State of the art in the use of long distance network

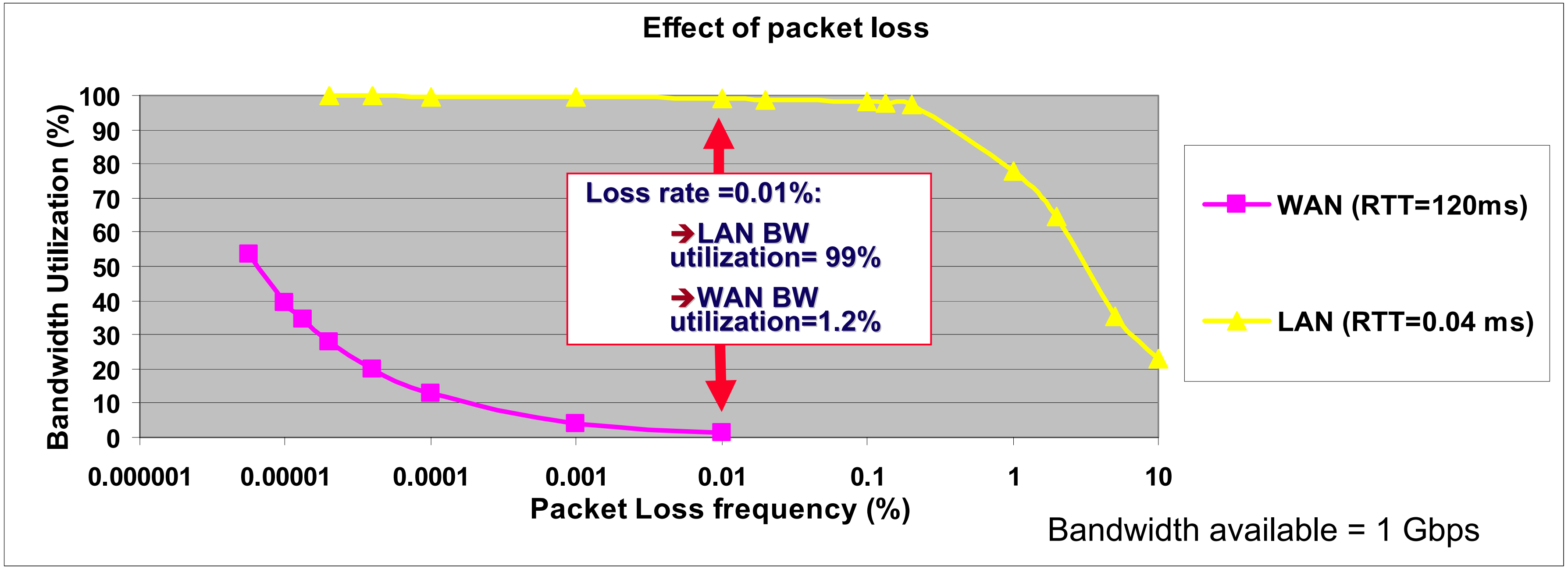
International ICFA Workshop on HEP Networking
Daegu, Korea
May 23, 2005

J. Bunn, D. Nae, H. Newman, S. Ravot, X. Su, Y. Xia
California Institute of Technology

AGENDA

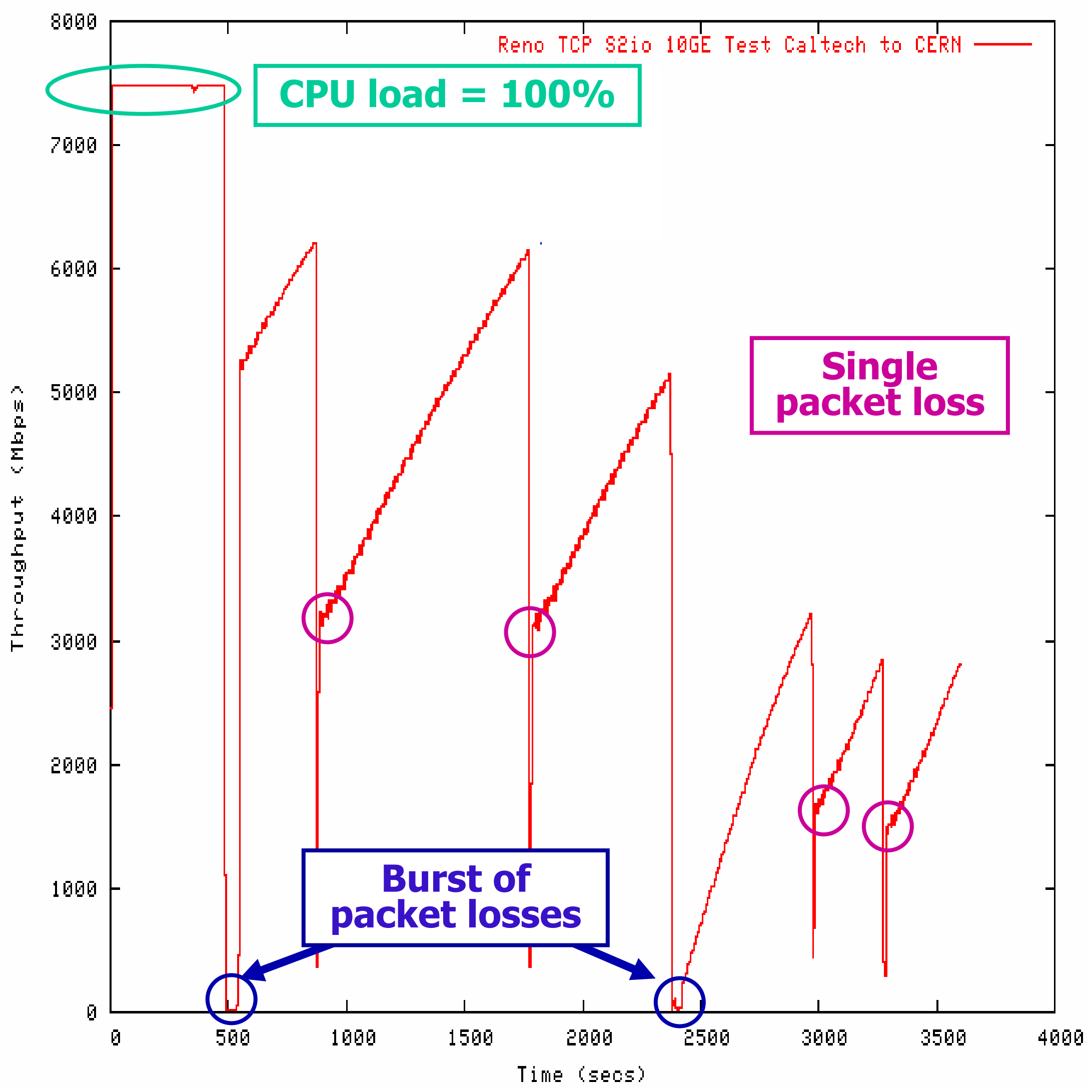
- TCP performance over high speed/latency networks
- Recent results
- End-Systems performance
- Next generation network
- Advanced R&D projects

Single TCP stream performance under periodic losses



- TCP throughput much more sensitive to packet loss in WANs than LANs
  - TCP's congestion control algorithm (AIMD) is not well-suited to gigabit networks
  - The effect of packets loss can be disastrous
- TCP is inefficient in high bandwidth*delay networks
- The future performance-outlook for computational grids looks bad if we continue to rely solely on the widely-deployed TCP RENO

Single TCP stream between Caltech and CERN



- Available (PCI-X) Bandwidth=8.5 Gbps
- RTT=250ms (16'000 km)
- 9000 Byte MTU
- 15 min to increase throughput from 3 to 6 Gbps
- Sending station:
  - Tyan S2882 motherboard, 2x Opteron 2.4 GHz , 2 GB DDR.
- Receiving station:
  - CERN OpenLab:HP rx4640, 4x 1.5GHz Itanium-2, zx1 chipset, 8GB memory
- Network adapter:
  - S2IO 10 GbE

Responsiveness

- Time to recover from a single packet loss:

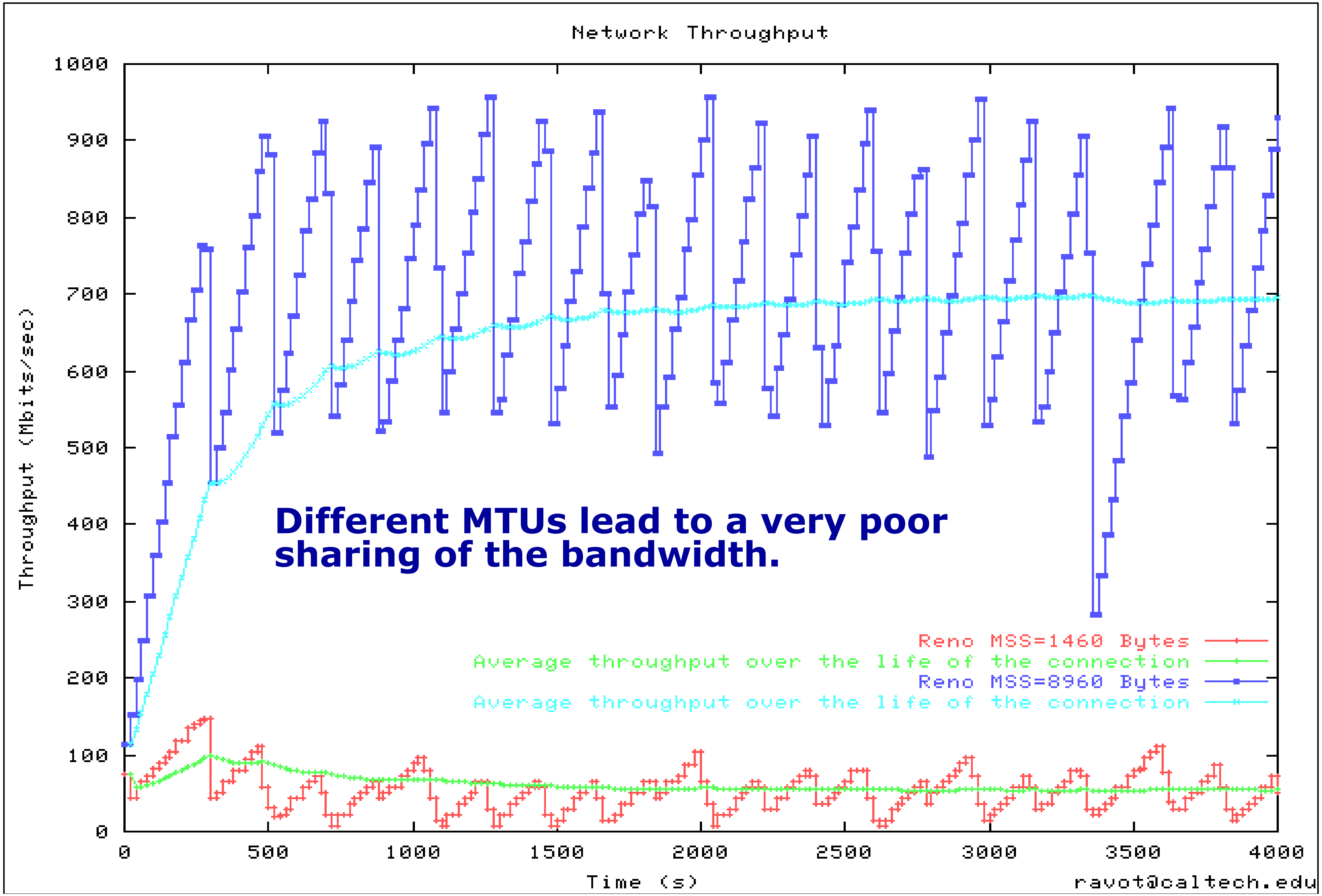
$$\rho = \frac{C \cdot RTT^2}{2 \cdot MSS}$$

C : Capacity of the link

Path	Bandwidth	RTT (ms)	MTU (Byte)	Time to recover
LAN	10 Gb/s	1	1500	430 ms
Geneva–Chicago	10 Gb/s	120	1500	1 hr 32 min
Geneva-Los Angeles	1 Gb/s	180	1500	23 min
Geneva-Los Angeles	10 Gb/s	180	1500	3 hr 51 min
Geneva-Los Angeles	10 Gb/s	180	9000	38 min
Geneva-Los Angeles	10 Gb/s	180	64k (TSO)	5 min
Geneva-Tokyo	1 Gb/s	300	1500	1 hr 04 min

- Large MTU accelerates the growth of the window
- Time to recover from a packet loss decreases with large MTU
- Larger MTU reduces overhead per frames (saves CPU cycles, reduces the number of packets)

Fairness: TCP Reno MTU & RTT bias

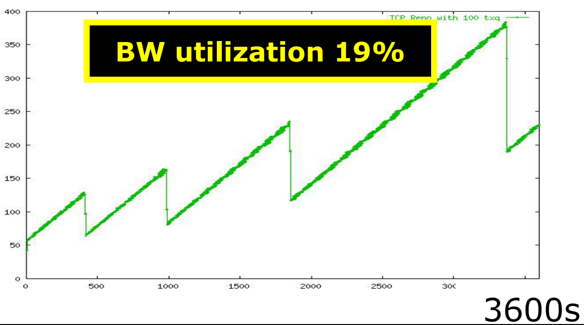


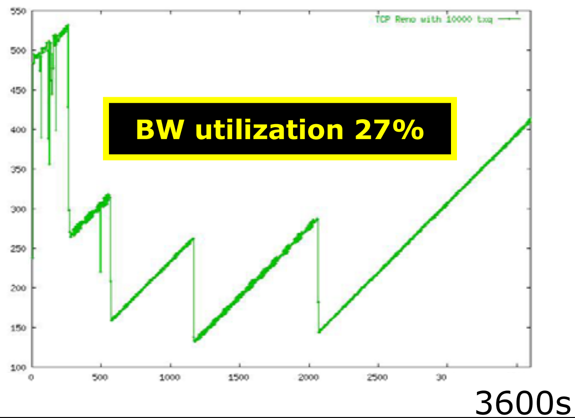
Different MTUs lead to a very poor sharing of the bandwidth.

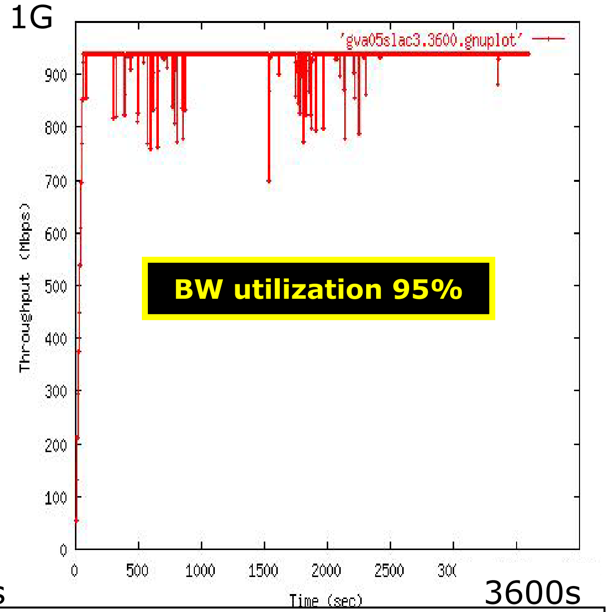
TCP Variants

- HSTCP, Scalable, Westwood+, Bic TCP, HTCP
- <u>FAST TCP</u>
  - Based on TCP Vegas
  - Uses end-to-end delay and loss to dynamically adjust the congestion window
  - Defines an explicit equilibrium
  - 7.3 Gbps between CERN and Caltech for hours
  - FAST vs RENO:

capacity = 1Gbps; 180 ms round trip latency; 1 flow





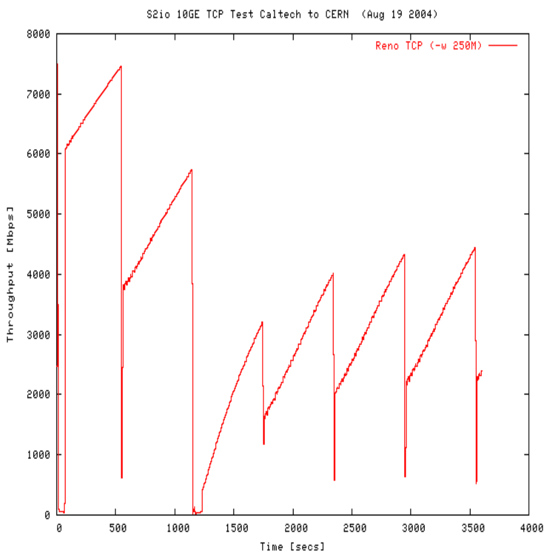


Linux TCP	Linux TCP (Optimized)	FAST

TCP Variants Performance

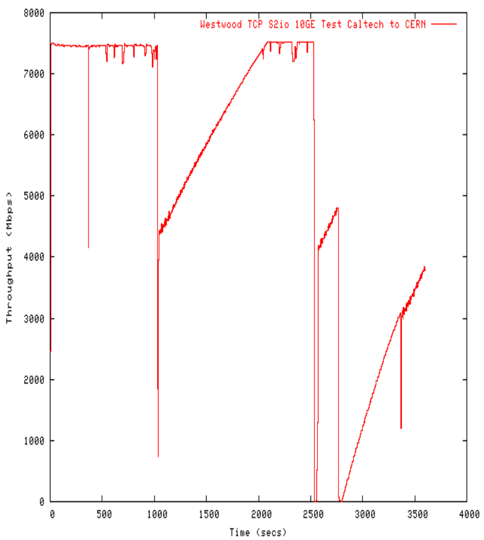
- Tests between CERN and Caltech
- Capacity = OC-192 9.5Gbps; 264 ms round trip latency; 1 flow
- Sending station: Tyan S2882 motherboard, 2x Opteron 2.4 GHz , 2 GB DDR.
- Receiving station (CERN OpenLab): HP rx4640, 4x 1.5GHz Itanium-2, zx1 chipset, 8GB memory
- Network adapter: Neterion 10 GE NIC

3.0 Gbps



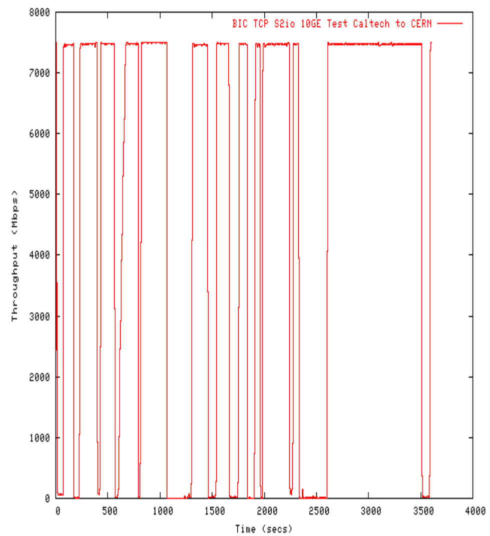
Linux TCP

4.1 Gbps



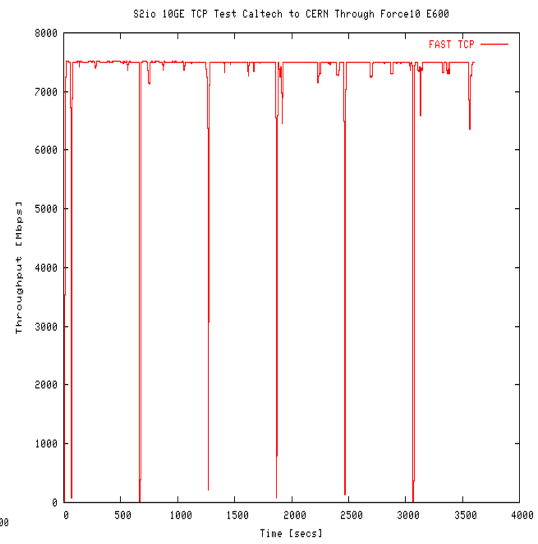
Linux Westwood+

5.0 Gbps



Linux BIC TCP

7.3 Gbps

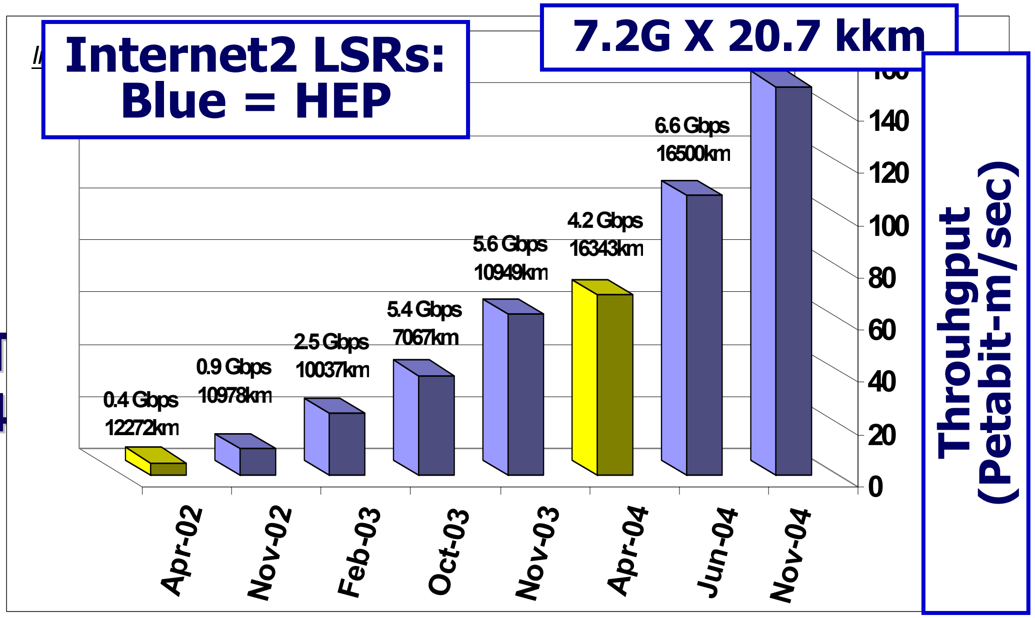


FAST TCP

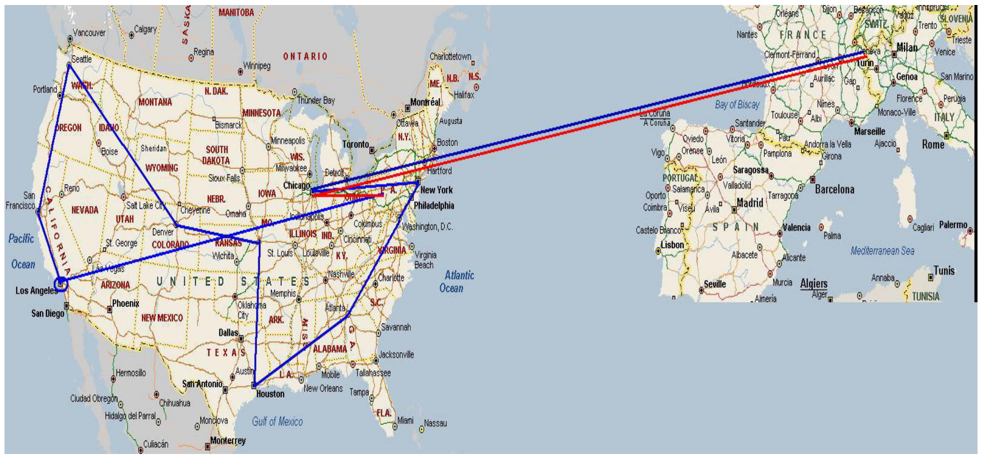
Internet 2 Land Speed Record (LSR)

Redefining the Role and Limits of TCP

- **Product of transfer speed and distance using standard Internet (TCP/IP) protocols.**
- **Single Stream 7.5 Gbps X 16 kkm with Linux: July 2004**
- **IPv4 Multi-stream record with FAST TCP: 6.86 Gbps X 27kkm: Nov 2004**
- **IPv6 record: 5.11 Gbps between Geneva and Starlight: Jan. 2005**
- ***Concentrate now on reliable Terabyte-scale file transfers***
  - ***Disk-to-disk Marks: 536 Mbytes/sec (Windows); 500 Mbytes/sec (Linux)***
  - ***Note System Issues: PCI-X Bus, Network Interface, Disk I/O Controllers, CPU, Drivers***



Nov. 2004 Record Network

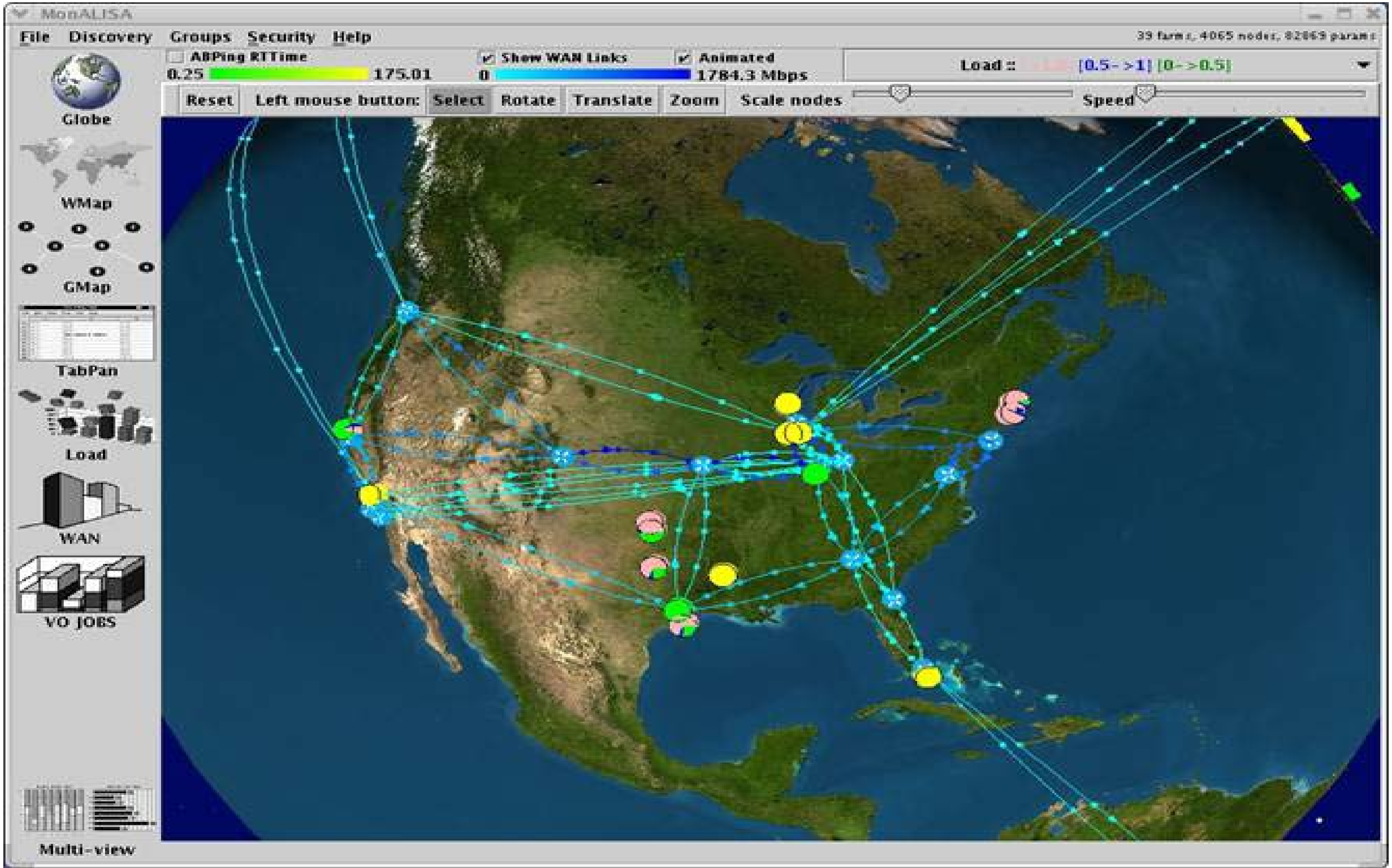


http://www.guinnessworldrecords.com/

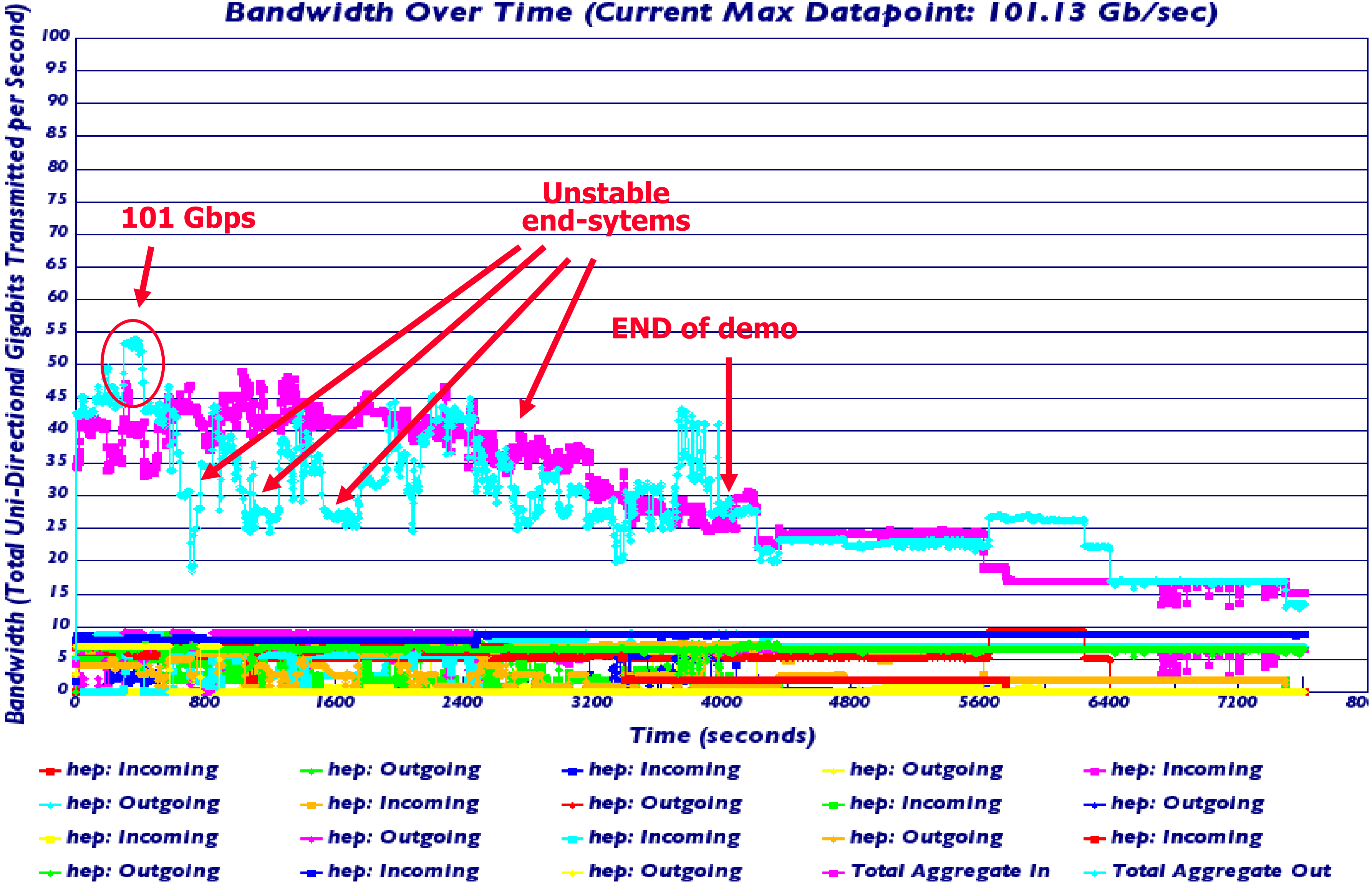
SC2004: High Speed TeraByte Transfers for Physics

- Demonstrating that many 10 Gbps wavelengths can be used efficiently over continental and transoceanic distances
- Preview of the globally distributed grid system that is now being developed in preparation for the next generation of high-energy physics experiments at CERN's Large Hadron Collider (LHC),
- Monitoring the WAN performance using the MonALISA agent-based system
- Major Partners : Caltech-FNAL-SLAC
- Major Sponsors:
  - Cisco, S2io, HP, Newysis
- Major networks:
  - NLR, Abilene, ESnet, LHCnet, Ampath, TeraGrid
- Bandwidth challenge award: 101 Gigabit Per Second Mark

SC2004 Network (I)

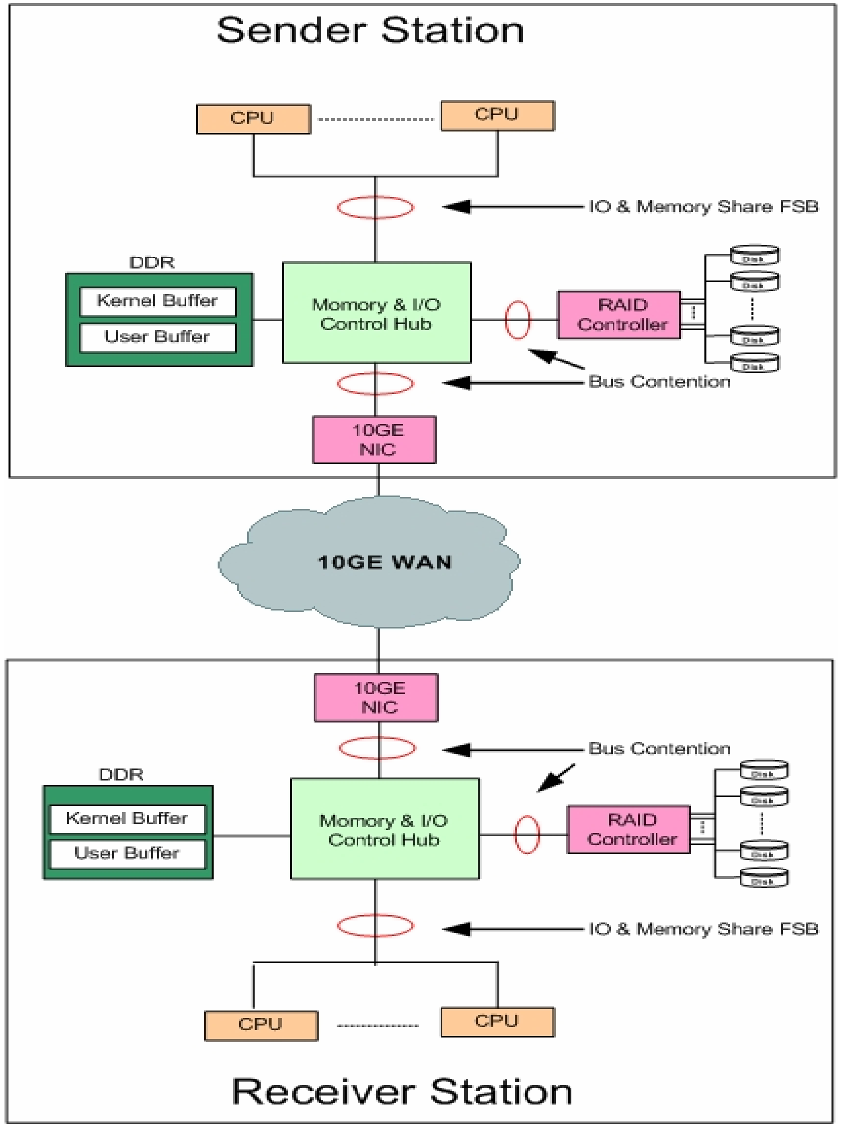


101 Gigabit Per Second Mark



Source: Bandwidth Challenge committee

High Throughput Disk to Disk Transfers: From 0.1 to 1GByte/sec



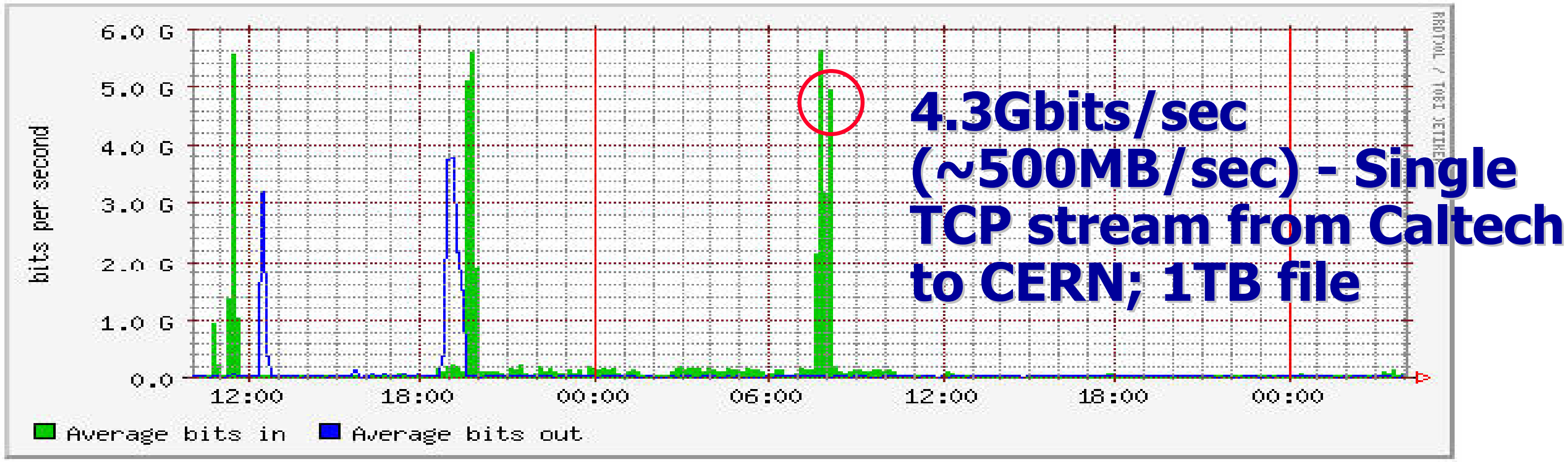
Server Hardware (Rather than Network) Bottlenecks:

- **Write/read and transmit tasks share the same limited resources: CPU, PCI-X bus, memory, IO chipset**
- **PCI-X bus bandwidth: 8.5 Gbps [133MHz x 64 bit]**

Performance in this range (from 100 MByte/sec up to 1 GByte/sec) is required to build a responsive Grid-based Processing and Analysis System for LHC

Transferring a TB from Caltech to CERN

- 3 Supermicro Marvell SATA disk controllers + 24 SATA 7200rpm SATA disks
  - Local Disk IO – 9.6 Gbits/sec (1.2 GBytes/sec read/write, with <20% CPU utilization)
- Neterion SR 10GE NIC
  - 10 GE NIC – 7.5 Gbits/sec (memory-to-memory, with 52% CPU utilization)
  - 2*10 GE NIC (802.3ad link aggregation) – 11.1 Gbits/sec (memory-to-memory)
- Memory to Memory WAN data flow, and local Memory to Disk read/write flow, are not matched when combining the two operations
- Quad Opteron AMD850 2.4GHz processors with 3 AMD-8131 chipsets: 4 64-bit/133MHz PCI-X slots.



New services aspirations

- **Circuit-based services**
  - Layer 1 & 2 switching, “the light path”
  - High bandwidth point-to-point circuits for big users (up to 10 Gbps)
  - Redundant paths
  - On-demand
  - Advance Reservation System;
    Authentication, Authorization and Accounting (AAA)
  - Control plane
    - GMPLS, UCLP, MonaLisa
- **New Initiatives/projects**
  - GEANT2, USNet, HOPI
  - GLIF (Global Lambda Integrated Facility)
  - OMNINET, Dragon, Cheetah
  - UltraLight and LambdaStation

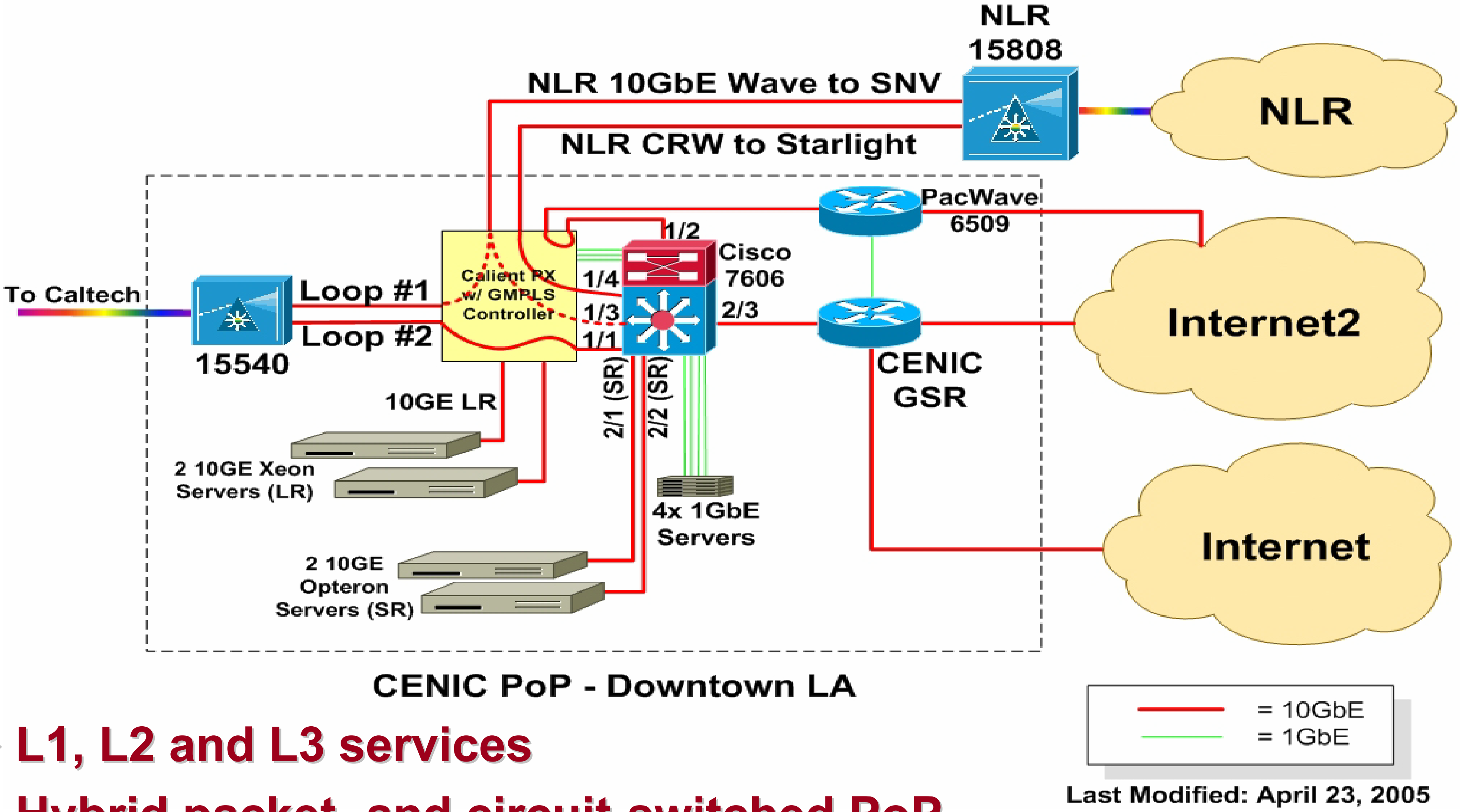
New Technology Candidates: Opportunities and Issues

- New standard for SONET infrastructures
  - Alternative to the expensive Packet-Over-Sonet (POS) technology currently used
  - May change significantly the way in which we use SONET infrastructures
- 10 GE WAN-PHY standard
  - Ethernet frames across OC-192 SONET networks
  - Ethernet as inexpensive linking technology between LANs and WANs
  - Supported by only a few vendors
- LCAS/VCAT standards
  - Point-to-point circuit-based services
  - Transport capacity adjustments according to the traffic pattern.
  - "Bandwidth on Demand" becomes possible for SONET network
  - "Intelligent" optical multiplexers (Available at the end of 2005)

UltraLight: Developing Advanced Network Services for Data Intensive HEP Applications

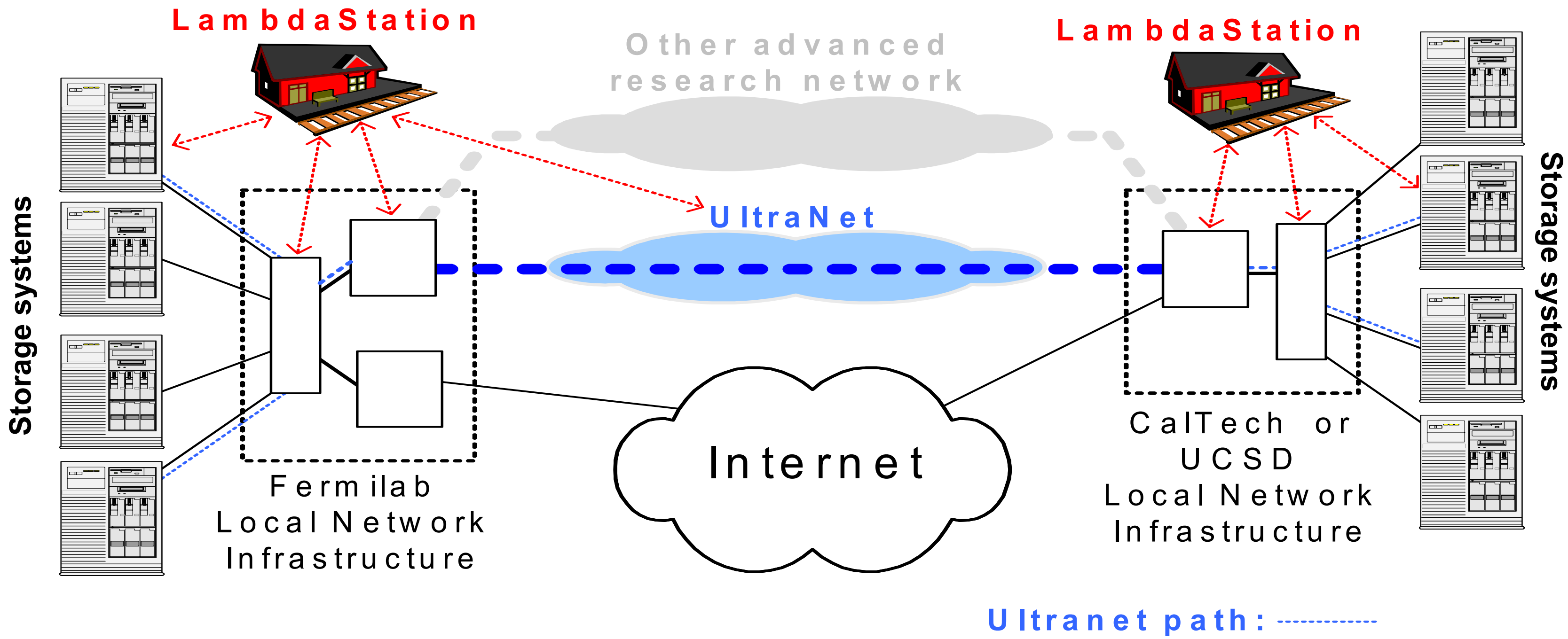
- ***UltraLight:*** a next-generation hybrid packet- and circuit-switched network infrastructure
  - **Packet switched:** cost effective solution; requires ultrascale protocols to share 10G λ efficiently and fairly
  - **Circuit-switched:** Scheduled or sudden "overflow" demands handled by provisioning additional wavelengths; Use path diversity, e.g. across the US, Atlantic, Canada,…
- Extend and augment existing grid computing infrastructures (currently focused on CPU/storage) to include the network as an integral component
  - Using MonALISA to monitor and manage global systems
- ***Partners:*** Caltech, UF, FIU, UMich, SLAC, FNAL, MIT/Haystack; CERN, NLR, CENIC, Internet2; Translight, UKLight, Netherlight; UvA, UCL, KEK, Taiwan
- Strong support from Cisco

Ultralight 10GbE OXC @ Caltech



- L1, L2 and L3 services
- Hybrid packet- and circuit-switched PoP
- Control plane is L3

LambdaStation



- A Joint Fermilab and Caltech project
- Enabling HEP applications to send high throughput traffic between mass storage systems across advanced network paths
- Dynamic Path Provisioning across Ultranet, NLR; Plus an Abilene "standard path"
- DOE funded

Summary

- For many years the Wide Area Network has been the bottleneck; this is no longer the case in many countries thus making deployment of a data intensive Grid infrastructure possible!
- Some transport protocol issues still need to be resolved; however there are many encouraging signs that practical solutions may now be in sight.
- 1GByte/sec disk to disk challenge.
  - Today: 1 TB at 536 MB/sec from CERN to Caltech
- Next generation network and Grid system: UltraLight and LambdaStation
  - The integrated, managed network
  - Extend and augment existing grid computing infrastructures (currently focused on CPU/storage) to include the network as an integral component.

Thanks

&

Questions