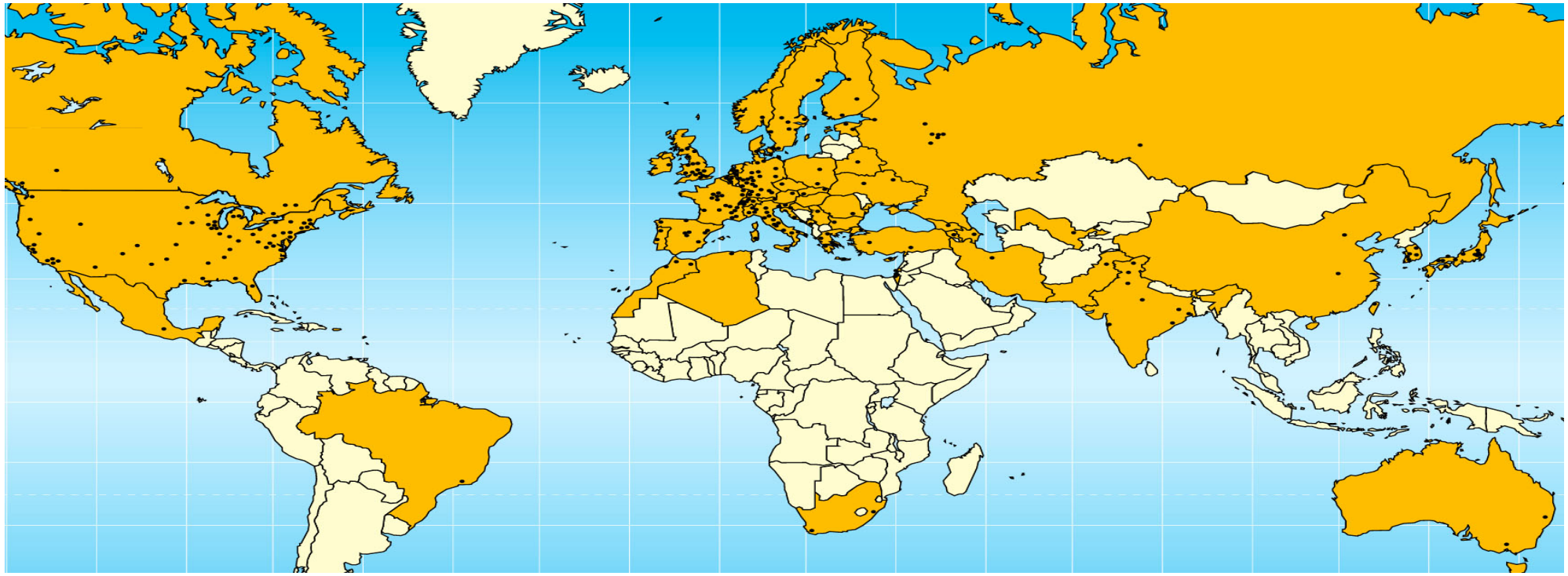
HENP Networks, ICFA SCIC and the Digital Divide



Harvey B. Newman

California Institute of Technology

AMPATH Workshop, FIU
January 30, 2003

Next Generation Networks for Experiments: Goals and Needs

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

- Providing rapid access to event samples, subsets and analyzed physics results from massive data stores
  - From Petabytes by 2002, ~100 Petabytes by 2007, to ~1 Exabyte by ~2012.
- Providing analyzed results with rapid turnaround, by coordinating and managing the large but *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, monitored, quantifiable high performance

ICFA Standing Committee on Interregional Connectivity (SCIC)

- Created by ICFA in July 1998 in Vancouver ; Following ICFA-NTF
- CHARGE:

  Make recommendations to ICFA concerning the connectivity between *the Americas*, Asia and Europe (and network requirements of HENP)

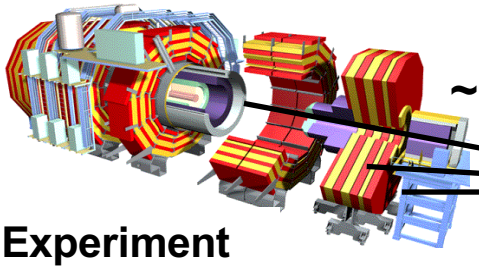
  - As part of the process of developing these recommendations, the committee should
    - Monitor traffic
    - Keep track of technology developments
    - Periodically review forecasts of future bandwidth needs, and
    - Provide early warning of potential problems
- Create subcommittees when necessary to meet the charge
- The chair of the committee should report to ICFA once per year, at its joint meeting with laboratory directors (Feb. 2003)
- Representatives: Major labs, ECFA, ACFA, NA Users, S. America

SCIC Sub-Committees

Web Page http://cern.ch/ICFA-SCIC/

- **Monitoring:** Les Cottrell (http://www.slac.stanford.edu/xorg/icfa/scic-netmon) With Richard Hughes-Jones (Manchester), Sergio Novaes (Sao Paolo); Sergei Berezhnev (RUHEP), Fukuko Yuasa (KEK), Daniel Davids (CERN), Sylvain Ravot (Caltech), Shawn McKee (Michigan)
- **Advanced Technologies:** Richard Hughes-Jones, With Vladimir Korenkov (JINR, Dubna), Olivier Martin(CERN), Harvey Newman
- **The Digital Divide:** Alberto Santoro (Rio, Brazil)
  - With V. Ilyin (MSU), Y. Karita(KEK), D.O. Williams (CERN)
  - Also Dongchul Son (Korea), Hafeez Hoorani (Pakistan), Sunanda Banerjee (India), Vicky White (FNAL)
- **Key Requirements:** Harvey Newman
  - Also Charlie Young (SLAC)

LHC Data Grid Hierarchy



Experiment

~PBytes/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

HPSS

Tier 1

~2.5 Gbps

IN2P3 Center

RAL Center

INFN Center

FNAL

2.5 Gbps

Tier 2

~2.5 Gbps

Tier2 Center

Tier 3

Institute ~0.25TIPS

Physics data cache

0.1 to 10 Gbps

Workstations

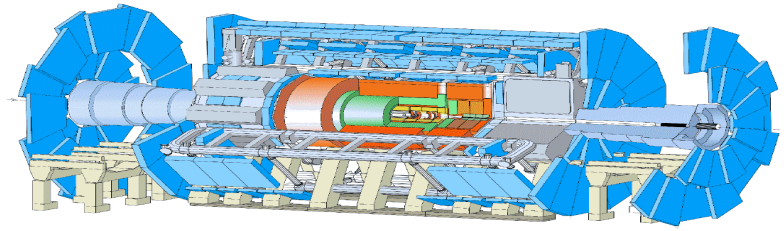
Tier 4

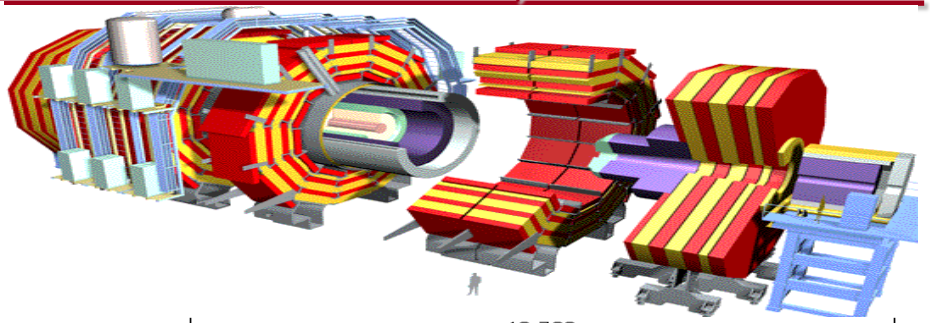
Tens of Petabytes by 2007-8.
An Exabyte within ~5 Years later.

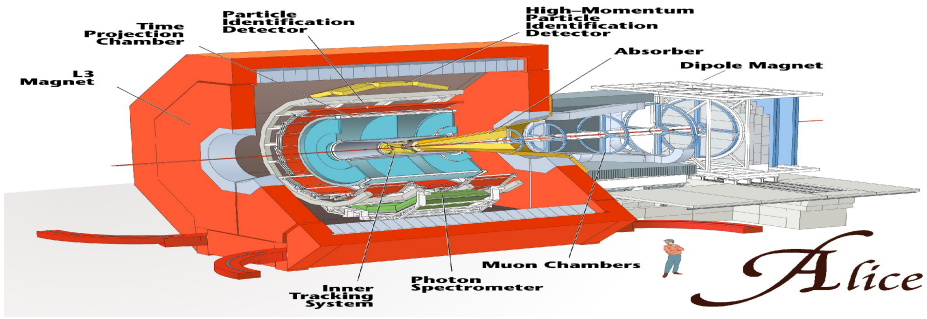
Four LHC Experiments: The Petabyte to Exabyte Challenge

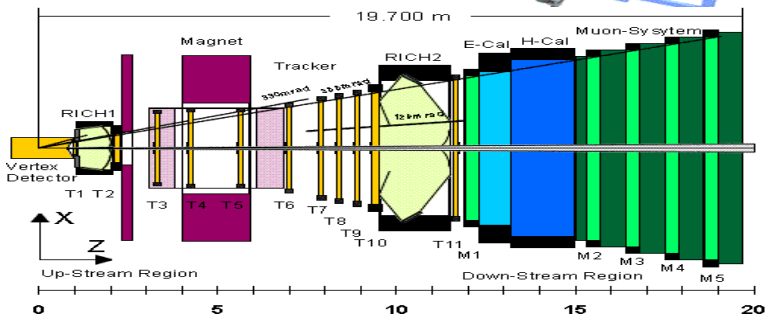
ATLAS, CMS, ALICE, LHCB

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

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	2500	5000	10000	20000

[*] BW Requirements Increasing Faster Than Moore's Law

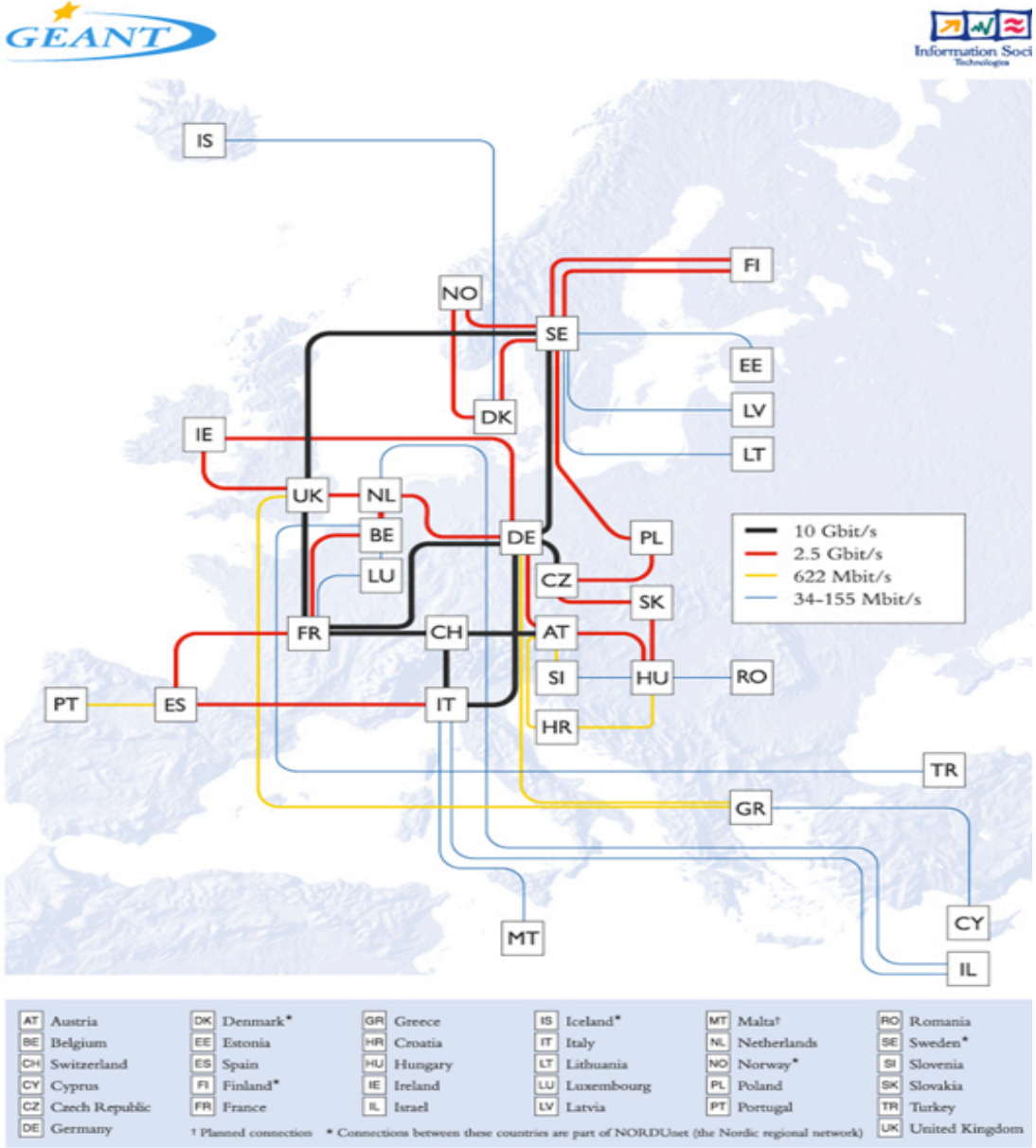
See http://gate.hep.anl.gov/lprice/TAN

ICFA SCIC: R&E Backbone and International Link Progress

- GEANT Pan-European Backbone (http://www.dante.net/geant)
  - Now interconnects >31 countries; many trunks 2.5 and 10 Gbps
- UK: SuperJANET Core at 10 Gbps
  - 2.5 Gbps NY-London, with 622 Mbps to ESnet and Abilene
- France (IN2P3): 2.5 Gbps RENATER backbone from October 2002
  - Lyon-CERN Link Upgraded to 1 Gbps Ethernet
  - Proposal for dark fiber to CERN by end 2003
- SuperSINET (Japan): 10 Gbps IP and 10 Gbps Wavelength Core
  - Tokyo to NY Links: 2 X 2.5 Gbps started; Peer with ESNet by Feb.
- CA*net4 (Canada): Interconnect customer-owned dark fiber nets across Canada at 10 Gbps, started July 2002
  - "Lambda-Grids" by ~2004-5
- GWIN (Germany): 2.5 Gbps Core; Connect to US at 2 X 2.5 Gbps; Support for SILK Project: Satellite links to FSU Republics
- Russia: 155 Mbps Links to Moscow (Typ. 30-45 Mbps for Science)
  - Moscow-Starlight Link to 155 Mbps (US NSF + Russia Support)
  - Moscow-GEANT and Moscow-Stockholm Links 155 Mbps

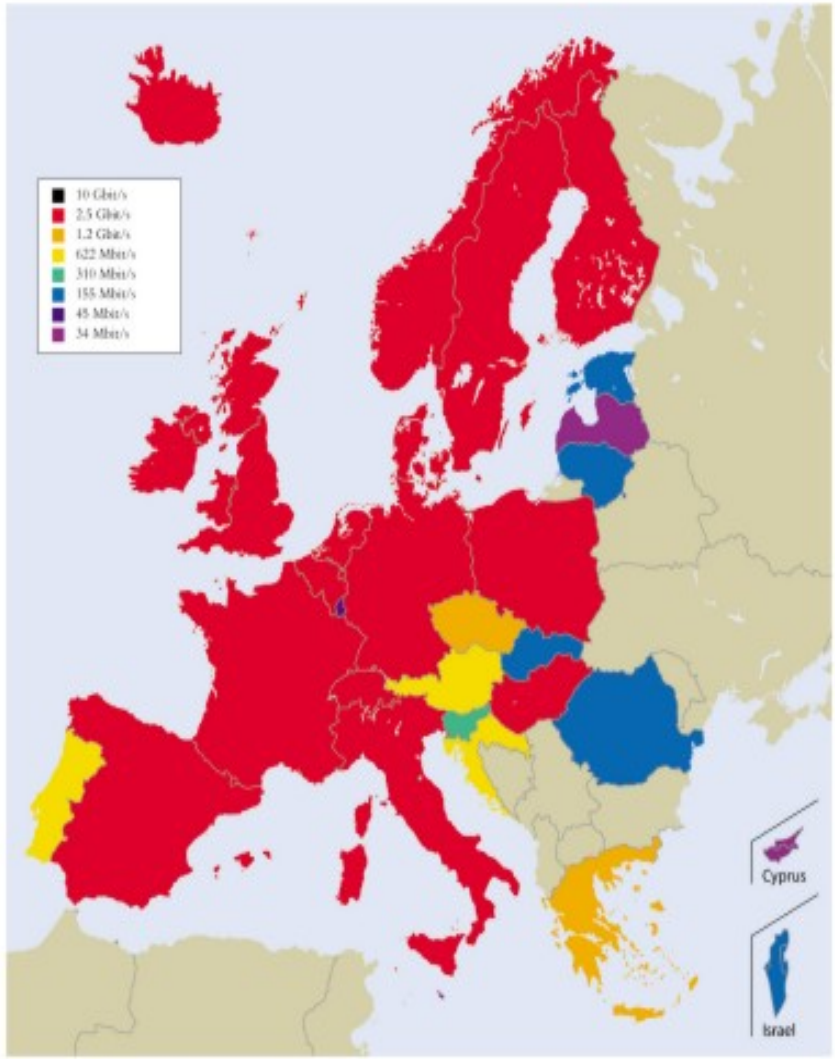
R&E Backbone and Int'l Link Progress

- Abilene (Internet2) Upgrade from 2.5 to 10 Gbps started in 2002
  - ➡ Encourage high throughput use for targeted applications; FAST
- ESNET: Upgrade: to 10 Gbps "As Soon as Possible"
- US-CERN
  - ➡ to 622 Mbps in August; Move to STARLIGHT
  - ➡ 2.5G Research Triangle from 8/02; STARLIGHT-CERN-NL; to 10G in 2003. [10Gbps SNV-Starlight Link Loan from Level(3)
- SLAC + IN2P3 (BaBar)
  - ➡ Typically ~400 Mbps throughput on US-CERN, Renater links
  - ➡ 600 Mbps Throughput is BaBar Target for Early 2003 (with ESnet and Upgrade)
- FNAL: ESnet Link Upgraded to 622 Mbps
  - ➡ Plans for dark fiber to STARLIGHT, proceeding
- NY-Amsterdam *Donation* from Tyco, September 2002: Arranged by IEEAF: 622 Gbps+10 Gbps Research Wavelength
- US National Light Rail Proceeding; Startup Expected this Year



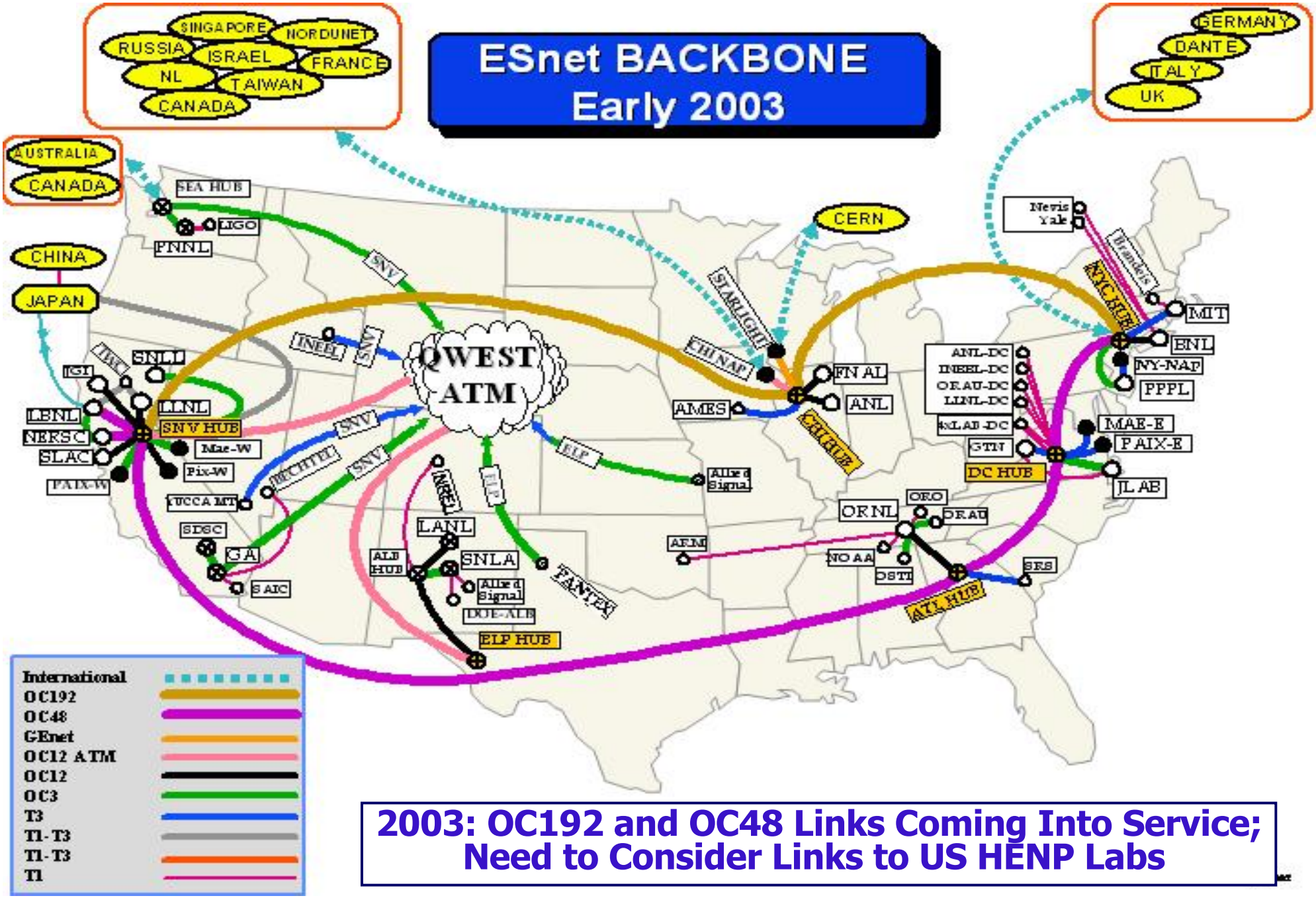
Multi-Gigabit pan-European Research Network
Backbone Topology December 2002

DANTE
www.dante.n



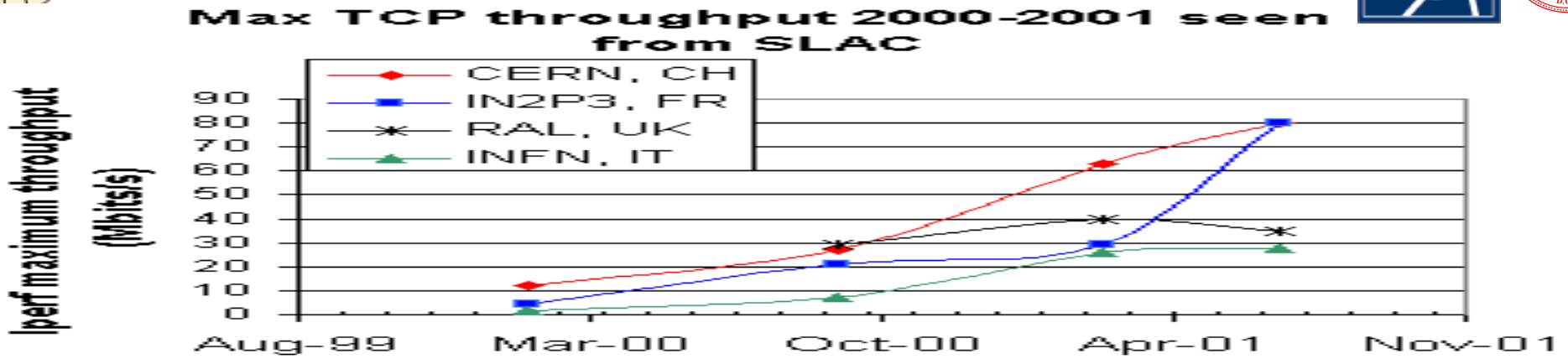
Multi-Gigabit pan-European Research Network
Backbone Access Speeds August 2002

DANTE
www.dante.net



2003: OC192 and OC48 Links Coming Into Service; Need to Consider Links to US HENP Labs

Progress: Max. Sustained TCP Thruput on Transatlantic and US Links

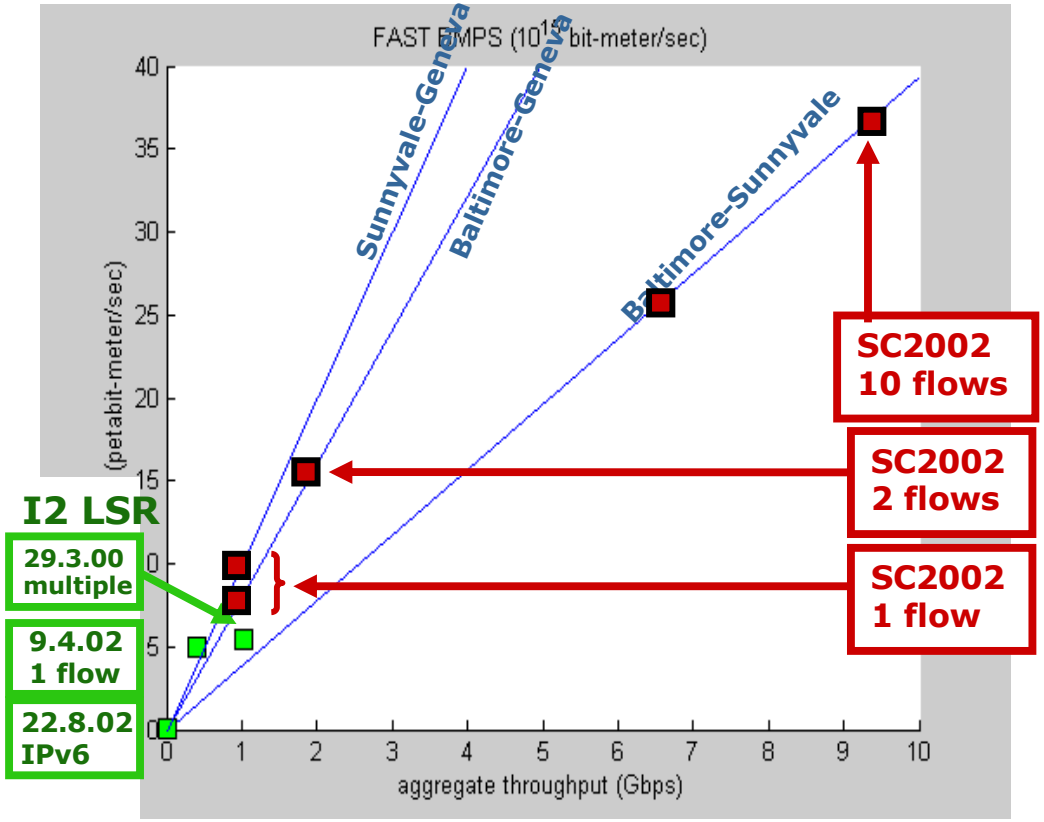


- 8-9/01 105 Mbps 30 Streams: SLAC-IN2P3; 102 Mbps 1 Stream CIT-CERN
- 11/5/01 125 Mbps in One Stream (modified kernel): CIT-CERN
- 1/09/02 190 Mbps for One stream shared on 2 155 Mbps links
- 3/11/02 120 Mbps Disk-to-Disk with One Stream on 155 Mbps link (Chicago-CERN)
- 5/20/02 450-600 Mbps SLAC-Manchester on OC12 with ~100 Streams
- 6/1/02 290 Mbps Chicago-CERN One Stream on OC12 (mod. Kernel)
- 9/02 850, 1350, 1900 Mbps Chicago-CERN 1,2,3 GbE Streams, OC48 Link
- 11-12/02 FAST: 940 Mbps in 1 Stream SNV-CERN; 9.4 Gbps in 10 Flows SNV-Chicago

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

FAST (Caltech): A Scalable, "Fair" Protocol for Next-Generation Networks: from 0.1 To 100 Gbps

SC2002 11/02



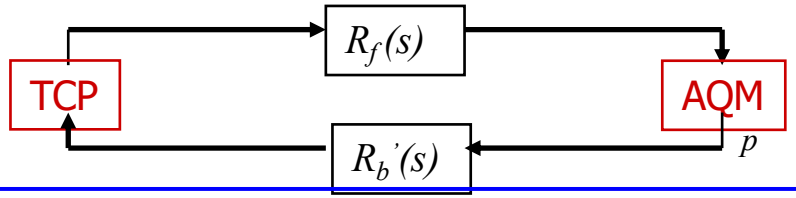
Highlights of FAST TCP

- Standard Packet Size
- 940 Mbps single flow/GE card
  - 9.4 petabit-m/sec
  - 1.9 times LSR
- 9.4 Gbps with 10 flows
  - 37.0 petabit-m/sec
  - 6.9 times LSR
- 22 TB in 6 hours; in 10 flows

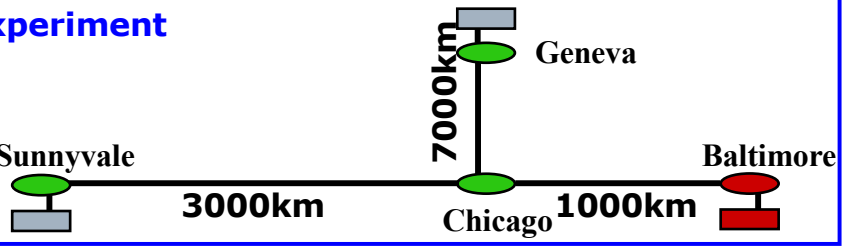
Implementation

- Sender-side (only) mods
- Delay (RTT) based
- Stabilized Vegas

Internet: distributed feedback system **Theory**



Experiment



URL: netlab.caltech.edu/FAST

Next: 10GbE; 1 GB/sec disk to disk

C. Jin, D. Wei, S. Low
FAST Team & Partners

HENP Major Links: Bandwidth Roadmap (Scenario) in Gbps

Year	Production	Experimental	Remarks
2001	0.155	0.622-2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH DWDM; GigE Integ.
2003	2.5	10	DWDM; 1 + 10 GigE Integration
2005	10	2-4 X 10	λ Switch; λ Provisioning
2007	2-4 X 10	~10 X 10; 40 Gbps	1st Gen. λ Grids
2009	~10 X 10 or 1-2 X 40	~5 X 40 or ~20-50 X 10	40 Gbps λ Switching
2011	~5 X 40 or ~20 X 10	~25 X 40 or ~100 X 10	2nd Gen λ Grids Terabit Networks
2013	~Terabit	~MultiTbps	~Fill One Fiber

Continuing the Trend: ~1000 Times Bandwidth Growth Per Decade; We are Rapidly Learning to Use and Share Multi-Gbps Networks

HENP Lambda Grids: Fibers for Physics

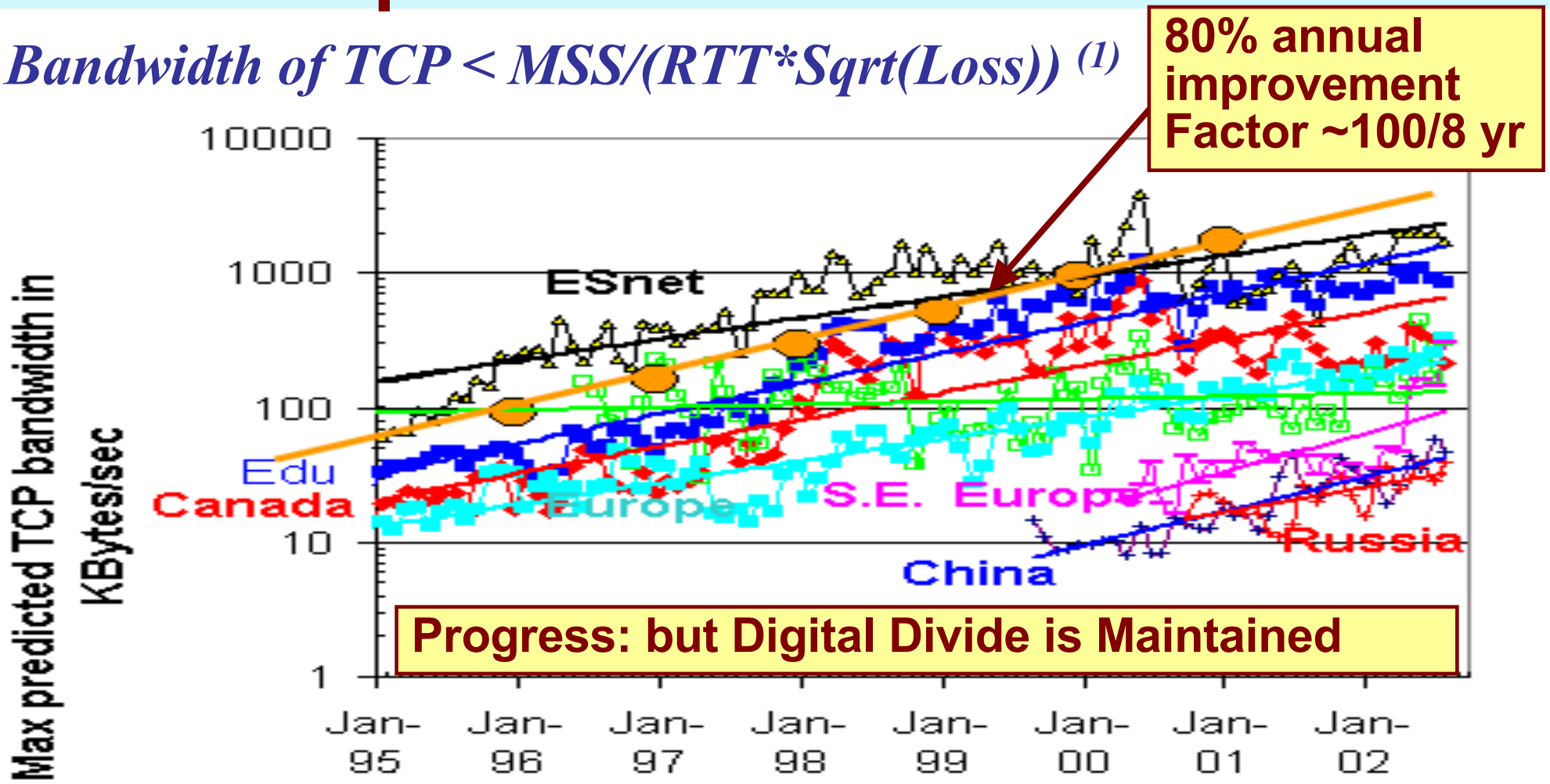
- Problem: Extract "Small" Data Subsets of 1 to 100 Terabytes from 1 to 1000 Petabyte Data Stores
- Survivability of the HENP Global Grid System, with hundreds of such transactions per day (circa 2007) requires that each transaction be completed in a relatively short time.
- Example: Take 800 secs to complete the transaction. Then

Transaction Size (TB)	Net Throughput (Gbps)
1	10
10	100
100	1000 (Capacity of Fiber Today)

- Summary: Providing Switching of 10 Gbps wavelengths within ~3-5 years; and Terabit Switching within 5-8 years would enable "Petascale Grids with Terabyte transactions", as required to fully realize the discovery potential of major HENP programs, as well as other data-intensive fields.

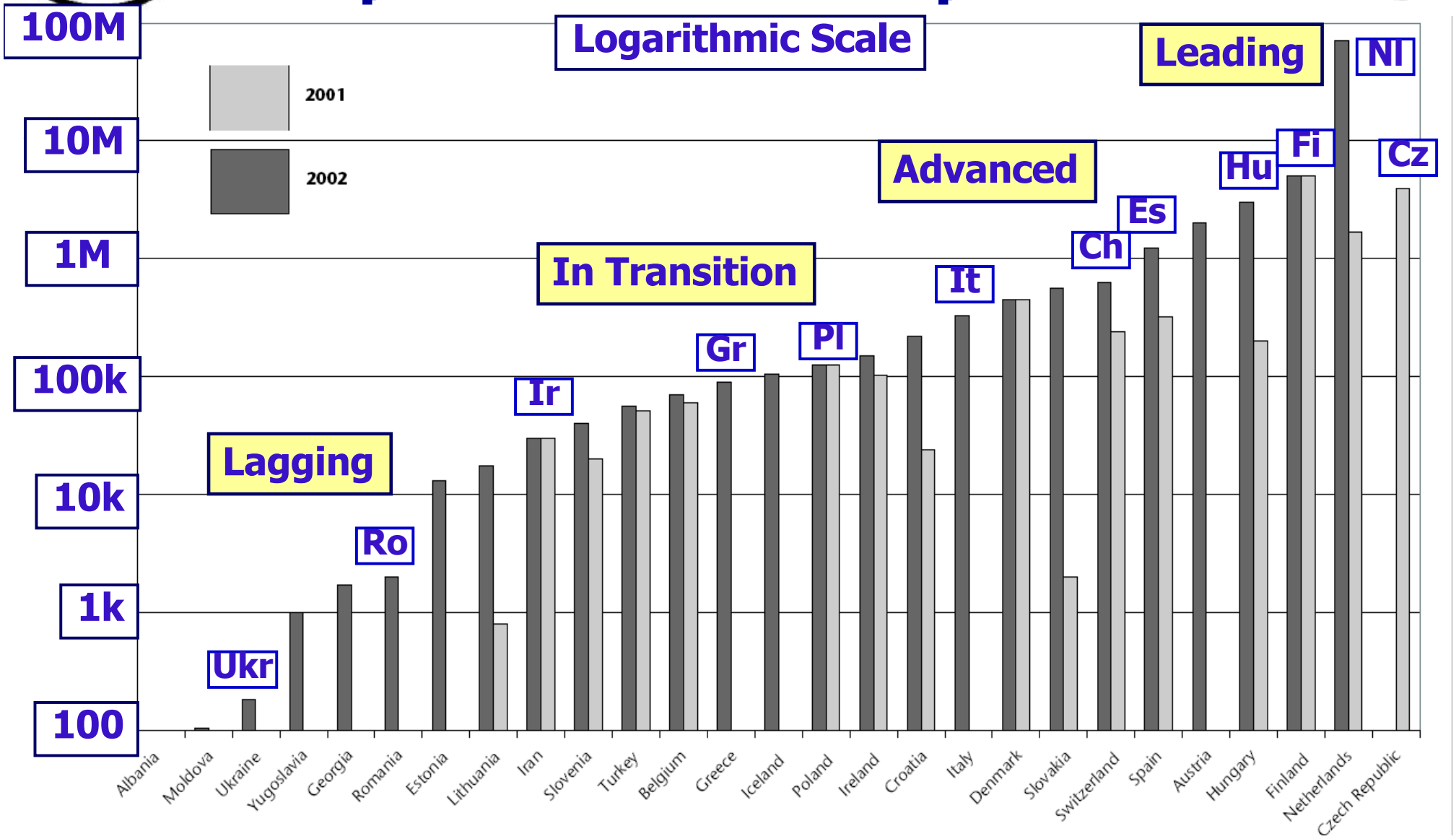
History - Throughput Quality Improvements from US

*Bandwidth of TCP < MSS/(RTT*Sqrt(Loss))* [1]



(1) *Macroscopic Behavior of the TCP Congestion Avoidance Algorithm*, Matthis, Semke, Mahdavi, Ott, Computer Communication Review 27(3), July 1997

NREN Core Network Size (Mbps-km):
http://www.terena.nl/compendium/2002



We Must Close the Digital Divide

Goal: To Make Scientists from All World Regions Full Partners in the Process of Search and Discovery

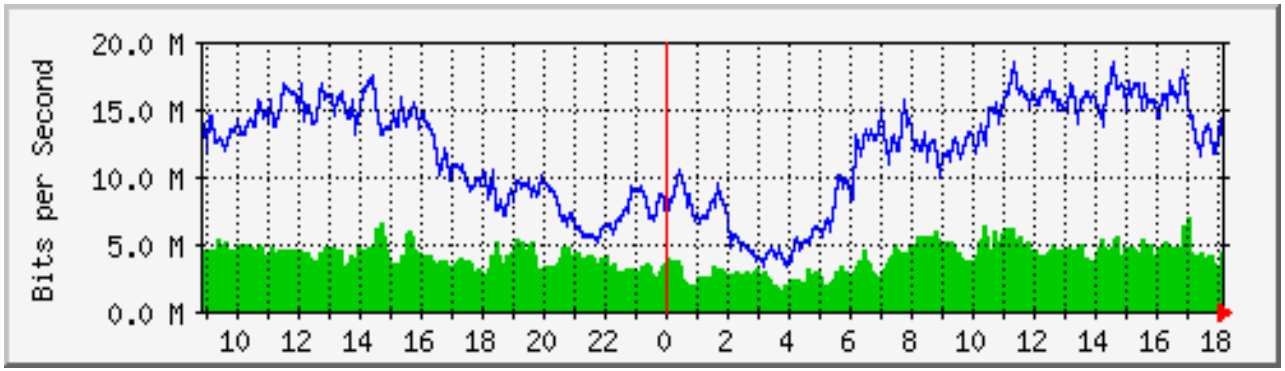
What ICFA and the HENP Community Can Do

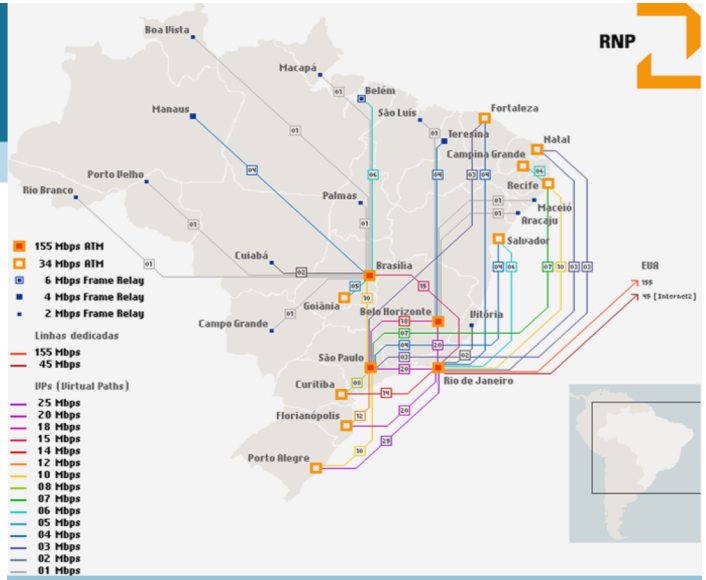
- Help identify and highlight specific needs (to Work On)
  - ➡ Policy problems; Last Mile problems; etc.
- Spread the message: ICFA SCIC is there to help; Coordinate with AMPATH, IEEAF, APAN, Terena, Internet2, etc.
- Encourage Joint programs [such as in DESY's Silk project; Japanese links to SE Asia and China; AMPATH to So. America]
  - ➡ NSF & LIS Proposals: US and EU to South America
- Make direct contacts, arrange discussions with gov't officials
  - ➡ ICFA SCIC is prepared to participate
- Help Start, or Get Support for Workshops on Networks (& Grids)
  - ➡ Discuss & Create opportunities
  - ➡ Encourage, help form funded programs
- Help form Regional support & training groups (requires funding)

AMPATH | noc

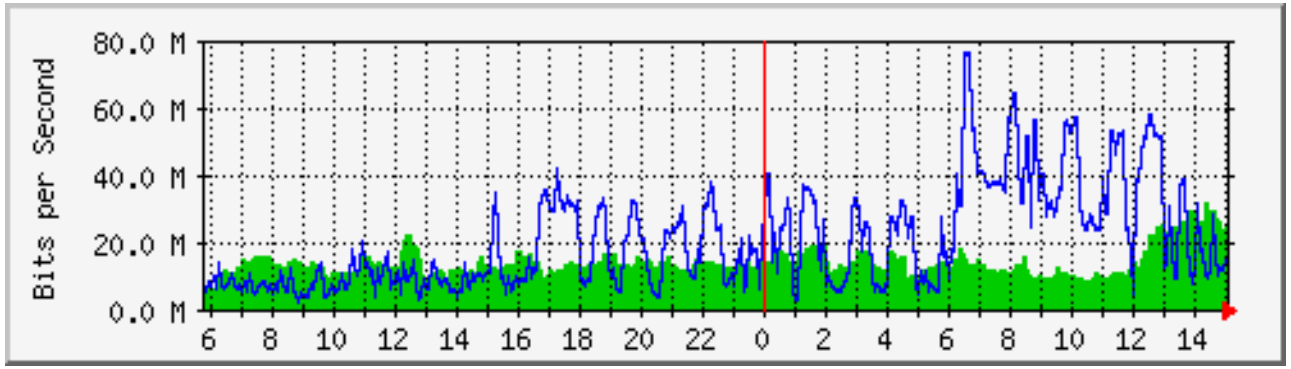
Abilene NOC | Euro-Link NOC | MIRnet NOC | STAR TAP NOC | TransPAC NOC | AMPATH Home

RNP Brazil





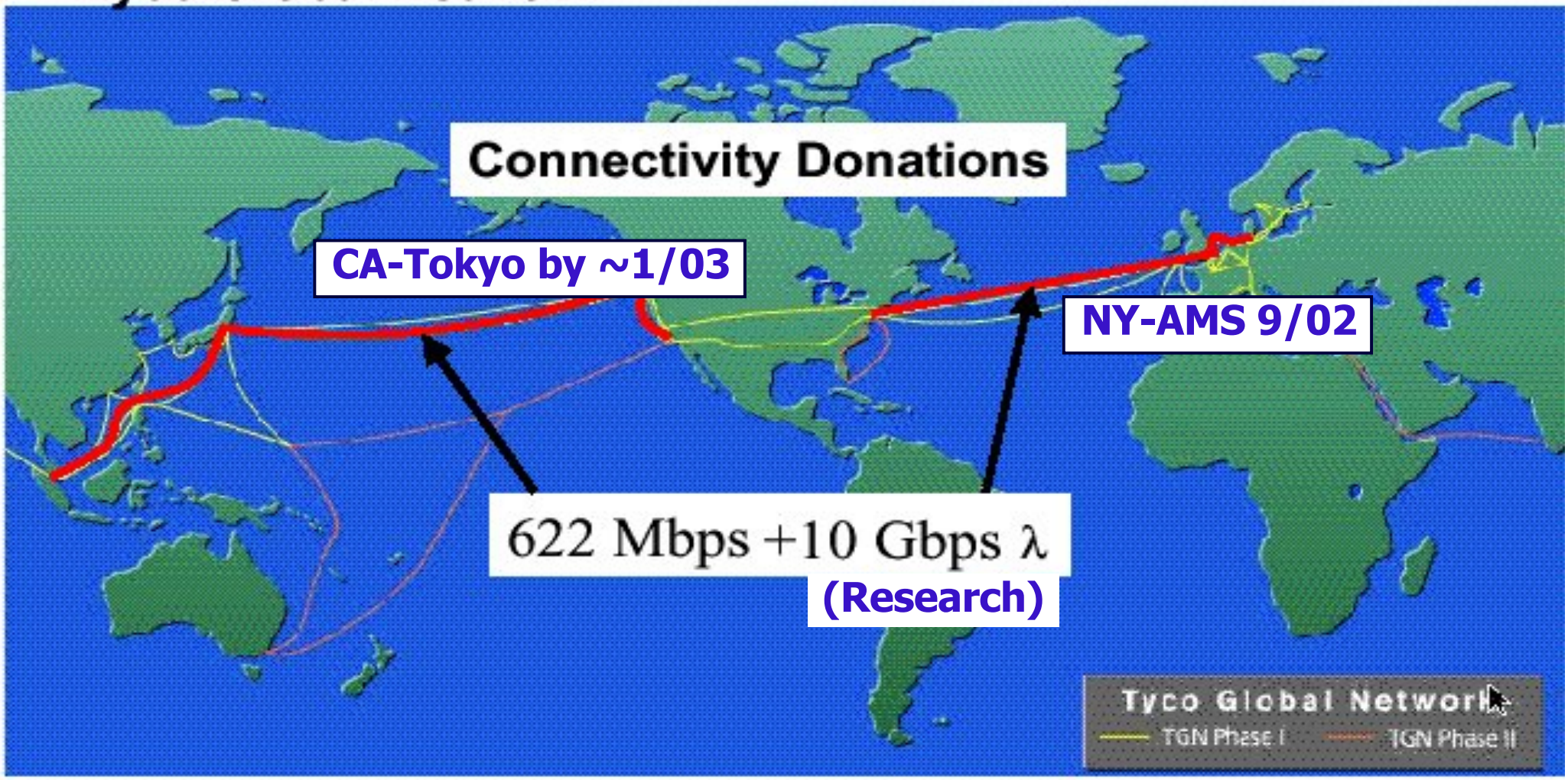
FIU Miami from So. America



Note: Auger (AG), ALMA (Chile), CMS-Tier1 (Brazil)

Internet Educational Equal Access Foundation

Tyco Global Network



Networks, Grids and HENP

- Current generation of 2.5-10 Gbps network backbones arrived in the last 15 Months in the US, Europe and Japan
  - ➡ Major transoceanic links also at 2.5 - 10 Gbps in 2003
  - ➡ Capability Increased ~4 Times, i.e. 2-3 Times Moore's
- Reliable high End-to-end Performance of network applications (large file transfers; Grids) is required. Achieving this requires:
  - ➡ End-to-end monitoring; a coherent approach
  - ➡ Getting high performance (TCP) toolkits in users' hands
- *Digital Divide: Network improvements are especially needed in SE Europe, So. America; SE Asia, and Africa:*
  - ➡ *Key Examples: India, Pakistan, China; Brazil; Romania*
- Removing Regional, Last Mile Bottlenecks and Compromises in Network Quality are now
  On the critical path, in all world regions
- Work in Concert with AMPATH, Internet2, Terena, APAN; DataTAG, the Grid projects and the Global Grid Forum