



# **How we quadrupled the network speed world record: *bandwidth Challenge at SC 2004***

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



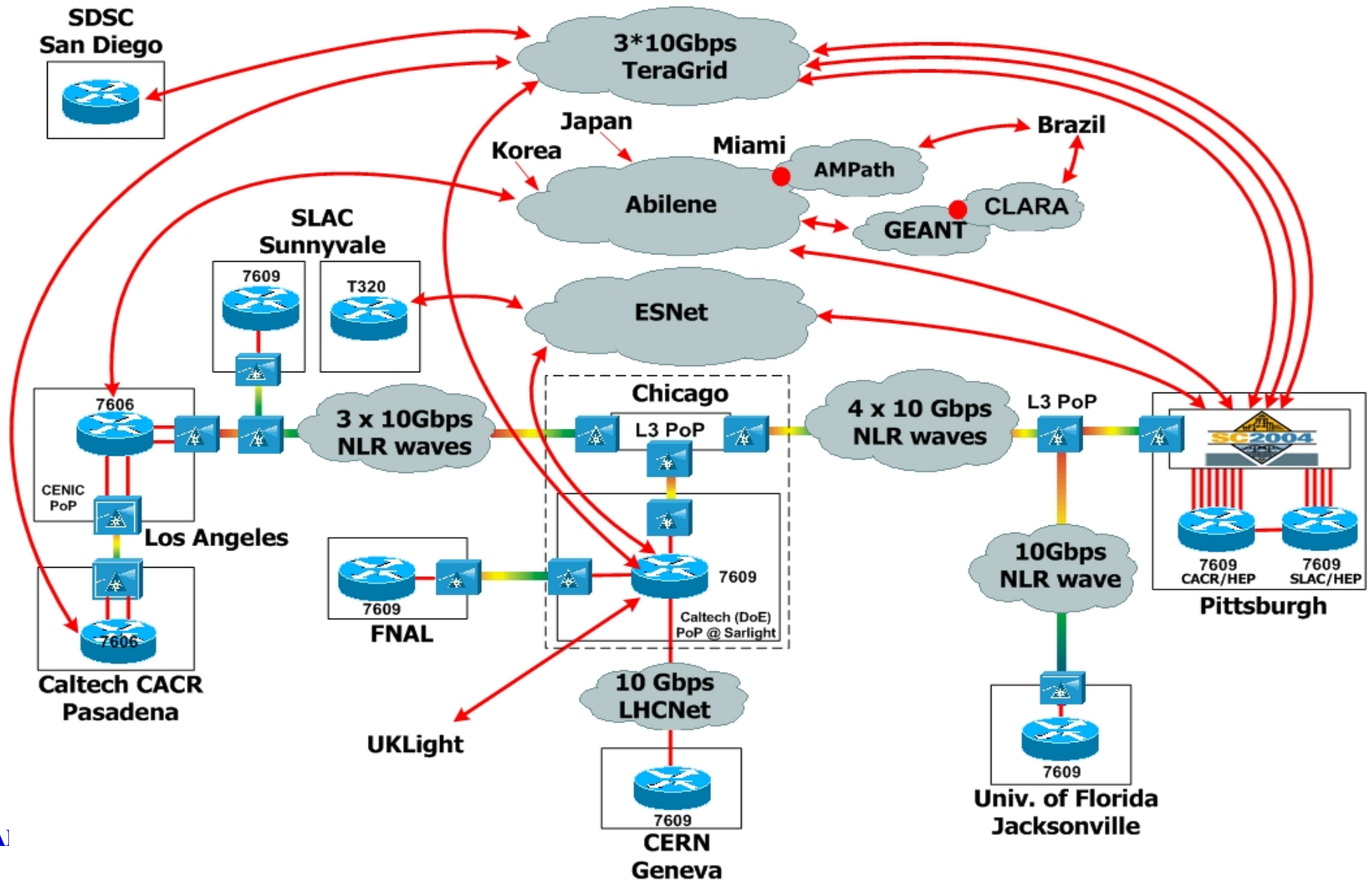
## Motivation and objectives



- ◆ **High Speed TeraByte Transfers for Physics: one-tenth of a terabits per second achieved (101 Gbps)**
- ◆ **Demonstrating that many 10 Gbps wavelengths can be used efficiently over various 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 MonALISA**
- ◆ **Major Partners : Caltech-FNAL-SLAC**
- ◆ **Major Sponsors:**
  - **Cisco, S2io, HP, Newysis**
- ◆ **Major networks:**
  - **NLR, Abilene, ESnet, LHCnet, AMPATH, TeraGrid**



# Network (I)



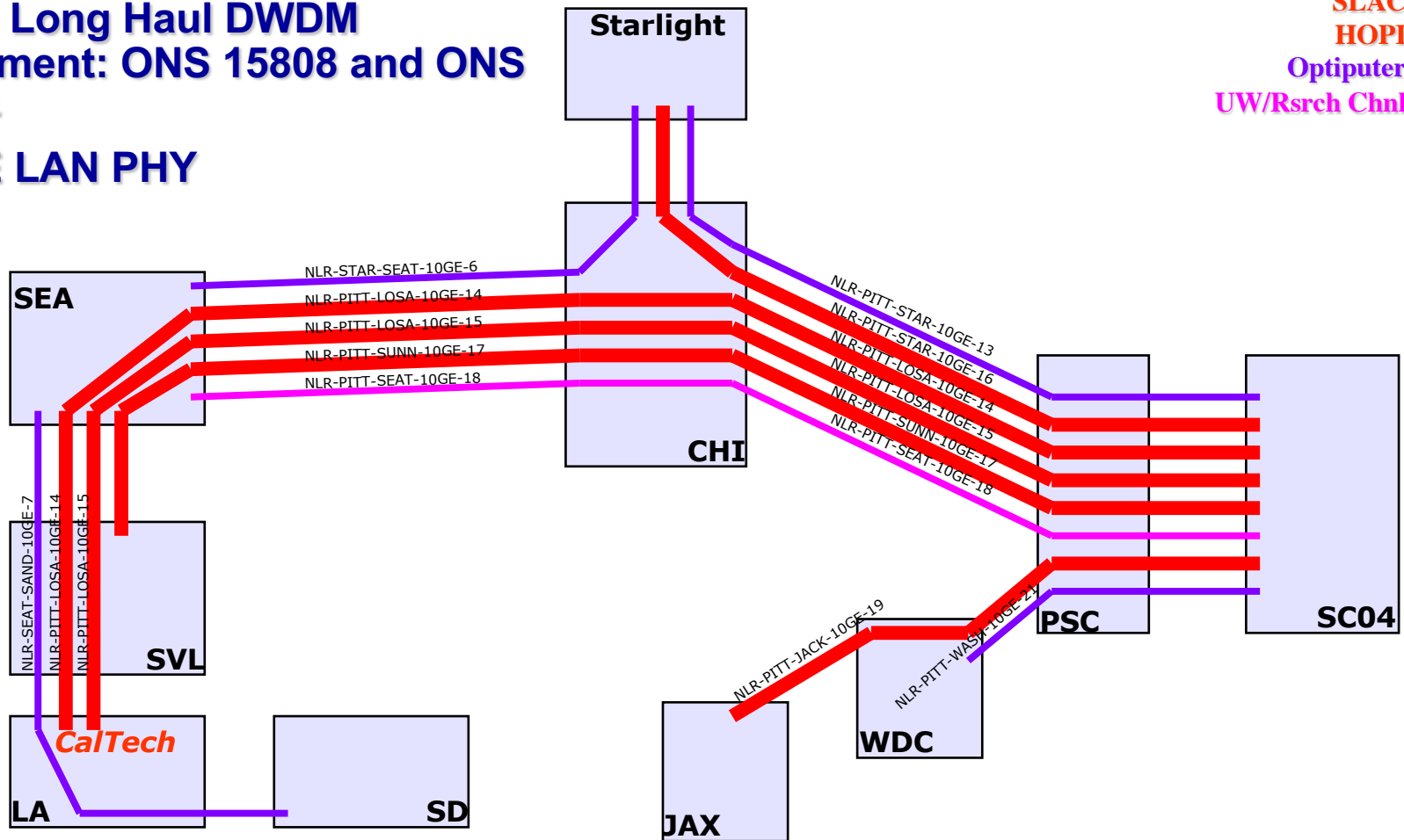


# 5 NLR Waves dedicated to HEP



- ☐ National Lambda Rail  
([www.nlr.net](http://www.nlr.net))
- ☐ Cisco Long Haul DWDM  
equipment: ONS 15808 and ONS 15454
- ☐ 10 GE LAN PHY

CalTech/Newman  
FL/Avery  
SLAC  
HOPI  
Optiputer  
UW/Rsrch Chnl





## Network (II)



- ◆ **80 10GE ports**
- ◆ **50 10GE network adapters**
- ◆ **10 Cisco 7600/6500 dedicated to the experiment**
- ◆ **Backbone for the experiment built in two days**
- ◆ **Four dedicated wavelengths of NLR to Caltech booth**
  - ❑ **Los Angeles (2 waves),**
  - ❑ **Chicago**
  - ❑ **Jacksonville**
- ◆ **One dedicated wavelengths of NLR to SLAC-FNAL booth from Sunnyvale**
- ◆ **Connections (L3 peerings) to showfloor via SCinet:**
  - ❑ **ESnet**
  - ❑ **TeraGrid**
  - ❑ **Abilene**
- ◆ **Balance traffic between WAN links**
  - ❑ **UltraLight Autonomous system dedicated to the experiment (AS32361)**
  - ❑ **L2 10 GE connections to LA, JACK and CHI**
  - ❑ **Dedicated LSPs (MPLS Label Switch Path) on Abilene**



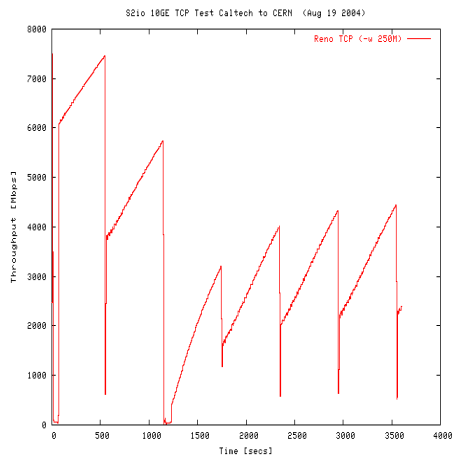


# Selecting the data Transport protocol



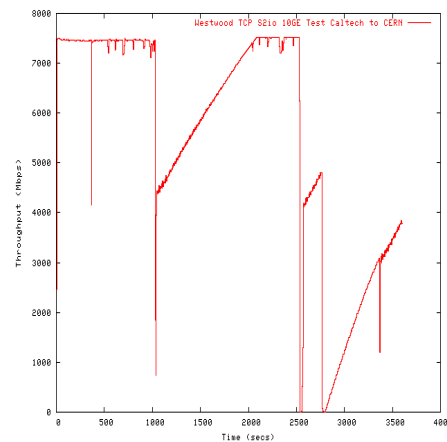
- ◆ Tests between CERN and Caltech
- ◆ Capacity = OC-192 9.5Gbps; 264 ms round trip latency; 1 flow
- ◆ ***Sending station:*** Tyan S2882 motherboard, 2 x Opteron 2.4 GHz, 2 GB DDR.
- ◆ ***Receiving station:*** (CERN OpenLab): HP rx4640, 4 x 1.5GHz Itanium-2, zx1 chipset, 8GB memory
- ◆ ***Network adapter:*** S2io 10 GE

3.0 Gbps



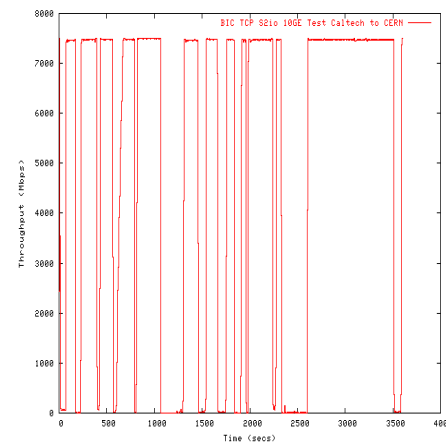
Linux TCP

4.1 Gbps



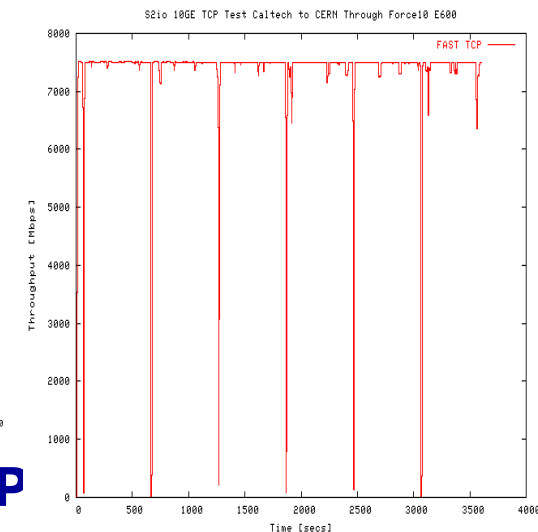
Linux Westwood+

5.0 Gbps



Linux BIC TCP

7.3 Gbps

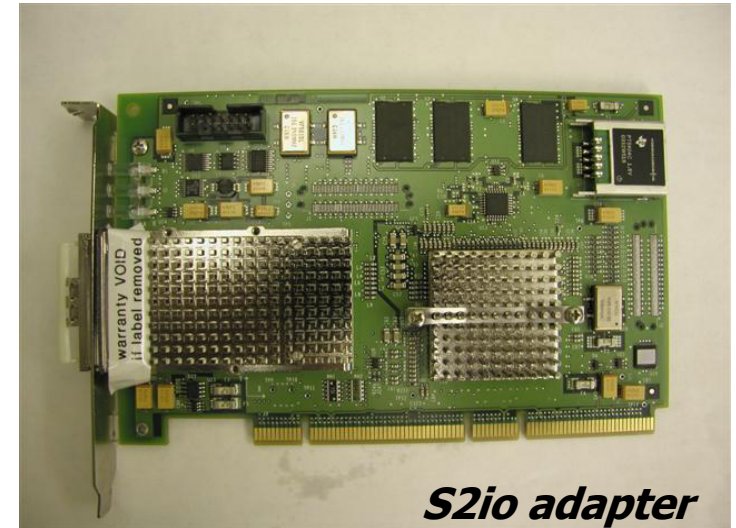




# Selecting 10 GE NICs



	<b>S2io Adapter</b>	<b>Intel Adapter</b>
<b>Max. b2b TCP throughput</b>	<b>7.5 Gbps</b>	<b>5.5 Gbps</b>
<b>Rx Frame Buffer Capacity</b>	<b>64MB</b>	<b>256KB</b>
<b>MTU</b>	<b>9600Byte</b>	<b>16114Byte</b>
<b>IPv4 TCP Large Send Offload</b>	<b>Max offload size 64k</b>	<b>Partial; Max offload size 32k</b>



*S2io adapter*

## ◆ TSO

- ◆ Must have hardware support in NIC.
- ◆ It allows TCP layer to send a larger than normal segment of data, e.g, 64KB, to the driver and then the NIC. The NIC then fragments the large packet into smaller ( $\leq$ mtu) packets.
- ◆ Benefits:
  - TSO can reduce CPU overhead by 10%~15%.
  - Increase TCP responsiveness.

$$p = (C * RTT * RTT) / (2 * MSS)$$

p: Time to recover to full rate

C: Capacity of the link

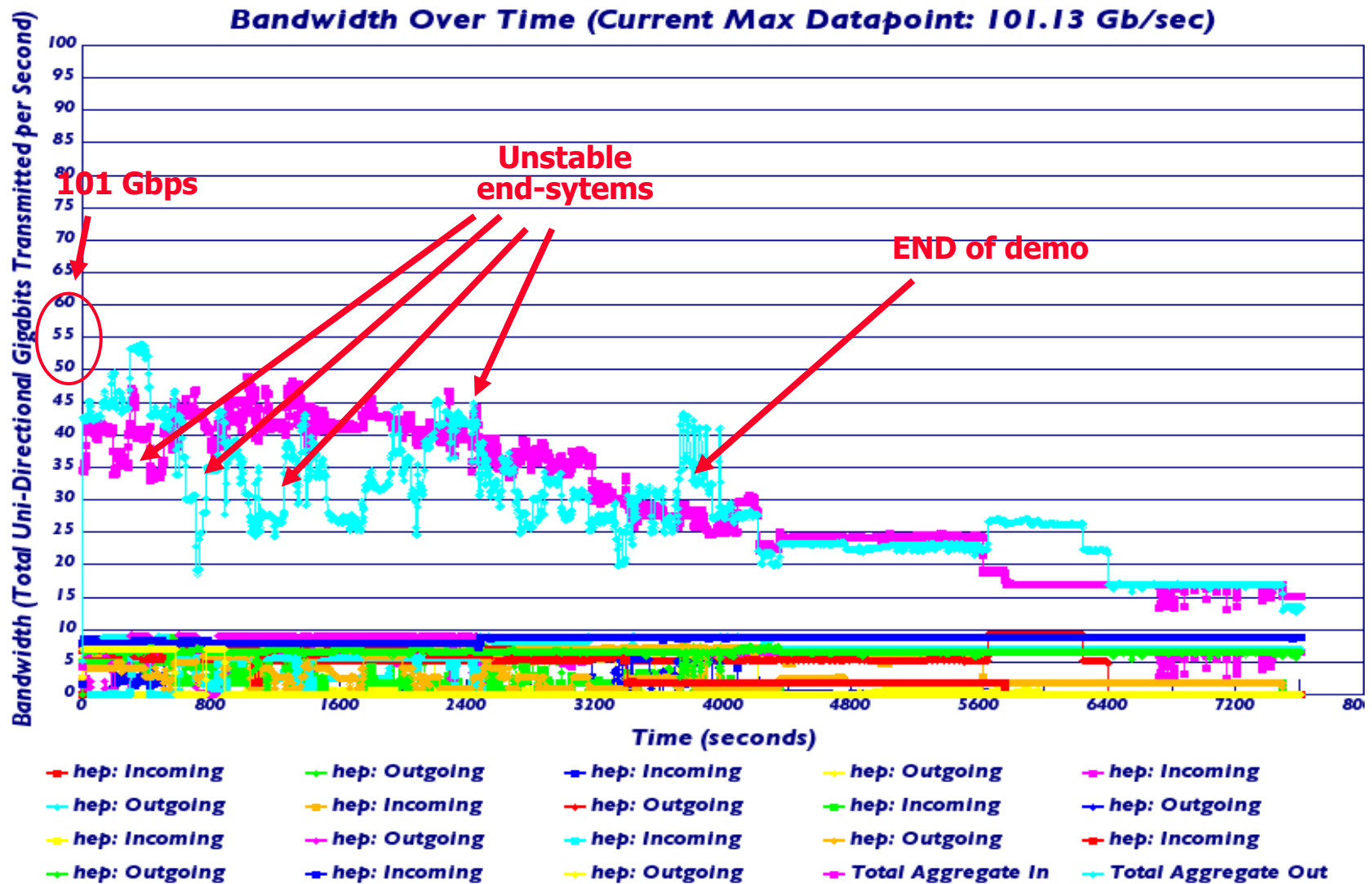
RTT: Round Trip Time

MSS: Maximum Segment Size





# 101 Gigabit Per Second Mark

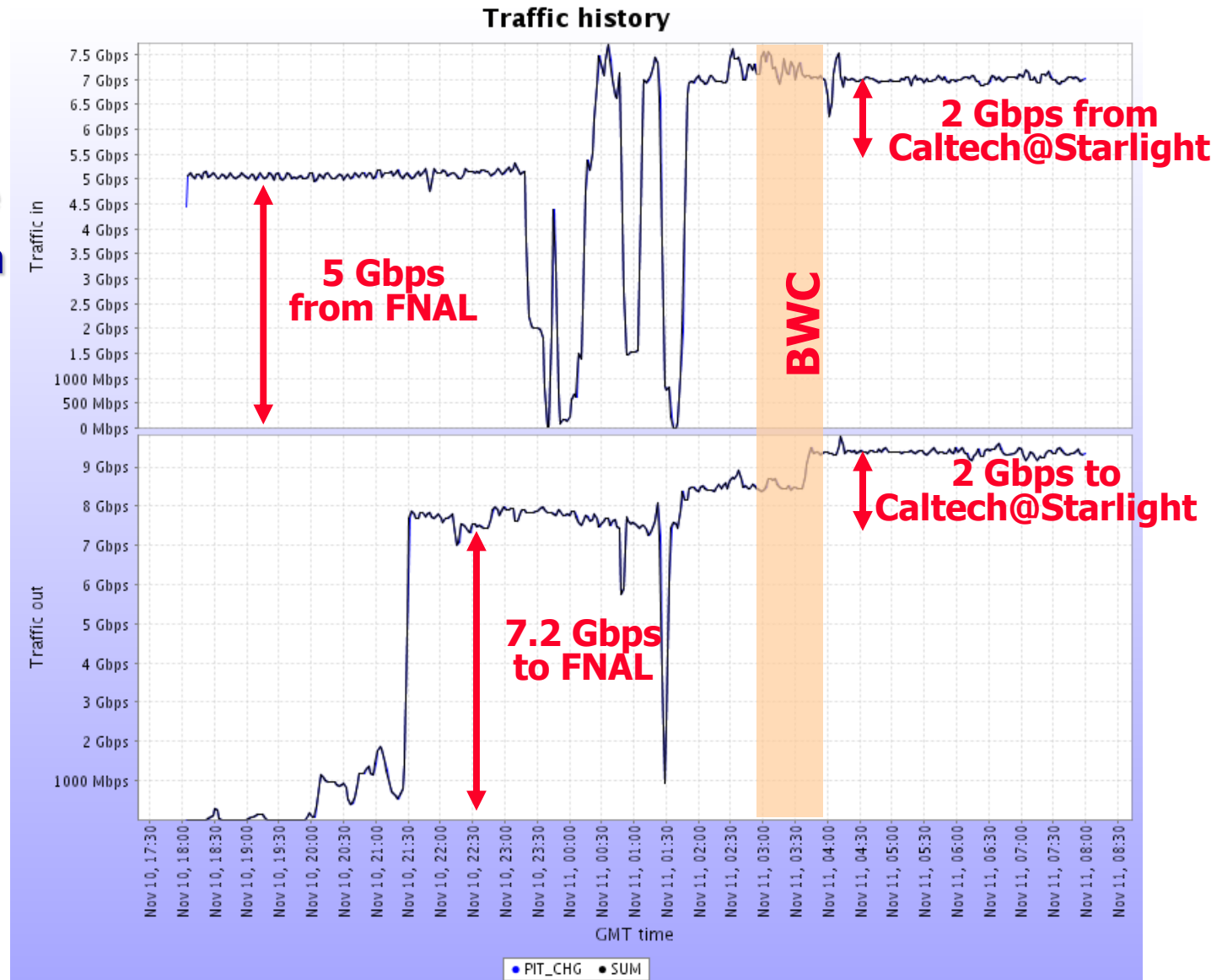




# PIT-CHI NLR wave utilization



- FAST TCP
- Performance to/from FNAL are very stable
- Bandwidth utilization during BWC = 80 %

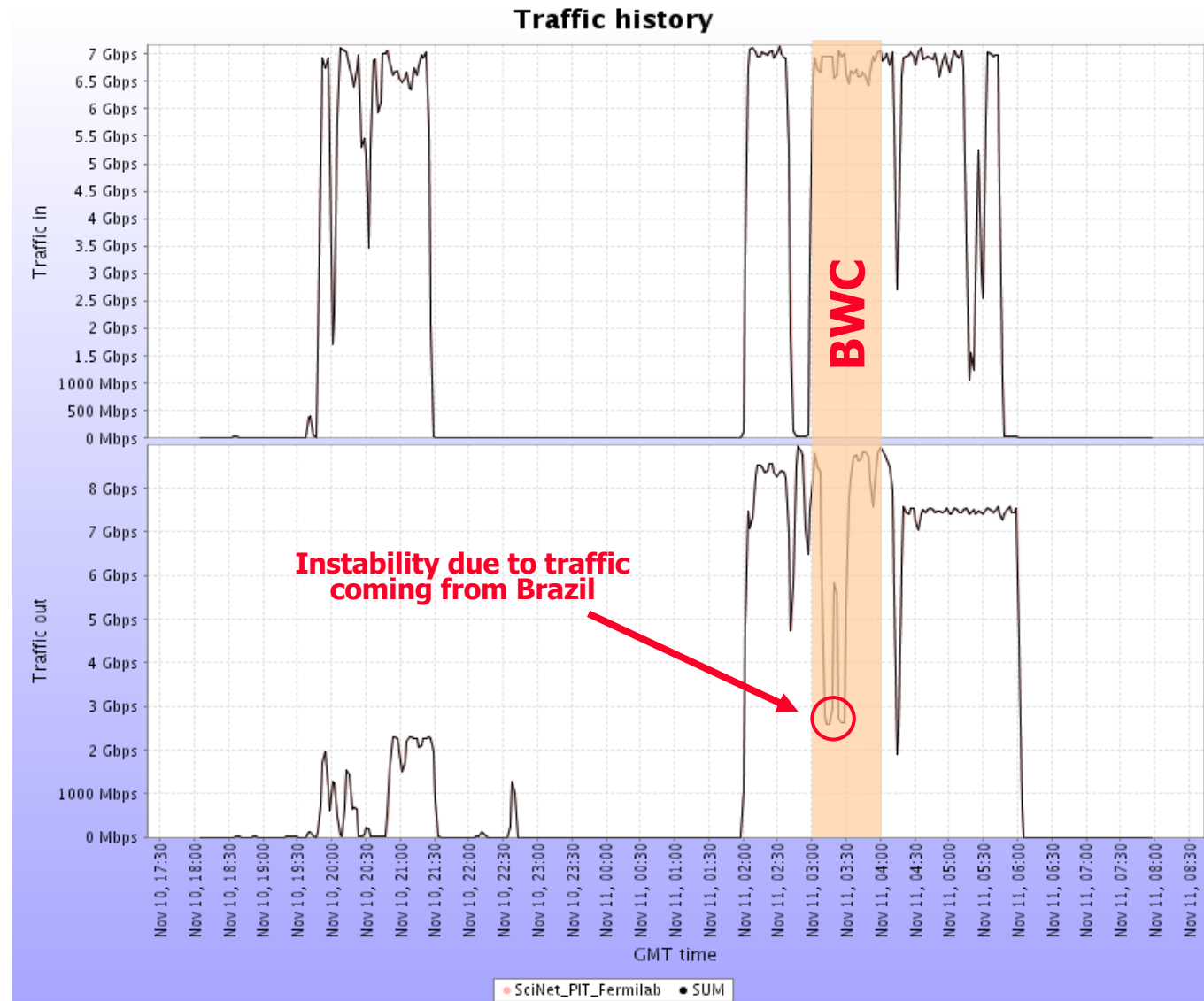




# Abilene Wave #1



- FAST TCP
- Traffic to CERN via Abilene & CHI
- Bandwidth utilization during BWC = 65 %

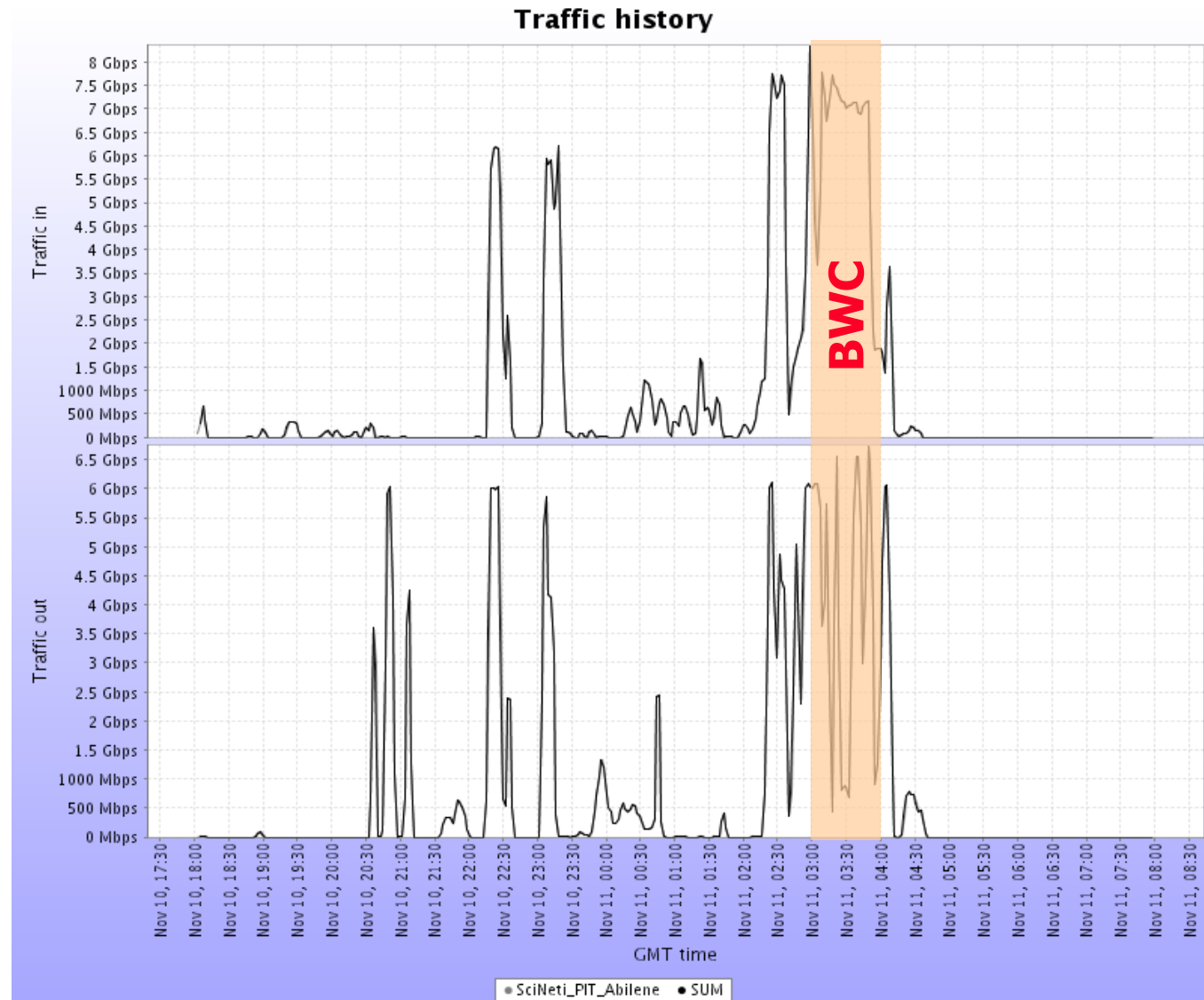




# Abilene Wave #2



- ❑ TCP Reno
- ❑ Traffic to LA via Abilene & NY
- ❑ Bandwidth utilization during BWC = 52.5 %
- ❑ Max. throughput is better with RENO but FAST is more stable
- ❑ Each time a packet is lost, the RENO flow has to be restarted

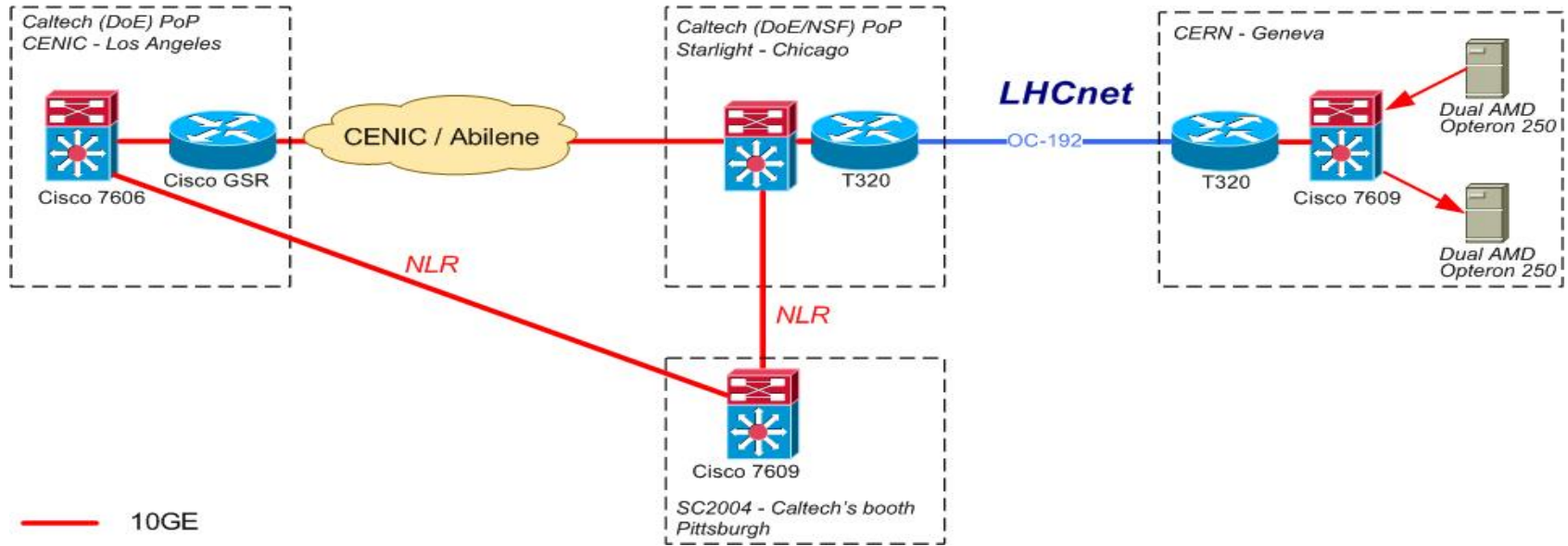
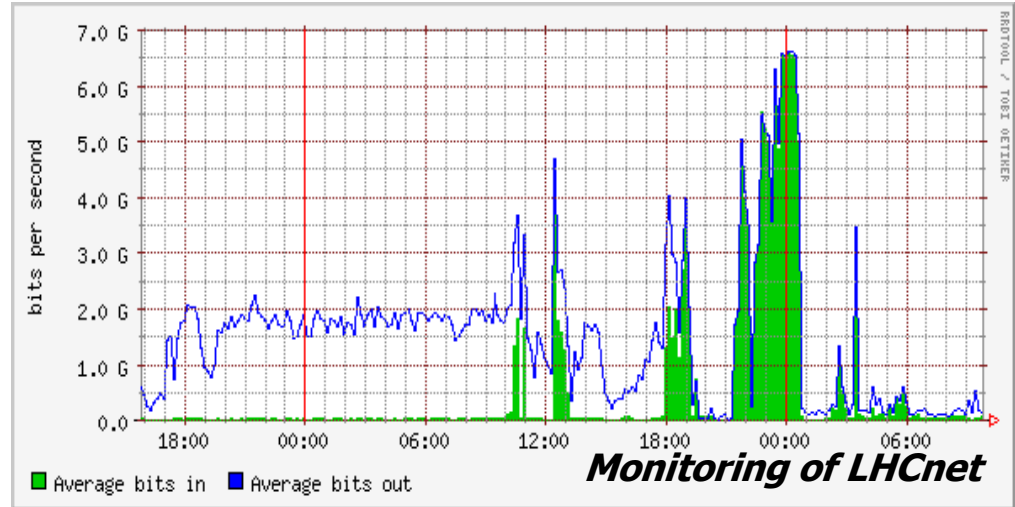




# Internet 2 Land Speed record submission

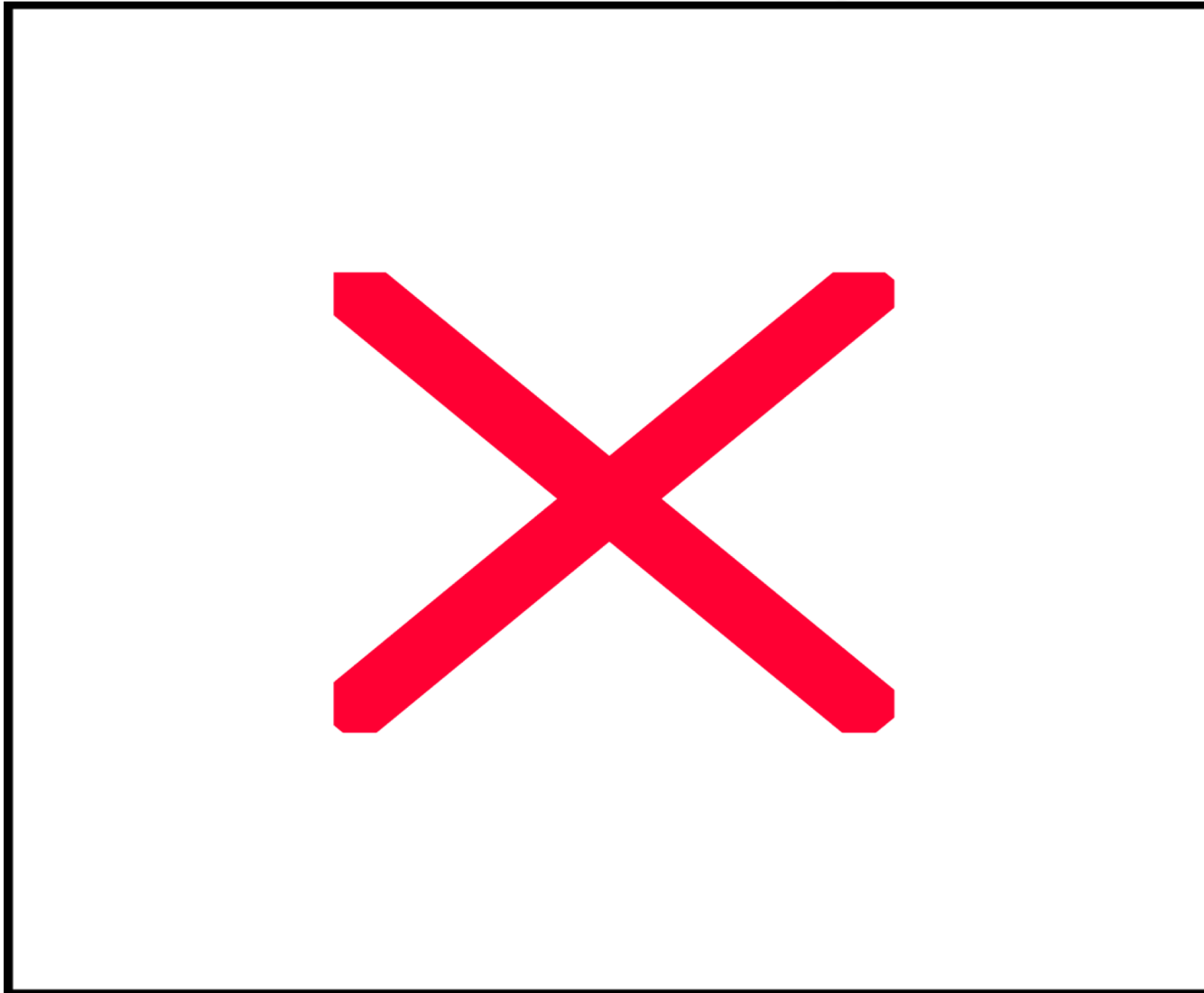


- ❑ Multi-Streams 6.86 Gbps X 27 kkm
- ❑ PATH: GVA-CHI-LA-PIT-CHI-GVA
- ❑ The LHCnet is crossed in both directions
- ❑ RTT = 466 ms
- ❑ FAST TCP (18 streams)





# An internationally collaborative effort: Brazilian example

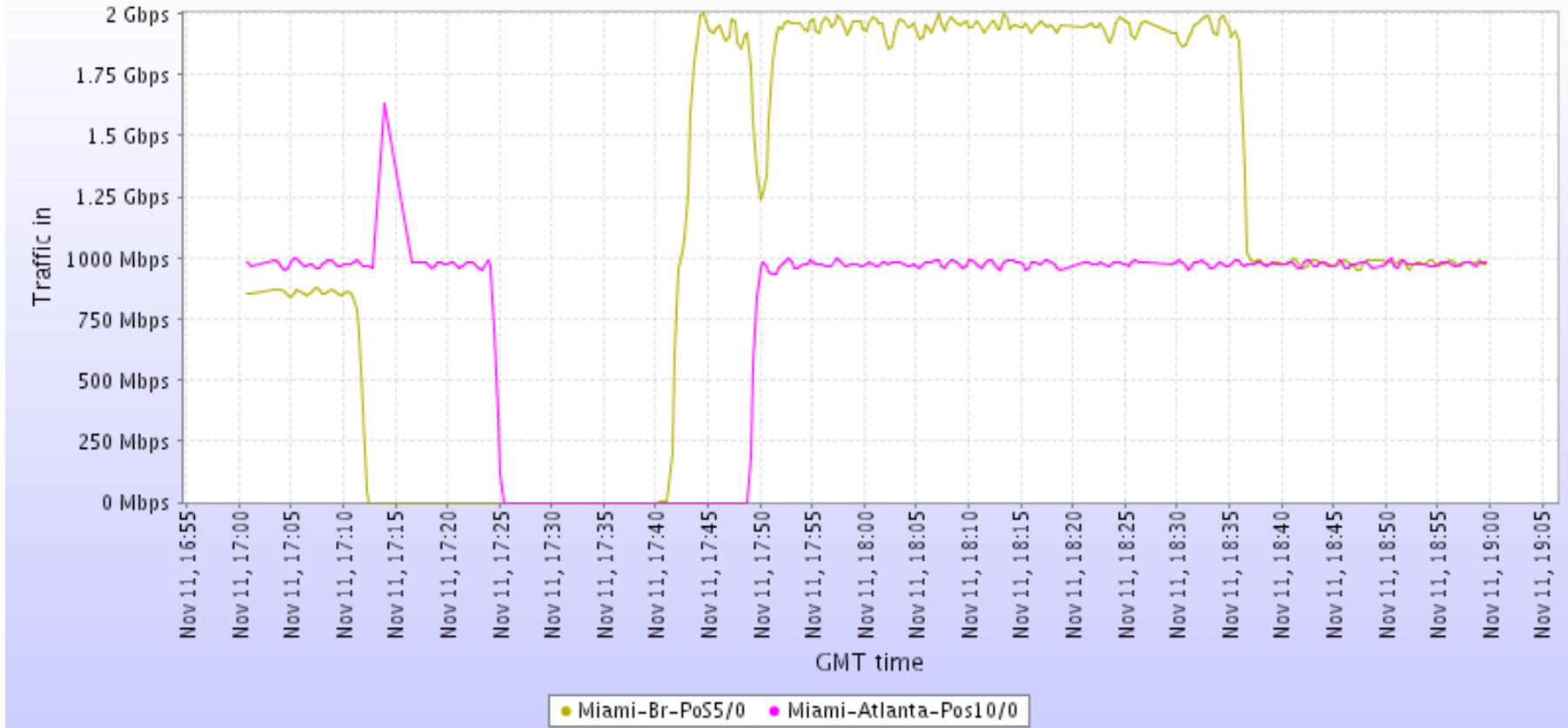




# 2 +1 Gbps network record on Brazillian network



### Traffic history





# A few lessons



- ◆ **Logistics: server crashes, NIC driver problems, Last minutes Kernel tuning, router misconfiguration, optics issues.**
- ◆ **Embarrassment of the rich: mix and match 3 10G SciNet drops with 2 Abilene and 3 TeraGrid waves. Better planning.**
- ◆ **Mixing small flows (1Gbps) with large flows (10 Gbps) and different RTTs can be tricky.**
  - ❑ Just filling in the leftover bandwidth?
  - ❑ Or negatively affect the large flows going over long distance?
  - ❑ Different level of FAST aggressiveness + packet losses on the path;
- ◆ **Efficient topological design**
  - ❑ instead of star-like topology to source-sink all traffic at the showfloor;
  - ❑ (routing traffic through a large router at showfloor but to remote servers.
- ◆ **Systematic tuning of the servers**
  - ❑ Server crashes main reason of performance problems;
  - ❑ 10GbE NIC drivers improvement;
  - ❑ Better ventilation at the show;
  - ❑ Using large default socket buffer versus setting large window by applications (iperf)





# Partners

