SNIA

Education

TCP/IP Inside the Data Center and Beyond

Dr. Joseph L White, Juniper Networks

SNIA Legal Notice

- 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.

Abstract

- TCP/IP Inside the Data Center and Beyond
  - This session provides an overview of TCP/IP high performance behavior for bulk data transfers in LAN, data center, and WAN environments. TCP/IP provides the underlying transport for iSCSI and FCIP. Within the data center understanding TCP/IP is important for converged networks not using FCoE. The effects of various drop patterns, the new Ethernet enhancements for the data center, and TCP/IP processing overhead will be explored. Outside the data center, TCP/IP provides the underlying transport for most of the bulk data transfers across wide area networks (WANs) including distance extension for block storage (iSCSI, FCIP, and iFCP), WAN acceleration for local TCP/IP sessions, and wide area file systems (WAFS) acceleration. The effects of high bandwidth, long latency, impaired, and congested networks as well as the TCP/IP modifications to mitigate these effects will be explored. A fundamental understanding of TCP/IP behaviors is essential for successful deployment of IP storage solutions.

TCP/IP based protocols are in your critical path for storage access

- NAS (CIFS/NFS)
- iSCSI
- WAN Acceleration
- SAN Distance Extension (FCIP)

- File System Access
- Direct Block storage access
- Data Backup and Recovery
- Remote Office Optimized Access

Demanding Data Center Requirements Must Still Be Satisfied

High Throughput

Low Latency

Robustness

Wide Scalability

High Availability

TCP/IP does not get a free pass!

Agenda

Education SNIA

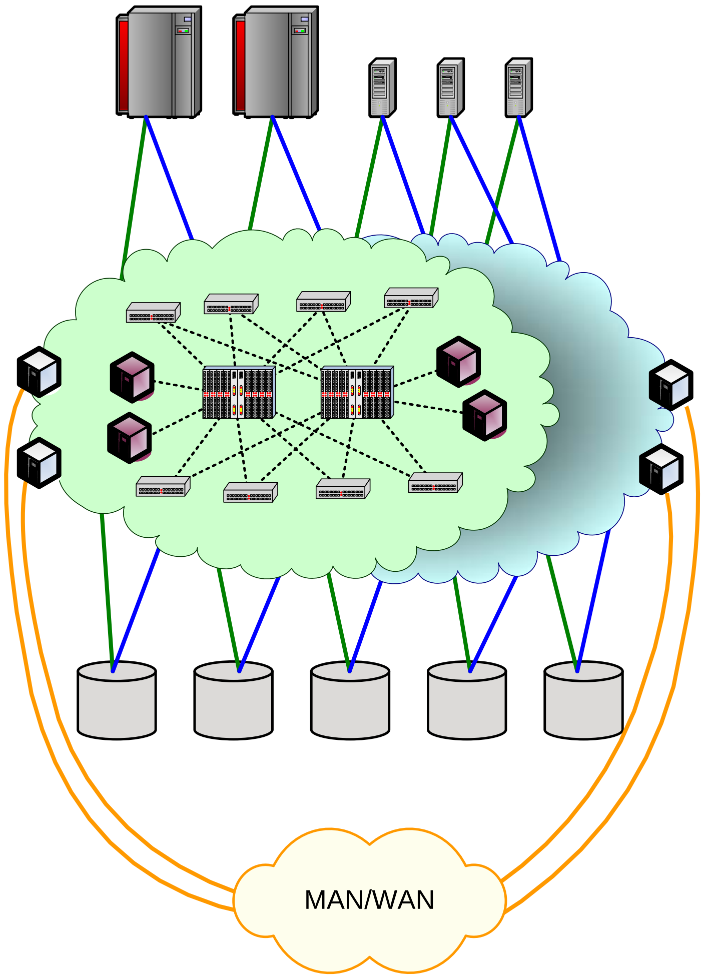
- **Distance Scales and Deployments**
  - LAN(Data Center)/MAN/WAN
- **TCP/IP Protocol Whirlwind Tour**
  - Characteristics of TCP/IP
  - Sliding Window
  - Sender Congestion Controls
  - Packet Loss and TCP/IP Retransmission
- **LAN (Data Center) and MAN Considerations**
  - Ethernet/CEE/DCB
  - Short Ranged TCP/IP over lossless networks
  - Pause vs Credit based Flow Control
- **WAN and MAN Considerations**
  - Packet Loss for Long Fat Networks (LFNs)
  - TCP/IP Modifications and Optimizations
  - Application Performance Droop

3 Distance Scales to Consider

LAN (Data Center)	<5 Km
MAN	1 Km to 200 Km
WAN	>100 Km

Each Distance scale places different constraints on TCP/IP

The LAN (FC, IB, Eth)



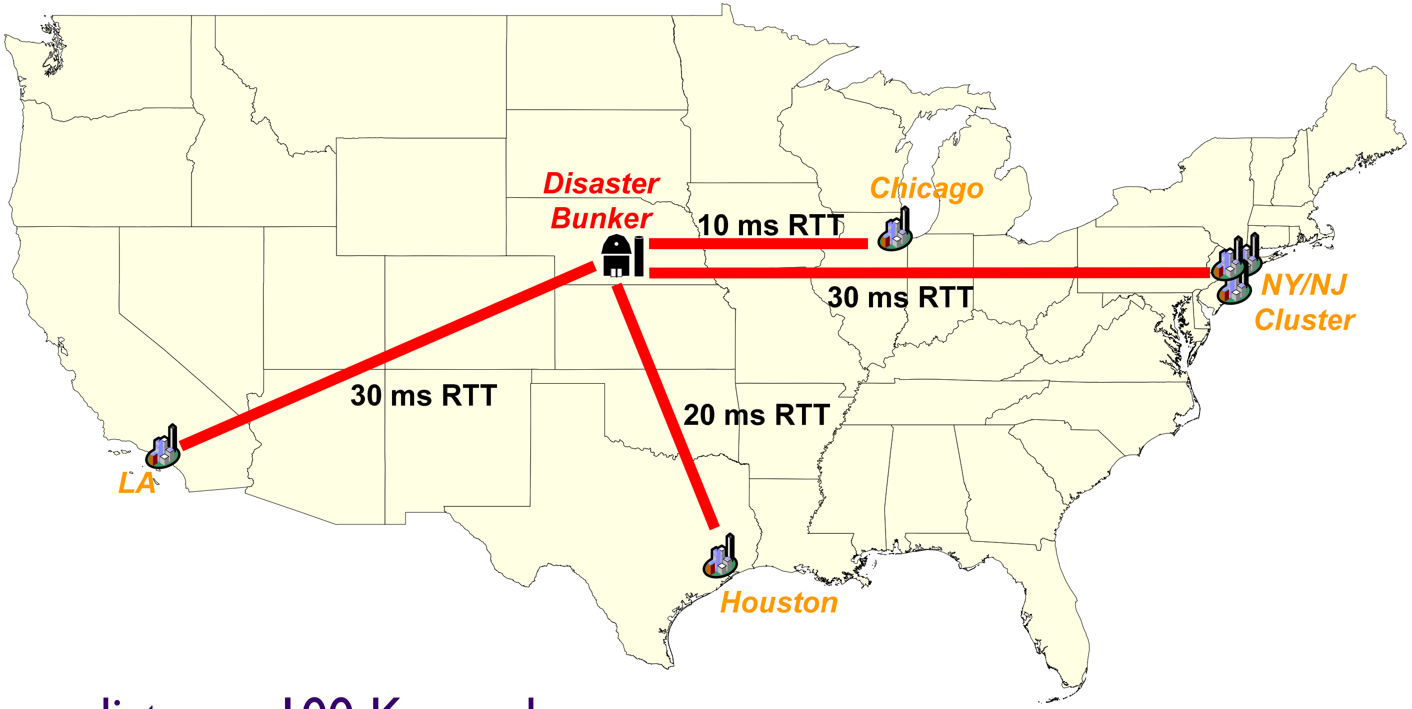
- Servers accessing Storage across switched Local Area Network
  - Remember **SANs** can span all distances
- 100s of meters max diameter
  - This is the effective range of direct access at high bandwidth and low latency
  - Direct cabling
  - Short range optics allowed
  - Can use copper interfaces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

The MAN

New Jersey
New York City
Philadelphia

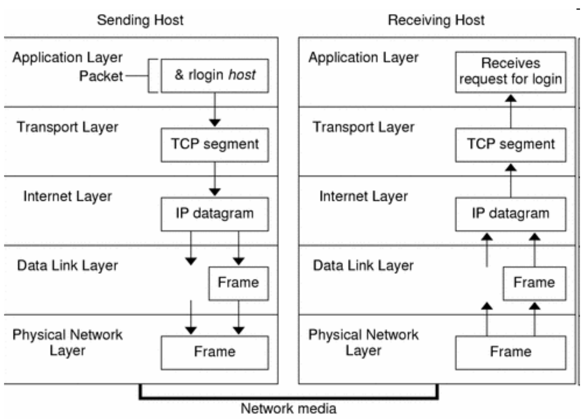
- 150-200 Km max diameter
  - Effective range of synchronous applications
  - Increasing longer range deployments (100Km+)
- Can be a short distance (1Km or so)
  - i.e. to the next building
- 5-10 Km separation between sites common
- Commonly used infrastructure
  - Long optics single hop (40-80Km max range)
  - Dark Fibre
  - DWDM/CWDM
  - SONET/SDH (TDM)
- Protocols
  - FC direct connect common at shorter ranges
  - FCIP comes in at longer ranges
  - iSCSI and NAS direct access

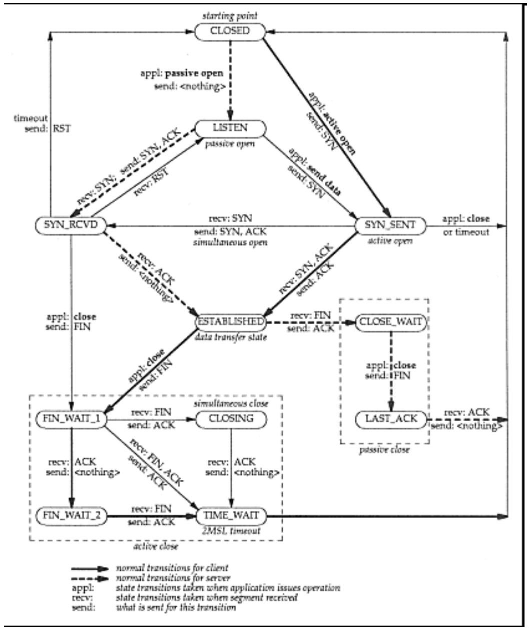
The WAN

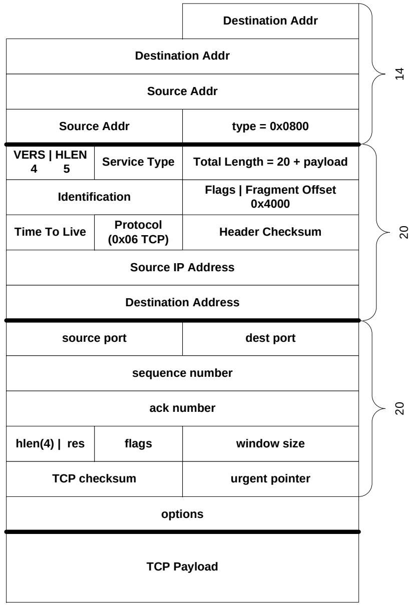


- Long distance 100 Km and up
  - High Latency
  - Usually lower bandwidth and fewer connections
- Data Backup and Replication usually Asynchronous
- Specialized gateways optimize traffic and provide distance offload

Tour of the TCP/IP Protocol







		Destination Addr		14
Destination Addr				
Source Addr				
Source Addr		type = 0x0800		
VERS | HLEN 4 5	Service Type	Total Length = 20 + payload		20
Identification		Flags | Fragment Offset 0x4000		
Time To Live	Protocol (0x06 TCP)	Header Checksum		
Source IP Address				
Destination Address				
source port		dest port		20
sequence number				
ack number				
hlen(4) | res	flags	window size		
TCP checksum		urgent pointer		
options				
TCP Payload				

Characteristics of TCP

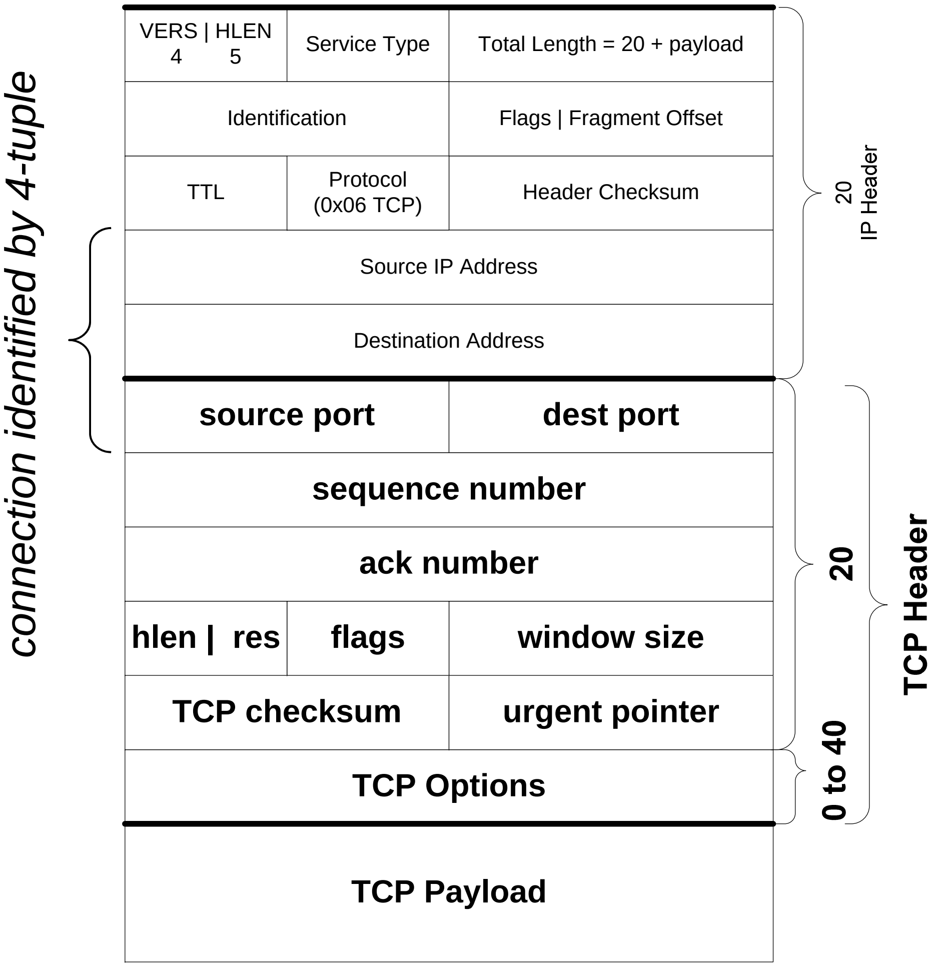
- For Storage Networking TCP is Critical (FCIP, iSCSI)
- Connection Oriented
  - Full Duplex
  - Byte Stream (to the application)
  - Port Numbers identify application/service endpoints within an IP address
  - Connection Identification: IP Address pair + Port Number pair ('4-tuple')
  - Well known port 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

The TCP/IP Protocols and Storage

- 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 and high bandwidth magnify the effects of drops
  - However, Many of TCP/IP drawbacks can be mitigated
    - Some changes only improve TCP behavior
      - Better resolution for TCP timers
      - SACK
    - Some have a possible negative effect on other traffic
      - Removing or altering congestion avoidance algorithms
- For Short latencies in flow controlled networks much of the TCP/IP protocol is not needed but it does not hurt much either
  - I will explain this as we go along

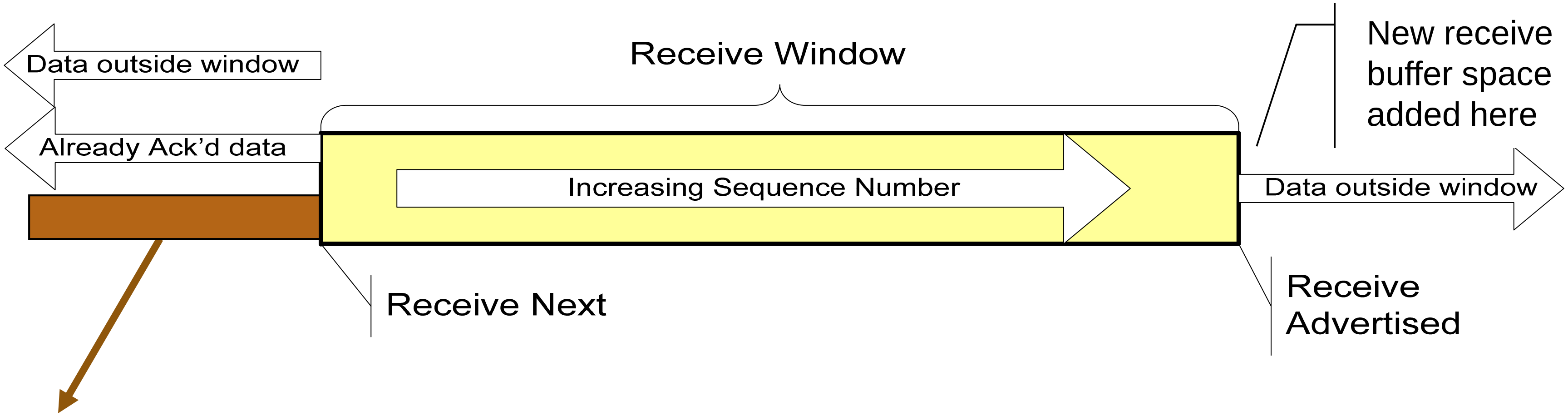
TCP Header

- Source Port Number
- Destination Port Number
- Sequence Number
- ACK Number
- Header Length
- Flags
  - SYN
  - FIN
  - RST
  - ACK
  - PSH
  - URG
- Window Size
- Checksum
- Urgent pointer



VERS | HLEN 4 5	Service Type	Total Length = 20 + payload
Identification		Flags | Fragment Offset
TTL	Protocol (0x06 TCP)	Header Checksum
Source IP Address		
Destination Address		
source port		**dest port**
sequence number		
ack number		
hlen | res	**flags**	**window size**
TCP checksum		**urgent pointer**
TCP Options		
TCP Payload		

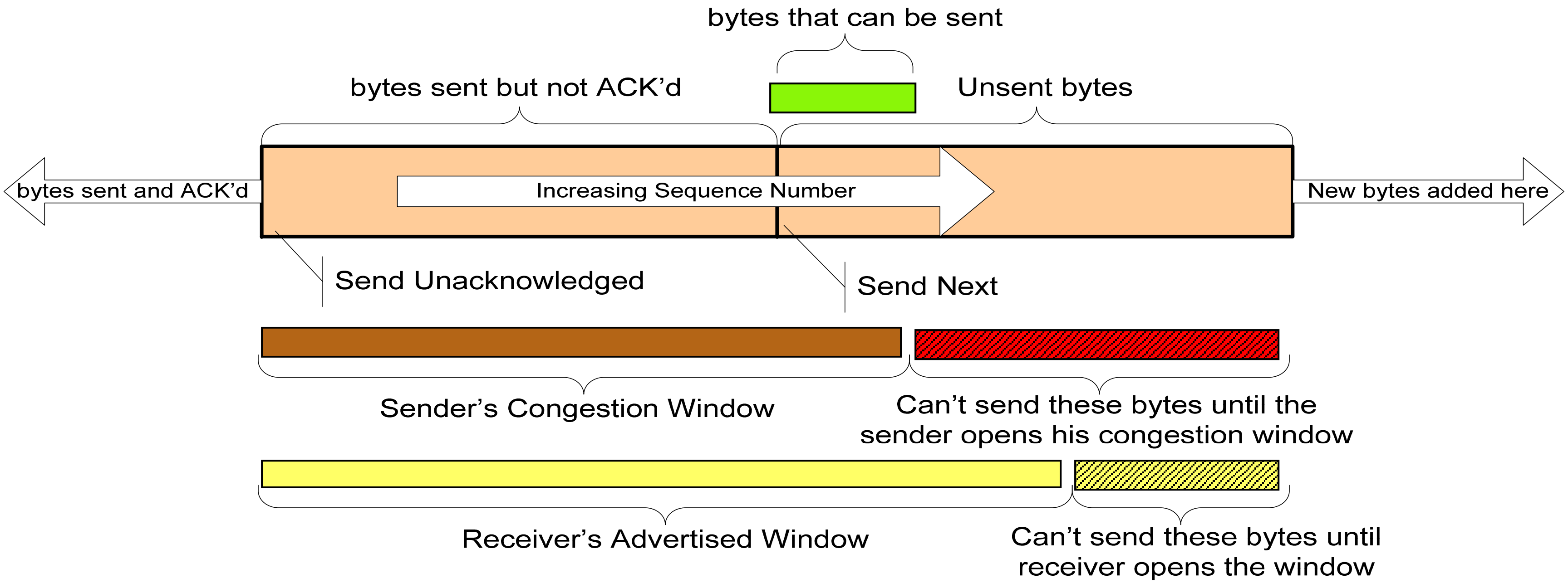
TCP Receive Window



These bytes have been sent to the application

- Sliding window protocol
  - Receiver advertises a window to the sender
  - Out of Order arrival and reassembly
  - Duplicate segments or out of window segments are discarded
  - Avoid the silly window syndrome: don't advertise too small windows
- Sender's Persist Timer generates Window Probes to recover from lost window updates

TCP Transmit Congestion Controls



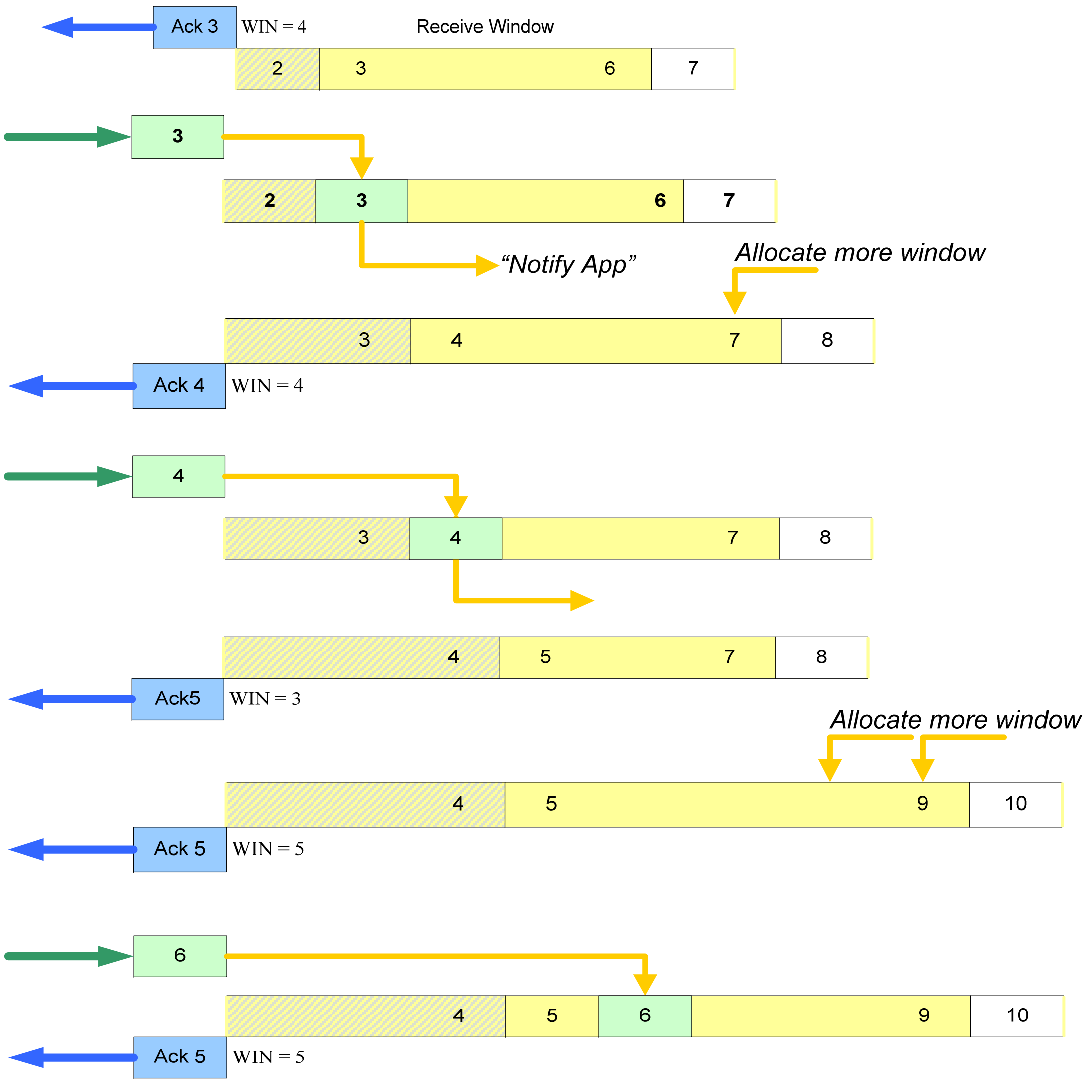
bytes that can be sent
bytes sent but not ACK'd
Unsent bytes
bytes sent and ACK'd
Increasing Sequence Number
New bytes added here
Send Unacknowledged
Send Next
Sender's Congestion Window
Can't send these bytes until the sender opens his congestion window
Receiver's Advertised Window
Can't send these bytes until receiver opens the window

- Data Stream chopped into Segments: Sent as IP Datagrams
- Congestion Window (cwnd) Uses AIMD (Additive Increase, Multiplicative Decrease)
- Nagle Algorithm (RFC 896): don't send less than one segment of data unless all sent data has been ACK'd

TCP ACK Schemes

- Duplicate ACK
  - Sent when segment is received out of order
  - May indicate a missing segment to the sender
- Delayed ACK
  - Do not send an ACK right away, wait a short time to see if Additional segments arrive which can also be ACK'd
- ACK/N
  - Wait for a specific number (N) of segments to arrive before sending an ACK
  - Send anyway after a short time interval

TCP Receiver Illustration



Everything previous ACK'd
And ready to receive…

TCP Segment Received

Receive Buffer allocated
and ACK sent

TCP Segment Received

Receive Buffer NOT allocated
But ACK still sent

Receive Buffer added
causes Window Update
(not considered a duplicate ACK)

Out of Order segment
Generates a Duplicate ACK

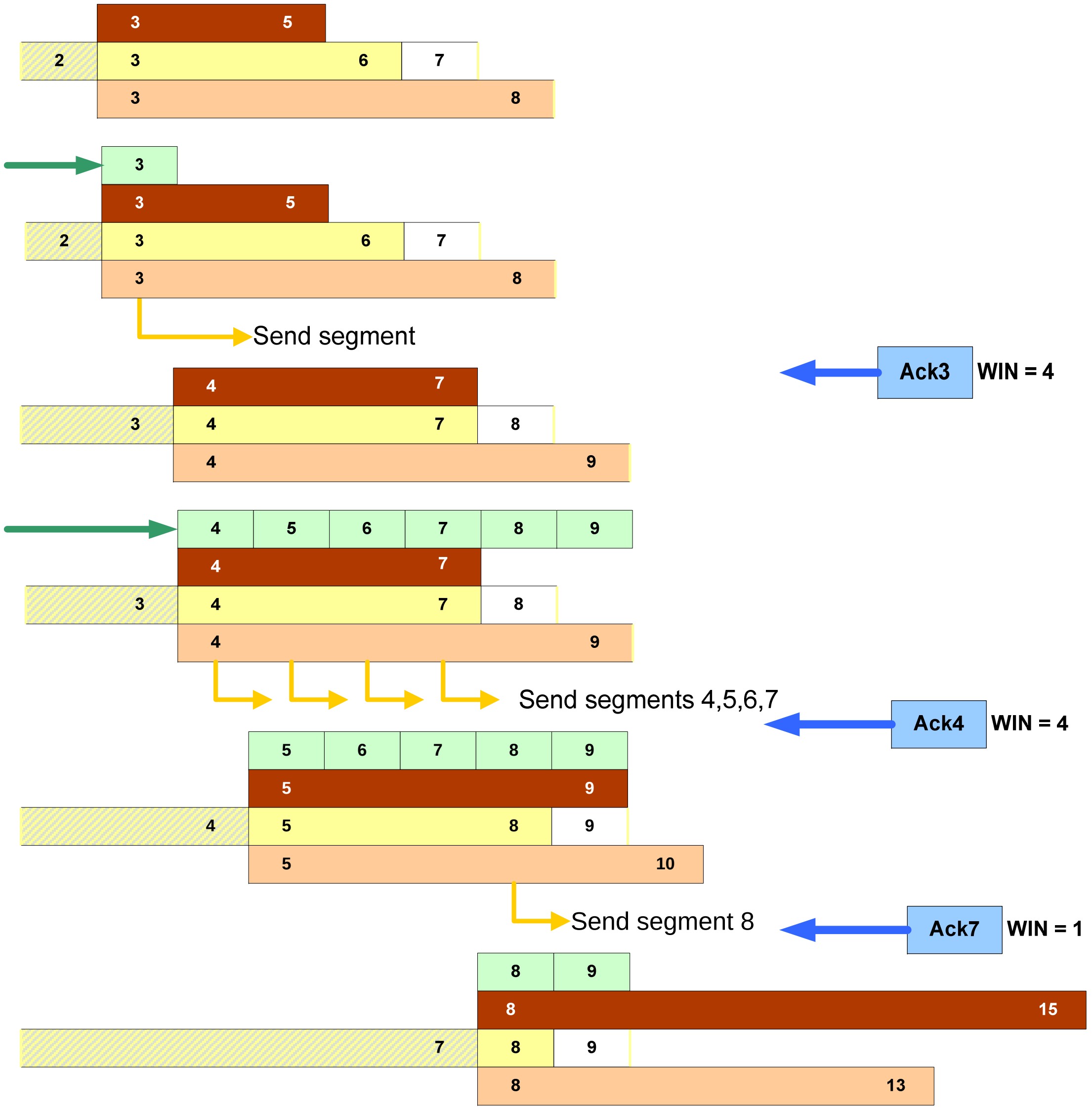
TCP Transmitter Illustration

🟥 *cwnd*

🟨 *Advertised receive window*

🟧 *Allowed transmit buffer*

🟩 *Data to send*



Send segment

Ack3 WIN = 4

Send segments 4,5,6,7

Ack4 WIN = 4

Send segment 8

Ack7 WIN = 1

Application has data

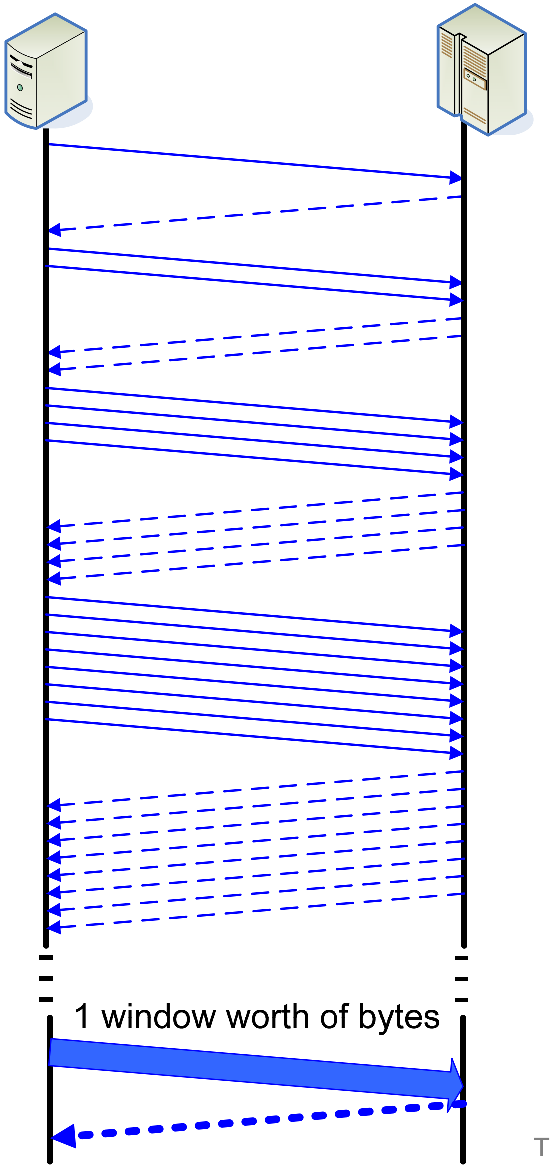
ACK Received

Application has more data
Can’t send all of it due to limits

ACK Received
Updates allow sending
But still unsent data

ACK Received but without window growing

Slow Start



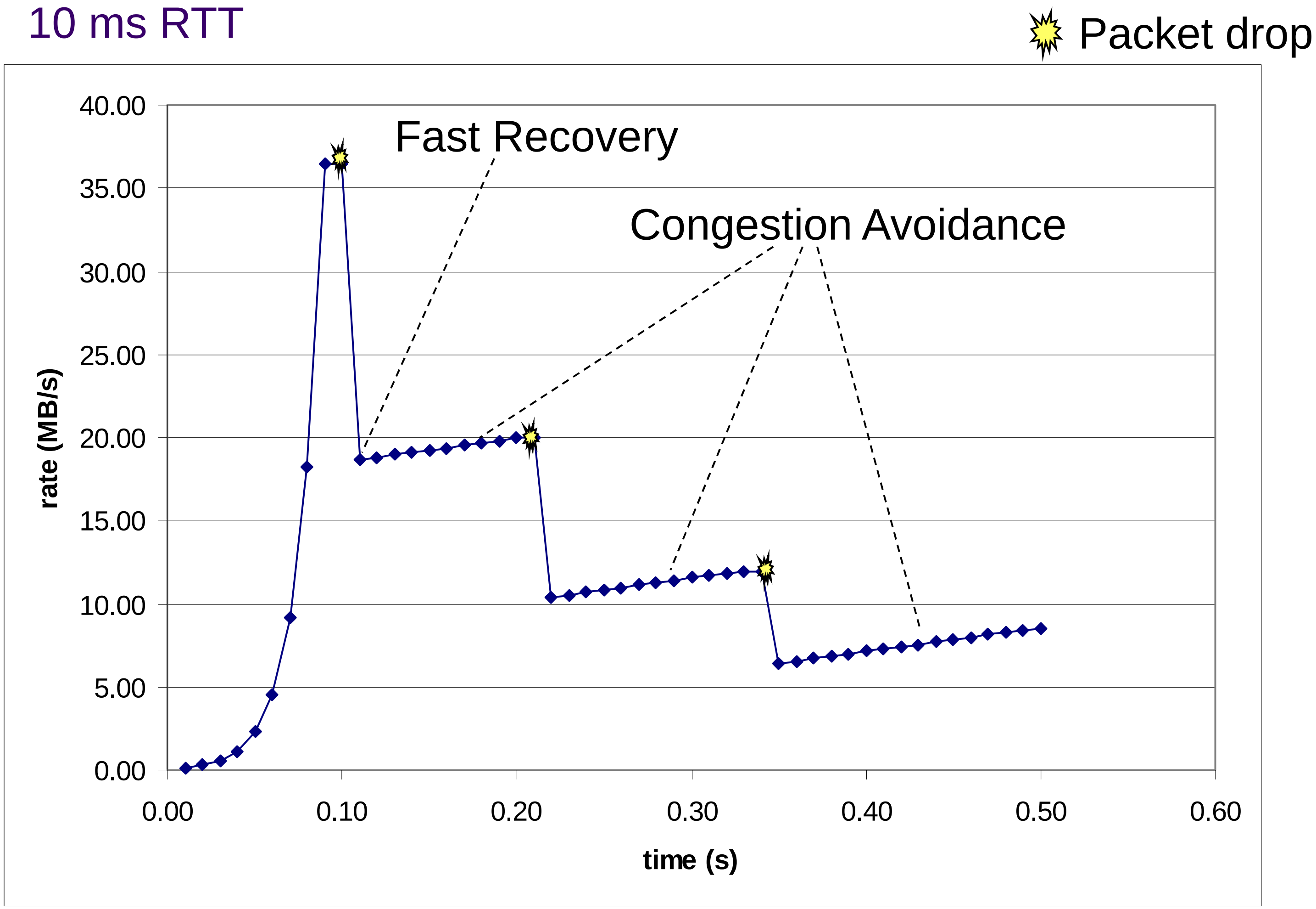
- Rate of packet injection into the network equals to rate which ACKs are received
- Leads to exponential sender cwnd ramp
- Exponential Ramp stops when
  - The limit of the receiver's advertised window is reached
    - (TCP can only move one window's worth of data per RTT)
  - The limit of the sender's un-acknowledged data buffering or outstanding data is reached
  - The limit of the network to send data is reached (network saturation)
  - A congestion event occurs

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
- Packet (segment) loss can occur for several reasons
  - Congestion
  - Ethernet/IP proactive flow control schemes (RED)
  - Faulty equipment
  - Uncorrectable bit errors

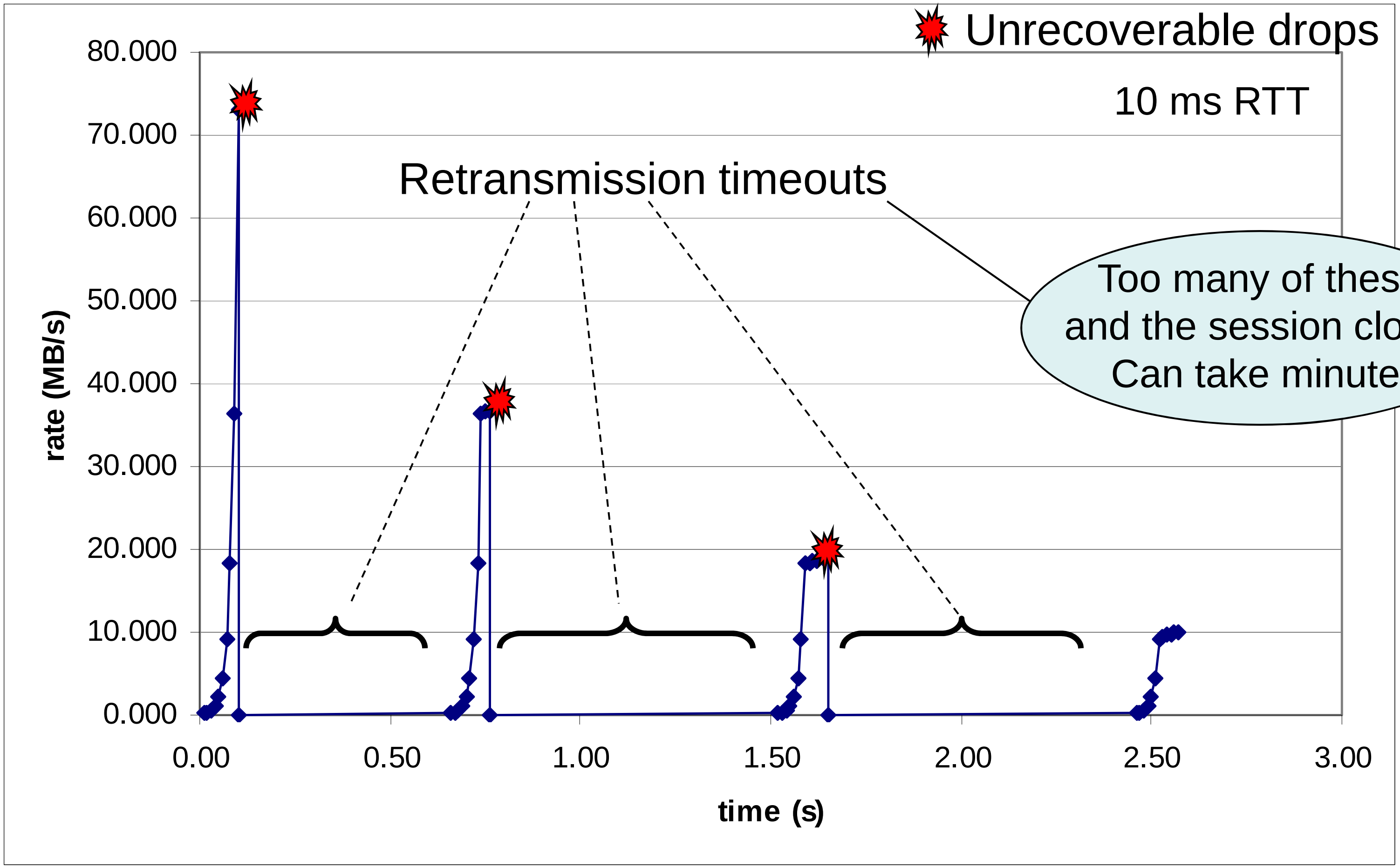
- When packet loss does occur it can be extremely detrimental to the throughput of the TCP connection
  - Extent of disruption determined by the pattern of drops
  - There are TCP features and modifications which mitigate effects of packet loss

TCP Fast Retransmit, Fast Recovery



- 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.

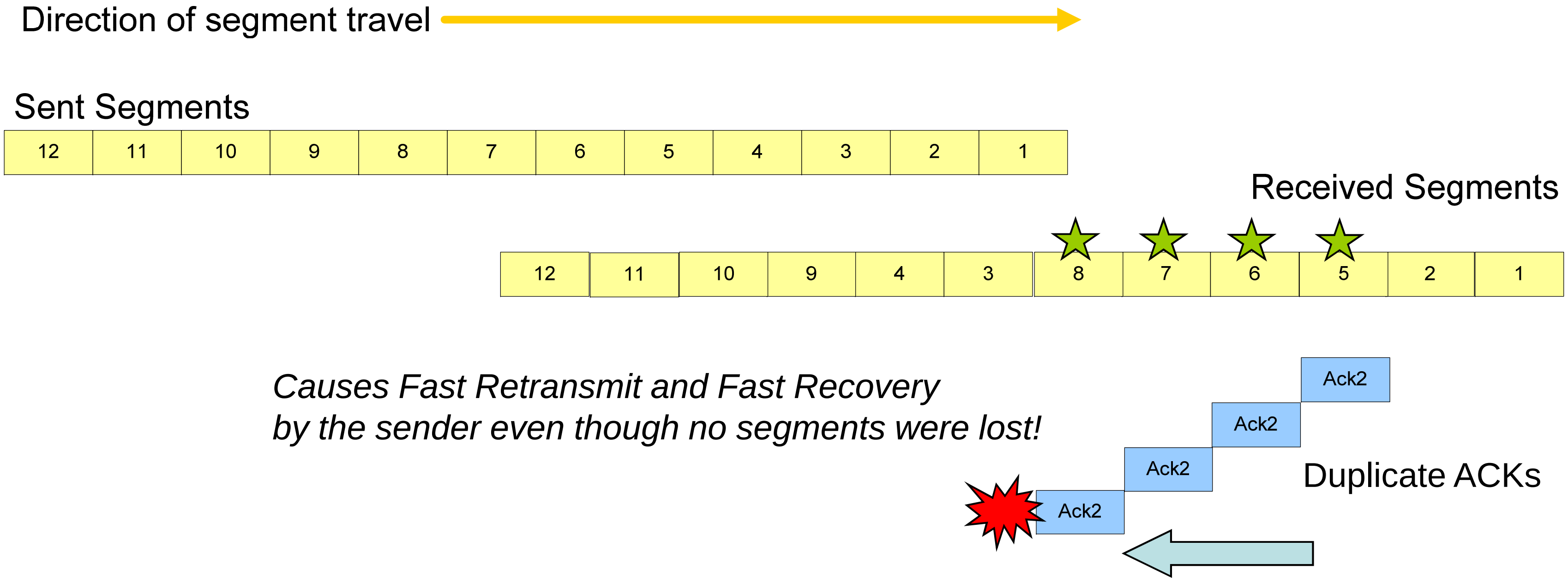
TCP Retransmission Timeout



- 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

Network Reordering

- Networks which dynamically load balance cause excessive false congestion events due to extensive reordering and TCP normally only uses a value of 3 for the duplicate ACK threshold



Direction of segment travel

Sent Segments

12	11	10	9	8	7	6	5	4	3	2	1

Received Segments

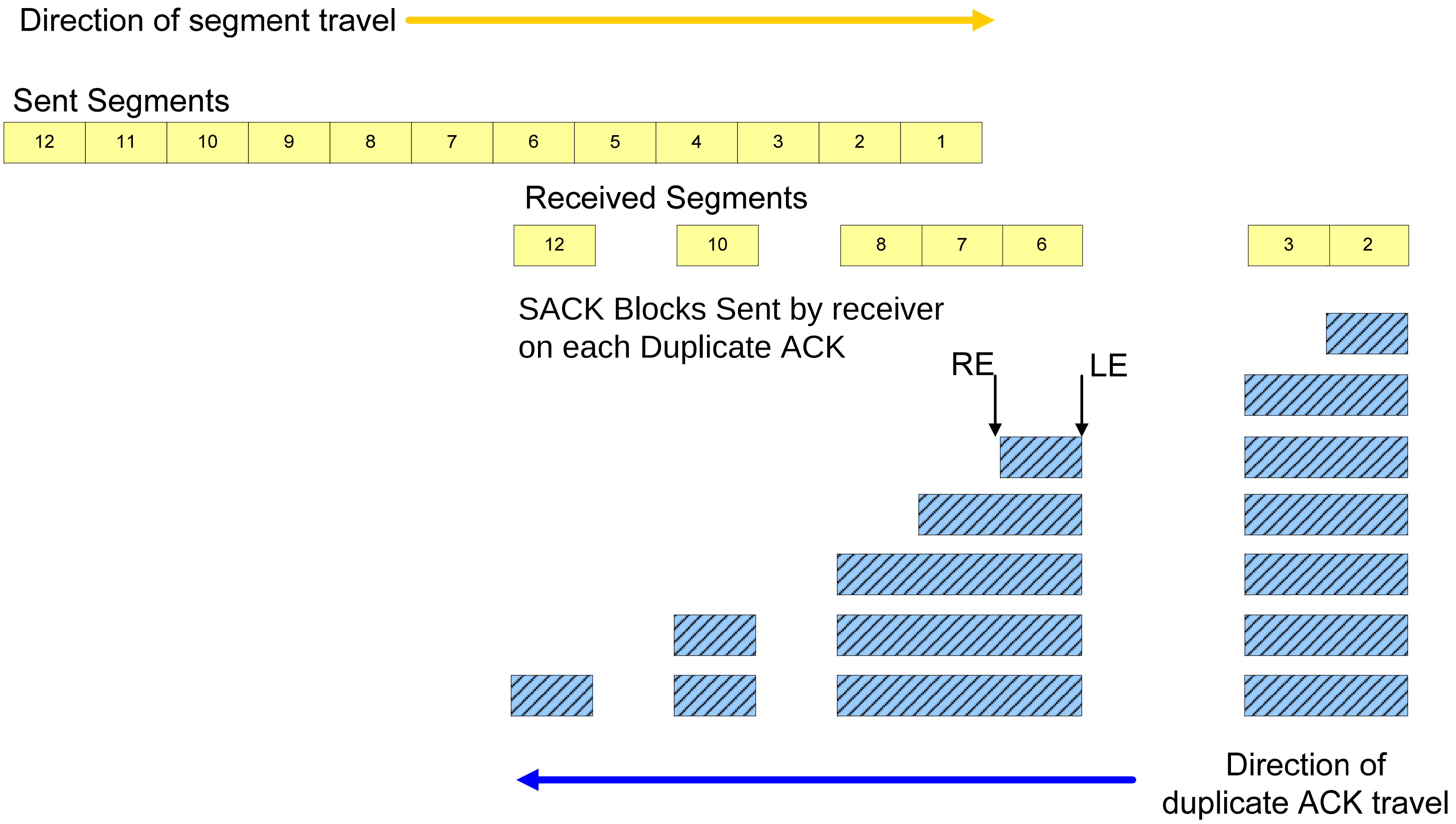
12	11	10	9	4	3	8	7	6	5	2	1

Causes Fast Retransmit and Fast Recovery by the sender even though no segments were lost!

Duplicate ACKs

- Can be helped by ignoring more duplicate ACKs before Fast Retransmit/Recovery
- Must be careful not to miss a retransmit that should have gone out

Selective Acknowledgement (SACK)



Direction of segment travel

Sent Segments

12 11 10 9 8 7 6 5 4 3 2 1

Received Segments

12 10 8 7 6 3 2

SACK Blocks Sent by receiver on each Duplicate ACK

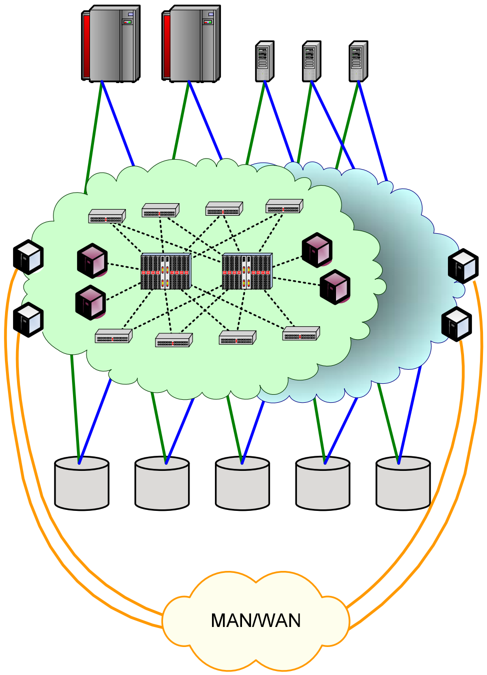
RE

LE

Direction of duplicate ACK travel

- Both sides track the current list of holes using additional TCP state
- Allows the sender to only retransmit 4, 5, 9, 11and he can fill in multi-segment holes!

TCP/IP in the LAN & MAN



New Jersey
New York City
Philadelphia

Ethernet Transport

- Layer 2 interconnect
- Speeds from Mb →multi-Gb
  - 100Mb, 1Gb, 10Gb, 40 Gb, and on
- Carries (for purposes of this discussion)
  - IP traffic (TCP, UDP)
    - iSCSI
    - NAS
    - server communication
  - FCoE

Protocols

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

CEE (Converged Enhanced Ethernet)

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

Effects of CEE on Storage Traffic

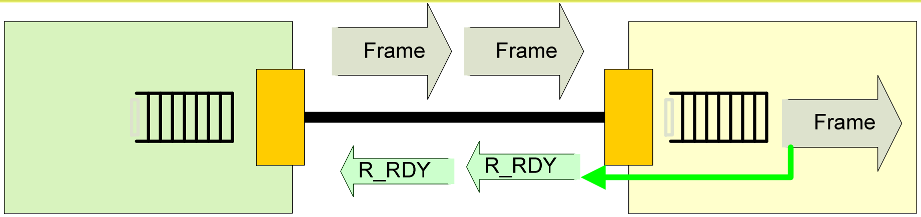
- Storage and non-storage traffic separated on same physical infrastructure
- Storage traffic can have link level flow control
- Some level of congestion management and load balancing by the network
- Capabilities exchange for interoperability

Effects of CEE on Storage Traffic

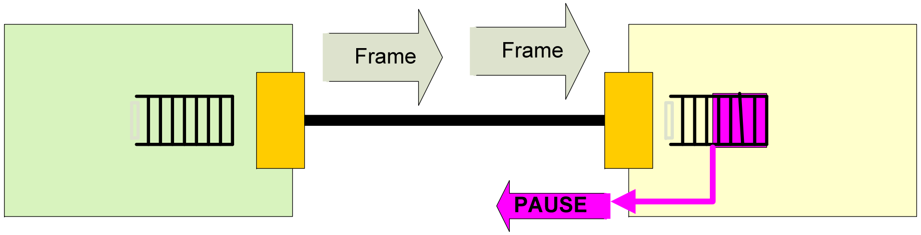
- Can make a Ethernet network Priority look like the equivalent Local FC network
  - link level flow control everywhere
  - Same performance advantages and disadvantages

- TCP/IP has no special considerations in a CEE based local flow controlled network other than a small amount of processing to deal with ACKs and slightly different header lengths
  - As far as the traffic behavior is concerned everything else is equivalent between SCSI over FC and SCSI over TCP/IP

Credit vs Pause Based Flow Control



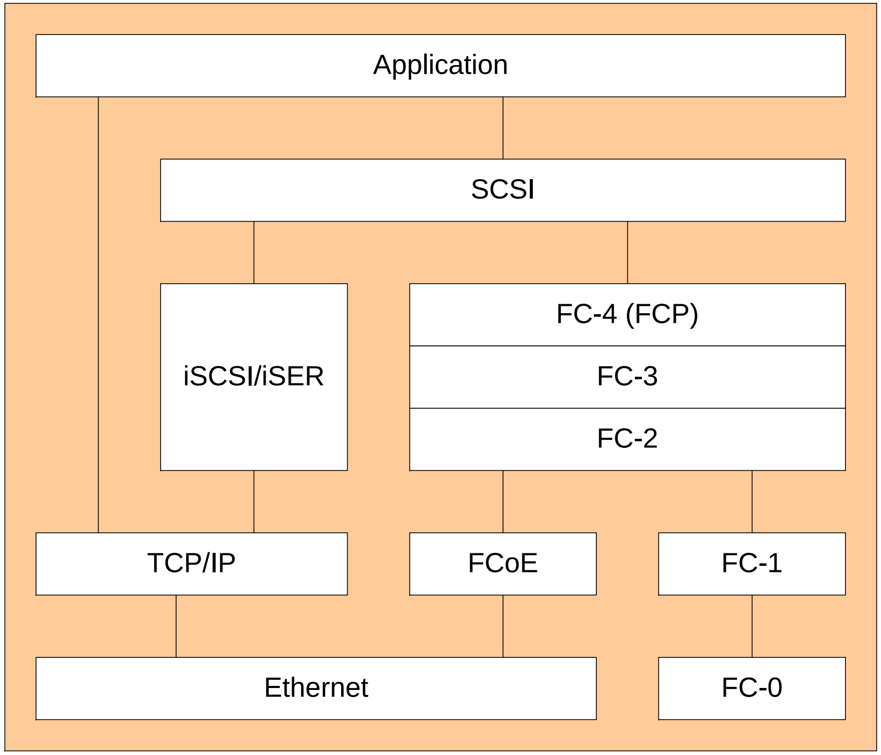
- **FC Credit based link level flow control**
  - A credit corresponds to 1 frame independent of size
    - (1 credit needed per 1 km separation for 2G FC)
  - Sender can only xmit frames if he has credit sent from the receiver (R_RDYs)
  - Amount of credit supported by a port with average frame size taken into account determines maximum distance that can be traversed
    - ***If the credit is exceeded the maximum possible sustained throughput will be reduced***



- **Pause Frame Based Link Level Flow control**
  - When the sender needs to be stopped the receiver sends a frame to notify the sender
  - For lossless behavior the receiver must absorb all the data in flight
  - This puts a hard limit based upon the receiver buffer on the distance for storage traffic across a direct connect Ethernet
    - ***If the buffer is overrun then frames can be dropped***

FC/FCoE vs iSCSI in the Server

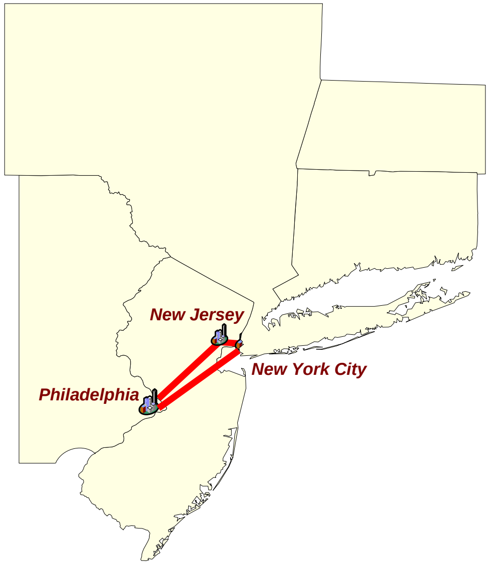
- Server Stack and Interfaces
  - Must compare apples to apples:
    - FC HBAs have command (FC Exchange) handling offload where a standard NIC does not, therefore CPU utilization on the server will be different
  - Cards that have TOE and iSCSI offload have equivalent properties (and cost) to an FC HBA
  - CNA’s (Converged Network Adaptors) have FCoE offload equivalent to an FC HBA.

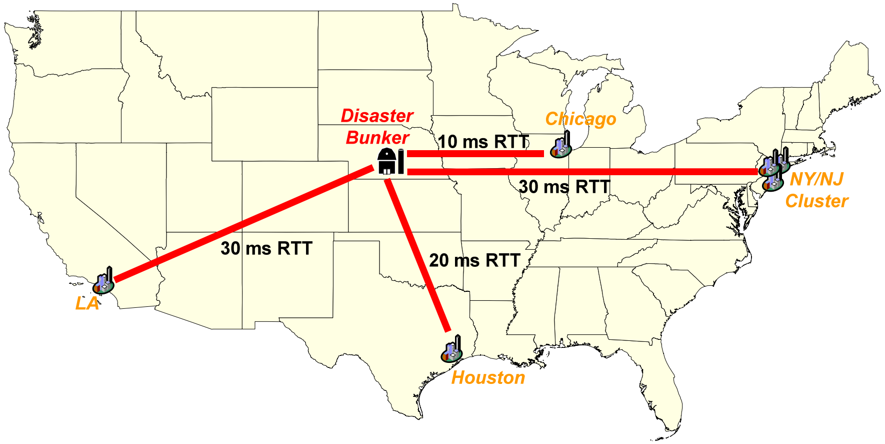


Server Software Layering

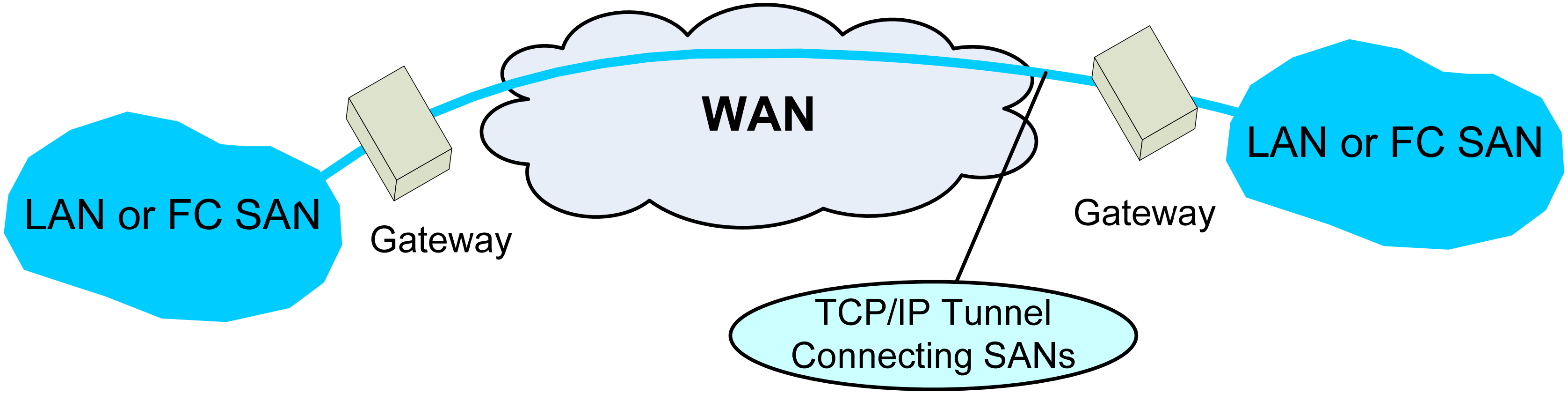
- Server Virtualization does not change the fundamental layering comparison in this picture. It can add additional layers for both or it can move where some of the pieces are done.

TCP/IP in the MAN and WAN





WAN Extension Gateways

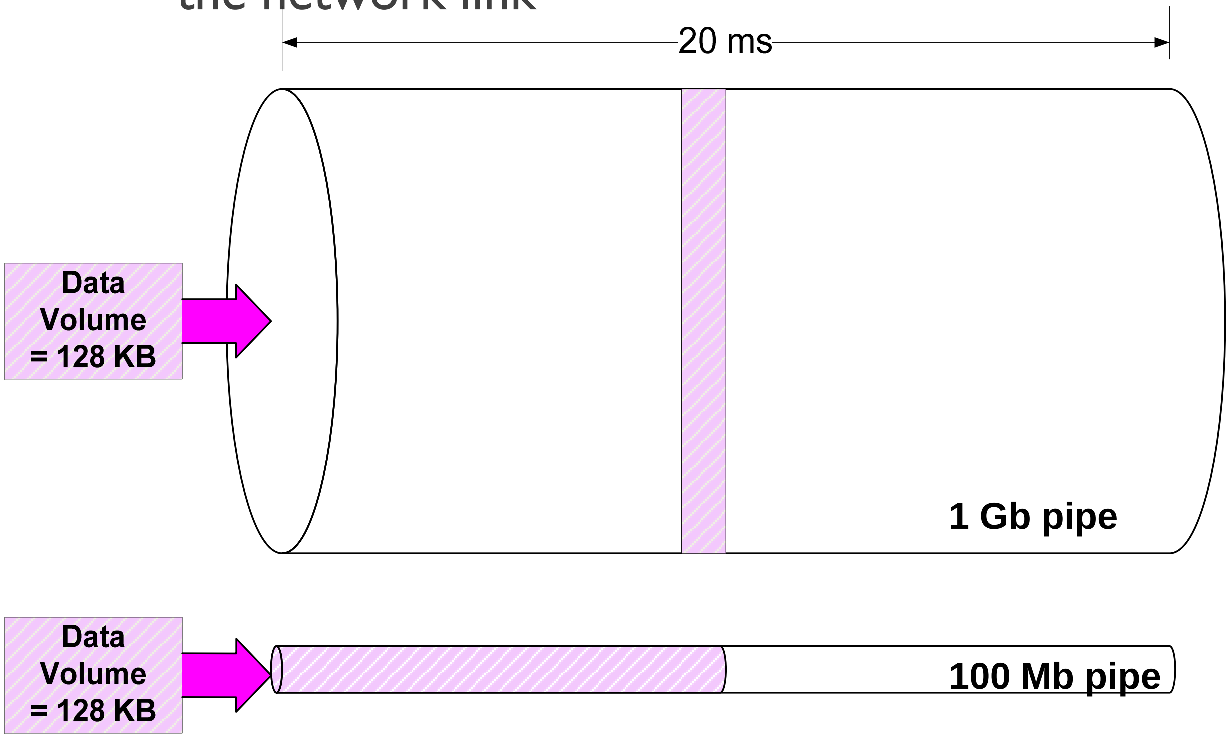


- FCIP for SAN extension
- TCP/IP to TCP/IP tunnel
- Application Acceleration/Compression Engine

- TCP/IP implementation and behavior important
  - Large buffers to implement long distance connections
  - Direct modifications to TCP/IP to better handle packet loss and ramp times
  - Optimizations such as compression and protocol acceleration

Long Fat Networks (LFNs)

- LFNs have a large bandwidth-delay product
  - Bandwidth-delay product = amount of data 'in flight' needed to saturate the network link

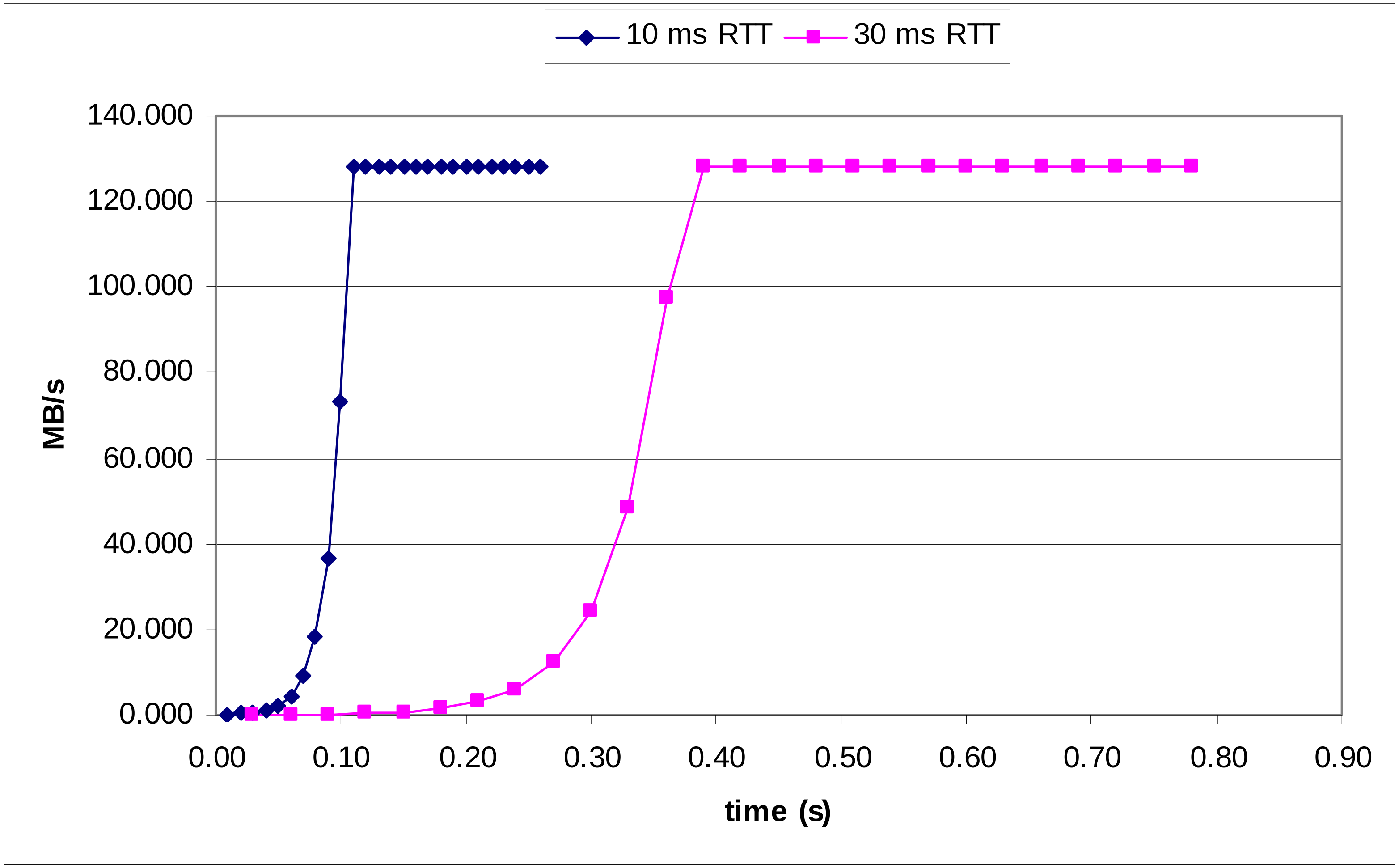


For this example we need 2.56 MB of both transmit data and receive window to sustain line rate

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

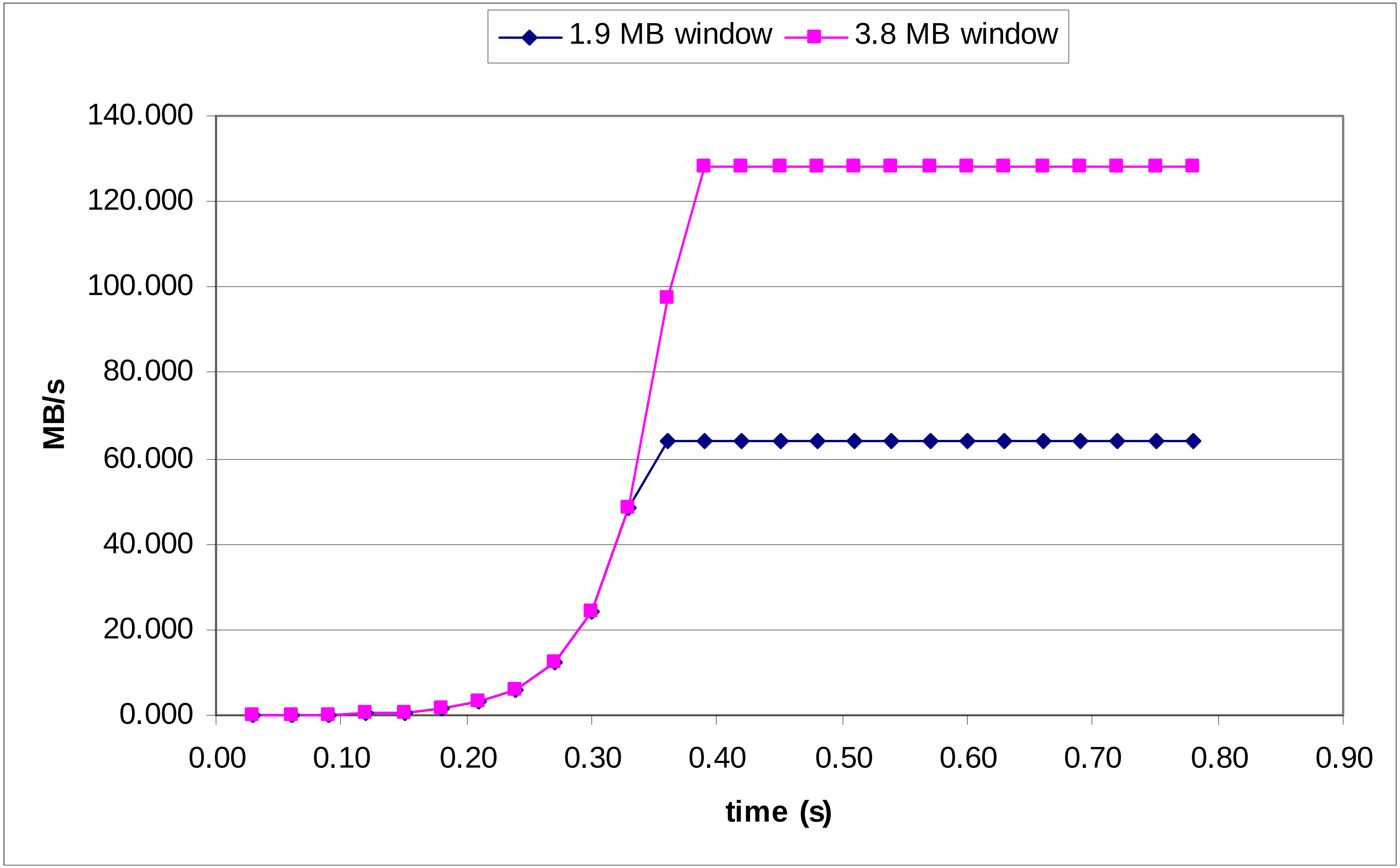
1 ms = 128 KB buffering at 1Gb/s
1 ms = 100 Km a maximum separation

RTT and Receive window plateau



- For window sizes big enough to support line rate

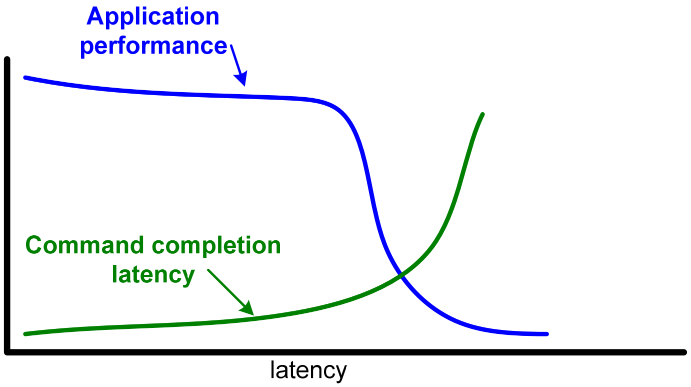
LFNs and Receive window plateau



- For 30 ms RTT

Latency

- Command Completion Time is important
- Contributing Factors: (sum 'em all up!)
  - Distance (due to 'speed of light') - latency of the cables
    - (2x10^8 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

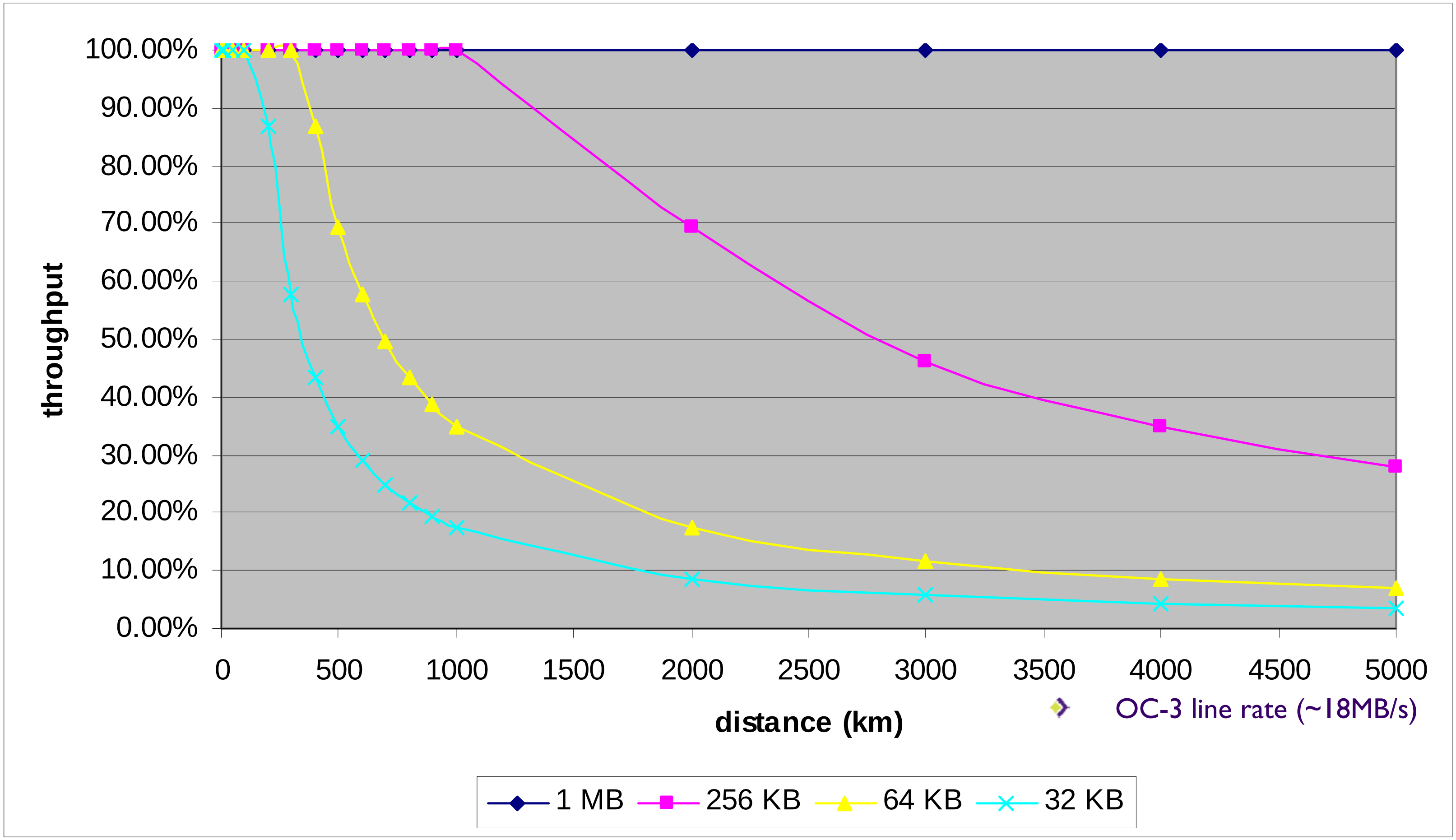


- A complicating factor is the I/O pattern and application configuration
  - Some patterns and applications hide latency much better than others
    - Good: File Servers and NAS
    - Bad: transactional database with heavy write to transaction logs

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

Throughput Droop due to distance



Curves represent varying amounts of buffer space or outstanding data

LFNs and Packet Loss

- Effect of packet drops in the network magnified
- Slow recovery due to large RTT
  - we'll explore this in detail
- Many more hops and greater variety of equipment increases chances of problems due to design flaws, incorrect configuration, failures

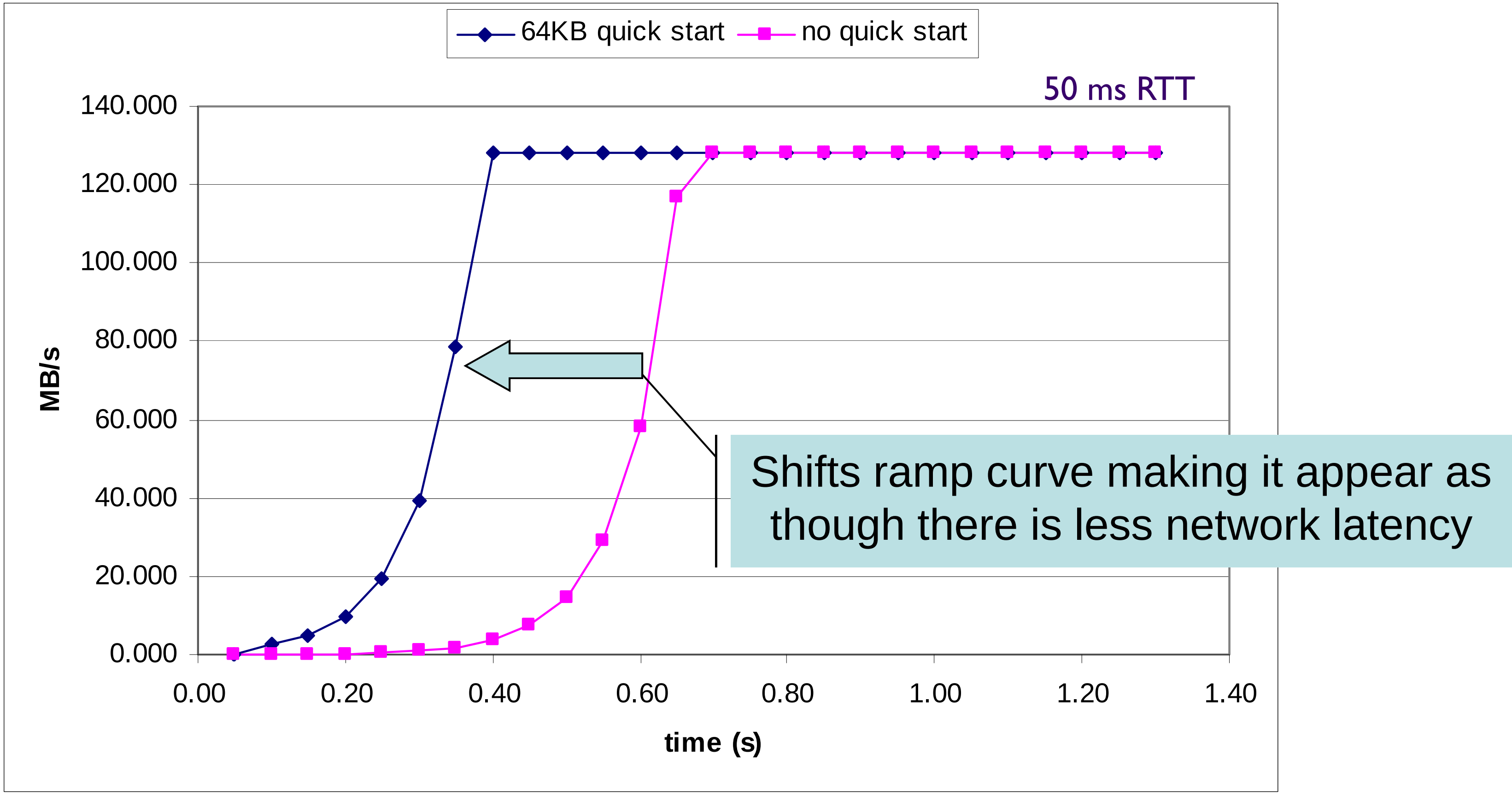
High Speed Long Distance TCP/IP Projects

- Industry Proprietary Implementations
  - WAN Acceleration Companies
  - Distance Extension for SAN across TCP/IP
  - Distance Extension for TCP/IP
- Scalable TCP (STCP)
- High Speed TCP (HS-TCP)
- FAST TCP
- H-TCP

Block Storage TCP/IP Across WAN

- Scaled receive windows
- Quick Start
- Modify Congestion Controls
- Deal with network reordering better
- 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

Quick Start



Shifts ramp curve making it appear as though there is less network latency

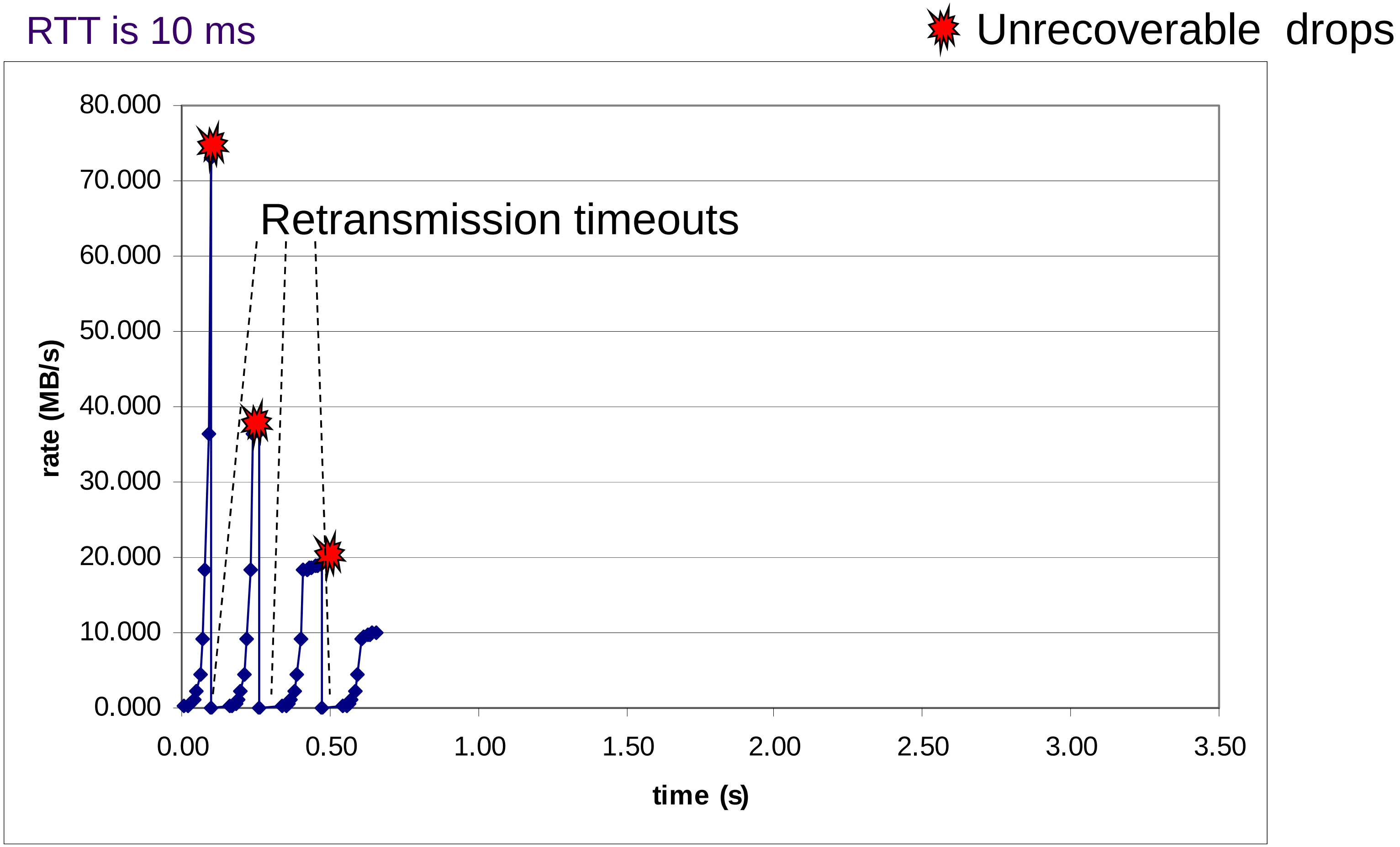
- Two common meanings
  - Quick Start means giving the sender cwnd a head start
  - Quick Start means giving the sender no cwnd limit at start

Congestion Control Modifications

- Don’t want to break TCP’s fundamental congestion behavior
- Modify Congestion Controls
  - Different ramp up or starting value for slow start (aka QuickStart)
  - Less reduction during fast recovery
  - Ignore or reduce the effects of congestion avoidance
  - Eifel Algorithm
- Modify Fast retransmit and Fast Recovery detection scheme
  - Ignore a larger fixed number of duplicate ACKs but backstop with a short timer
  - RFC 4653 – TCP-NCR (non-congestion robustness)
    - Retransmission detection is based upon cwnd of data leaving network instead of a fixed limit

Improve Retransmission Timeouts

RTT is 10 ms

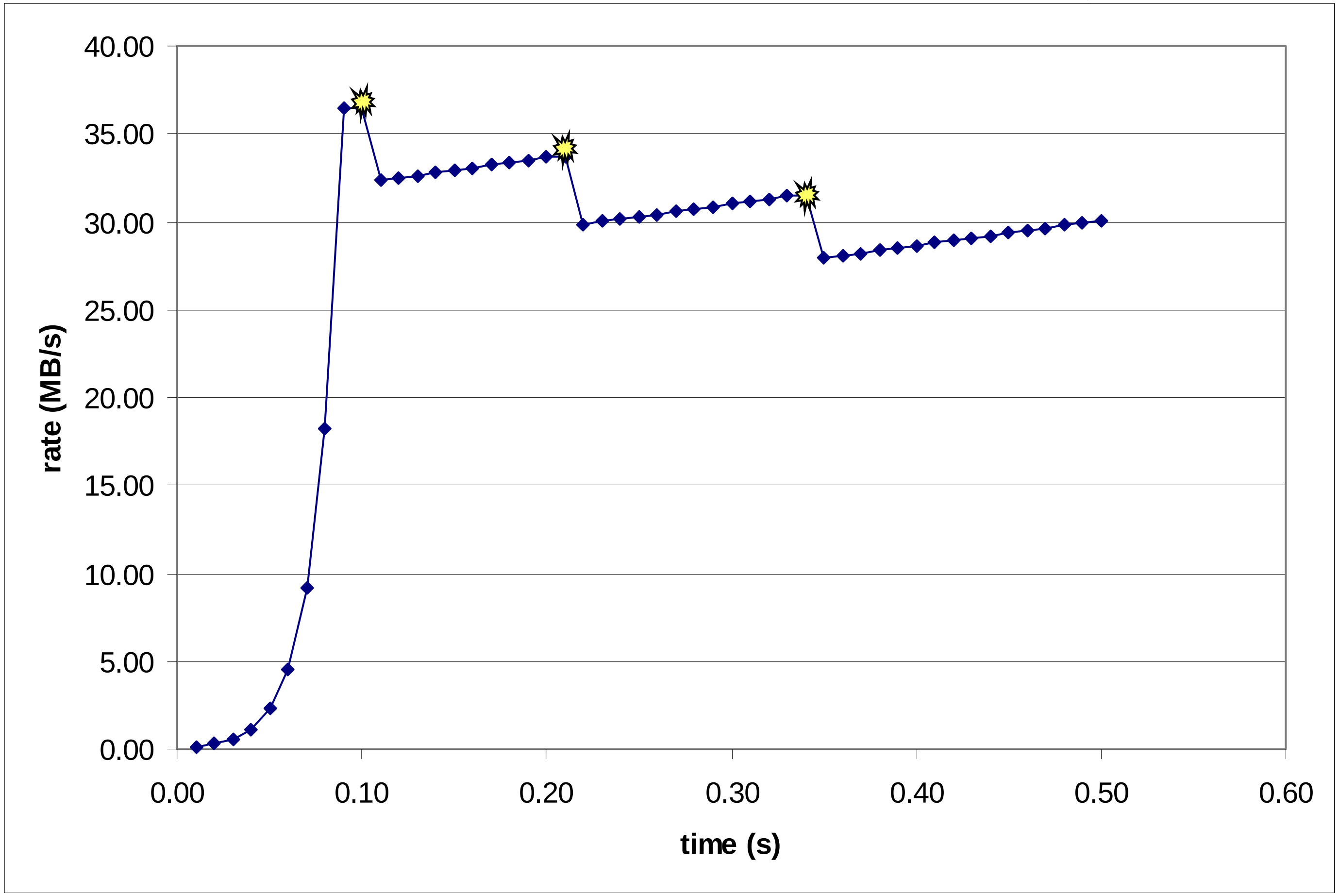


- If we have a finer grained TCP clock we can estimate timeouts with much better accuracy
- Instead of 500 ms to 1s timeouts we could for example have 50-60 ms timeouts
- Also helps the long TCP connection close times (reduced from minutes to seconds)

Change Fast Recovery Threshold

10 ms RTT

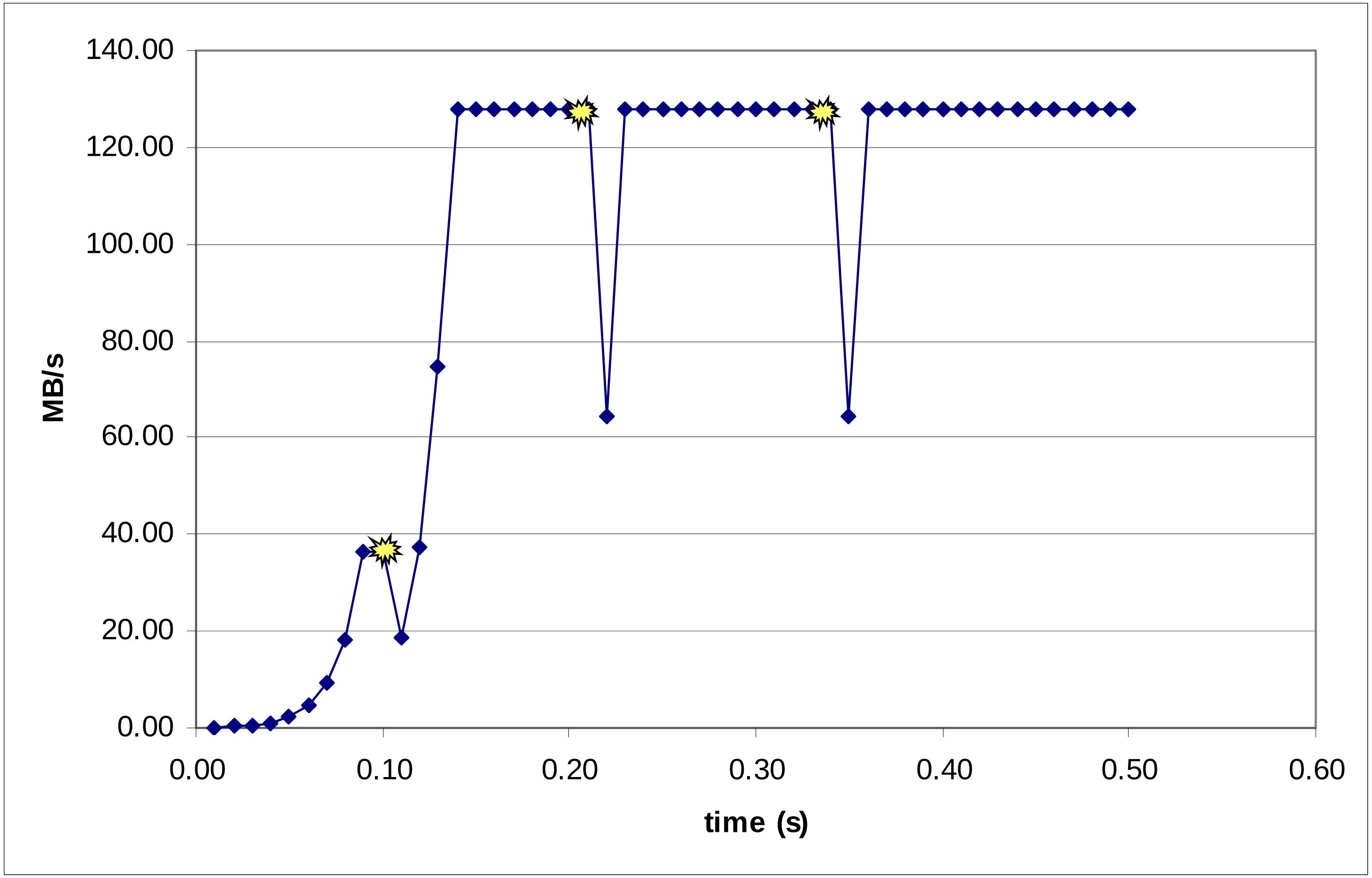
✸ Packet drop



- During fast recovery reduce sender cwnd by 1/8 instead of 1/2

Remove Congestion Avoidance

✸ Packet drop



- Retain exponential ramp even after congestion avoidance should kick in
- 10 ms RTT

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

- For the CEE based LAN
  - TCP/IP based storage networks will behave as the equivalent FC Network

- For the MAN and WAN
  - TCP/IP's congestion controls and recovery of lost segments can cause problems for block storage
  - However, Many of TCP/IP drawbacks can be mitigated
    - Some changes improve TCP behavior without side effects
      - SACK
    - Some changes have a possible negative effect on other traffic
      - For example removing congestion avoidance completely

Q&A / Feedback

- 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

Appendix: Relevant Internet RFCs

- RFC 793 – Transmission Control Protocol
- RFC 896 – Congestion control in IP/TCP internetworks
- RFC 1122 – Requirements for Internet Hosts - Communication Layers
- RFC 1323 – TCP Extensions for High Performance
- RFC 2018 – TCP Selective Acknowledgment Options
- RFC 2140 – TCP Control Block Interdependence
- RFC 2581 – TCP Congestion Control
- RFC 2861 – TCP Congestion Window Validation
- RFC 2883 – An Extension to the Selective Acknowledgement (SACK) Option for TCP
- RFC 2988 – Computing TCP's Retransmission Timer
- RFC 3042 – Enhancing TCP's Loss Recovery Using Limited Transmit
- RFC 3124 – The Congestion Manager
- RFC 3155 – End-to-end Performance Implications of Links with Errors
- RFC 3168 – The Addition of Explicit Congestion Notification (ECN) to IP
- RFC 3390 – Increasing TCP's Initial Window
- RFC 3449 – TCP Performance Implications of Network Path Asymmetry
- RFC 3465 – TCP Congestion Control with Appropriate Byte Counting (ABC)
- RFC 3517 – A Conservative Selective Acknowledgment based Loss Recovery Algorithm for TCP
- RFC 3522 – The Eifel Detection Algorithm for TCP
- RFC 3649 – HighSpeed TCP for Large Congestion Windows
- RFC 3742 – Limited Slow-Start for TCP with Large Congestion Windows
- RFC 3782 – The NewReno Modification to TCP's Fast Recovery Algorithm
- RFC 4015 – The Eifel Response Algorithm for TCP
- RFC 4138 – Forward RTO-Recovery (F-RTO)
- RFC 4653 – Improving the Robustness of TCP to Non-Congestion Events
- RFC 4782 – Quick-Start for TCP and IP
- RFC 4828 – TCP Friendly Rate Control (TFRC): The Small-Packet (SP) Variant

Appendix: Miscellaneous TCP Features

- MTU Discovery
  - Packets probe the network path to determine maximum packet size
- Timestamp Option
  - enables PAWS
  - allows RTT calculation on each ACK instead of once per window
  - produces better smoothed averages
  - Timer rate guidelines: 1 ms <= period <= 1 second
- PAWS: protection against wrapped sequences
  - In very fast networks where data can be held, protects against old sequence numbers accidentally appearing as though they are in the receiver’s valid window
  - uses timestamp as 32-bit extension to the sequence number
  - requires that the timestamp increment at least once per window

Appendix: TCP Options

Kind	*Len*	*Values...*

- End of option list

0

- No operation

Options are usually 4 byte aligned with leading NOPs

1

- Maximum (*Receive*) Segment Size [SYN only]

2	4	MSS

- Window Scale Factor [SYN only]

NOP	3	3	shift

- Timestamp

NOP	NOP	8	10	Timestamp Value	Timestamp Echo Reply

- Selective ACK Permitted [SYN packet only]

NOP	NOP	4	2

- Selective ACK block

NOP	NOP	5	L	Left Edge	Right Edge

...

L = 2 + N * 8, N is number of left-right pairs