



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 individual members may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced in their entirety without modification
 - ◆ The SNIA must be acknowledged as the 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, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant 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

Agenda

- Why extend SANs across MANs and WANs
- General Background
- Discussion on MAN
- Discussion on WAN
- Discussion on Applications + Throughput Droop

Why is Distance Important?

It's about Data Protection!

BC/DR

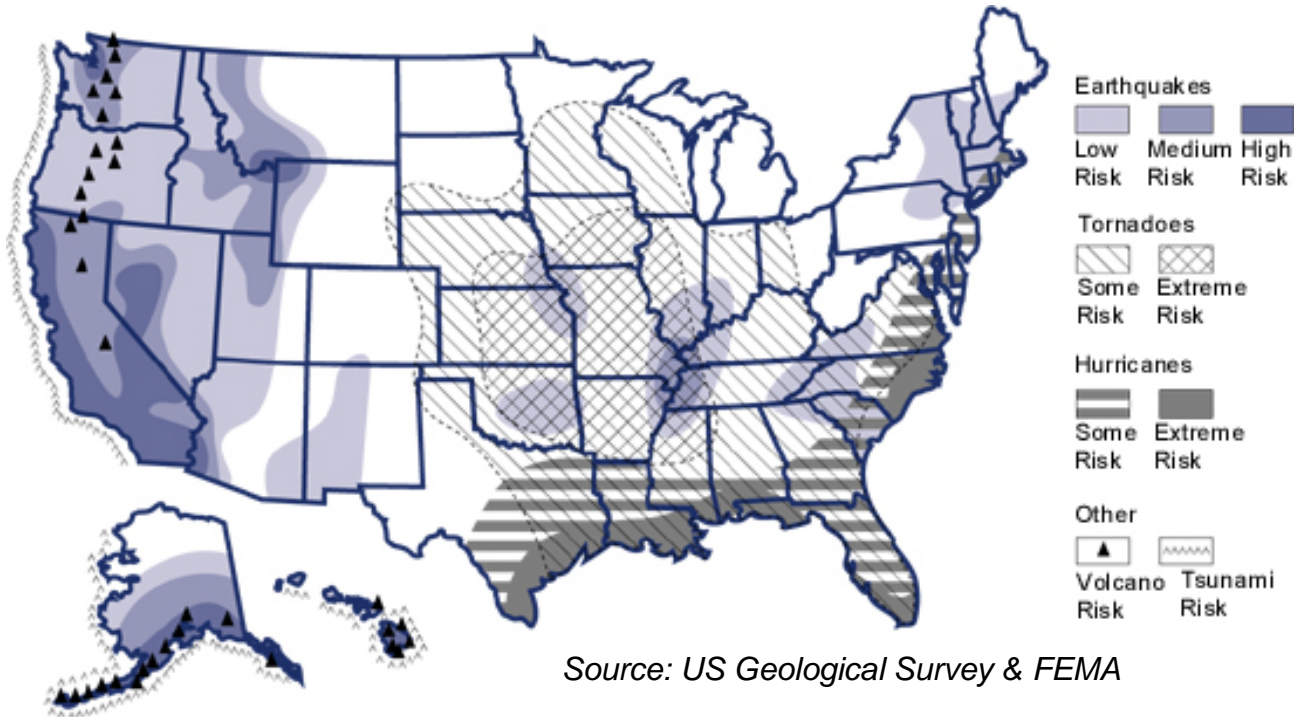
- ◆ Human
- ◆ HW/SW
- ◆ Power Outages
- ◆ Nature

Business

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

Regulatory

- ◆ HIPAA
- ◆ SoX
- ◆ Finance

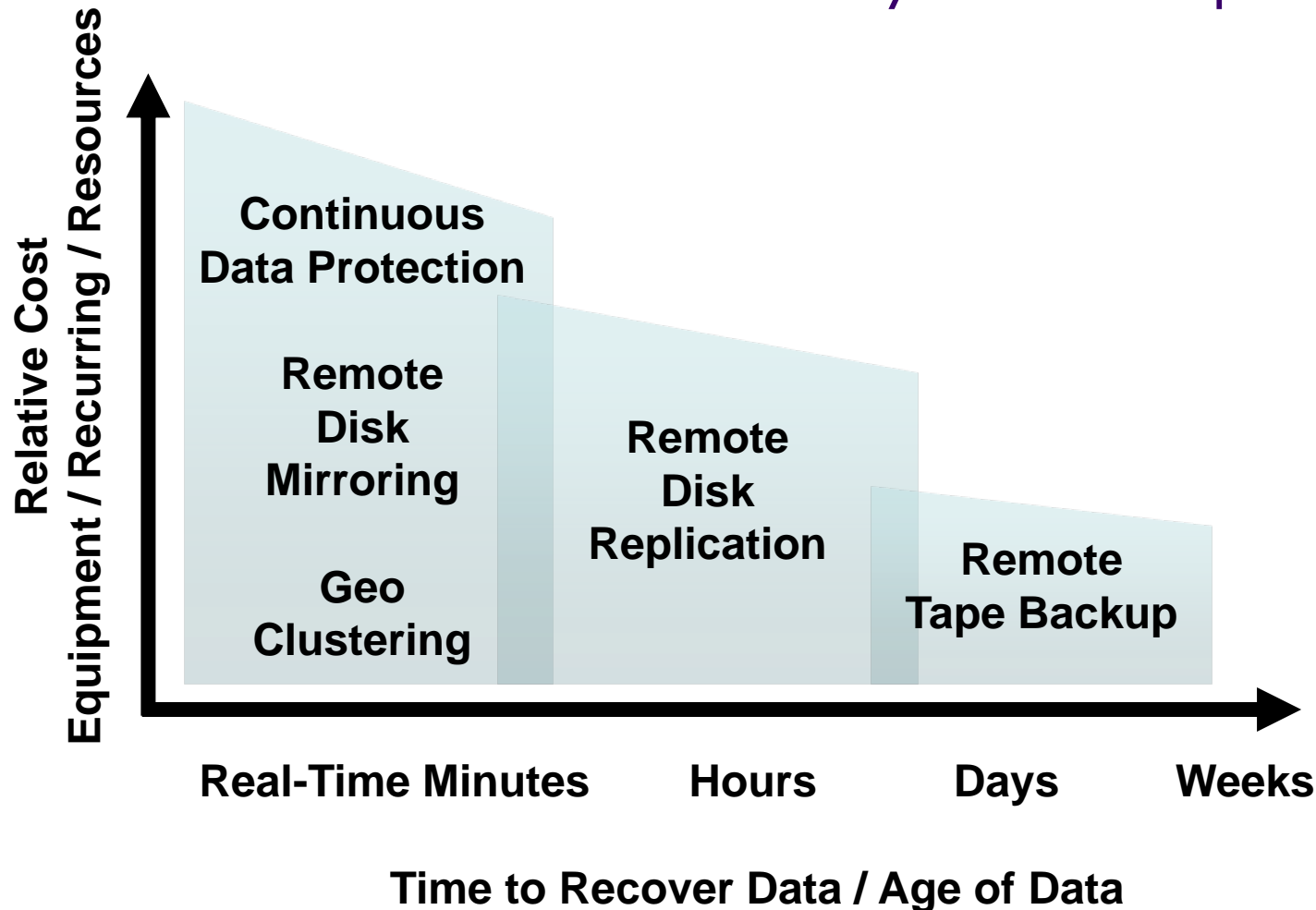


Minimize Risk from a Single Threat Source

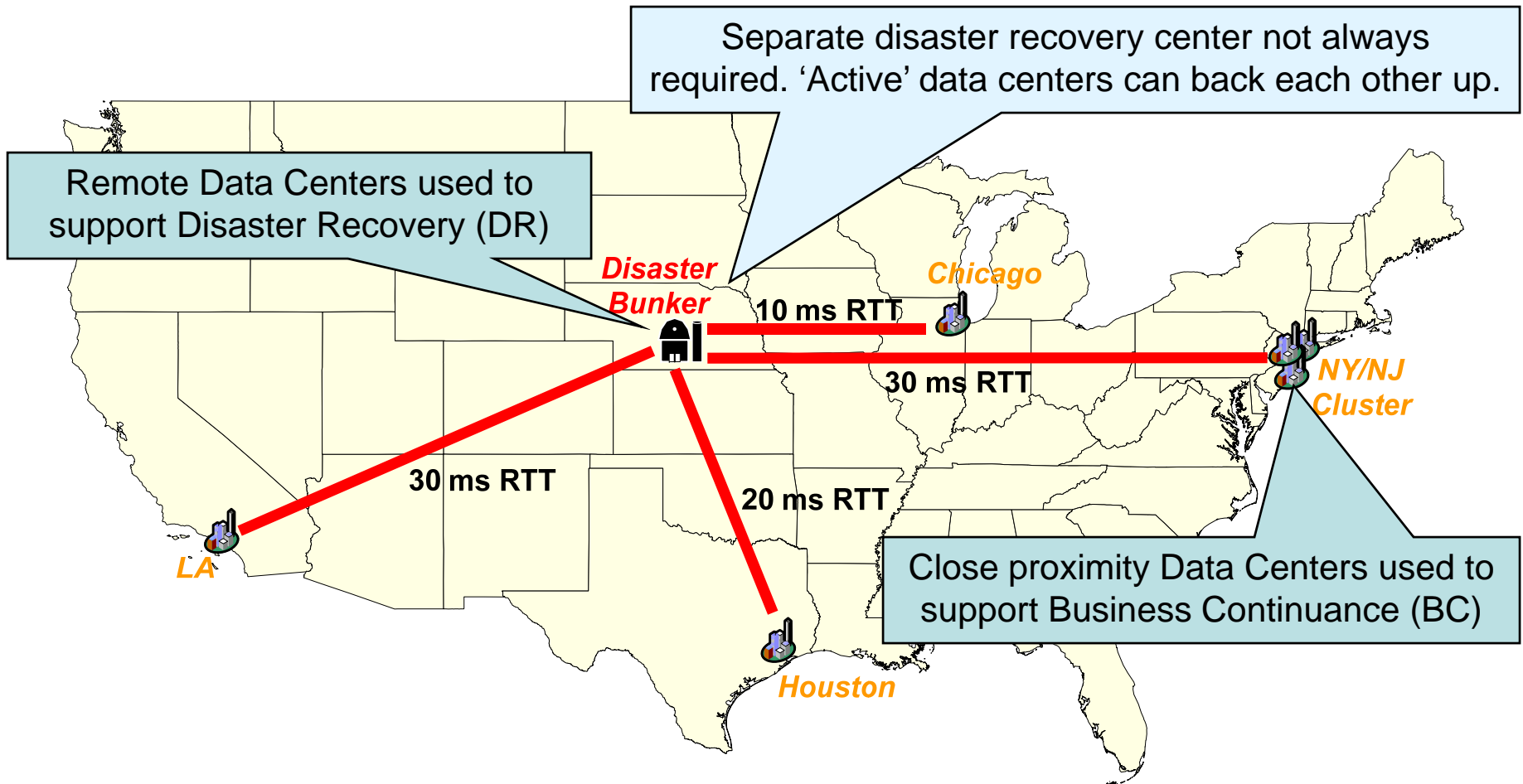


What SAN apps run over distance?

➤ Choice is Driven by business requirements



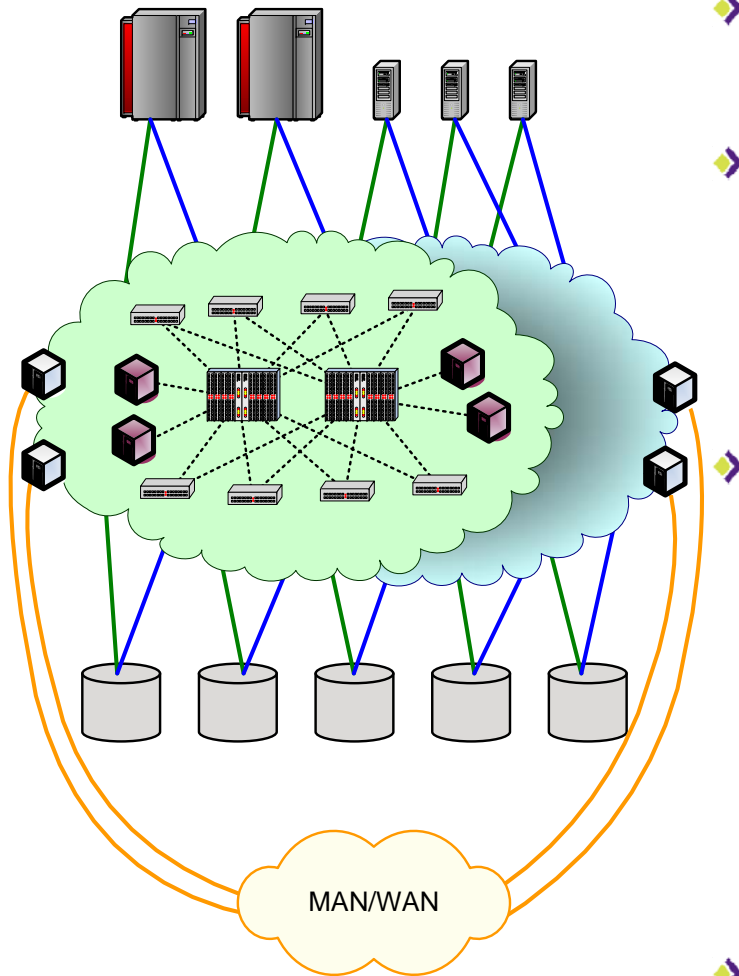
The WAN



- 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)



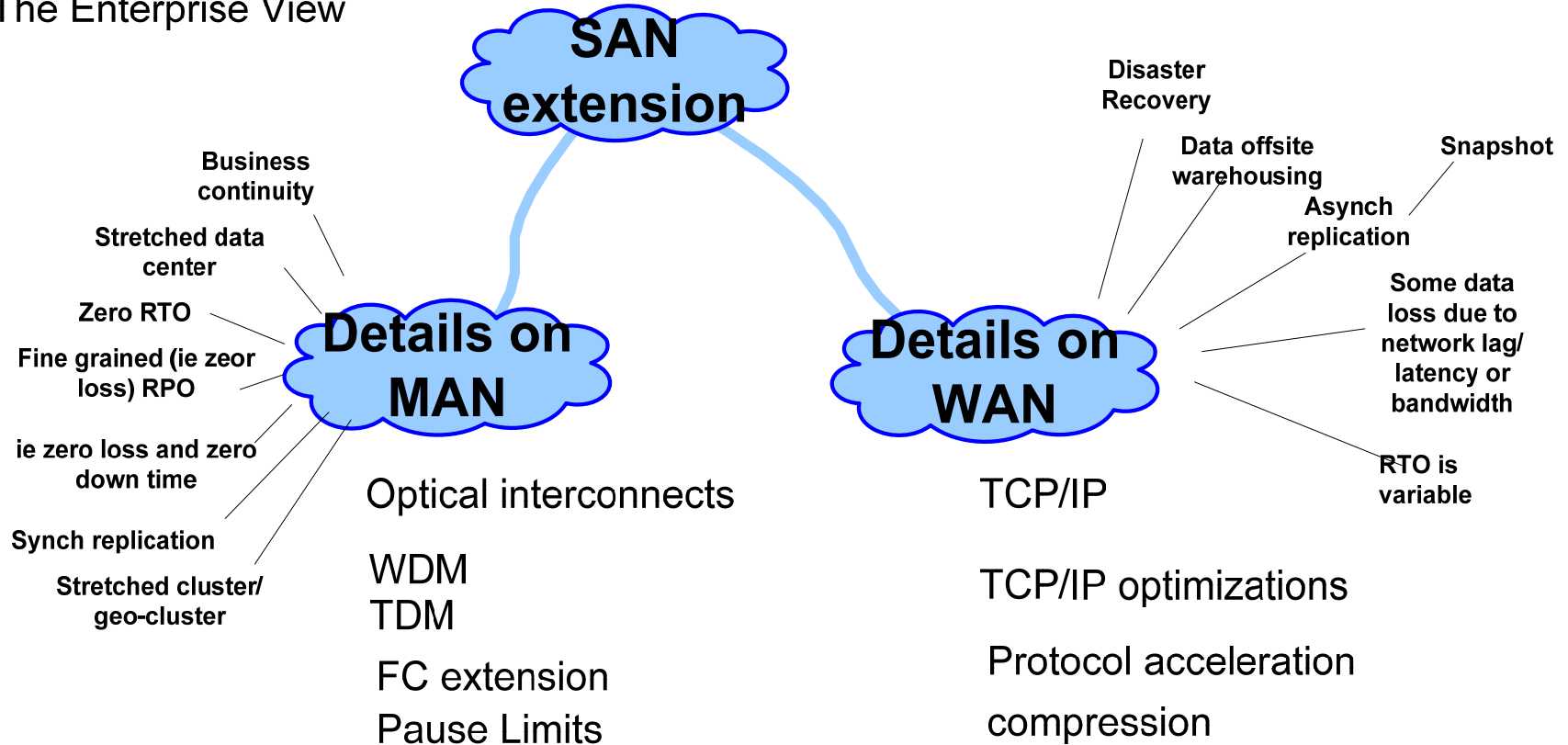
- 150-200 Km max diameter
 - ◆ Effective range of synchronous applications
 - ◆ Increasing longer range deployments (100Km+)
- Can be as short as a few 100 meters
 - ◆ i.e. to the next building
- 5-10 Km separation between sites common
 - ◆ Older installs + newer SMBs
- Long range optics
 - ◆ 40-80 km 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



- Servers accessing Storage across switched Local Area Network
 - ◆ Remember **SANs** can span all distances
- 100s of meters max diameter
 - ◆ This is the effective range of direct access at high bandwidth and low latency
 - ◆ Direct cabling
 - ◆ Short range optics allowed
 - ◆ Can use copper interfaces as well
- **Conventionally**
 - ◆ Multiple SAN Islands deployed as pairs for full dual rail redundancy
 - ◆ Islands provide isolation and limit scale
 - ◆ Each Island is 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.
 - ◆ Appliances can be attached to provide data services (block virtualization, encryption, etc)
- Gateways attached to provide WAN access

SAN Extension Characteristics and Uses

The Enterprise View



➤ For the mid-range

- ◆ MAN tends to be Ethernet + IP
- ◆ Distances tend to be limited to 25 KM and single Telco-POP

A comparison of the distances

- SAN (Data Center) has completely deterministic bandwidth and latency

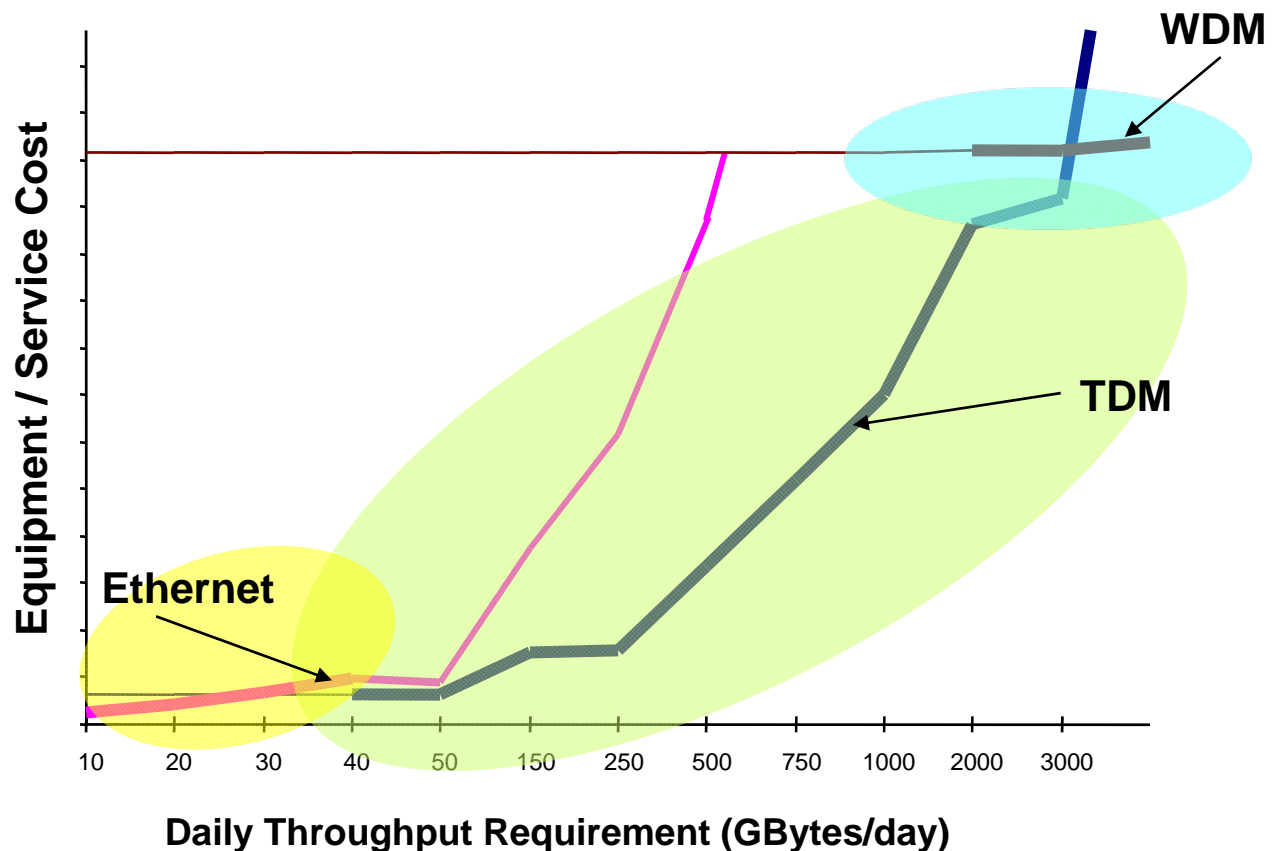
- MAN tends to be deterministic due to single POP access and simple topology for bandwidth and latency
 - ◆ Economics are 1 x GE <25km to single pop is fairly inexpensive <<public numbers>>

- WAN tends to be complex topology and can have non-deterministic behavior
 - ◆ WAN cost goes up significantly with distance and cross region and cross country

Transport Options

Many considerations:

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



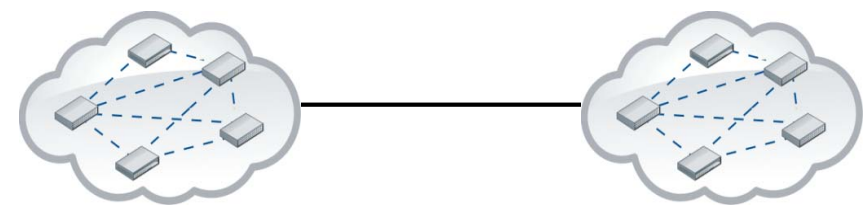
- In general data in transit needs to be secured whenever it traverses an exposed network segment...
 - ◆ This can be lots of places but generally it is where the network leaves a secure data center
 - ◆ Technologies include IPSec, FC_SP, etc

- Will not discuss this in detail
- There is an entire track on the security topic

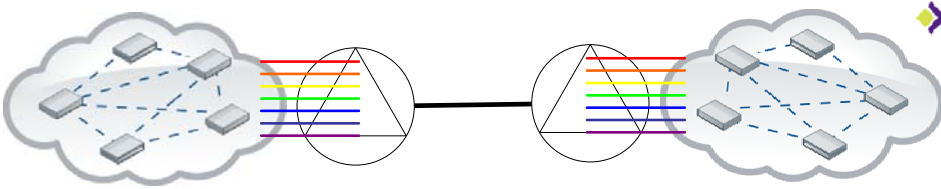


**Check out SNIA Tutorial
Track: Security**

Interconnect Technology

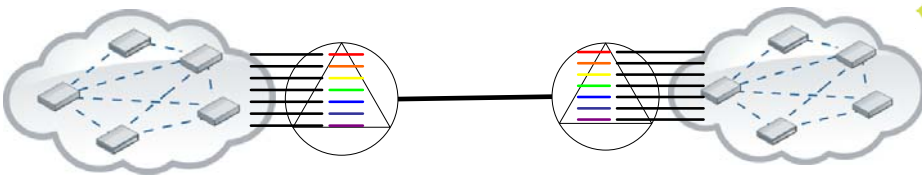


➤ Direct Optical Interconnect



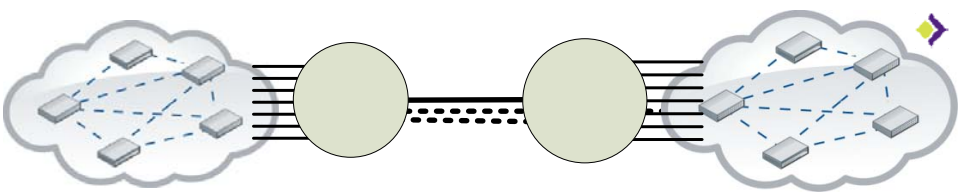
➤ WDM Interconnect 1

- ◆ “Colored Optics”
- ◆ External box is mux only



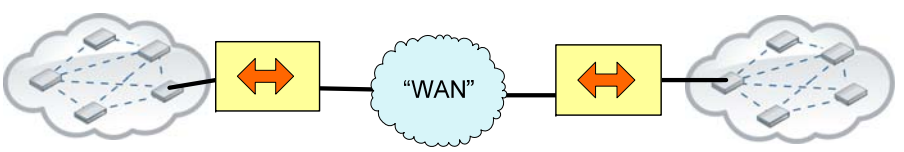
➤ WDM Interconnect 2

- ◆ 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...

FC Transport

- 1/2/4/8G + 10G speeds
 - ◆ (higher speeds on roadmap...)

- 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
 - ◆ 1 credit gives maximum line rate at 1 km separation for 2G FC speed assuming maximum sized frames and only distance latency

- Virtual Channels (VC_RDY)

- Virtual Fabric Tagging

- Class-2 Service (acknowledged frames)

Ethernet Transport

- Layer 2 interconnect
- Speeds from Mb → multi-Gb
 - ◆ 100Mb, 1Gb, 10Gb +roadmap
- Carries
 - ◆ IP traffic (TCP, UDP)
 - ◆ FCoE
- CEE is set of Data Center Enhancements

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

CEE

- 802.1Qbb Priority-based flow Control (PFC)
- 802.1Qaz Enhanced Transmission Selection (DCBX) Data Center Bridging Exchange
- 802.1Qau Protocol Congestion Management
- TRILL (IETF) L2 multipath
- 802.1aq Shortest Path Bridging

➤ Pause frames and distance...

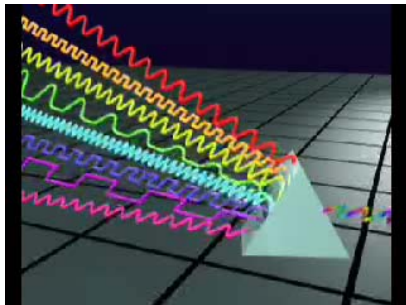
- ◆ When the sender needs to be stopped the receiver sends a frame to notify the sender ...
 - ***If the buffer is overrun then frames can be dropped***
- ◆ This puts a hard limit on the distance for storage traffic across a direct connect Ethernet

Wavelength Division Multiplexing

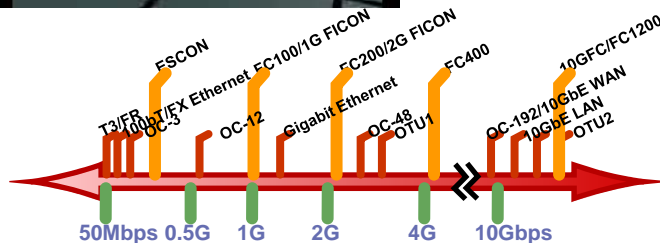


© New Line Productions, Inc

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.



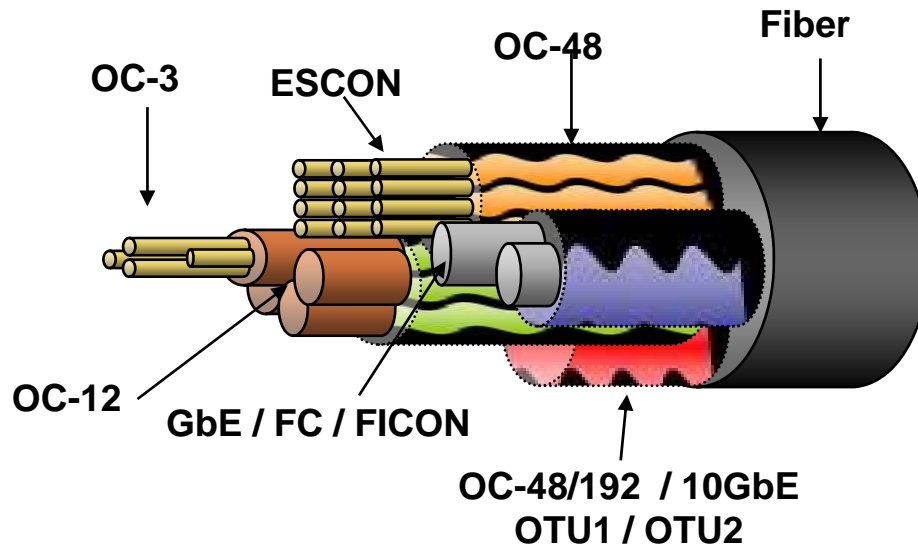
...with each wavelength carrying up to an input connection of “full-rate” throughput...

DWDM: Dense WDM

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

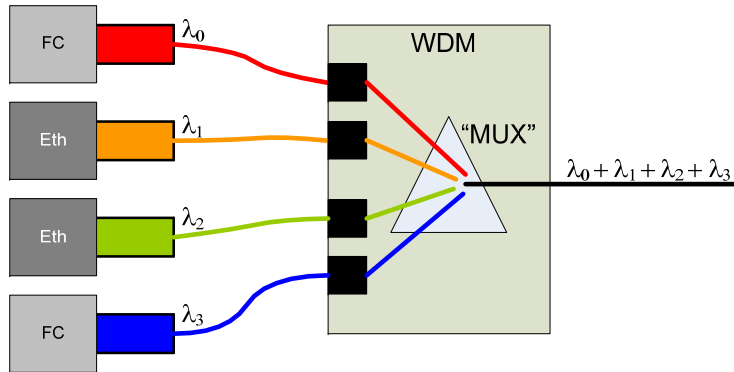
CWDM: Coarse WDM

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

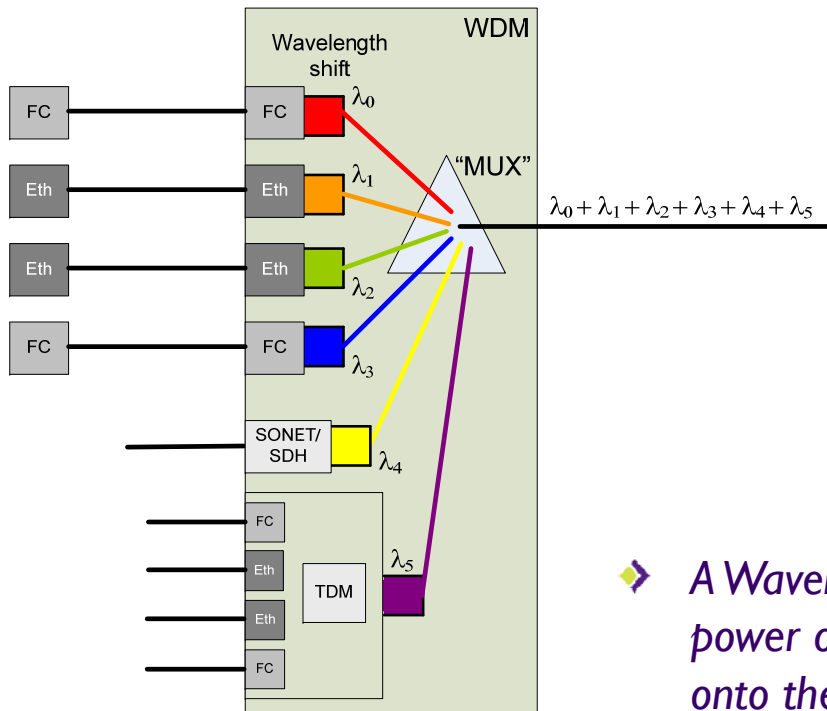


Each wavelength (aka lambda) can utilize its full bandwidth capacity for multiple services

WDM Infrastructure



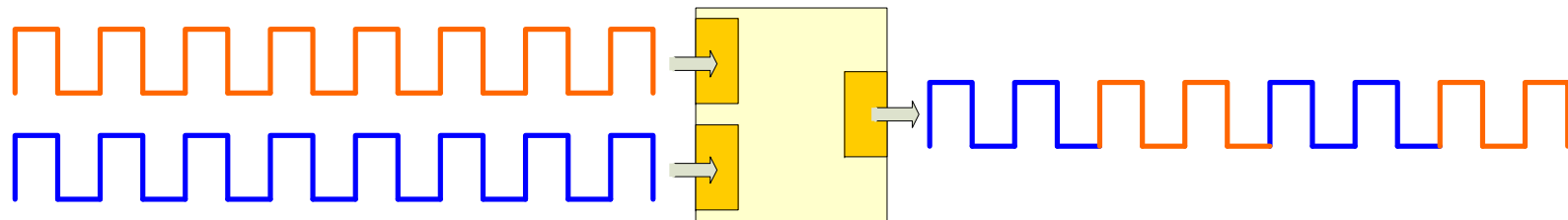
- Colored optics inserted into device
- WDM combines light
- MUX is prism or electronic



- 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.

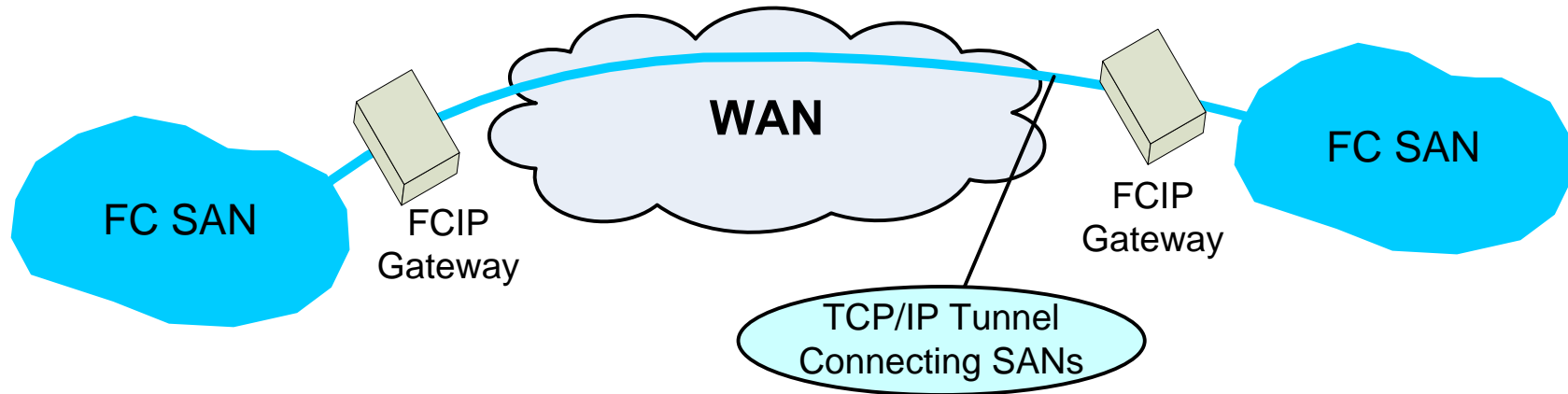
➤ *A Wavelength or 'lambda' is really tight range. Resolving power of equipment determines how many lambdas 'fit' onto the fiber*

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.

WAN Extension Gateways



➤ Predominantly FCIP

- ◆ FC run across TCP/IP tunnel between Gateways
- ◆ Connects local FC Fabrics
- ◆ FC devices & fabric services used as-is
- ◆ *SAN routing can be used to isolate FC fabrics*

➤ TCP/IP implementation and behavior important

➤ Optimizations such as compression and protocol acceleration important

- TCP/IP is both good and bad for block storage traffic
- TCP/IP's fundamental characteristics are good
- TCP/IP's congestion controls and lost segment recovery can cause problems for block storage
 - ◆ Large latencies CAN occur when drops are happening
- 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

- 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 and Packet Loss

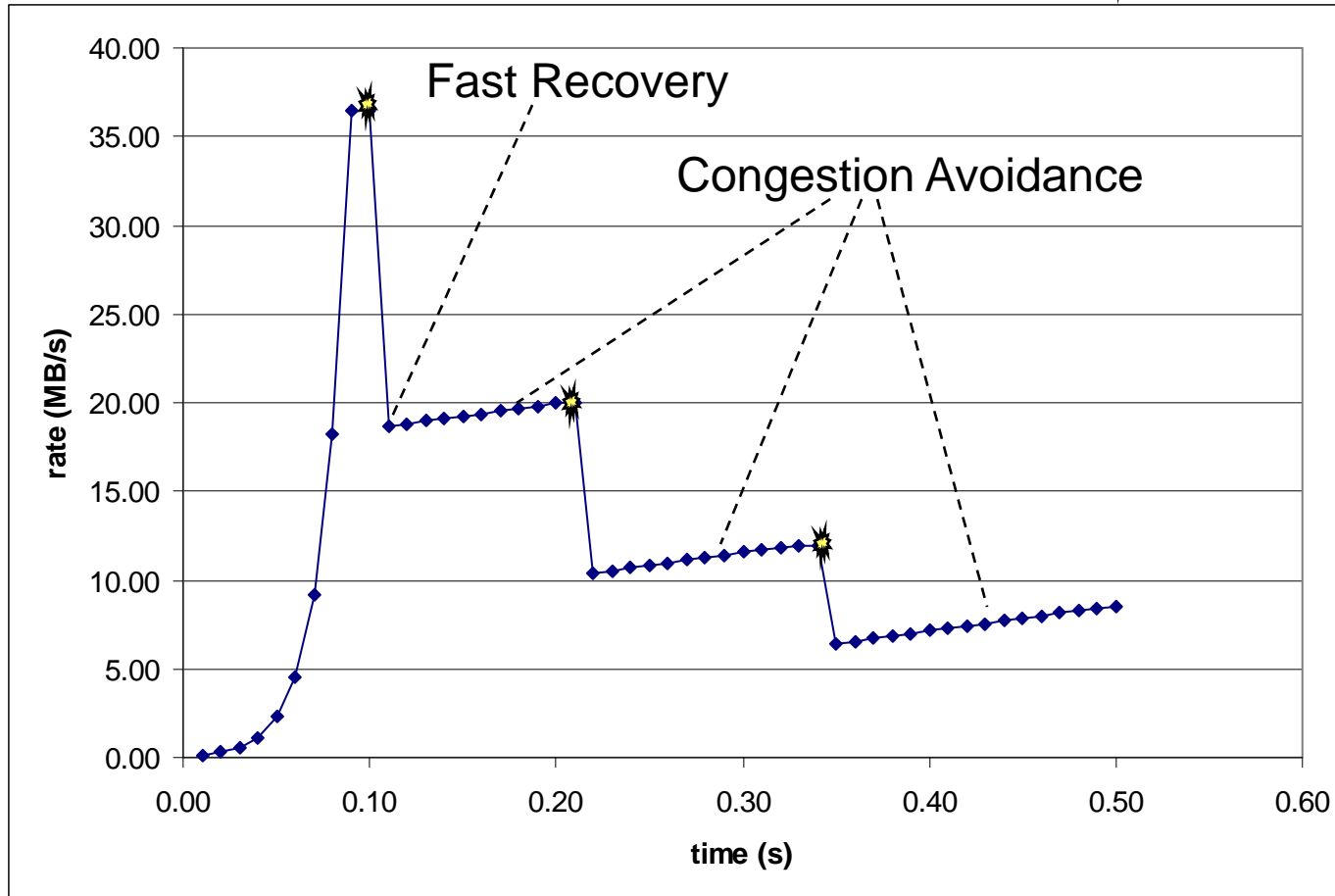
➤ TCP sources of packet loss

- ◆ QoS schemes
 - Strict priority
 - RED, WRED
- ◆ Faulty equipment
- ◆ Inappropriate configuration setting that otherwise has no effect
 - eg PAUSE should be on but it was forgotten or was never previously required.
- ◆ Buffer overrun along the path
 - Typically Due to burst transmit with speed mis-match or other traffic causing congestion

TCP Fast Retransmit, Fast Recovery

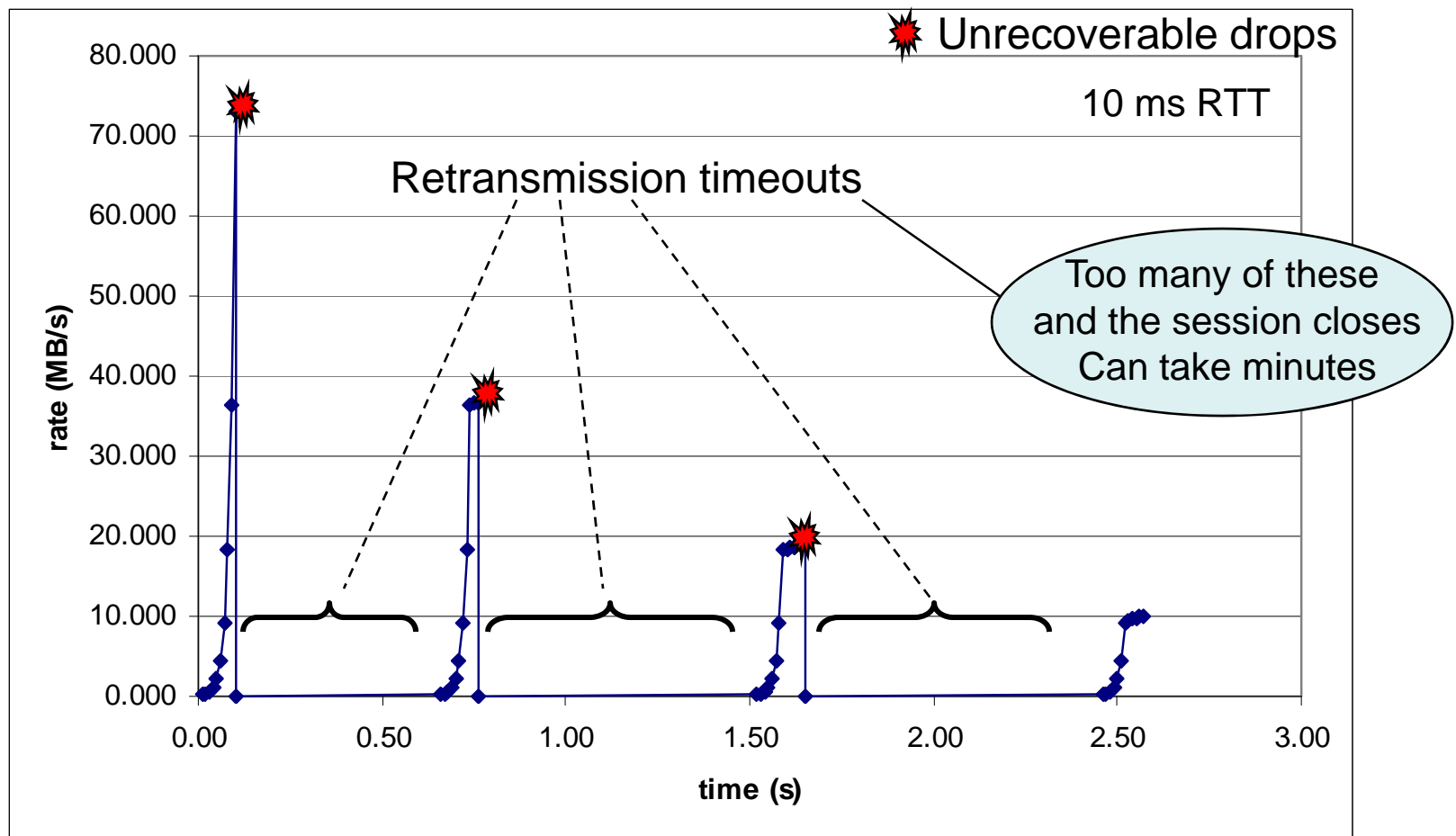
10 ms RTT

☀ 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.

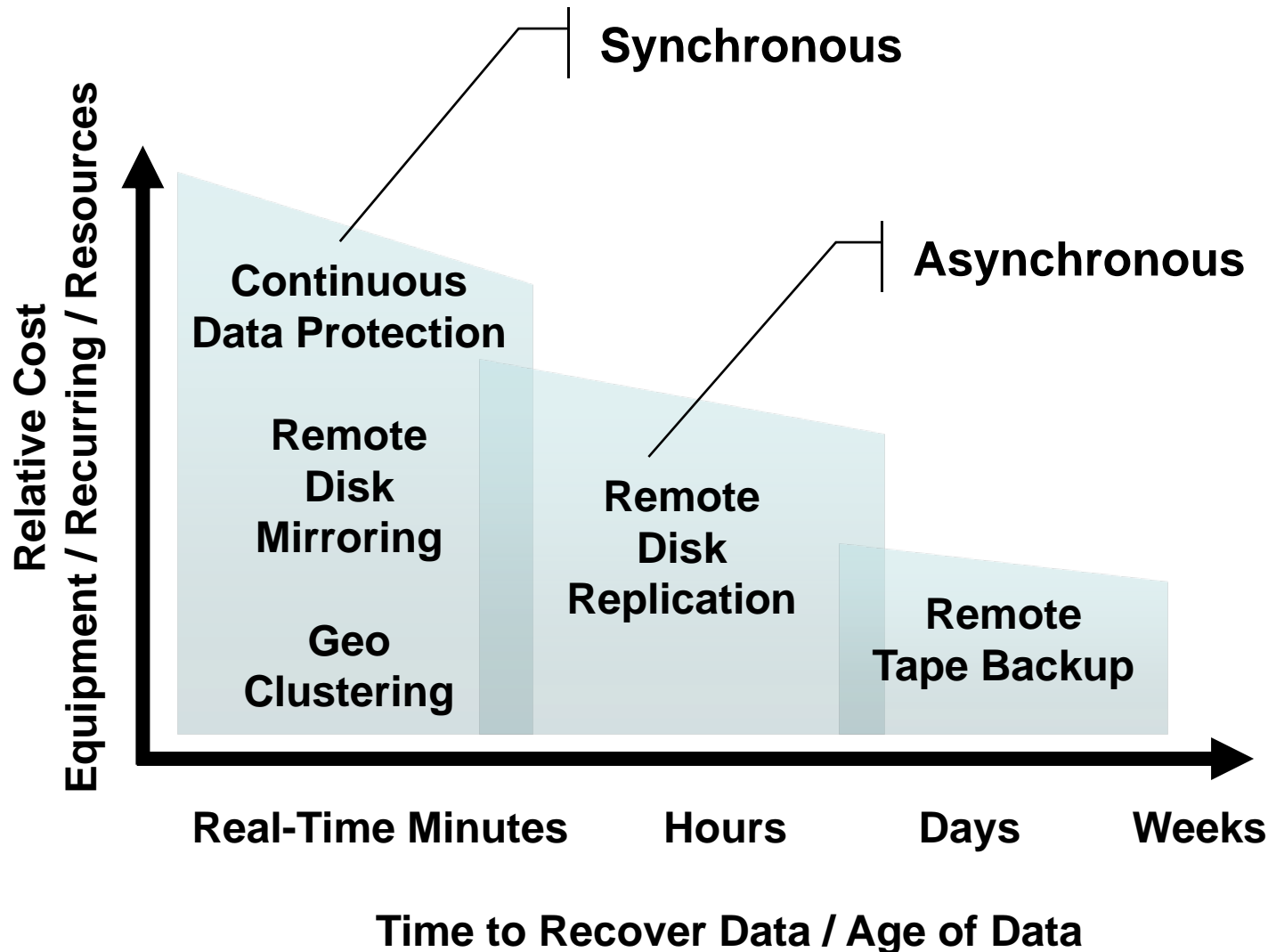
TCP Retransmission Timeout



- 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

- 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

Application Types (again)



Application-Storage Interaction

- 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.
- Not all applications are created equal even at the same distance separation/latency
- Some Apps work better configured for sync replication some for async replication
- RPO/RTO Service Level Agreements can vary significantly

Understand your actual throughput needs:

Changed data size ÷ by backup window = data rate

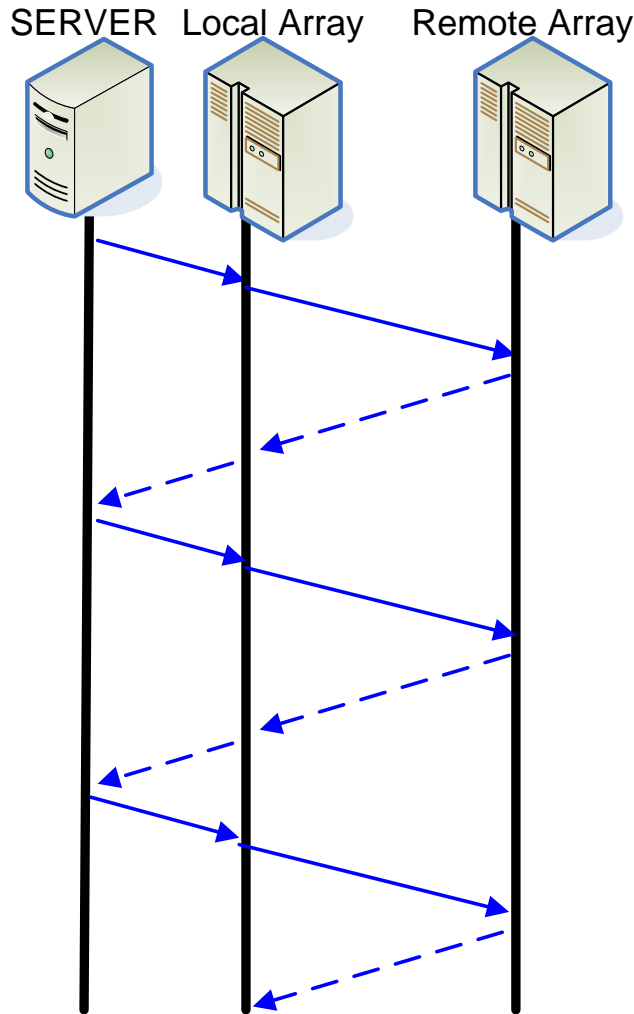
Networks and Performance

- What matters for WAN usage is average sustained throughput sufficient to maintain desired RPO and not affect application performance
 - ◆ RTO here is heavily dependent upon application recovery and restart

- What matters for MAN usage is application performance and responsiveness even under maximum load
 - ◆ Clustering tends to couple RTO to RPO over MANs

Synchronous Replication

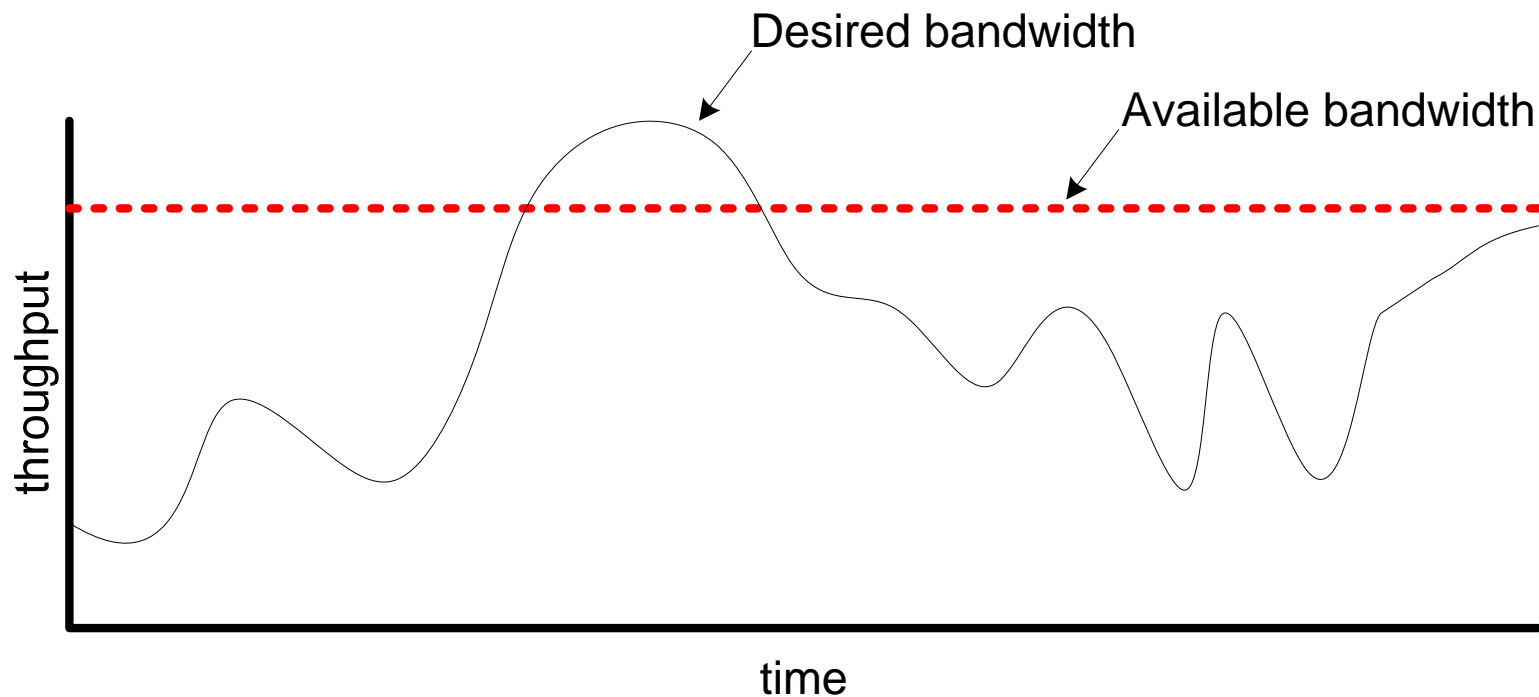
each command must be completed remotely before it is completed locally



- Synchronous replication allows 'zero-time'/'zero-loss' RTO/RPO
- Mostly across MAN distances (<200km)
- Using synchronous replication constrains the network by
 - ◆ Latency – because of application tolerance to command completion rate
 - ◆ Bandwidth – network connections must be sized for short term peaks or can get congestion related latency
- CDP tends to have the same demands and restrictions

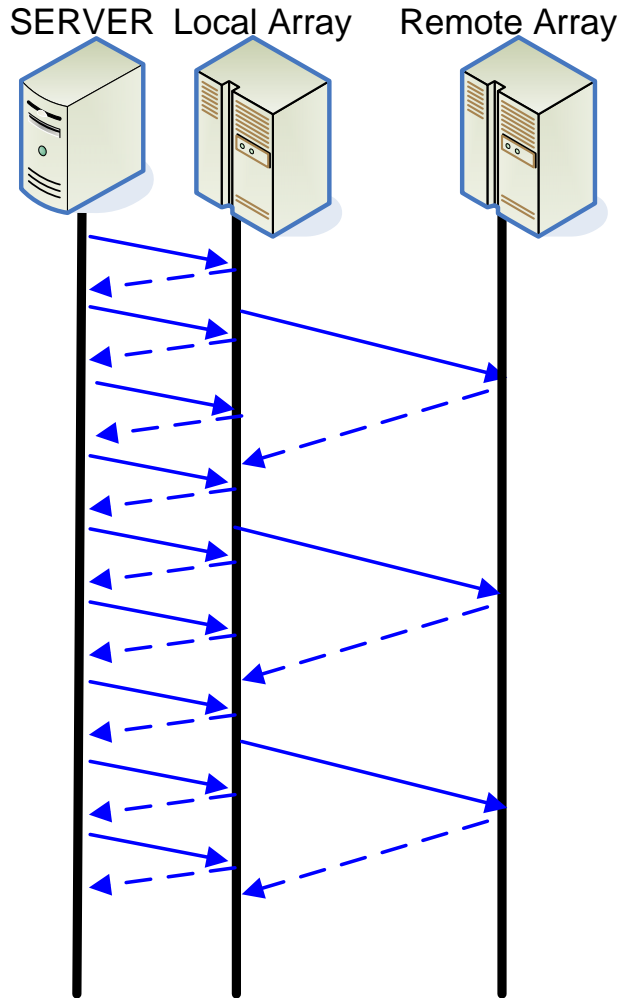
Synchronous Replication

- If the desired bandwidth goes over the available bandwidth, application performance suffers due to the increased latency.
- If the time over bandwidth is long enough the application may break and require intervention to recover.



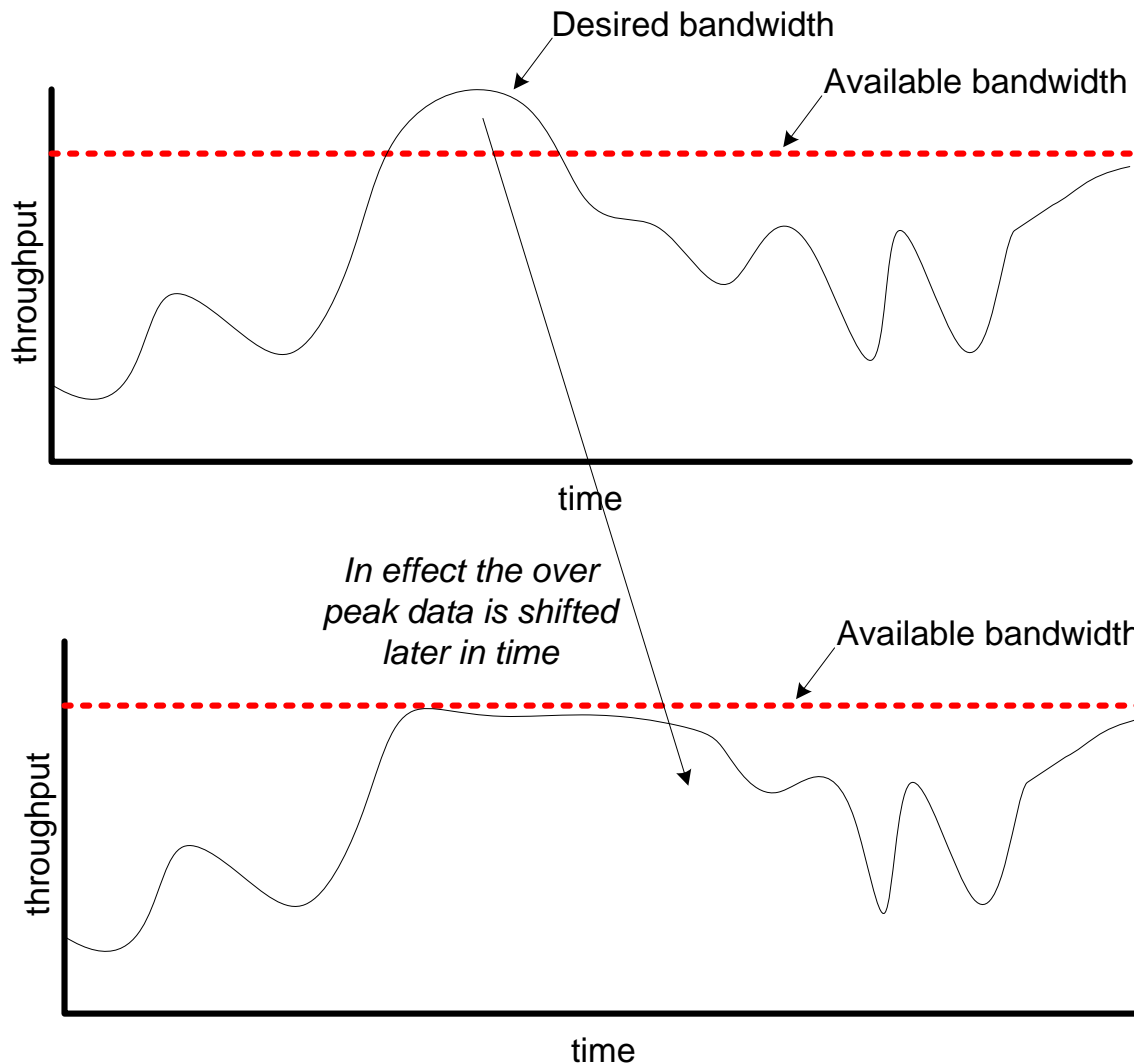
Asynchronous Replication

Commands are completed locally. Data written remotely in the background



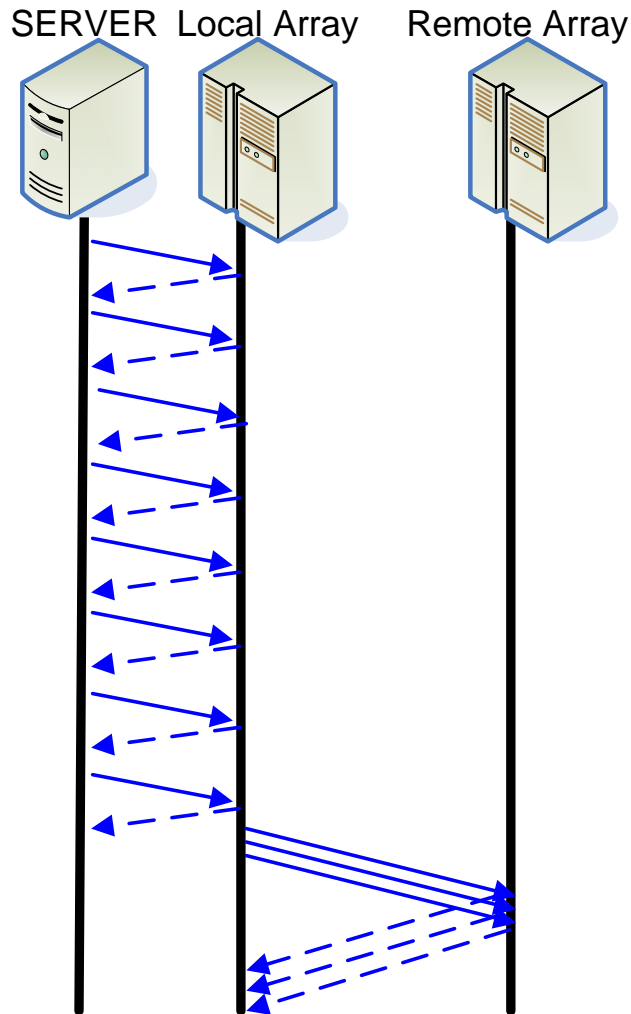
- Asynchronous replication allows WAN distances without hurting application
- Using asynchronous replication constrains the network by
 - ◆ Average throughput – must be sufficient to support moving the data changes from the local to the remote array without falling permanently behind

Asynchronous Replication



- Asynch replication allows data rate to be smoothed without hurting the application performance
- Asynch replication might also reduce the bandwidth needed due to aggregating changes

Snapshot Replication



- The local storage device does a bulk backup of the data instead of continuous writing to remote storage as the data changes
 - ◆ 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.
 - ◆ Characteristics otherwise similar to asynchronous replication

Throughput Droop

- **Physical Network Limit**
 - ◆ Bandwidth-Delay product

- **Transport Buffering Limit**
 - ◆ Number of credits
 - ◆ TCP transmit and receive buffering

- **Available Data Limit**
 - ◆ Outstanding commands (SCSI, NAS, etc)
 - ◆ Individual Command request size

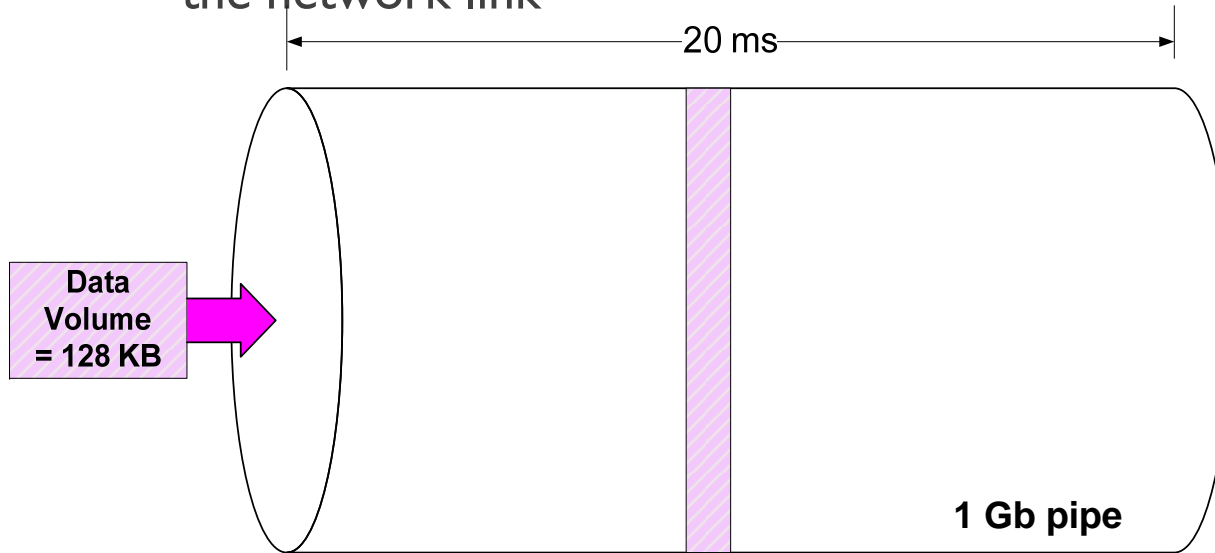
- **Actual throughput is min of these three...**
 - ◆ i.e. must do equivalent bandwidth delay at each protocol level

- **Also have Protocol handshakes or limitations**
 - ◆ For example, transfer ready in FCP write command

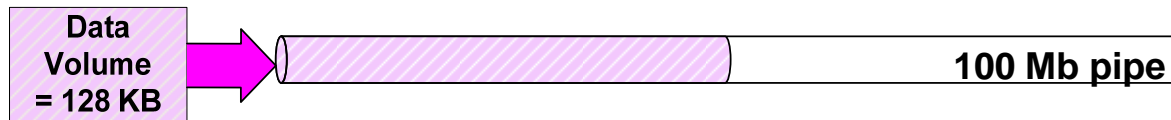
Bandwidth Delay Product

➤ 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

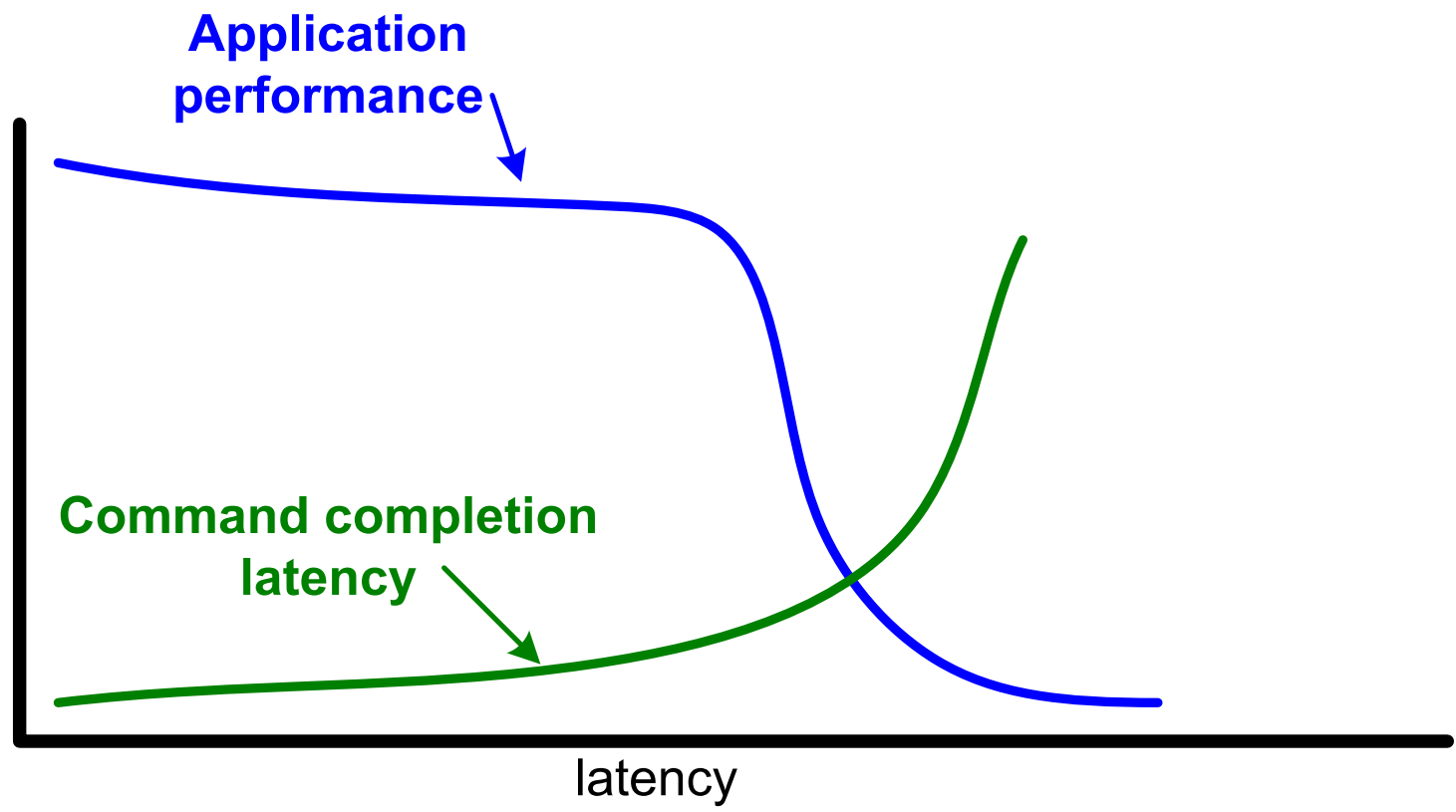
1 ms = 128 KB buffering at 1Gb/s
1 ms = 100 Km a maximum separation

Latency

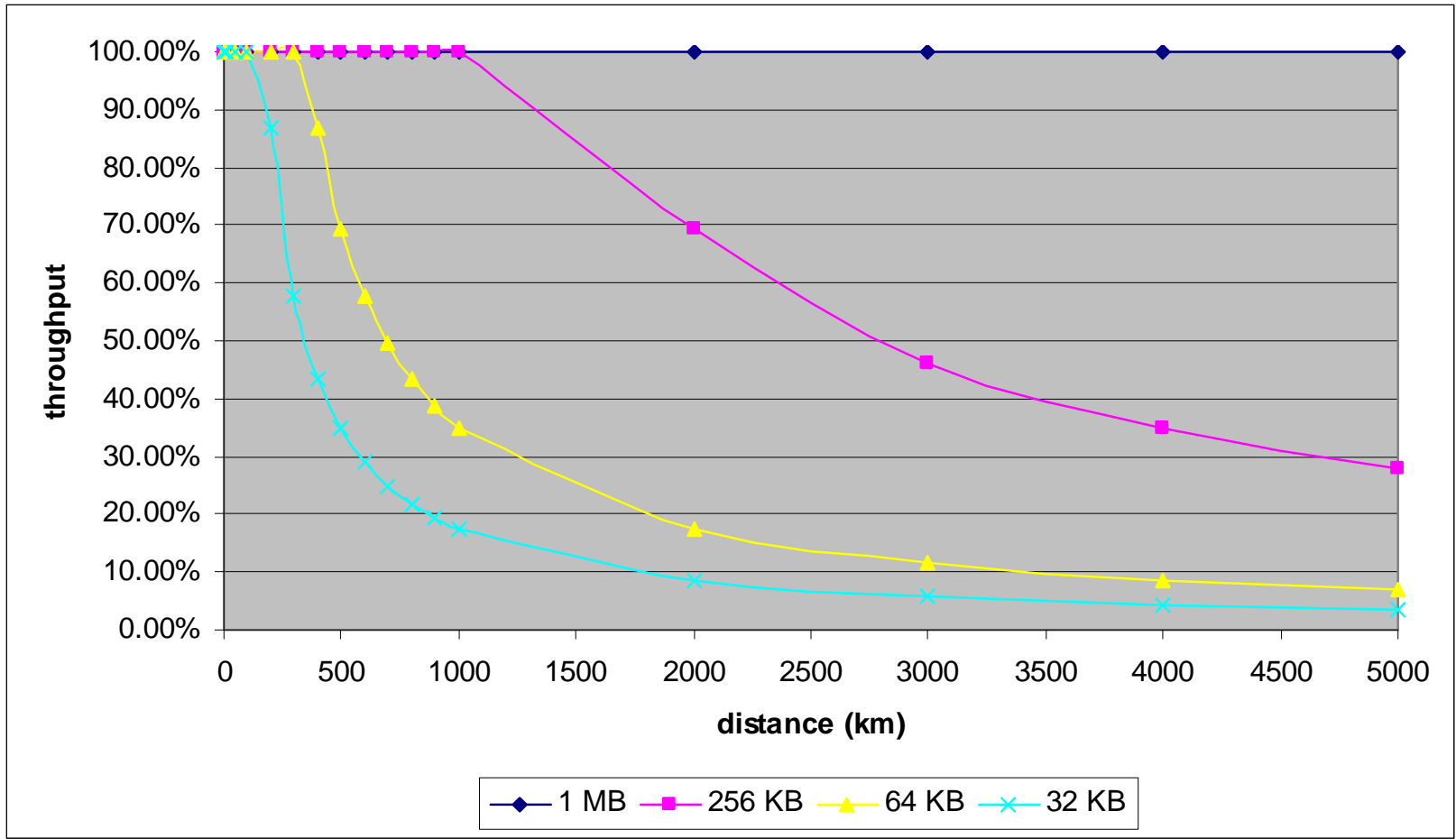
- Command Completion Time is important
- Contributing Factors: (sum 'em all up!)
 - ◆ Distance (due to 'speed of light') - latency of the cables
 - › (2×10^8 m/s gives 1 ms RTT per 100Km separation)
 - ◆ 'Hops' – latency through the intermediate devices
 - ◆ Queuing delays due to congestion
 - ◆ Protocol handshake overheads
 - ◆ Target response time
 - ◆ Initiator response time
- A complicating factor is the I/O pattern and application configuration
 - ◆ Some patterns and applications hide latency much better than others
 - › Good: File Servers and NAS
 - › Bad: transactional database with heavy write to transaction logs

Latency

➤ Can have non-linear application effects...

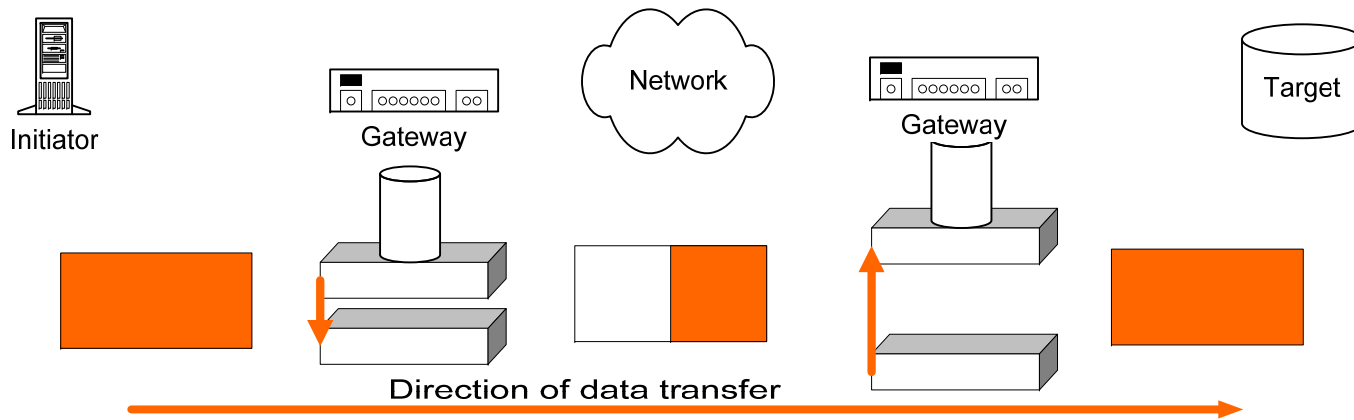


Throughput Droop due to distance



Lines represent varying sizes of buffer space or outstanding data

➤ OC-3 line rate (~18MB/s)



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

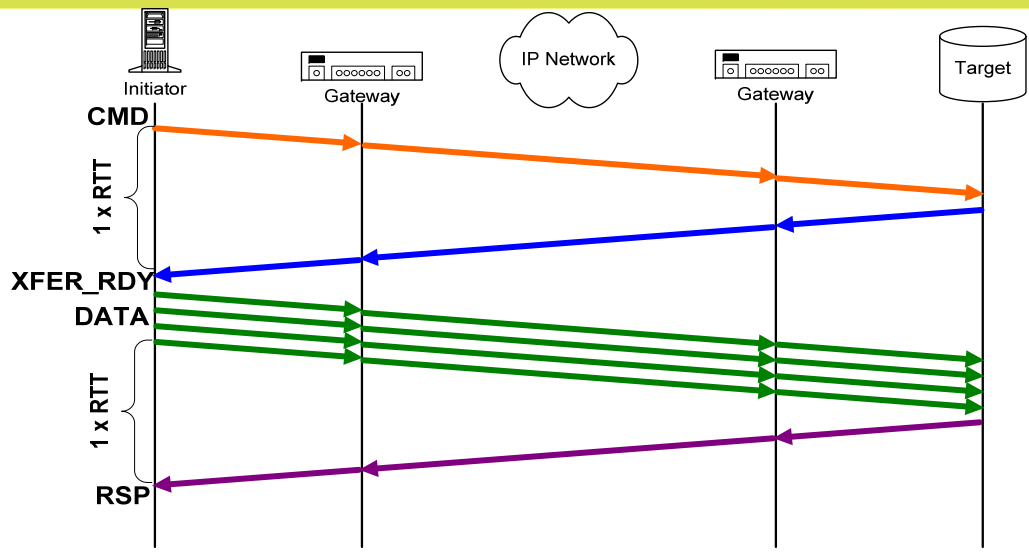
➤ Compression Rate

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

➤ Many algorithms

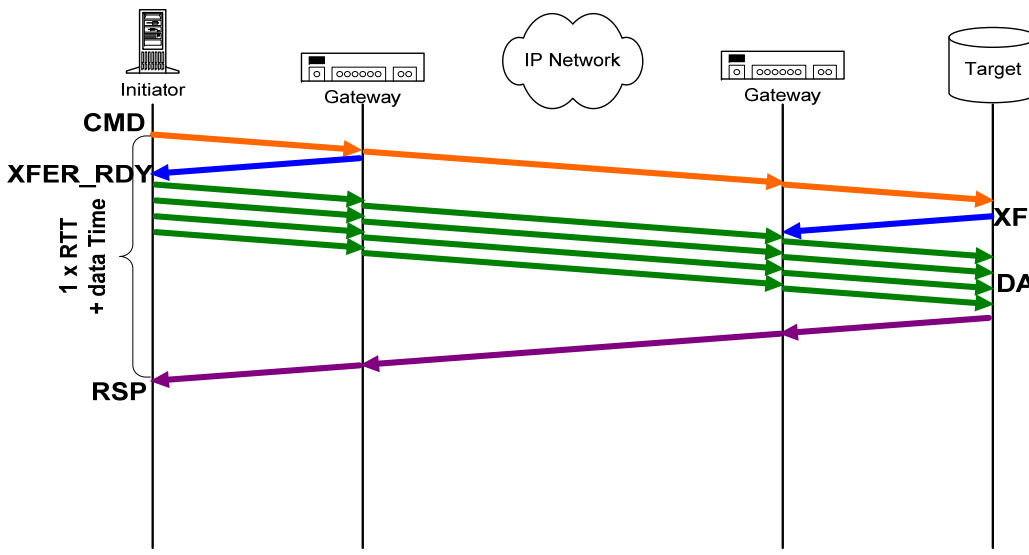
- Higher compression ratios generally require more processing power to achieve the same throughput
- Encrypted data incompressible
- Latency added by compression not usually significant on MAN or WAN time scales (adds about a frame delay)

Write Acceleration (*Fast Write*)



Write Command

➤ 2 x RTT + response times



With write acceleration

➤ 1 x RTT + response times

Some protocols and applications can do this trick directly with immediate or unsolicited data
 Notably iSCSI

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
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: 155583113