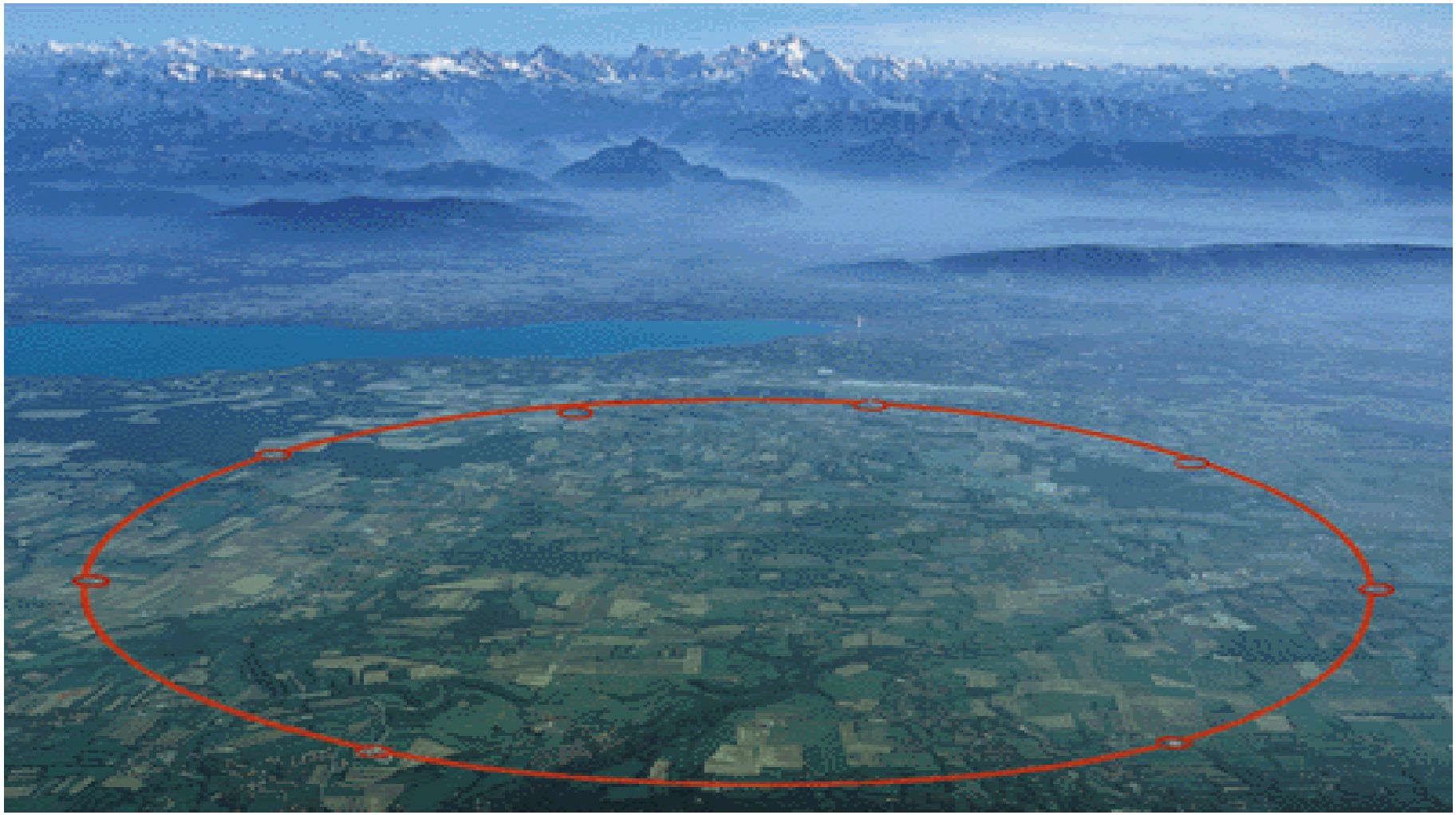
Wide Area Networking Performance Challenges



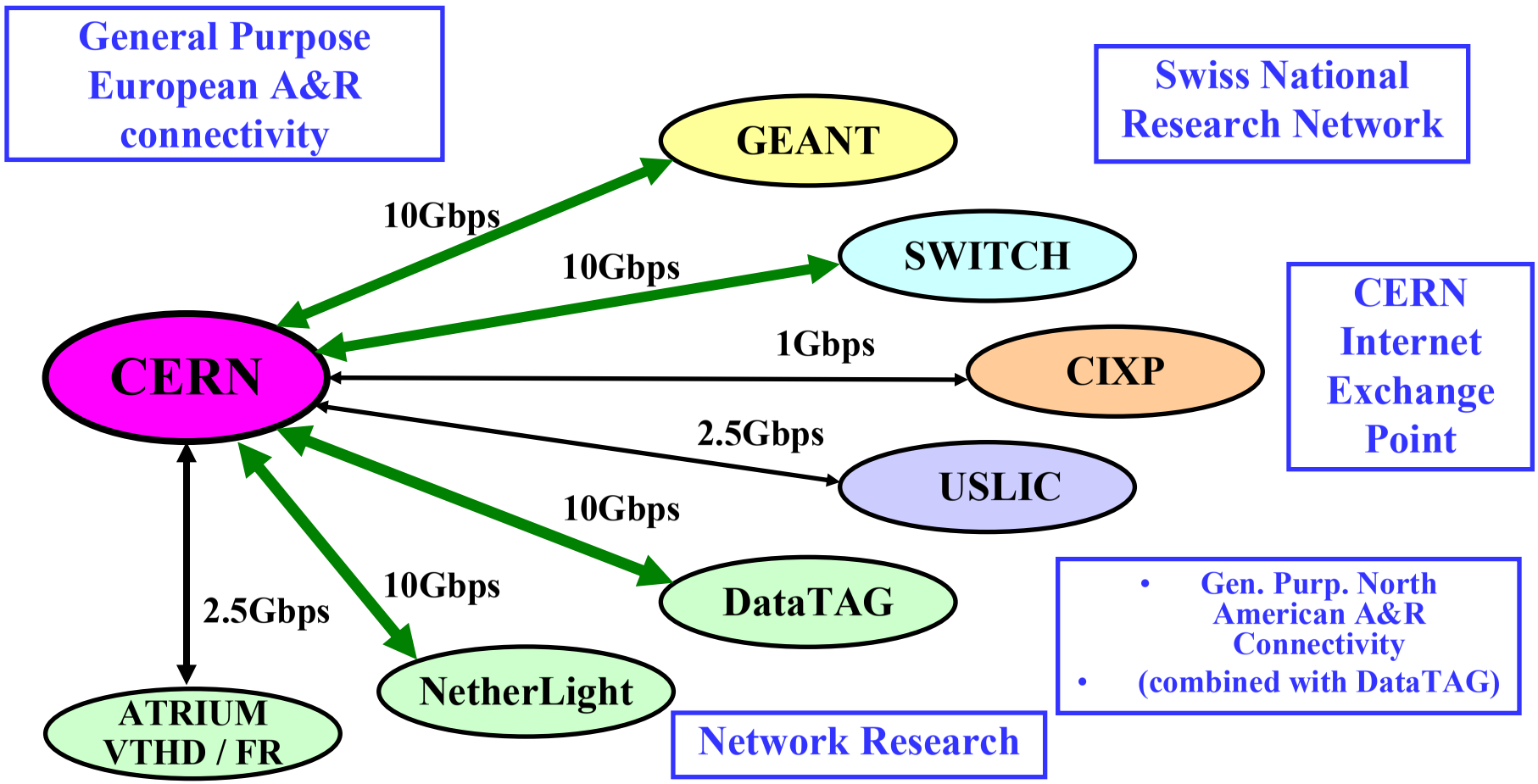
Olivier Martin, CERN

UK DTI visit

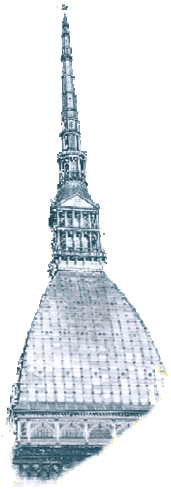
Presentation Outline

- CERN's connectivity to the Internet
- DataTAG project overview
- Wide Area Networking challenges
  - Where do we want to be by the start of the LHC in 2007?
  - Where are we now?

CERN External Networking
Main Internet Connections



Project partners



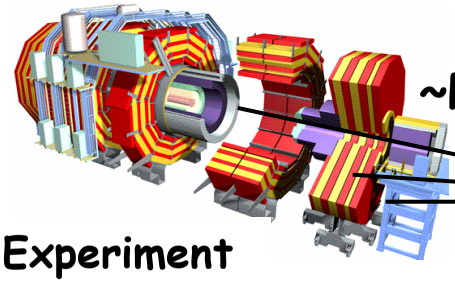
http://www.datatag.org

DataTAG Mission

TransAtlantic Grid

- EU ⇔ US Grid network research
  - High Performance Transport protocols
  - Inter-domain QoS
  - Advance bandwidth reservation
- EU ⇔ US Grid Interoperability
- Sister project to EU DataGRID

LHC Data Grid Hierarchy



Experiment

~PByte/sec

Online System

~100-400 MBytes/sec

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

Tier 0 +1

CERN 700k SI95
~1 PB Disk;
Tape Robot

10 Gbps

Tier 1

IN2P3 Center

RAL Center

INFN Center

FNAL: 200k SI95; 600 TB

2.5/10 Gbps

Tier 2

Tier2 Center

~2.5 Gbps

Tier 3

Institute ~0.25TIPS

Physics data cache

0.1–1 Gbps

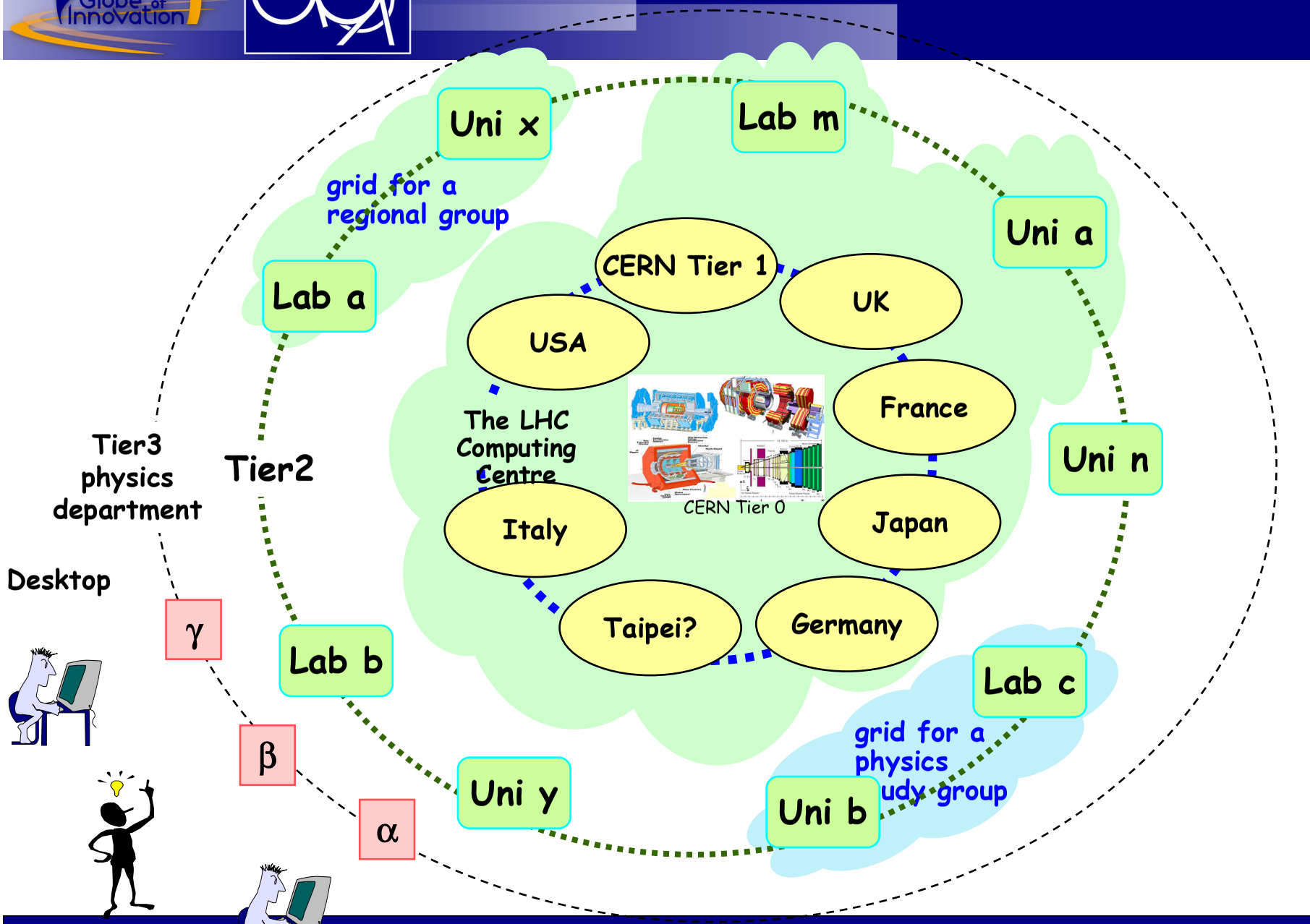
Tier 4

Workstations

Physicists work on analysis "channels"

Each institute has ~10 physicists working on one or more channels

Deploying the LHC Grid



les.robertson@cern.ch

Main Networking Challenges

- Fulfill the, yet unproven, assertion that the network can be « nearly » transparent to the Grid
- Deploy suitable Wide Area Network infrastructure (50-100 Gb/s)
- Deploy suitable Local Area Network infrastructure (matching or exceeding that of the WAN)
- Seamless interconnection of LAN & WAN infrastructures
  - firewall?
- End to End issues (transport protocols, PCs (Itanium, Xeon), 10GigE NICs (Intel, S2io)
  - where are we today:
    - memory to memory: 6.5Gb/s
    - memory to disk: 1.2MB (Windows 2003 server/NewiSys)
    - disk to disk: 400MB (Linux), 600MB (Windows)

Main TCP issues

Does not scale to some environments
- High speed, high latency
- Noisy

Unfair behaviour with respect to:
- Round Trip Time (RTT)
- Frame size (MSS)
- Access Bandwidth

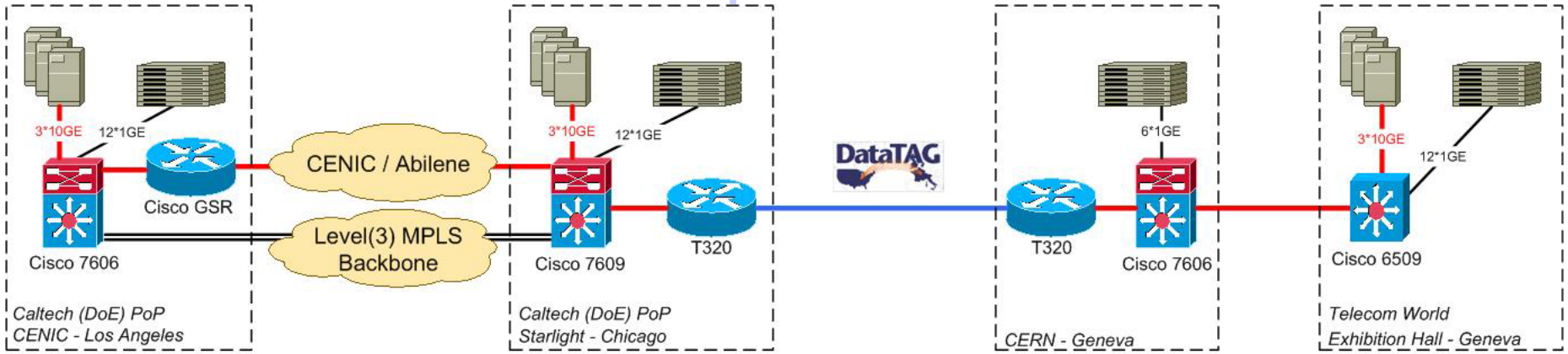
Widespread use of multiple streams in order to compensate for inherent TCP/IP limitations (e.g. Gridftp, BBftp):
- Bandage rather than a cure

New TCP/IP proposals in order to restore performance in single stream environments
- Not clear if/when it will have a real impact
- In the mean time there is an absolute requirement for backbones with:
  - Zero packet losses,
  - And no packet re-ordering

TCP dynamics (10Gbps, 100ms RTT, 1500Bytes packets)

- Window size (W) = Bandwidth*Round Trip Time
  - Wbits = 10Gbps*100ms = 1Gb
  - Wpackets = 1Gb/(8*1500) = 83333 packets
- Standard Additive Increase Multiplicative Decrease (AIMD) mechanisms:
  - W=W/2 (halving the congestion window on loss event)
  - W=W + 1 (increasing congestion window by one packet every RTT)
- Time to recover from W/2 to W (congestion avoidance) at 1 packet per RTT:
  - RTT*Wp/2 = 1.157 hour
  - In practice, 1 packet per 2 RTT because of delayed acks, i.e. 2.31 hour
- Packets per second:
  - RTT*Wpackets = 833’333 packets

10G DataTAG testbed extension to Telecom World 2003 and Abilene/Cenic



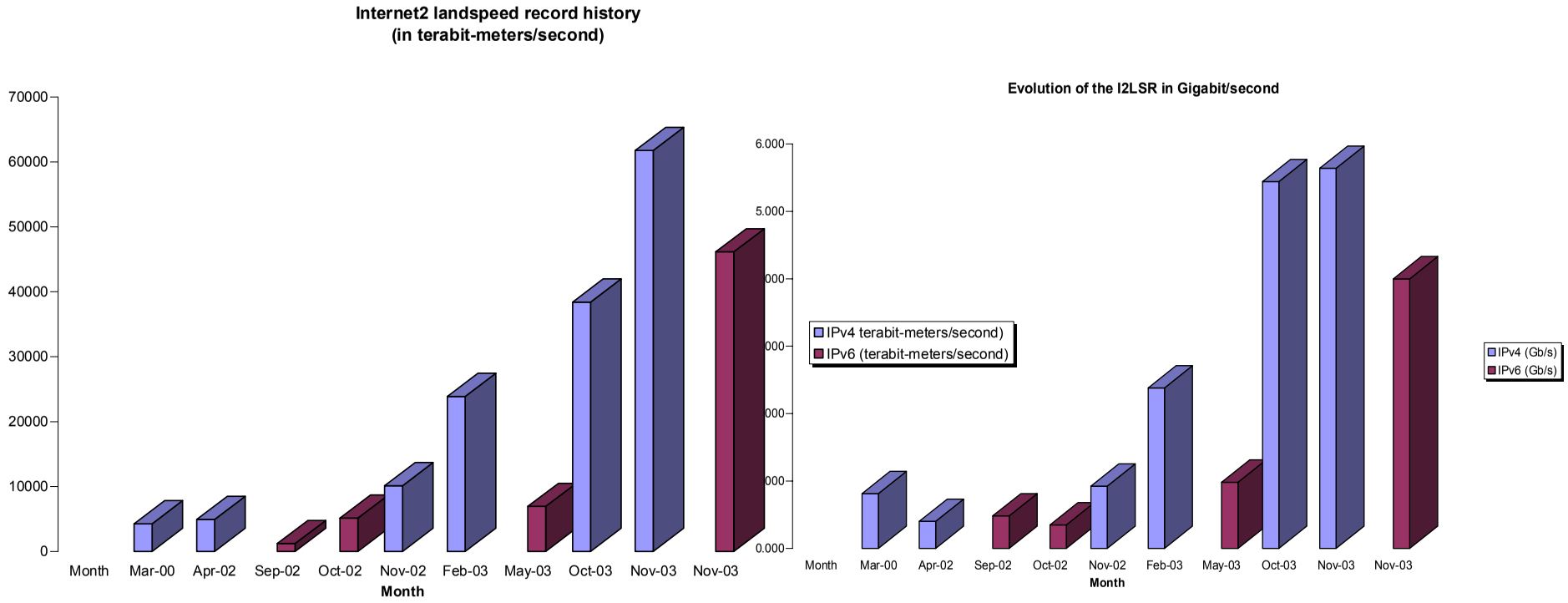
1G ethernet
10G ethernet
OC-192

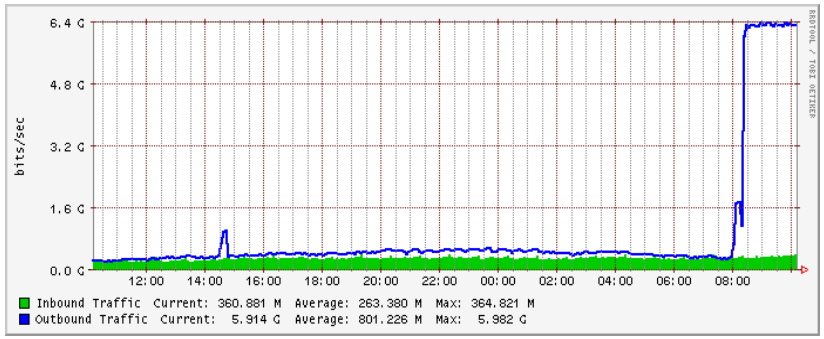
Disk servers (5-6 TBytes)

Linux Farm

On September 15, 2003, the DataTAG project was the first transatlantic testbed offering direct 10GigE access using Juniper's VPN layer2/10GigE emulation.

Internet2 land speed record history (IPv4 & IPv6) period 2000-2003





Impact of a single multi-Gb/s flow on the Abilene backbone

Internet2 land speed record history (IPv4 & IPv6) period 2000-2004