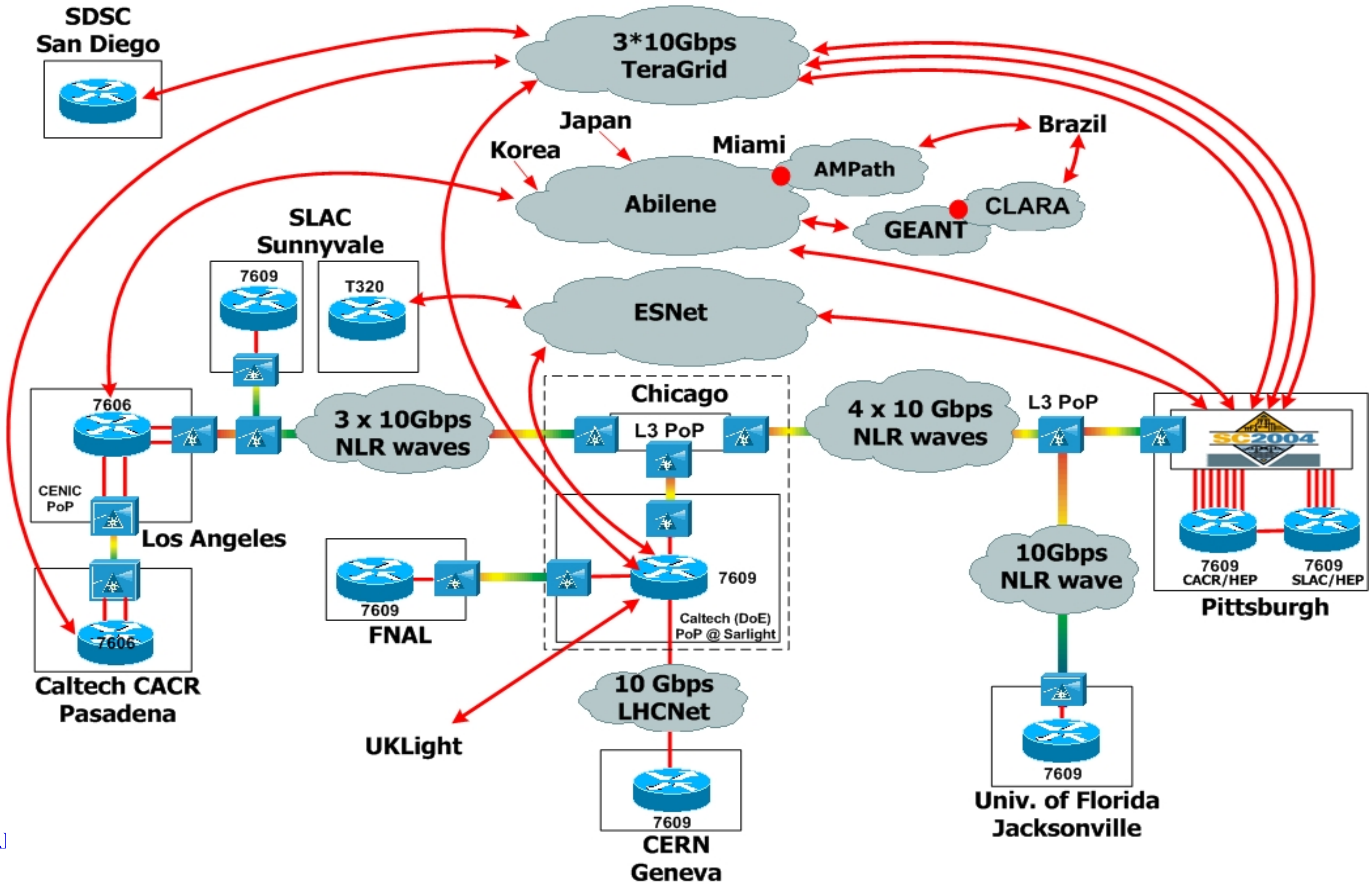
How we quadrupled the network speed world record: *bandwidth Challenge at SC 2004*

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)

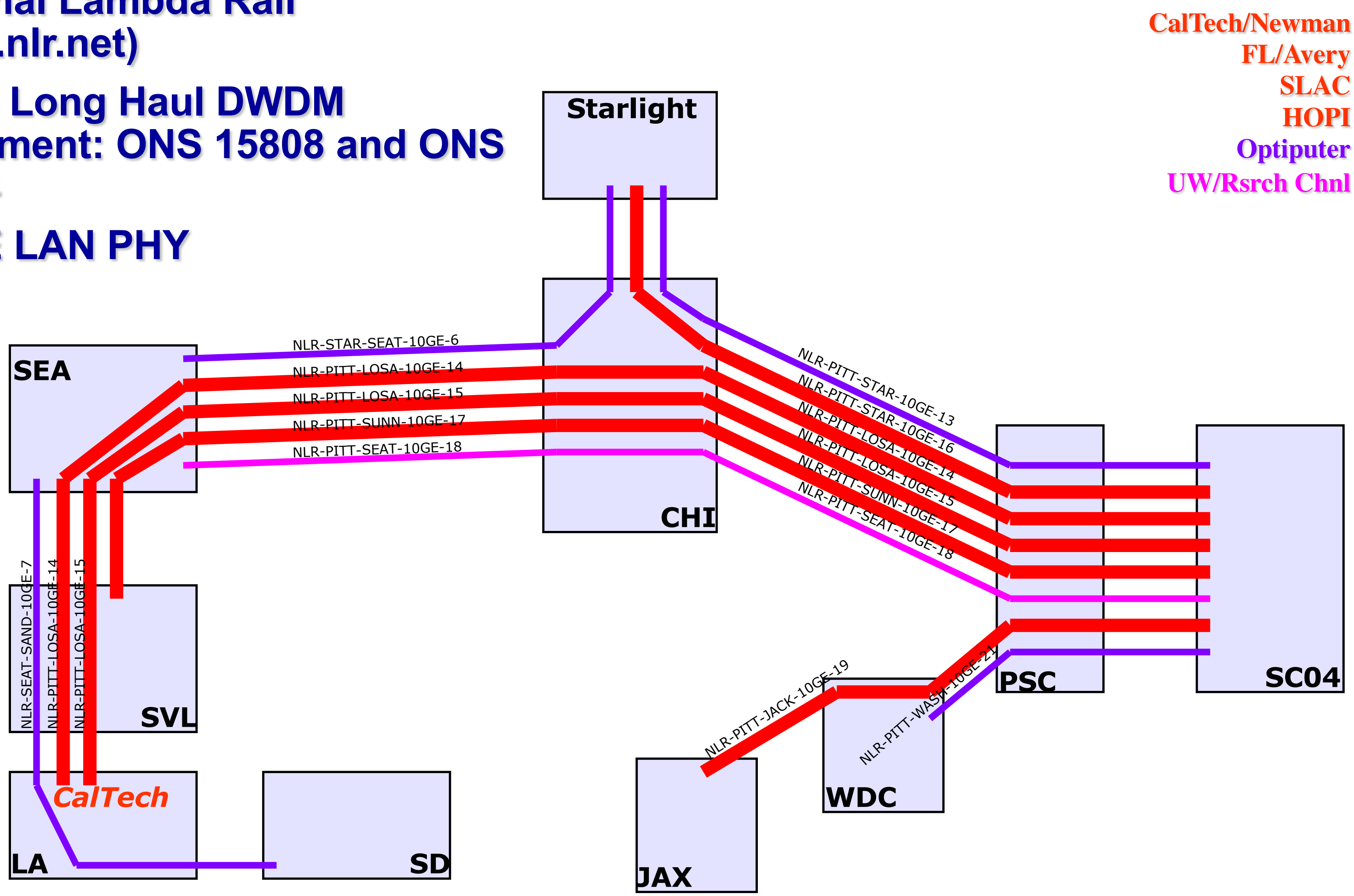


SDSC
San Diego
3*10Gbps
TeraGrid
Japan
Korea
Miami
AMPath
Brazil
Abilene
CLARA
GEANT
SLAC
Sunnyvale
7609
T320
ESNet
7606
CENIC
PoP
Los Angeles
3 x 10Gbps
NLR waves
Chicago
L3 PoP
4 x 10 Gbps
NLR waves
L3 PoP
SC2004
7609
CACR/HEP
7609
SLAC/HEP
Pittsburgh
7606
Caltech CACR
Pasadena
7609
FNAL
7609
Caltech (DoE)
PoP @ Sarlight
10Gbps
NLR wave
UKLight
10 Gbps
LHCNet
7609
CERN
Geneva
7609
Univ. of Florida
Jacksonville

CA

5 NLR Waves dedicated to HEP

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

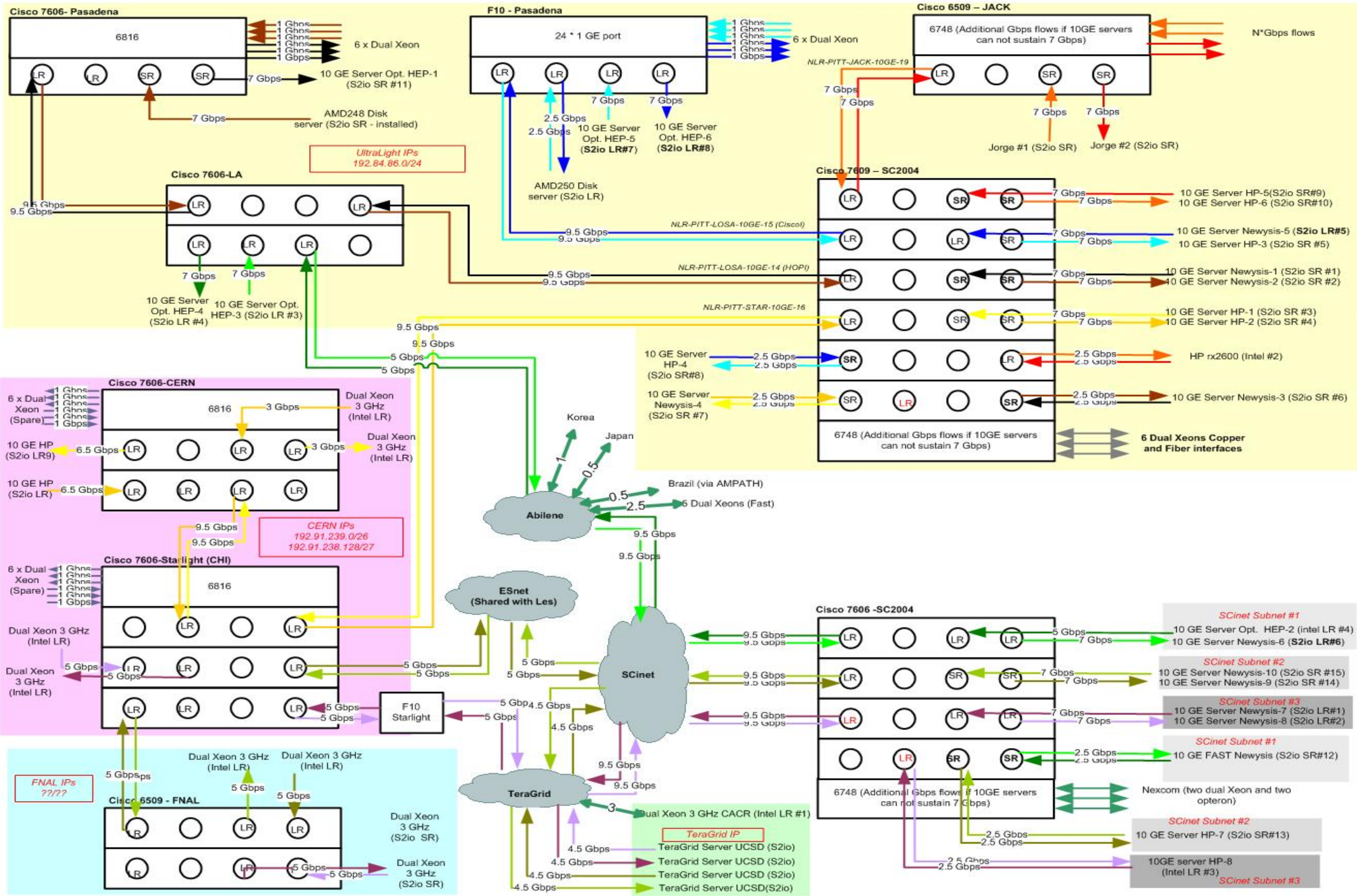


All lines 10GE

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

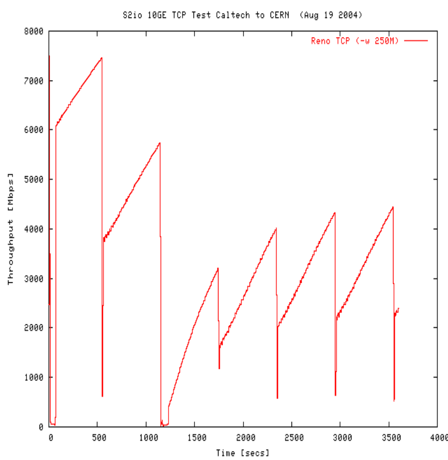
Flows Matrix



Selecting the data Transport protocol

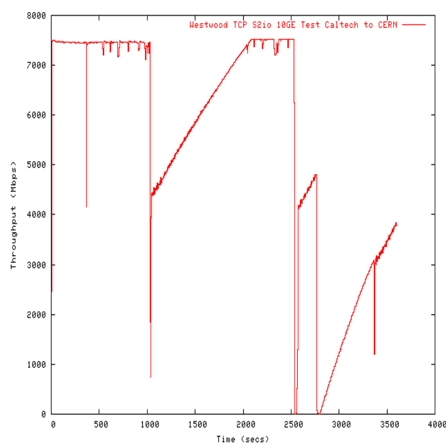
- Tests between CERN and Caltech
- Capacity = OC-192 9.5Gbps; 264 ms round trip latency; 1 flow
- *Sending station:* Tyan S2882 motherboard, 2 x Opteron 2.4 GHz, 2 GB DDR.
- *Receiving station:* *(CERN OpenLab):* HP rx4640, 4 x 1.5GHz Itanium-2, zx1 chipset, 8GB memory
- *Network adapter:* S2io 10 GE

3.0 Gbps



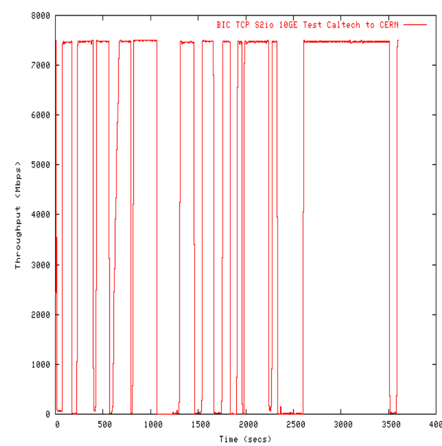
Linux TCP

4.1 Gbps



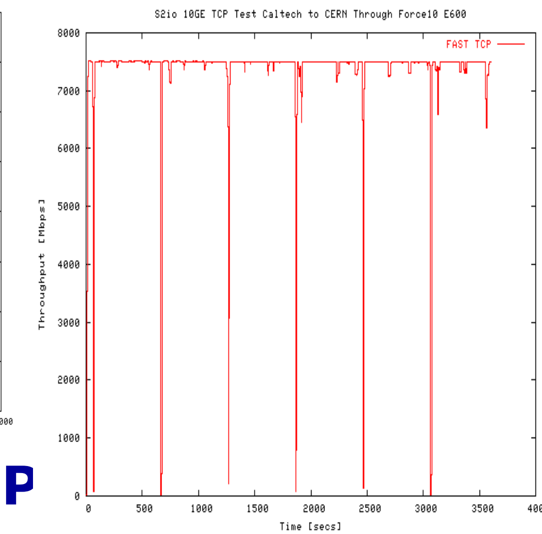
Linux Westwood+

5.0 Gbps



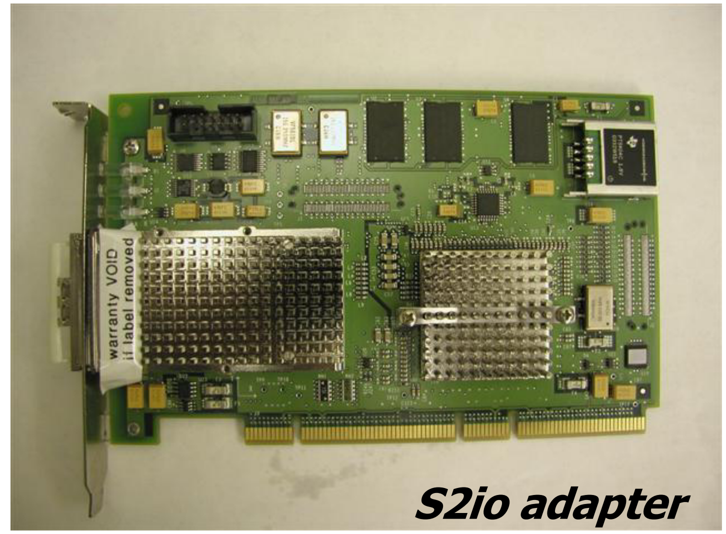
Linux BIC TCP

7.3 Gbps



Selecting 10 GE NICs

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



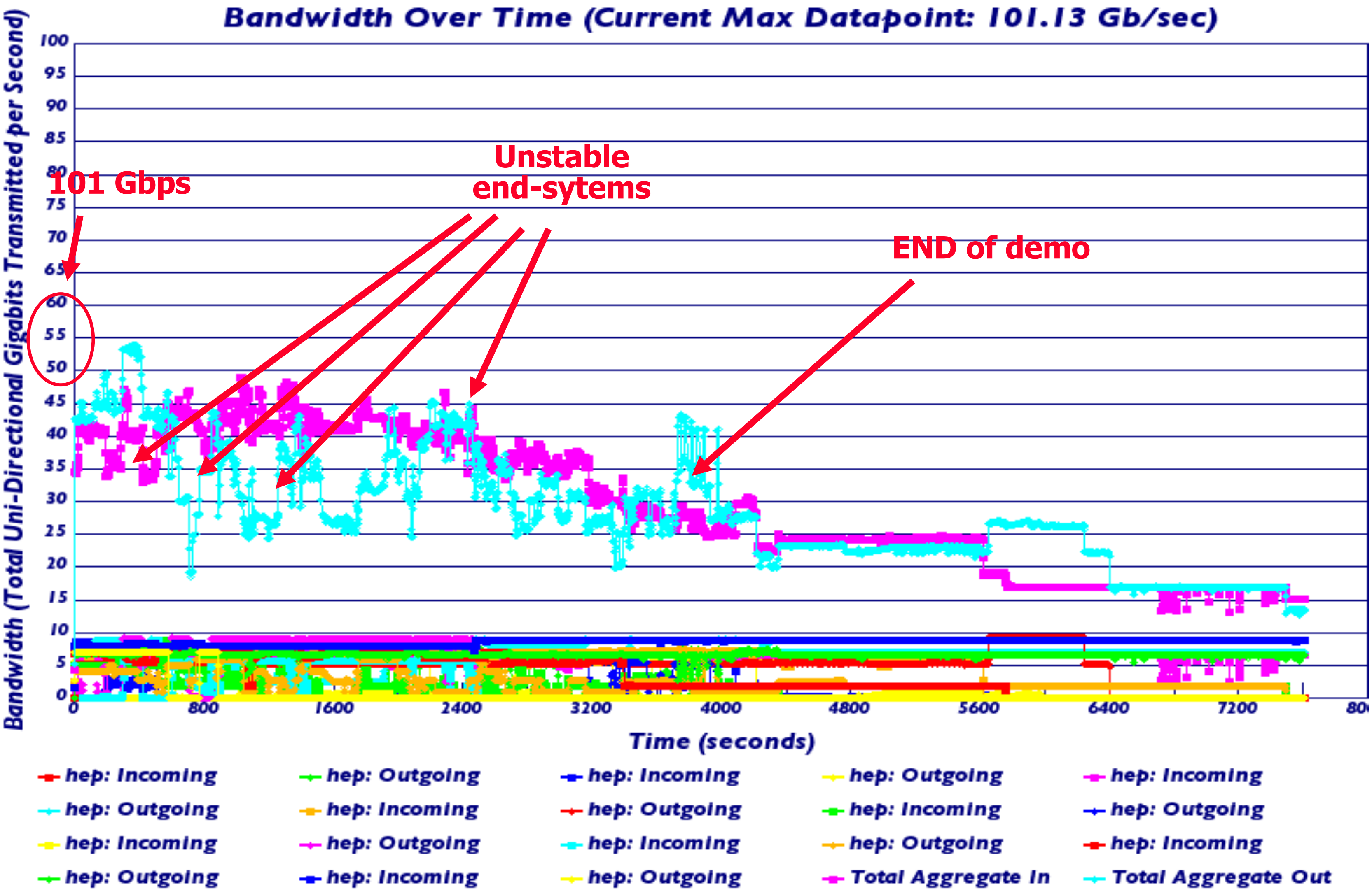
S2io adapter

- **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 (<=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

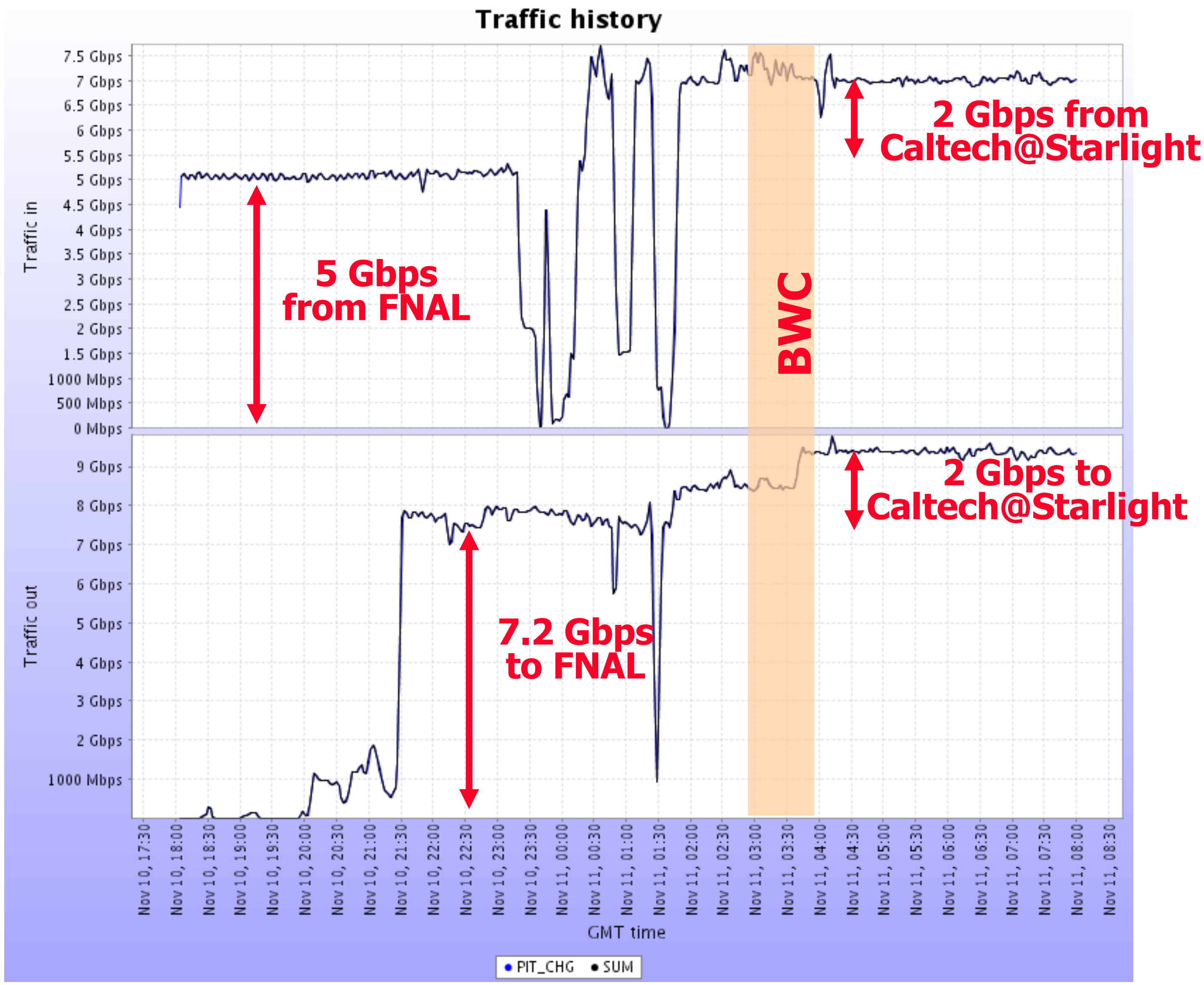
101 Gigabit Per Second Mark



Source: Bandwidth Challenge committee

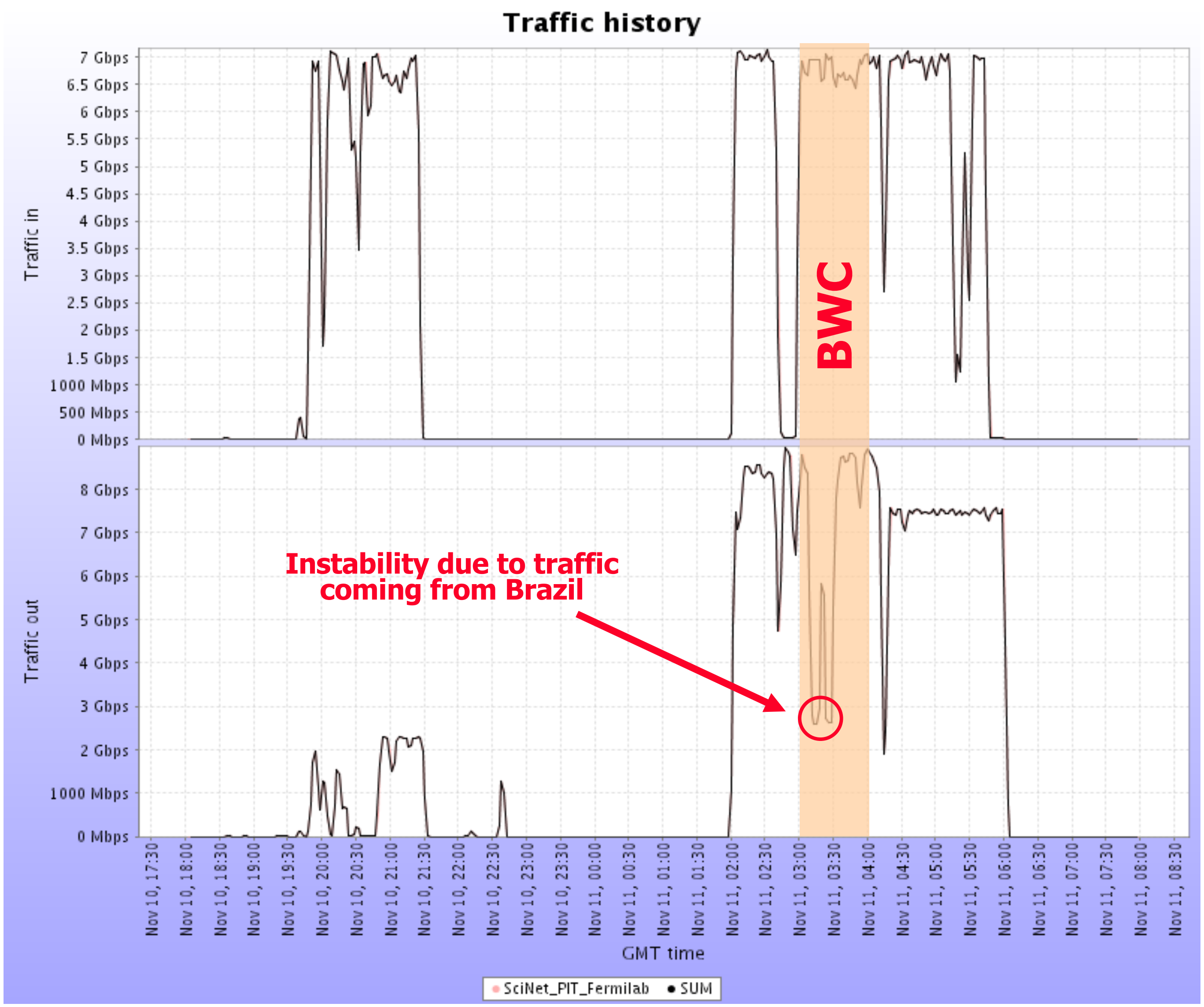
PIT-CHI NLR wave utilization

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



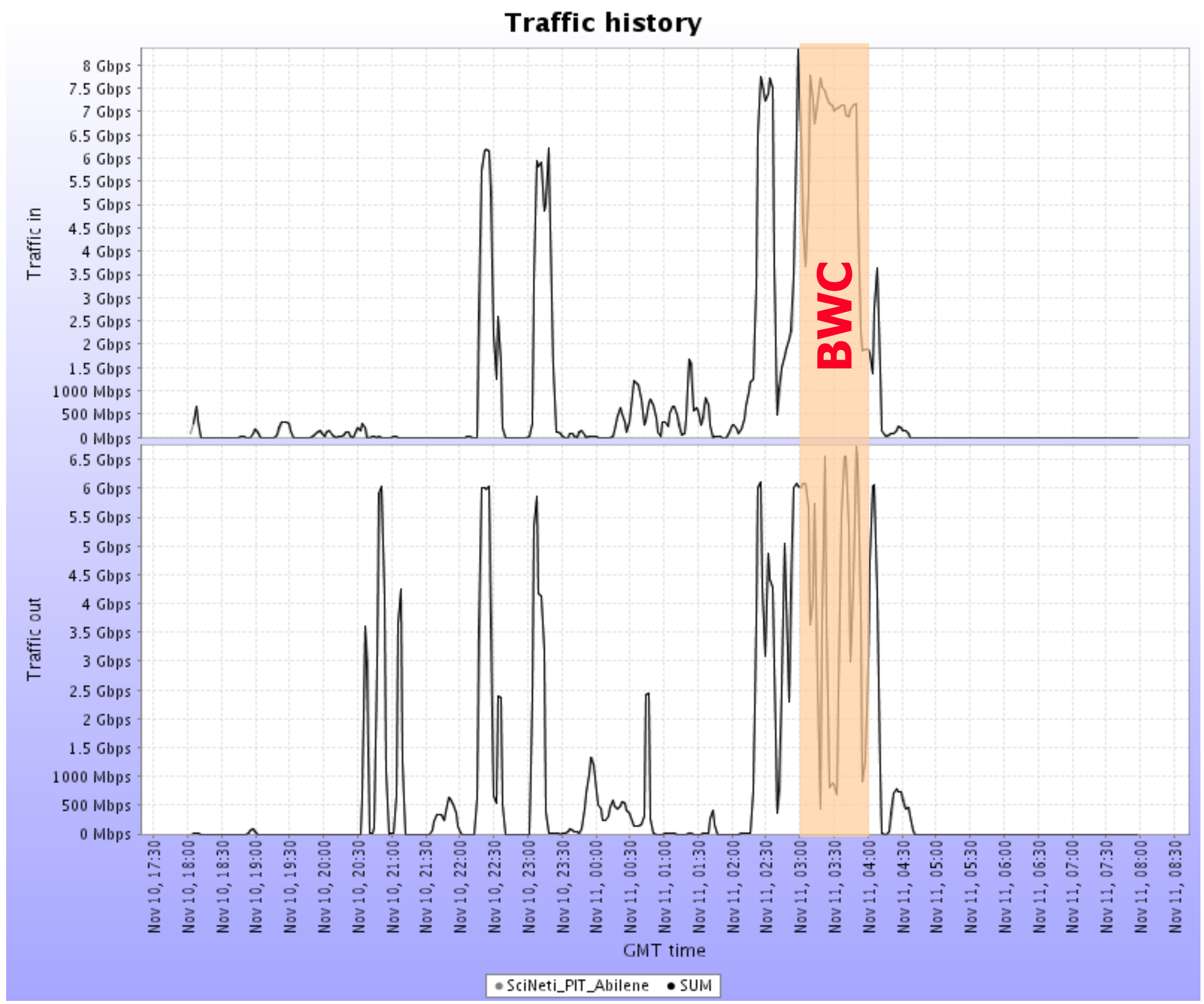
Abilene Wave #1

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



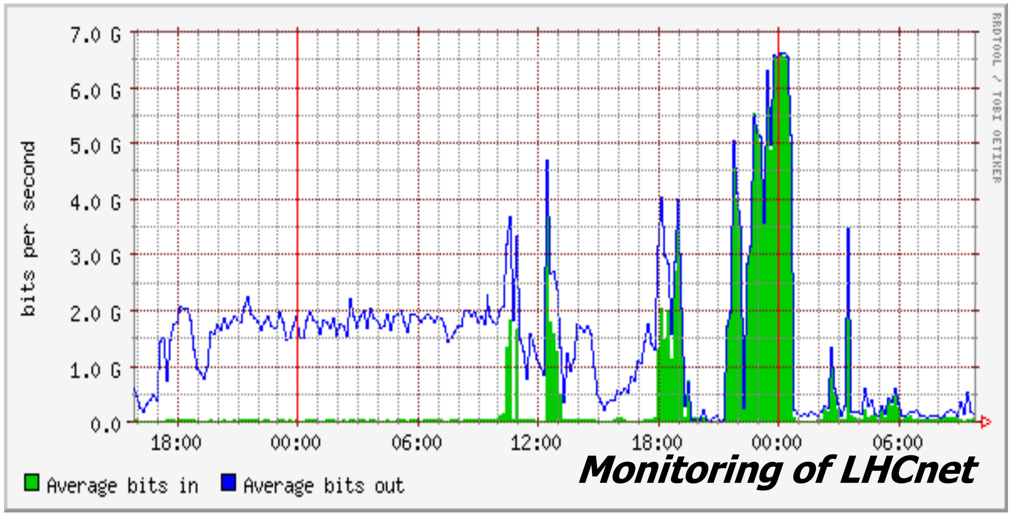
Abilene Wave #2

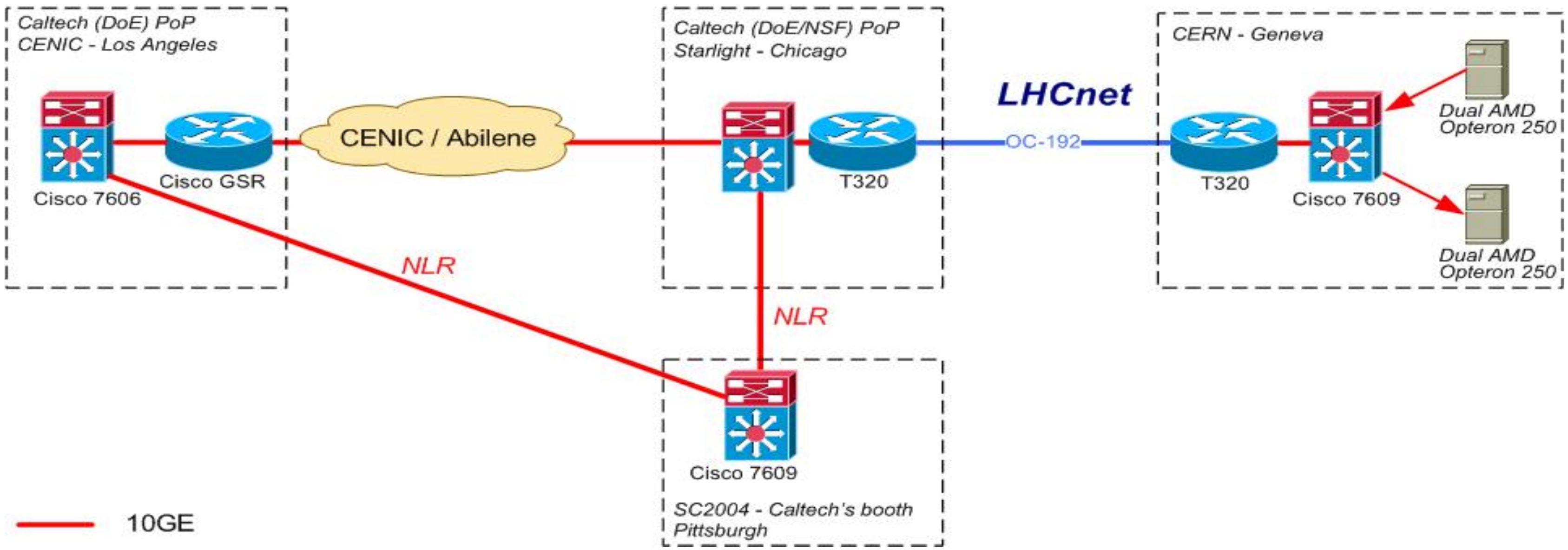
- TCP Reno
- Traffic to LA via Abilene & NY
- Bandwidth utilization during BWC = 52.5 %
- Max. throughput is better with RENO but FAST is more stable
- Each time a packet is lost, the RENO flow has to be restarted



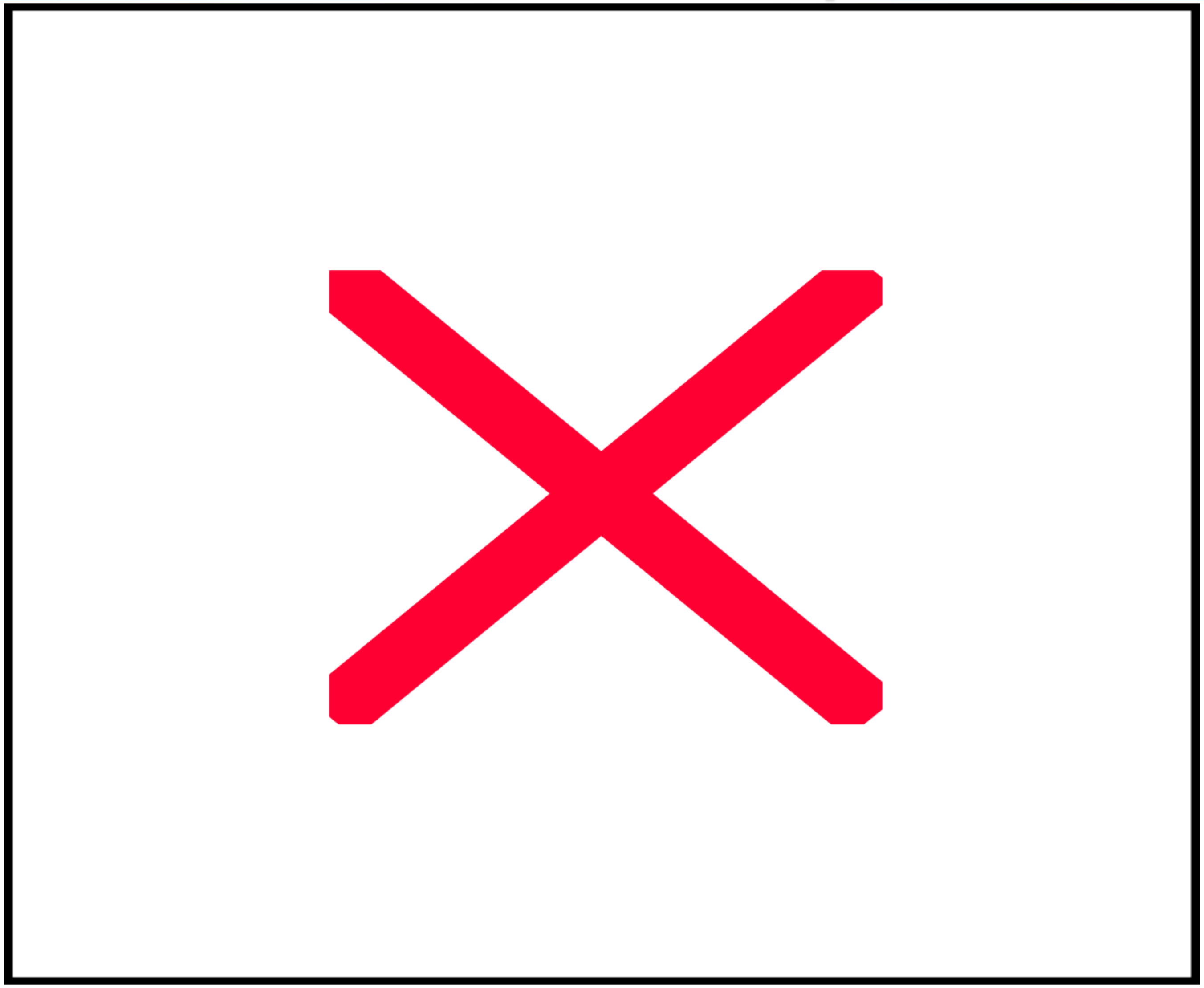
Internet 2 Land Speed record submission

- 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)

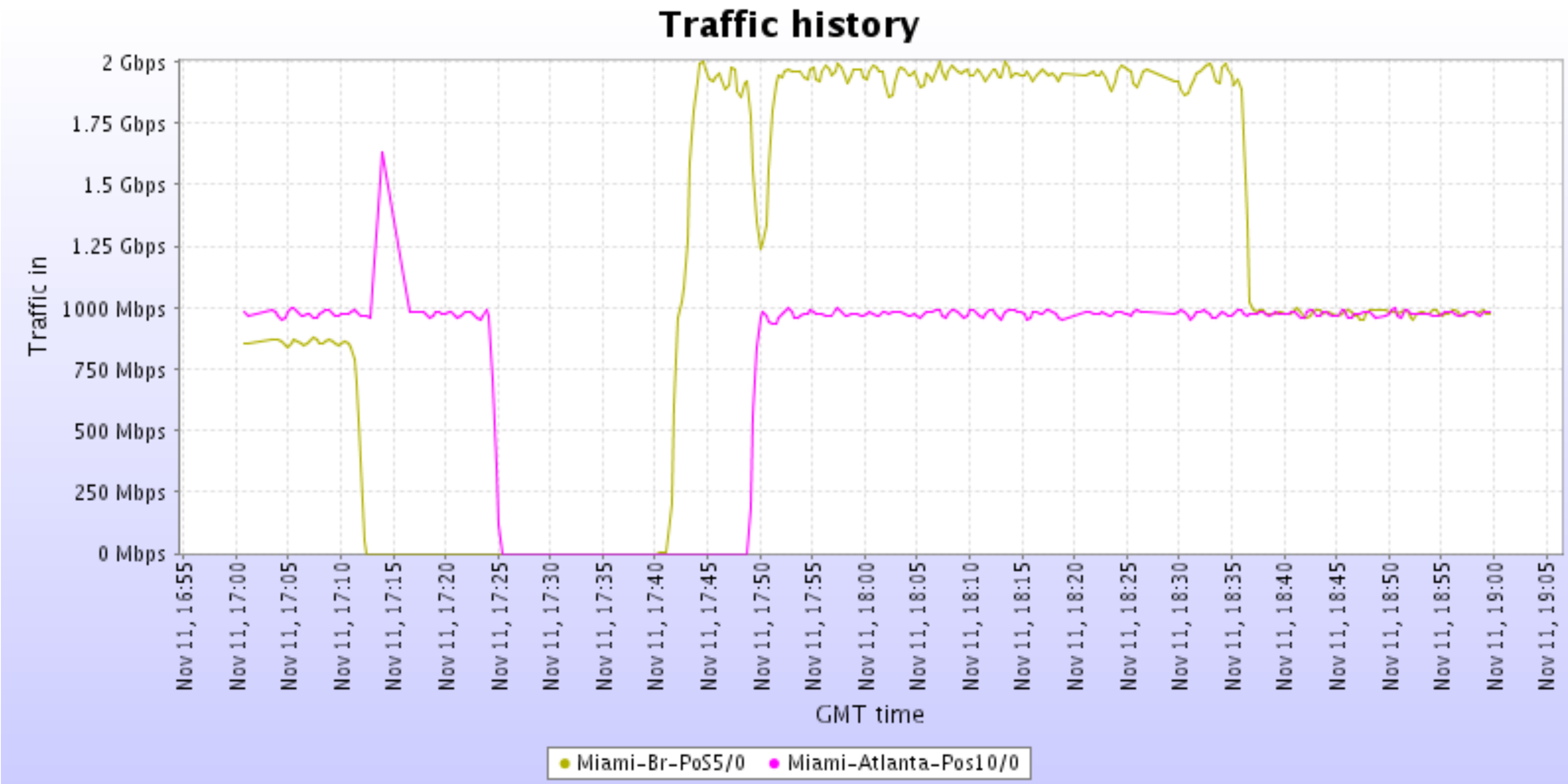




An internationally collaborative effort: Brazillian example



2 +1 Gbps network record on Brazillian network



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

GEANT
Rio de Janeiro State University
UKLight
Cisco Systems
NSF
Kyungpook National University
Telemar
RNP
StarLight℠ The Optical STAR TAP™
NLR light the future
DANTE www.dante.net
The University of Manchester
hp invent
PPDG
University of Florida
Internet2
S2io
intelig Telecom
Ampath
California Institute of Technology 1891
RedCLARA
HEP grid Brasil
AMD
TeraGrid
Stanford Linear Accelerator Center
Lambda Station
FINEP Financiadora de Estudos e Projetos
CPqD Telecom & IT Solutions
Abilene
Advanced networking for leading-edge research and education
HEP
Newisys A Sanmina-SCI Company
CERN
ICFA SCIC
iVDGL
ALICE
Microsoft
ESnet
USP Universidade de São Paulo BRASIL
KAIST
NetLogger
GriPhyN Data Intensive Science
Fermilab
unesp
intel