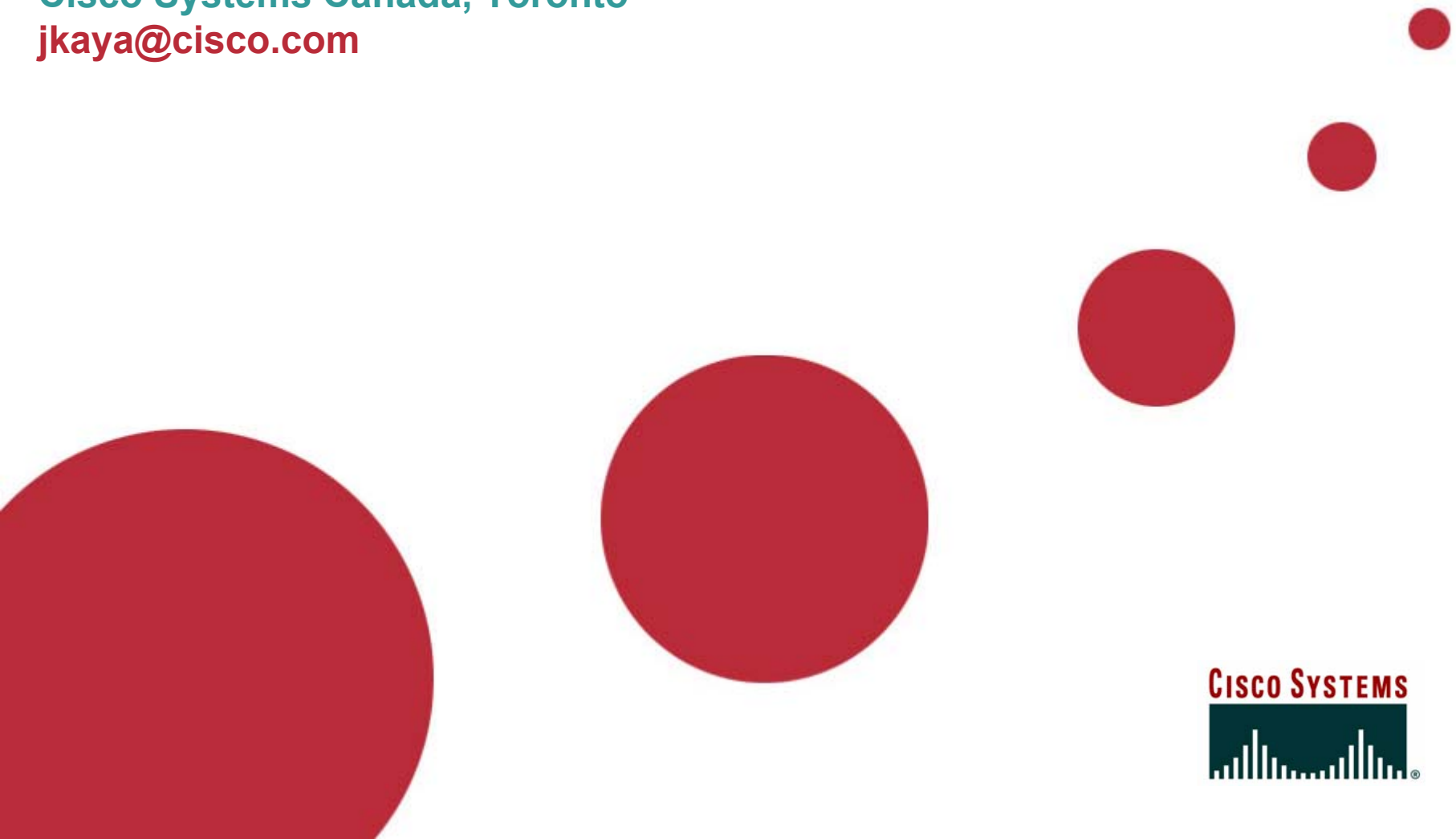
Technical Symposium

Storage Networking across the MAN and WAN

Joshua Kaya, Consulting Systems Engineer
Cisco Systems Canada, Toronto
jkaya@cisco.com



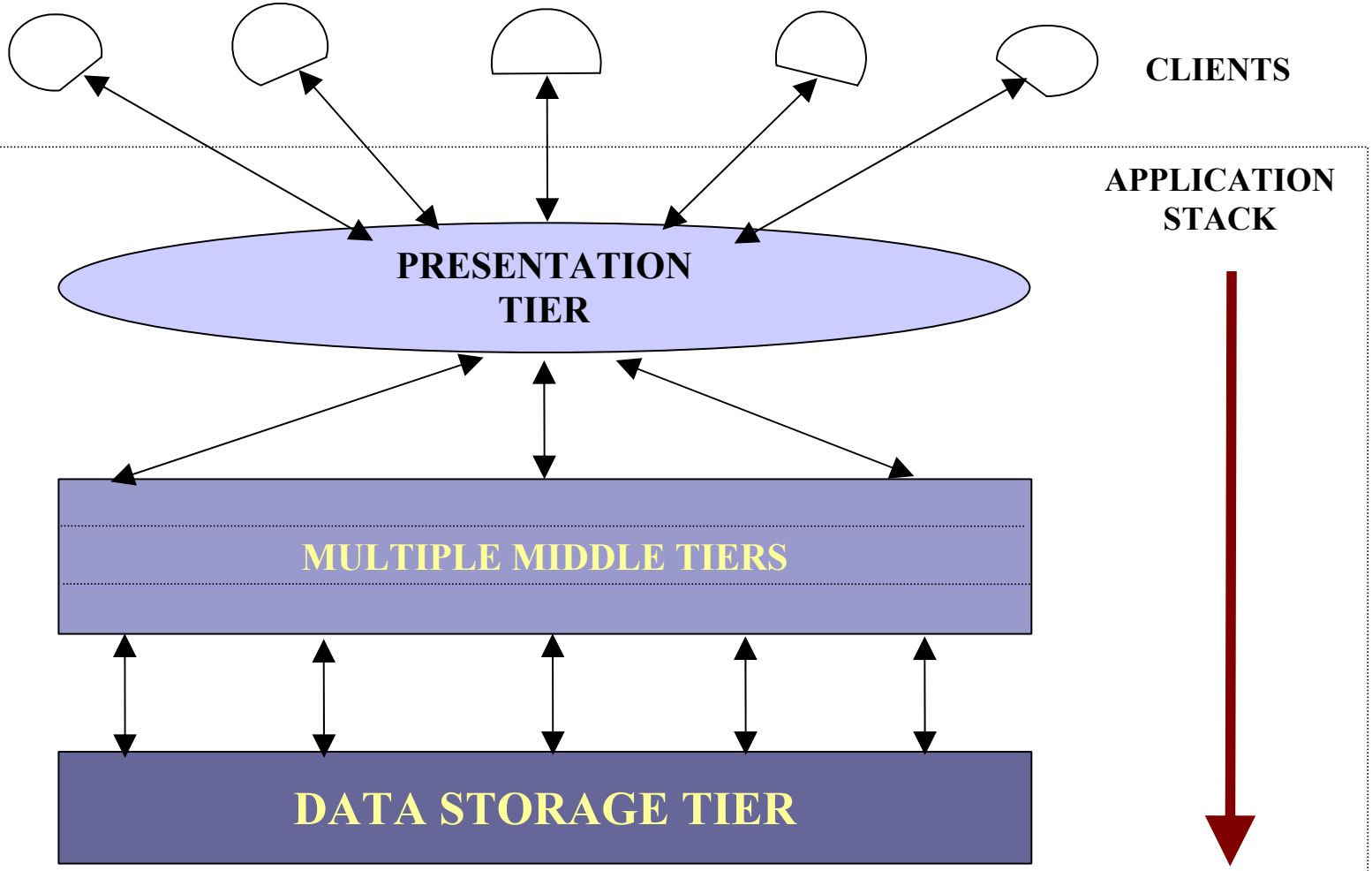
Cisco Systems

Agenda:

- Review: Replication
- Overview: Storage Extension
  - FCIP Storage Extension
  - Optical Storage Extension
- SAN Extension Case Study

Review: Replication

Application Environment / Stacks



Disaster Recovery

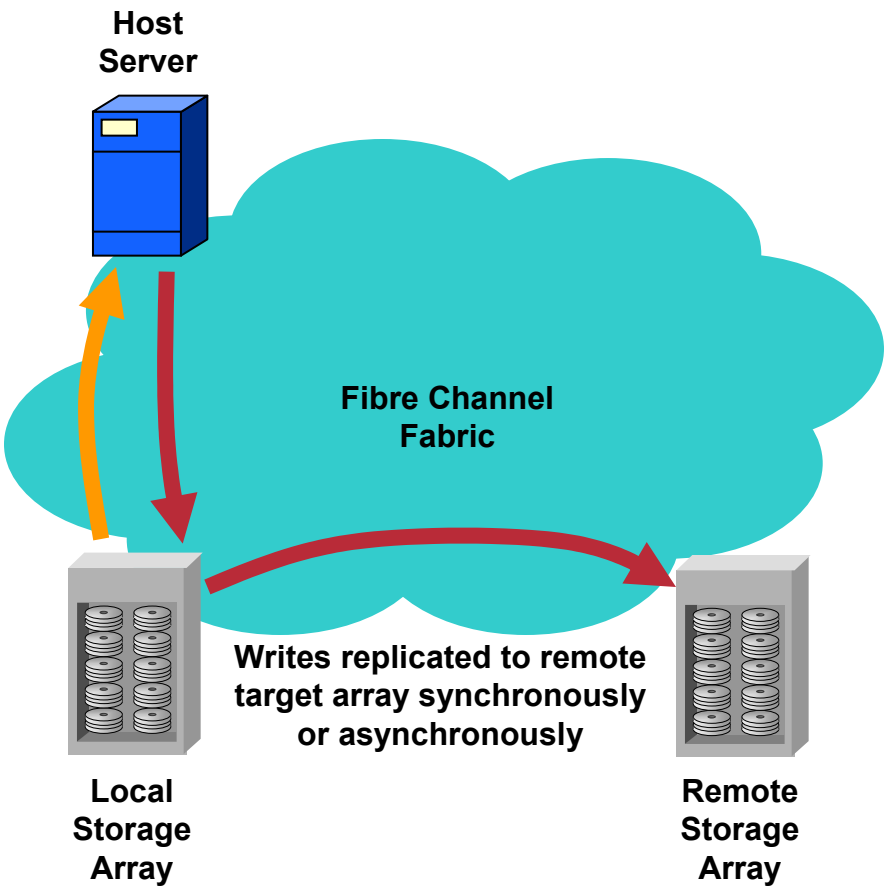
Application Stack

- **Presentation Tier**
  - **WEB GUI, Java GUI, Proprietary GUI**
- **Middle Tier**
  - **Message switching, transaction manager**
- **Data Storage Tier**
  - **Databases, mail stores**

Replication Overview

- Replication is the process of making copies of information across different devices
- The main goal is to provide for Disaster Recovery and High Availability
  - Hardware vs. Software Replication
  - Synchronous vs. Asynchronous
  - File Level vs. Block Level

Review: Replication



- Two arrays located on extended Fibre Channel Fabric
- Read only from local array
- Writes I/Os replicated to remote array
  - – replication managed by software in storage arrays
  - – Host server is unaware of replication
  - – Implementations are <u>proprietary</u>
    - SRDF, Truecopy, DRM,…

Replication Techniques

- There are three principal techniques used for data replication
  - Synchronous
  - Asynchronous
  - Snapshot (Used in Mirroring within the Disk Array)

Synchronous Replication

- **In Synchronous mode, data between primary and secondary are copied, validated and committed all at the same time.**
  - **Reliability is high**
  - **Applications might be in waiting while the secondary disk is “synching up”**

Asynchronous Replication

- In Asynchronous mode, data between primary and secondary are copied but the commits are done separately between primary and secondary systems
  - Over extended distances, application performance has less impacts than Synchronous mode
  - Potential data integrity issue if the secondary has not committed its write process during a failure at the primary host

Replication Implementations

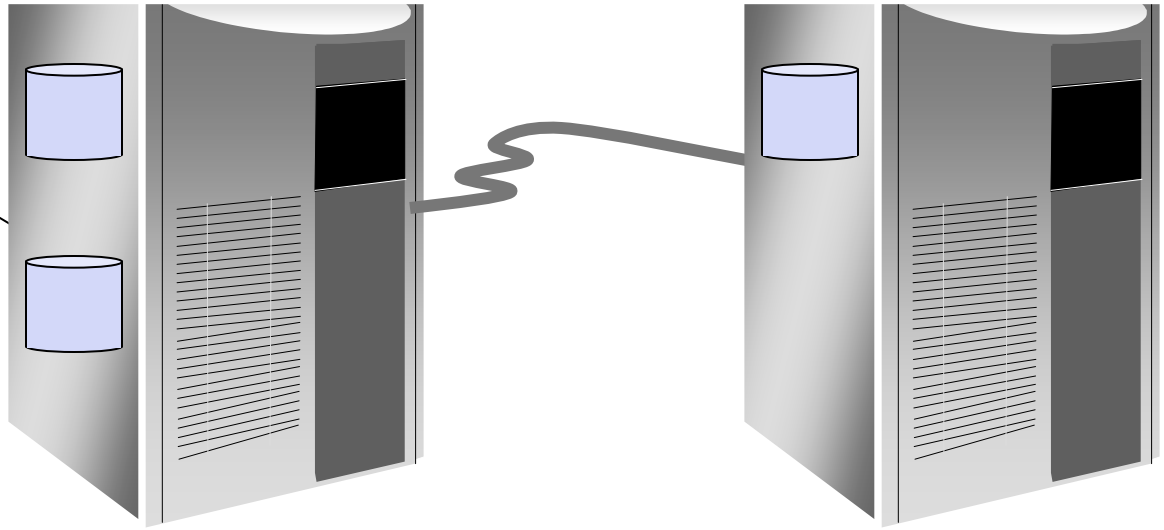
- **EMC Symmetrix Remote Data Facility (SRDF)**
  - Synchronous or “Semi-synchronous” replication across frames
- **Hitachi TrueCopy**
  - Hitachi Synchronous or Asynchronous replication within or across storage subsystems
- **IBM Peer-to-Peer Remote Copy (PPRC)**
  - Synchronous replication, Asynchronous with PPRC-XD
- **HP-Compaq Data Replication Manager (DRM)**
  - Synchronous or Asynchronous replication

FC Replication Distances

Host

Maximum Distances

- **With Switching approx. 100km (Single Mode with repeaters)**
- **With FCIP – No theoretical limit**

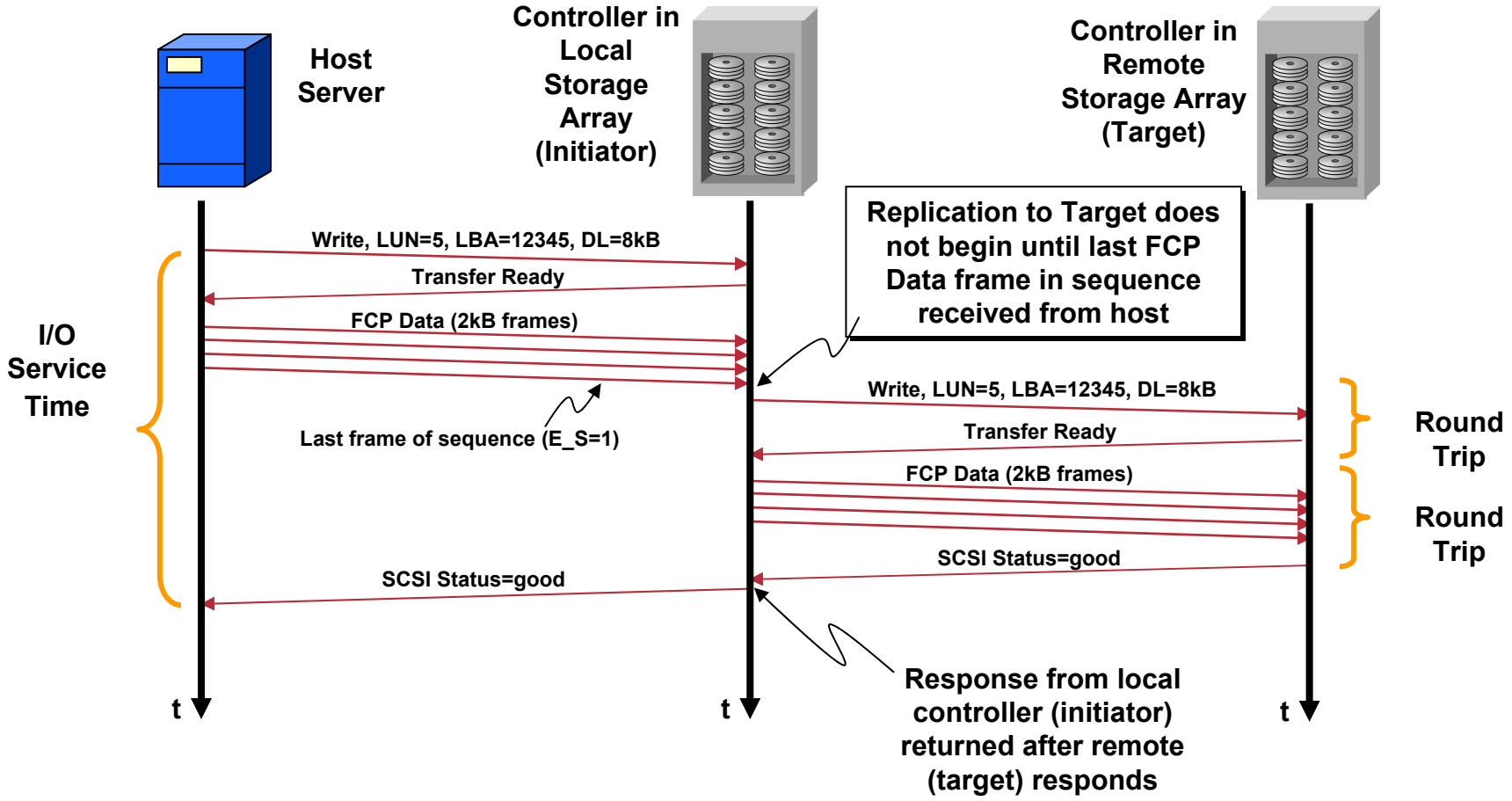


Storage Subsystem

Storage Subsystem

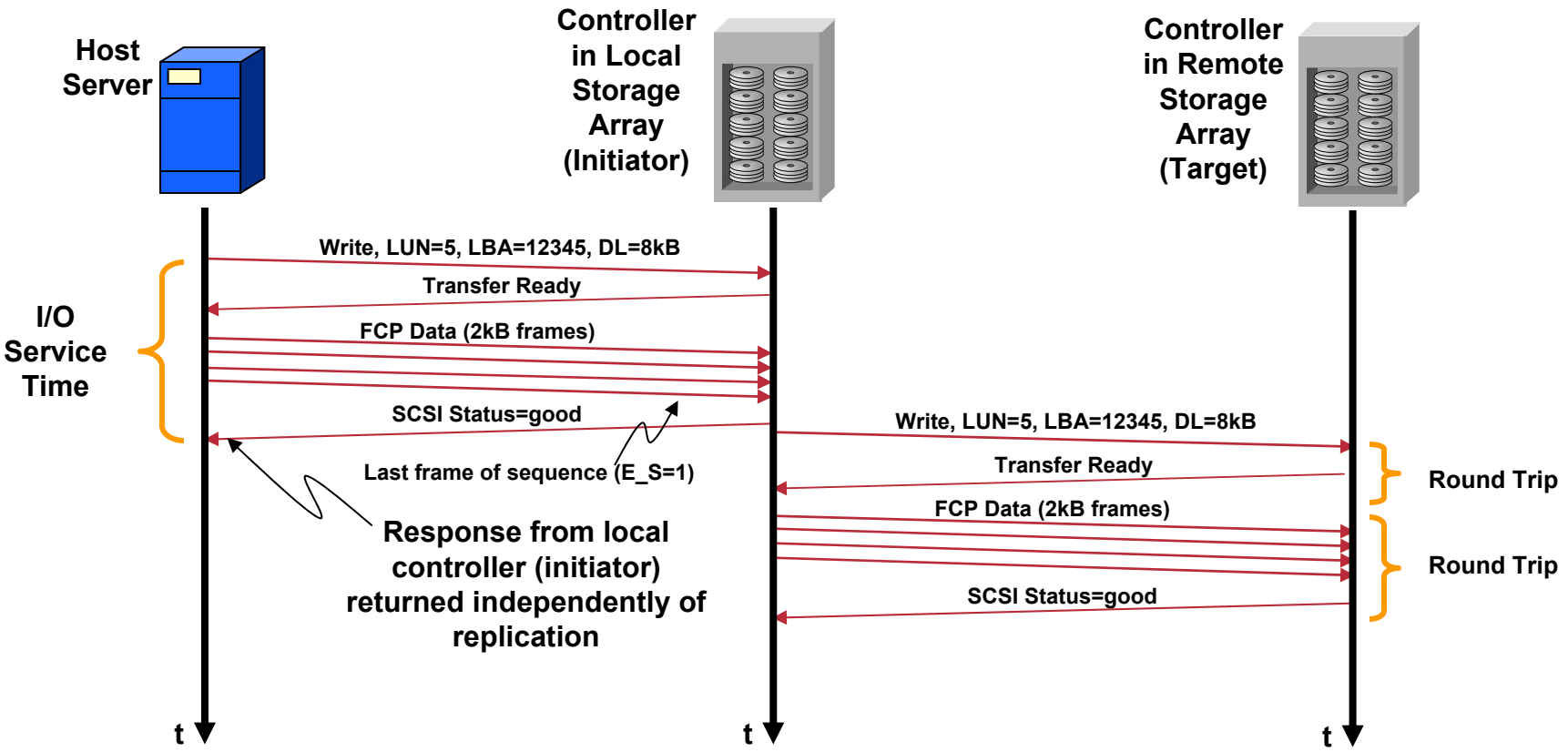
Synchronous Replication: I/O Detail

Example: HP DRM

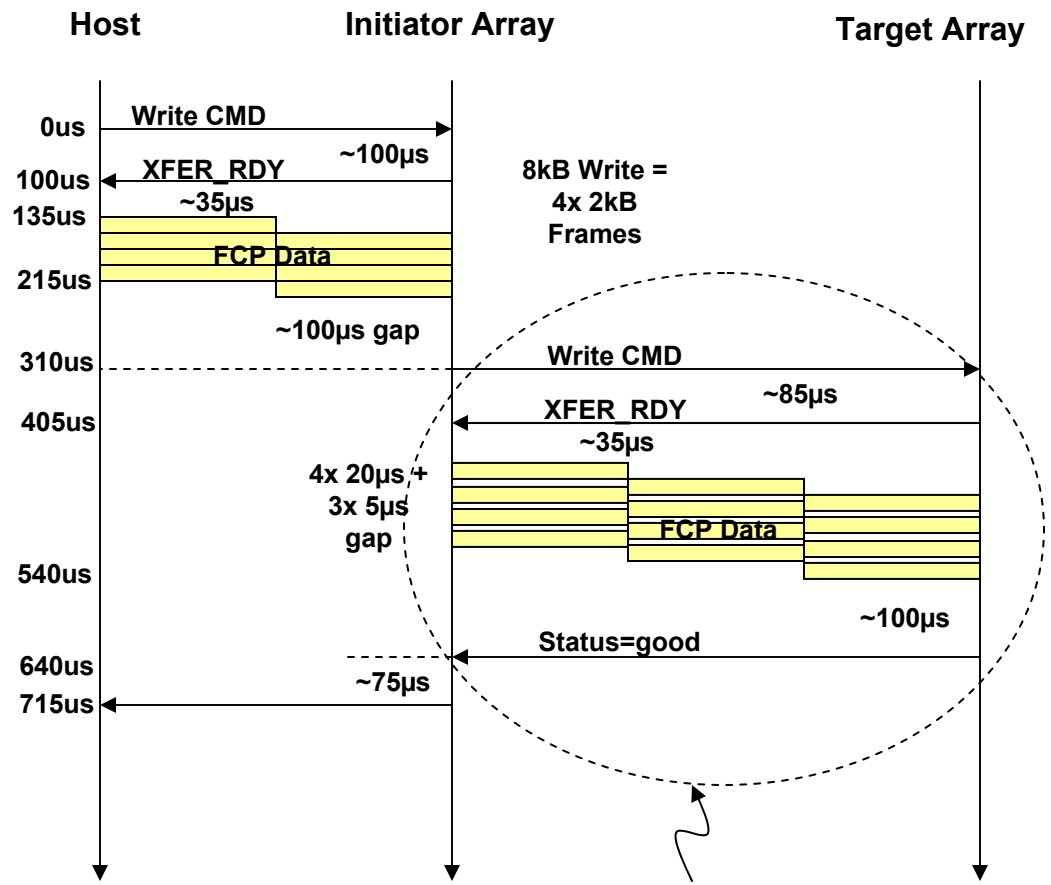


Asynchronous Replication: I/O Detail

Example: HP DRM



HP DRM Synchronous I/O Timings



HBA, Controller Ports and FC Switches all at 1Gbps

Introducing Distance between Initiator and Target will impact this part of the I/O with serialization delays, queuing and latency

- Average replicated 8kB Write I/O completed in around ~700-800µs (lightly loaded at around 80 IOPS)
  - local (Initiator) accounts for around half this (~300-400µs)
  - sync replication (Init -> Target) accounts for the remaining response time
- Most time is spent waiting for controller to respond (~75-100µs)
  - 2Gbps FC would have little impact in this situation

Overview: Storage Extension

What is SAN Extension?

- Extending Storage Area Network over metro or WAN distances.
- Interconnect Customer Data Centers
- Connection ranging from DS1 to OC48

SAN Extension
What options are available?

SAN extension Options	Customer	Protocol	Network	Distance/ Bandwidth
Leased Fiber	Lease Fiber	Native FC	Point-to-point Optical Loop	80km Full FC speed
Leased Line/ Lambda	Lease OC-3, OC-12, OC-48, Lambda	FC over SONET/SDH, DWDM	Point-to-Point Ring	Unlimited Distance in some cases FC speed or Sub-rate
Leased Bandwidth	Lease Data connection	FC over IP	Multi-point Dynamic provisioning	Depends on IP buffering, QoS, Latency

Limitation to extend SANs

- Physical Limitations
- Protocol Limitations
- Application Limitations

Physical Limitations

The maximum distance transported depends upon:

- **Transmitter power**
- **Receiver sensitivity**
- **BER**
- **Type of fiber plant**
- **Optical signal to noise ratio**

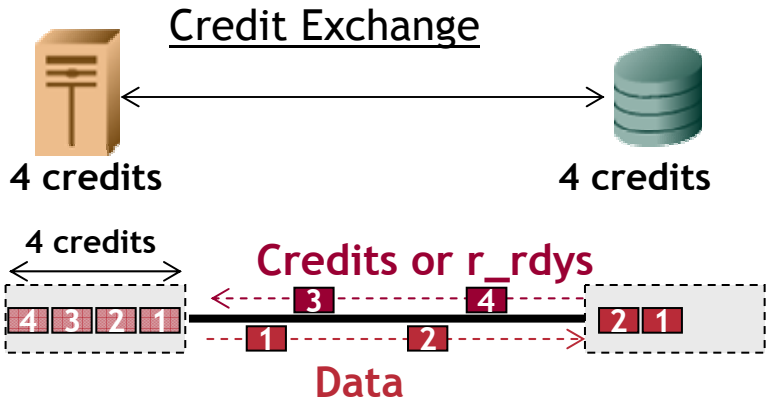
With standard FC GBIC the maximum reach is approximately 10 KM (~2.5 dB)

With extended FC GBIC the maximum reach is approximately 80 KM (~20 dB)

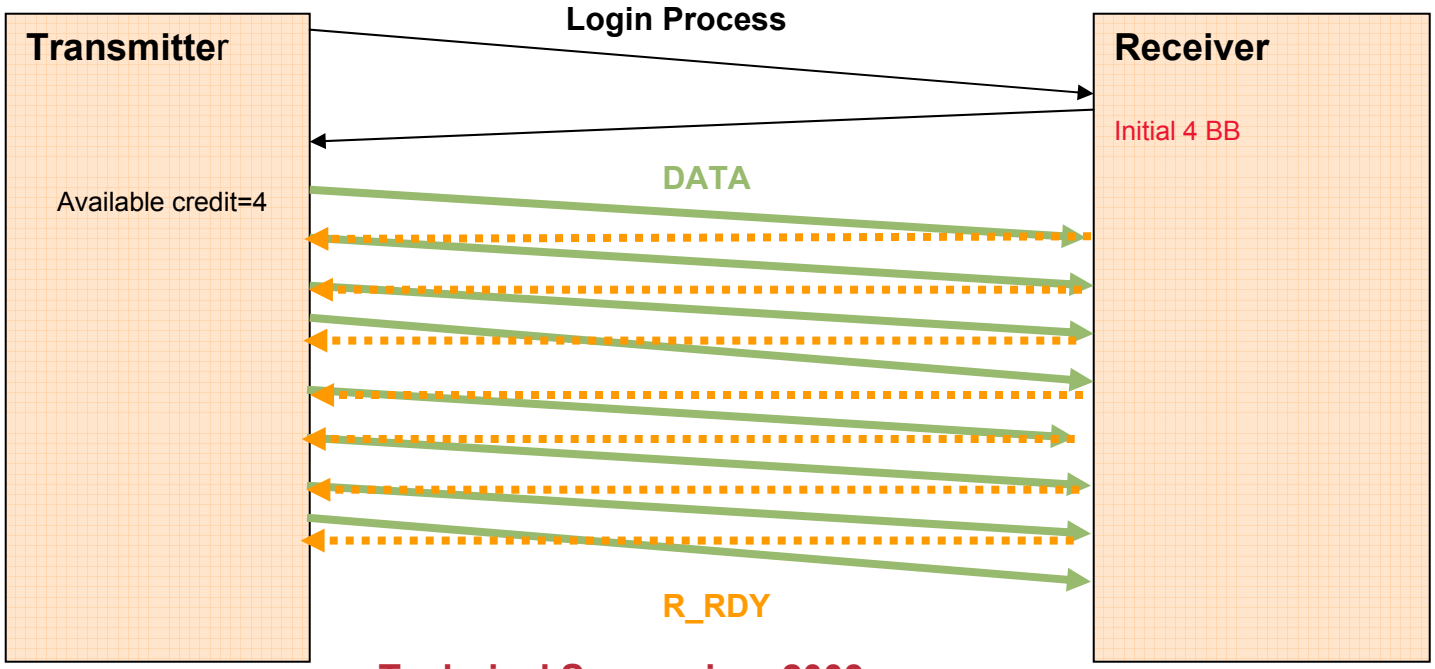
FC protocol Limitations

- Every frame sent is assumed to occupy far end interface buffer capacity until an R_RDY is returned.
- Full throughput cannot be supported beyond a distance based on the buffer capacity.
- Number of frames in flight cannot exceed the far end buffer capacity.
- Near end system must wait for R_RDY's to continue data transmission.
- Industry rule of thumb is one buffer credit can support up to 2 km of distance at full throughput. (for 1 G FC)
- Beyond this distance maximum sustainable throughput degrades

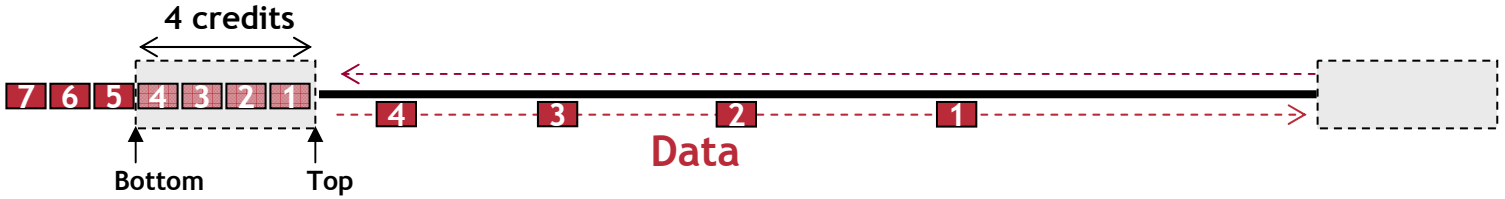
Fiber Channel Flow Control



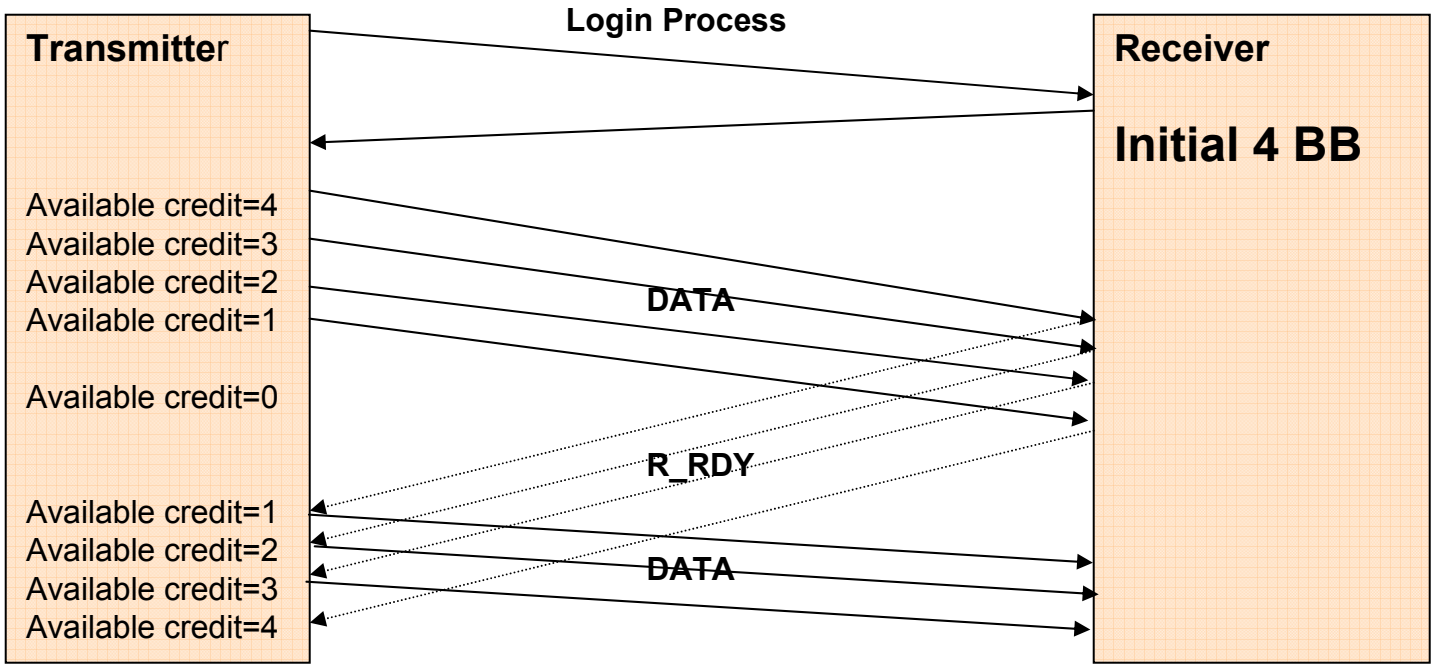
- Each credit(r_rdy) carries one FC frame
- A FC frame cannot be sent until an r_rdy is received
- The receiving device allocates the credits to the sending device



FC protocol Limitations



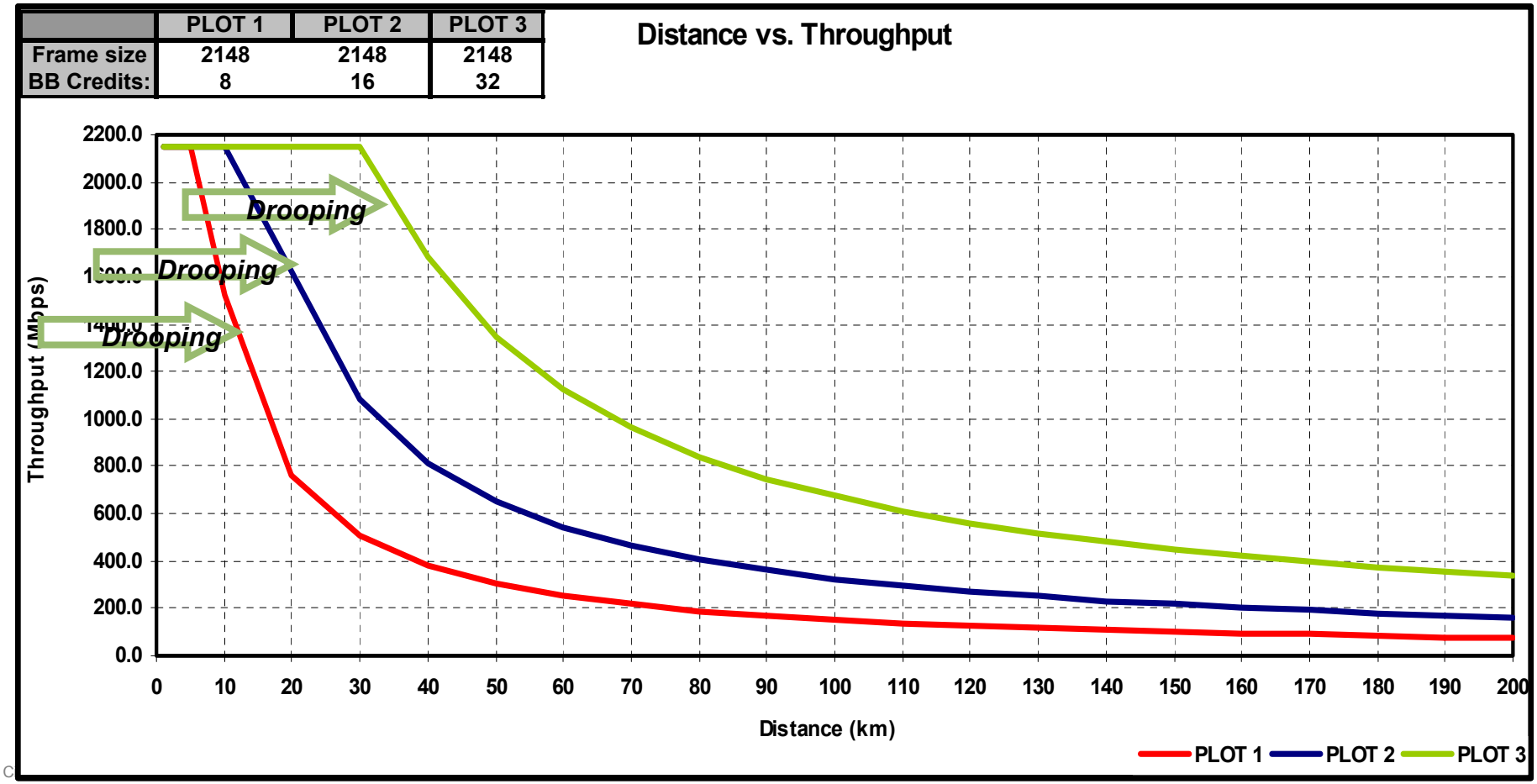
- As the distance is increased, a set number of credits can't fill the pipe
- Throughput degrades significantly as distance is increased



Drooping

"Drooping begins when the distance reaches a point where the time the light takes to make one round trip equal to the time it takes to transmit the number of bytes that matches the receiver buffer".

	PLOT 1	PLOT 2	PLOT 3
Frame size	2148	2148	2148
BB Credits:	8	16	32



Drooping

Drooping begins if: $BB\_Credit < \frac{RTT}{SF}$

Where RTT = Round Trip Time SF = Serialization delay for a data frame

Round Trip delay is affected by the serialization, propagation, processing and transmission delay. In the calculation that follows, processing delay is considered to be minimal compare to other delays.

Round Trip Time = Serialization delay of a data frame + Propagation delay of a data frame + Serialization of R_RDY + Propagation Delay of R_RDY.

SF = Serialization of a data frame
PF = Propagation Delay of data frame
SR = Serialization of R_RDY
PR = Propagation Delay of R_RDY

Drooping (Example)

Distance = **72** kilometers
Fibre Channel Speed: **2** G FC

SF = ((Frame size + 24 bytes for IDLE) * 8 bits/bytes *1.25 for 8b/10b encoding)
/ (FC Speed))* 1000000
= ((2148+24)*10 / (2125000000))*1000000
= (21720 / 2125000000) * 1000000
= 10.221 microsecond

PF = PR=72*5=360 Micro Seconds

SR = ((R_RDY Frame size + 8 bytes for 2 IDLE) * 8 bits/bytes *1.25 for 8b/10b encoding)
/ (FC Speed))* 1000000
= ((4+8)*10 /2125000000)*1000000
= 0.056 microsecond

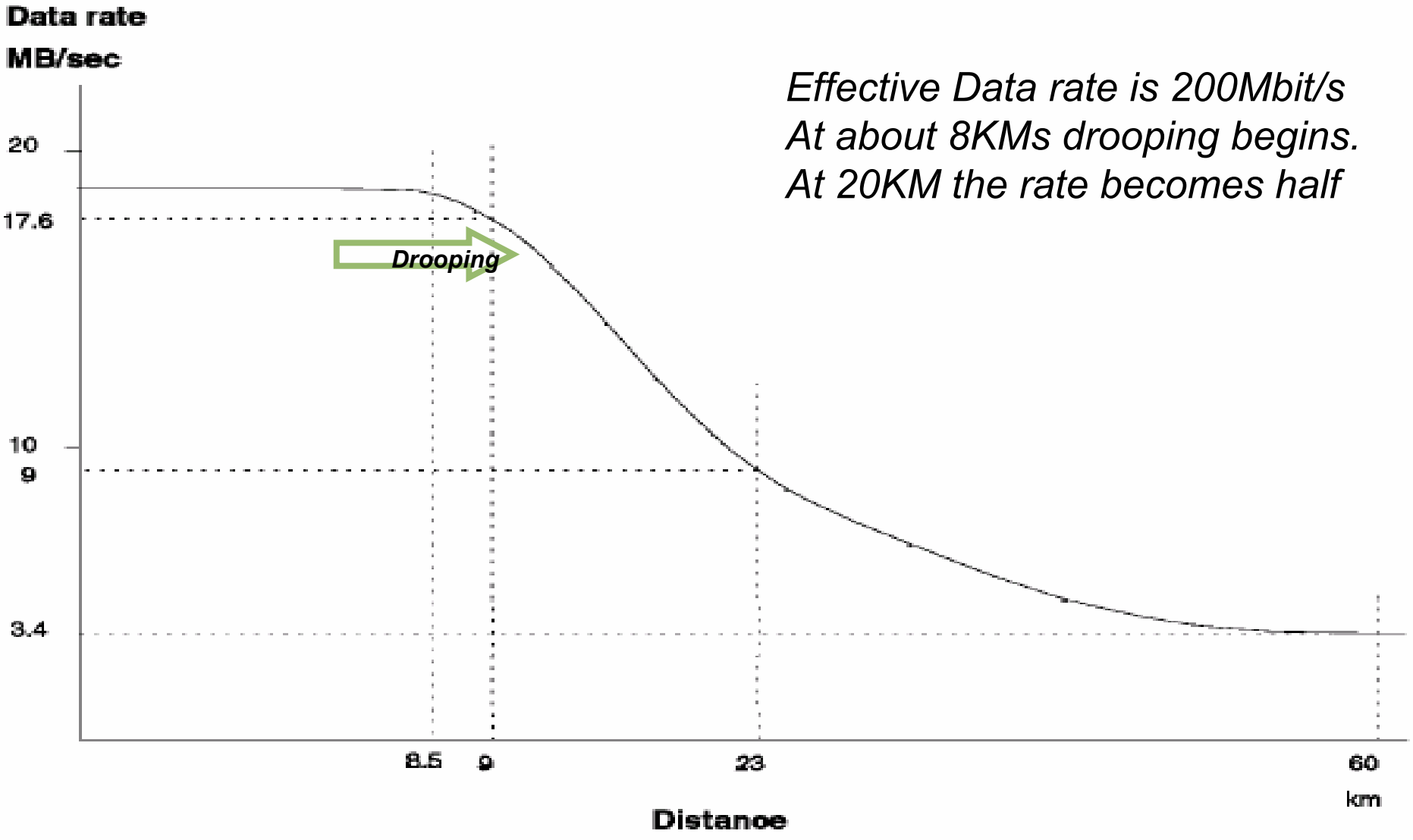
RTT = SF+SR+PR+PF = 10.221 + 360 + 0.056 +360 = 730.277 MicroSecond

RTT/SF = 730.277/10.221 = 71.44 ~= 72
Hence **2 G FC system requires 72 BB Credits to reach 72 KMs without drooping**.

Enterprise System CONnection ESCON

- Introduced in 1990 by IBM
- Bidirectional Serial Bit Transmission
- Primarily used for IBM S/390 solutions(replacement for old ‘bus and tag’ technology)
- Max speed of 20 MB/s (nominally 17 MB/s or *200Mbit/s)*
- Supports IBM S/390 channel command words (CCW)
- Point to Point Connections
- ESCON directors (switches) allow dynamic connection capabilities
- Transmission over fiber optic cable
- Max distance of 3 Km for each ESCON segment

ESCON



Effective Data rate is 200Mbit/s
At about 8KMs drooping begins.
At 20KM the rate becomes half

What are the Solutions? Spoofing

- Does not involve termination of FC link
  - Only terminates protocol error monitoring and flow control
  - End systems interoperate through solution transparently
- R_RDY's terminated locally and not part of flow control so they are not wasting WAN bandwidth
  - Supports maximum FC throughput independent of distance
- IDLE frames terminated locally and regenerated at far end allowing underutilized services equivalent performance over less bandwidth

What are the Solution? Compression

Compression is the reduction in size of data in order to save space or transmission time*.*

Compress the data content or on the entire transmission unit (including header data) *(depending on a number of factors).*

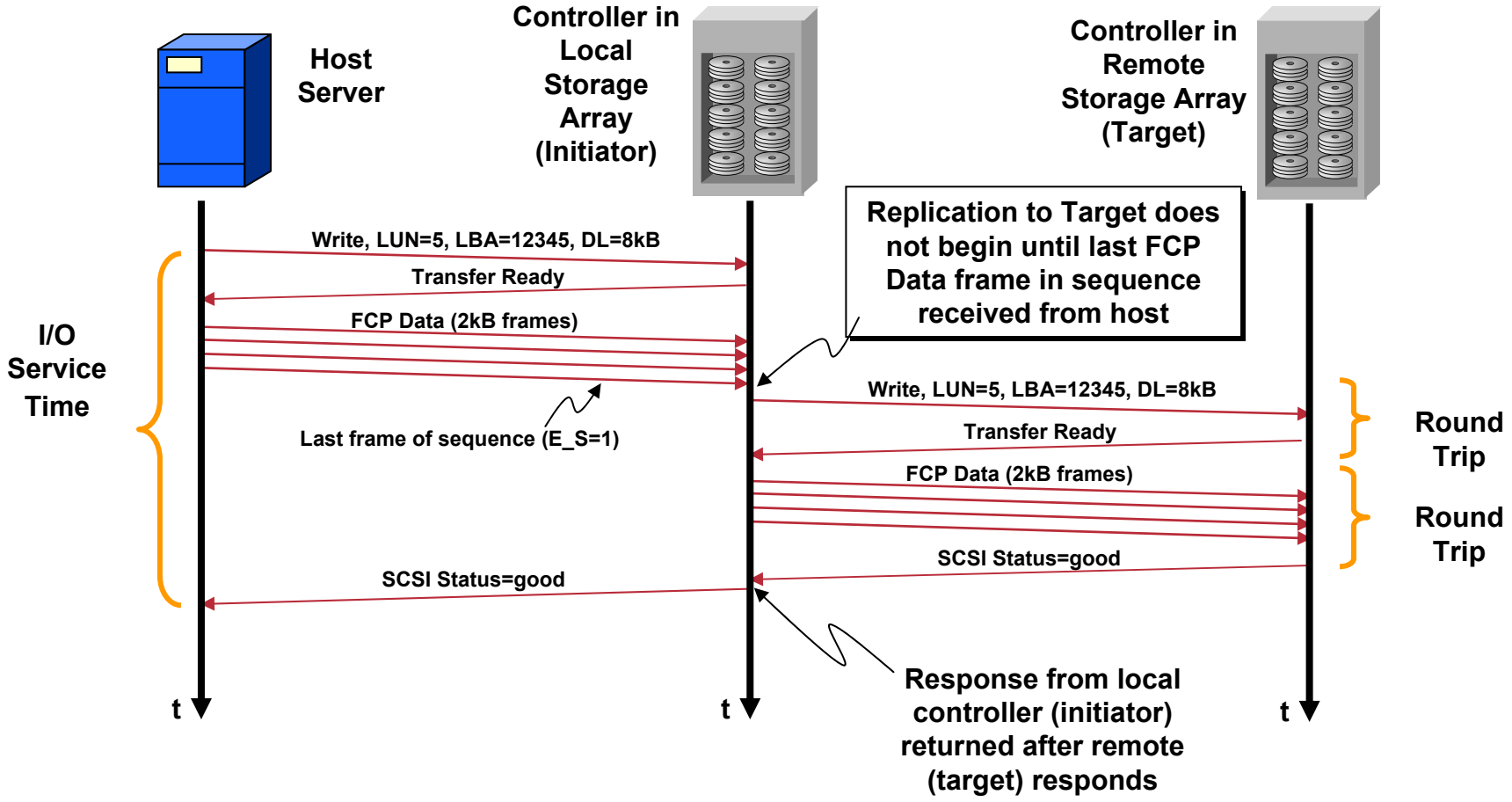
Content compression *:*
removing all extra space characters,
inserting a single repeat character to indicate a string of repeated characters,
and substituting smaller bit strings for frequently occurring characters.

can reduce a text file to 50 percent of its original size.

The distance we can send data is often gaited by the amount of space or memory that it takes up, so less information or data to send often means longer distances are achievable.

Synchronous Replication: I/O Detail

Example: HP DRM

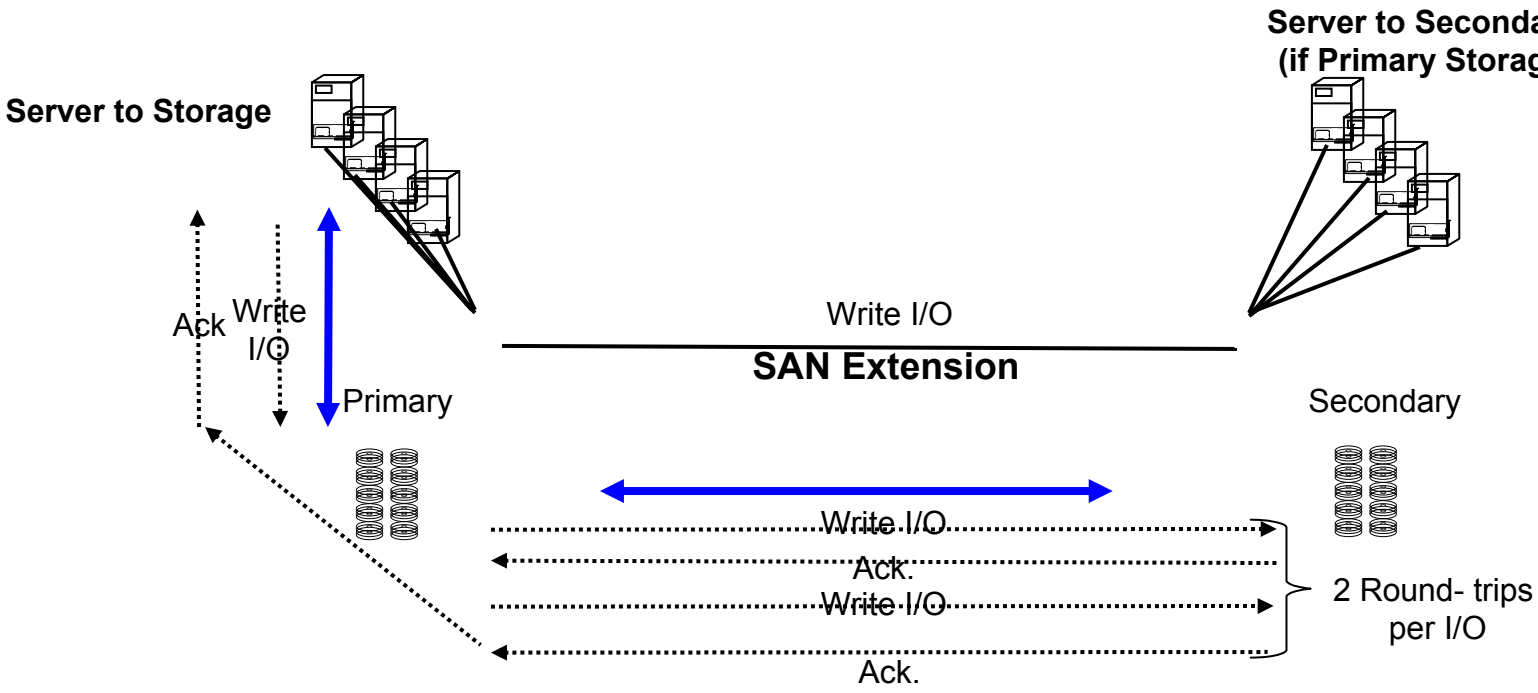


Application Limitation Synchronous Replication/Mirroring

- Distance limitation directly impacts I/O's response time.
- I/O not completed until written on both images of the mirror.
- With Synchronous Replication (versus Mirroring), I/O Write sequence are being serialized
- Two round trips per write = >20µs/km additional latency

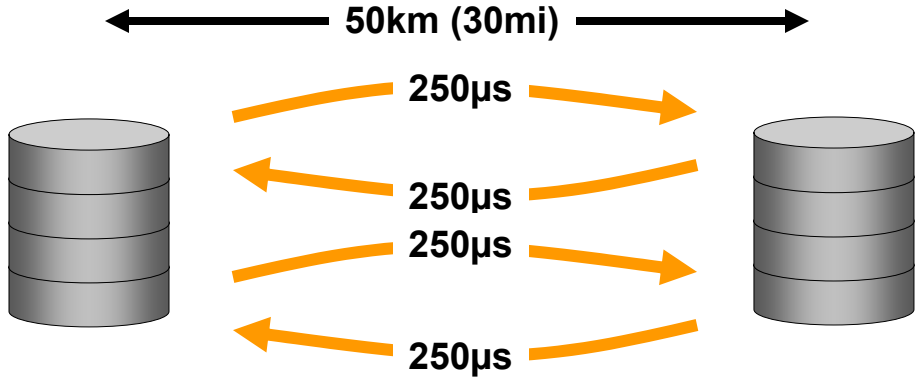
  At 50 kilometers, 1 millisecond added per I/O
- Requires dual path between the two sites.

Application limitation Synchronous Replication



- Only Write I/O's traverse the SAN extension, Read I/O are local
- 2 Round Trip Time per I/O

Latency & Synchronous Replication



Speed of Light

c = 3 x 10^8m/s (vacuum) ≈ 3.3µs/km

Speed through fiber ≈ ⅔ c ≈ 5µs/km

Two Round Trips between source and destination arrays per write I/O

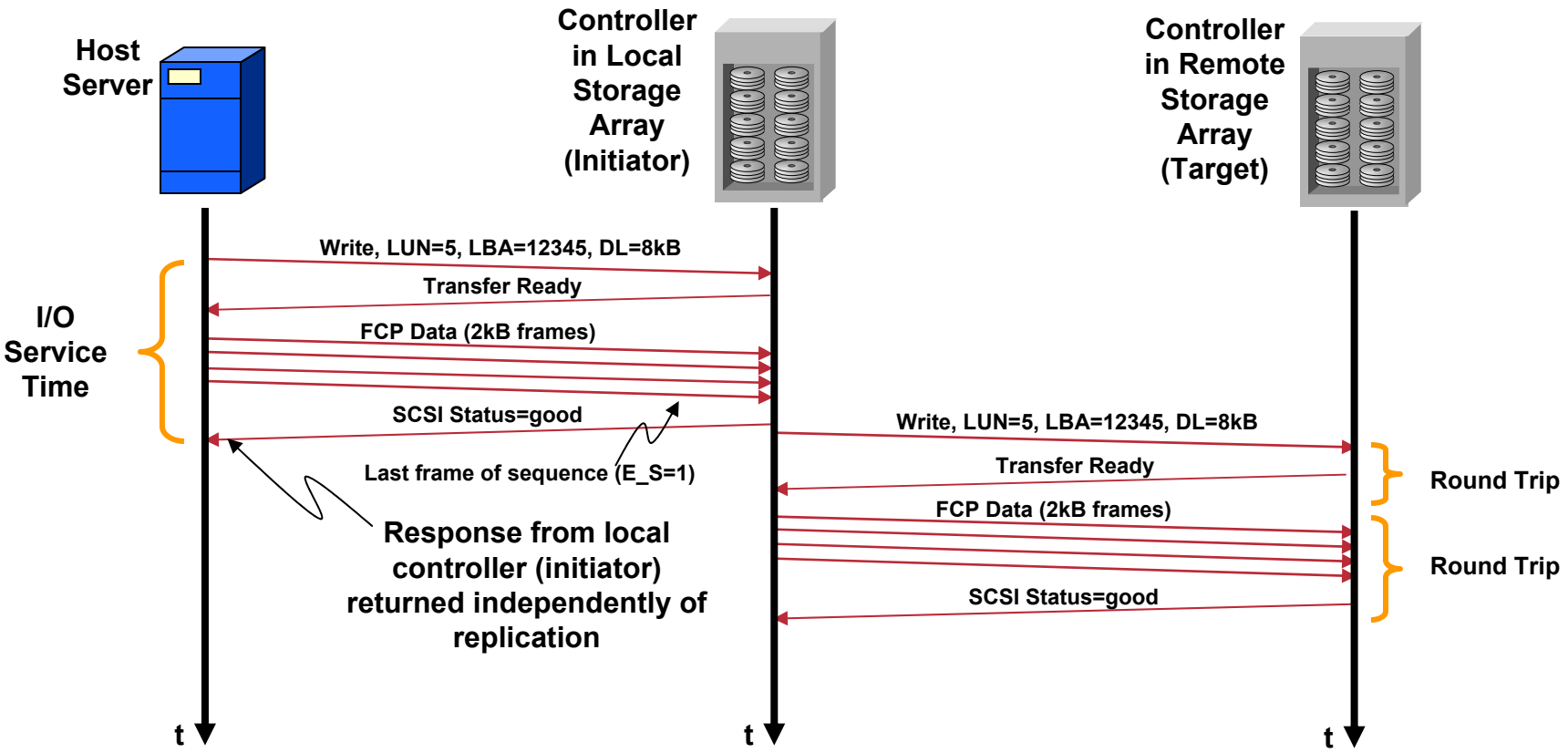
→ 2 x 2 x 5µs/km = 20µs/km additional latency

e.g. at 50km → additional 1000µs (1ms) I/O Service time (write) with Synchronous replication

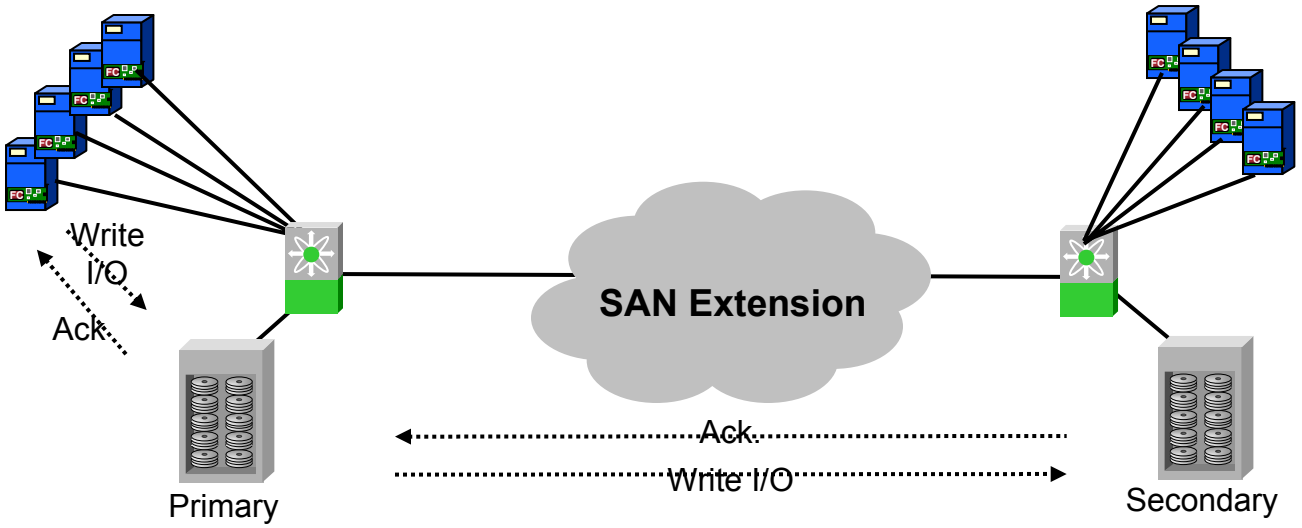
- Implementation dependent (2RTT for SRDF, DRM)

Asynchronous Replication: I/O Detail

Example: HP DRM



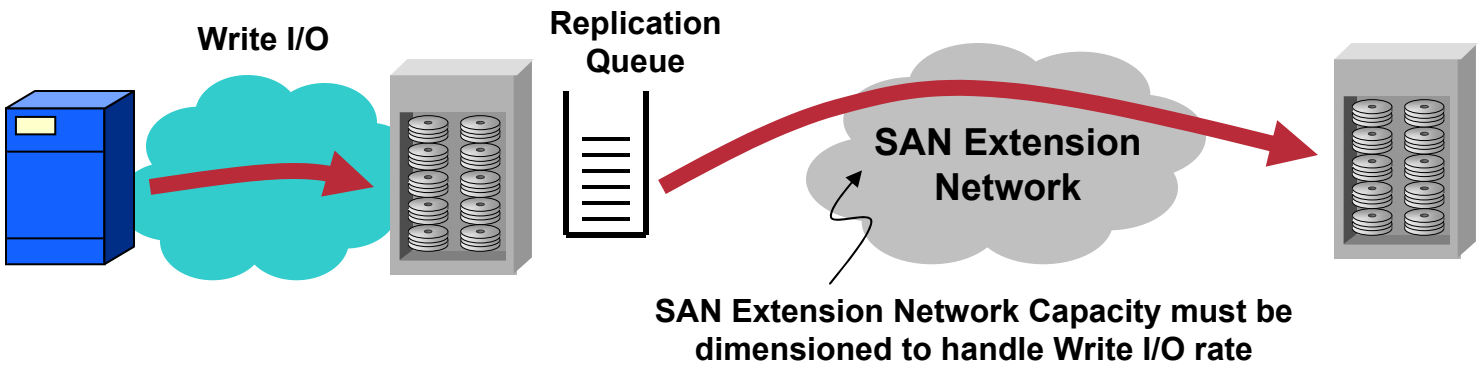
Application limitations Asynchronous Replication



- **Distance is not a limitation as Write I/O acknowledgement is local.**
- **If the Bandwidth provisioned is too low, Write I/O's start being queued at the primary storage array.**
- **Application automatically switch to synchronous replication if too many outstanding Write I/O's.**

Asynchronous Replication Considerations (2)

- **If the configured "lag" limit is reached, replication changes to Synchronous Mode to clear the backlog (e.g. SRDF & DRM)**
  - **→Instantly raises Write I/O response time**
  - **- Occurs if: Write I/O Rate > SAN Extension capacity**



FCIP Storage Extension

Cisco Storage Vision

IP Protocol Encapsulation

- IP SANs carry block I/O traffic on top of IP

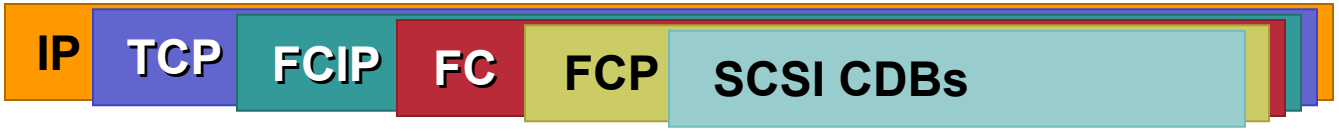
  Leverage Gigabit Ethernet performance for local traffic

  Use TCP: A reliable transport for delivery in MAN/WANs

- Two primary protocols:

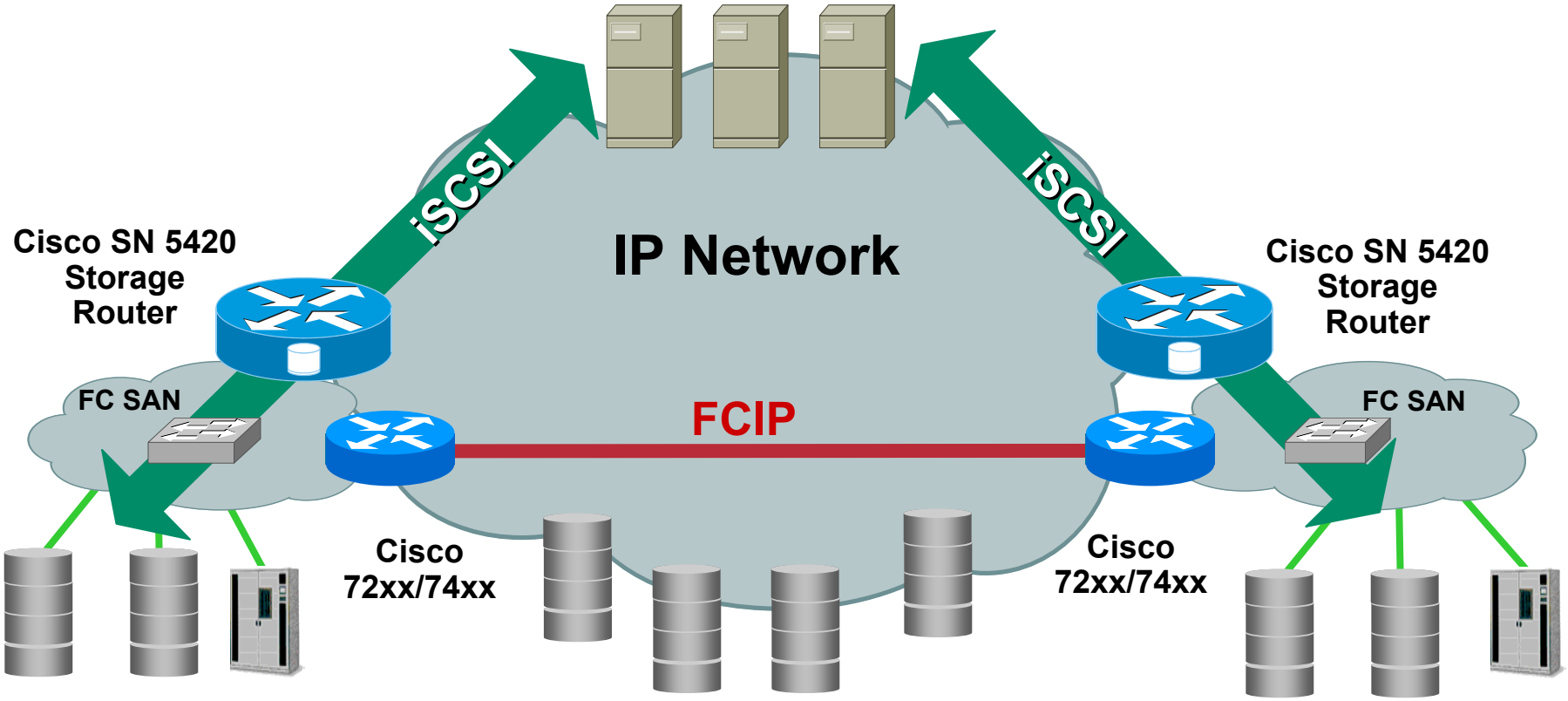
  iSCSI—"IP-SCSI" IP-native transport of SCSI CDBs and data within TCP/IP connections

  FCIP—"Fibre-Channel-over-IP"— Tunneling of Fibre Channel frames within TCP/IP connections, including FC fabric management frames



FCIP and iSCSI – Complementary Standards

- FCIP: SAN-to-SAN over IP
- iSCSI: Host to Storage over IP



The Status Of The FCIP Standard With IETF

- **IETF Internet Draft (Standards-track)**
  http://www.ietf.org/internet-drafts/draft-ietf-ips-fcovertcpip-12.txt
- **Latest Update – February, 2003**

  Protocol originally introduced by Lucent Technologies, Gadzoox Networks, and Brocade Communications

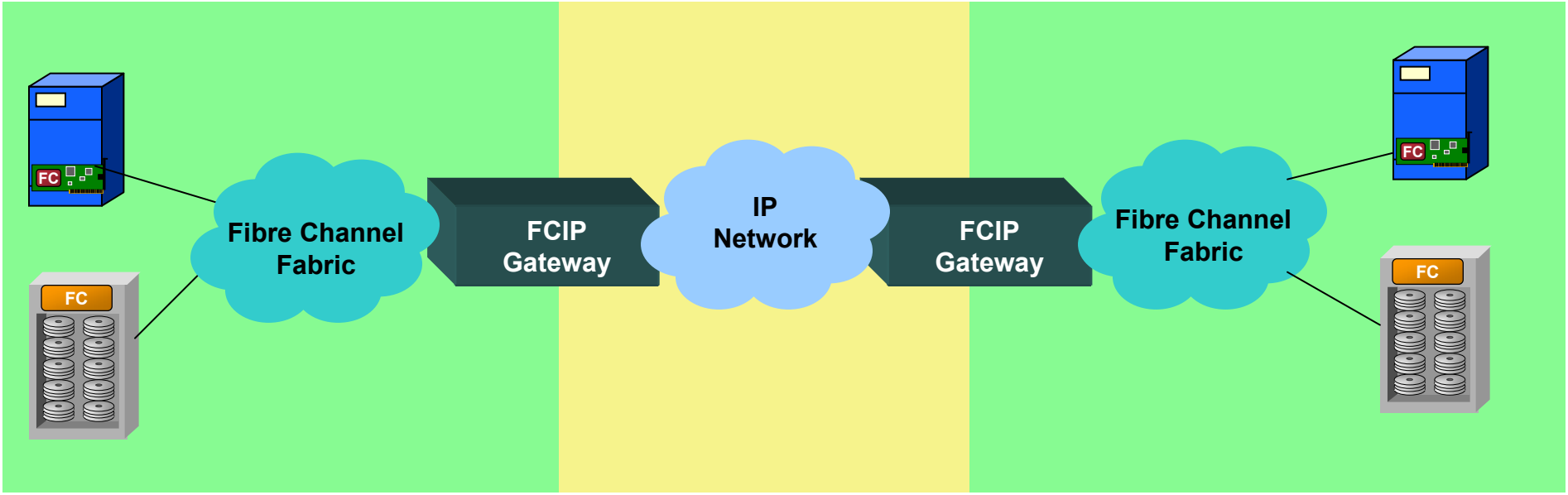
  Additional co-authors of the evolving standard include: Lightsand Communications, Vixel, McData, Qlogic, Aarohi Communications, Compaq, Pirus Networks, Cisco Systems, Rhapsody Networks, and SAN Valley

  Numerous vendors including Cisco have developed FCIP products and continue to follow FCIP developments

  Next step is for ratification to an Internet Standard

What is FCIP (Fibre Channel over IP)?

IT creates one logical fabric between remote SANS, and the switches think they are connected. IP is only used for tunneling through the WAN.

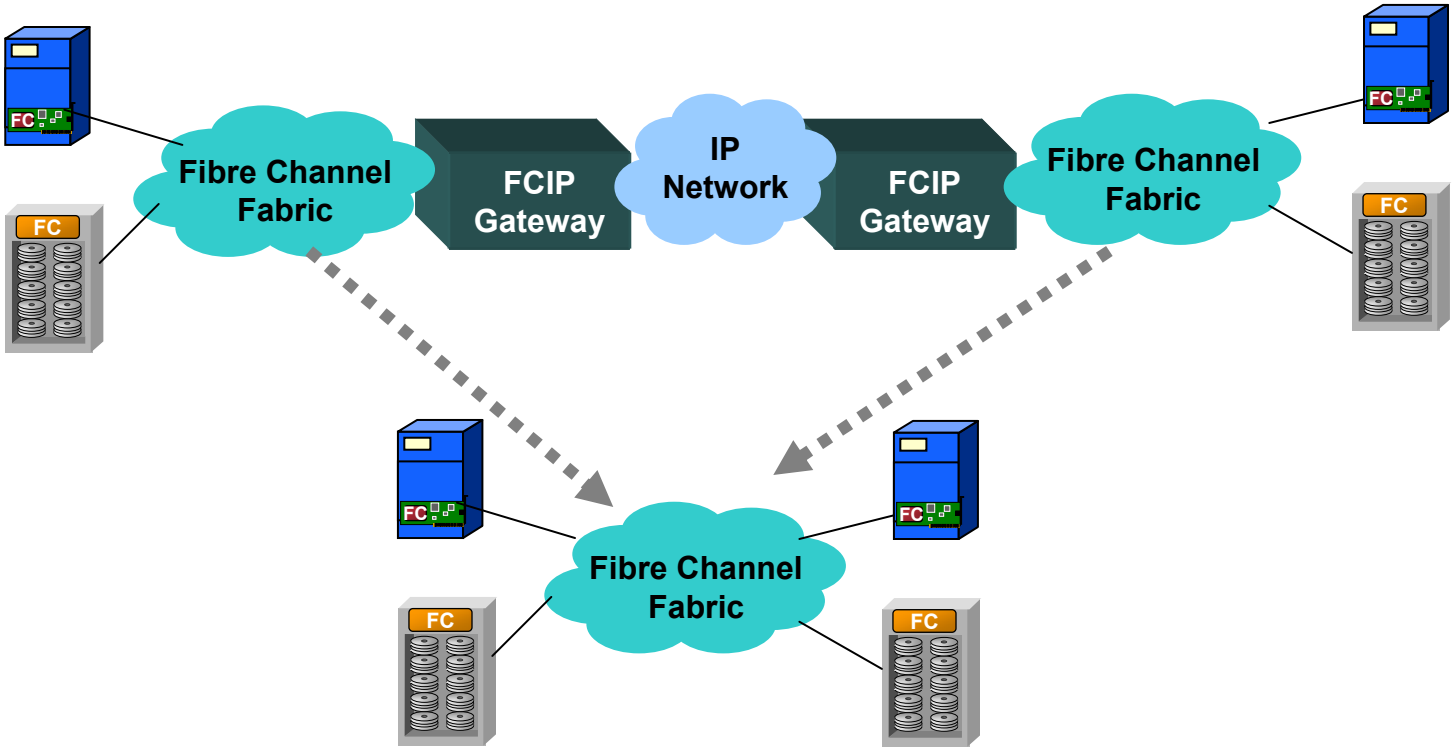


What is FCIP (Fibre Channel over IP) cont…

Remote FC resources are viewed as local

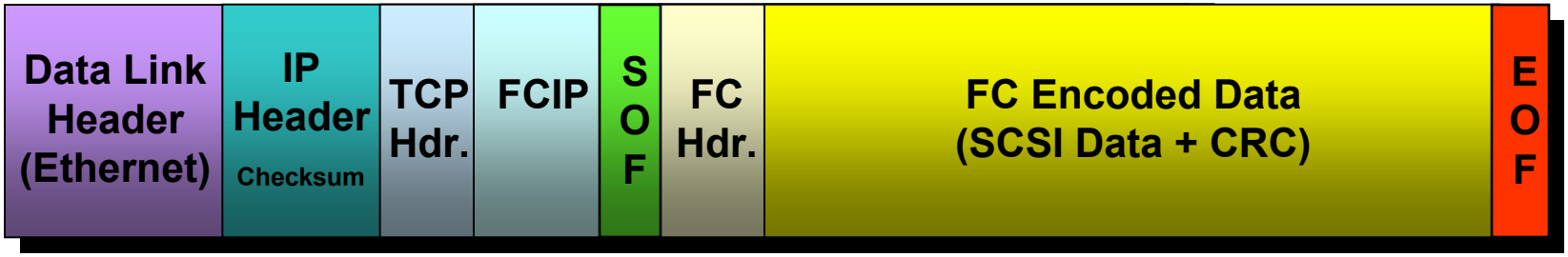
FCIP creates a Virtual FC Inter-Switch Link (ISL)

Fabric service information is extended across the FCIP ISLs



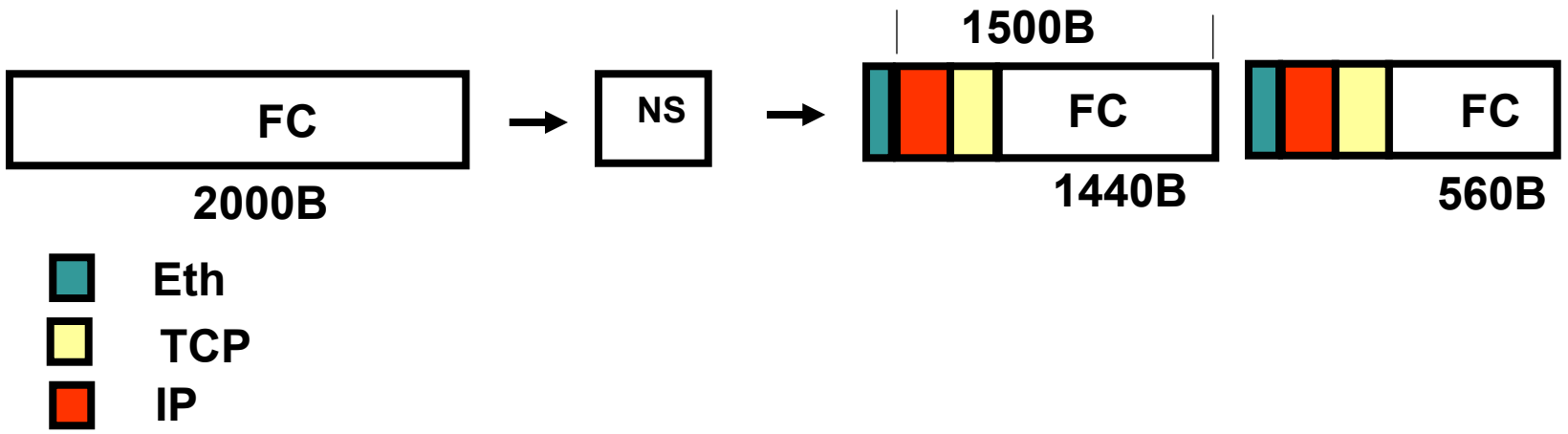
FCIP Packet Structure

- TCP/IP is the underlying transport protocol
  - Flow Control / Retransmission during Network Congestion
  - In-order packet delivery of error-free data
- Fibre Channel fabric domains are mapped to IP addresses
- All classes of FC frames are treated the same as Datagrams
- Blocks will typically be fragmented and re-assembled in-order
- IP is unaware of the Fibre Channel Payload and the Fibre channel fabric is unaware of the IP network

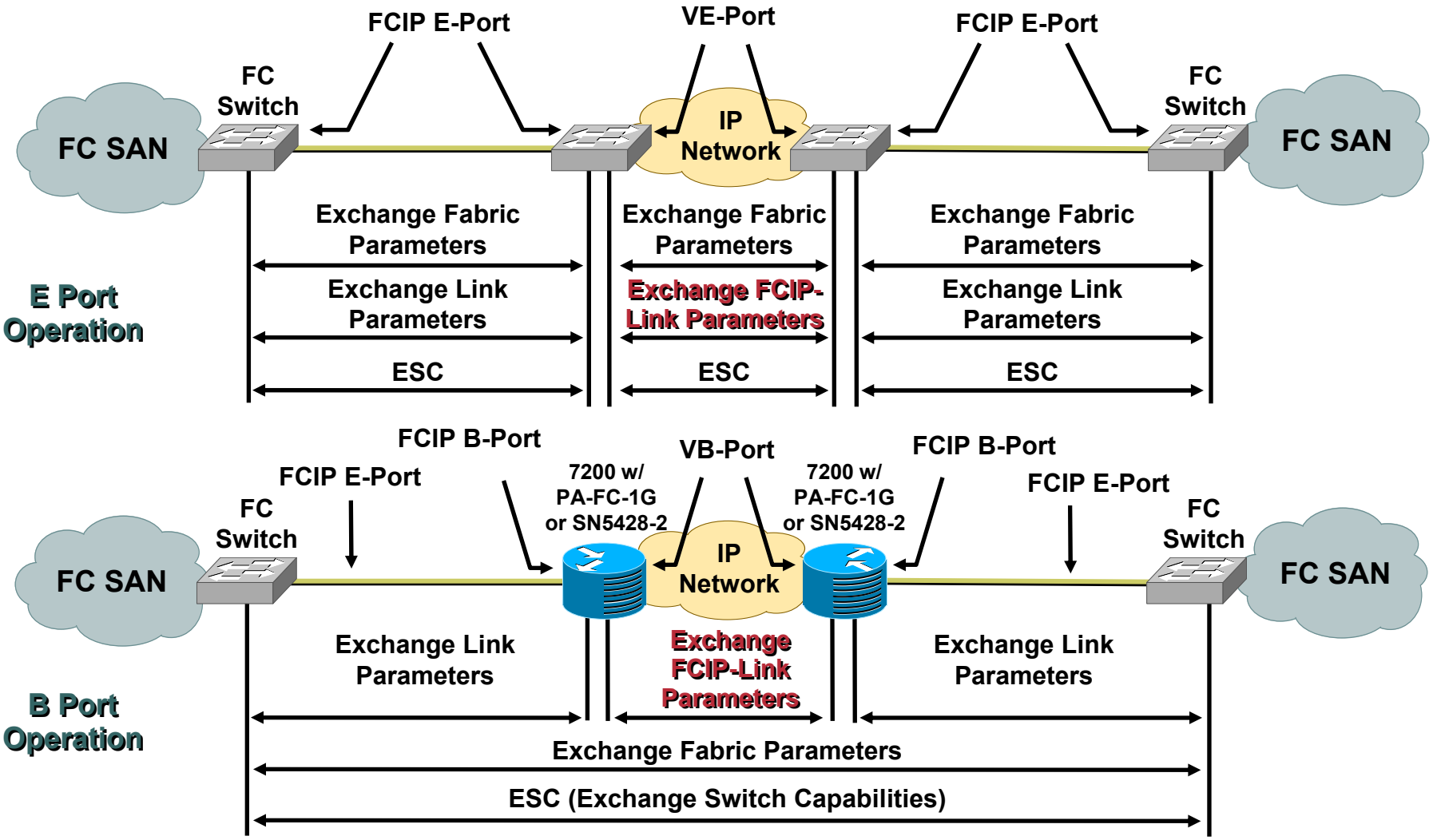


MTU and Max Segment Size (MSS)

- The MSS is derived from the configured MTU value on the FCPA interface
- MSS = MTU – 60B (IP + TCP + TCP options)
- FC frame greater that MSS value will be segmented by NorthStar.

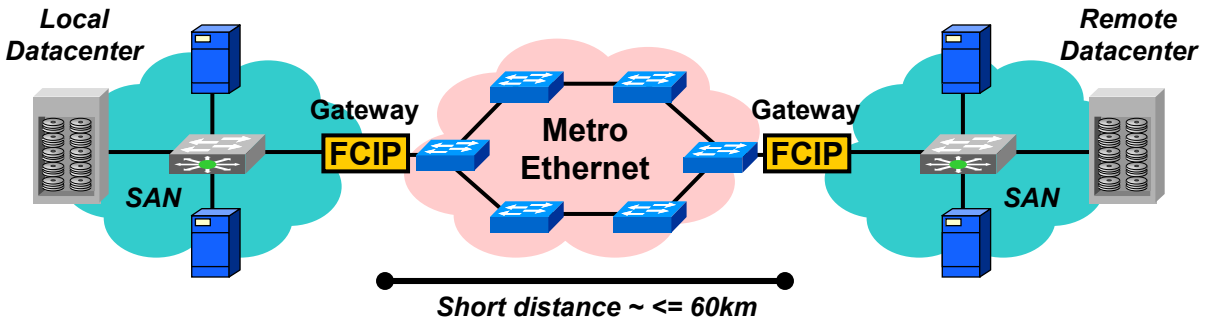


FCIP E-Port and B-Port Relationships

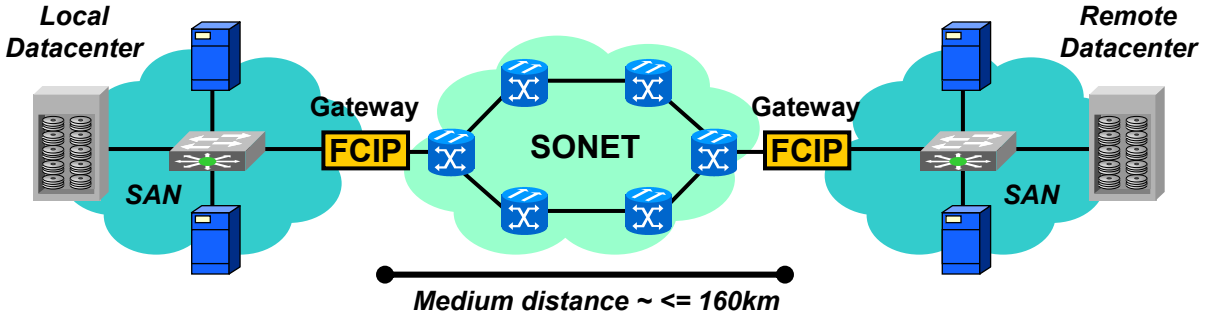


Potential FCIP Environments

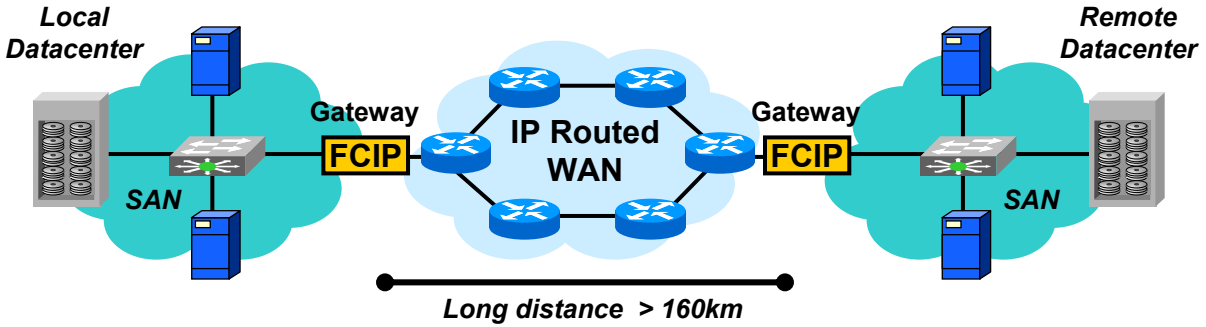
- 1Gb->OC48 or Higher
- Relatively low latency
- Synch/Asynch Applications



- Typical OC3 / OC12
- Relatively low latency
- Mainly asynchronous
- Suitable for some synchronous apps



- Low speed (T1 – DS3)
- Higher latency
- Longer distance
- Mainly asynchronous



MDS Series w/ Multiprotocol Support

Cisco.com

MDS 9500 Multilayer Directors

MDS 9216 Multilayer Fabric Switch

MDS 9216

MDS 9506

MDS 9509

MDS 9513

MDS 9000 Modules

Supervisor 1

16-port FC

32-port FC

8-port IP

Mgmt

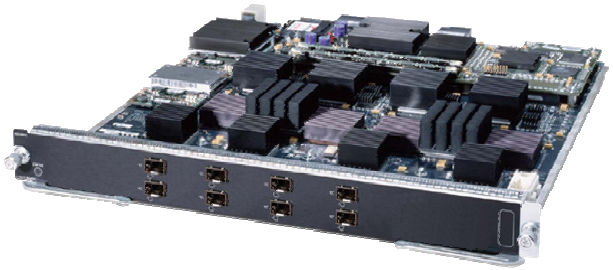
Cisco Fabric Manager

OS

MDS 9000 Family-OS

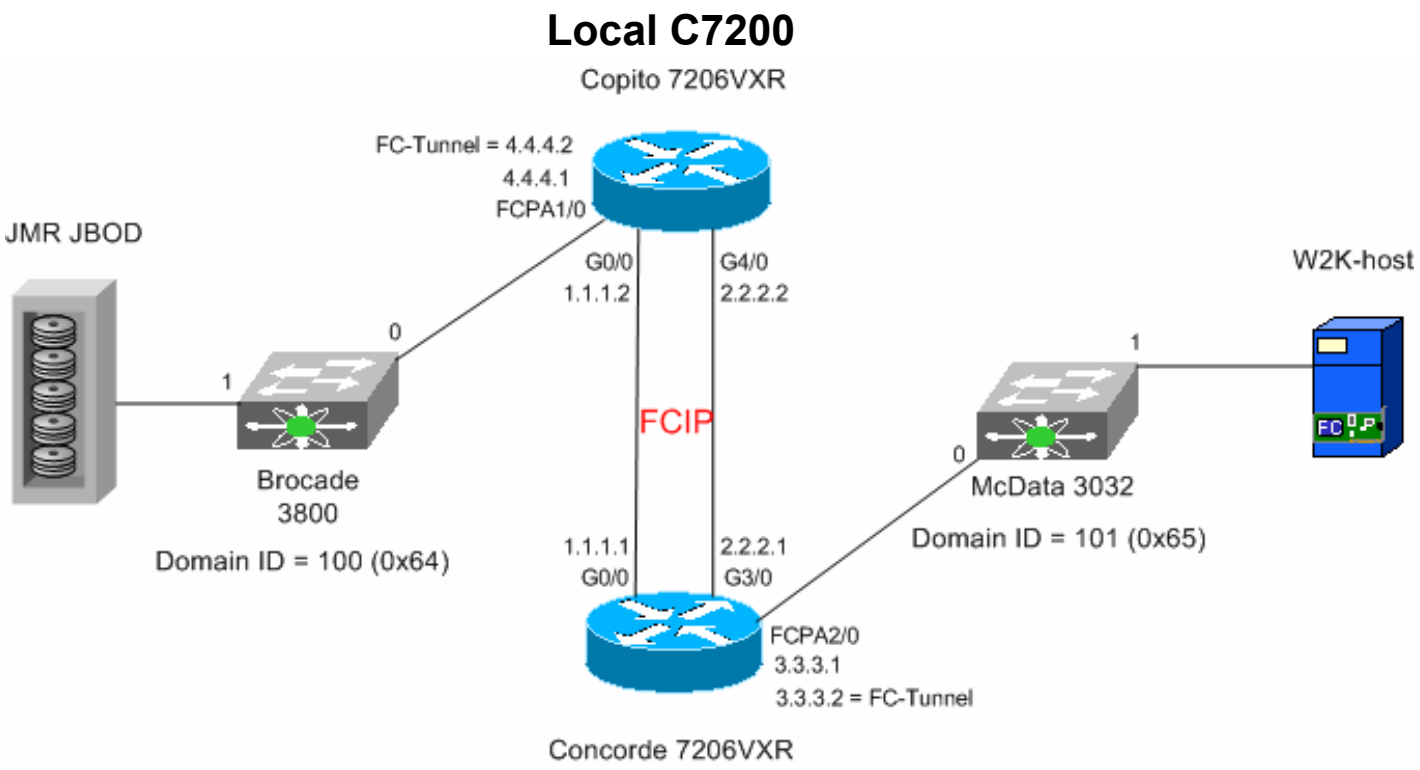
IP Storage Services Module Basic Features

- **Interfaces**
  - 8-port 1 Gbps Ethernet with SFP/LC optical interfaces
- **iSCSI Feature Highlights**
  - iSCSI Initiator-Fibre Channel Target
  - Transparent view of all allowed hosts/targets
  - iSCSI to Fibre Channel zone mapping
- **FCIP**
  - Up to 3 FCIP tunnels per port on all ports (24 tunnels per line card)
- **Fibre Channel Features**
  - All standard Fibre Channel line card features (interfaces N/A)
  - Leverages Fibre Channel interfaces on other switch modules
- **Performance**
  - Line rate performance on all ports for all protocols



7200/7400-series PA-FC-1G FCIP

- **Example of FCIP connectivity between McData and Brocade using the PA-FC-1G Port Adapter**



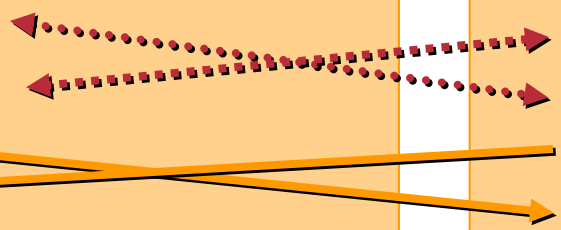
FCPA Configuration

Local Side Tunnel

```
interface Fcpa1/0
 ip address 4.4.4.1 255.255.255.0
 fc-tunnel Local
   src-ip 4.4.4.2
   dest-ip 3.3.3.2
   src-port 3000
   dest-port 2000
   tcp sack
   tcp mws 256
   tcp kad 7200
   ip tos 0
   inservice
```

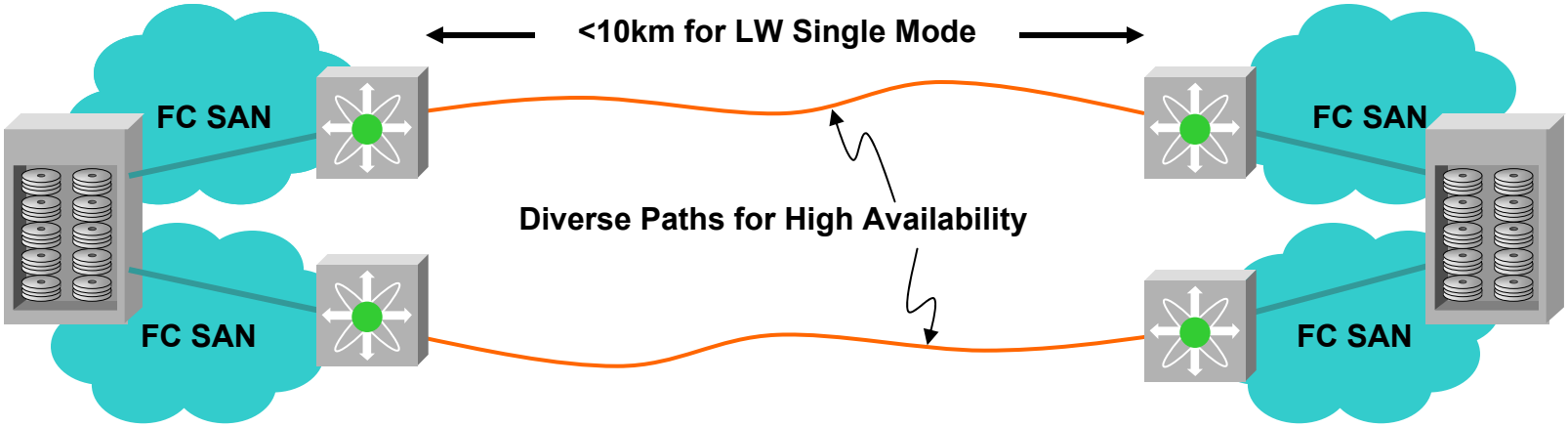
Remote Side Tunnel

```
interface Fcpa2/0
 ip address 3.3.3.1 255.255.255.0
 fc-tunnel Remote
   src-ip 3.3.3.2
   dest-ip 4.4.4.2
   src-port 2000
   dest-port 3000
   tcp sack
   tcp mws 256
   tcp kad 7200
   ip tos 0
   inservice
```



Optical Storage Extension

Dark Fiber



- Single 1 or 2 Gbps link per fiber pair
  - SW (850nm) 300m over 62.5/125µm Multimode
  - SW (850nm) 500m over 50/125µm Multimode
  - LW (1310nm) 10km over 9/125µm Single Mode
- "Client Protection" – ULP (SAN or Application) responsible for failover protection

What is SONET?

- SONET is a physical layer network technology designed to carry large volumes of traffic over relatively long distances on fiber optic cabling. SONET was originally designed by ANSI for the public telephone network in the mid-1980s.

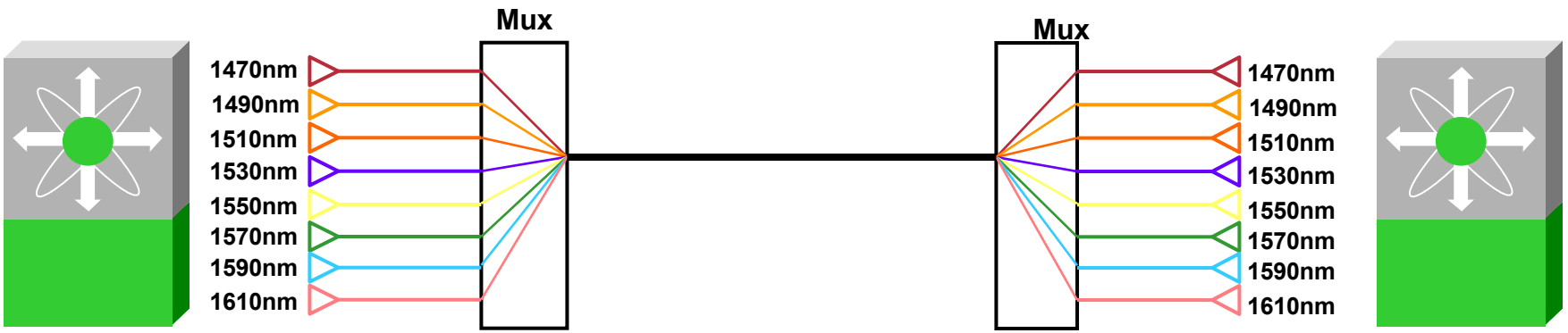
What is DWDM?

- DWDM is a technique for increasing the bandwidth of optical network communications.
- DWDM allows dozens of different data signals to be transmitted simultaneously over a single fiber.
- To keep the signals distinct, DWDM manipulates wavelengths of light to keep each signal within its own narrow band.

So What's the Difference?

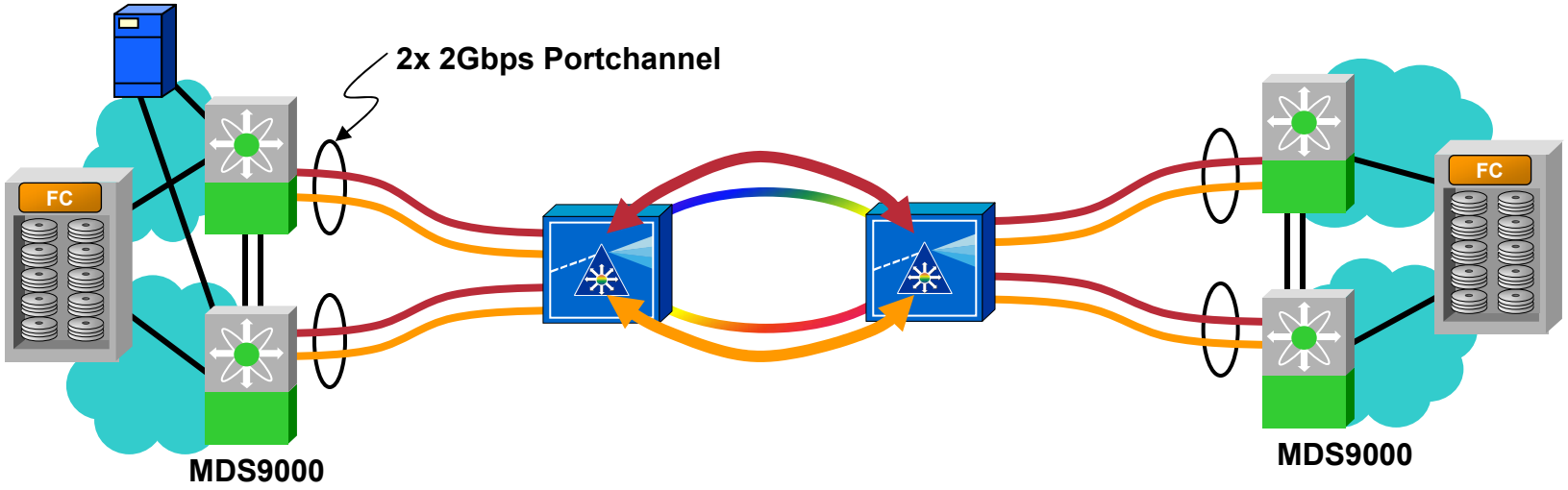
- Early DWDM equipment was SONET-only!
- SONET-only DWDM equipment DWDM was also known as closed interface DWDM while later any-protocol DWDM equipment was referred to as open interface DWDM
- Most metro gear will carry native ESCON, Ethernet, FDDI as well as SONET and is respectively referred to as 'DWDM'

CWDM – Course Wave Division Multiplexing



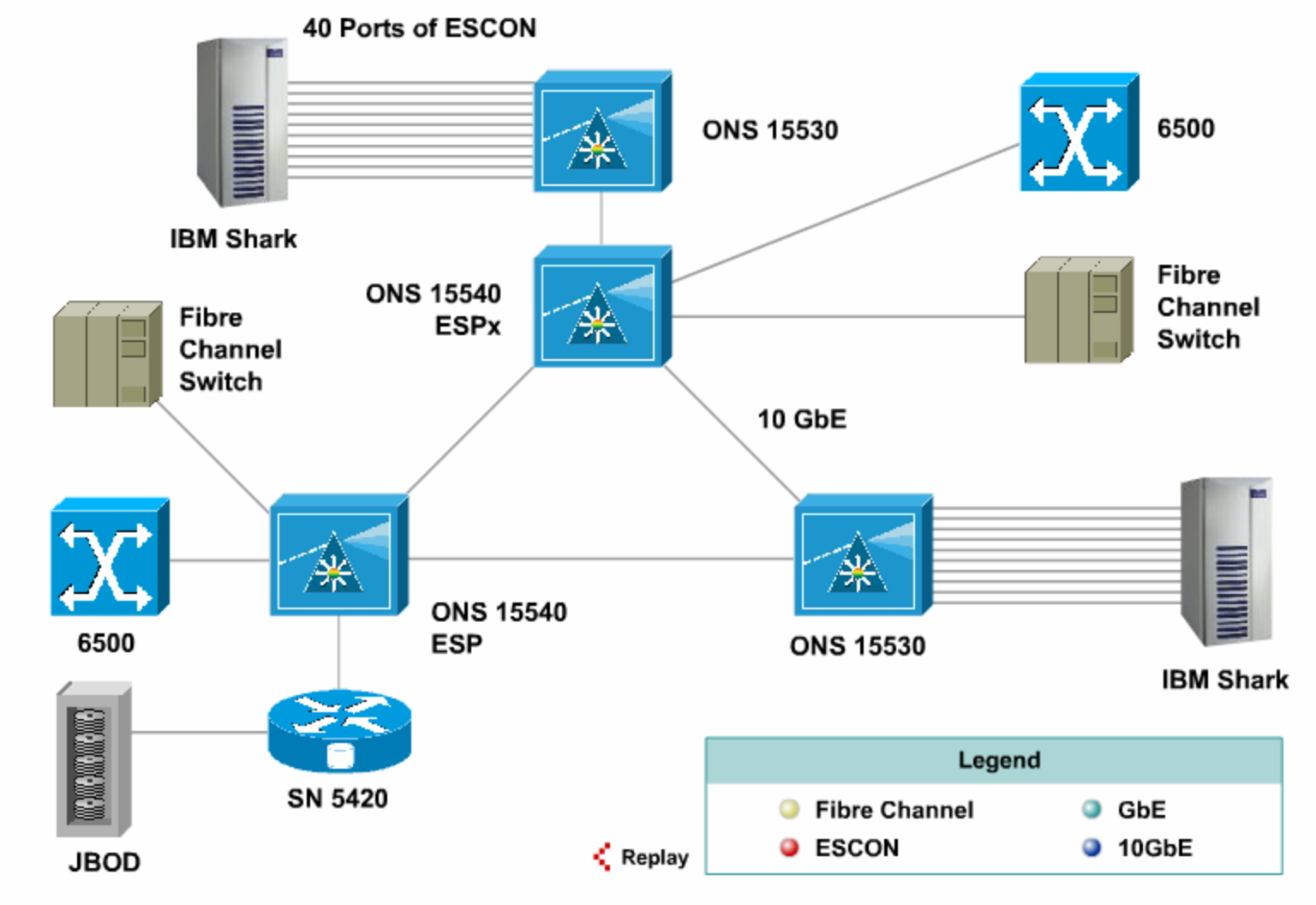
- 8-channel WDM at 20nm spacing (cf DWDM at <1nm spacing)
  - 1470, 1490, 1510, 1530, 1550, 1570, 1590, 1610nm
- Special "Colored" SFPs (or GBICs) used in FC Switches
- Muxing done in CWDM OADM (Optical Add/drop Multiplexer)
  - passive (unpowered) device – just mirrors and prisms
- 30dBm power budget (36dBm typical) on SM fiber
  - ~90km Point-to-point or ~40km ring
- Not EDFA amplifiable

DWDM HA Storage Network Topology

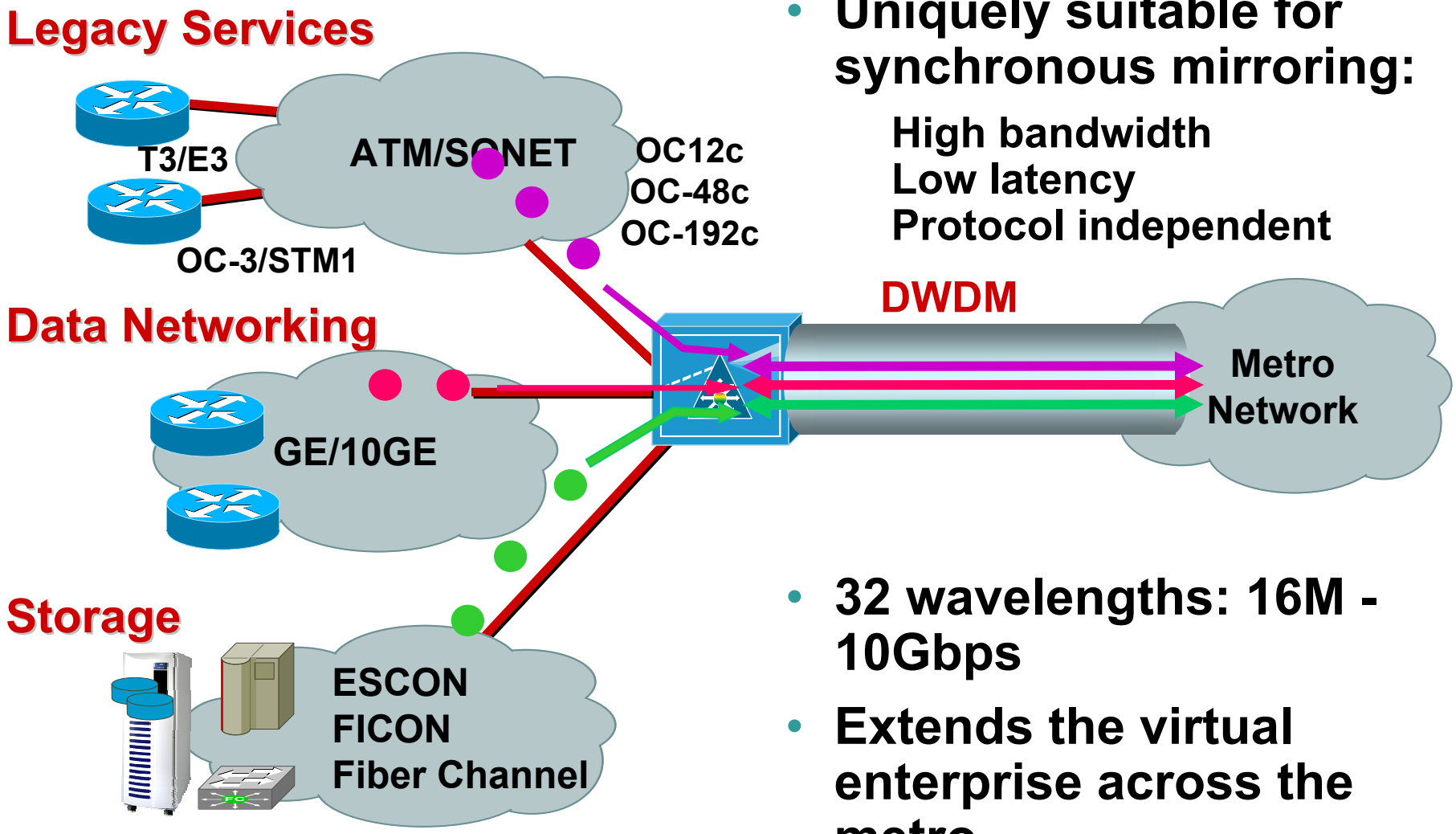


- Client Protection Recommended – Fabric and Application responsible for failover recovery
- Portchannel provides resilience
  - Portchannel members follow diverse paths
  - Single fiber cut will not affect Fabric (no RSCNs, etc…)
  - Use “Src/Dst” hash for load balancing (rather than “Src/Dst/Oxid” per Exchange) for each extended VSAN

DWDM HA Storage Network Topology



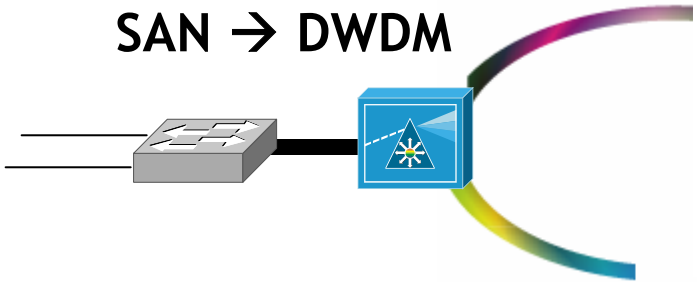
DWDM Multi-Service Network Topology



- **Uniquely suitable for synchronous mirroring:**
  - High bandwidth
  - Low latency
  - Protocol independent
- **32 wavelengths: 16M - 10Gbps**
- **Extends the virtual enterprise across the metro**

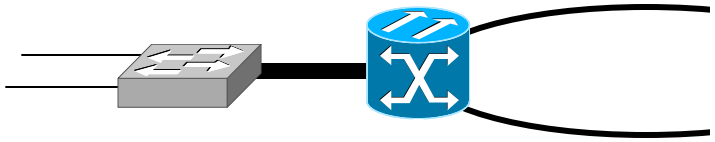
Alternatives for SAN Extension

SAN → DWDM



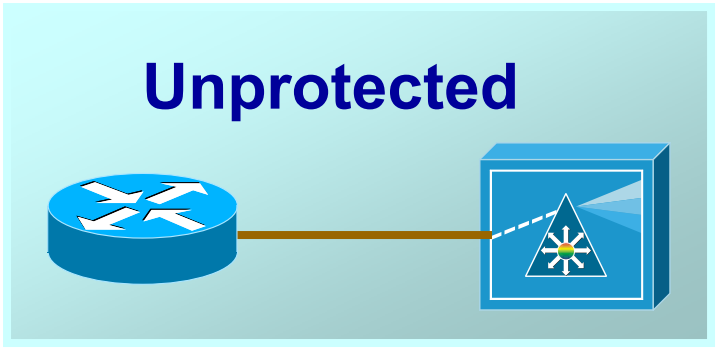
- Fortune 200 coverage
- ~2,000 metro DWDM rings deployed
- High density/High Bandwidth deployments
  - Stringent RTO/RPO requirements
  - Tier 1 DR/BC services

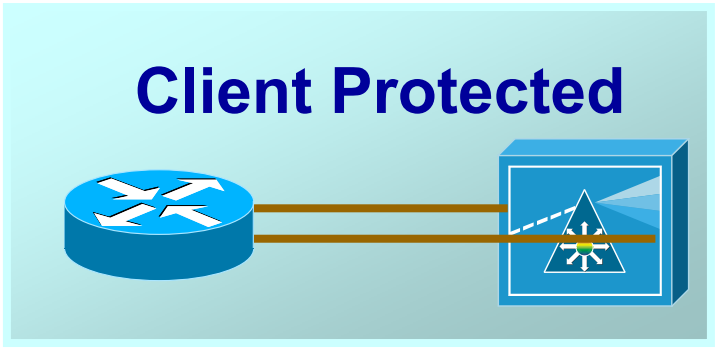
SAN → SONET

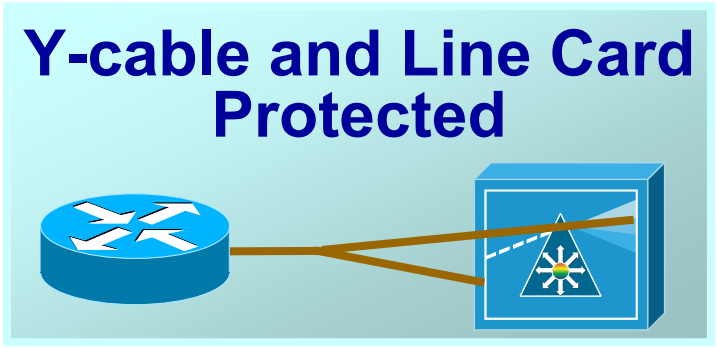


- Fortune 5000 coverage
- SONET Ubiquity ⇒ Enterprise Connectivity
  - Over 150,000 SONET rings deployed. source:RHK
- Seamless MAN/WAN connectivity
  - DS3, NxSTS-1 bandwidth
  - Well-understood operations model

Optical Protection Schemes

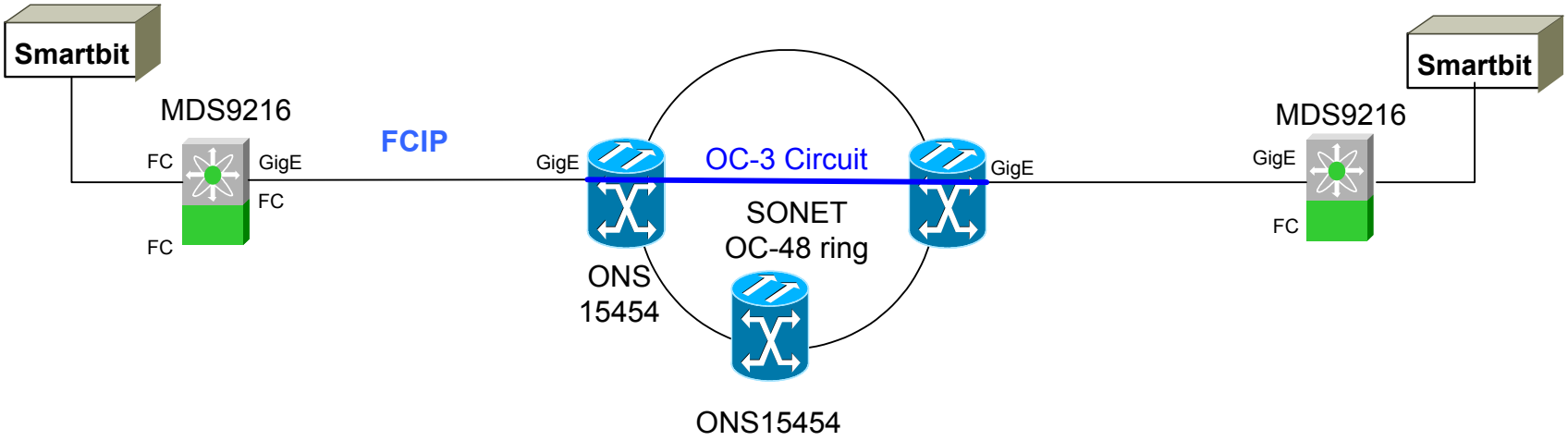






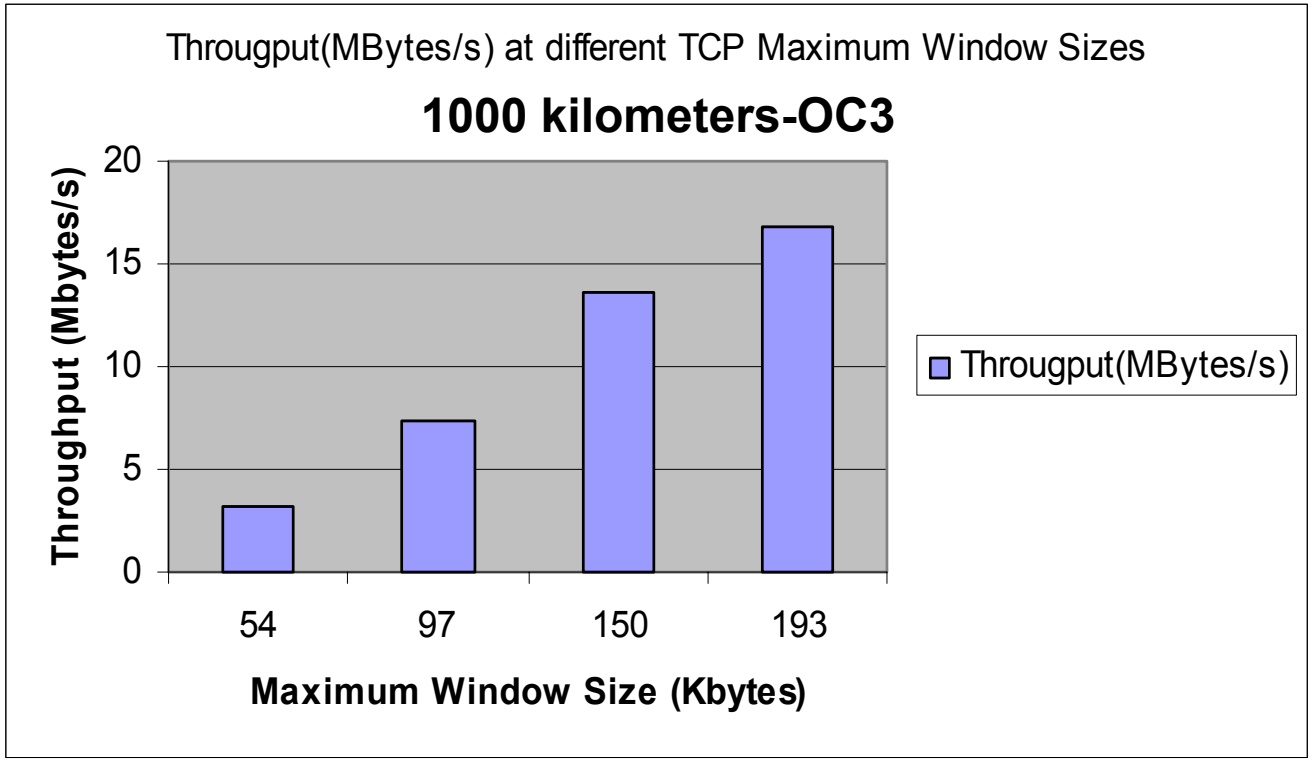
SAN Extension Case Study

SAN Extension Case Study



- ONS 15454 in an OC48 SONET ring
- OC3 circuit provisioned
- G series card on the ONS15454
- IPS card on the MDS 9216

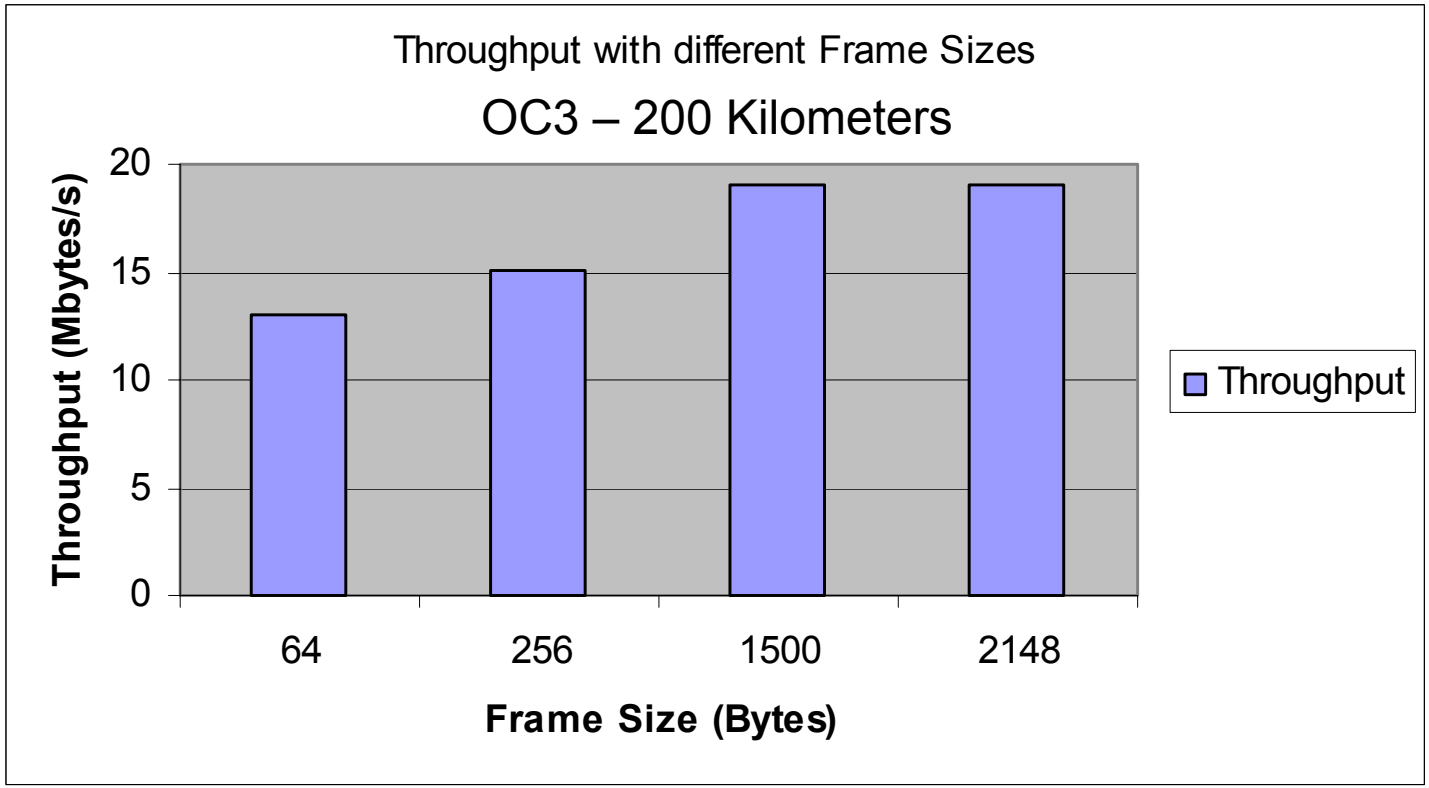
Throughput function of TCP Maximum Window Size (MWS)



OC3 link
Frame size=2148
MTU=3000
Delay introduced within the Network: 5ms
Equivalent Distance(km): 1000
Bidirectional Traffic

Optimal TCP MWS calculation:
Optimal MWS(Bytes)= Smallest rate * Round Trip delay *8
= 155000000 * 0.010 * 8
= 193 KBytes

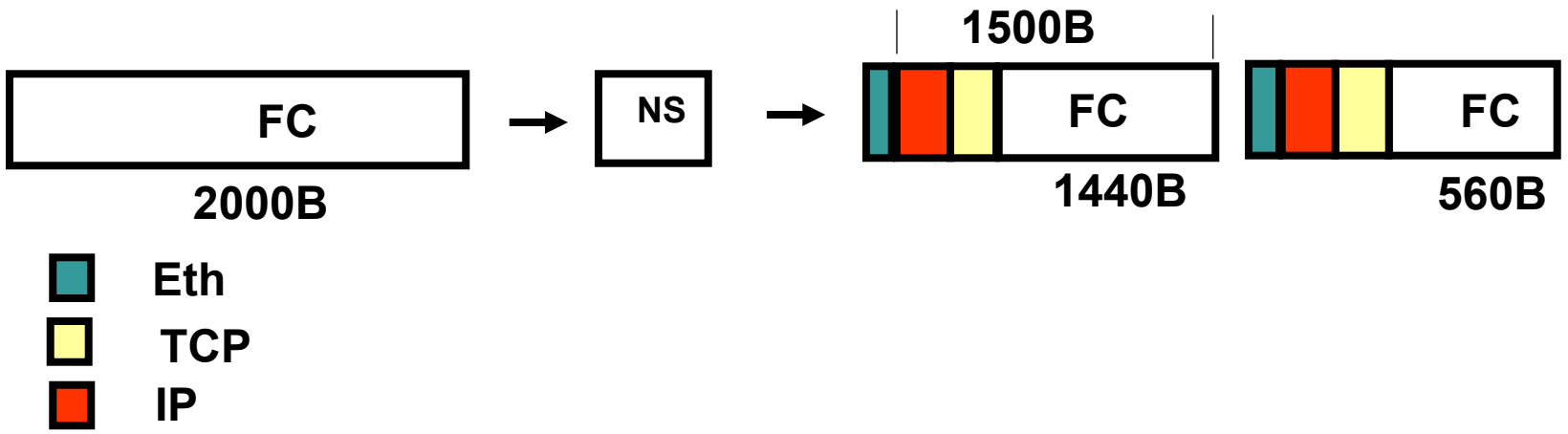
Throughput with different Frame Size



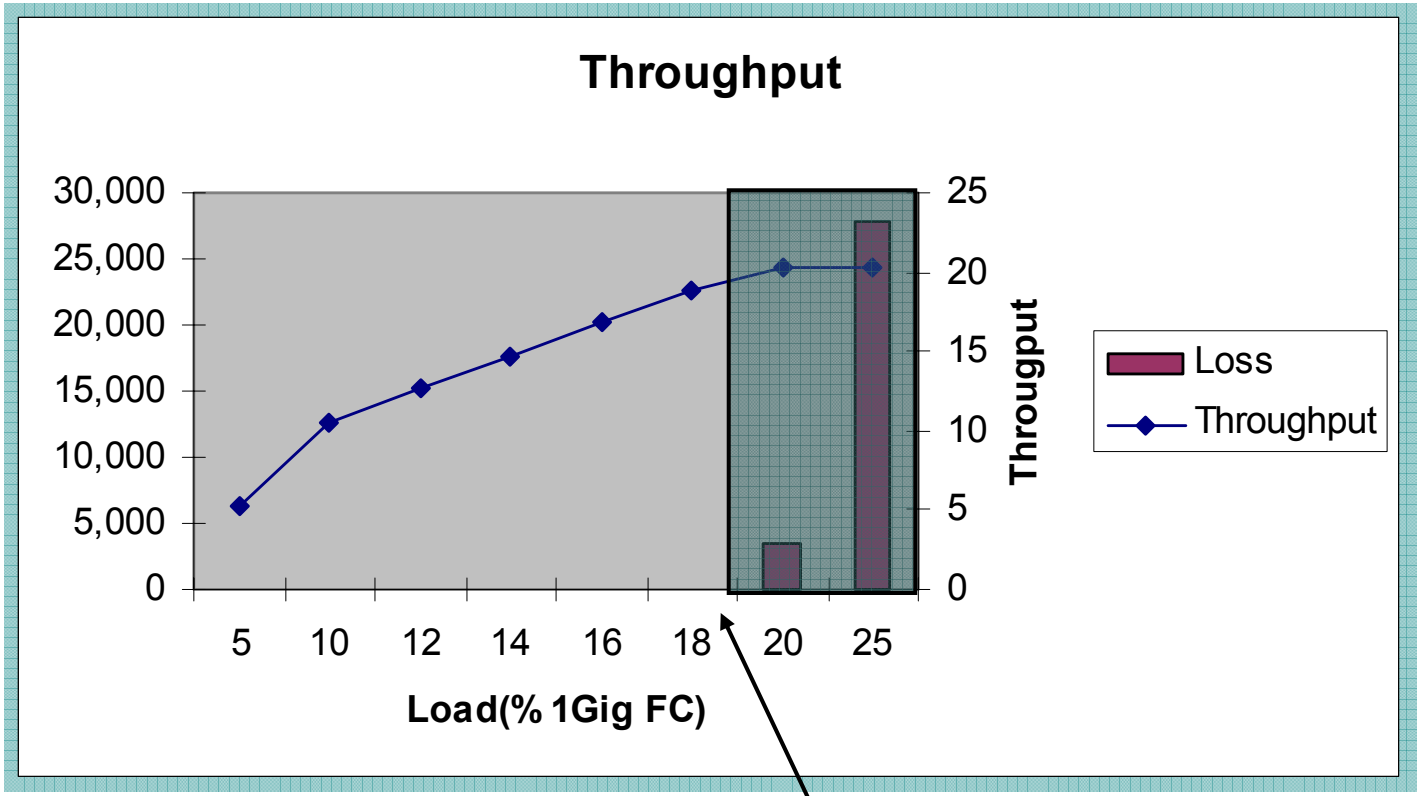
MTU: 3000
Max. window Size 58
Introduced Latency: 1 ms
Distance(km) 200

TCP Overhead

- The MSS is derived from the configured MTU value on the FCPA interface
- MSS = MTU – 60B (IP + TCP + TCP options)
- FC frame greater that MSS value will be segmented



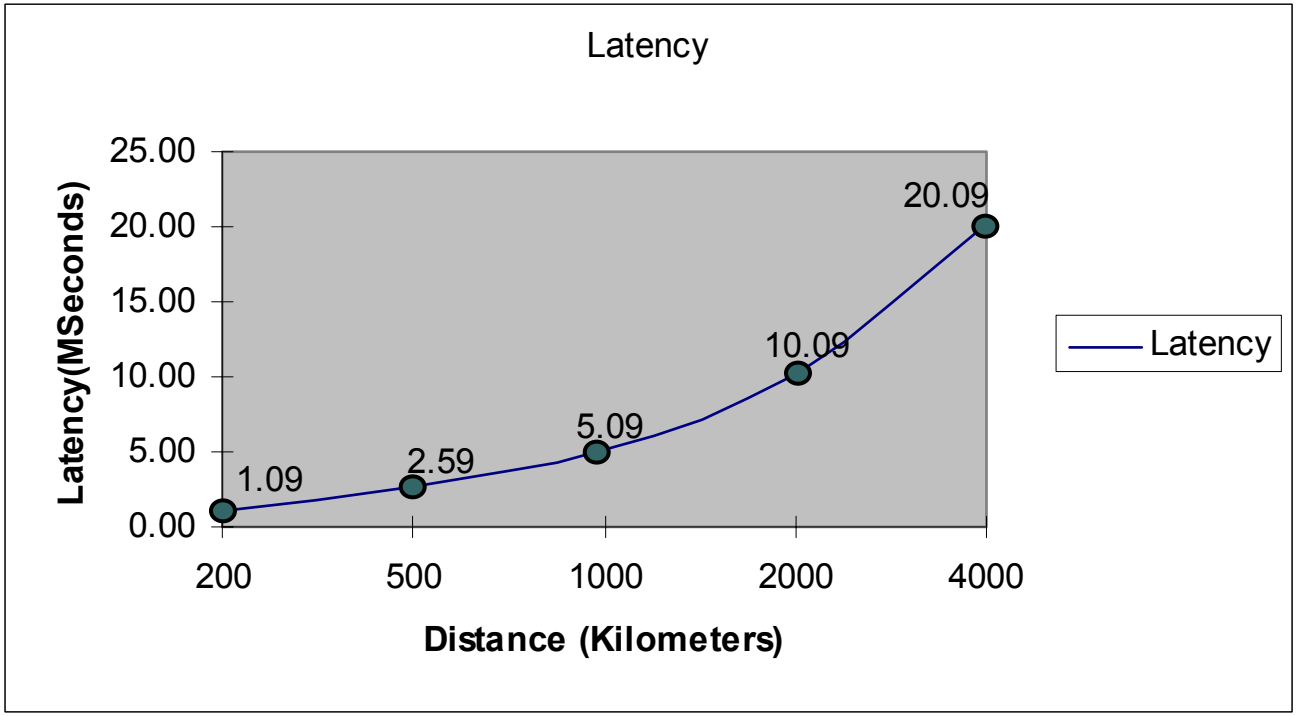
Throughput
OC3 – 200 Kilometers



Frame size=2148
MTU=3000
Maximum Window Size=39
Introduced Latency(ms)= 1
Distance(km)=200

Maximum Throughput=18Mbytes/s

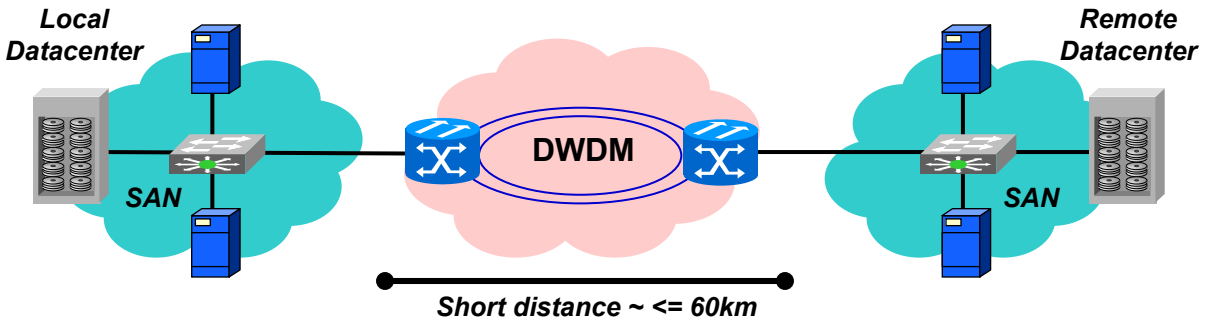
Latency versus Distance



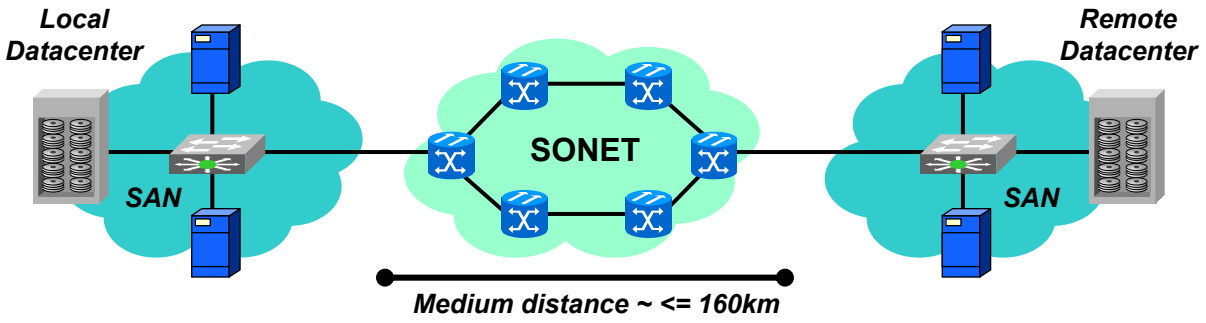
Frame size=2148
MTU=3000

Storage Extension Environments

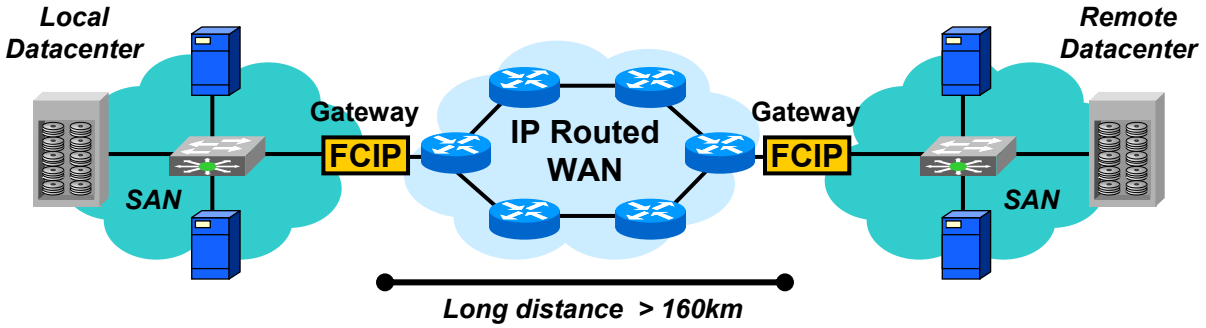
- 1Gb->2Gb or Higher
- Relatively low latency
- Synchronous apps



- Typical OC3 / OC12
- Relatively low latency
- Mainly asynchronous
- Suitable for some synchronous apps



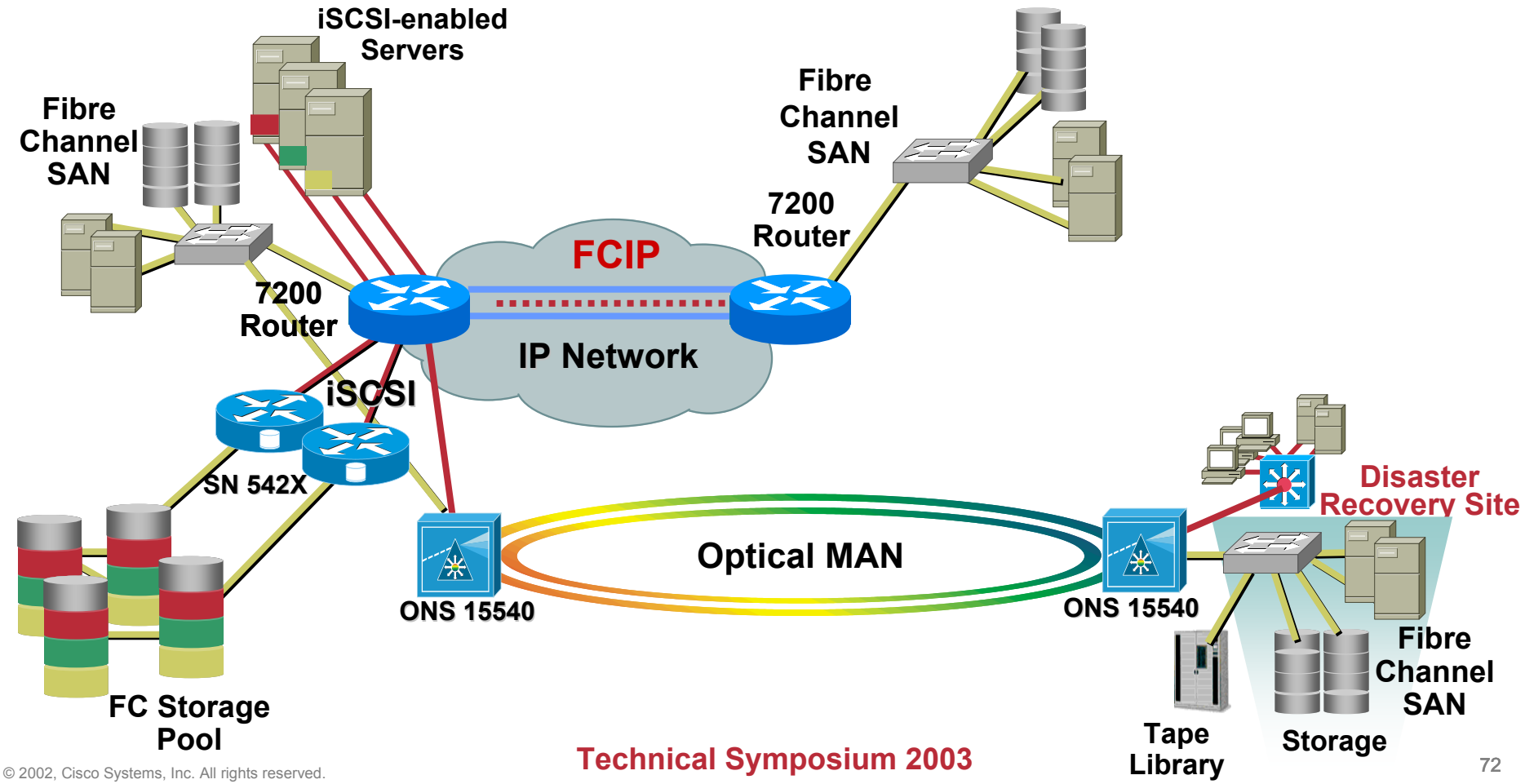
- Low speed (T1 – DS3)
- Higher latency
- Longer distance
- Mainly asynchronous



iSCSI, FCIP and Metro DWDM

iSCSI: Lowering storage TCO through increased utilization and simplified management

FCIP and Metro DWDM: Offering business continuity over the MAN and the WAN



Useful Information

- **Cisco Storage Networking**
  http://www.cisco.com/go/storagenetworking
- **Cisco AVVID Storage Networking Partner Program**
  http://www.cisco.com/warp/public/779/largeent/partner/esap/storage.html
- **Cisco Storage Router Product Information**
  http://www.cisco.com/go/storagenetworking
- **Cisco Metro Optical Product Information**
  http://www.cisco.com/go/comet
- **Storage Network Industry Association (SNIA)**
  http://wwwin.snia.org
- **Internet Engineering Task Force – IP Storage**
  http://www.ietf.org/html.charters/ips-charter.html
- **ANSI T11 – Fibre Channel**
  http://www.t11.org/index.htm