HENP Grids and Networks
Global Virtual Organizations

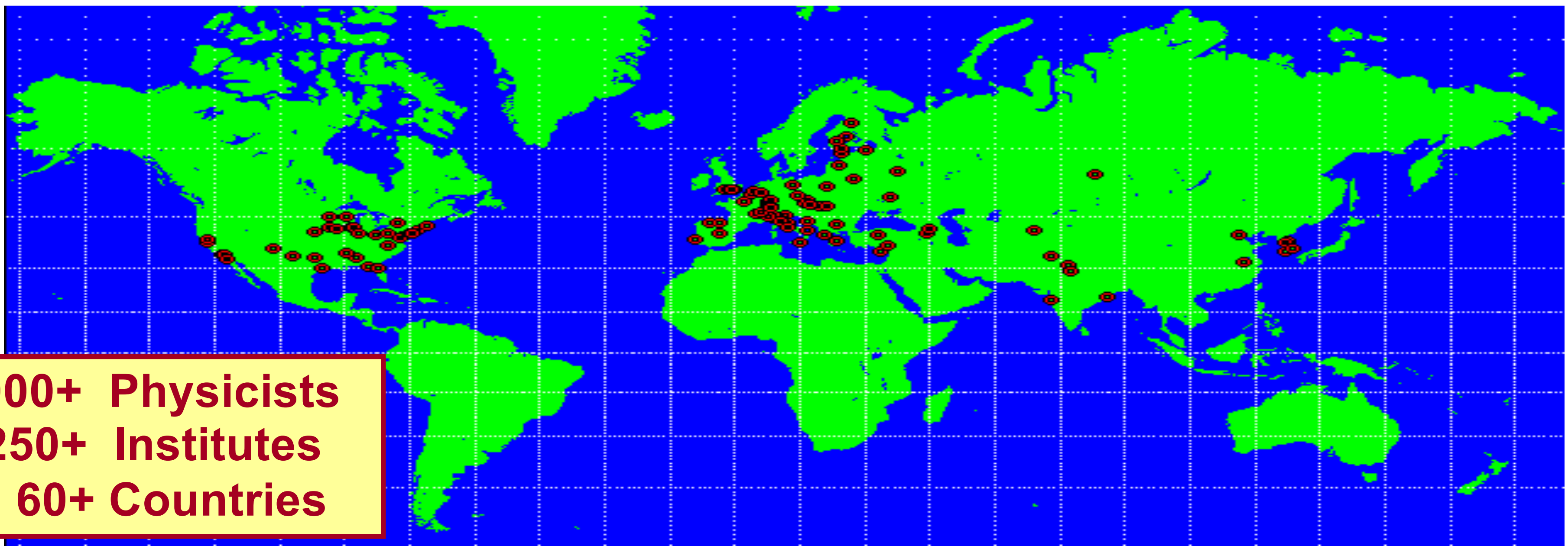
Harvey B. Newman, Caltech

First AMPATH International Conference
Valdivia, Chile
April 12, 2002

http://l3www.cern.ch/~newman/HENPGridsNets_AMPATHChile041202.ppt

Computing Challenges: Petabyes, Petaflops, Global VOs

- **Geographical dispersion:** of people and resources
- **Complexity:** the detector and the LHC environment
- **Scale:** Tens of Petabytes per year of data



5000+ Physicists
250+ Institutes
60+ Countries

Major challenges associated with:

Communication and collaboration at a distance
Managing globally distributed computing & data resources
Remote software development and physics analysis
R&D: New Forms of Distributed Systems: Data Grids

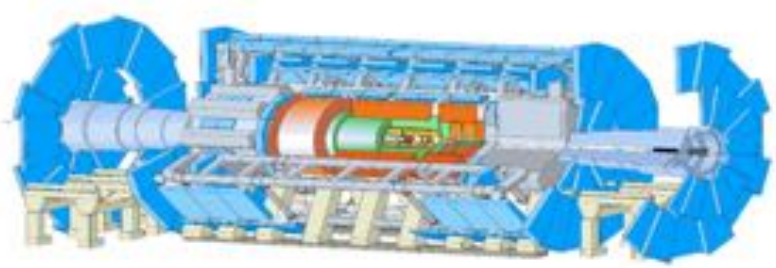
Next Generation Networks for Experiments: Goals and Needs

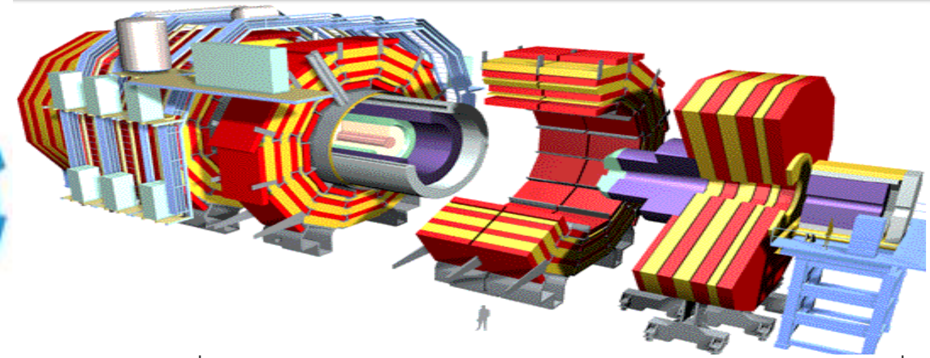
- Providing rapid access to event samples and subsets from massive data stores
  - From ~400 Terabytes in 2001, ~Petabytes by 2002, ~100 Petabytes by 2007, to ~1 Exabyte by ~2012.
- Providing analyzed results with rapid turnaround, by coordinating and managing the *LIMITED* computing, data handling and *NETWORK* resources effectively
- Enabling rapid access to the data and the collaboration
  - Across an ensemble of networks of varying capability
- Advanced integrated applications, such as Data Grids, rely on seamless operation of our LANs and WANs
  - With reliable, quantifiable (monitored), high performance
  - For “Grid-enabled” event processing and data analysis, and collaboration

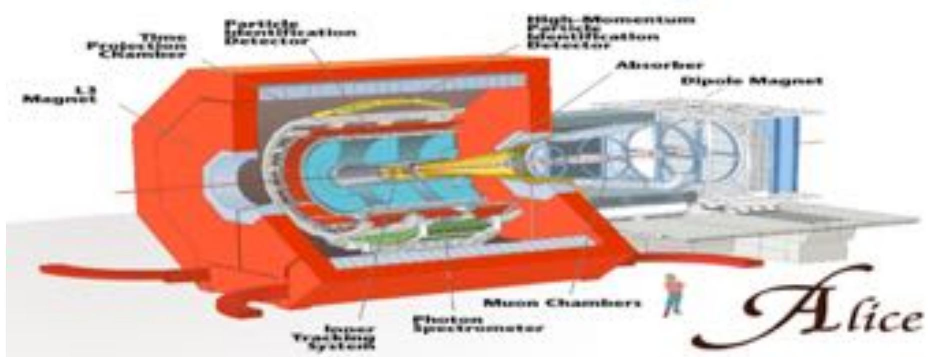
Four LHC Experiments: The Petabyte to Exabyte Challenge

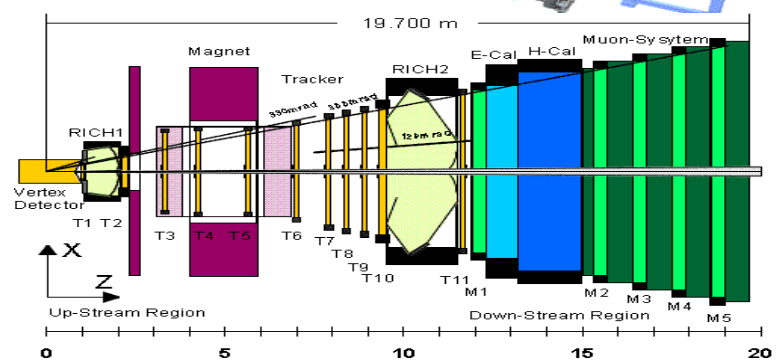
ATLAS, CMS, ALICE, LHCB

Higgs + New particles; Quark-Gluon Plasma; CP Violation





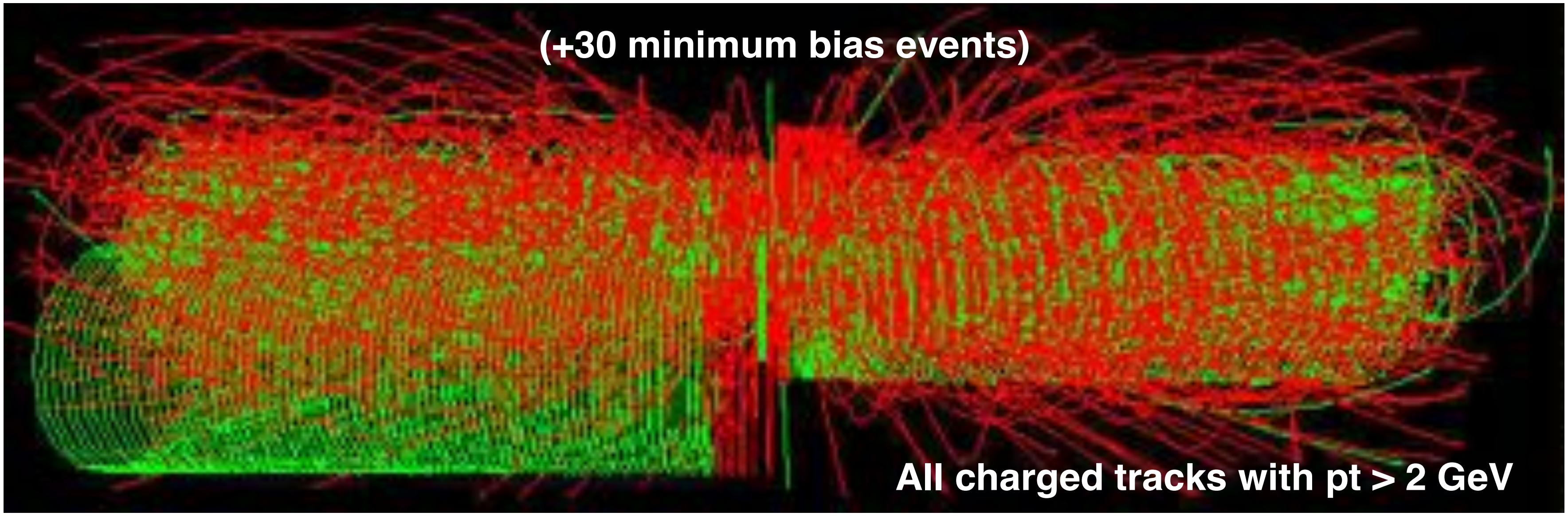


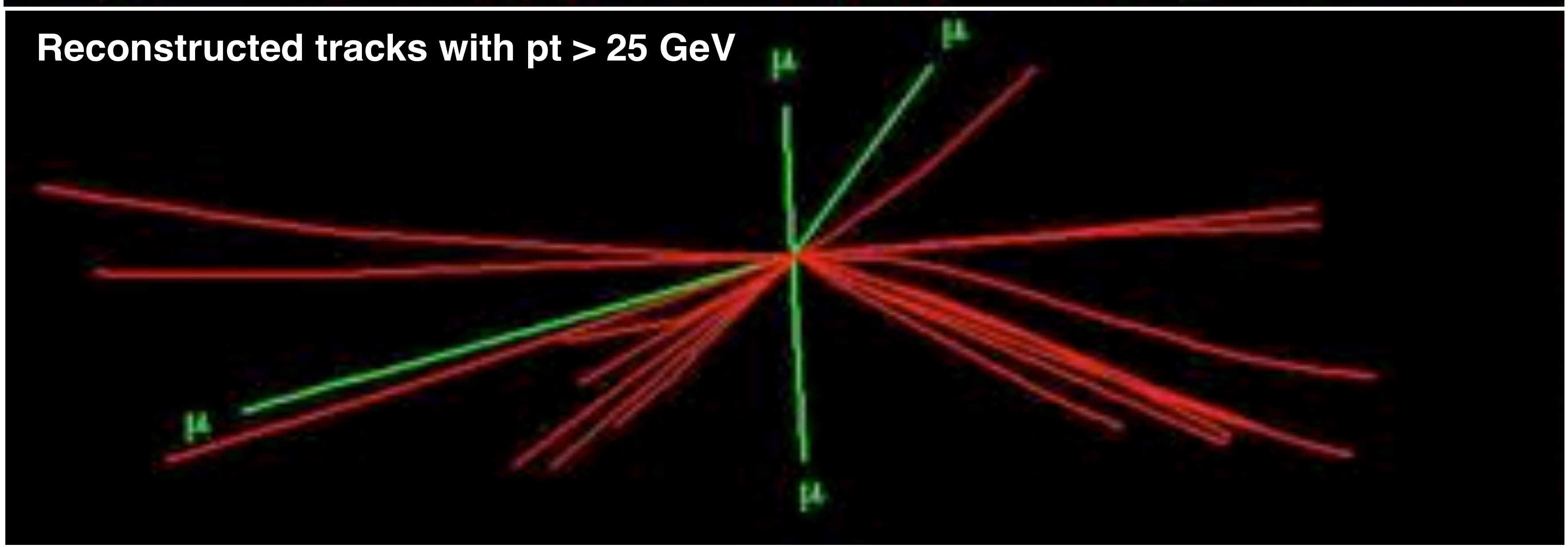


Data stored ~40 Petabytes/Year and UP;
CPU 0.30 Petaflops and UP

0.1 to 1 Exabyte (1 EB = 10^{18} Bytes)
(2007) (~2012 ?) for the LHC Experiments

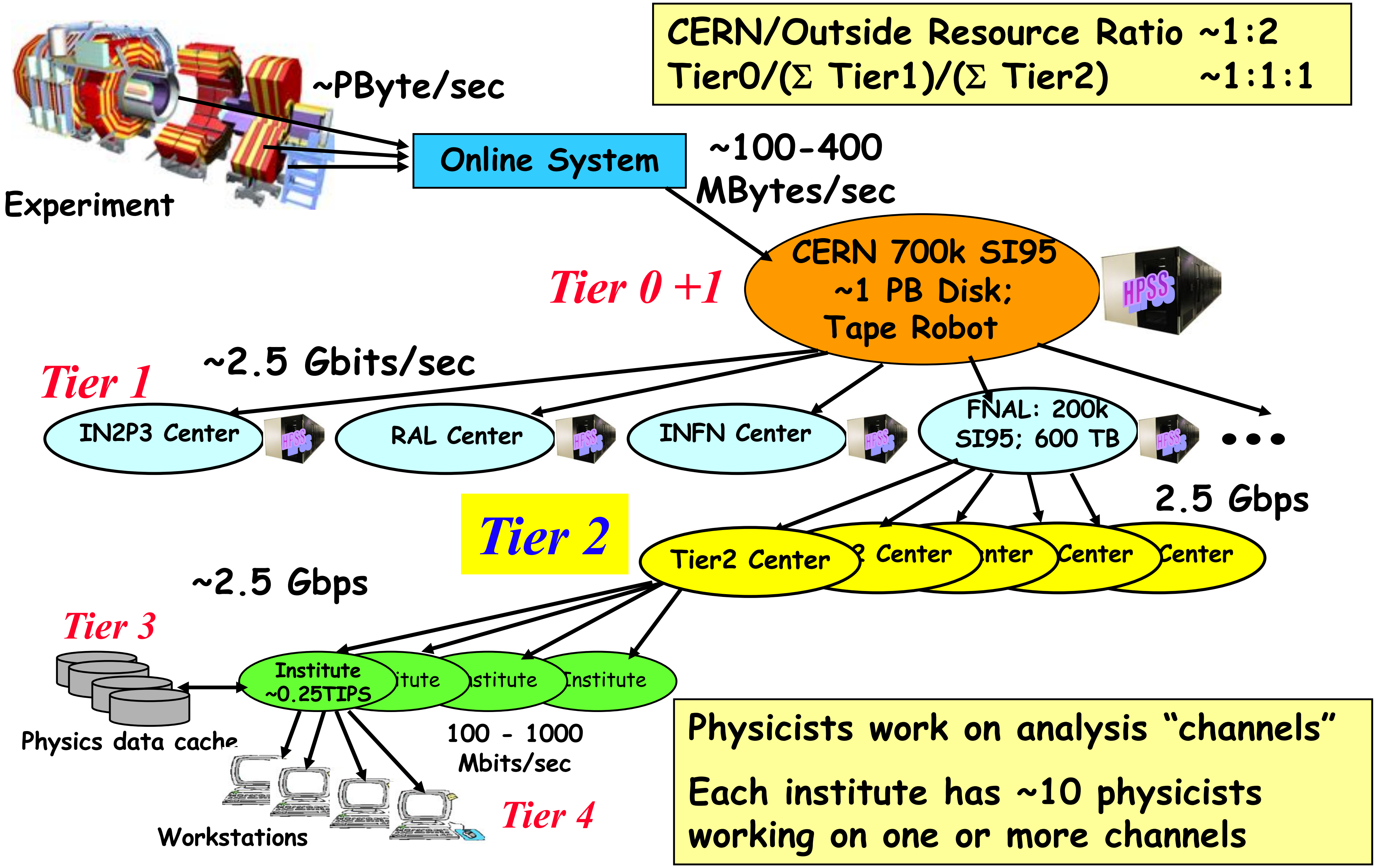
LHC: Higgs Decay into 4 muons (Tracker only); 1000X LEP Data Rate





10^9 events/sec, selectivity: 1 in 10^{13} (1 person in a thousand world populations)

LHC Data Grid Hierarchy



Transatlantic Net WG (HN, L. Price) Bandwidth Requirements [*]

	2001	2002	2003	2004	2005	2006
CMS	100	200	300	600	800	2500
ATLAS	50	100	300	600	800	2500
BaBar	300	600	1100	1600	2300	3000
CDF	100	300	400	2000	3000	6000
D0	400	1600	2400	3200	6400	8000
BTeV	20	40	100	200	300	500
DESY	100	180	210	240	270	300
CERN BW	155-310	622	1250	2500	5000	10000

[*] Installed BW. Maximum Link Occupancy 50% Assumed

See http://gate.hep.anl.gov/lprice/TAN

HENP Related Data Grid Projects

Projects

PPDG I	USA	DOE	$2M	1999-2001
GriPhyN	USA	NSF	$11.9M + $1.6M	2000-2005
EU DataGrid	EU	EC	€10M	2001-2004
PPDG II (CP)	*USA*	*DOE*	*$9.5M*	*2001-2004*
iVDGL	*USA*	*NSF*	*$13.7M + $2M*	*2001-2006*
DataTAG	*EU*	*EC*	*€4M*	*2002-2004*
GridPP	*UK*	*PPARC*	*>$15M*	*2001-2004*
LCG (Ph1)	*CERN*	*MS*	*30 MCHF*	*2002-2004*

Many Other Projects of interest to HENP

- Initiatives in US, UK, Italy, France, NL, Germany, Japan, …
- US and EU networking initiatives: *AMPATH, I2, DataTAG*
- US Distributed Terascale Facility:
  ($53M, 12 TeraFlops, 40 Gb/s network)

CMS Production: Event Simulation and Reconstruction

	Simulation	Digitization		GDMP	Common Prod. tools (IMPALA)
		No PU	PU		
CERN				✓	✓
FNAL				✓	✓
Moscow				✓	In progress
INFN (9)				✓	✓
Caltech				✓	✓
UCSD				✓	✓
UFL				✓	✓
Imperial College				✓	✓
Bristol				✓	✓
Wisconsin				✓	✓
IN2P3				✓	✓
Helsinki				✓	✓

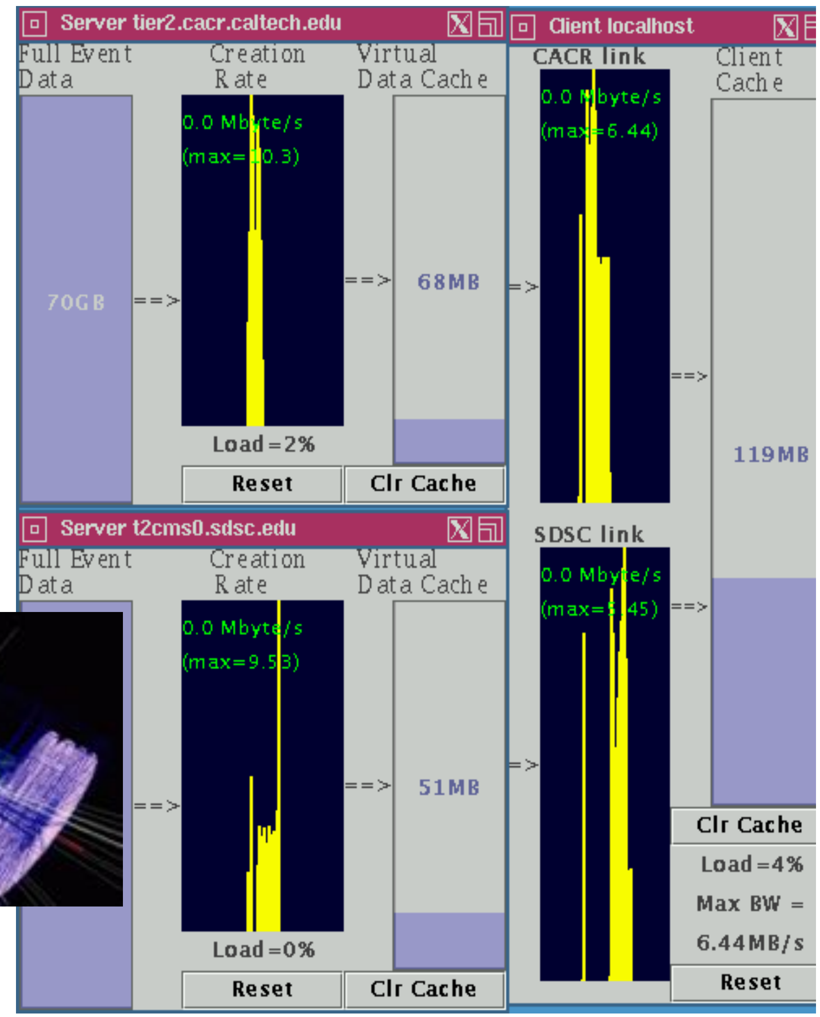
Fully operational

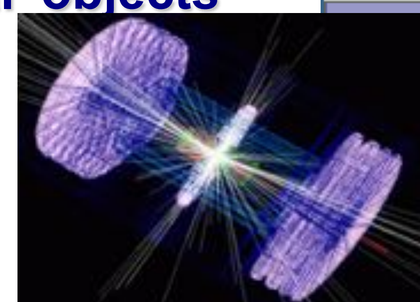
Worldwide Production at 20 Sites

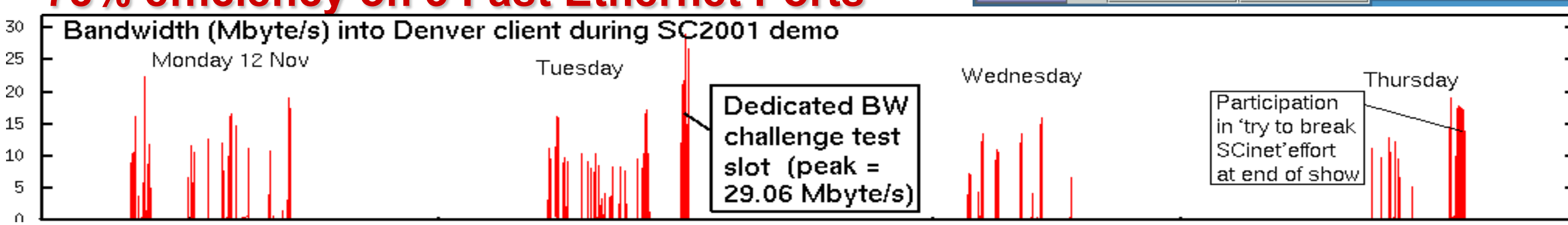
"Grid-Enabled" Automated

Grid-enabled Data Analysis: SC2001 Demo by K. Holtman, J. Bunn (CMS/Caltech)

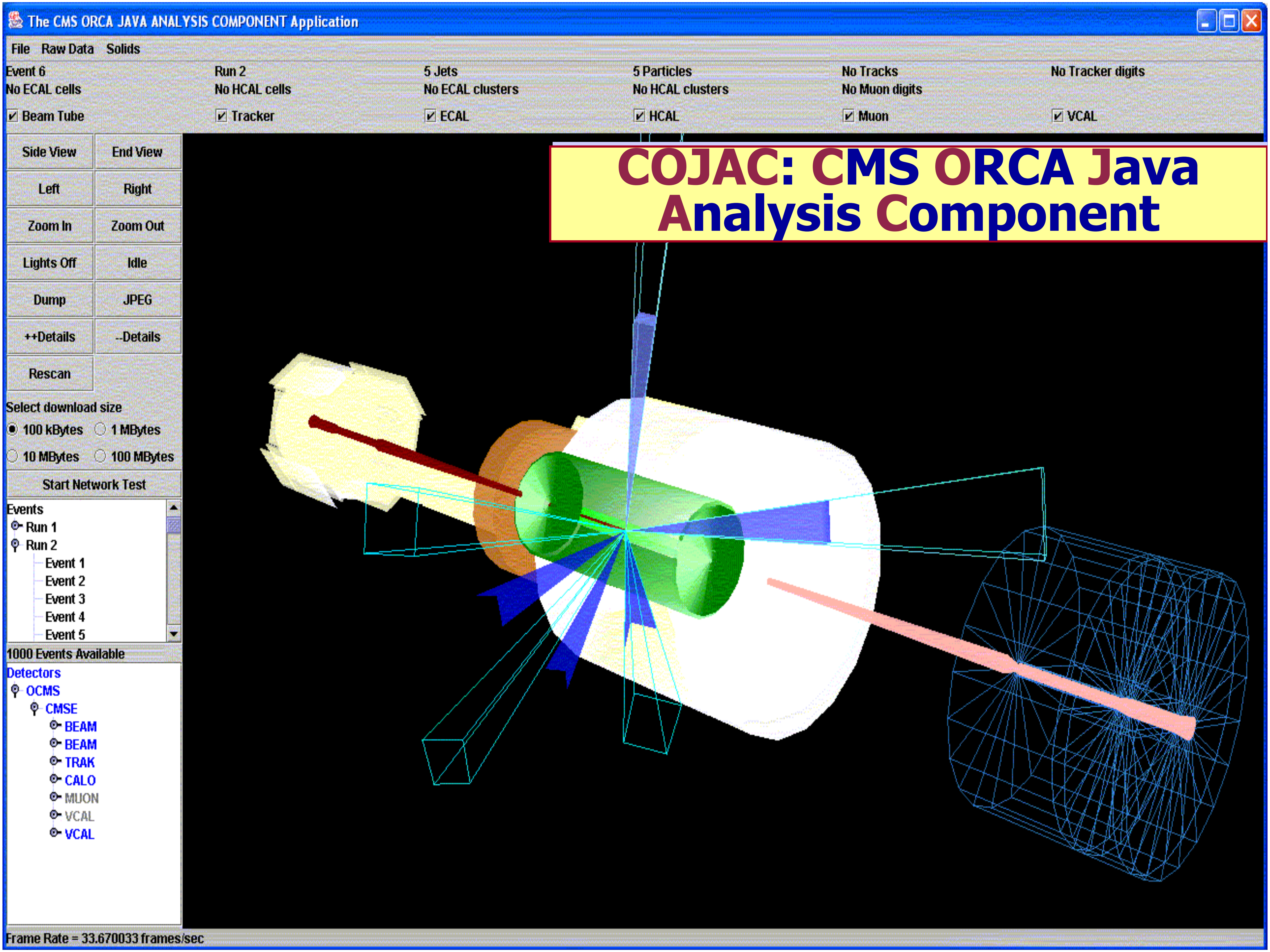
- **Demonstration of Virtual Data technology for interactive CMS physics analysis at Supercomputing 2001, Denver**
  - Interactive subsetting and analysis of 144,000 CMS QCD events (105 GB)
  - Tier 4 workstation (Denver) gets data from two tier 2 servers (Caltech and San Diego)
- **Prototype tool showing feasibility of these CMS computing model concepts:**
  - Navigates from tag data to full event data
  - Transparently accesses `virtual' objects through Grid-API
  - Reconstructs On-Demand *(=Virtual Data materialisation)*
  - Integrates object persistency layer and grid layer
- **Peak throughput achieved: 29.1 Mbyte/s; 78% efficiency on 3 Fast Ethernet Ports**







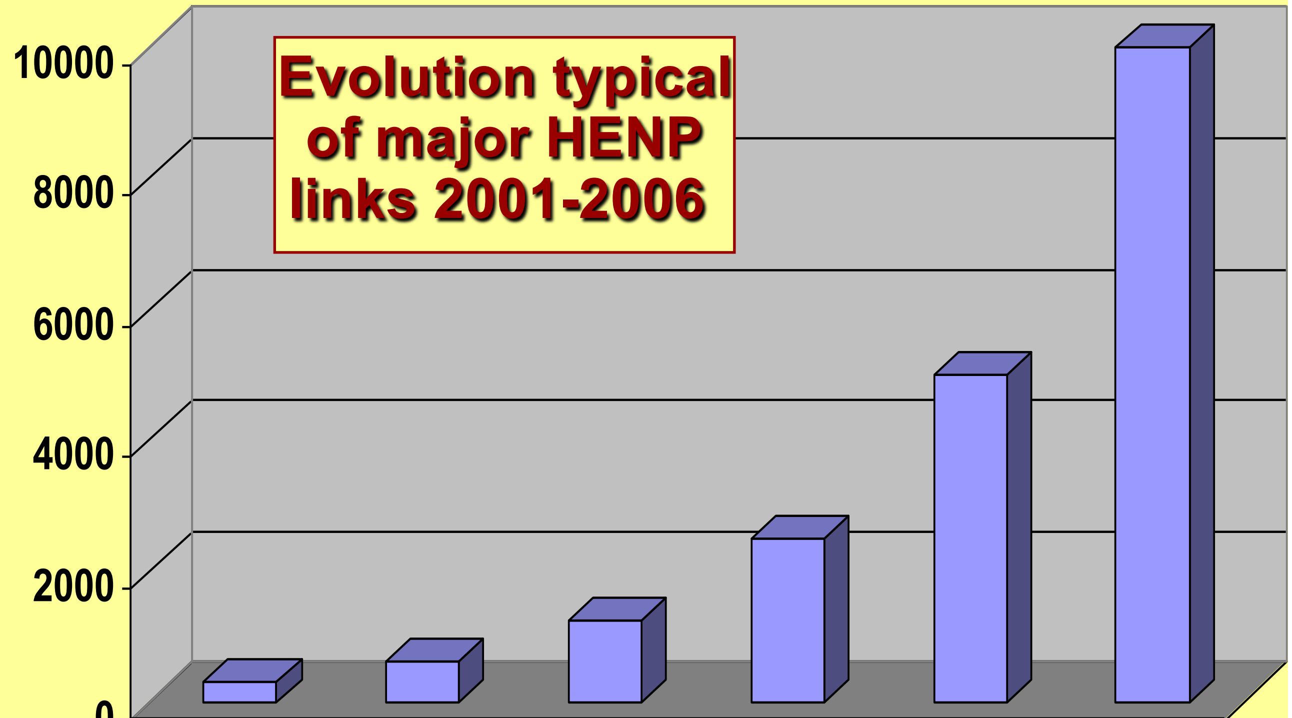
COJAC: CMS ORCA Java Analysis Component



Baseline BW for the US-CERN Link: HENP Transatlantic WG (DOE+NSF)

Transoceanic Networking
Integrated with the Abilene, TeraGrid, Regional Nets and Continental Network Infrastructures in US, Europe, Asia, South America

Link Bandwidth (Mbps)

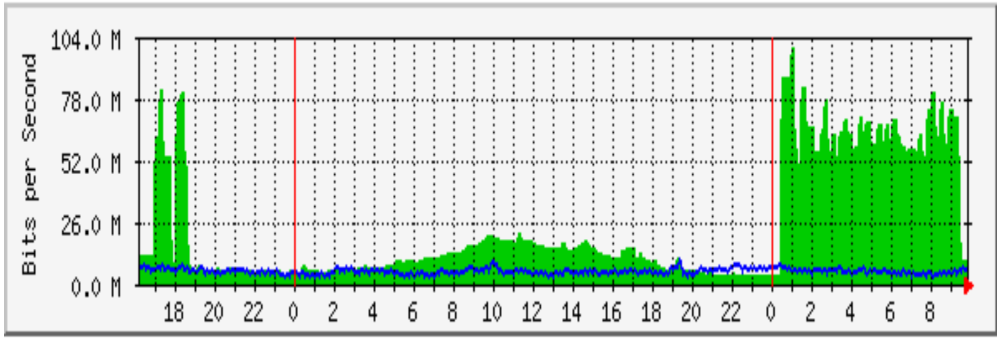


	FY2001	FY2002	FY2003	FY2004	FY2005	FY2006
BW (Mbps)	310	622	1250	2500	5000	10000

US-CERN Link: 2 X 155 Mbps Now;
Plans: 622 Mbps in April 2002;
DataTAG 2.5 Gbps Research Link in Summer 2002;
10 Gbps Research Link in ~2003 or Early 2004

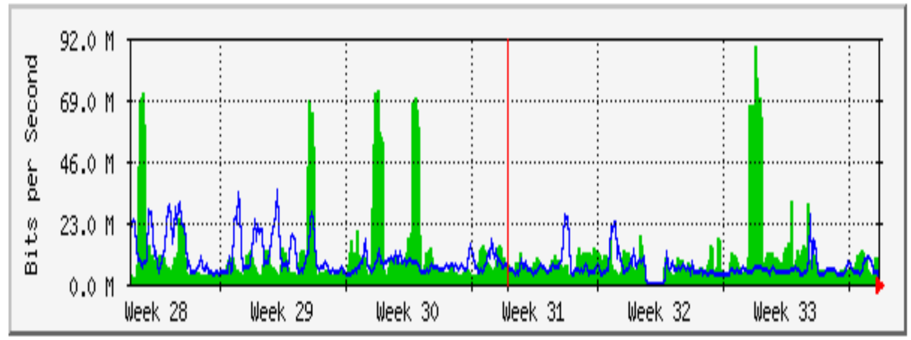
Daily, Weekly, Monthly and Yearly Statistics on 155 Mbps US-CERN Link

`Daily' Graph (5 Minute Average)



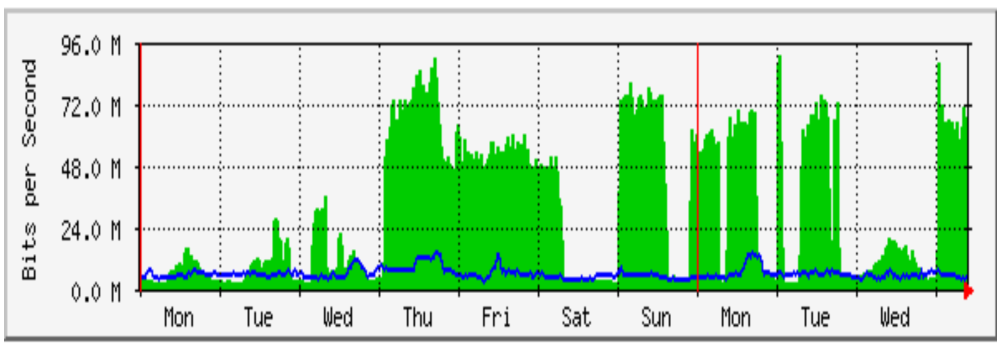
Max In: 100.1 Mb/s (64.6%) Average In: 24.2 Mb/s (15.6%) Current In: 11.2 Mb/s (7.2%)
Max Out: 10.9 Mb/s (7.0%) Average Out: 6008.5 kb/s (3.9%) Current Out: 6576.8 kb/s (4.2%)

`Monthly' Graph (2 Hour Average)



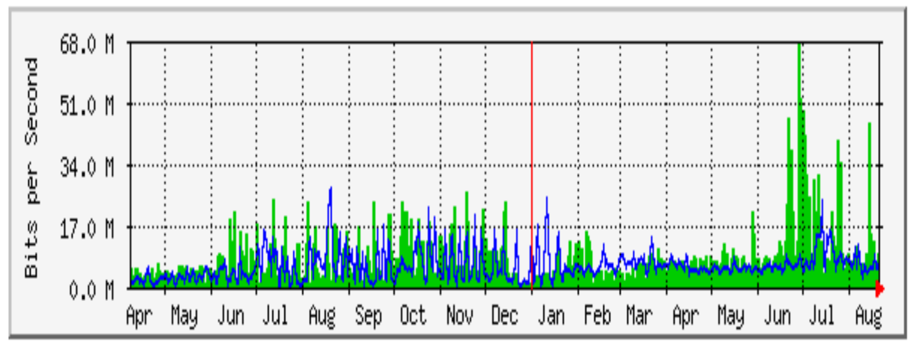
Max In: 89.4 Mb/s (57.7%) Average In: 11.6 Mb/s (7.5%) Current In: 10.8 Mb/s (7.0%)
Max Out: 35.4 Mb/s (22.8%) Average Out: 8581.8 kb/s (5.5%) Current Out: 5521.1 kb/s (3.6%)

`Weekly' Graph (30 Minute Average)



Max In: 92.1 Mb/s (59.4%) Average In: 31.5 Mb/s (20.3%) Current In: 60.1 Mb/s (38.8%)
Max Out: 15.3 Mb/s (9.9%) Average Out: 6679.1 kb/s (4.3%) Current Out: 6020.2 kb/s (3.9%)

`Yearly' Graph (1 Day Average)

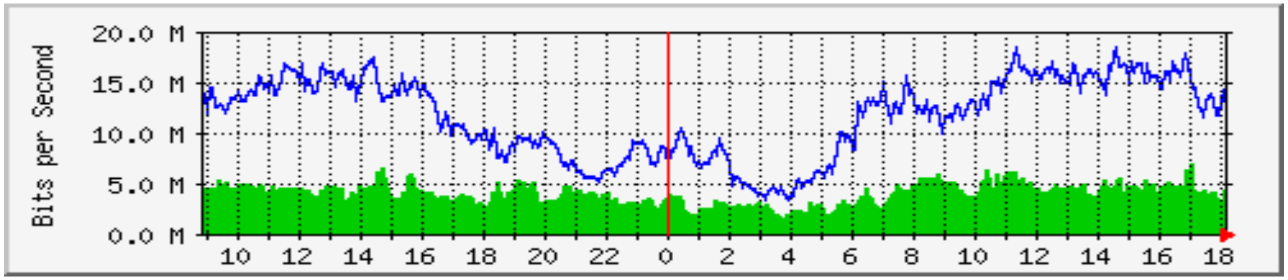


Max In: 67.8 Mb/s (43.8%) Average In: 8357.7 kb/s (5.4%) Current In: 9907.8 kb/s (6.4%)
Max Out: 27.7 Mb/s (17.9%) Average Out: 6205.2 kb/s (4.0%) Current Out: 6953.1 kb/s (4.5%)

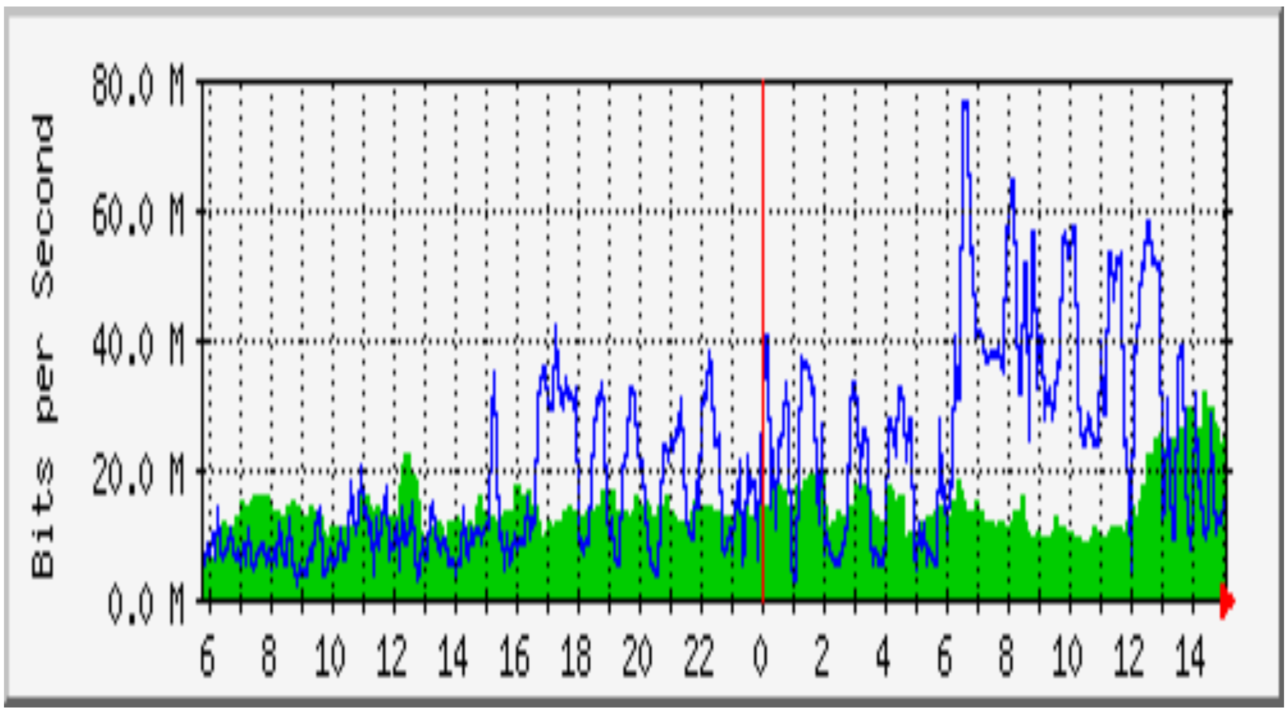
20 - 100 Mbps Used Routinely in '01
BaBar: 600Mbps Throughput in '02

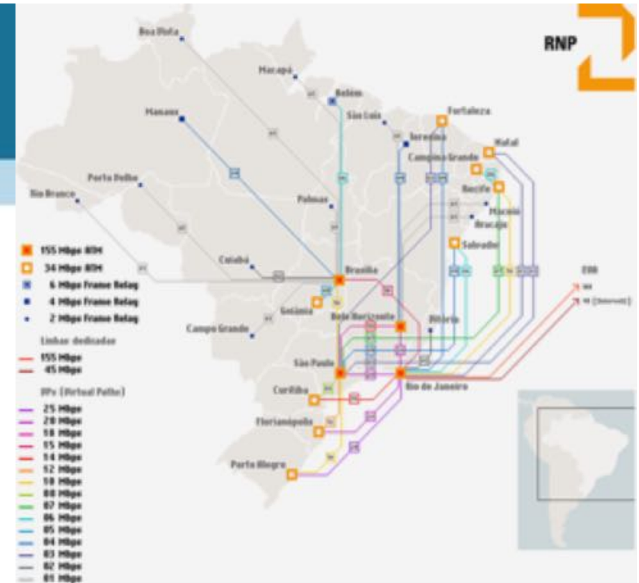
BW Upgrades Quickly Followed by Upgraded Production Use

RNP Brazil (to 20 Mbps)

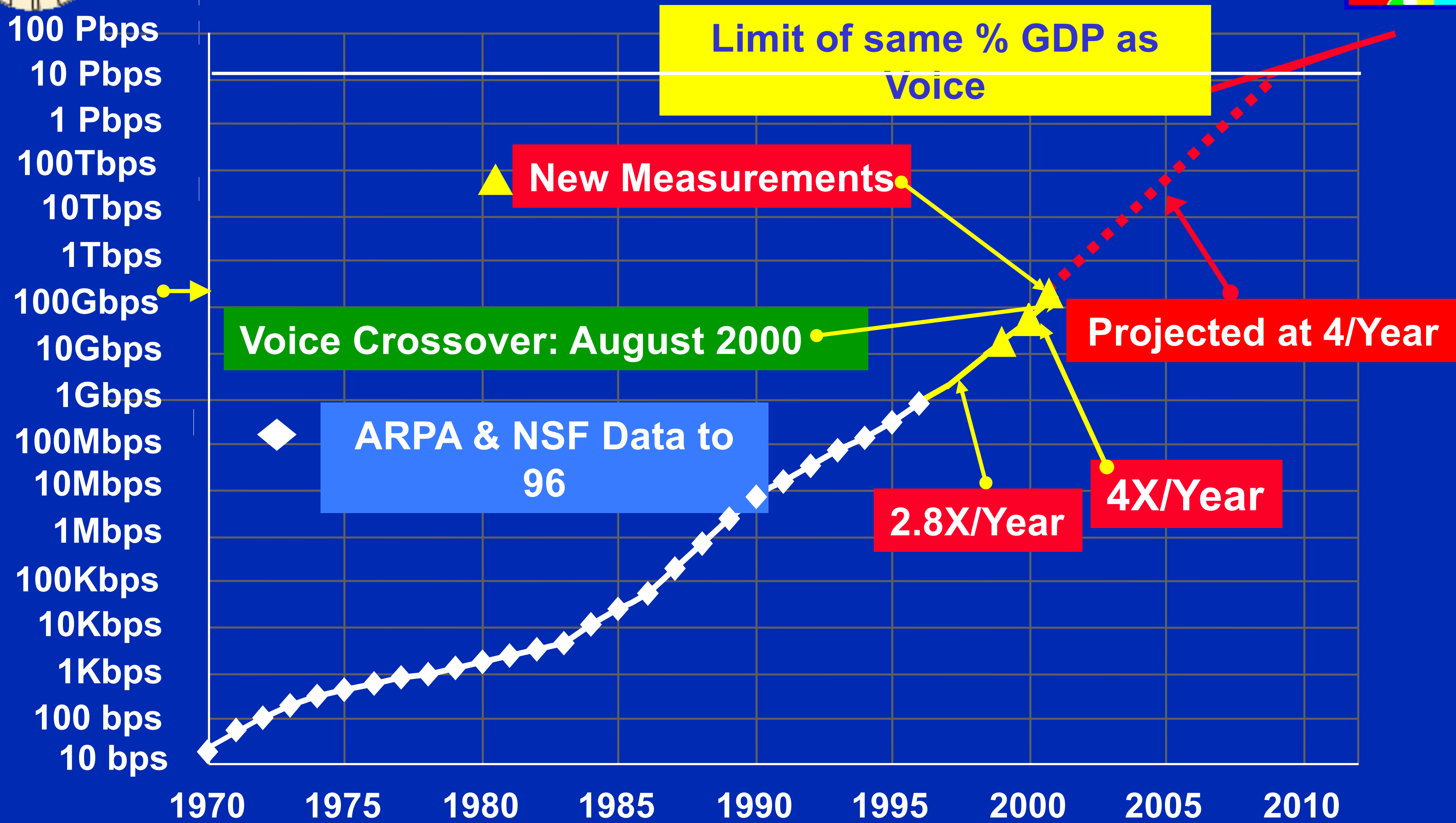


AMPATH Miami (to 80 Mbps)





Total U.S. Internet Traffic



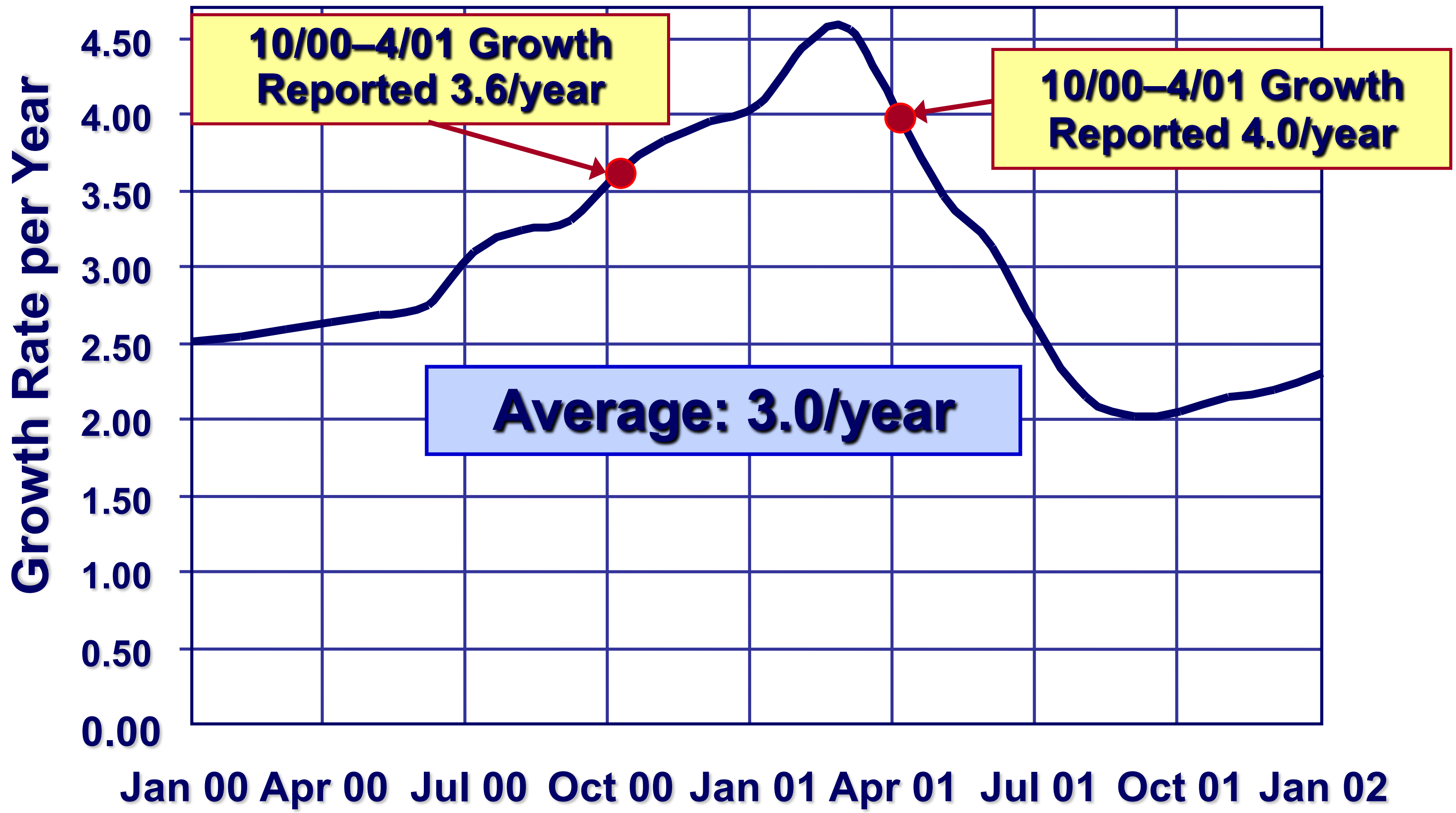
U.S. Internet Traffic

Source: Roberts et al., 2001

Internet Growth Rate Fluctuates Over Time

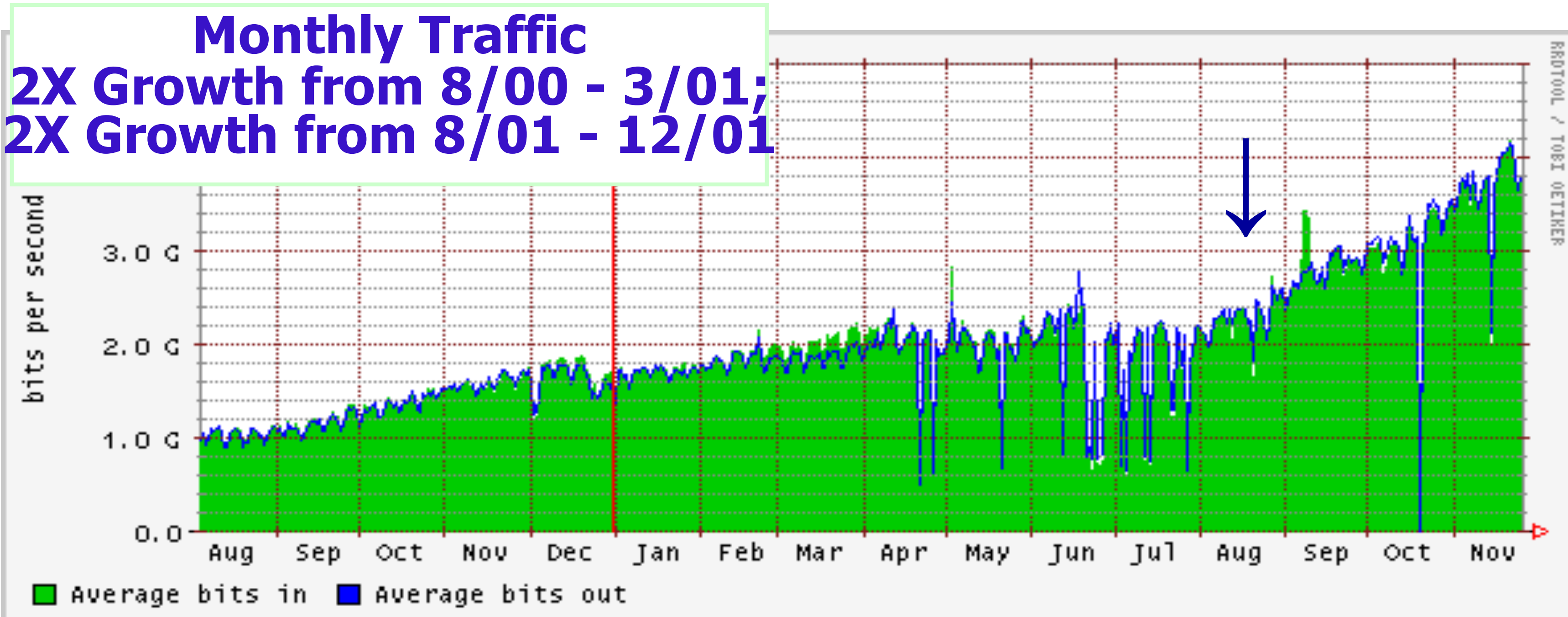
U.S. Internet Edge Traffic Growth Rate

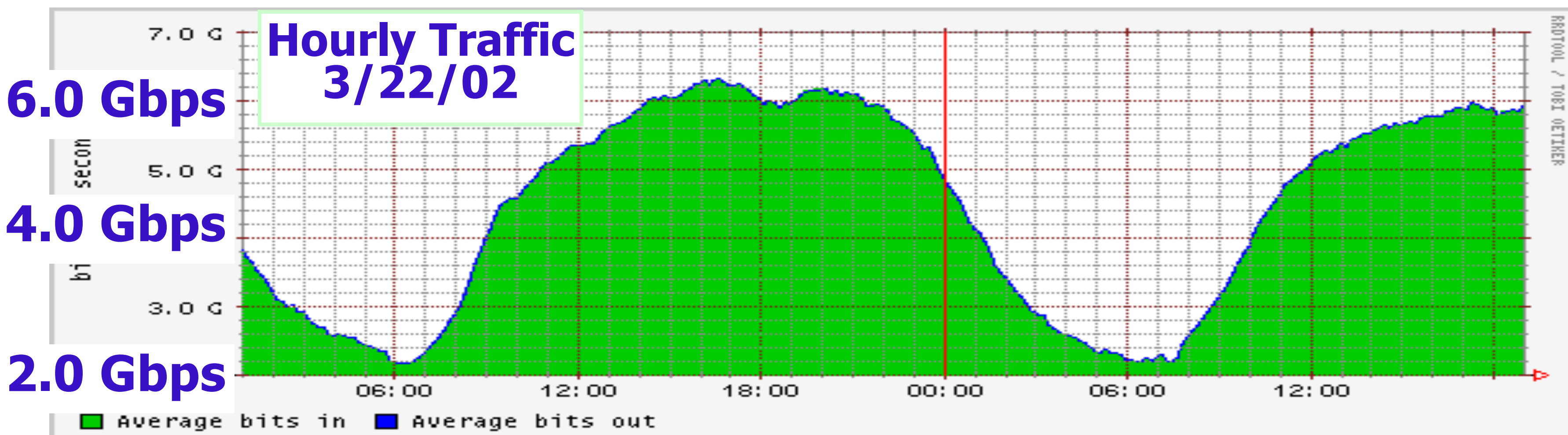
6 Month Lagging Measure



Source: Roberts et al., 2002

AMS-IX Internet Exchange Throughpu
Accelerating Growth in Europe (NL)





ICFA SCIC 12/01 – 3/02: Backbone and International Link Progress

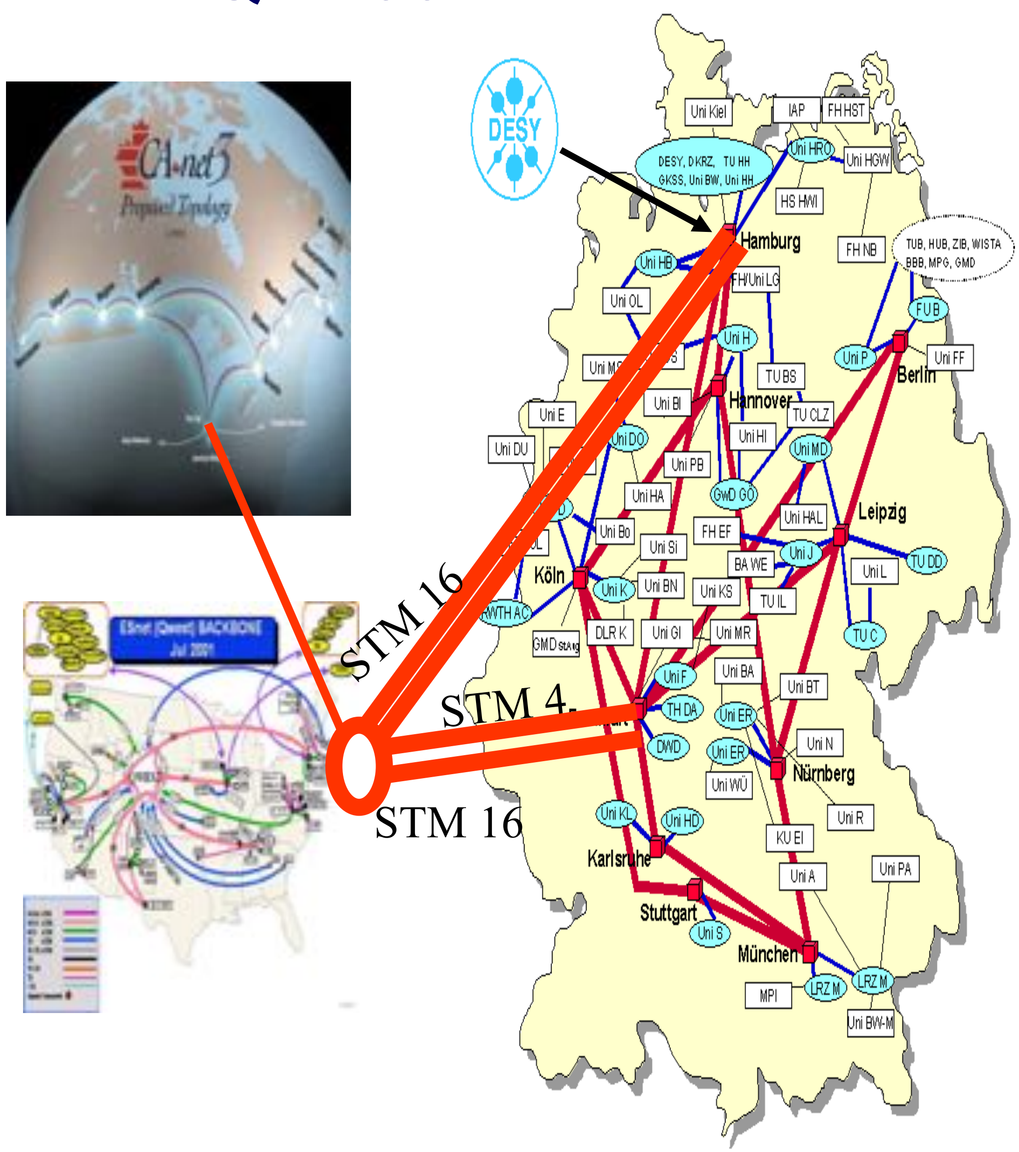
- GEANT Pan-European Backbone (http://www.dante.net/geant)
  - Now interconnects 31 countries
  - Includes many trunks at 2.5 and 10 Gbps
- UK
  - 2.5 Gbps NY-London, with 622 Mbps to ESnet and Abilene; Commodity Internet peering in London instead of NY
- SuperSINET (Japan): 10 Gbps IP and 10 Gbps Wavelength
  - Upgrade to Two 0.6 Gbps Links, to Chicago and Seattle
  - Plan upgrade to 2 X 2.5 Gbps Connection to US West Coast by 2003
- CA*net4 (Canada): Interconnect customer-owned dark fiber nets across Canada at 10 Gbps, starting July 2002
  - "Lambda-Grids" by ~2004-5
- GWIN (Germany): Connection to Abilene to 2 X 2.5 Gbps in 2002
- Russia
  - Start 10 Mbps link to CERN and ~90 Mbps to US Now

ICFA SCIC Meeting March 9 at CERN: Updates from Members

- **Abilene Upgrade** from 2.5 to 10 Gbps
  - Additional scheduled lambdas planned for targeted applications
- **US-CERN**
  - Upgrade On Track: 2 X 155 to 622 Mbps in April; Move to STARLIGHT
  - 2.5G Research Lambda by this Summer: STARLIGHT-CERN
  - 2.5G Triangle between STARLIGHT (US), SURFNet (NL), CERN
- **SLAC + IN2P3 (BaBar)**
  - Getting 100 Mbps over 155 Mbps CERN-US Link
  - 50 Mbps Over RENATER 155 Mbps Link, Limited by ESnet
  - 600 Mbps Throughput is BaBar Target for this Year
- **FNAL**
  - Expect ESnet Upgrade to 622 Mbps this Month
  - Plans for dark fiber to STARLIGHT, could be done in ~6 Months; Railway and Electric Co. providers considered

National R&E Network Example

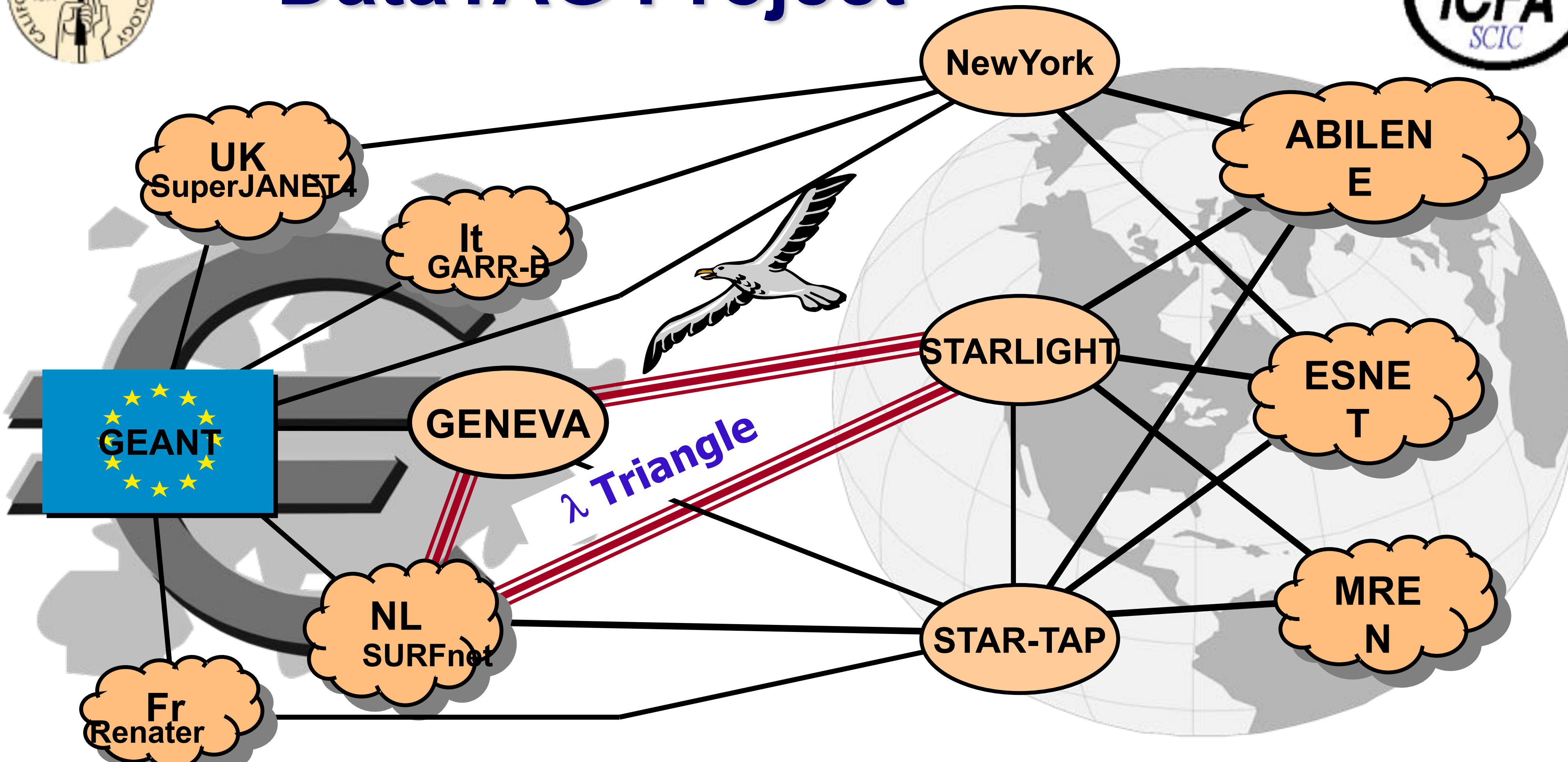
Germany: DFN TransAtlanticConnectivity Q1 2002



- 2 X OC12 Now: NY-Hamburg and NY-Frankfurt
- ESNet peering at 34 Mbps
- Upgrade to 2 X OC48 expected in Q1 2002
- Direct Peering to Abilene and Canarie expected
- UCAID will add (?) another 2 OC48's; Proposing a Global Terabit Research Network (GTRN)

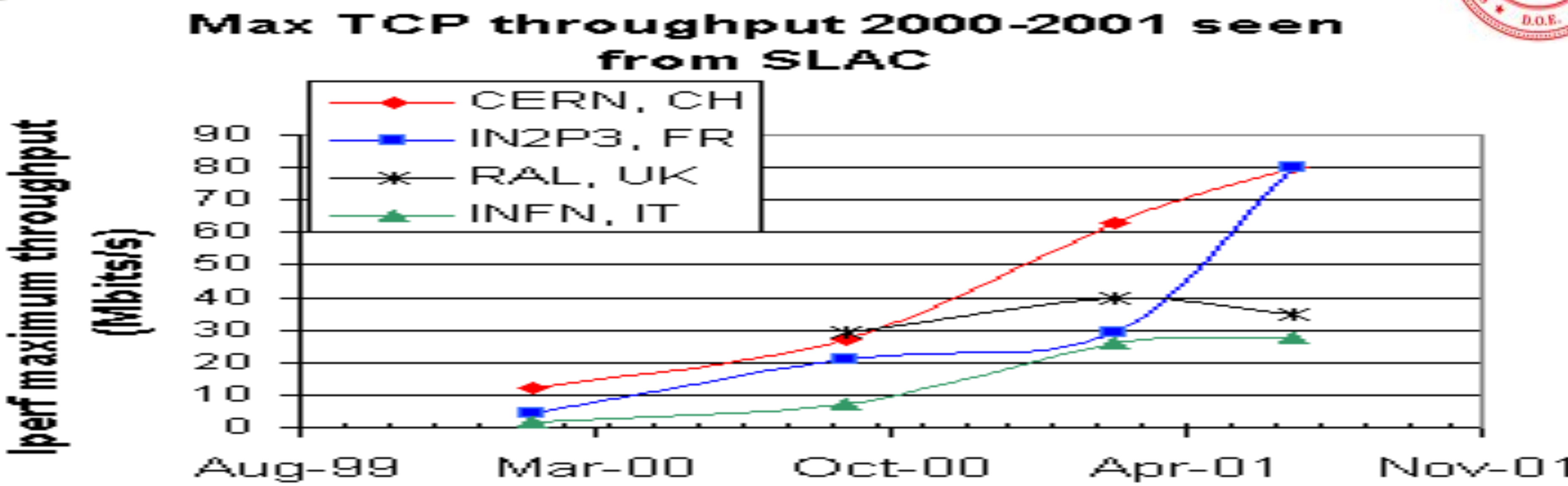
- FSU Connections via satellite: Yerevan, Minsk, Almaty, Baikal
  - Speeds of 32 - 512 kbps
- SILK Project (2002): NATO funding
  - Links to Caucasus and Central Asia (8 Countries)
  - Currently 64-512 kbps
  - Propose VSAT for 10-50 X BW: NATO + State Funding

DataTAG Project



- EU-Solicited Project. CERN, PPARC (UK), Amsterdam (NL), and INFN (IT); and US (DOE/NSF: UIC, NWU and Caltech) partners
- Main Aims:
  - Ensure maximum interoperability between US and EU Grid Projects
  - Transatlantic Testbed for advanced network research
- 2.5 Gbps wavelength-based US-CERN Link 7/02 (10 Gbps in 2003 or 2004)

Maximum Throughput on Transatlantic Links (155 Mbps)



- 8/01 105 Mbps reached with 30 Streams: SLAC-IN2P3
- 9/1/01 102 Mbps in One Stream: CIT-CERN
- 11/5/01 125 Mbps in One Stream (modified kernel): CIT-CERN
  135 Mbps in One Stream (modified kernel): CIT-Chicago
- 1/09/02 190 Mbps for One stream shared on 2 155 Mbps links
- 3/11/02 120 Mbps Disk-to-Disk with One Stream on a 155 Mbps link (Chicago-CERN)

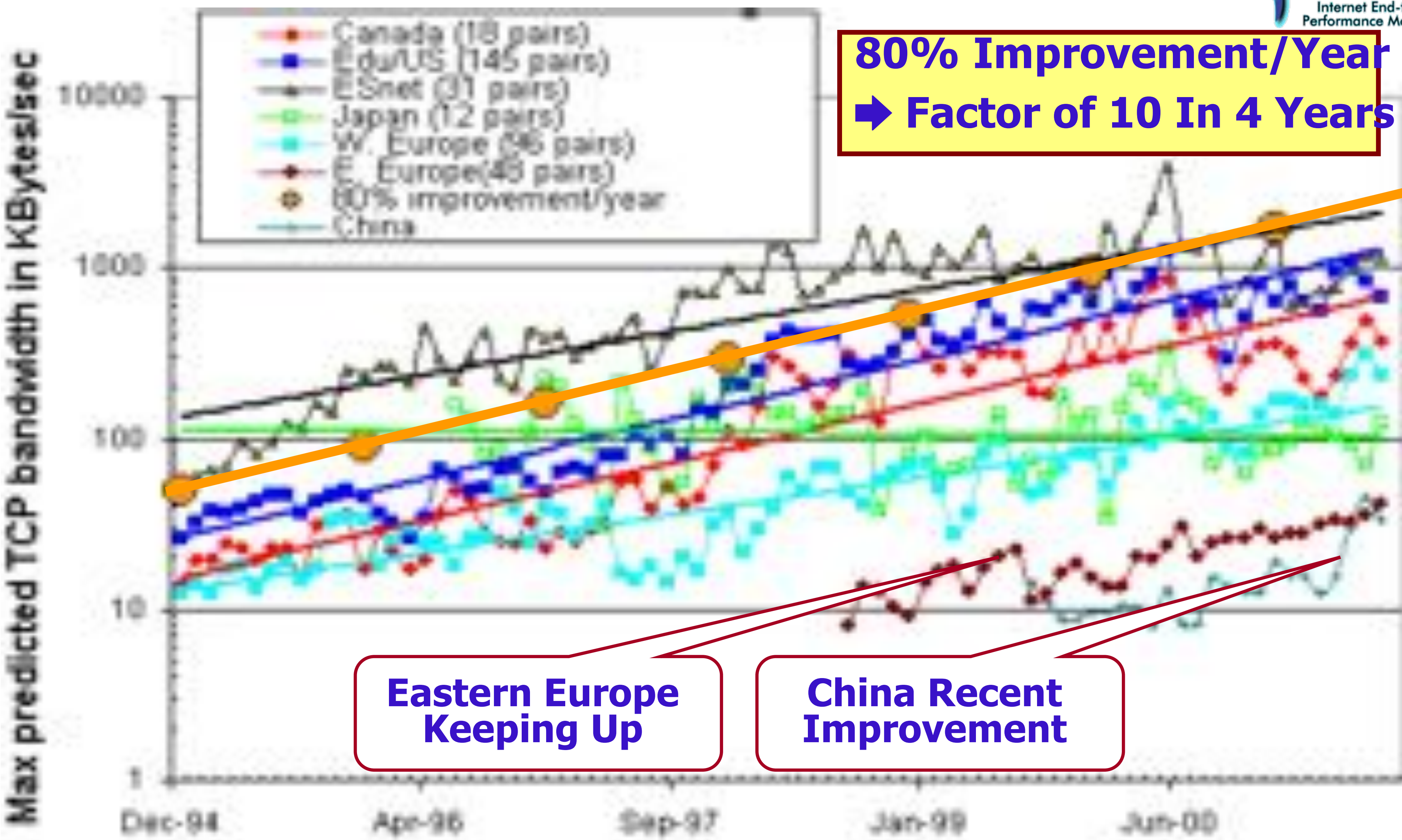
Also see http://www-iepm.slac.stanford.edu/monitoring/bulk/; and the Internet2 E2E Initiative: http://www.internet2.edu/e2e

Key Network Issues & Challenges

- **Net Infrastructure Requirements for High Throughput**
  - *Packet Loss must be ~Zero (well below 0.01%)*
    - I.e. No "Commodity" networks
    - Need to track down uncongested packet loss
  - No Local infrastructure bottlenecks
    - Gigabit Ethernet "clear paths" between selected host pairs are needed now
    - To 10 Gbps Ethernet by ~2003 or 2004
  - TCP/IP stack configuration and tuning Absolutely Required
    - Large Windows; Possibly Multiple Streams
    - New Concepts of *Fair Use* Must then be Developed
  - Careful Router, Server, Client, Interface configuration; monitoring
    - Sufficient CPU, I/O and NIC throughput sufficient
  - *End-to-end* monitoring and tracking of performance
  - Close collaboration with local and "regional" network staffs

TCP Does Not Scale to the 1-10 Gbps Range

Throughput quality improvements: $BW_{TCP} < MSS/(RTT*sqrt(loss))$ [*]



[*] See "Macroscopic Behavior of the TCP Congestion Avoidance Algorithm," Matthis, Semke, Mahdavi, Ott, Computer Communication Review 27(3), 7/1997

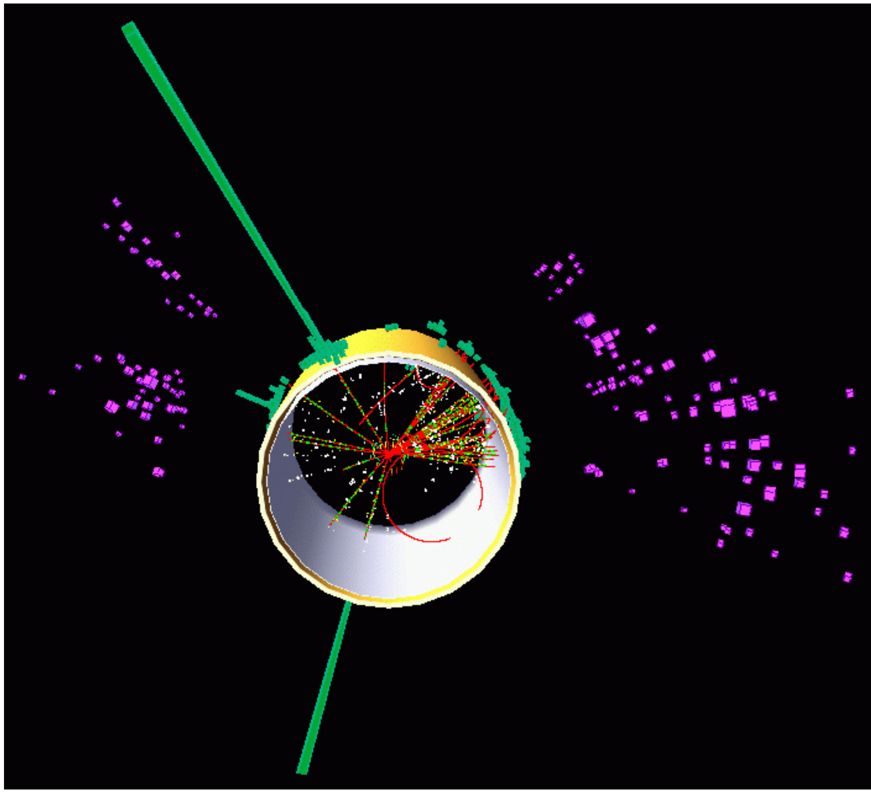
Networks, Grids and HENP

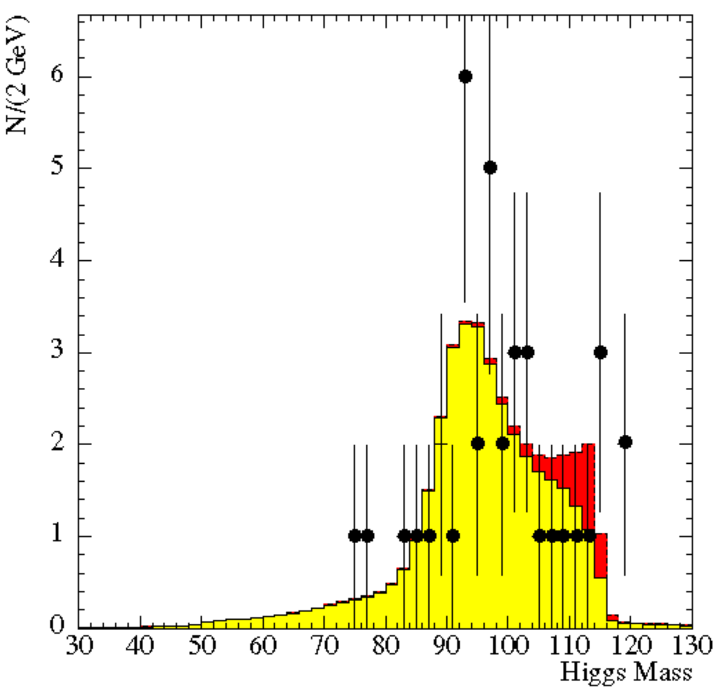
- **Grids are changing the way we do science and engineering**
  - **Successful use of Grids relies on high performance national and international networks**
- **Next generation 10 Gbps network backbones are almost here: in the US, Europe and Japan**
  - **First stages arriving in 6-12 months**
- **Major transoceanic links at 2.5 - 10 Gbps within 0-18 months**
- ***Network improvements are especially needed in South America; and some other world regions. Leading Examples:***
  - ***Brazil, Chile; India, Pakistan, China; Southeastern Europe; Africa***
- **Removing regional, last mile bottlenecks and compromises in network quality are now *all on the critical path***
- **Getting high (reliable) Grid performance across networks means!**
  - **End-to-end monitoring; a coherent approach**
  - **Getting high performance (TCP) toolkits in users' hands**
  - **Working in concert with AMPATH, Internet E2E, I2 HENP WG, DataTAG; Working with the Grid projects and GGF**

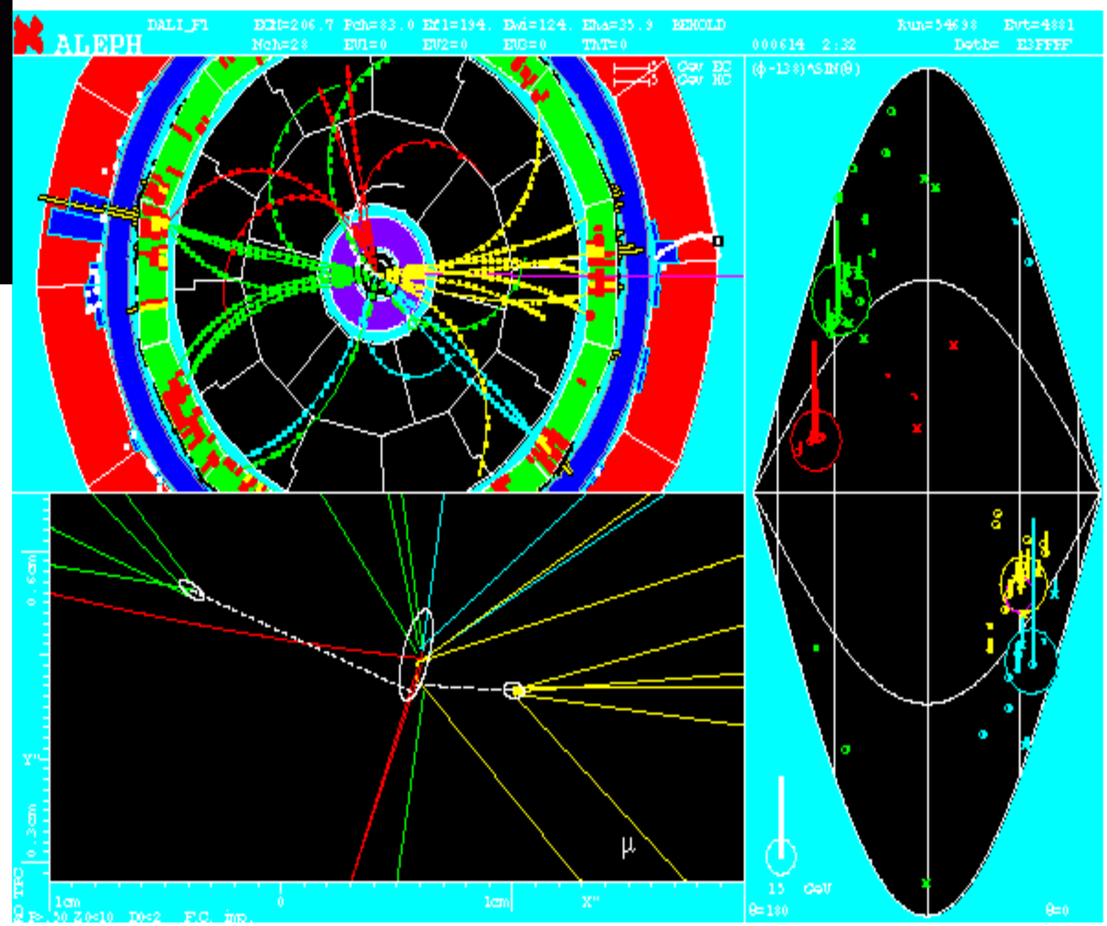
Some Extra Slides Follow

Evidence for the Higgs at LEP at M~115 GeV

The LEP Program Has Now Ended







The Large Hadron Collider (2006-)

- **The Next-generation Particle Collider**
  - ➡ **The largest superconductor installation in the world**
- **Bunch-bunch collisions at 40 MHz, Each generating ~20 interactions**
  - ➡ **Only one in a trillion may lead to a major physics discovery**
- ***Real-time data filtering: Petabytes per second to Gigabytes per second***
- ***Accumulated data of many Petabytes/Year***

Large data samples explored and analyzed by thousands of globally dispersed scientists, in hundreds of teams

Rapid Advances of Nat'l Backbones: Next Generation Abilene

- Abilene partnership with Qwest extended through 2006
- Backbone to be upgraded to 10-Gbps in phases, to be Completed by October 2003
  - GigaPoP Upgrade started in February 2002
- Capability for flexible λ provisioning in support of future experimentation in optical networking
  - In a multi- λ infrastructure

US CMS TeraGrid Seamless Prototype

- Caltech/Wisconsin Condor/NCSA Production
- Simple Job Launch from Caltech
  - Authentication Using Globus Security Infrastructure (GSI)
  - Resources Identified Using Globus Information Infrastructure (GIS)
- CMSIM Jobs (Batches of 100, 12-14 Hours, 100 GB Output) Sent to the Wisconsin Condor Flock Using Condor-G
  - Output Files Automatically Stored in NCSA Unitree (Gridftp)
- ORCA Phase: Read-in and Process Jobs at NCSA
  - Output Files Automatically Stored in NCSA Unitree
- Future: Multiple CMS Sites; Storage in Caltech HPSS Also, Using GDMP (With LBNL's HRM).
- Animated Flow Diagram of the DTF Prototype:

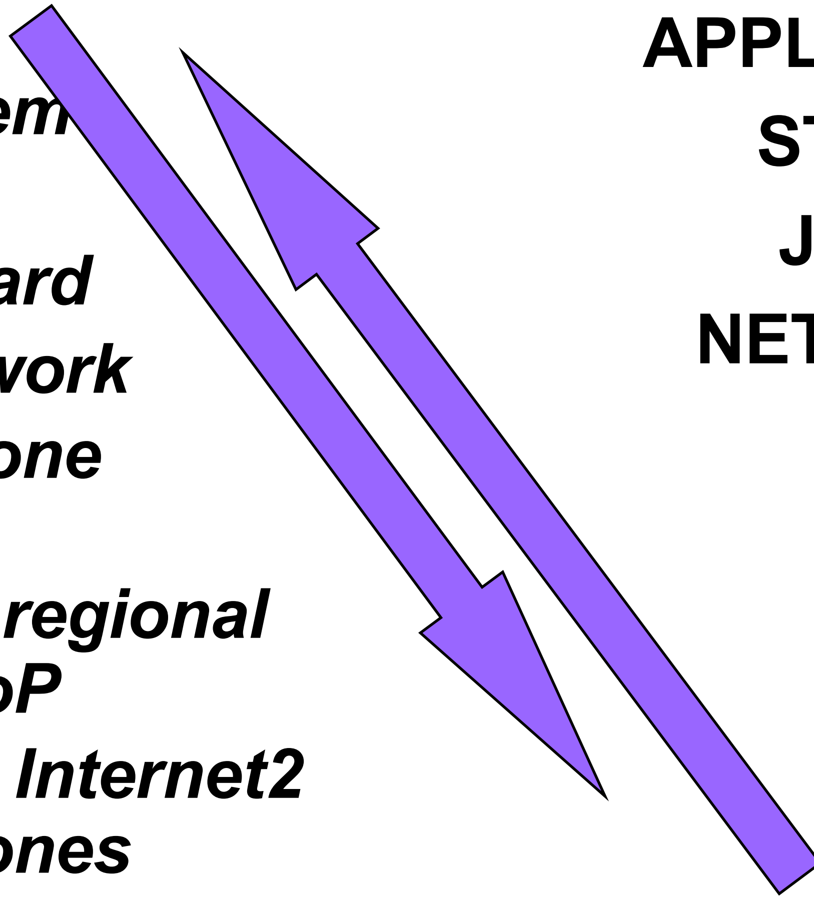
http://cmsdoc.cern.ch/~wisniew/infrastructure.html

Internet2 HENP WG [*]

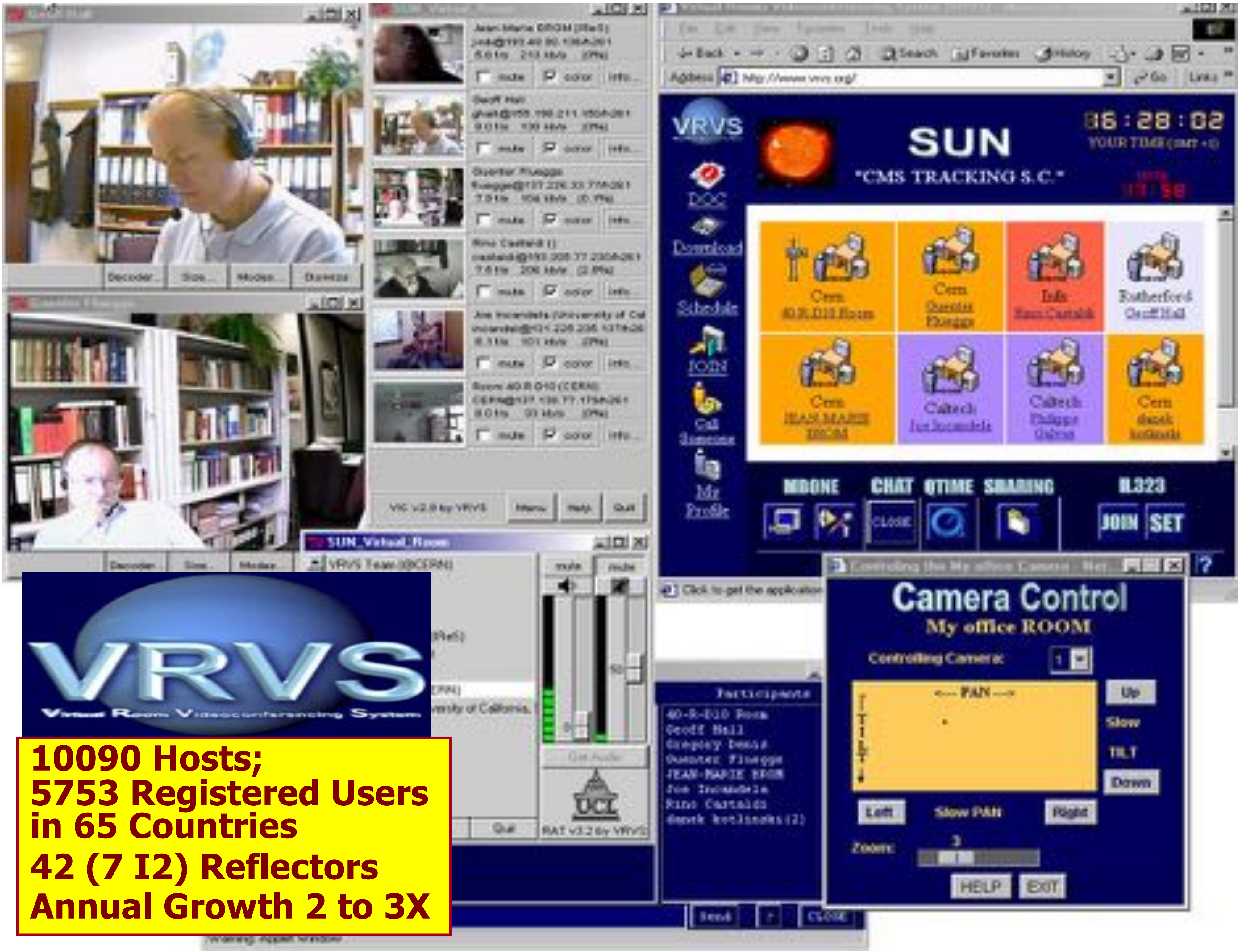
- Mission: To help ensure that the required
  - National and international network infrastructures (end-to-end)
  - Standardized tools and facilities for high performance and end-to-end monitoring and tracking, and
  - Collaborative systems
- are developed and deployed in a timely manner, and used effectively to meet the needs of the US LHC and other major HENP Programs, as well as the at-large scientific community.
  - To carry out these developments in a way that is broadly applicable across many fields
- Formed an Internet2 WG as a suitable framework: Oct. 26 2001
- [*] Co-Chairs: S. McKee (Michigan), H. Newman (Caltech); Sec'y J. Williams (Indiana
- *Website: http://www.internet2.edu/henp; also see the Internet2 End-to-end Initiative: http://www.internet2.edu/e2e*

True End to End Experience

- *User perception*
- *Application*
- *Operating system*
- *Host IP stack*
- *Host network card*
- *Local Area Network*
- *Campus backbone network*
- *Campus link to regional network/GigaPoP*
- *GigaPoP link to Internet2 national backbones*
- *International connections*



EYEBALL
APPLICATION
STACK
JACK
NETWORK
. . .
. . .
. . .
. . .



10090 Hosts;
5753 Registered Users in 65 Countries
42 (7 I2) Reflectors
Annual Growth 2 to 3X