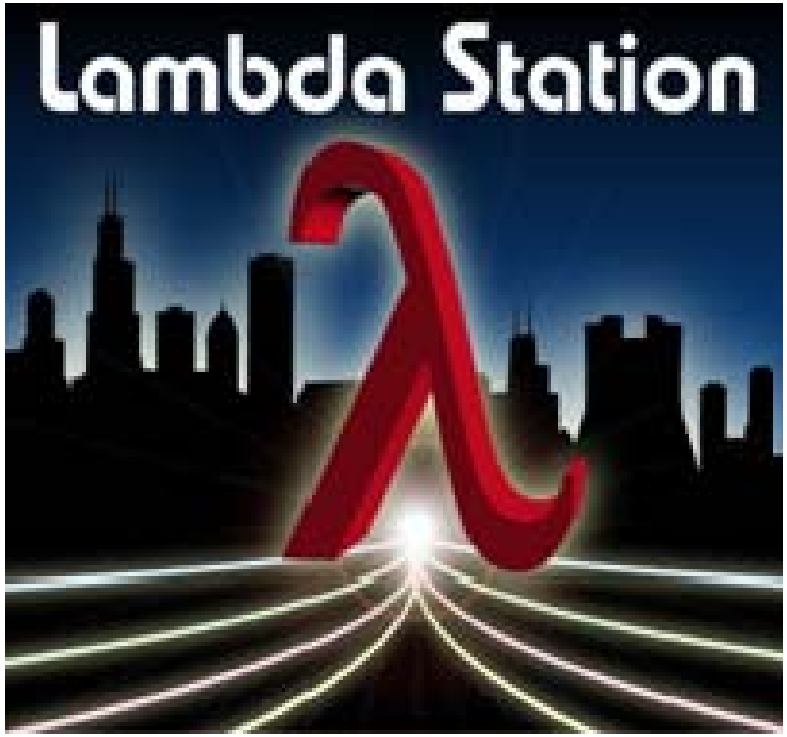
LambdaStation



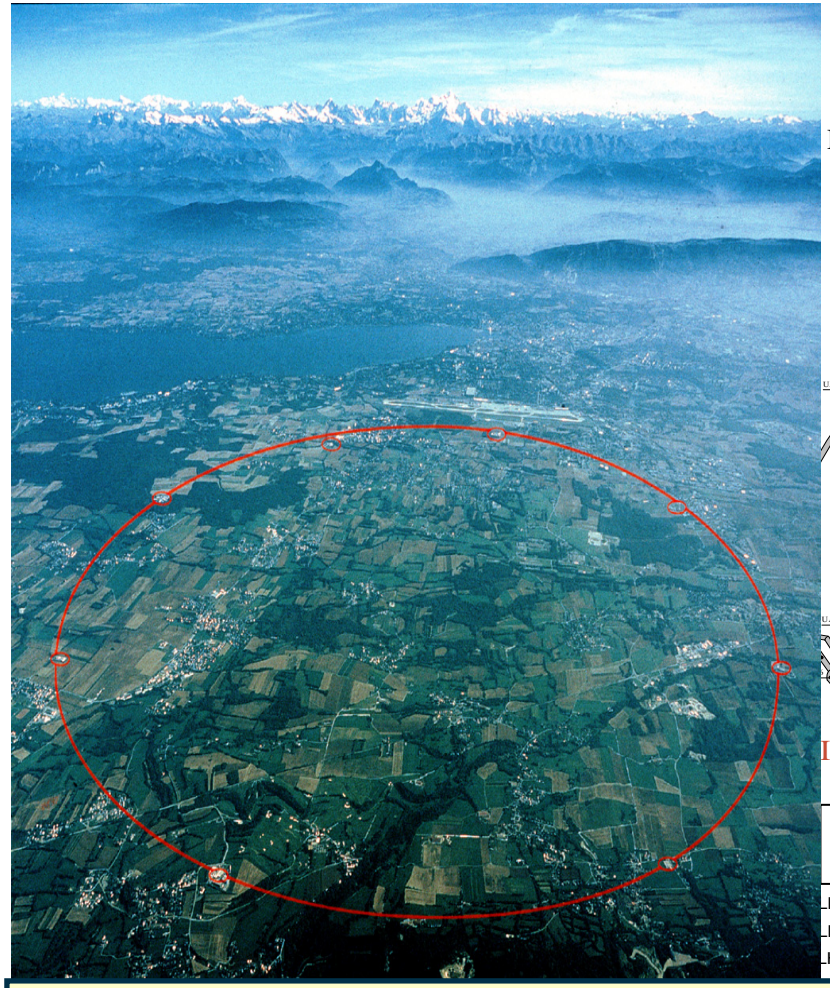
DoE PI meeting
September 30, 2005

Sylvain Ravot

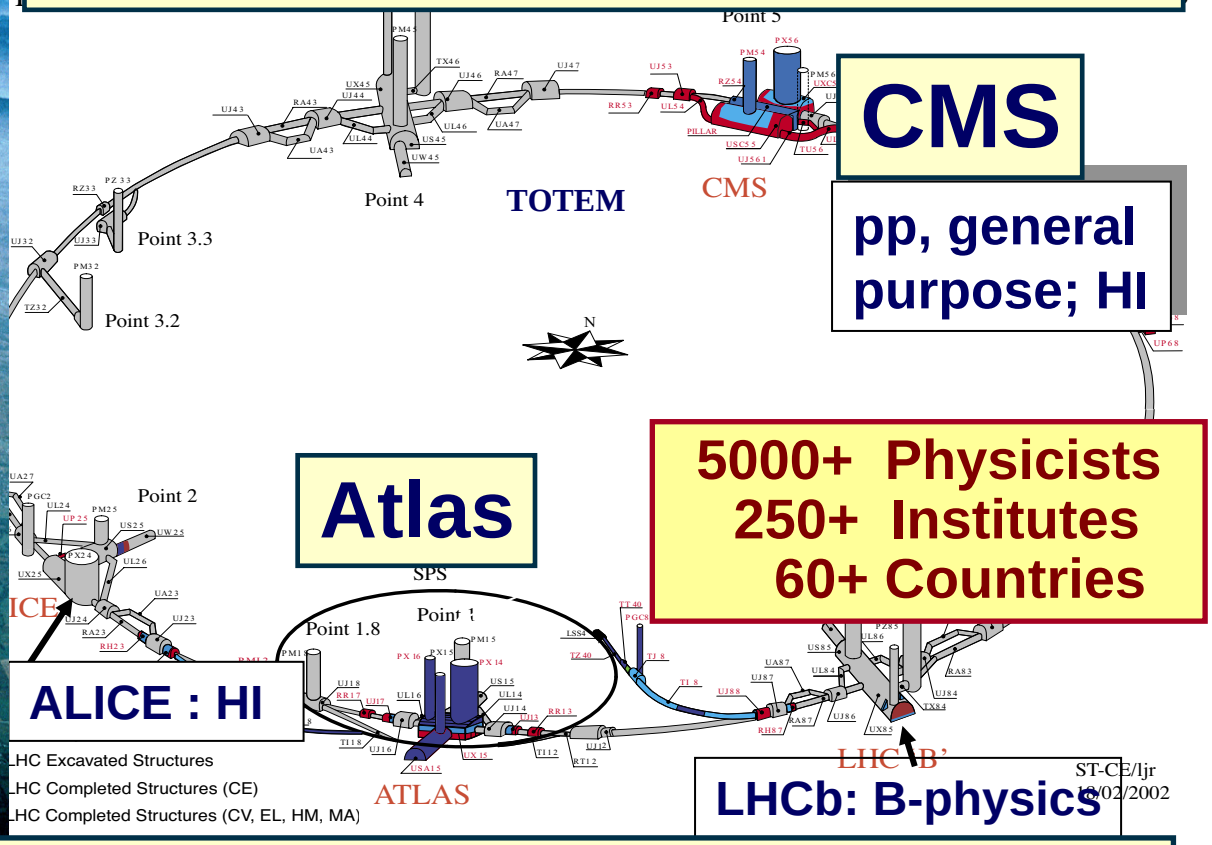
Agenda

- Motivation
- Building a wide-area testbed infrastructure
- Monalisa
- Data transport performance
- Next Generation LHCNet

Large Hadron Collider CERN, Geneva: 2007 Start



- pp $\sqrt{s}$ =14 TeV L=10^{34} cm^{-2} s^{-1}
- 27 km Tunnel in Switzerland & France



CMS

pp, general purpose; HI

Atlas

5000+ Physicists
250+ Institutes
60+ Countries

ALICE : HI

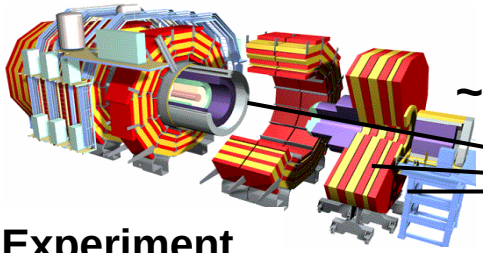
LHCb: B-physics

ST-CE/ljr
15/02/2002

Higgs, SUSY, Extra Dimensions, CP Violation, QG Plasma, …

the Unexpected

LHC Data Grid Hierarchy: Developed at Caltech



Experiment

~PByte/sec

Online System

~100-1500 MBytes/sec

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

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

Tier 1

10 - 40 Gbps

IN2P3 Center

RAL Center

BNL Center

FNAL Center

...

~10 Gbps

Tier 2

Tier2 Center

Center

nter

Center

Center

Tier 3

~1-10 Gbps

Institute

itute

nstitute

Institute

Physics data cache

1 to 10 Gbps

Workstations

Tier 4

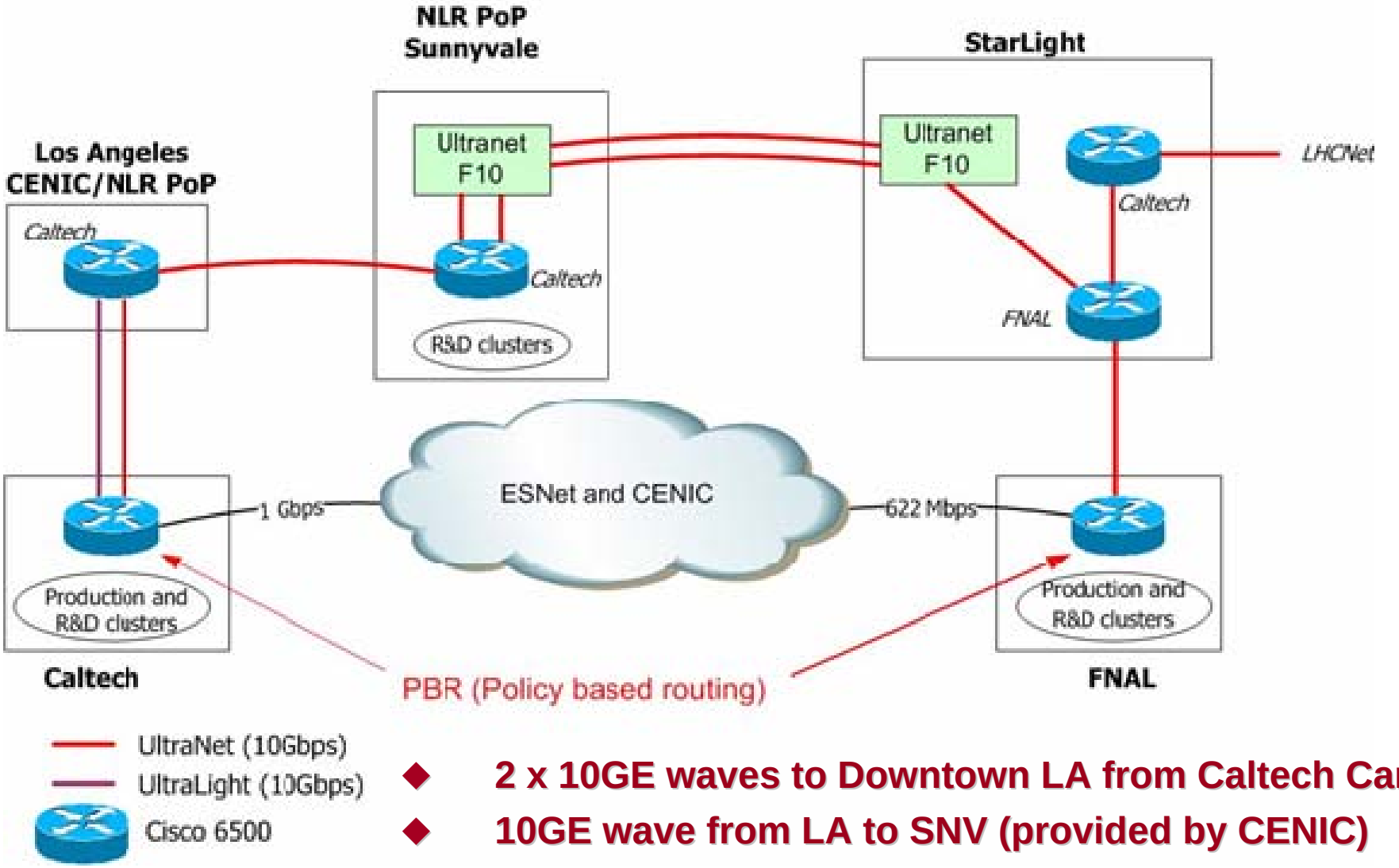
Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

Emerging Vision: A Richly Structured, Global Dynamic System

Wide-Area TestBed Infrastructure

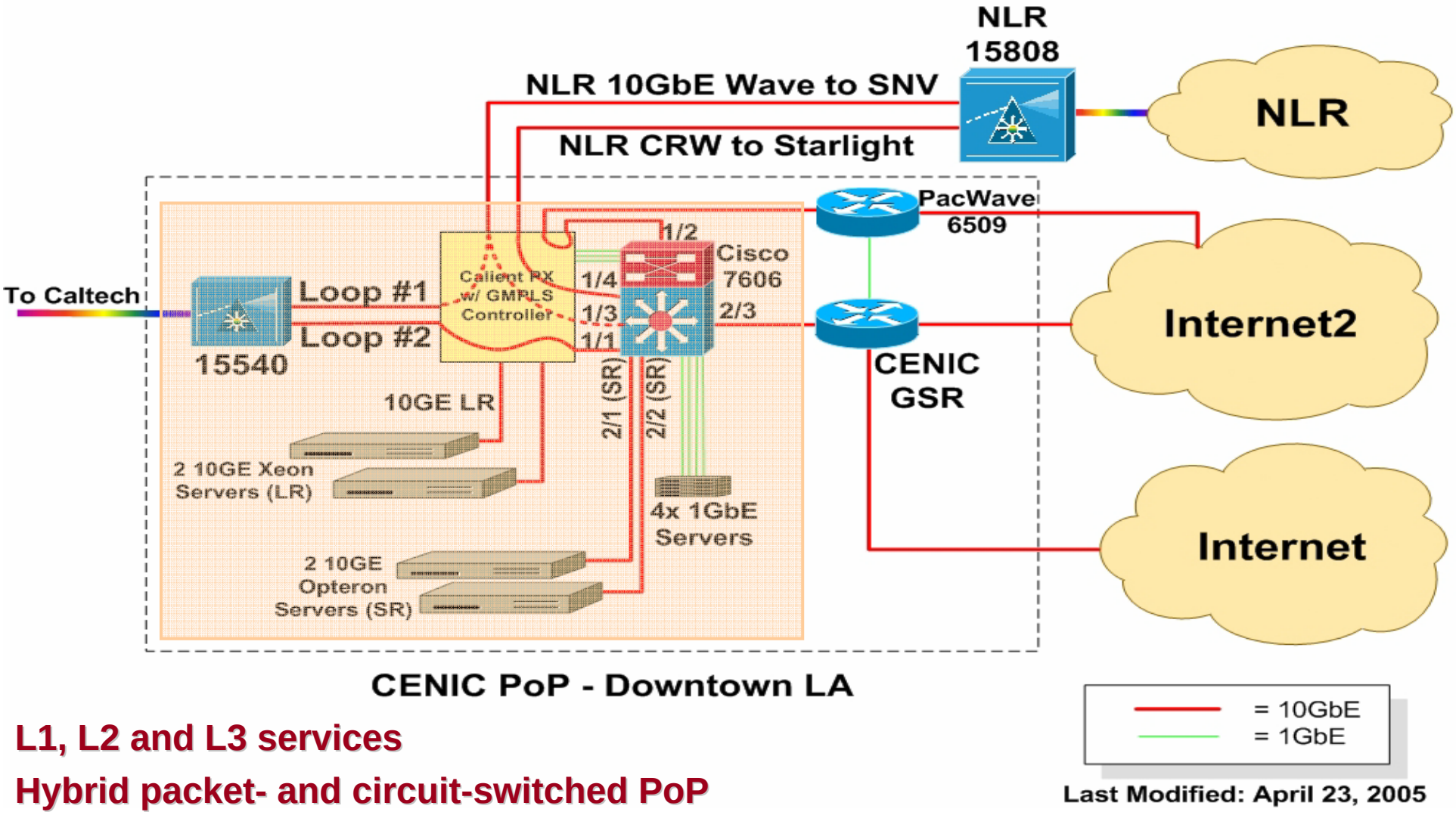
- Two sites: Fermilab and Caltech,
  - 10GBps connectivity via UltraScienceNet (USN)
  - Interconnecting hubs at StarLight/Chicago and Sunnyvale
- At Fermilab: 8 servers available for software development and bandwidth tests including 4 machines with 10Gb NICs (IntelPro/10GbE and Neterion).
- At Caltech: 2 machines equipped with 10Gb Neterion cards)
  - dual Opteron 250 2.4GHz; 4GB RAM
  - Neterion 10GbE card
  - 3 3ware 8500 controllers with 12 250GB SATA drives
- Each sites supports policy based routing to USN, jumbo frames and DSCP tagging.
- Californian sites support MPLS

FNAL-Caltech testbed



- 2 x 10GE waves to Downtown LA from Caltech Campus
- 10GE wave from LA to SNV (provided by CENIC)
- Cisco 6509 at Caltech, LA and SNV
- Extension of the testbed to CERN (Fall 2005)

Caltech 10GbE OXC



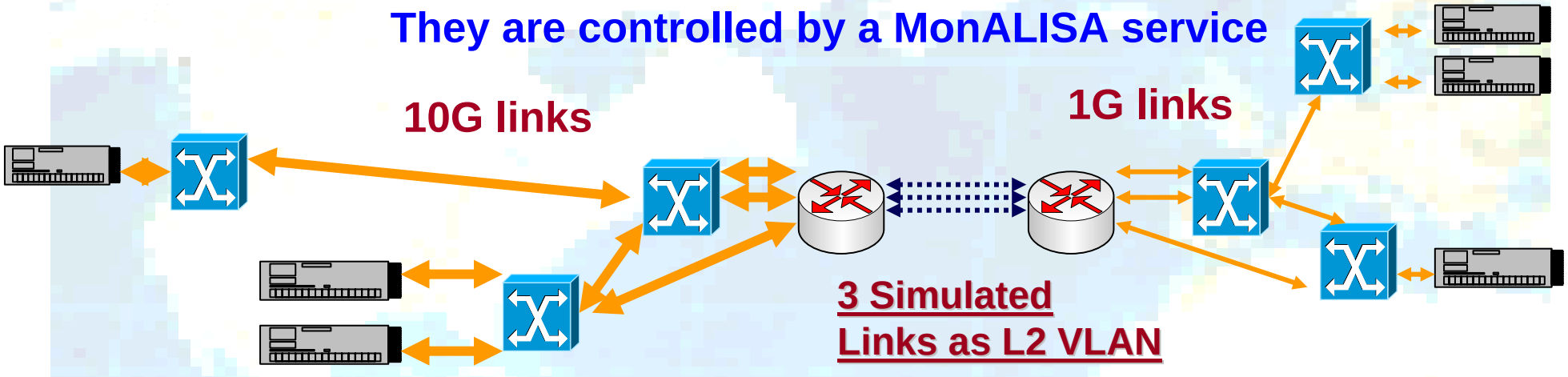
- L1, L2 and L3 services
- Hybrid packet- and circuit-switched PoP
- Photonic Switch
- Control plane is L3 (Monalisa)

Test Setup for Controlling Optical Switches

CALIENT (LA)

Glimmerglass (GE)

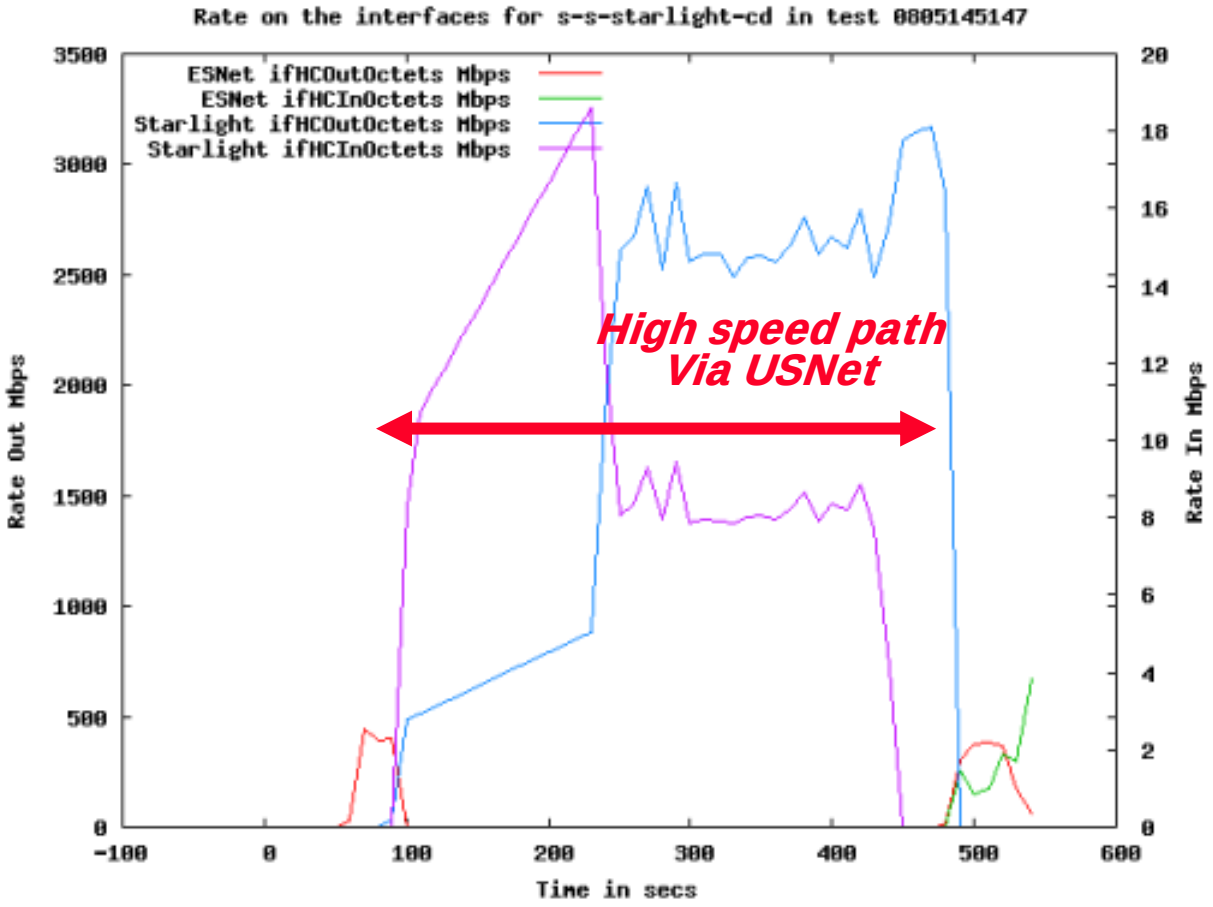
3 partitions on each switch
They are controlled by a MonALISA service



- Monitor and control switches using TL1
- Interoperability between the two systems
- End User access to service
- GMPLS

Can easily be adapted to GFP-based products!!!

Data transport performance

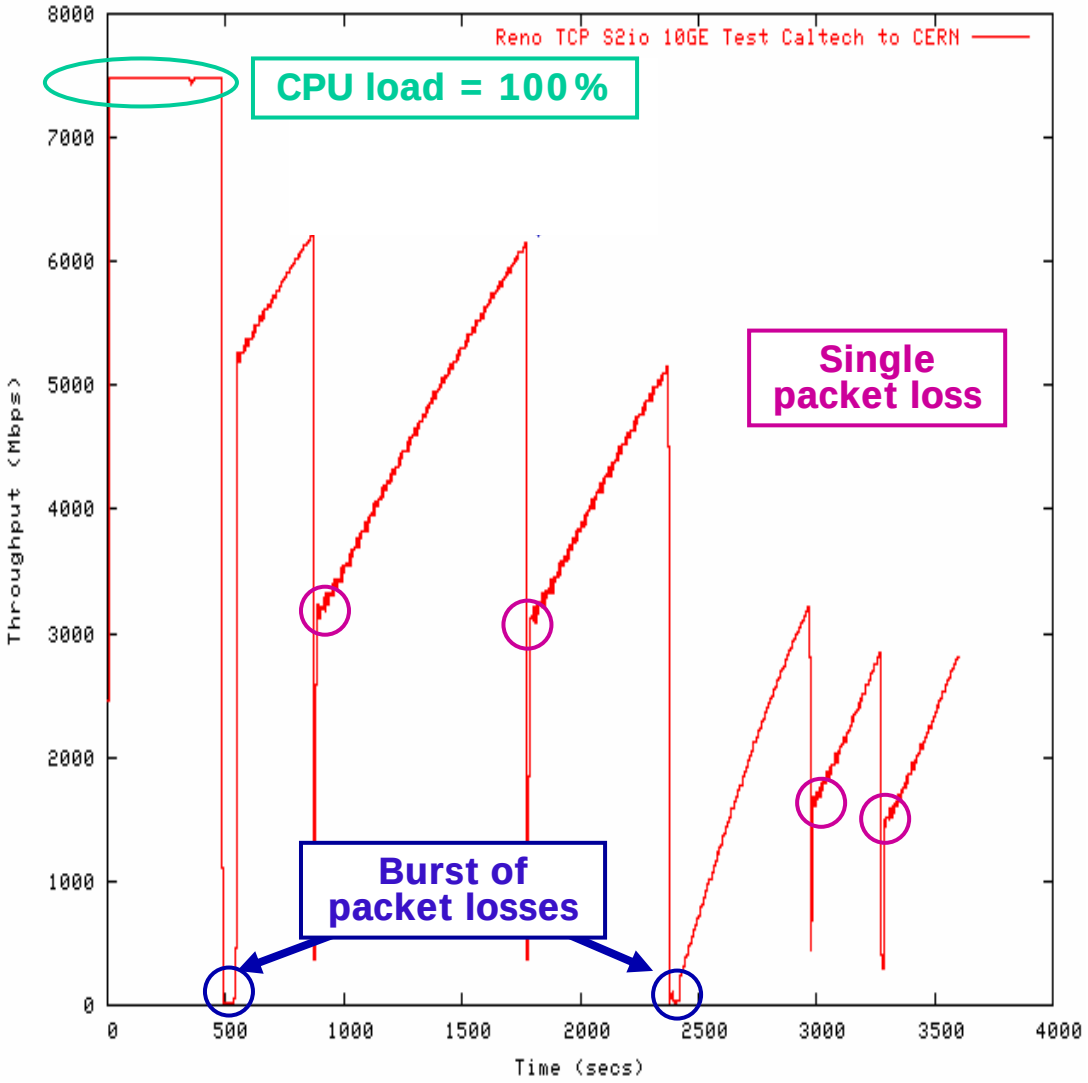


- Five parallel TCP streams
  - Data: Orange and light blue
  - Ack: Green and purple

Sophisticated systems to provision bandwidth

….. but TCP performance issues (AIMD, MTU …)

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:
  - Neterion 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

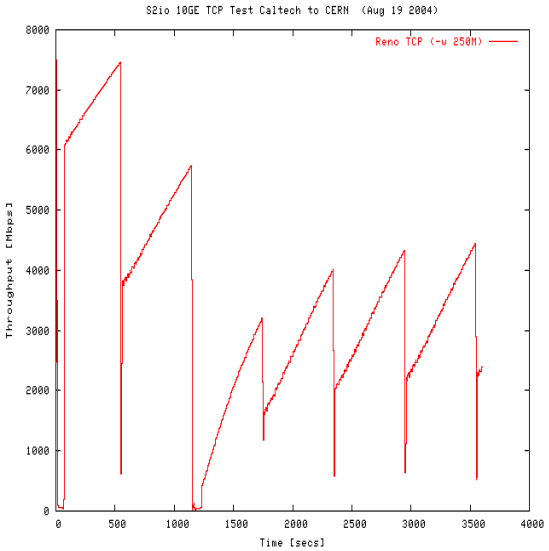
Path	Bandwidth	RTT (ms)	MTU (Byte)	Time to recover
LAN	10 Gb/s	1	1500	430 ms
Geneva–Chicago	10 Gb/s	120	1500	1 hr 32 min
Geneva-Los Angeles	1 Gb/s	180	1500	23 min
Geneva-Los Angeles	10 Gb/s	180	1500	3 hr 51 min
Geneva-Los Angeles	10 Gb/s	180	9000	38 min
Geneva-Los Angeles	10 Gb/s	180	64k (TSO)	5 min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

TCP Variants Performance

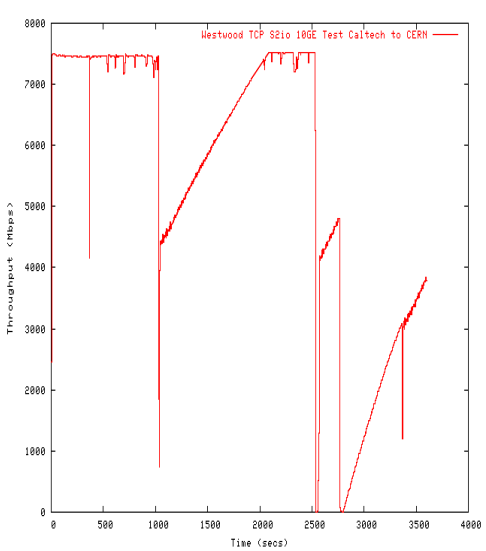
- Tests between CERN and Caltech
- Capacity = OC-192 9.5Gbps; 264 ms round trip latency; 1 flow
- Sending station: Tyan S2882 motherboard, 2x Opteron 2.4 GHz , 2 GB DDR.
- Receiving station (CERN OpenLab): HP rx4640, 4x 1.5GHz Itanium-2, zx1 chipset, 8GB memory
- Network adapter: Neterion 10 GE NIC

3.0 Gbps



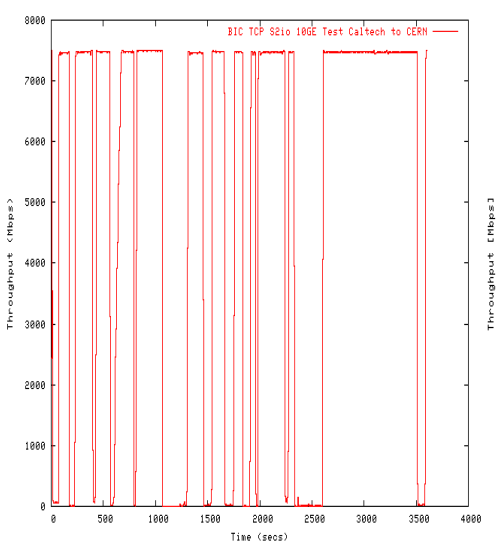
Linux TCP

4.1 Gbps



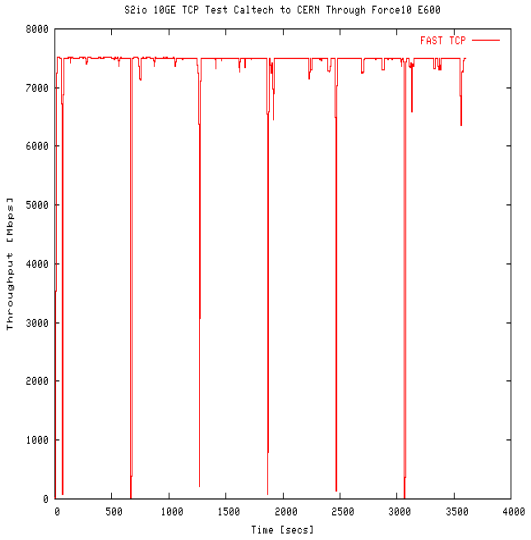
Linux Westwood+

5.0 Gbps



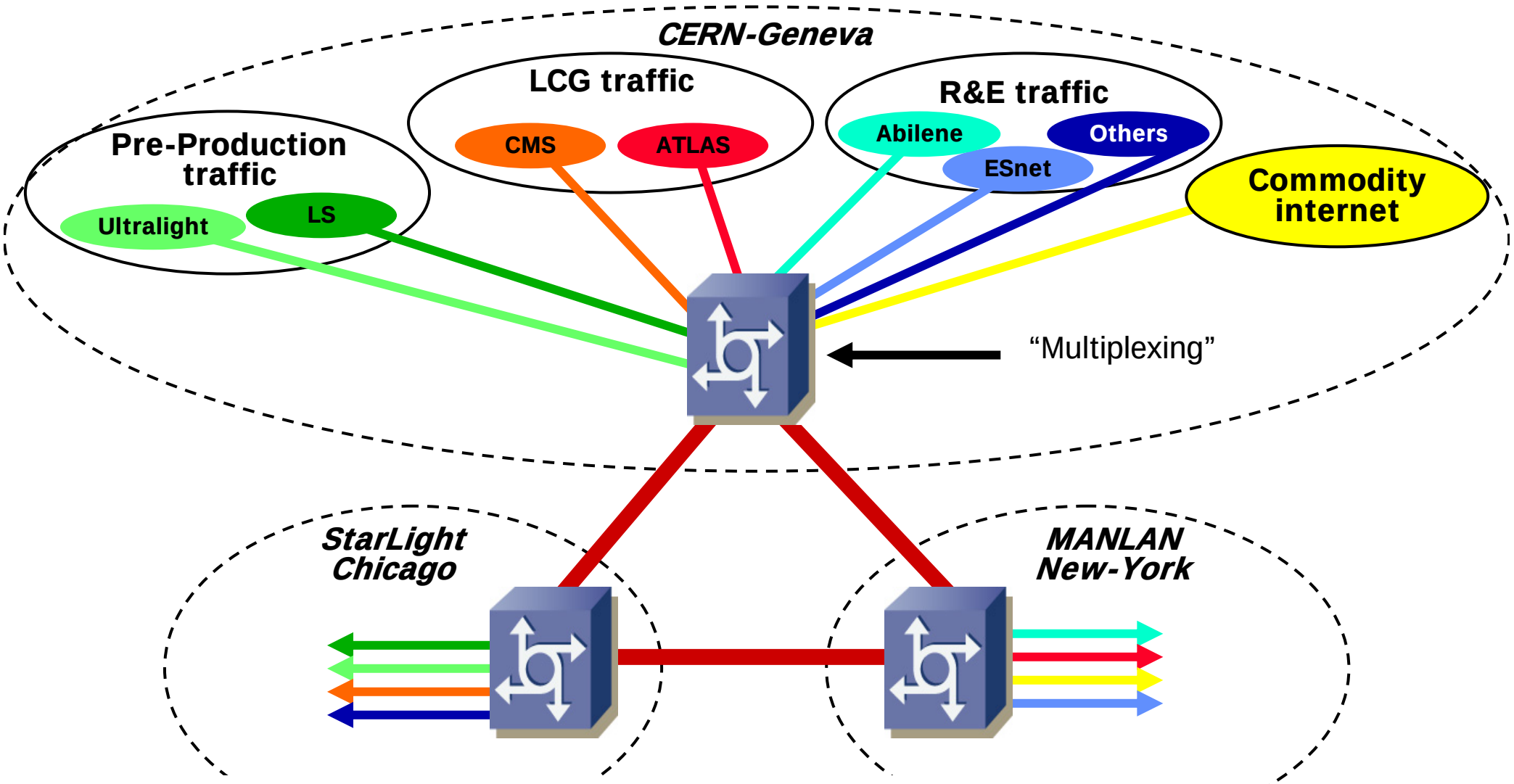
Linux BIC TCP

7.3 Gbps



FAST TCP

LHCNet Design

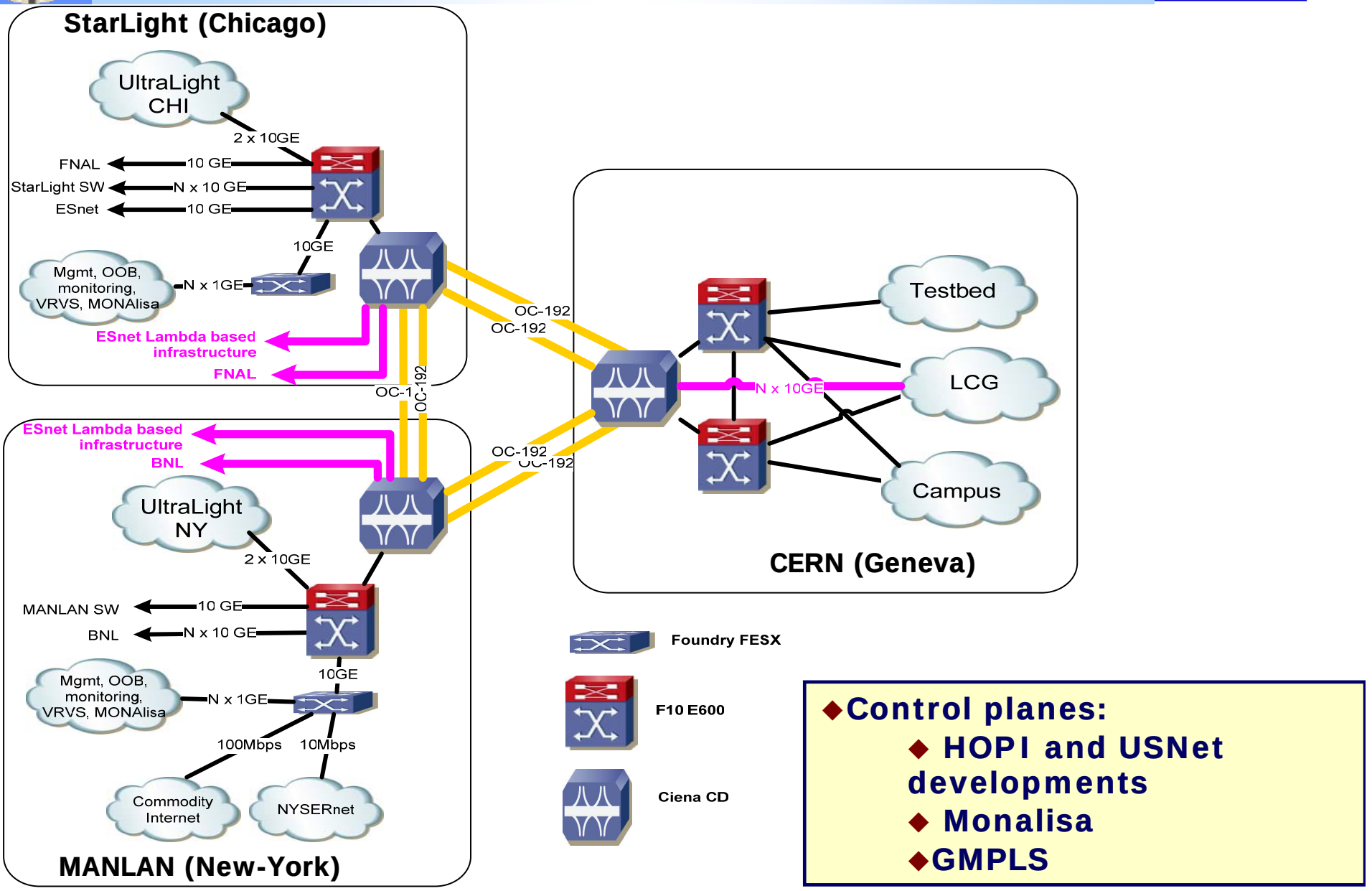


- October 2005: L2 switching with VLAN; One tag per type of traffic
- October 2006: Circuit switching (GFP-based products)

New Technology Candidates: Opportunities and Issues

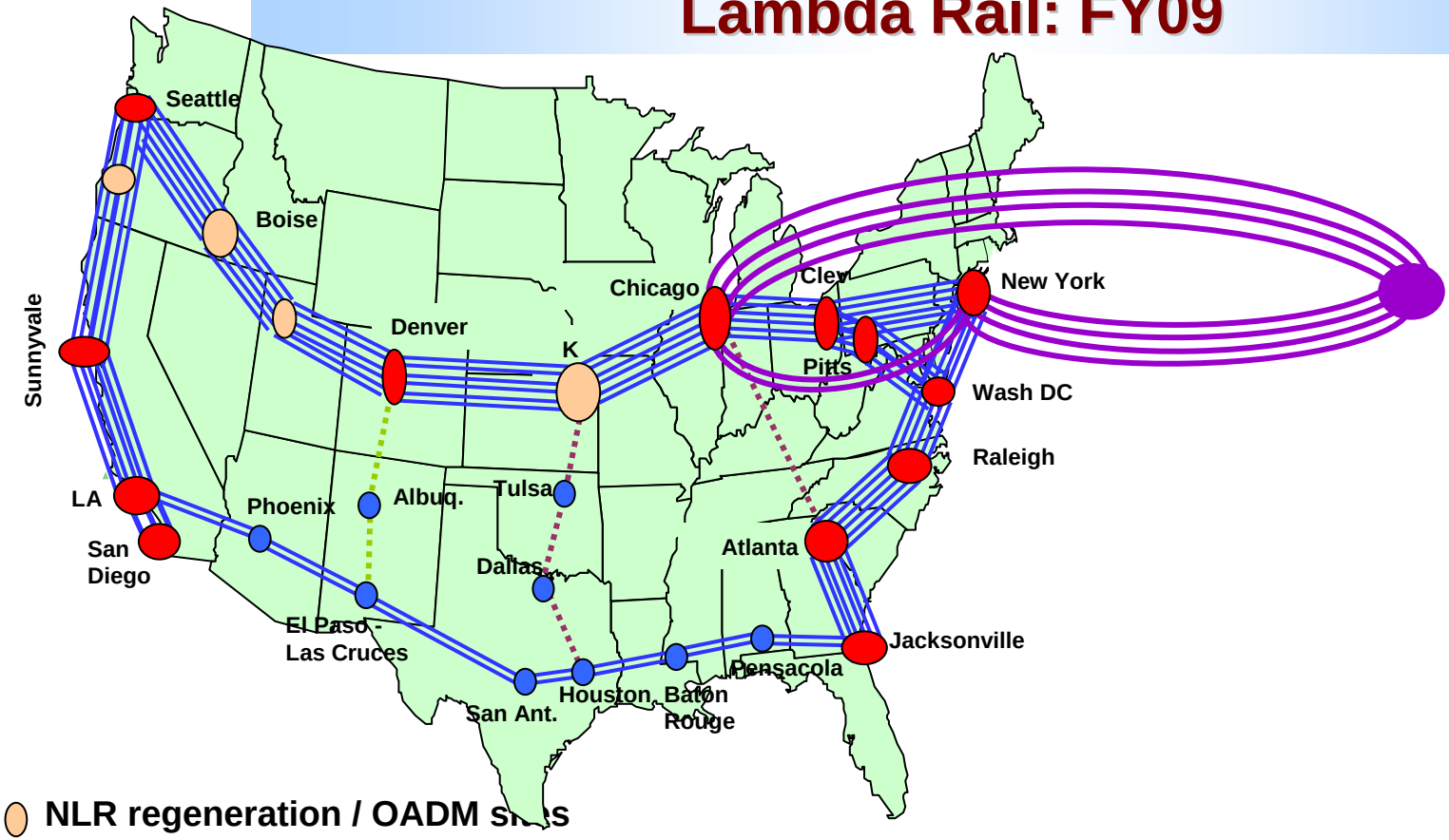
- **New standard for SONET infrastructures**
  - Alternative to the expensive Packet-Over-Sonet (POS) technology currently used
  - May change significantly the way in which we use SONET infrastructures
- **10 GE WAN-PHY standard**
  - Ethernet frames across OC-192 SONET networks
  - Ethernet as inexpensive linking technology between LANs & WANs
  - Supported by only a few vendors
- **GFP/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
- **NEW standards and NEW hardware**
  - Intensive evaluation and validation period
  - WAN-PHY tests in July 2005

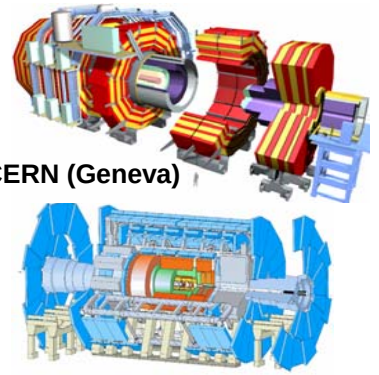
2006-2007 (Circuit Oriented Services)



- **Control planes:**
  - **HOPI and USNet developments**
  - **Monalisa**
  - **GMPLS**

LHCNet connection to Proposed ESnet Lambda Infrastructure Based on National Lambda Rail: FY09





NLR regeneration / OADM sites

NLR wavegear sites

ESnet via NLR (10 Gbps waves)

LHCNet (10 Gbps waves)

- LHCNet: To ~80 Gbps by 2009
- Routing + Dynamic managed circuit provisioning

Summary and conclusion

- The WAN infrastructure is operational
- Continue research of flow-based switching and its effect on applications, performance, disk-to-disk and storage-to-storage tests
- Evaluation of new data transport protocols
- US LHCnet: Circuit oriented services by 2006/2007 (GFP-based products)