

Wide Area Networking

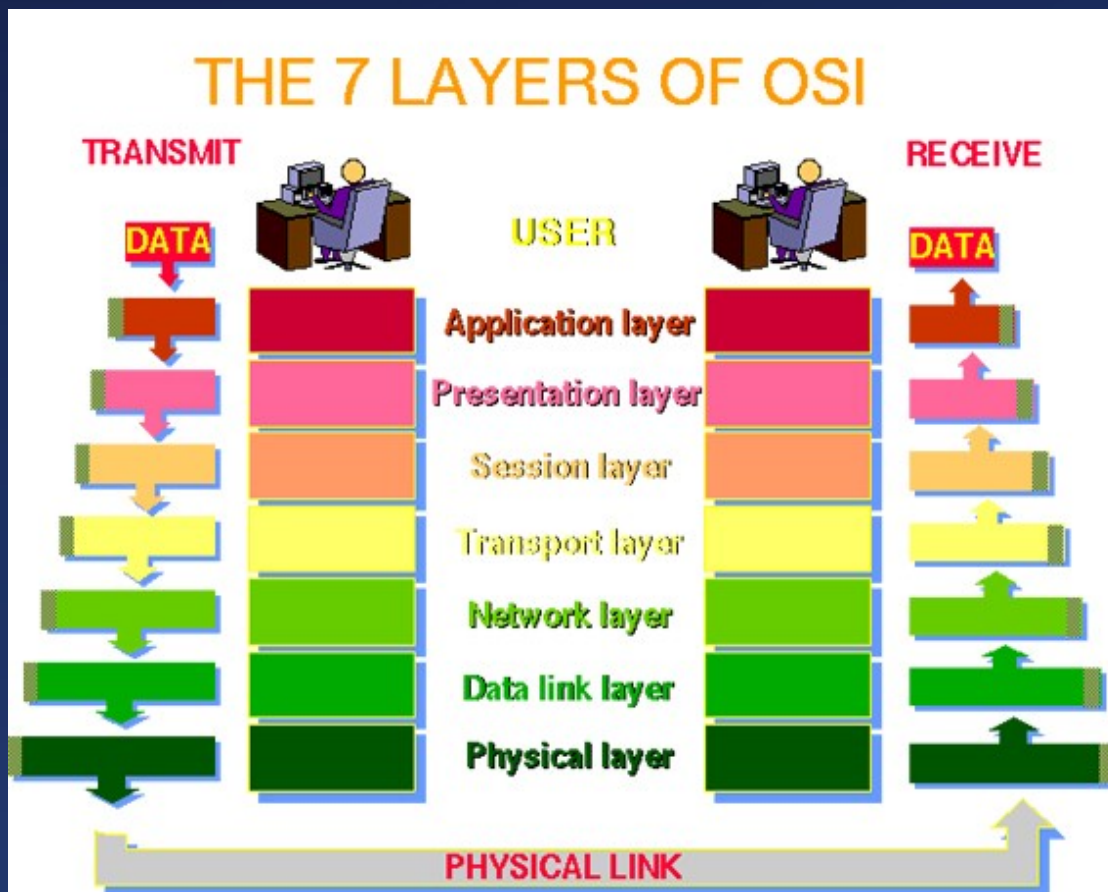
A short introduction to
High-Speed Wide-Area-Networking

Wide Area Networking

- Quick introduction to the OSI model
- Quick introduction to TCP/IP
- Problems of TCP on a high speed WAN link
- Measurements

The OSI Model

- OSI – Open Systems Interface
- Defines a networking framework in seven layers



- Each layer provides interface to the layer above
- Each layer adds a header (some also a trailer)
- Lowest layer transmits the message

The OSI Model

- Physical Layer
 - Concerned with transmission of *bits*
 - Standardized protocol for electrical, mechanical and signaling interfaces
- DataLink Layer
 - Groups bits into *frames* and ensures correct delivery
 - Handles errors in physical layer
 - Adds bits (head/tail) + checksum (receiver verifies checksum)
 - Sublayers: LLC – Logical Link Control and MAC – Medium Access Control
- Network Layer (“Packet” layer)
 - Transmission of *packets* and choosing best path for the *packet* (routing)
 - IP – Internet Protocol
 - Connectionless; IP *packet* can be send without a connection being established
 - Each *packet* gets routed *independently* to it's destination

OSI Model

- Transport Layer
 - Ensures reliable service (network layer does not deal with lost messages)
 - Breaks message into *packets*, assigns a sequence number and sends them
 - Builds reliable network connection on top of **IP** (or other protocols)
 - In case of **IP**, packets arriving out of order must be reordered
 - **TCP** – Transport Control Protocol (TCP/IP widely used protocols)
 - **UDP** – Universal Datagram Protocol (connectionless)
- Session Layer
 - Establishes, maintains and terminates sessions across networks
 - Examples: interactive login and file transfer connections
- Presentation Layer
 - Translates application network format + De-/Encryption, Compression...
- Application Layer
 - DNS, FTP, SMTP, NFS, ...

A bit more about TCP/IP + Ethernet I

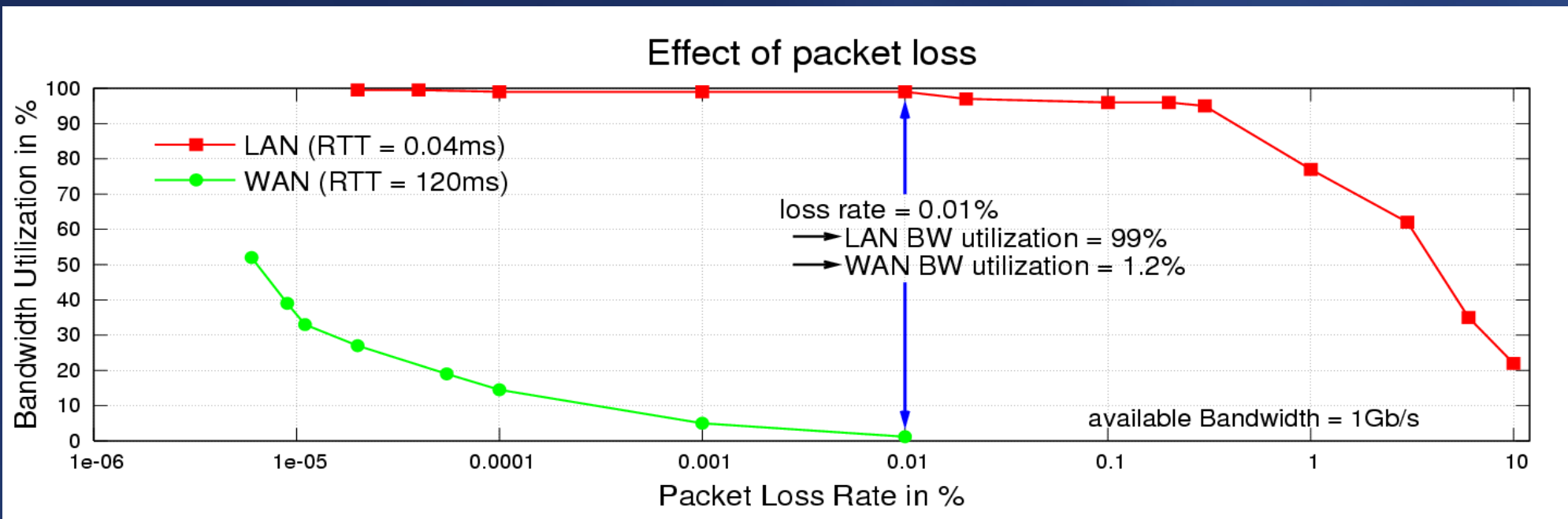
- Designed for slow and unreliable networks (1970's)
- The TCP Window
 - Amount of outstanding data a sender can send before it gets an ACK back from the receiver.
 - Why do we need it? Congestion control
 - Network has a bottleneck somewhere sender too fast
packet loss TCP Window throttles the transmission speed
down no packet loss
 - Min. window for max. bandwidth = bandwidth * delay
(10Gb and 100ms delay: min. TCP window = 128 Mbyte)
 - Standard TCP Window (*nix): 32kByte - 256kByte

A bit more about TCP/IP + Ethernet II

- The MTU – Maximum Transfer Unit
 - Chunk size the data gets chopped into (*frame* size)
(+ Headers and Trailers)
 - The bigger the MTU, the smaller the overhead
(... the more efficient the transfer...)
 - Ethernet standard: 1500 byte
(remnant from unreliable networks...)
 - High end equipment supports up to 9216 byte
(Intel 10Gb NICs support 16114 byte MTU !!)
 - Very difficult to build switches/router for bigger MTU
(large fast buffers, checksums, etc.)

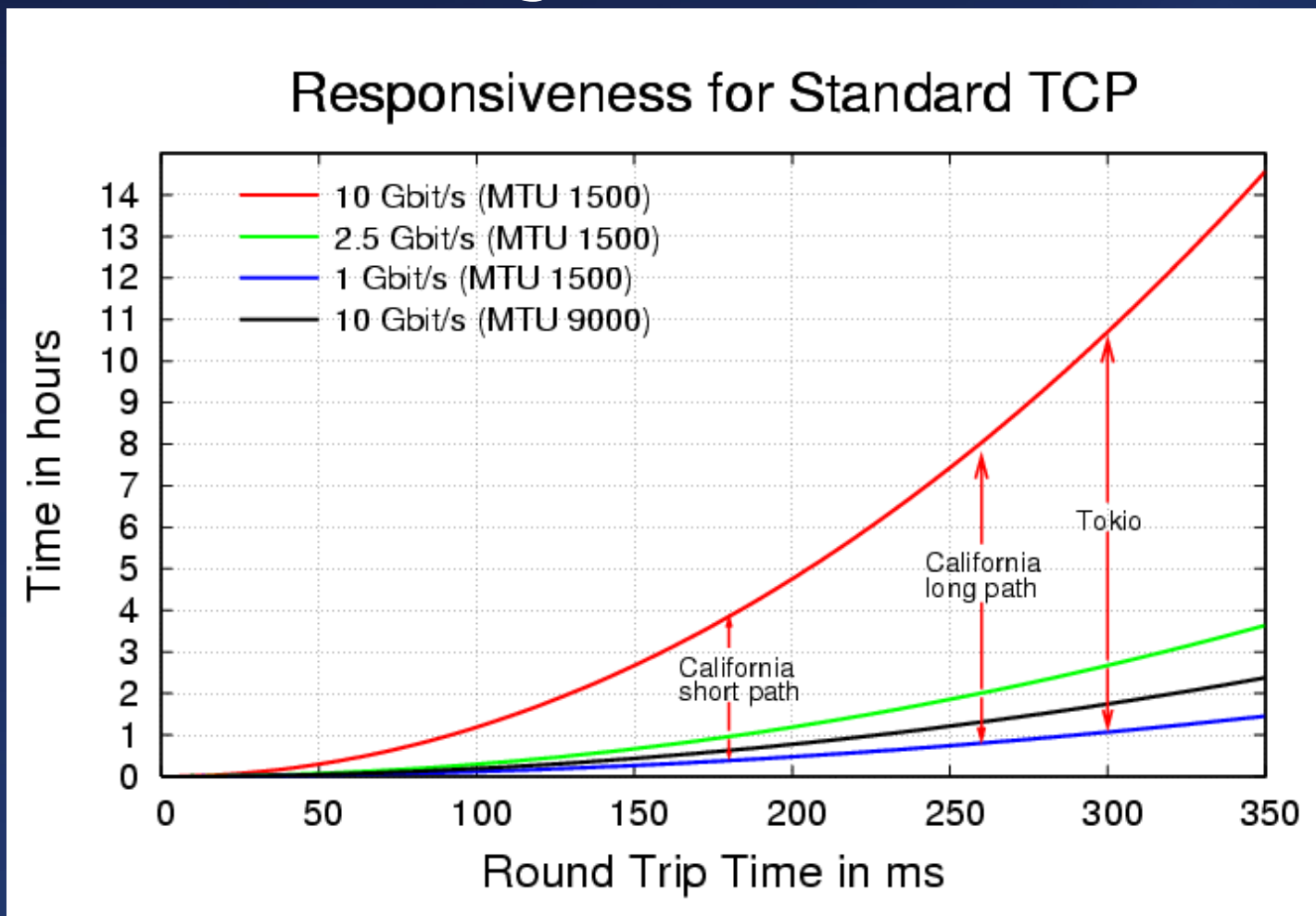
A bit more about TCP/IP + Ethernet III

- Recovery mechanism after a packet loss is too slow.
 - TCP window is cut in half after a packet loss
 - Current recovery algorithm increases window size only linearly with time



A bit more about TCP/IP + Ethernet IV

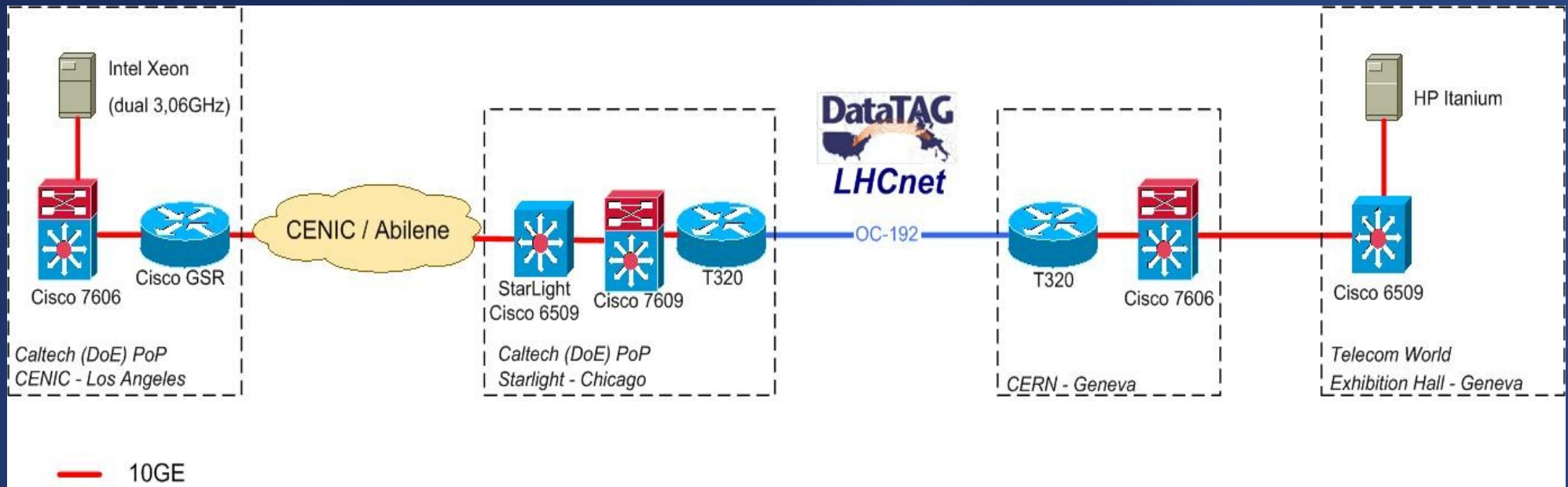
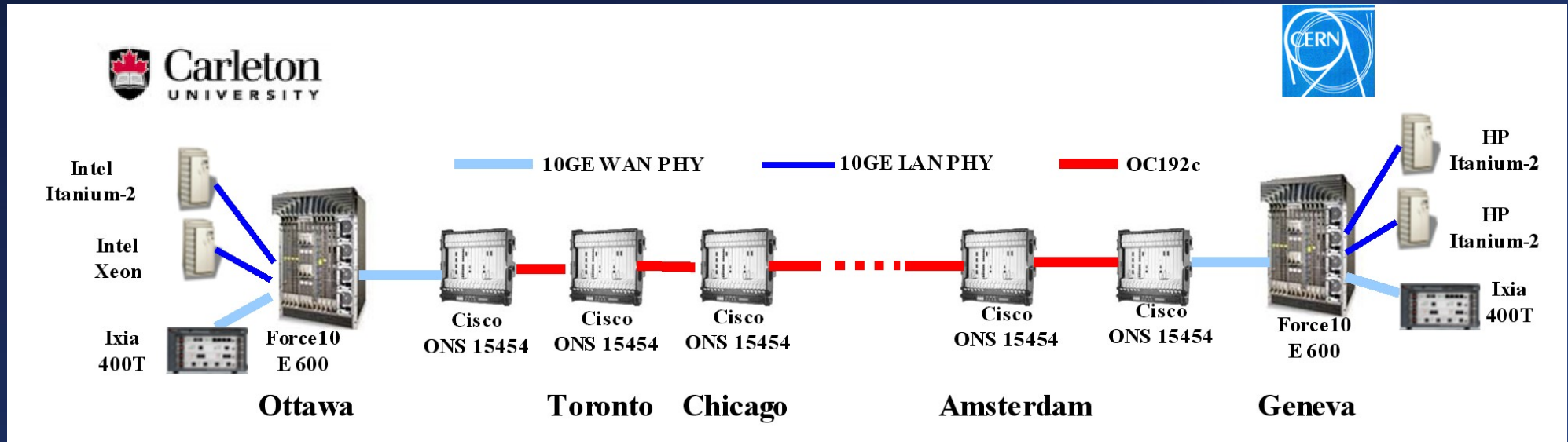
- Responsiveness ρ measures how quickly the connection goes back to full bandwidth after a packet



$$\rho = \frac{C * RTT^2}{2 * MSS}$$

C – Capacity of the link
RTT – Round Trip Time
MSS – Message size
(MTU - 40Bytes)

How does a WAN link look like?

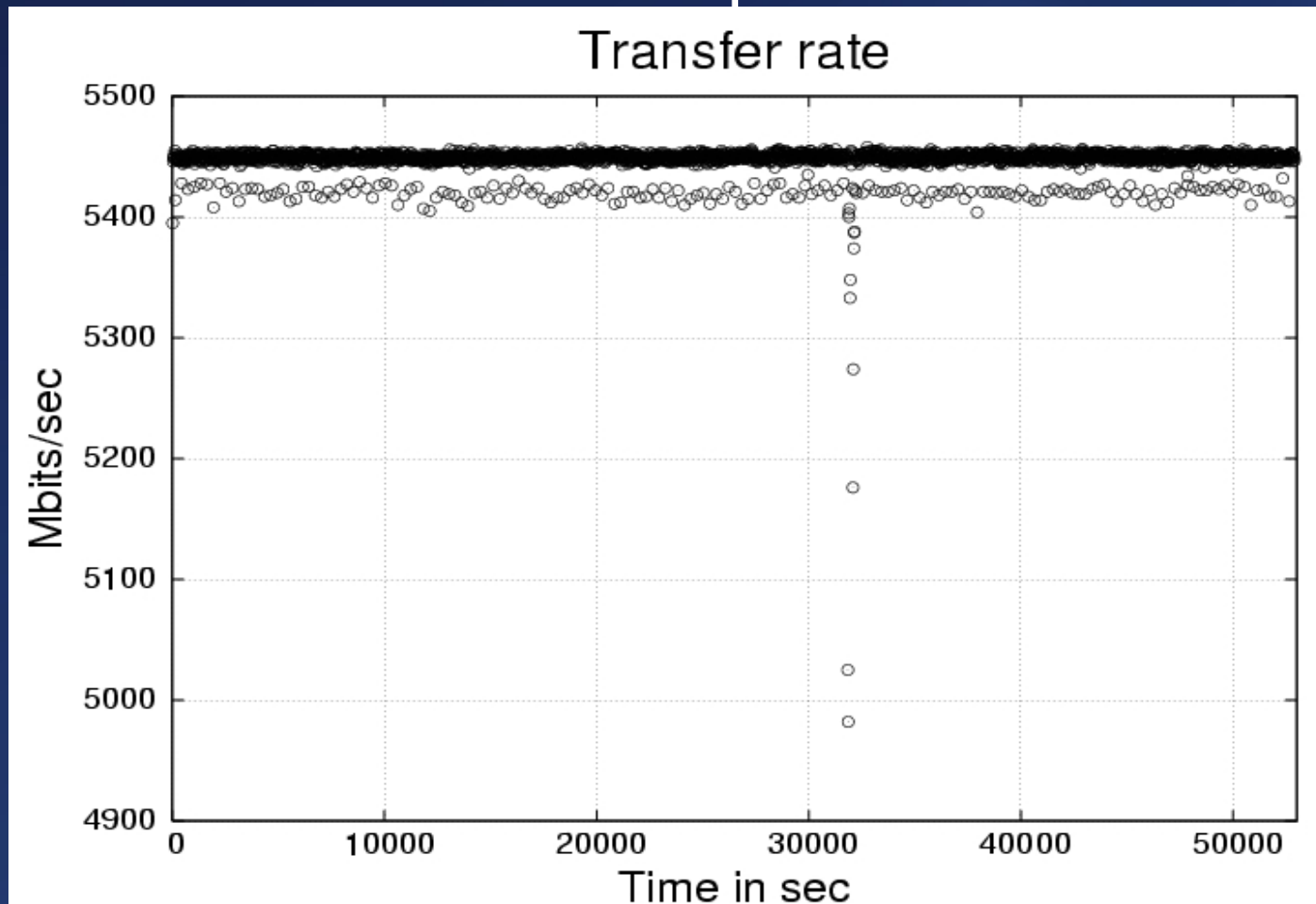


Measurements

- With ATLAS TDAQ group
 - Ethernet over WAN equipment
 - Amsterdam and Ottawa
- With DataTag
 - “Standard” WAN
 - Chicago and California
- All measurements with improved recovery algorithms!!
- All measurements were memory-to-memory transfers!

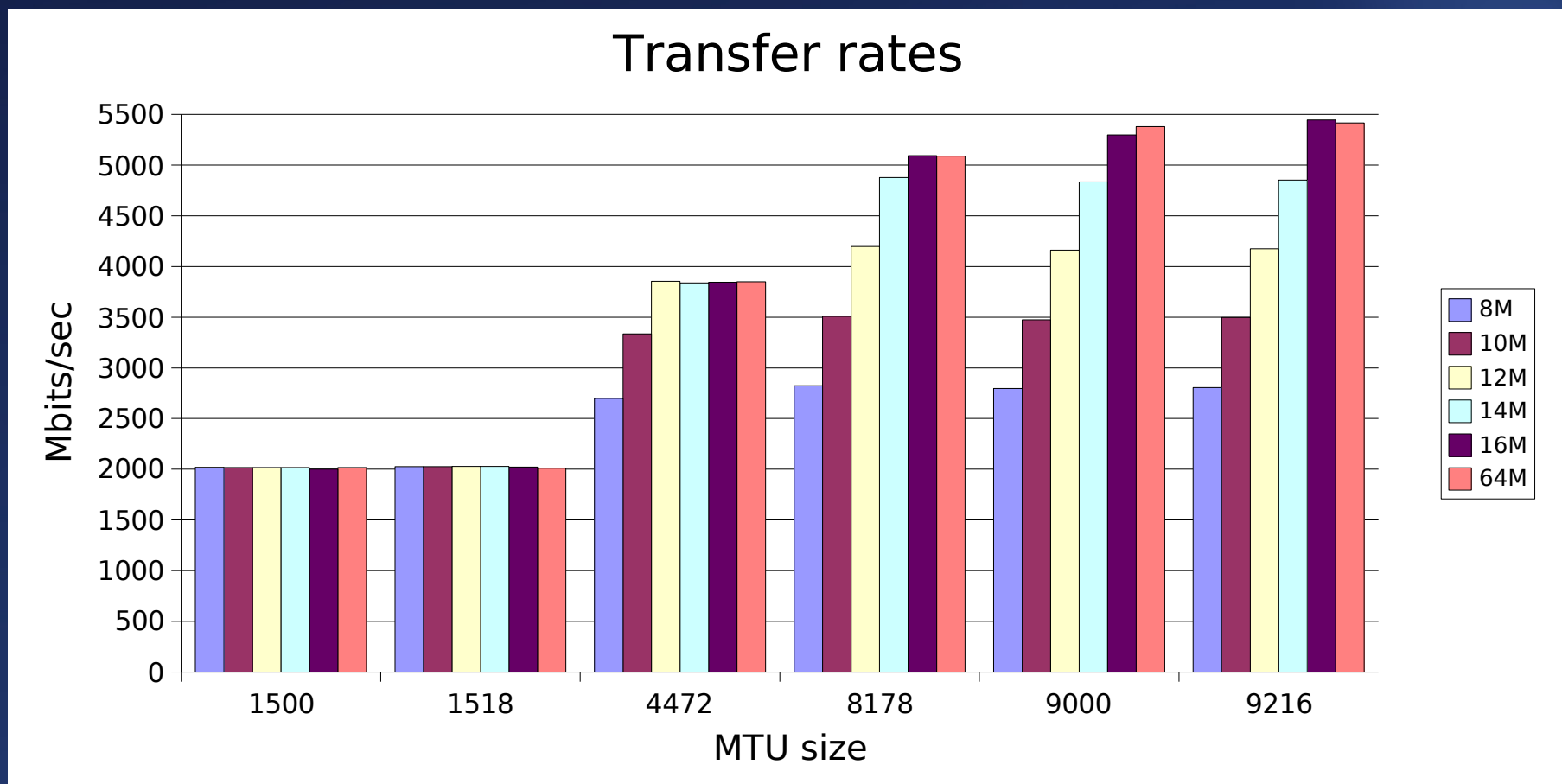
Measurements – ATLAS

- ~15 hours with only 2-3 packet losses
- Factor >100 better than Spec



Measurements – ATLAS

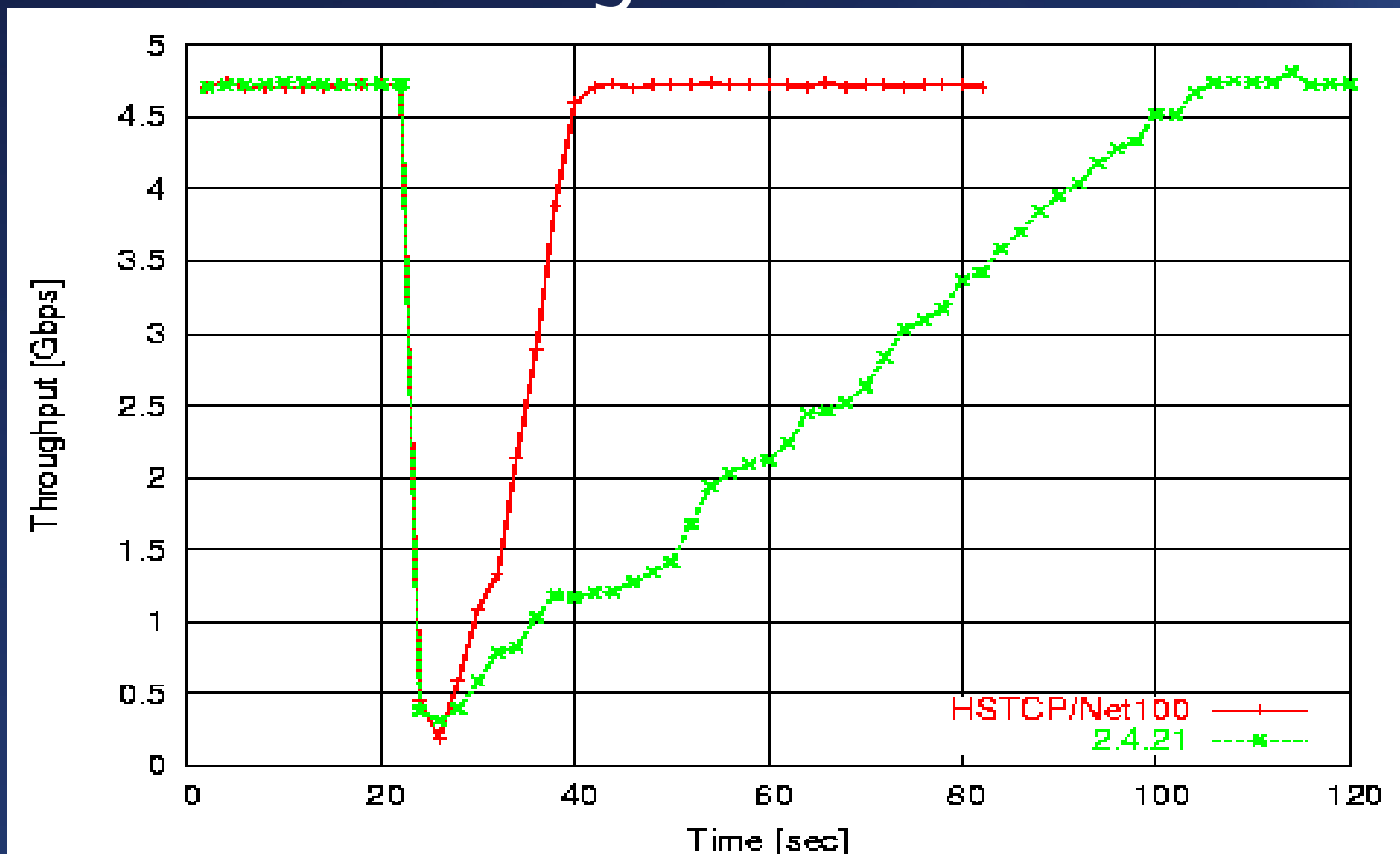
The effect of TCP window size and MTU



Bandwidth Delay Product (min. window size): $10\text{Gb/s} * 17\text{ms} = 20\text{MB}$

Measurements – ATLAS

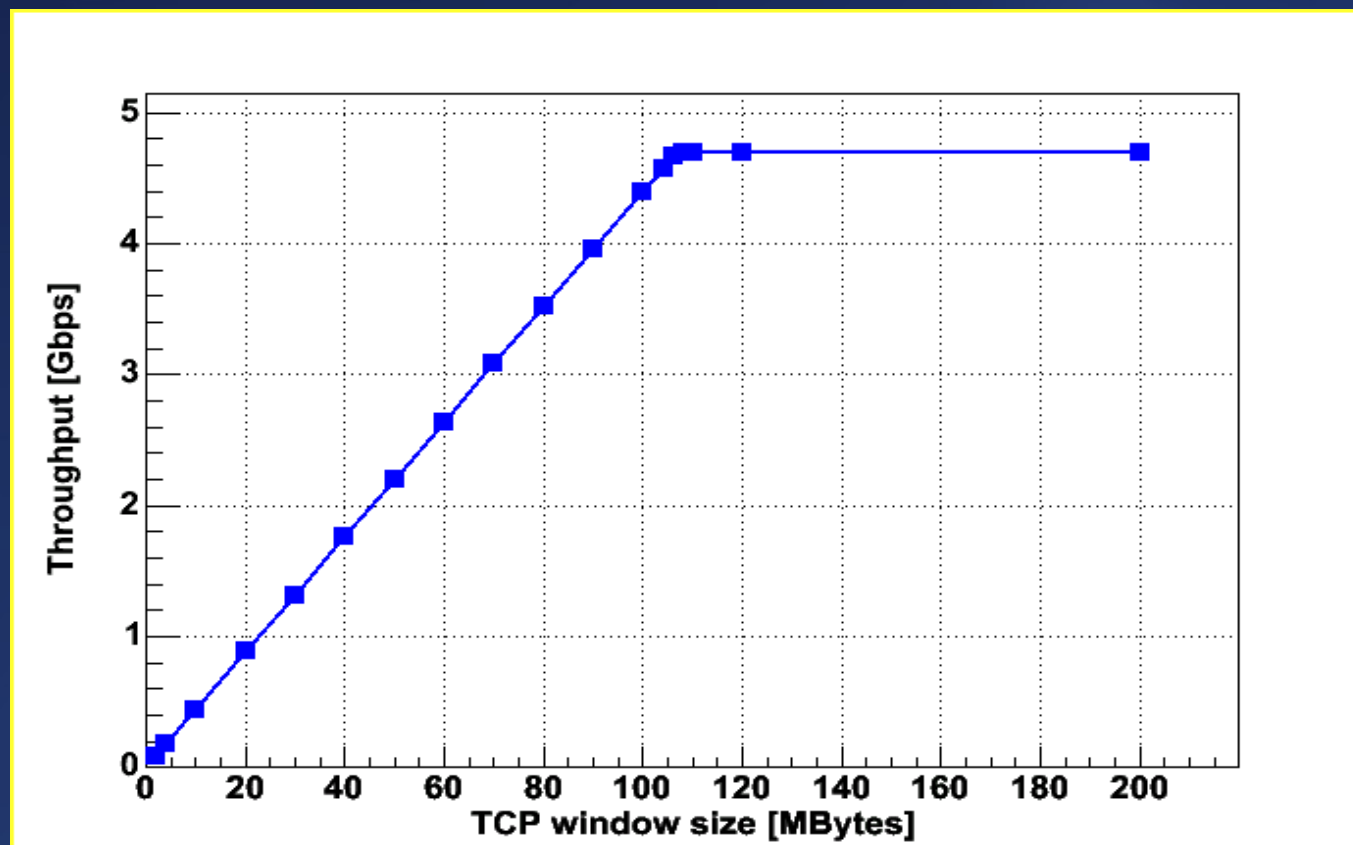
TCP Recovery time for different algorithms



Measurements - ATLAS

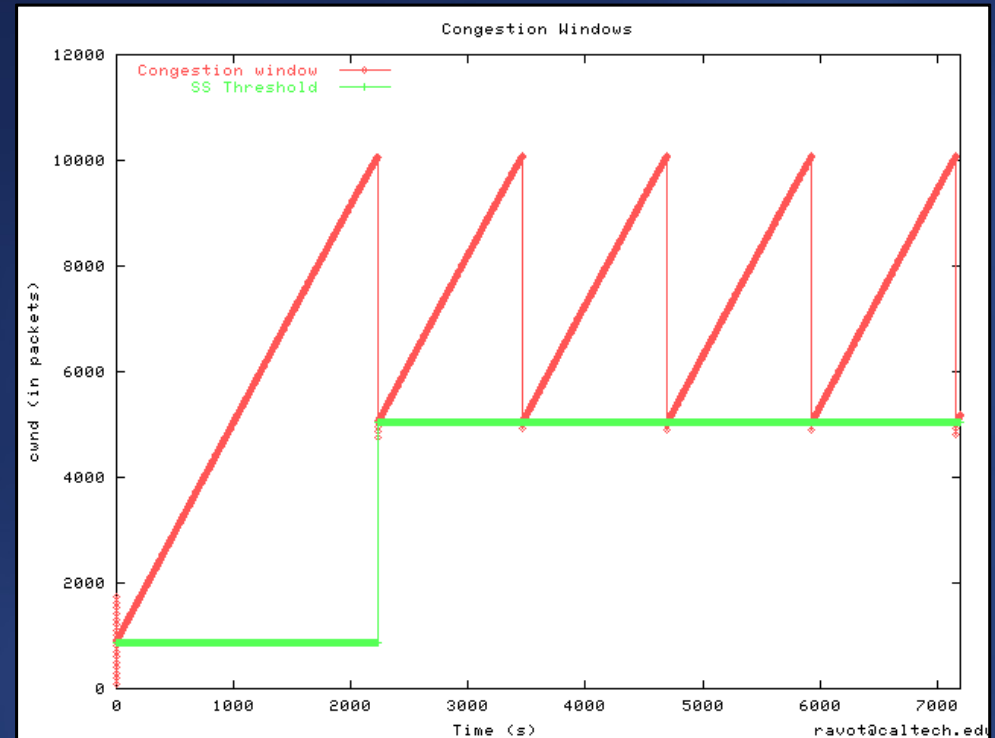
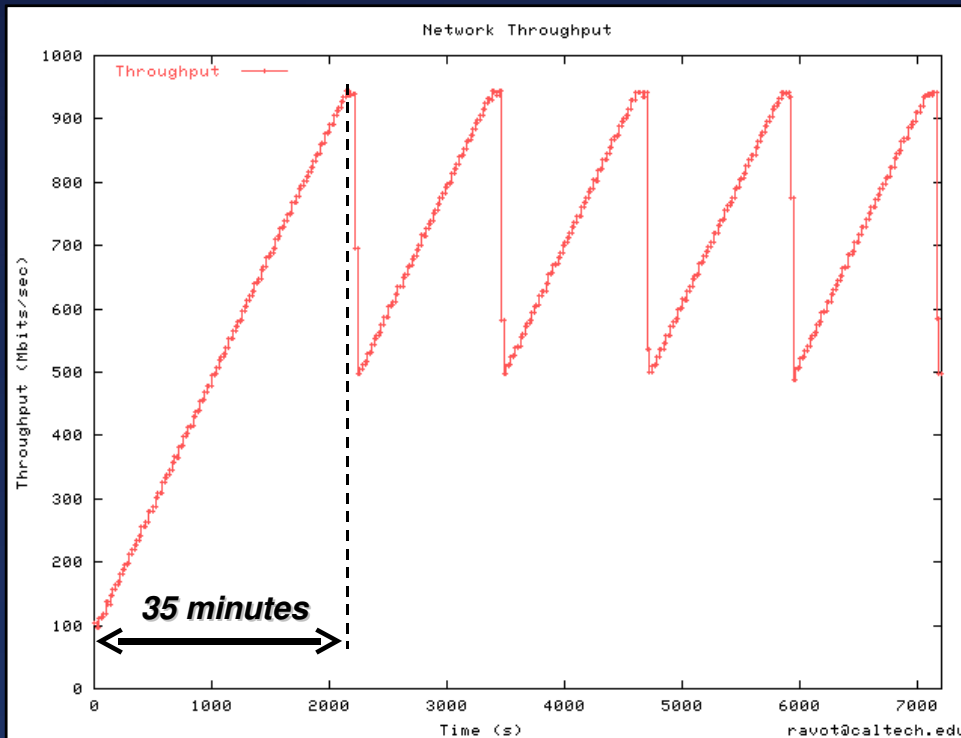
- First transatlantic Ethernet connection!!

Figure 7 - TCP single stream throughput versus the client TCP window size



Measurements – DataTag

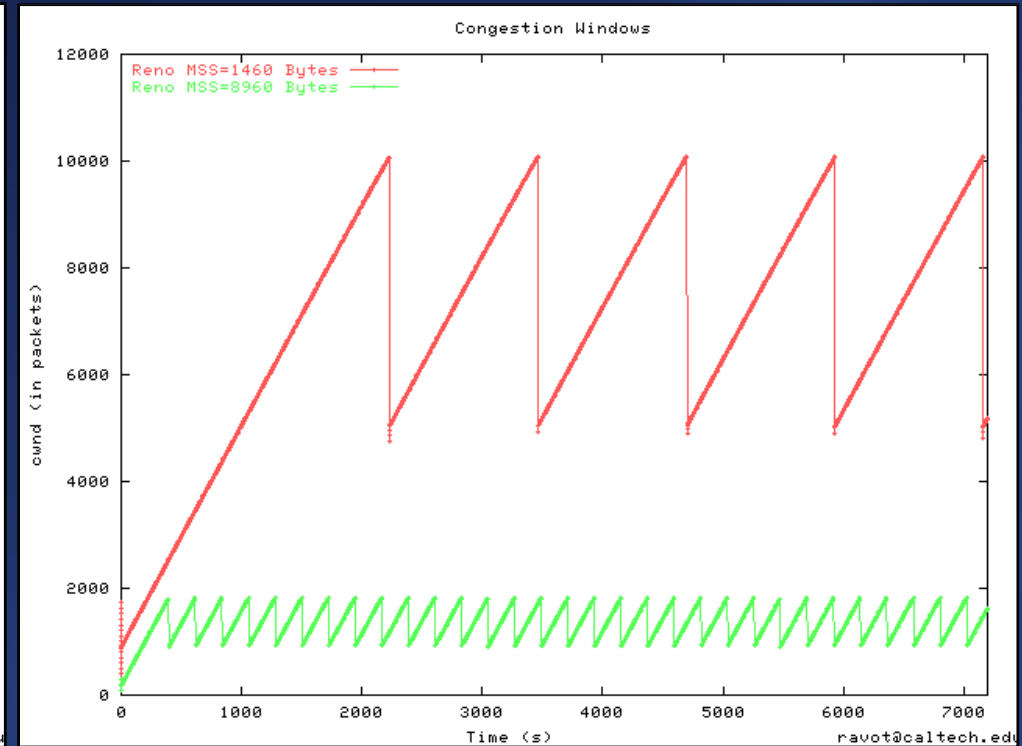
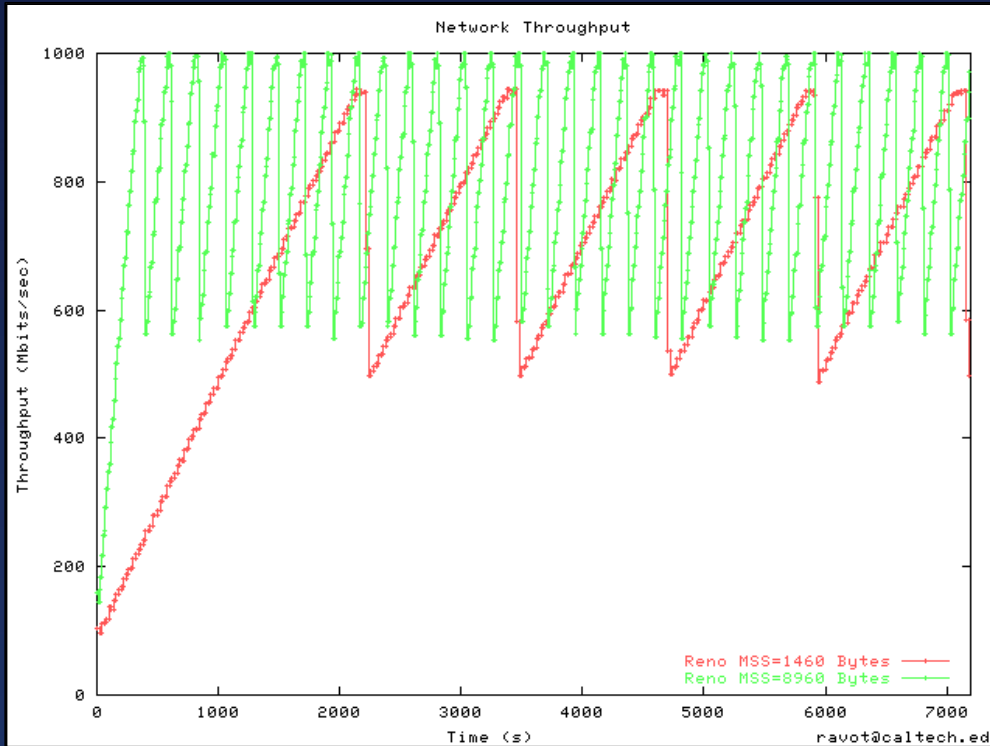
Measurements with Standard TCP recovery algorithm



- Geneva Chicago: $C = 1\text{Gbit/s}$, $\text{MTU} = 1500\text{byte}$, $\text{RTT} = 120\text{ms}$
- Packet loss occurs, when throughput approaches pipe size
- On average 75% bandwidth utilisation


Measurements – DataTag

The effect of different MTU sizes



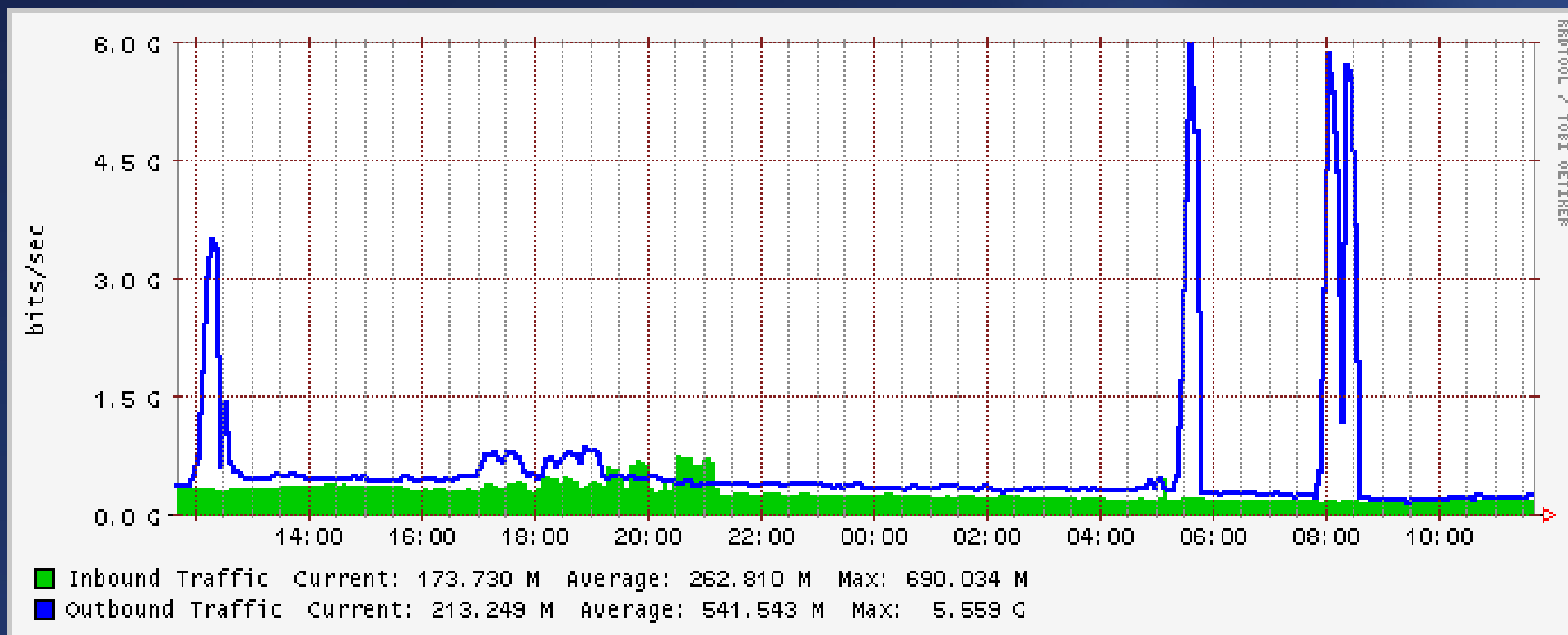
- ~75% link utilisation in both cases
- Large MTU sizes accelerate the growth of the window size
- Time to recover after a loss decreases (significantly)

Measurements – DataTag

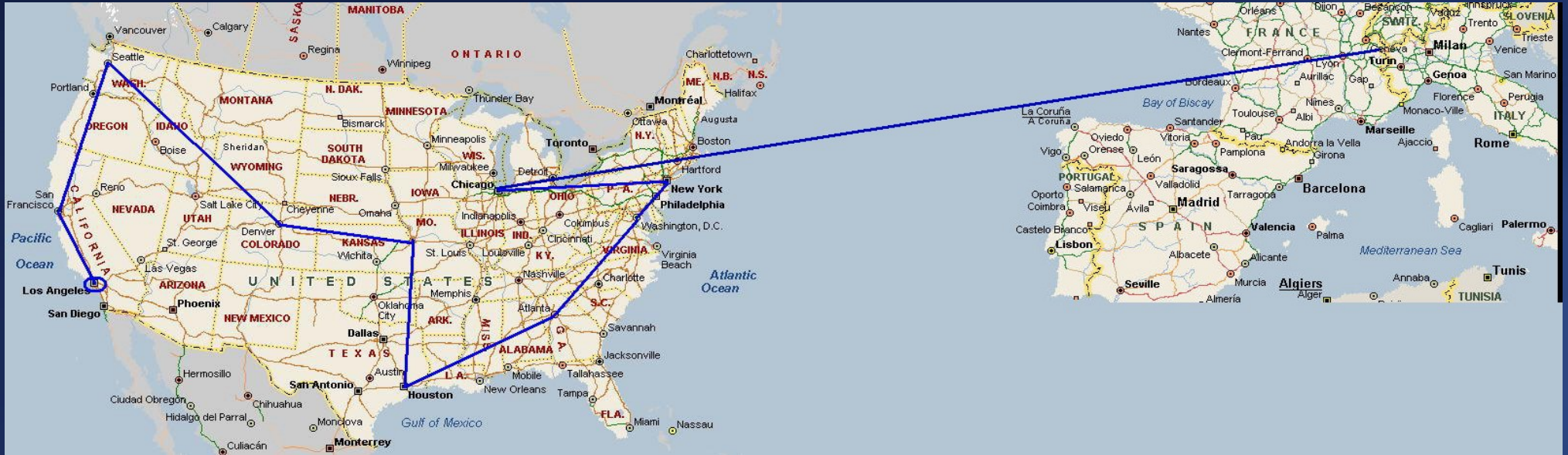
- ~5.6 Gb/s to Chicago (Telecom 2003)
 - Dedicated link
 - Sustained for hours
- ~6.6 Gb/s to California
 - Shared link between Chicago and California
 - Sustained only for ~10min
 - New Land Speed Record
- ~7.4 Gb/s to California
 - Sustained only for 2-4min 

Measurements – DataTag

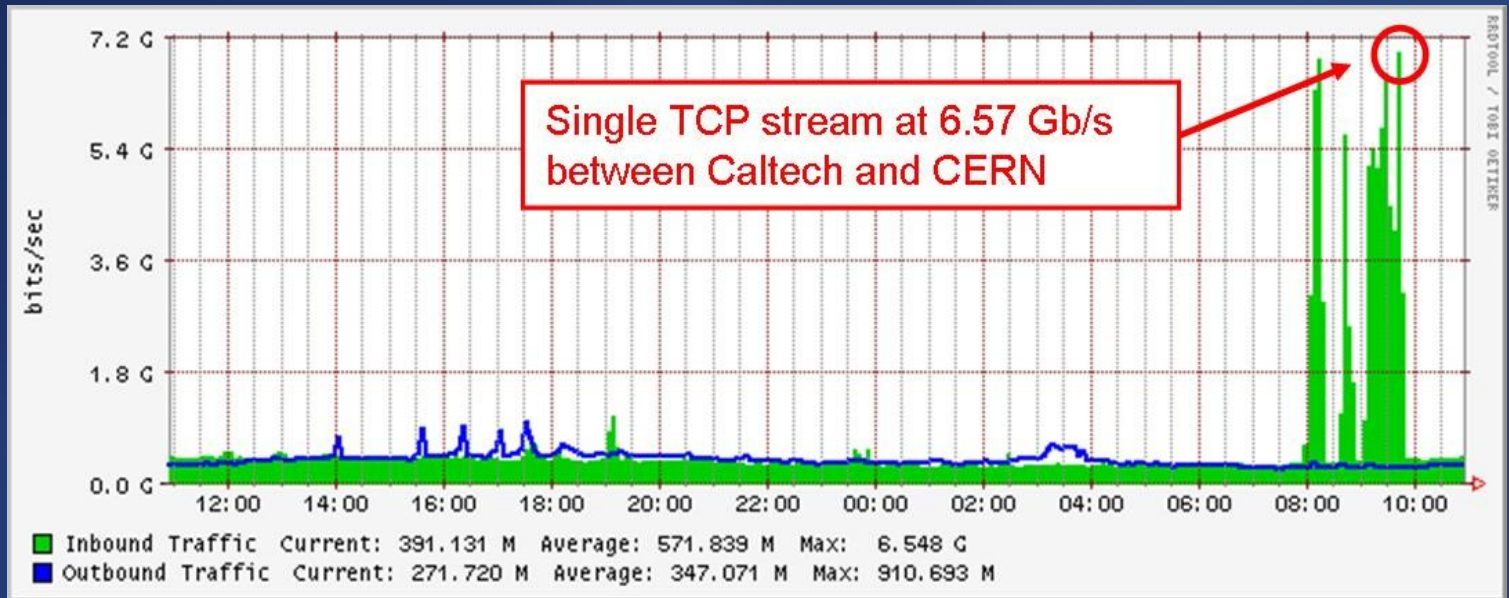
- Land Speed Record at Telecom 2003: 5.65 Gb/s
- Geneva → Chicago



Latest Land Speed Record (submitted ;-)



- 15766 km
 - 6.57 Gb/s
- =103583Tbm/s



Outlook

- Start disk-to-disk transfers
 - Sustained (low-number-)multistream connections
 - ~400-500MB/s for months
 - Aggregation of 1Gb links into 10Gb WAN
- Direct 10Gb connection for disk-to-disk transfers
 - First step: ~350MB/s disk-to-memory with RFIO (home grown protocol) via 10Gb LAN
- No tests up to now many Unknowns