



Education

SANs across MANs and WANs

Joseph L White, Juniper Networks

- The material contained in this tutorial is copyrighted by the SNIA.
- Member companies and individuals may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced without modification
 - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA Education Committee.
- Neither the Author nor the Presenter is an attorney and nothing in this presentation is intended to be nor should be construed as legal advice or opinion. If you need legal advice or legal opinion please contact an attorney.
- The information presented herein represents the Author's personal opinion and current understanding of the issues involved. The Author, the Presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.
NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

➤ **SANs across MANs and WANs**

- ◆ Extending storage networks across distance is essential to BC/DR (Business Continuity/Disaster Recovery), compliance, and data center consolidation. This tutorial will provide both an overview of available techniques and technologies for extending storage networks into the Metro and Wide area networks and a discussion of the applications and scenarios where distance is important. Transport technologies and techniques discussed will include SONET, CWDM, DWDM, Metro Ethernet, TCP/IP, FC credit expansion, data compression, and FCP protocol optimizations (Fast Write, etc). Scenarios discussed will include disk mirroring (both synchronous and asynchronous), remote backup, and remote block access.

➤ **Learning Objectives**

- ◆ Overview of transport technologies used in Metro and Wide area networks
- ◆ Overview of protocol and transport optimizations for Metro and Wide area networks including data compression and fast write
- ◆ Overview of deployment scenarios and business drivers for extending storage networks across metro and wide are networks

Outline

- Motivation
- Basic definitions
 - ◆ SAN
 - ◆ MAN
 - ◆ WAN
- Protocols
 - ◆ SCSI
 - › FCP
 - › FCoE
 - › iSCSI
 - › FCIP
 - › iFCP
 - ◆ FICON
- Transport
 - ◆ FC
 - ◆ TCP/IP
 - ◆ Ethernet
 - ◆ WDM
 - ◆ Transport TDM (SONET/SDH)
- Effects of Distance
 - ◆ Sources of Latency
 - ◆ Performance Droop
 - › Buffers and Data
 - › Bandwidth-delay product
- Application Behavior
 - ◆ Synch vs Asynch
 - ◆ Continuous vs Snapshot/Backup
- Optimizations
 - ◆ Compression
 - ◆ Acceleration
 - › (eg 'fast write', 'tape acceleration')

Why is Distance Important?

It's about Data Protection!

▶ BC/DR

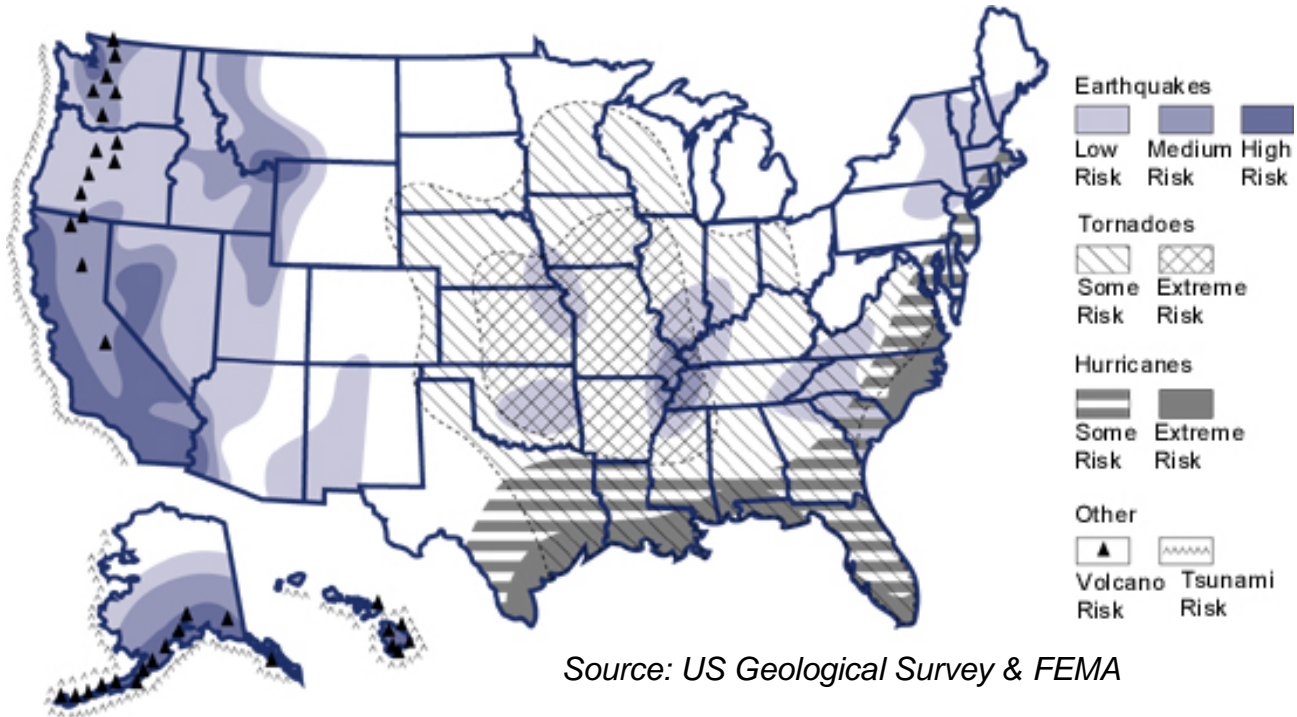
- ◆ Human
- ◆ HW/SW
- ◆ Power Outages
- ◆ Nature

▶ Business

- ◆ Consolidation
- ◆ Virtualization
- ◆ Security
- ◆ "Lost Tapes"

▶ Regulatory

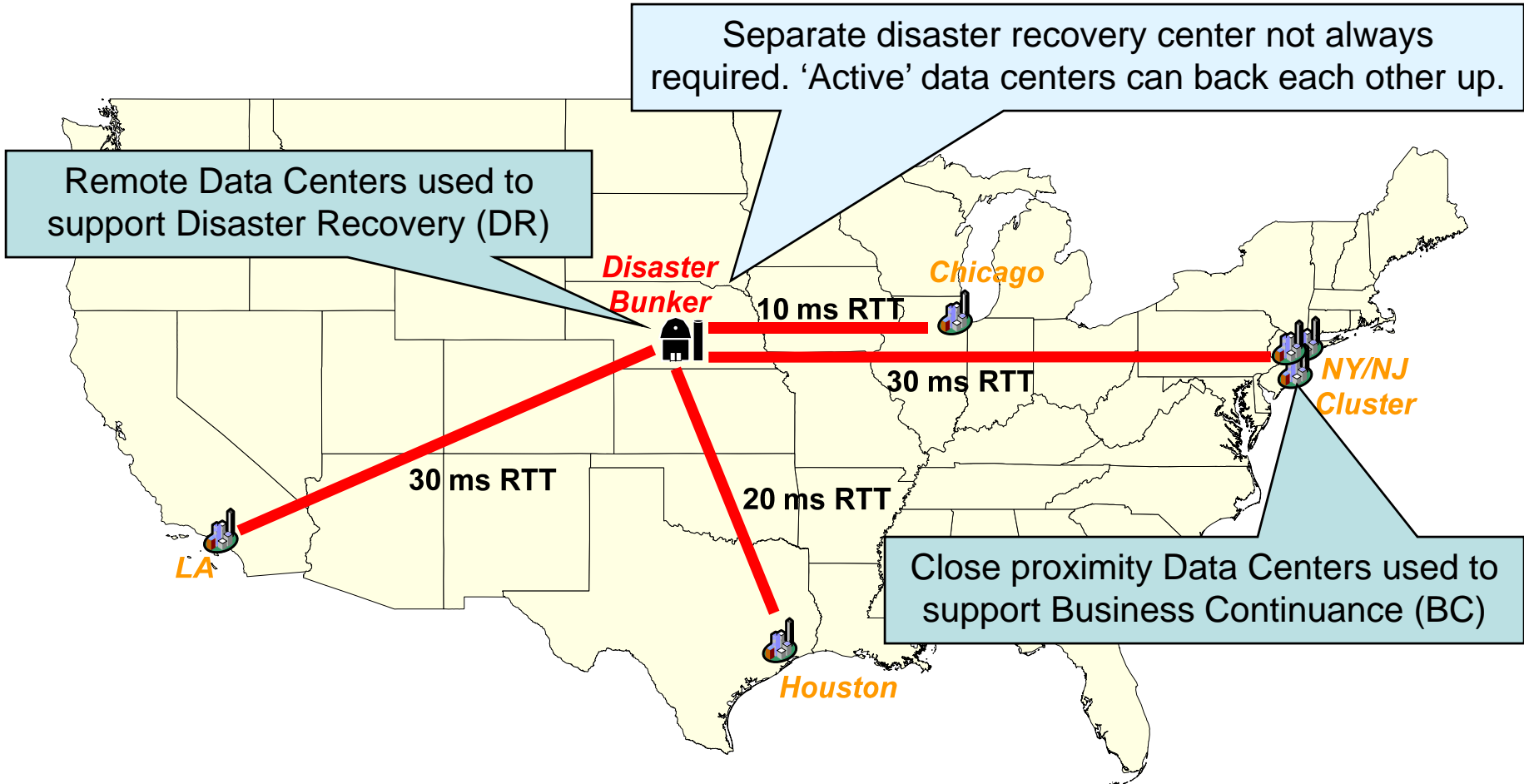
- ◆ HIPAA
- ◆ SoX
- ◆ Finance



Minimize Risk from a Single Threat Source



The WAN as seen by Storage



- Synch vs. Asynch replication applications is a separate distinction from BC/DR
- Must determine sites, distances, applications, etc by data classification and risk analysis while considering Recovery Time Objectives (RTO) or Recovery Point Objectives (RPO)

The MAN as seen by Storage



- 150-200 Km max diameter
 - ◆ Effective range of synchronous applications
 - ◆ Increasing longer range deployments (100Km+)
- Can be as short as a few 100 meters
 - ◆ i.e. to the next building
- 5-10 Km separation between sites common
 - ◆ Older installs + newer SMBs
- Long range optics
 - ◆ 40-80 km reach for direct connect
- Commonly used infrastructure
 - ◆ Direct Fibre (may have been 'Dark' previously)
 - ◆ DWDM/CWDM
 - ◆ SONET/SDH (TDM)
- FC direct connect common at shorter ranges
- FCIP comes in at longer ranges

The SAN

SANs *always* deployed full dual rail

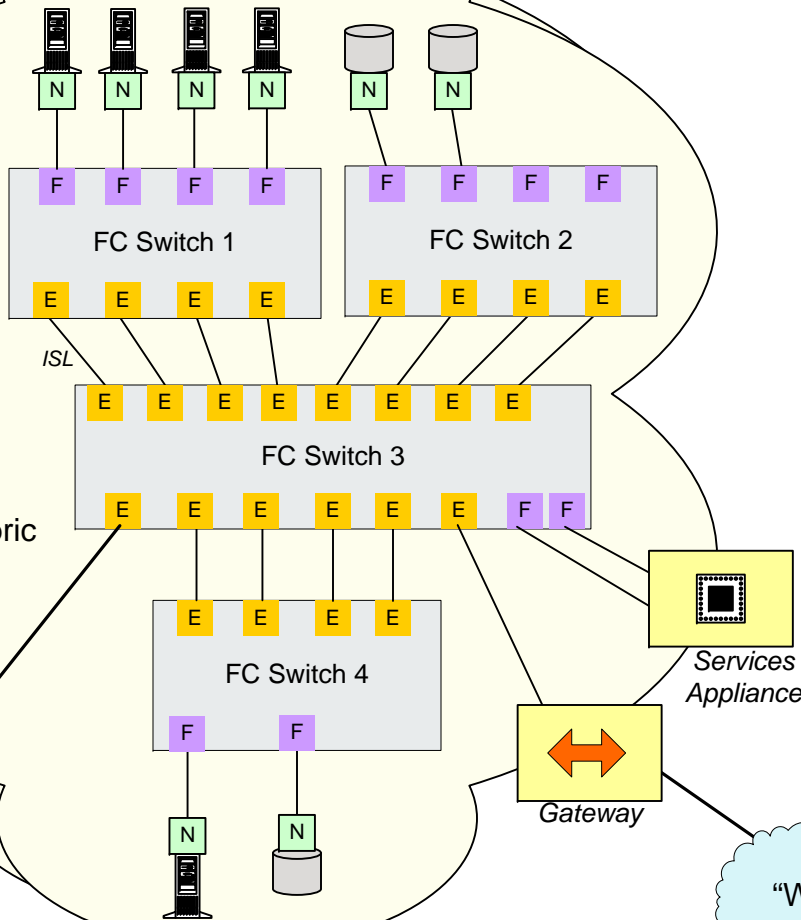
FC Services

Configuration database

Network Management

Simple FC Fabric

“MAN”

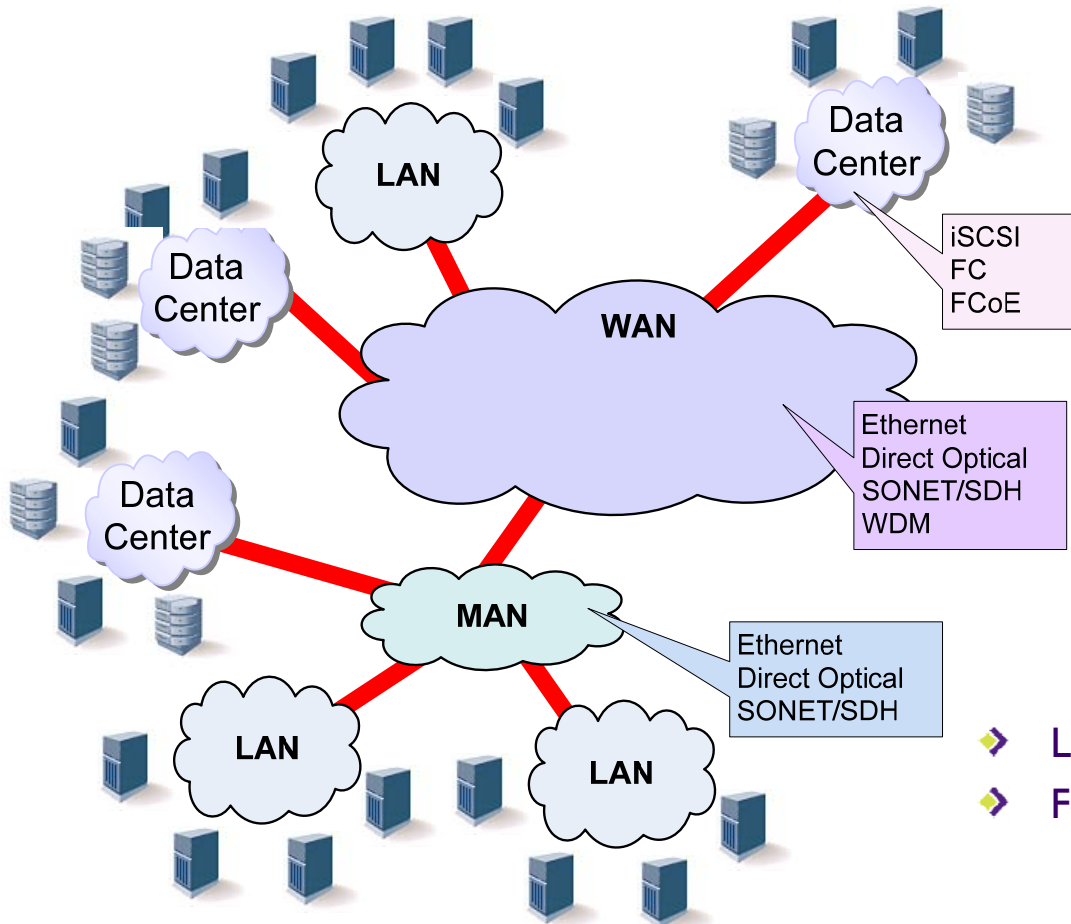


- Conventionally a collection of FC switches operating together as a Fabric supporting a set of FC services and allowing servers and storage devices (disk, tape, arrays) to communicate with each other using block protocols.
- MAN Access by direct connect
- Appliances can be attached to provide data services (block virtualization, encryption, etc)
- Gateways can be attached to provide WAN access

SANs across MANs and WANs

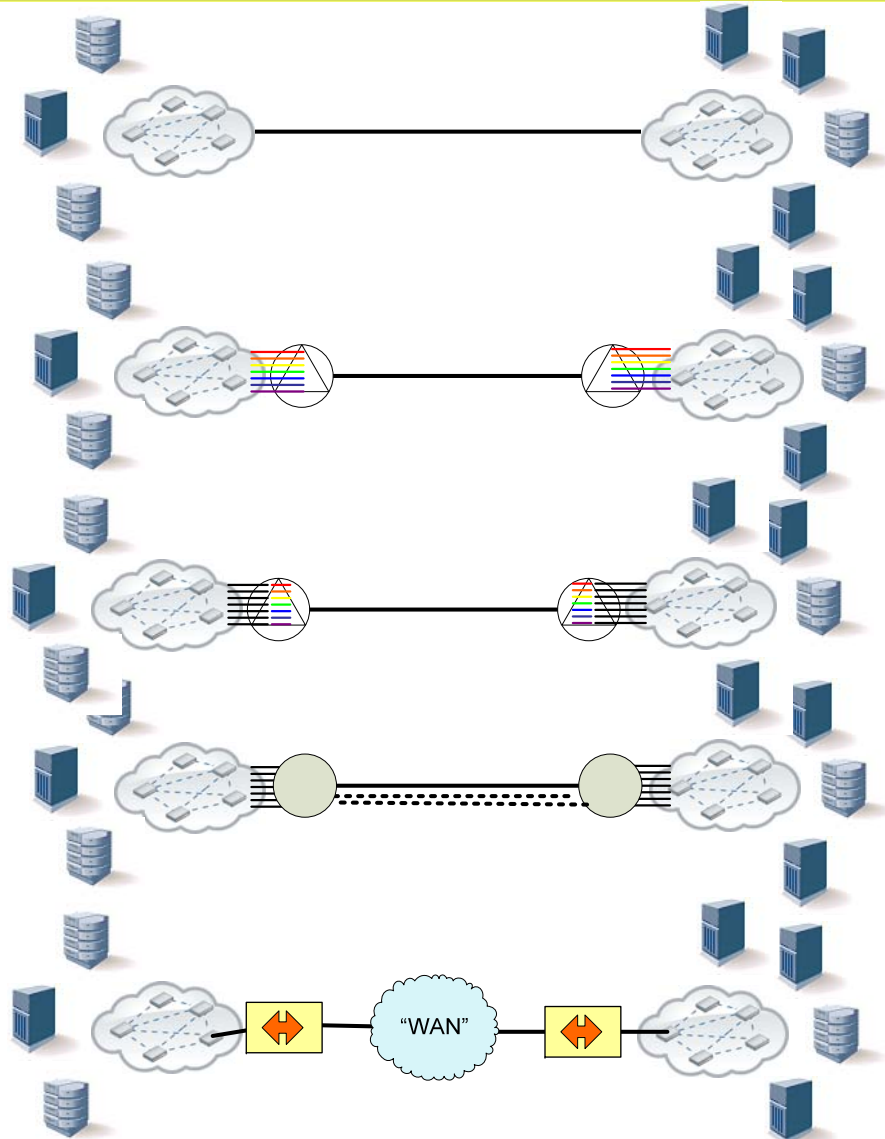
© 2008 Storage Networking Industry Association. All Rights Reserved.

Layers



- Remote Office, Central Office, Data Centers all linked
 - MAN and WAN carry already carry converged traffic
 - WAN traffic is largely TCP/IP
 - MAN traffic is mixed
 - Gateways used to connect FC SANs to MAN/WAN
 - WAN accelerators also used for optimized remote office access
 - Lots of physical interconnect options
-
- Lots of layering possible
 - For example: if talking FC we could have
 - ◆ FC over SONET/SDH
 - ◆ FC over IP over Ethernet
 - ◆ FC over native optical
 - ◆ FC over WDM
 - ◆ etc

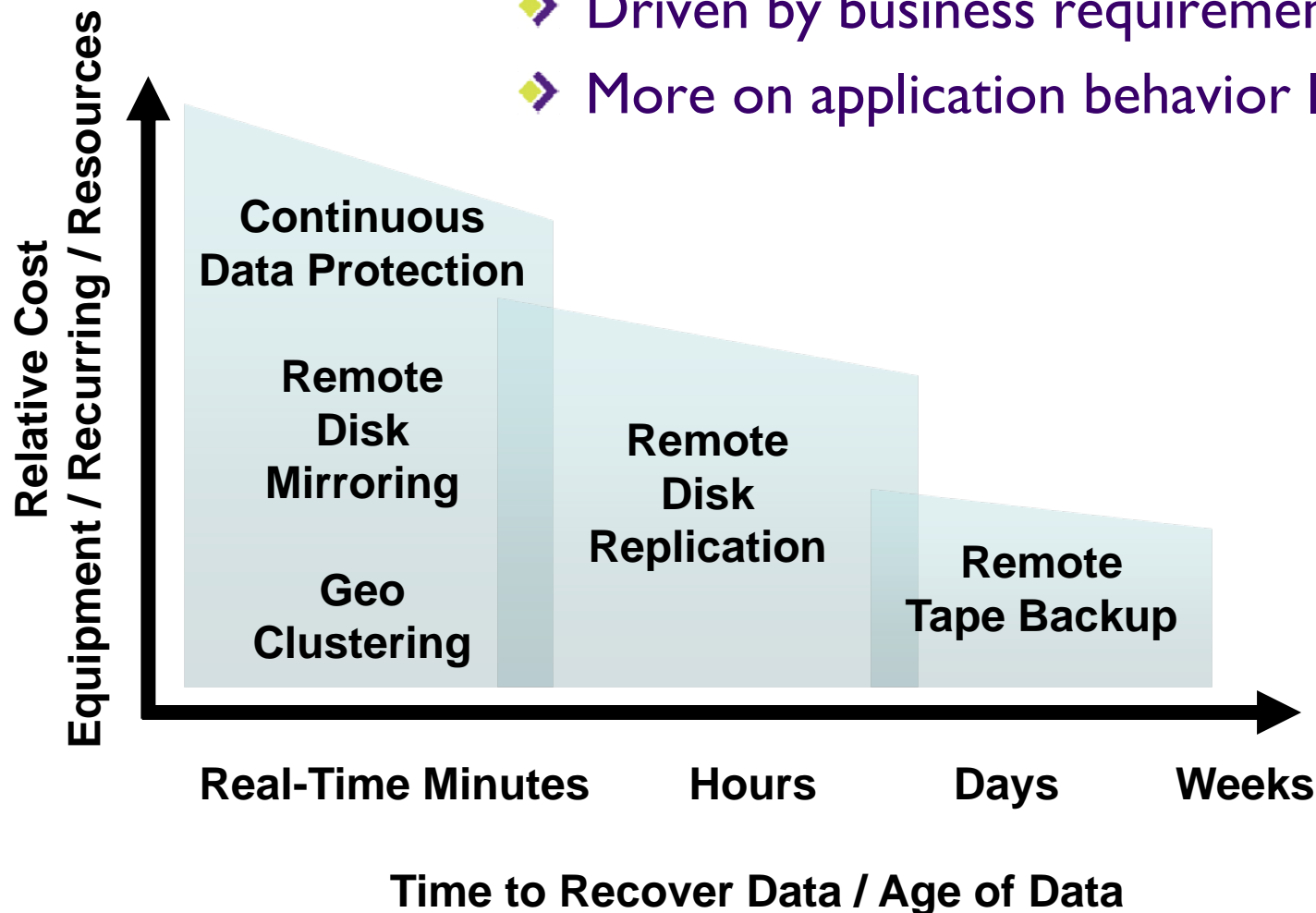
Interconnect Topology/Technology



- Direct Optical Interconnect
- WDM Interconnect 1
 - ◆ “Colored Optics”
 - ◆ External box is mux only
- WDM Interconnect 2
 - ◆ Native interface locally
 - ◆ External box does wavelength shifting
- TDM Interconnect
 - ◆ Bit level protocol dependencies (inter-frame gap etc)
- Gateway Interconnect across other “WAN” infrastructure
 - ◆ FC and above dependencies

protocol agnostic...

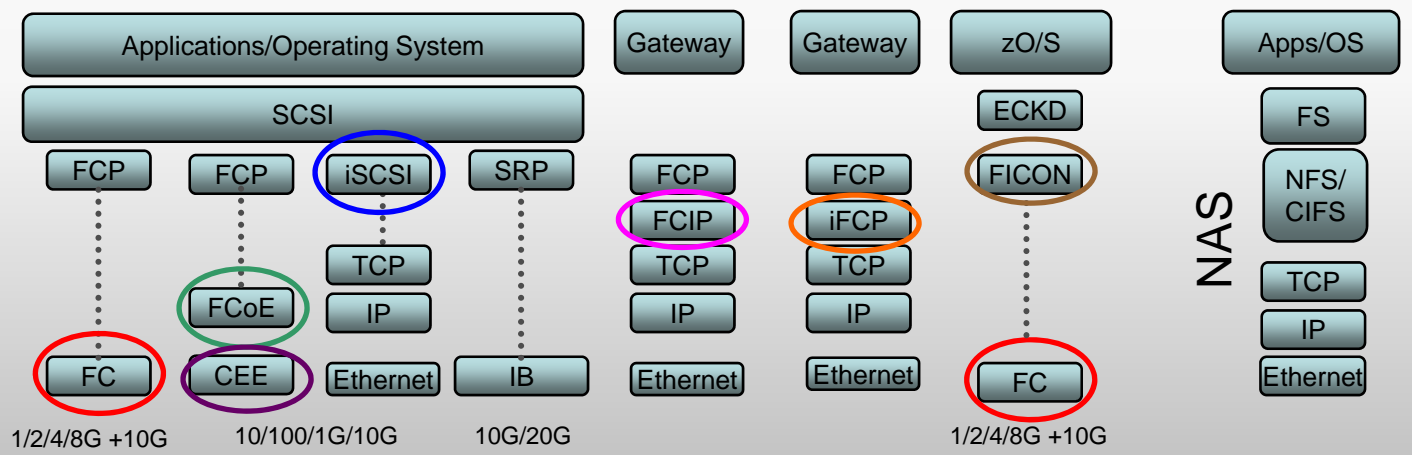
- Driven by business requirements
- More on application behavior later...



Storage Networking Protocols

'SCSI' is the protocol and command set for 'block' storage access.

Multiple SCSI Transports



FC

- No-drop, credit flow control
- Fabric Services
- Switched Network

FCoE

- Transport FC over Ethernet, while maintaining FC operational model
- Consolidate I/O for SAN, LAN fabrics

CEE

- Converged Enhanced Ethernet
- Makes Ethernet directly suitable for storage traffic

iSCSI

- Direct Transport of SCSI over TCP/IP
- Fabric Services (iSNS)
- Runs over existing infrastructure

FCIP

- Tunnels FC using TCP/IP
- Mainly a long-distance solution
- Interconnects FC infrastructure into distributed SANs

iFCP

- Interconnects FC devices across an IP network
- Local FC infrastructure isolated from remote infrastructure

FICON

- Transport of ECKD across FC infrastructure
- IBM Mainframe
- Replaced ESCON

Infiniband

- Mostly HPC environment

NAS

- File access semantics (instead of block)
- Shown here for context

Local FCP: FC, FCoE

- FCP is the serialization of SCSI commands across Fibre Channel transport
 - ◆ Uses the FC exchange, sequence, frame structures
 - ◆ Maps SCSI task management to FC constructs
 - ◆ Generally the term FC applies to FCP/FC plus the FC Fabric Services
- FCoE refers to replacing the FC-I and FC-0 layers with Ethernet as a transport
 - ◆ Realistically CEE – Converged Enhanced Ethernet



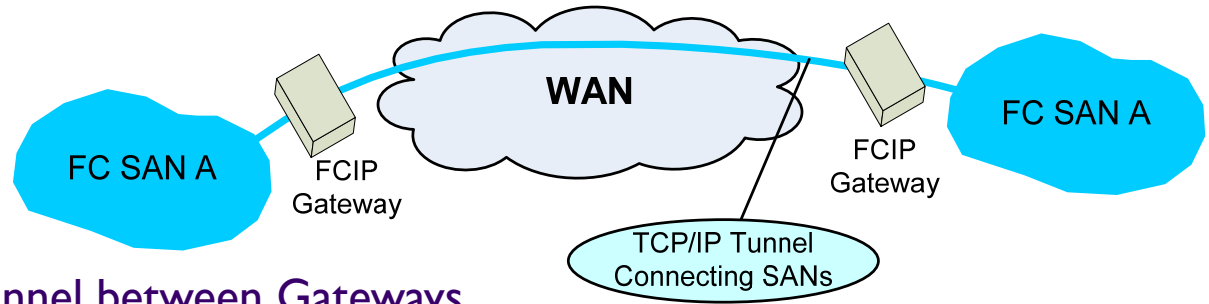
Check out SNIA Tutorial:

Fibre Channel Over Ethernet (FCoE)

Distance FCP: FCIP, iFCP

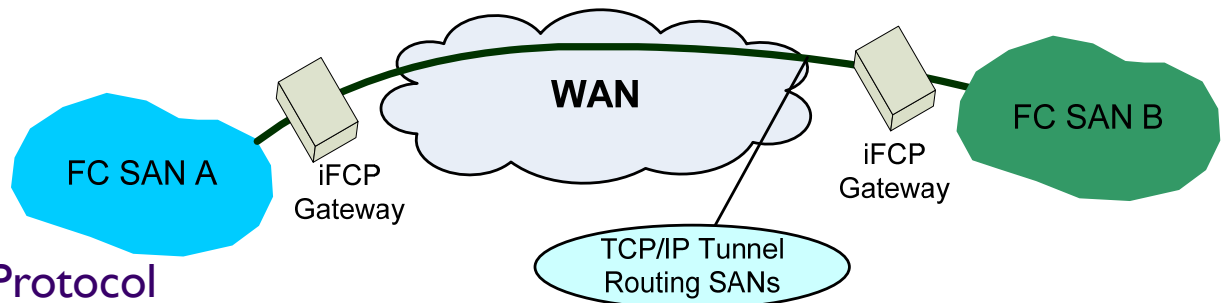
➤ FCIP

- ◆ FC over IP
- ◆ FC run across TCP/IP tunnel between Gateways
- ◆ Connects FC SAN segments into one SAN
- ◆ FC devices & fabric services used as-is
- ◆ *SAN routing can be used to isolate FC fabrics just like for iFCP*

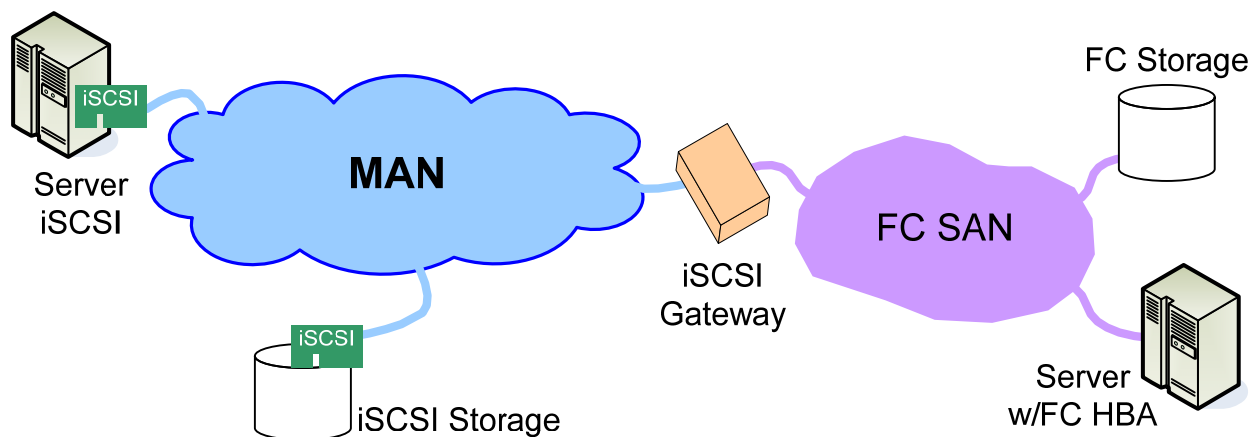


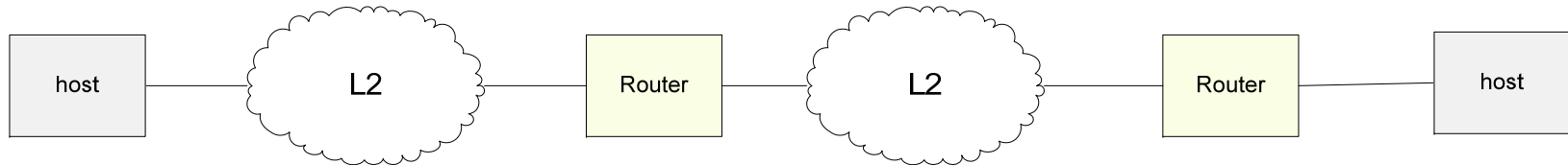
➤ iFCP

- ◆ Internet-Fibre Channel Protocol
- ◆ FCP over TCP/IP
- ◆ Provides isolation of local FC SANs
- ◆ In practice used like FCIP
 - › Native iFCP devices would be allowed but none were implemented



- iSCSI directly implements a SAN across an IP network using TCP/IP
 - ◆ Traditionally for SME or SMB market
 - ◆ A iSCSI-FC gateway can be used to access native FC devices
 - FC Storage, FC Initiator, iSCSI Initiator, iSCSI Storage in any combination
 - Usual deployment is FC Storage and iSCSI storage accessed by iSCSI Servers
 - ◆ iSNS provides fabric services (name, zone, config, SCN)
 - ◆ Servers can contain a HBA, TOE, or standard NIC
 - ◆ Direct access across local and metro distances
 - ◆ Could access WAN devices since IP is a fully routed protocol but most implementations would suffer significant performance degradation

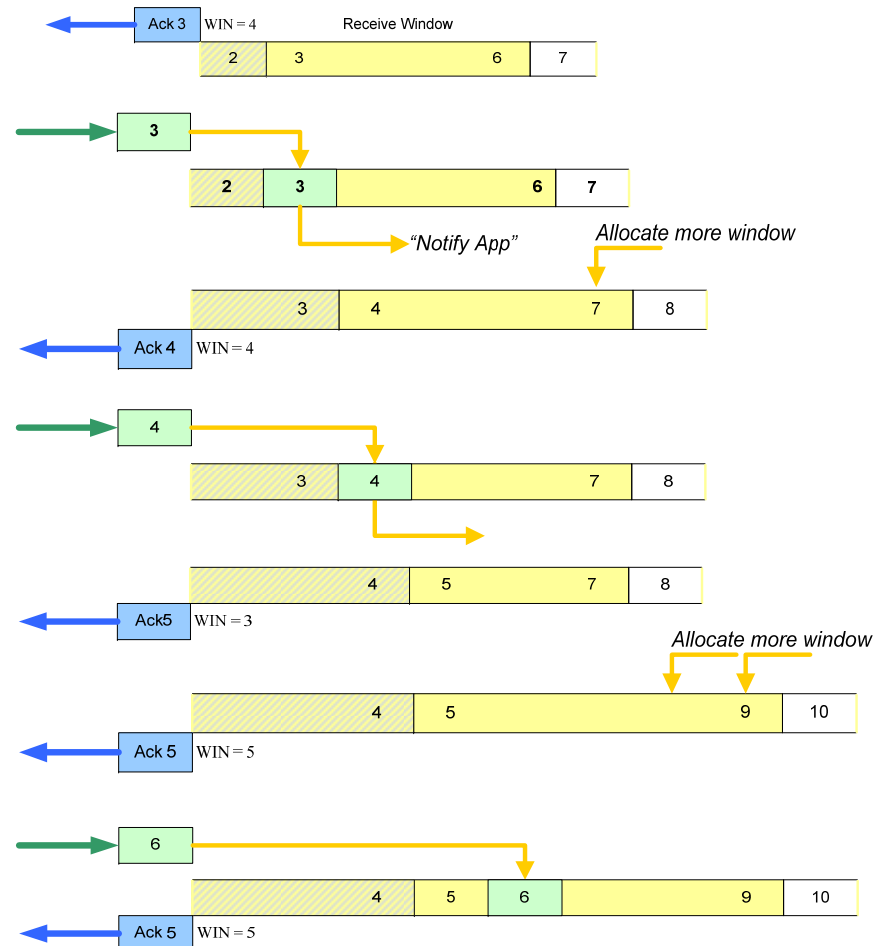
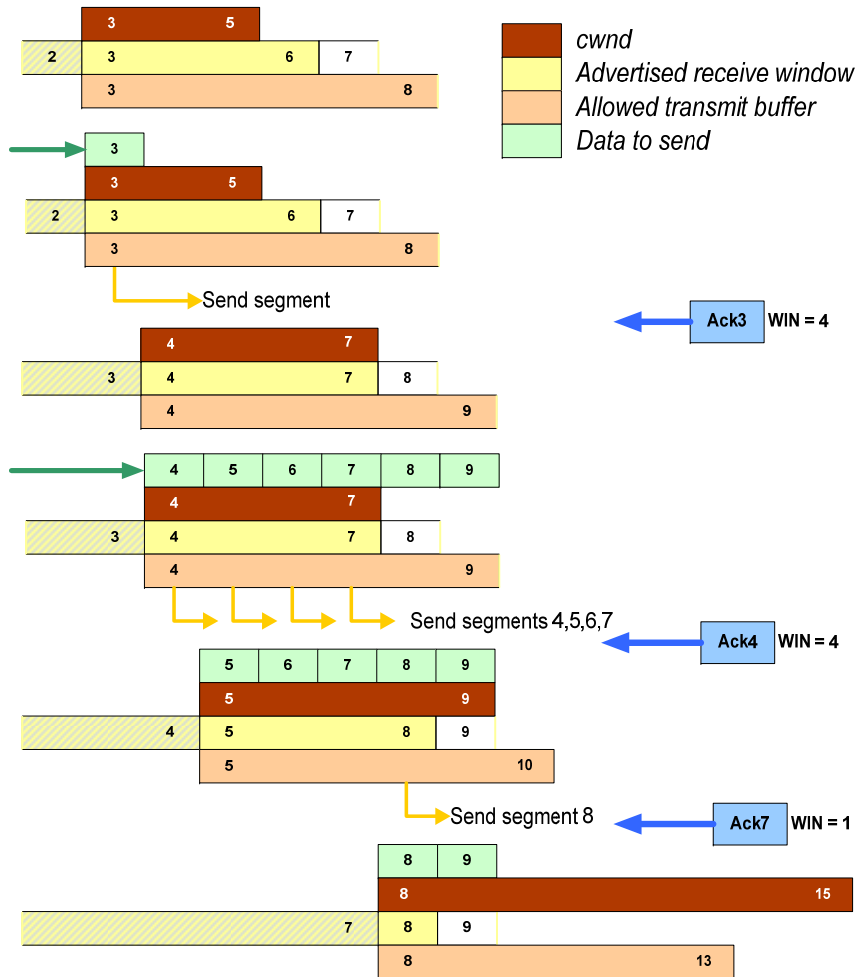




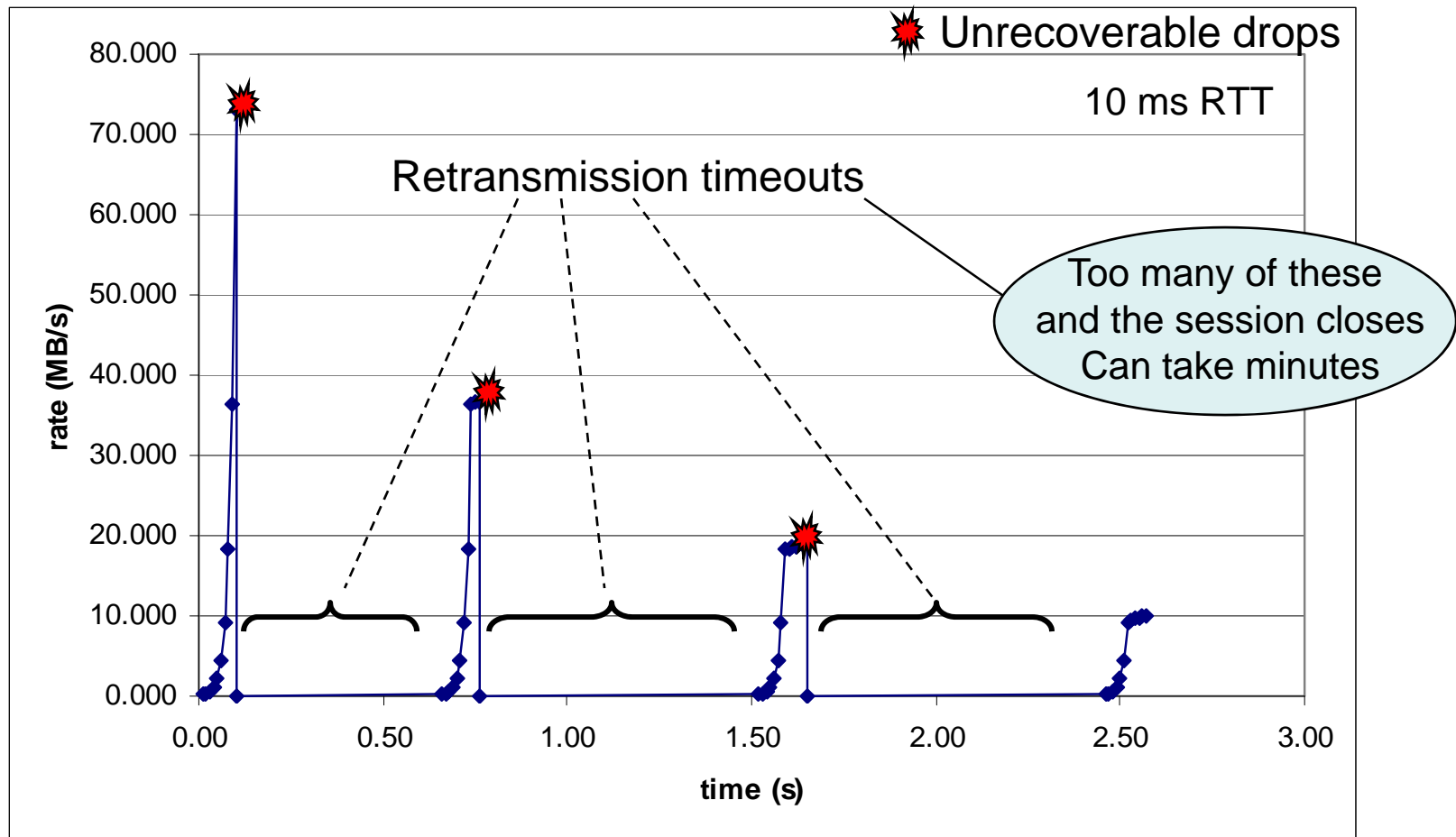
- For LAN IP rides over Ethernet to *hop* across the network
- For MAN/WAN IP *rides* over SONET/SDH, WDM
 - ◆ OR over native Ethernet that is in turn carried over SONET/SDH, WDM
- IP and Ethernet generally carries TCP or UDP traffic
 - ◆ Under I/O Convergence Ethernet also carries Storage traffic
- Ethernet + IP (will also apply to FCoE)
 - ◆ Well understood and accepted in IT world
 - ◆ Low service cost points for *best-effort* services
 - › Short-term bursty, file-based, small “packets”, connectionless
 - › Congestion common, retransmits, variable/high latency
 - ◆ Services available
 - › Ethernet Private Line, MPLS, RPR, Carrier Ethernet

- For WAN networking TCP is Critical (FCIP, iSCSI, iFCP)
- Connection Oriented
 - ◆ Full Duplex
 - ◆ Byte Stream (to the application)
 - ◆ Port Numbers identify application/service endpoints within an IP address
 - ◆ Connection Identification: IP Address pair + Port Number pair ('4-tuple')
 - ◆ Well known port numbers for some services
 - ◆ Reliable connection open and close
 - ◆ Capabilities negotiated at connection initialization (TCP Options)
- Reliable
 - ◆ **Guaranteed In-Order Delivery**
 - ◆ Segments carry sequence and acknowledgement information
 - ◆ Sender keeps data until received
 - ◆ Sender times out and retransmits when needed
 - ◆ Segments protected by checksum
- Flow Control and Congestion Avoidance
 - ◆ **Flow control is end to end (NOT port to port over a single link)**
 - ◆ Sender Congestion Window
 - ◆ Receiver Sliding Window

TCP Send/Receive Illustration



TCP Retransmission Timeout

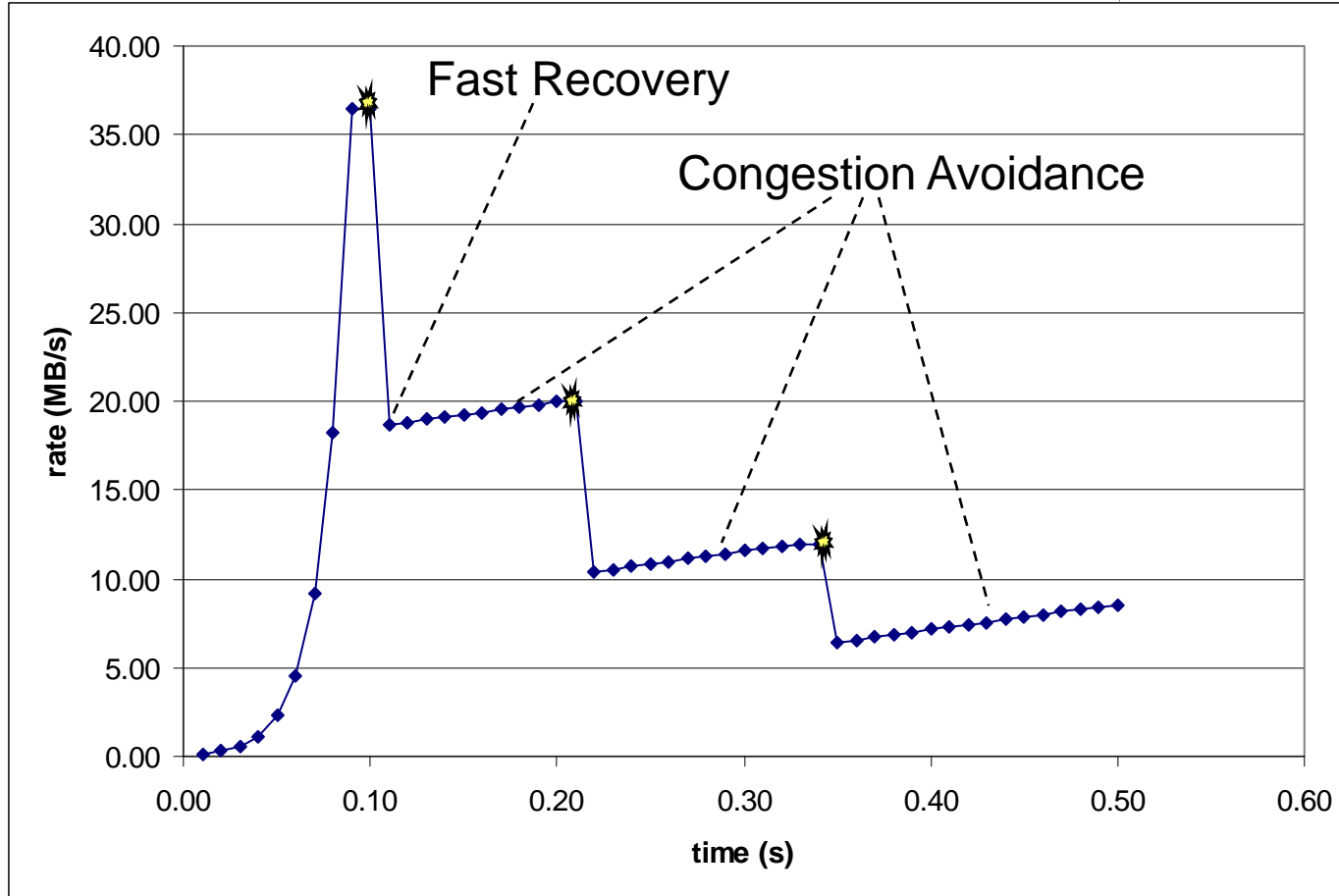


- time oldest sent, unacknowledged data
- Requires RTT estimation for connection (typically 500 ms resolution TCP clock)
- Retransmission timeouts are 500 ms to 1 s with exponential back-off as more timeouts occur

TCP Fast Retransmit, Fast Recovery

10 ms RTT

☀ Packet drop

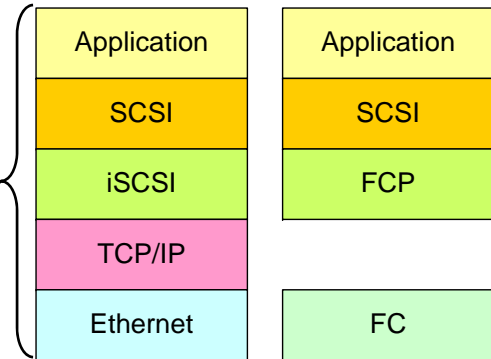


- ◆ Dropped frames can be detected by looking for duplicate ACKs
- ◆ 3 dup ACKs frames triggers Fast Retransmit and Fast Recovery
- ◆ With Fast Retransmit there is no retransmission timeout.

- Scaled receive windows
- Quick Start
- Modify Congestion Controls
- Deal with network reordering
- Detect retransmission timeouts faster
- Implement Selective Acknowledgement (SACK)
- Reduce the amount of data transferred (compression)
- Aggregate multiple TCP/IP sessions together
- Bandwidth Management, Rate Limiting, Traffic Shaping

TCP/IP Summary

- TCP/IP is both good and bad for block storage traffic
- TCP/IP's fundamental characteristics are good
 - ◆ Connection oriented
 - ◆ full duplex
 - ◆ guaranteed in-order delivery
 - ◆ Basic latency is not significant when compared to native FC



- TCP/IP's congestion controls and lost segment recovery can cause problems for block storage
 - ◆ Large latencies CAN occur when drops are happening (this is bad)
- However, Many of TCP/IP drawbacks can be mitigated
 - ◆ Some changes only improve TCP behavior
 - › For example better resolution TCP timers leading to more precise
 - › Or SACK
 - ◆ Some have a possible negative effect on other traffic
 - › For example removing congestion avoidance completely

Physical Interconnects

- FC
- Ethernet
- Protocol Agnostic
 - ◆ WDM (Wave Division Multiplexing)
 - > CWDM, DWDM
 - ◆ TDM
 - > SONET/SDH
 - ◆ ATM and legacy
 - ◆ dark fiber
 - > *Optical fiber in place but not used (i.e. unlit)*

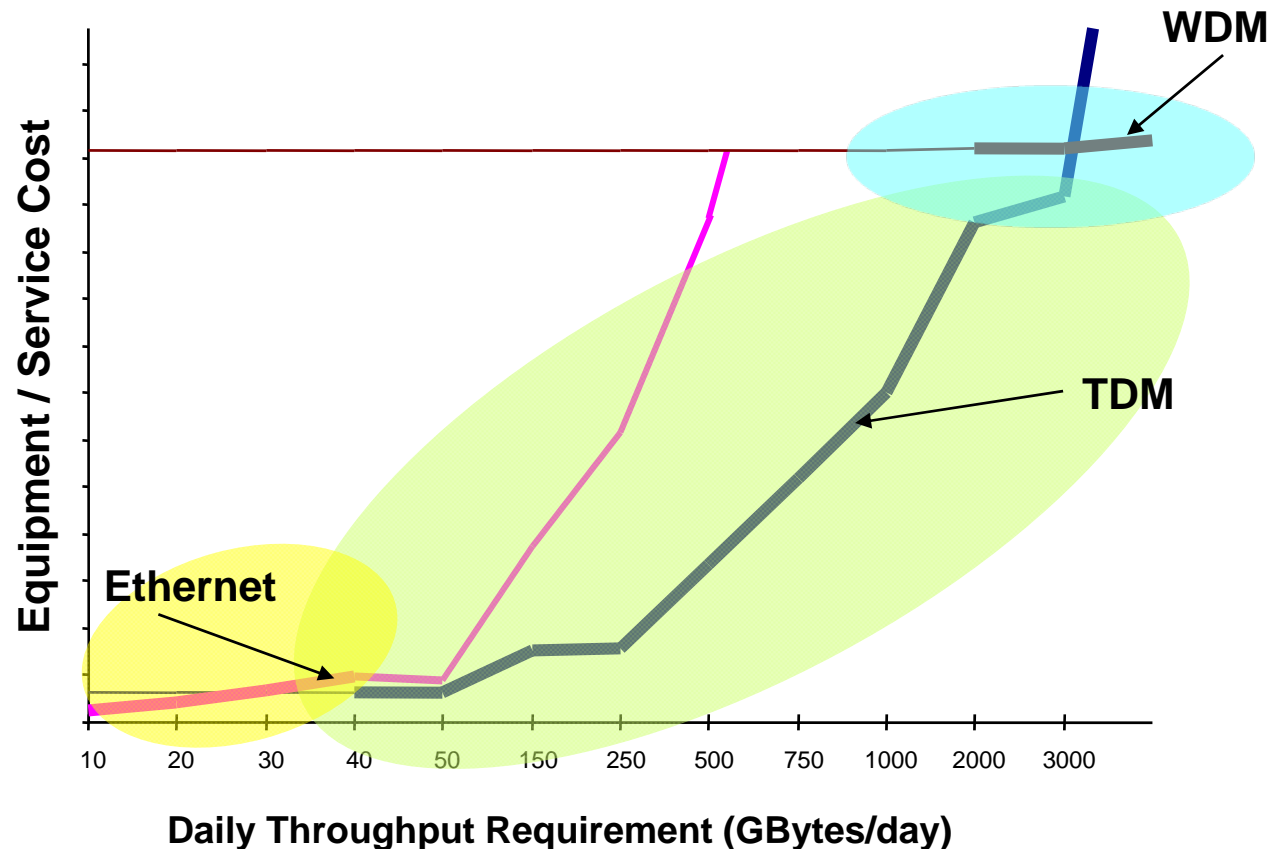
MAN/WAN Transport Options

Understand your actual throughput needs:

Changed data size ÷ by backup window = data rate

Many considerations:

- **Application**
- **Performance**
- **Latency**
- **Bandwidth**
- **Security**
- **Protection**
- **Distance**
- **Availability**
- **Cost**



Fibre Channel

- Switched network protocol
 - ◆ Loop can still apply to disk array internal interconnects
- 1/2/4/8G + 10G speeds
- Provides transport with credit based link level flow control
 - ◆ A credit corresponds to 1 frame independent of size
 - ◆ Amount of credit supported by a port with average frame size taken into account determines maximum distance that can be traversed
- Switches organized into fabrics with Fabric Services
- FC has fabric and services related frames (Basic and Extended Link Services) in addition to transporting FCP or FICON
 - ◆ FC can transport other protocols including IP but this is not generally done



Check out SNIA Tutorial:

Fibre Channel Technologies: Current and Future

Ethernet

- Layer 2 interconnect
- 10/100/1000 (1GE) /10000 (10GE)
- Carries
 - ◆ IP traffic (TCP, UDP)
 - ◆ FCoE

Protocol Features

802.3x: Flow Control (PAUSE)
802.1d/802.1w: STP/RSTP
802.3ad: Link Aggregation
802.1p: Class of Service
802.1q: VLAN

- Pause frames and distance...
 - ◆ When the sender needs to be stopped the receiver sends a frame to notify the sender
... If the buffer is overrun then frames can be dropped
 - ◆ This puts a hard limit on the distance for storage traffic
 - › Unlike the case for FC using BB Credits
- Pause can also cause extensive congestion spreading

CEE: Converged Enhanced Ethernet

- Expand Ethernet such that it is better suited to converged networks
 - ◆ Proposals to present to TII and IEEE and IETF
 - ◆ Owed by IEEE and IETF (See standards organizations web pages for details)



Check out SNIA Tutorial:

Ethernet Enhancements for Storage: Deploying FCoE

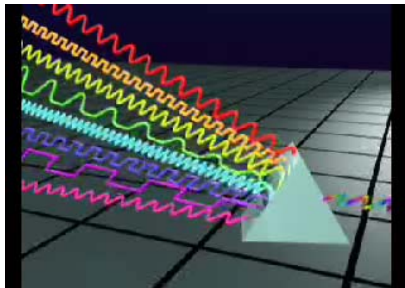
- Priority-based flow Control (PFC) 802.1Qbb
 - ◆ Provides no packet-drop behavior
- Enhanced Transmission Selection (ETS) 802.1Qaz
 - ◆ Multiple priority groups with bandwidth guarantees. Strict priority.
- Data Center Bridging Exchange protocol (DCBX)
 - ◆ Uses LLDP (802.1AB) to advertise connectivity and management information between two link peers
- Congestion Management (CM) 802.1Qau
 - ◆ Provides link level congestion management and notification
- TRILL (IETF)
 - ◆ Transparent Interconnect of Lots of Links – allows L2 multipath
- Shortest Path Bridging 802.1aq
 - ◆ Eliminate spanning tree for L2 topologies and allows L2 multipath

Wavelength Division Multiplexing

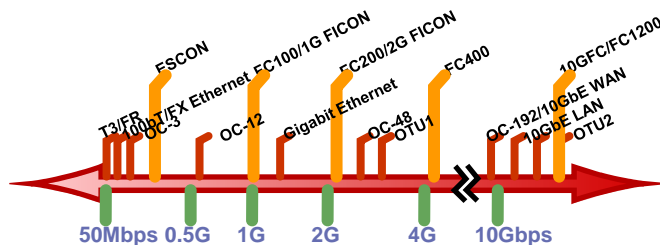


© New Line Productions, Inc

Multiple Lasers each shooting light of a particular wavelength through a single fiber allow multiple streams of data traffic to be carried simultaneously!

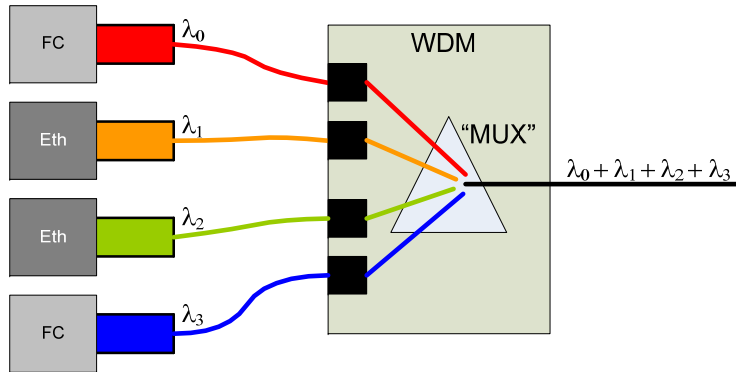


Prisms or their electronic equivalent combine and split the light at each end of the long haul optical link.

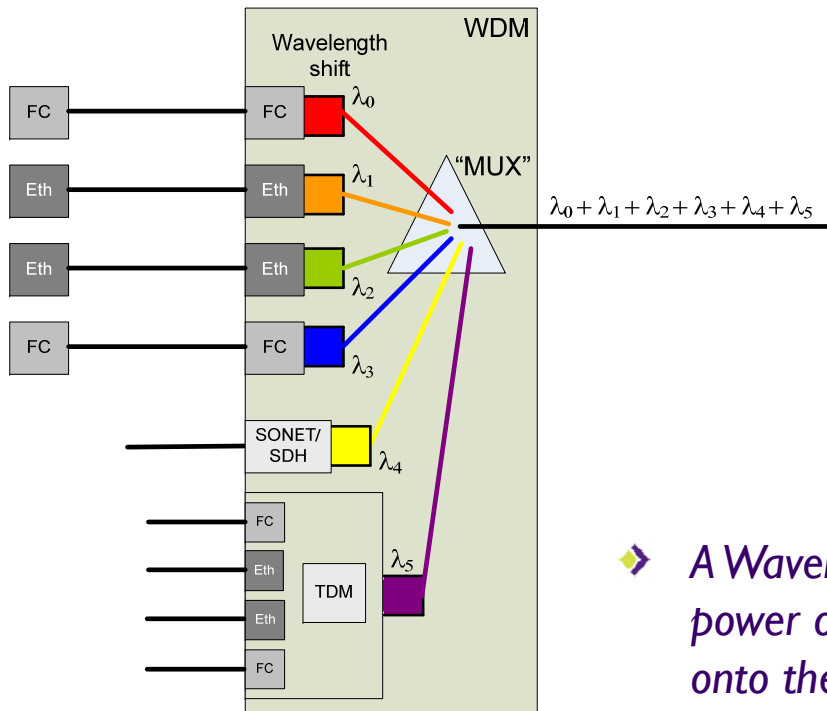


...with each wavelength carrying up to an input connection of “full-rate” throughput...

WDM Infrastructure



- Colored optics inserted into device
- WDM combines light
- MUX is prism or electronic



- Standard interface used in device
- WDM 'shifts' wavelength
- MUX still combined signals
- Input can also be TDM
- Can have multi-input TDM card to put several standard interfaces onto single wavelength.

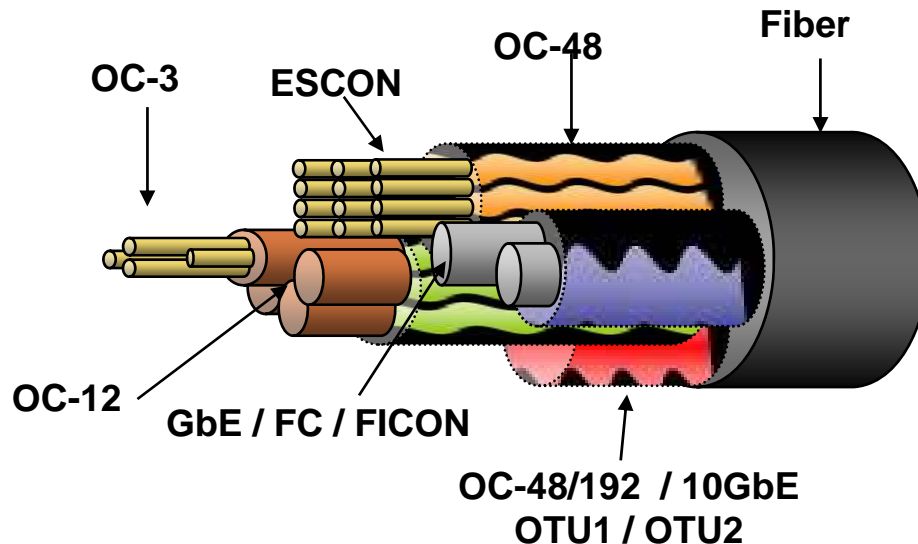
➤ *A Wavelength or 'lambda' is really tight range. Resolving power of equipment determines how many lambdas 'fit' onto the fiber*

DWDM: Dense WDM

- 8-40+ waves per fiber
- 500mile reach with amplification
- 2.5Gbps & 10Gbps common
- Optical protection
- Optics experience needed

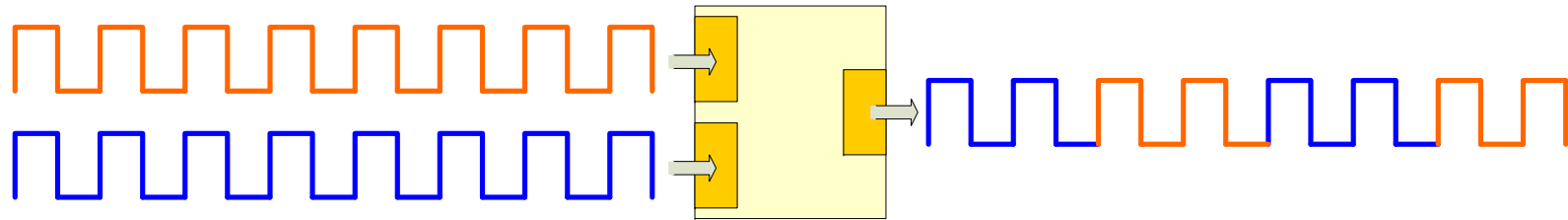
CWDM: Coarse WDM

- 4-8 waves per fiber
- 50mile reach
- 2.5Gbps
- Optical protection
- Lower cost with passive optics



Each wavelength (aka lambda) can utilize its full bandwidth capacity for multiple services

TDM – Time Division Multiplexing SONET/SDH (OC-1+/T1+/E1+/DS1+/etc)



- Well established and widely available
- Any distance support from Metro to Wide area
- Connection based with predictable low latency
- Highly reliable with path protection
- SDH is the international equivalent of SONET
- Some extension gateways have direct SONET/SDH interfaces
- Used to aggregate slower traffic onto faster links
 - This applies to combining ‘fast’ into ‘superfast’ links for example stretched data centers across metro distances.

Latency

- Command Completion Time is important

- Contributing Factors: (sum 'em all up!)
 - ◆ Distance (due to 'speed of light') - latency of the cables
 - › (2×10^8 m/s gives 1 ms RTT per 100Km separation)
 - ◆ 'Hops' – latency through the intermediate devices
 - ◆ Queuing delays due to congestion
 - ◆ Protocol handshake overheads
 - ◆ Target response time
 - ◆ Initiator response time

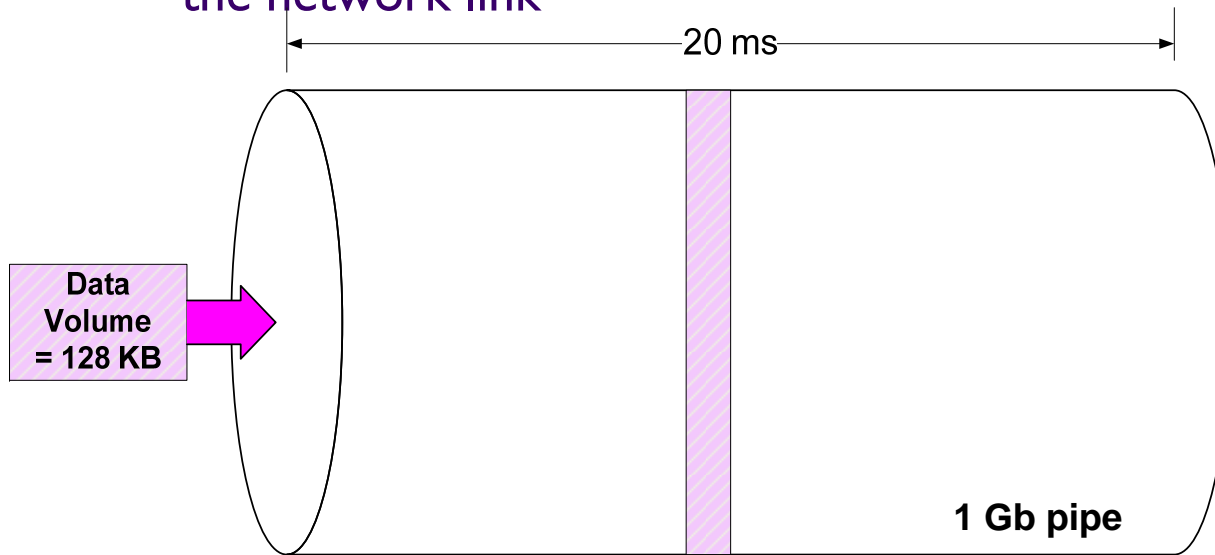
Performance Droop

- Many of sources of performance droop
- Transport Buffer relative to bandwidth delay product
 - ◆ Number of credits
 - ◆ TCP transmit and receive buffering
- Available Data relative to bandwidth delay product
 - ◆ Outstanding commands
 - ◆ Command request size
 - ◆ i.e. must do bandwidth delay at each protocol level
- Protocol handshakes or limitations
 - ◆ For example, transfer ready in FCP write command

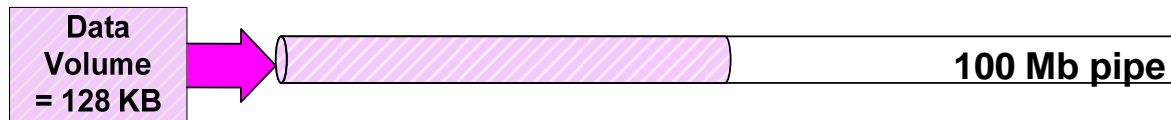
Bandwidth Delay Product

➤ Long Fat Networks have a large bandwidth-delay product

- Bandwidth-delay product = amount of data 'in flight' needed to saturate the network link



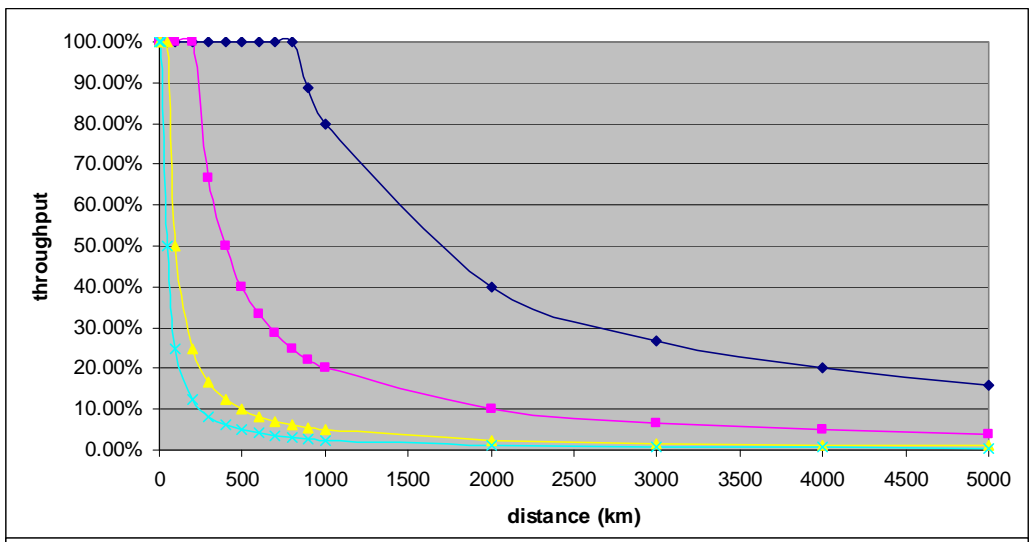
For this example we need 2.56 MB of both transmit data and receive window to sustain line rate



...but for this example only 256KB is needed to sustain line rate

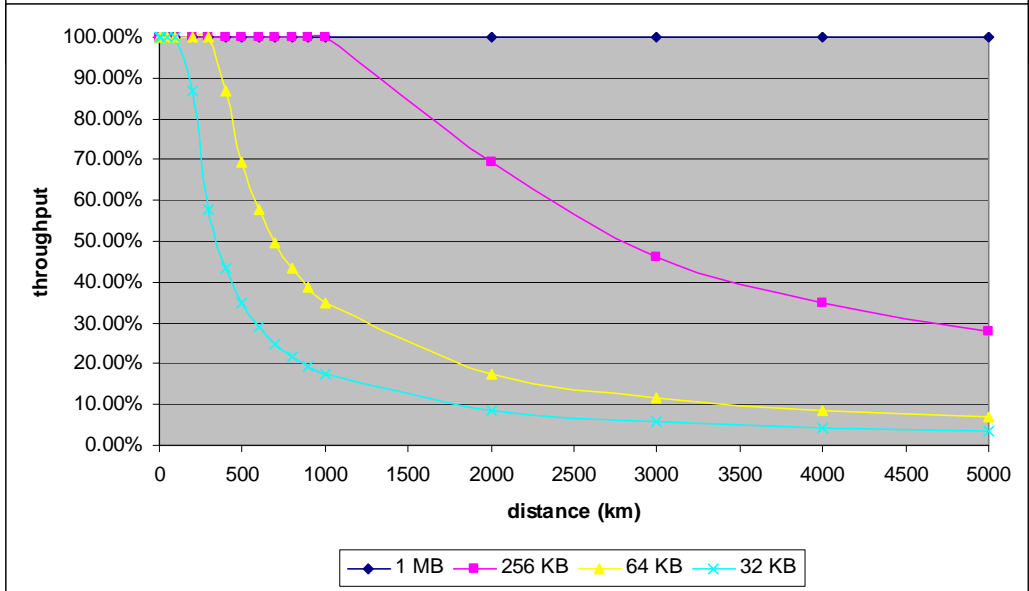
1 ms = 128 KB buffering at 1Gb/s
1 ms = 100 Km a maximum separation

Performance Droop due to distance



➤ Droop at GE line rate (125 MB/s)

Lines represent varying sizes of buffer space or outstanding data

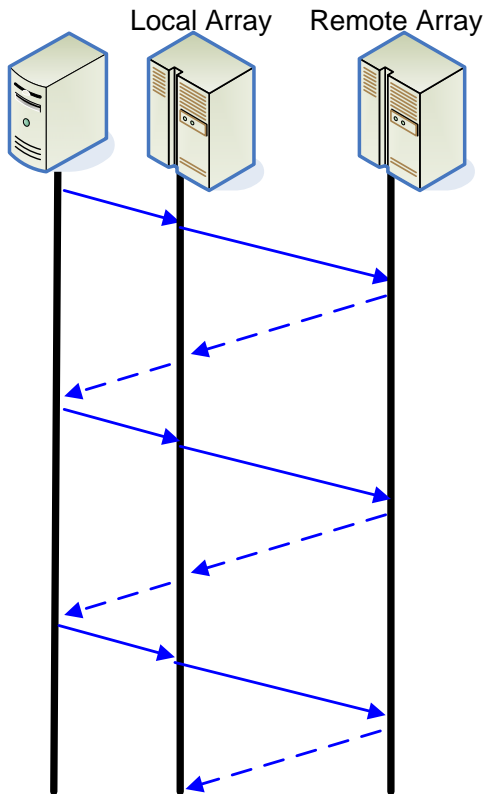


➤ Droop at OC-3 line rate (~18MB/s)

Application-Storage Interaction

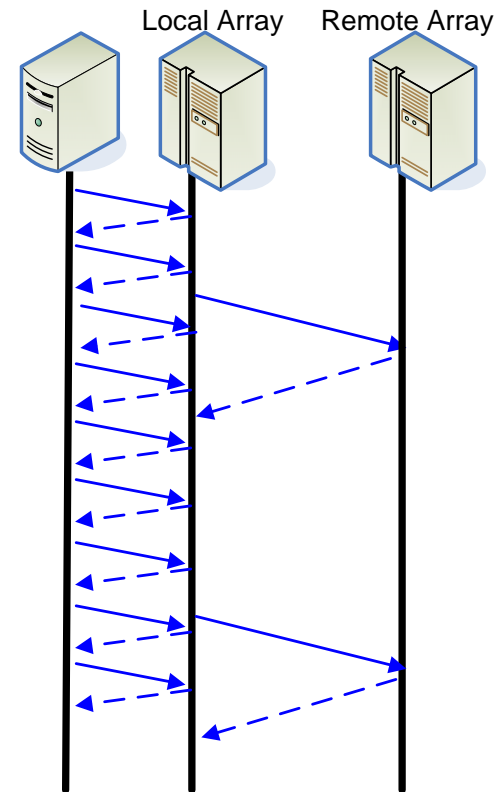
Synchronous Replication

each command must be remotely completed before it is locally completed



Asynchronous Replication

*commands are completed locally as they happen
 Data written remotely in the background*



- Distance between the local and remote array has a large effect on execution of the application. Most Synchronous replication has about a 200 Km range.

Application-Storage Interaction

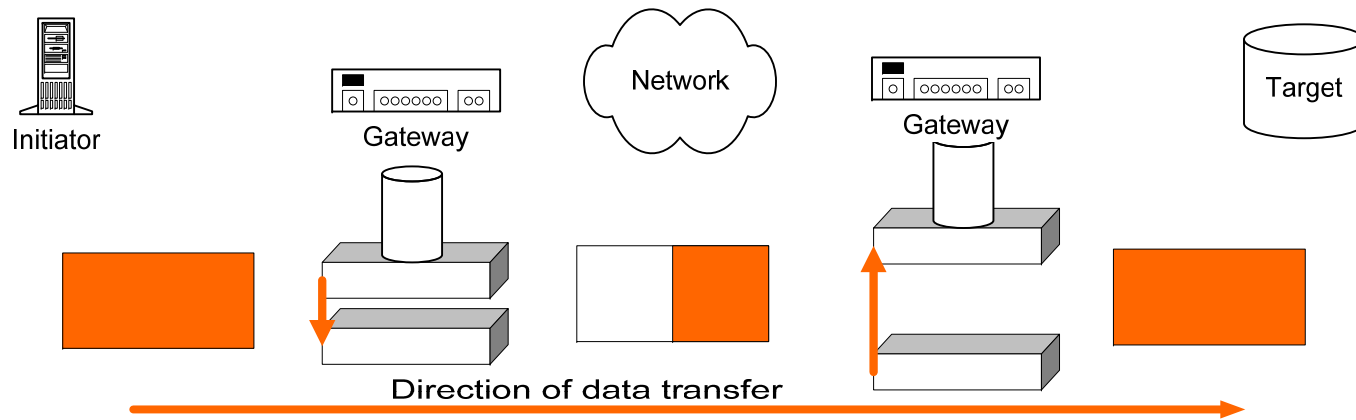
- The storage device or server could do a snapshot or backup of the data instead of continuous writing to remote storage
- In this case a relatively large block of data must be moved across the network. This may have to happen in a specific backup window.
- It is important to understand the behavior of the applications and storage devices SAN to know what demands this places upon the MAN or WAN network.

Optimization Examples

- Compression
- Write Acceleration
- Tape Acceleration

Transport Accelerations such as for TCP/IP already discussed

Compression



Increases the Effective network capacity by the compression ratio.

➤ Compression Ratio

- ◆ The size of the incoming data divided by the outgoing data
- ◆ Determined by the data pattern and algorithm
- ◆ History buffers help the compression ratio since they retain more data for potential matches

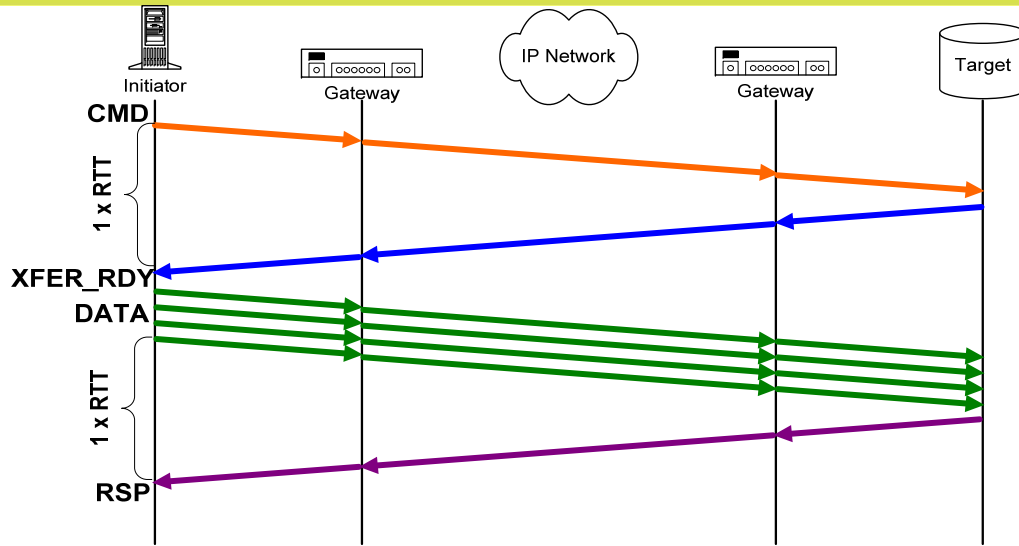
➤ Compression Rate

- ◆ Speed of incoming data processing
- ◆ Different algorithms need different processing power

➤ Many algorithms

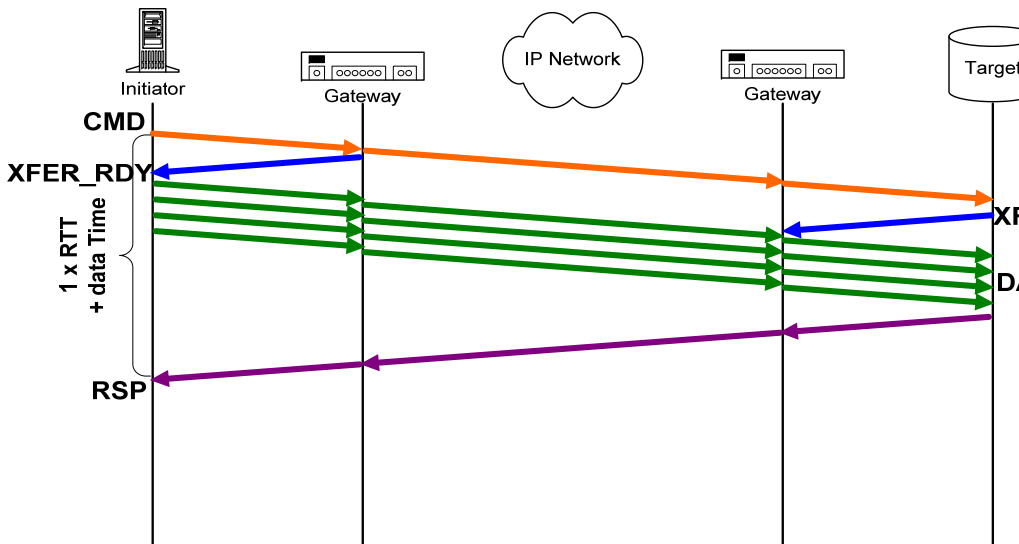
- Higher compression ratios generally require more processing power to achieve the same throughput
- Encrypted data incompressible
- Latency added by compression not usually significant on MAN or WAN time scales (adds about a frame delay)

Write Acceleration (*Fast Write*)



NO write acceleration

➤ $2 \times \text{RTT} + \text{response times}$

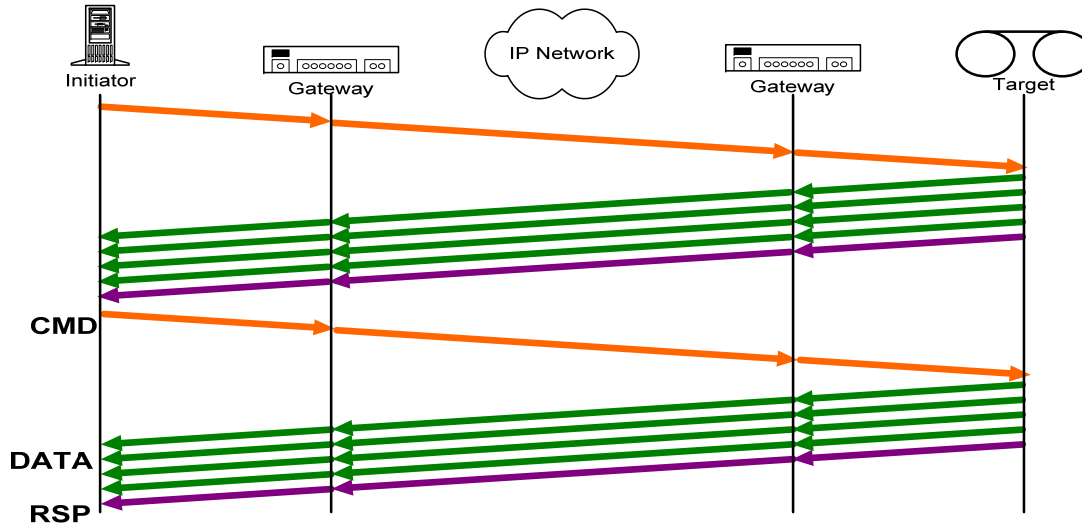


With write acceleration

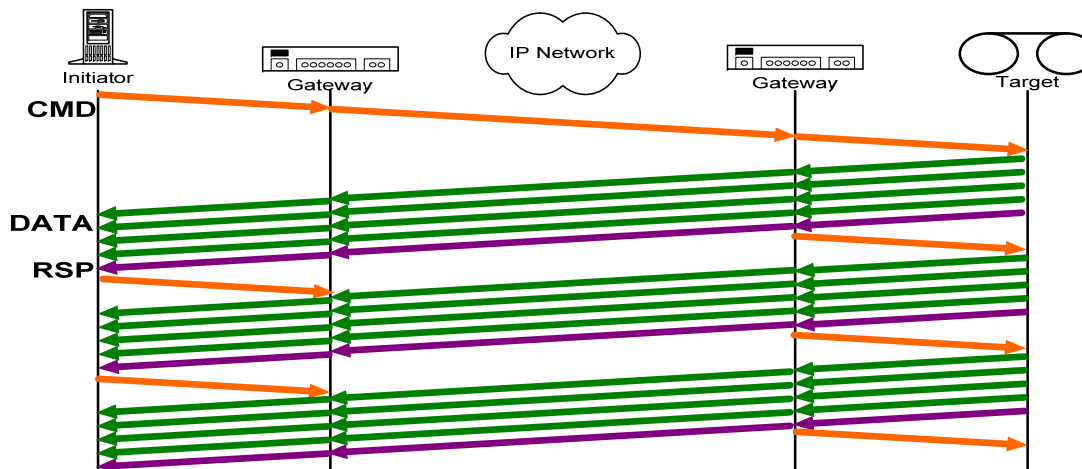
iSCSI can do this trick with immediate data and unsolicited data

➤ $1 \times \text{RTT} + \text{response times}$

Tape Read Acceleration



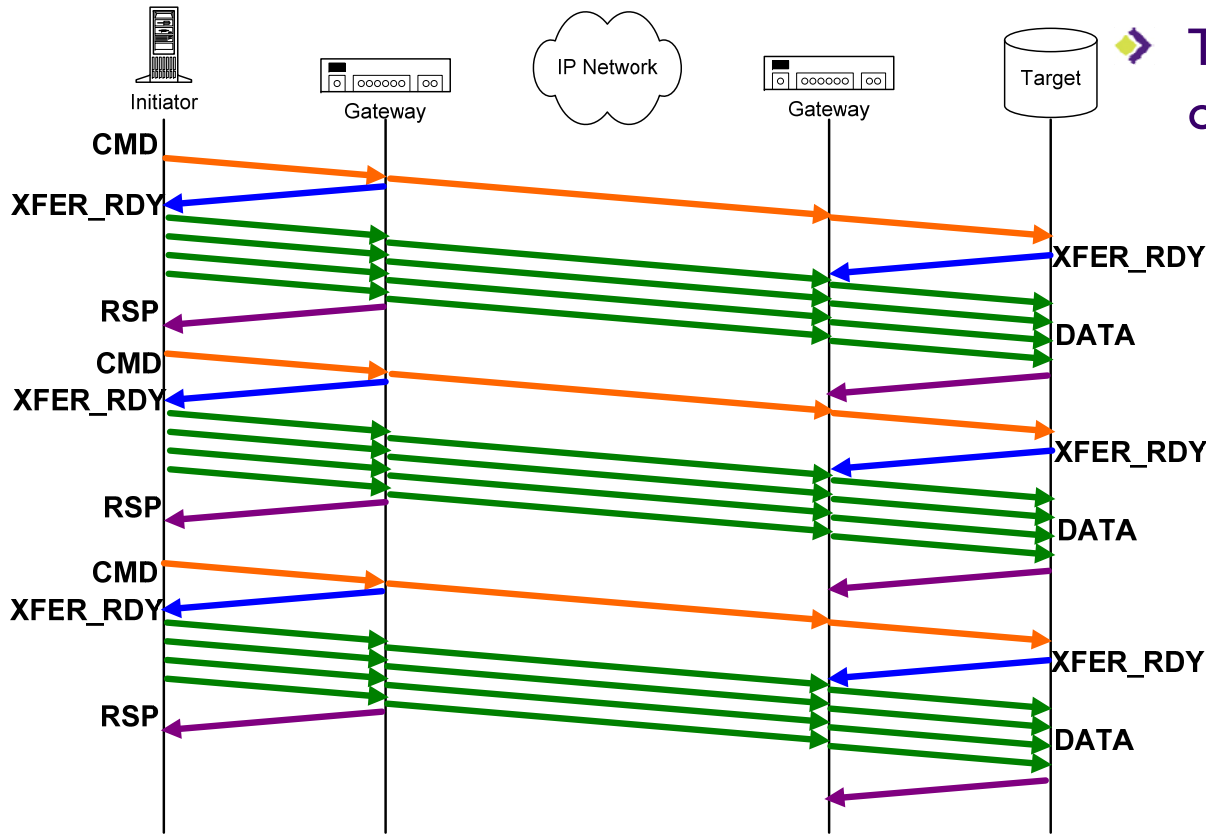
➤ Tape devices only allow 1 outstanding command.



➤ Remote Gateway reads ahead by issuing commands itself

- ◆ Data sent to and buffered by Local gateway until command received
- ◆ Works well because a tape is a sequential access device

Tape Write Acceleration



➤ Tape devices only allow 1 outstanding command.

➤ Apply both Write Acceleration and early response to allow pipelined commands

- Didn't discuss this at all but clearly important!
- There is an entire track on the security topic



**Check out SNIA Tutorial
Track: Security**

- In general data in transit needs to be secured whenever it traverses an exposed network segment...
 - ◆ This can be lots of places but generally it is where the network leaves a secure data center
 - ◆ Technologies include IPSec, FC_SP, etc

The End

- MAN and WAN storage networking is a big topic
- Lots of diverse technologies
- Once the technologies are chosen
 - ...There are still lots of ‘moving parts’ to worry about
 - ◆ **Must design SAN to match MAN/WAN**
 - AND**
 - ◆ **Must design MAN/WAN to match SAN**
- This world overlaps with WAN accelerators, remote file system access, grids & clouds, etc, etc, etc

- Please send any questions or comments on this presentation to SNIA: tracknetworking@snia.org

**Many thanks to the following individuals
for their contributions to this tutorial.**

- SNIA Education Committee

**Joseph L White
Simon Gordon
Viswesh Ananthakrishnan
Howard Goldstein
Walter Dey**

**Based upon the presentation by
Stephen Barr
Greg Schulz**

Appendix: References

- Resilient Storage Networks - Designing Flexible Scalable Data Infrastructures
Greg Schulz – Elsevier/Digital Press Books ISBN: 1555583113