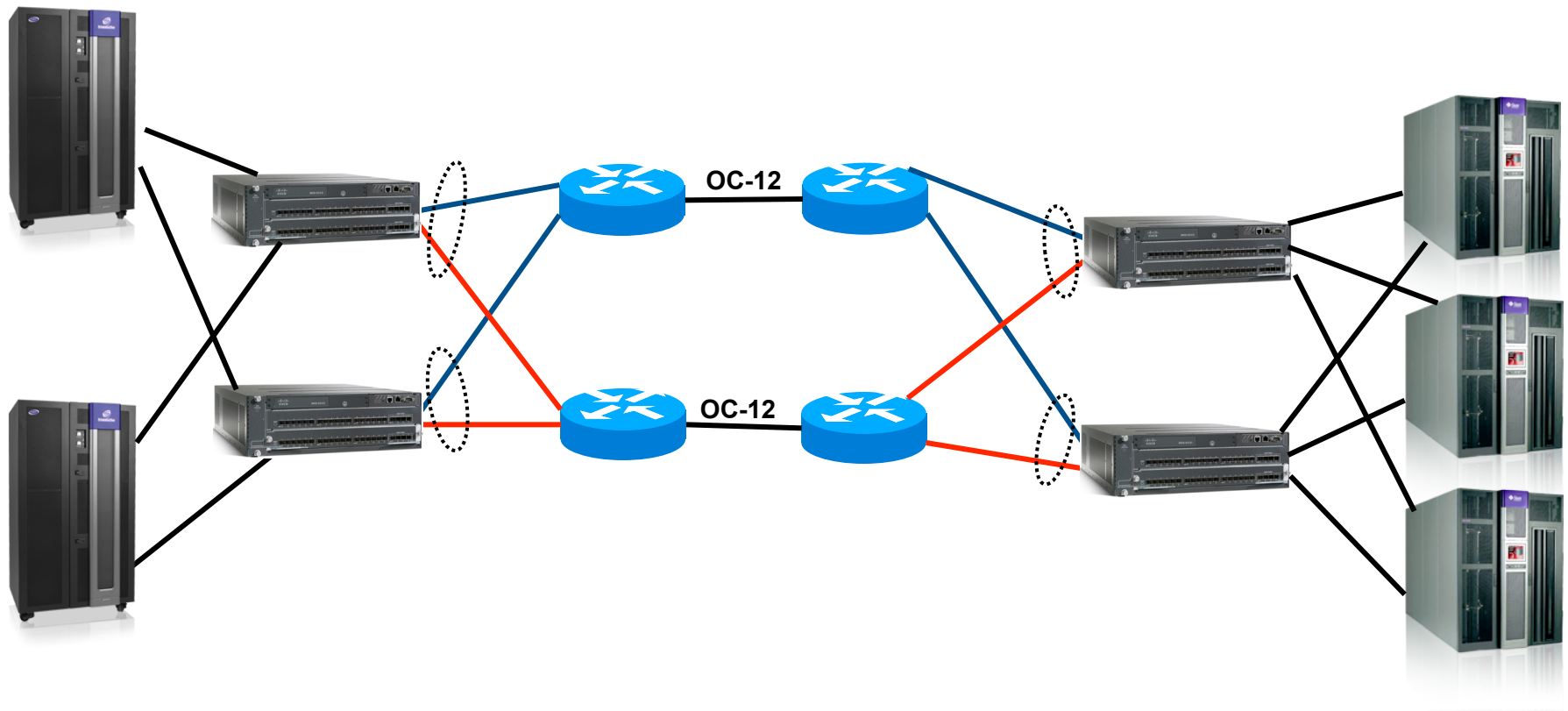


1000 Km – 10ms  
5000 Km – 50ms

Site-Site Delay

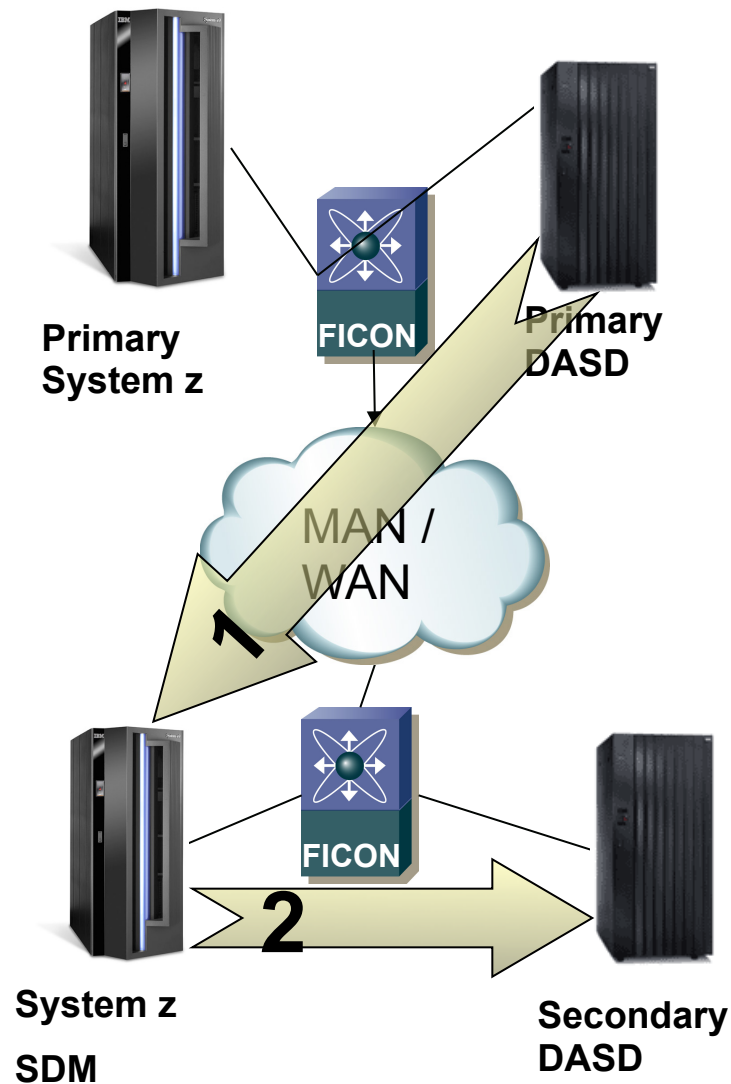


# VSM - RTD Customer Example - EMEA





# 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 “**S**ystem **D**ata **M**over” (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
- Until now, only one solution for Channel extension
  - Cisco is now supporting XRC!
    - Will support up to 20,000 Km with XRC Acceleration feature

# XRC Acceleration – How It Works

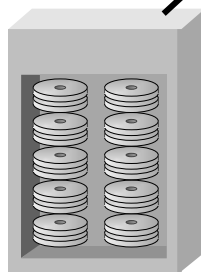
- Acceleration of RRS channel programs

- Read Record Sets (ie. Updates)

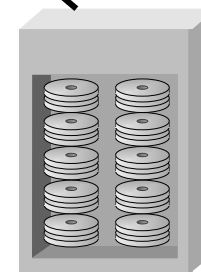
- In DSO (first command in RRS set) we know how to pre-read for the whole chain of data

- We pre-read the data and send across .. thereby filling the pipe.

- This works around the IU limitations in the FICON architecture.



PRIMARY DASD



SECONDARY DASD

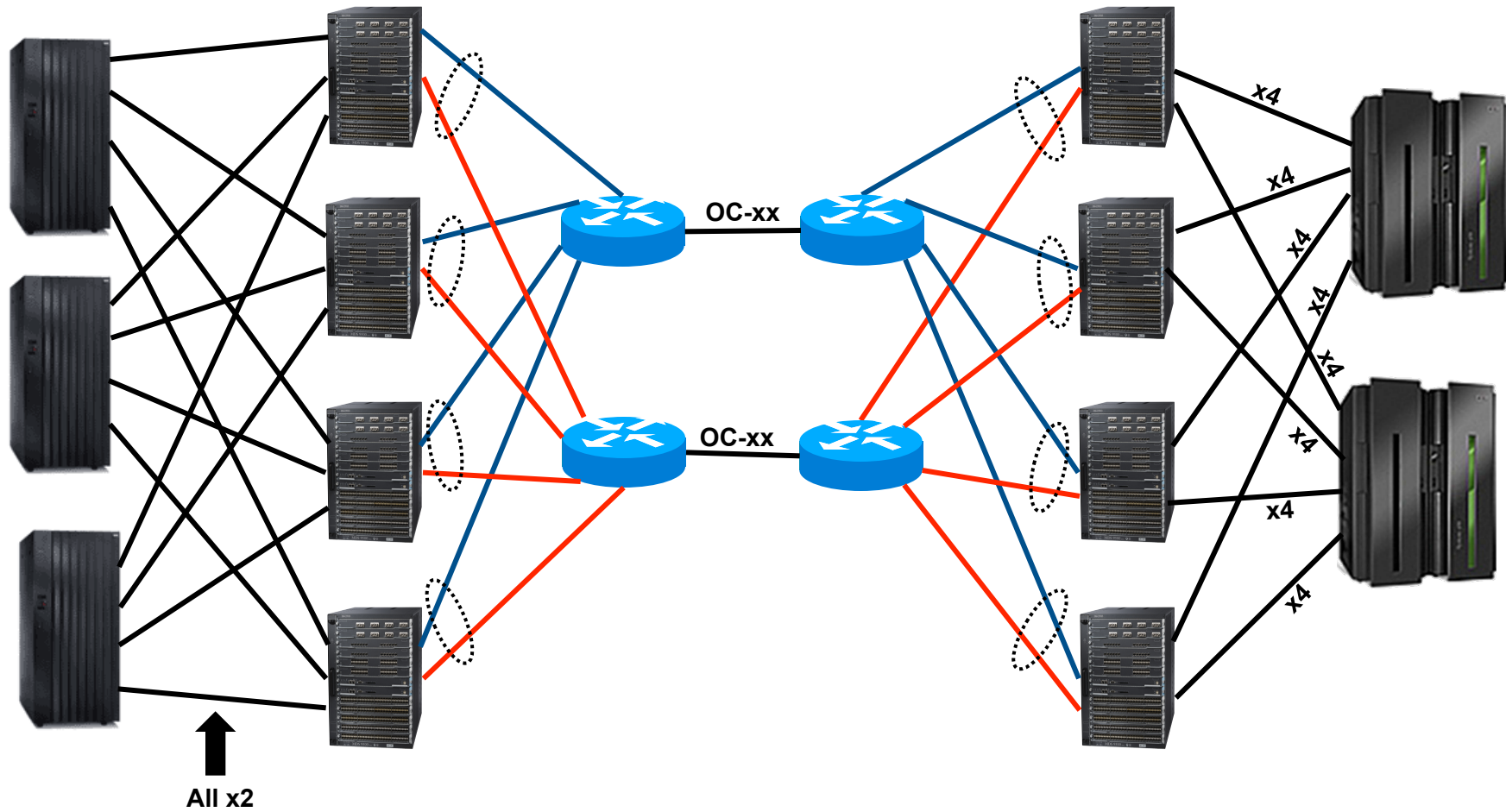
## How Fast is it?

- Some performance testing results:
  - Vs. no Acceleration:
    - Almost 5x faster at 1600 km
    - Almost 9x faster at 3200 km
  - vs. IBM' s new Extended Distance FICON feature (**z10 and above ONLY**)
    - Comparable up to 3200 km; gets better over longer distance

## 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 – Hi Redundancy



## FICON – A Few last comments

- FTA Read and Write supported at the same time
  - Seems obvious but wanted to make sure
- FTA and XRCA are mutually exclusive on an FCIP link
  - Can be on the same physical link but not same virtual interface
- Remember FTA only supports 1 FCIP link between sites
  - Per FICON VSAN
- Any Questions ?!

## Session 9853





# FICON Device Support Matrix (4.2.7b)

## Without Acceleration

Traffic Type	Transport	Supported	Max Distance
XRC	FC ISL (CWDM optics)	Yes	100 km / 40 km
XRC	FC ISL (DWDM optics)	Yes	200-300 km
XRC	FCIP ISL	Yes	200-300 km
FICON	FC ISL (CWDM optics)	Yes	100 km
FICON	FC ISL (DWDM optics)	Yes	300 km
FICON	FCIP ISL	Yes	300 km
Tape - Host to Oracle*** VSM 4/5	FC ISL (CWDM optics)	Yes	100 km / 40 km
Tape - Host to Oracle VSM 4/5	FC ISL (DWDM optics)	Yes	300 km
Tape - Host to Oracle VSM 4/5	FCIP ISL	Yes	300 km
Tape - Oracle VSM 4/5 to RTD	FC ISL (CWDM optics)	Yes	100 km / 40 km
Tape - Oracle VSM 4/5 to RTD	FC ISL (DWDM optics)	Yes	300 km
Tape - Oracle VSM 4/5 to RTD	FCIP ISL	Yes	300 km
Tape - Host to Oracle Real Tape *	FC ISL (CWDM optics)	Yes	100 km / 40 km
Tape - Host to Oracle Real Tape *	FC ISL (DWDM optics)	Yes	300 km
Tape - Host to Oracle Real Tape *	FCIP ISL	Yes	300 km
Tape - Host to IBM Real Tape **	FC ISL (CWDM optics)	Yes	100 km / 40 km
Tape - Host to IBM Real Tape **	FC ISL (DWDM optics)	Yes	300 km
Tape - Host to IBM Real Tape **	FCIP ISL	Yes	300 km
Tape - Host to IBM TS77x0	FC ISL (CWDM optics)	Yes	100 km / 40 km
Tape - Host to IBM TS77x0	FC ISL (DWDM optics)	Yes	300 km
Tape - Host to IBM TS77x0	FCIP ISL	Yes	300 km

\* - includes T9840C, T9840D, T9940B, T10K-A, T10K-B

\*\* - includes all FICON supported real tape models

\*\*\* - Oracle products are those formerly sold by Sun and StorageTek (STK)

# FICON Acceleration Distance Support (4.2.7b)



Traffic Type	Transport	Max Distance
XRC (z/OS Global Mirror)	FCIP w/ XRCA (XRC Acceleration) ***	20,000 km
Tape - Host to Oracle VSM 4/5	FCIP w/ FTA (FICON Tape Acceleration)*	5,000 km
Tape - Oracle VSM 4/5 to RTD	FCIP w/ FTA (FICON Tape Acceleration)*	5,000 km
Tape - Host to Oracle Real Tape	FCIP w/ FTA (FICON Tape Acceleration)*	5,000 km
Tape - Host to IBM Real Tape	FCIP w/ FTA (FICON Tape Acceleration)*	3,000 km **
Tape - Host to IBM TS77x0	FCIP w/ FTA (FICON Tape Acceleration)*	3,000 km **

\* - Includes both Read and Write acceleration on the MSM-18/4 card and MDS 9222i, Write only for 14+2. SSN-16 not supported (yet)

\*\* - Longer distances can be supported via RPQ

\*\*\* - Supported on the MSM-18/4 card, SSN-16 not supported (yet)