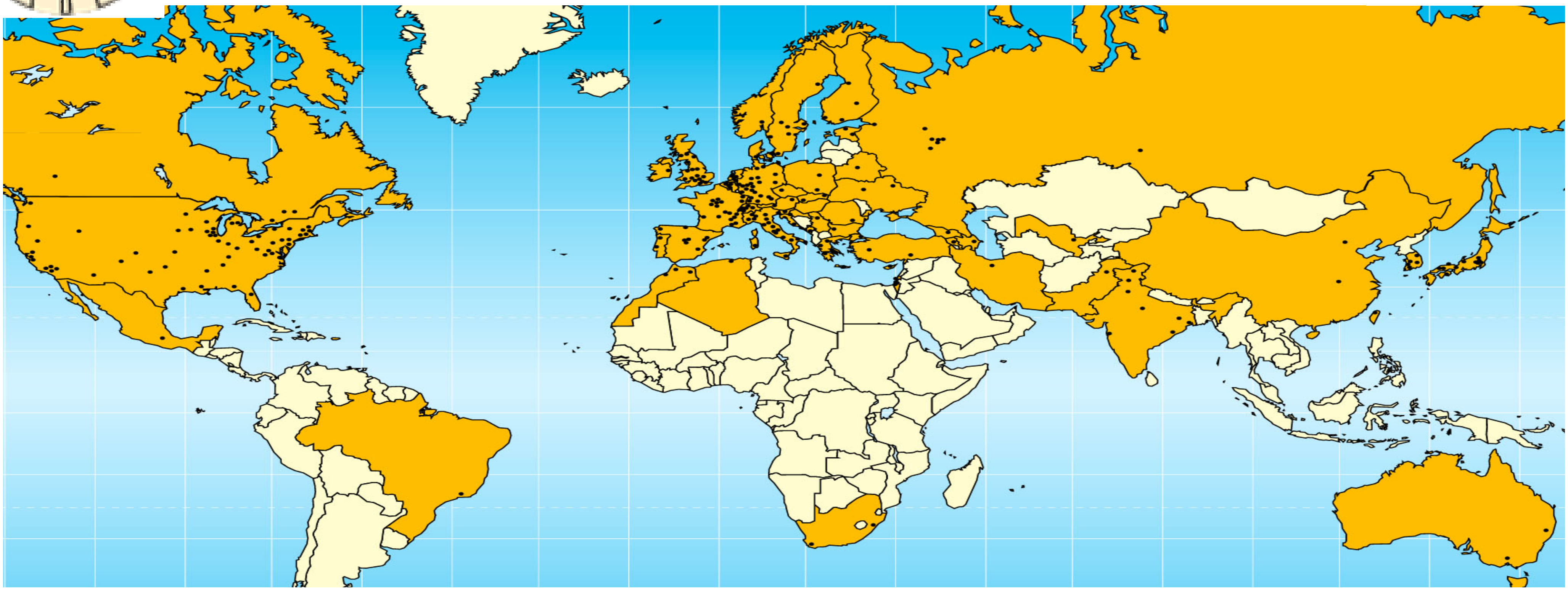
Virtual Organizations for Global Science: Network and Grid Drivers



Harvey B. Newman

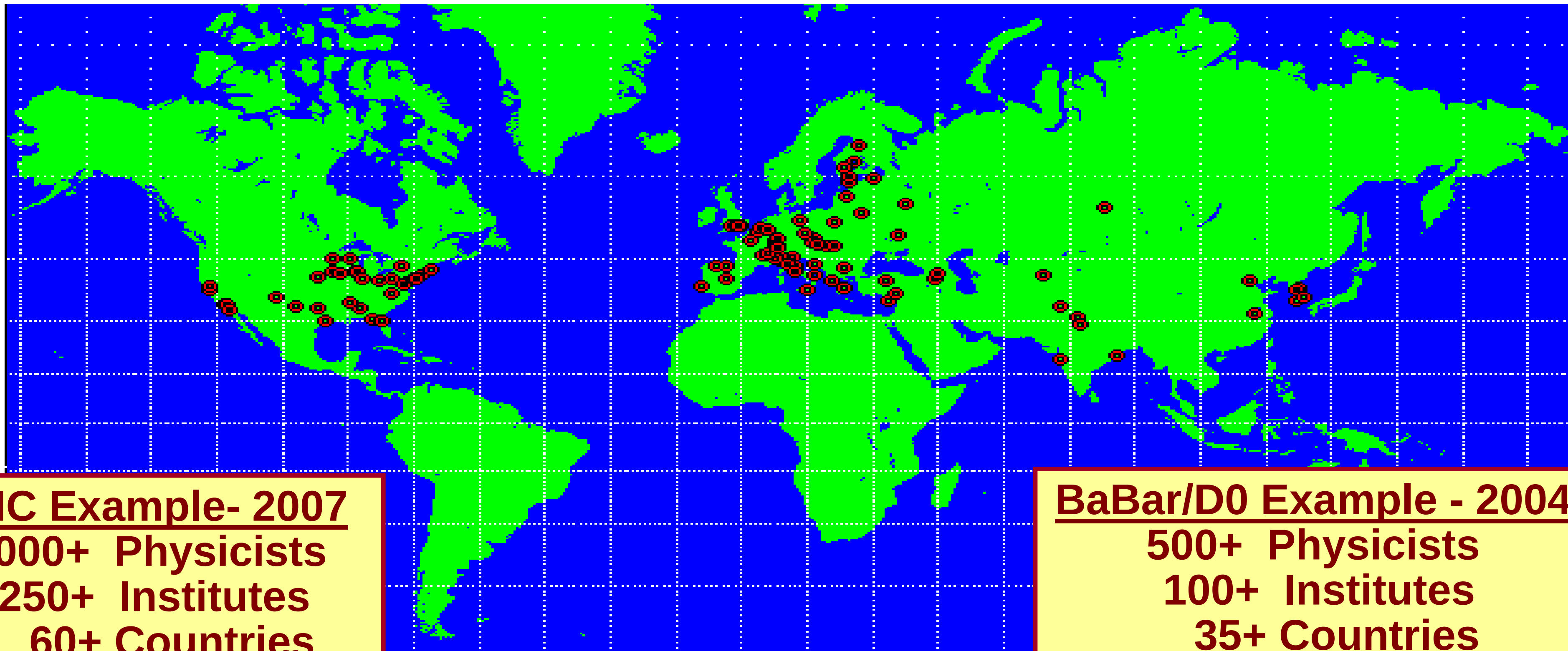
California Institute of Technology

JET Roadmap Workshop, Jefferson Lab

Newport News, VA, April 13, 2004

ICFA SCIC

Challenges for GlobalHENP Experiments



LHC Example- 2007
5000+ Physicists
250+ Institutes
60+ Countries

BaBar/D0 Example - 2004
500+ Physicists
100+ Institutes
35+ Countries

Major Challenges (Shared with ITER and Other Projects)

- Worldwide Communication and Collaboration
- Managing Globally Distributed Computing & Data Resources
- Cooperative Software Development and Data Analysis

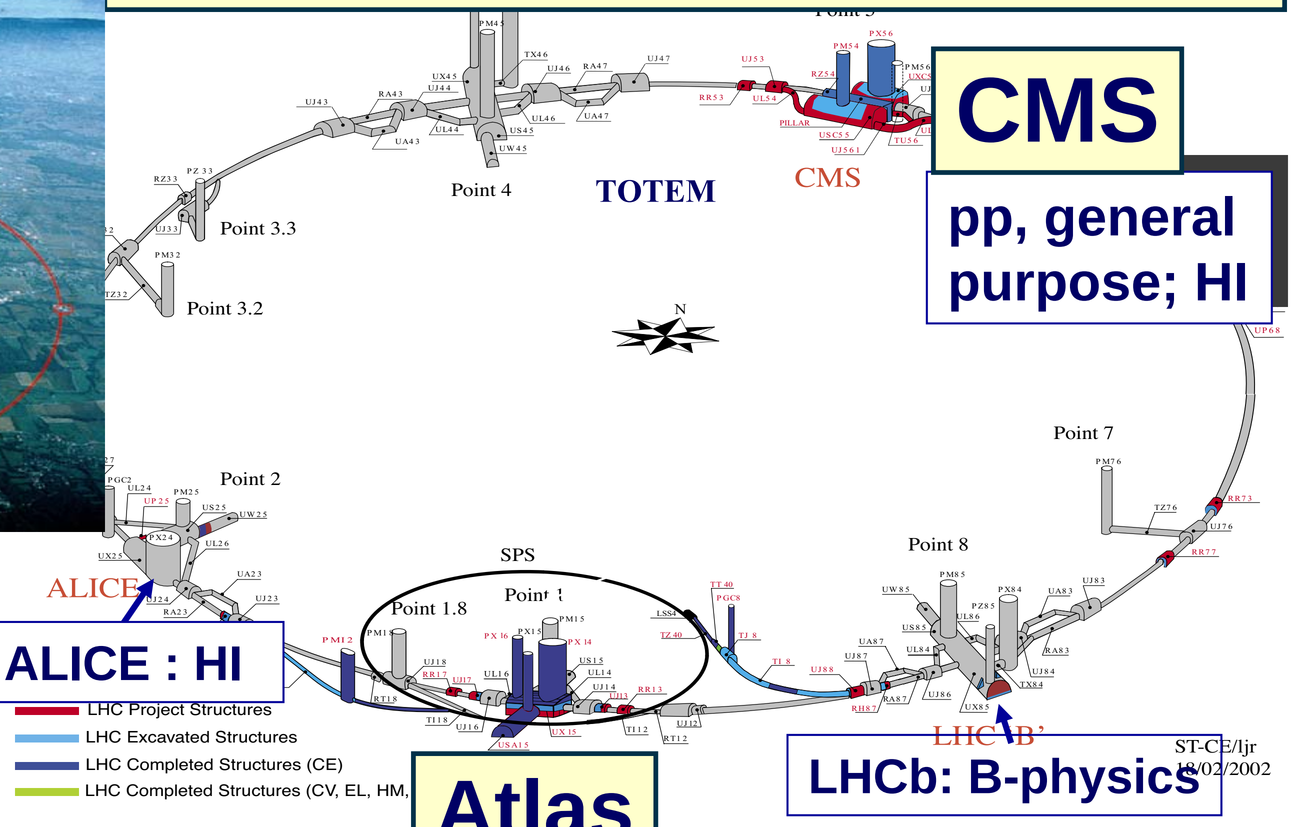
Challenges of Next Generation Science in the Information Age

Petabytes of complex data explored and analyzed by 1000s of globally dispersed scientists, in hundreds of teams

- **Flagship Applications**
  - **High Energy & Nuclear Physics, AstroPhysics Sky Surveys:** TByte to PByte "block" transfers at 1-10+ Gbps
  - **eVLBI:** Many real time data streams at 1-10 Gbps
  - **Fusion Energy:** Time Critical Burst-Data Distribution; Distributed Plasma Simulations, Visualization and Analysis; Preparations for Fusion Energy Experiment
  - **BioInformatics, Clinical Imaging:** GByte images on demand
- **Provide results with rapid turnaround, coordinating large but limited computing and data handling resources, over networks of varying capability in different world regions**
- **Advanced integrated applications, such as Data Grids, rely on seamless operation of our LANs and WANs**
  - **With reliable, quantifiable high performance**

Large Hadron Collider (LHC)
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

ALICE : HI

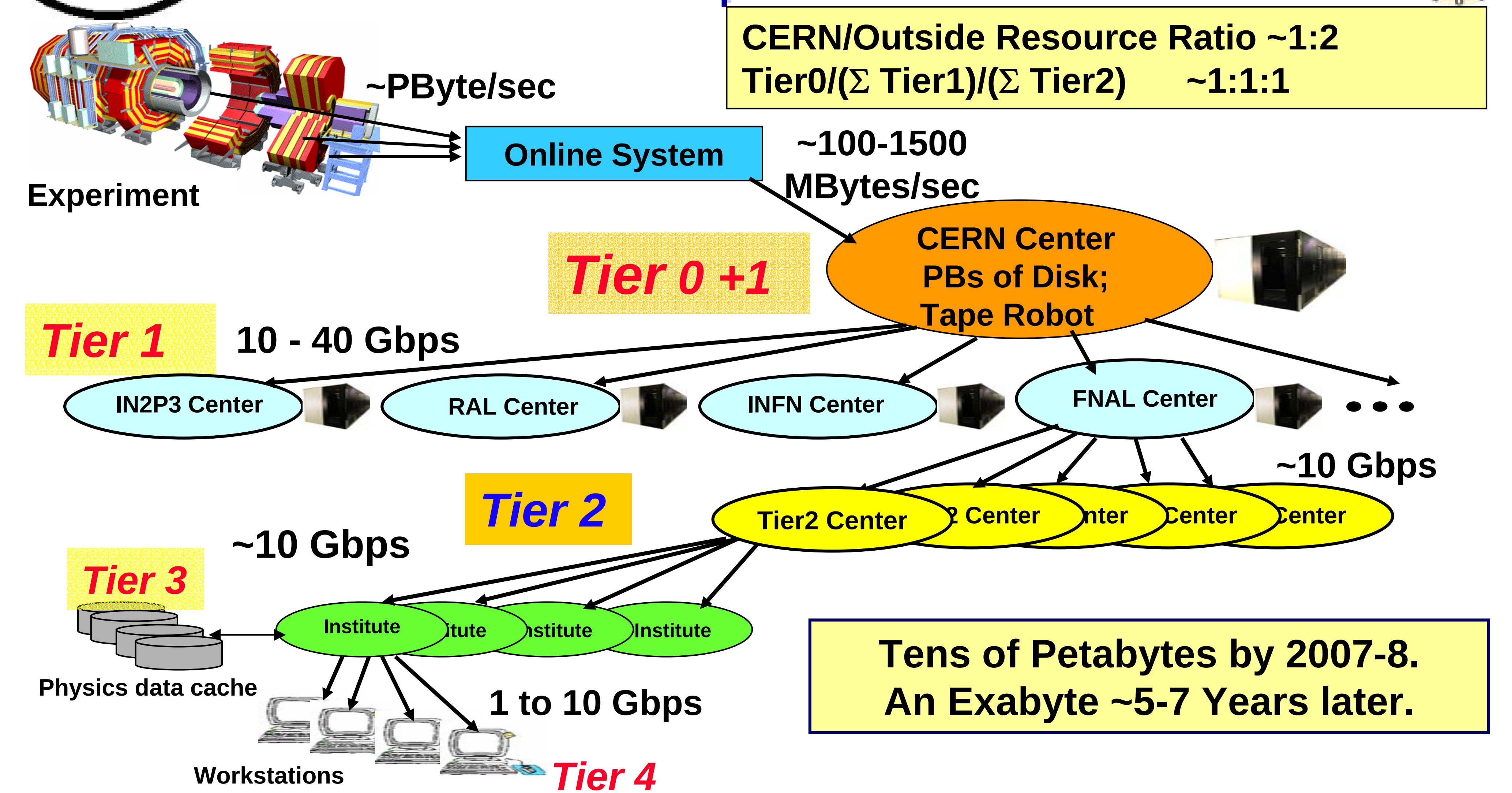
Atlas

LHCb: B-physics

First Beams: April 2007
Physics Runs: from Summer 2007

Higgs, SUSY, QG Plasma, CP Violation, *… the Unexpected*

LHC Data Grid Hierarchy: Developed at Caltech



Emerging Vision: A Richly Structured, Global Dynamic System

ICFA and Global Networks for Collaborative Science

- ***National and International Networks, with sufficient (rapidly increasing) capacity and seamless end-to-end capability, are essential for***
  - ***The daily conduct of collaborative work in both experiment and theory***
  - ***Experiment development & construction on a global scale***
  - ***Grid systems supporting analysis involving physicists in all world regions***
  - ***The conception, design and implementation of next generation facilities as "global networks"***
- ***"Collaborations on this scale would never have been attempted, if they could not rely on excellent networks"***

ICFA Standing Committee on Interregional Connectivity (SCIC)

- Created by ICFA in July 1998 in Vancouver
- <u>CHARGE:</u>

  Make recommendations to ICFA concerning the connectivity between *the Americas*, Asia and Europe
  - 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
- Representatives: Major labs, ECFA, ACFA, North and South American Physics Community

SCIC in 2003-2004
http://cern.ch/icfa-scic

- **Strong Focus on the Digital Divide Since 2002**

Three 2004 Reports; Presented to ICFA Feb. 13

- ***Main Report: "Networking for HENP"*** **[H. Newman et al.]**
  - **Includes Brief Updates on Monitoring, the Digital Divide and Advanced Technologies [*]**
  - **A World Network Overview (with 27 Appendices):**
    Status and Plans for the Next Few Years of National and Regional Networks, and Optical Network Initiatives
- ***Monitoring Working Group Report*** **[L. Cottrell]**
- ***Digital Divide in Russia*** **[V. Ilyin]**

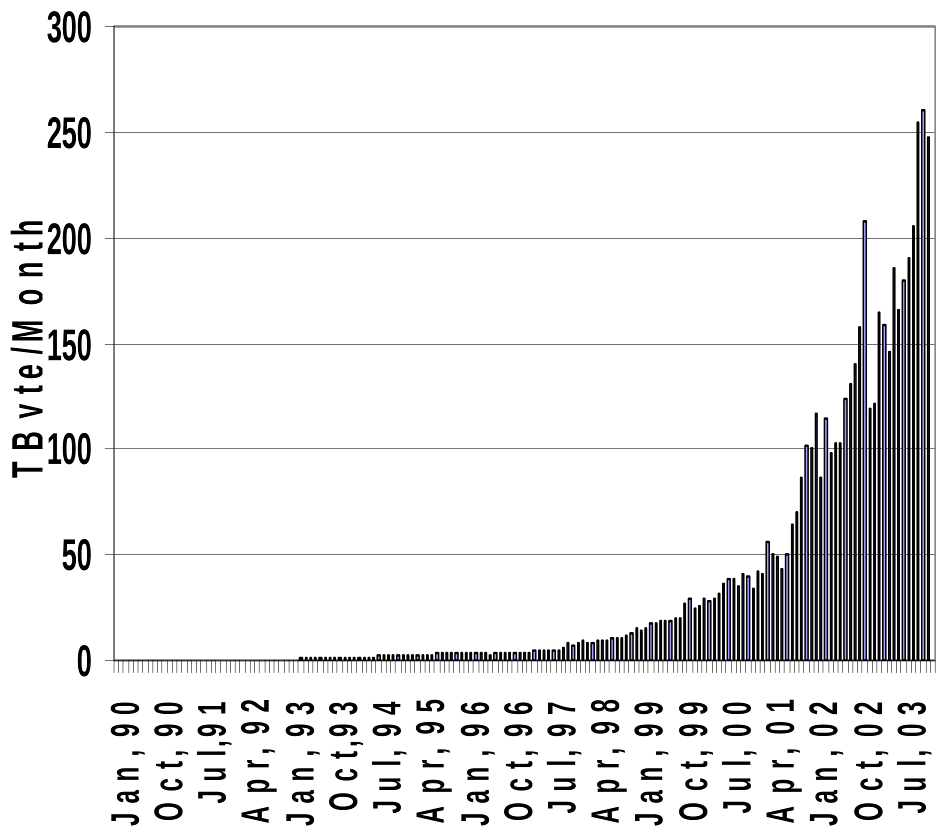
[*] Also See the 2003 SCIC Reports of the Advanced Technologies and Digital Divide Working Groups

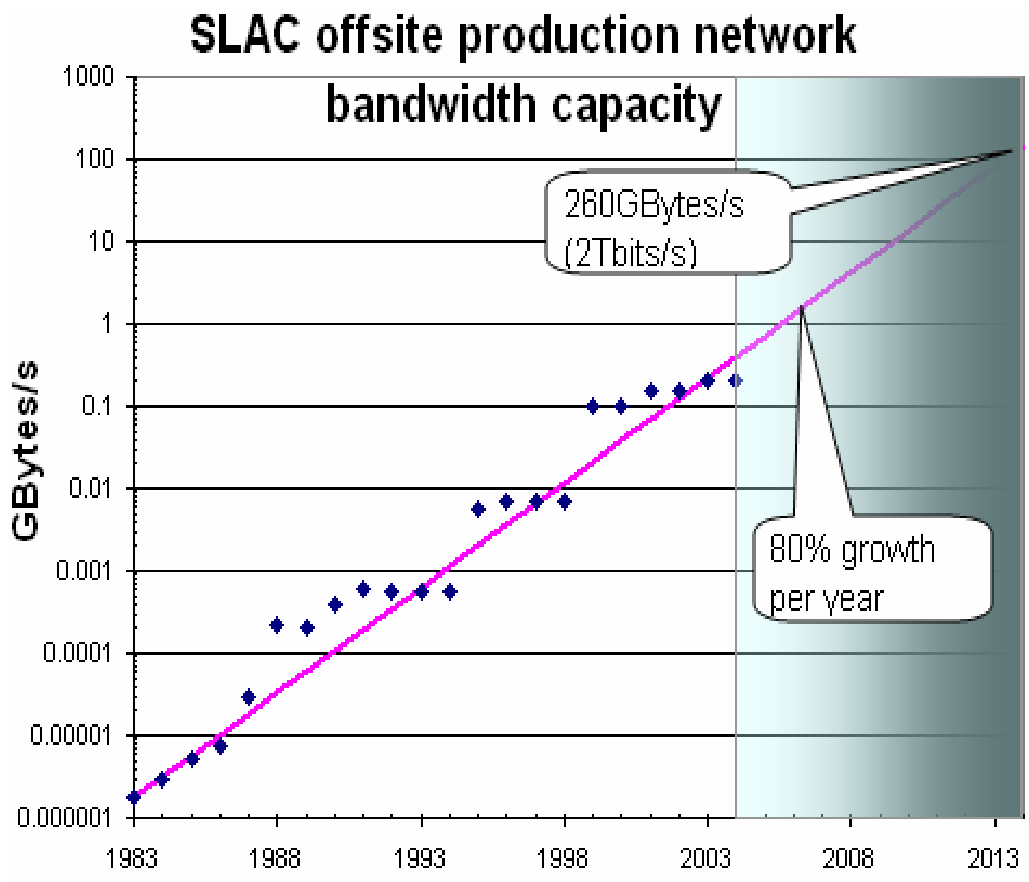
ICFA Report: Networks for HENP
General Conclusions (1)

- Bandwidth Usage Continues to Grow by 80-100% Per Year
- Current generation of 2.5-10 Gbps backbones and major Int'l links used by HENP arrived in the last 2 Years [US+Europe+Japan+Korea]
  - Capability Increased from ~4 to several hundred times, i.e. much faster than Moore's Law
  - This is a direct result of the continued precipitous fall of network prices for 2.5 or 10 Gbps links in these regions
- Technological progress may drive BW higher, unit price lower
  - More wavelengths on a fiber; Cheap, widespread Gbit Ethernet
- Grids may accelerate this growth, and the demand for seamless high performance
- *Some regions are moving to owned or leased dark fiber*
- *The rapid rate of progress is confined mostly to the US, Europe, Japan , Korea, and the major TransAtlantic and Pacific routes*
  - *This may worsen the problem of the Digital Divide*

History of Bandwidth Usage – One Large Network; One Large Research Site

ESnet Accepted Traffic 1/90 – 1/04
Exponential Growth Since '92; Annual Rate Increased from 1.7 to 2.0X Per Year In the Last 5 Years





SLAC Traffic ~300 Mbps; ESnet Limit Growth in Steps: ~ 10X/4 Years Projected: ~2 Terabits/s by ~2014

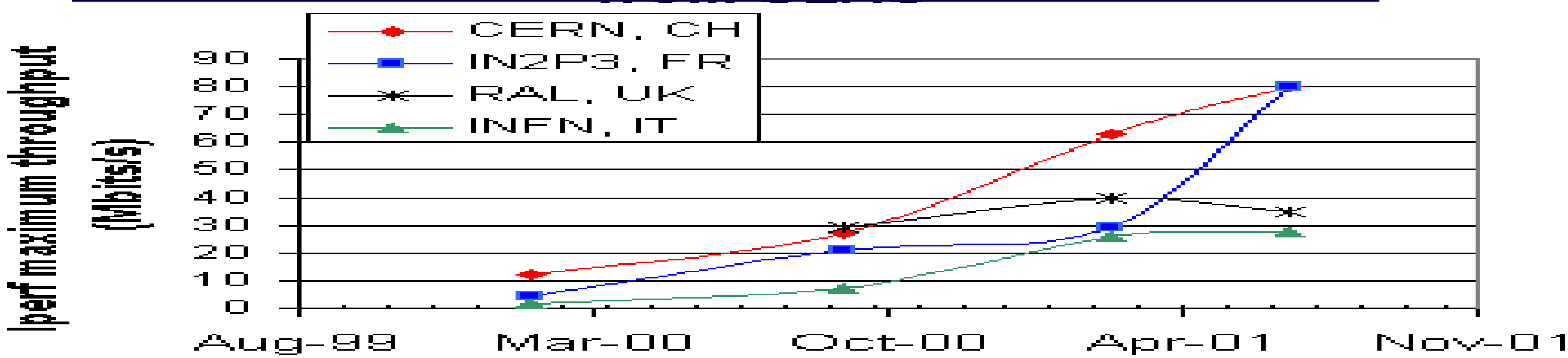
Bandwidth Growth of Int'l HENP Networks (US-CERN Example)

- ***Rate of Progress >> Moore's Law. (US-CERN Example)***

9.6 kbps Analog	(1985)	
64-256 kbps Digital	(1989 - 1994)	[X 7 – 27]
1.5 Mbps Shared	(1990-3; IBM)	[X 160]
2 -4 Mbps	(1996-1998)	[X 200-400]
12-20 Mbps	(1999-2000)	[X 1.2k-2k]
155-310 Mbps	(2001-2)	[X 16k – 32k]
622 Mbps	(2002-3)	[X 65k]
2.5 Gbps λ	(2003-4)	[X 250k]
10 Gbps λ	(2005)	[X 1M]

- ***A factor of ~1M over a period of 1985-2005 (a factor of ~5k during 1995-2005)***
- ***HENP has become a leading applications driver, and also a co-developer of global networks;***

HEP is Learning How to Use Gbps Networks Fully: Factor of ~500 Gain in Max. Sustained TCP Thruput in 4 Years, On Some US+Transoceanic Routes



- **9/01 105 Mbps 30 Streams: SLAC-IN2P3; 102 Mbps 1 Stream CIT-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**
- **9/02 850, 1350, 1900 Mbps Chicago-CERN 1,2,3 GbE Streams, 2.5G Link**
- **11/02 [LSR] 930 Mbps in 1 Stream California-CERN, and California-AMS**
- ***FAST TCP* 9.4 Gbps in 10 Flows California-Chicago**
- **2/03 [LSR] 2.38 Gbps in 1 Stream California-Geneva (99% Link Use)**
- **5/03 [LSR] 0.94 Gbps IPv6 in 1 Stream Chicago- Geneva**
- **TW & SC2003 [LSR]: 5.65 Gbps (IPv4), 4.0 Gbps (IPv6) GVA-PHX (11 kkm)**
- ***3/04 [LSR] 6.25 Gbps (IPv4) in 8 Streams LA-CERN***

Transatlantic Ultraspeed TCP Transfers
Throughput Achieved: X50 in 2 years

Terabyte Transfers by the Caltech-CERN Team:

- ❒ *Across Abilene (Internet2) Chicago-LA, Sharing with normal network traffic*

- ◆ Oct 15: 5.64 Gbps IPv4 Palexpo-L.A. (10.9 kkm)
  - ❒ Peaceful Coexistence with a Joint Internet2-Telecom World VRVS Videoconference
- ◆ Nov 18: 4.00 Gbps IPv6 Geneva-Phoenix (11.5 kkm)
- ◆ March 2004: 6.25 Gbps in 8 Streams (S2IO, Windows)

Nov 19: 23+ Gbps TCP: Caltech, SLAC, CERN, LANL, UvA, Manchester

Juniper, HP Level(3) Telehouse

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; DOE Science Network Roadmap: Compatible

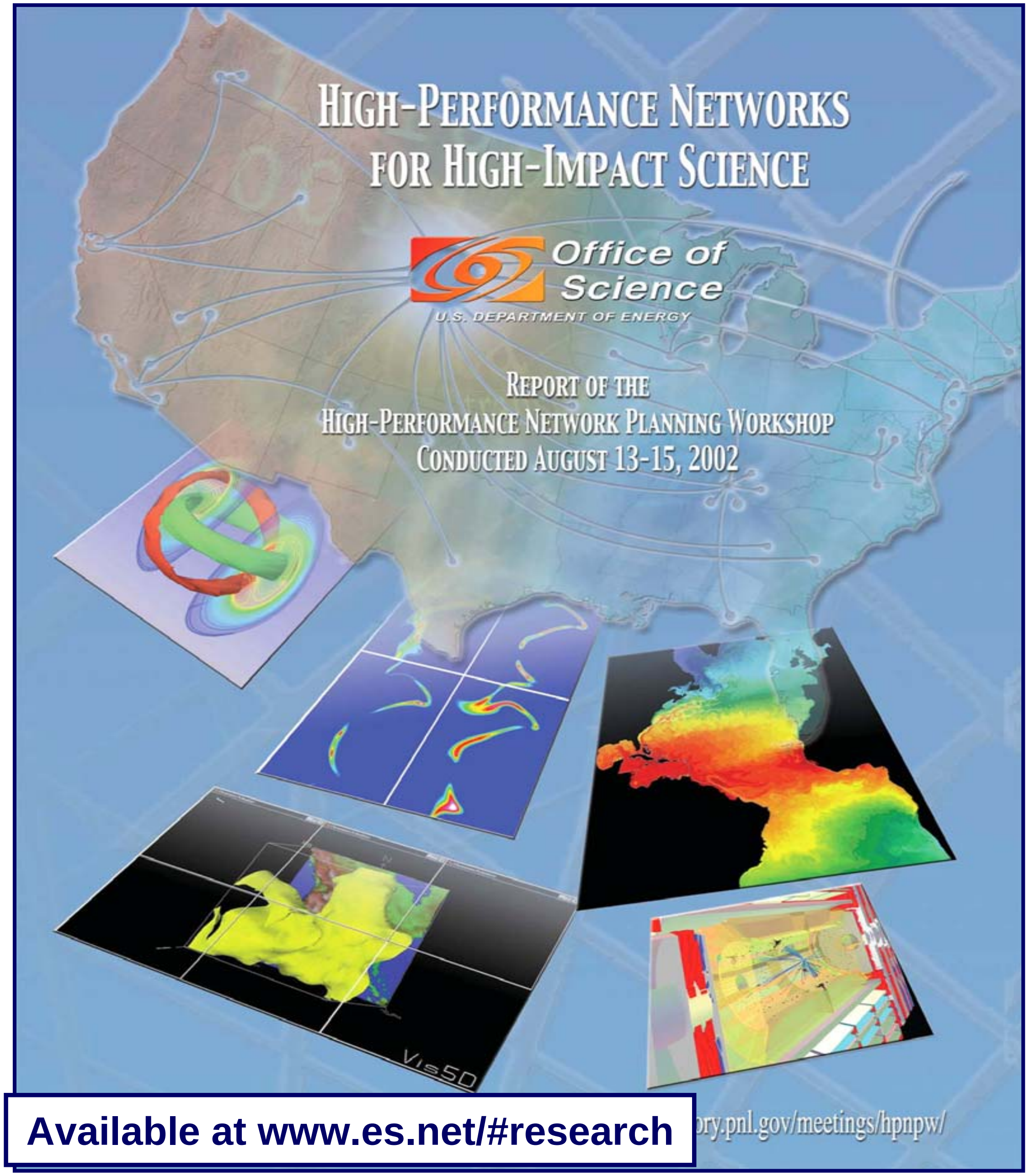
HENP Lambda Grids: Fibers for Physics

- **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 ~2-4 years; and Terabit Switching within 5-8 years would enable "Petascale Grids with Terabyte transactions", to fully realize the discovery potential of major HENP programs, as well as other data-intensive research.**

ESnet: Driven by the Needs of DOE Science



Available at www.es.net/#research

High Perf. Network Planning Workshop, August 13-15, 2002

Focus on Science Requirements that Drive

- ***Advanced Network Infrastructure***
- ***Middleware Research***
- ***Network Research***
- ***Network Governance Model***

Organized by Office of Science

Mary Anne Scott, Chair
Dave Bader
Steve Eckstrand
Marvin Frazier
Dale Koelling
Vicky White

Panel Chairs

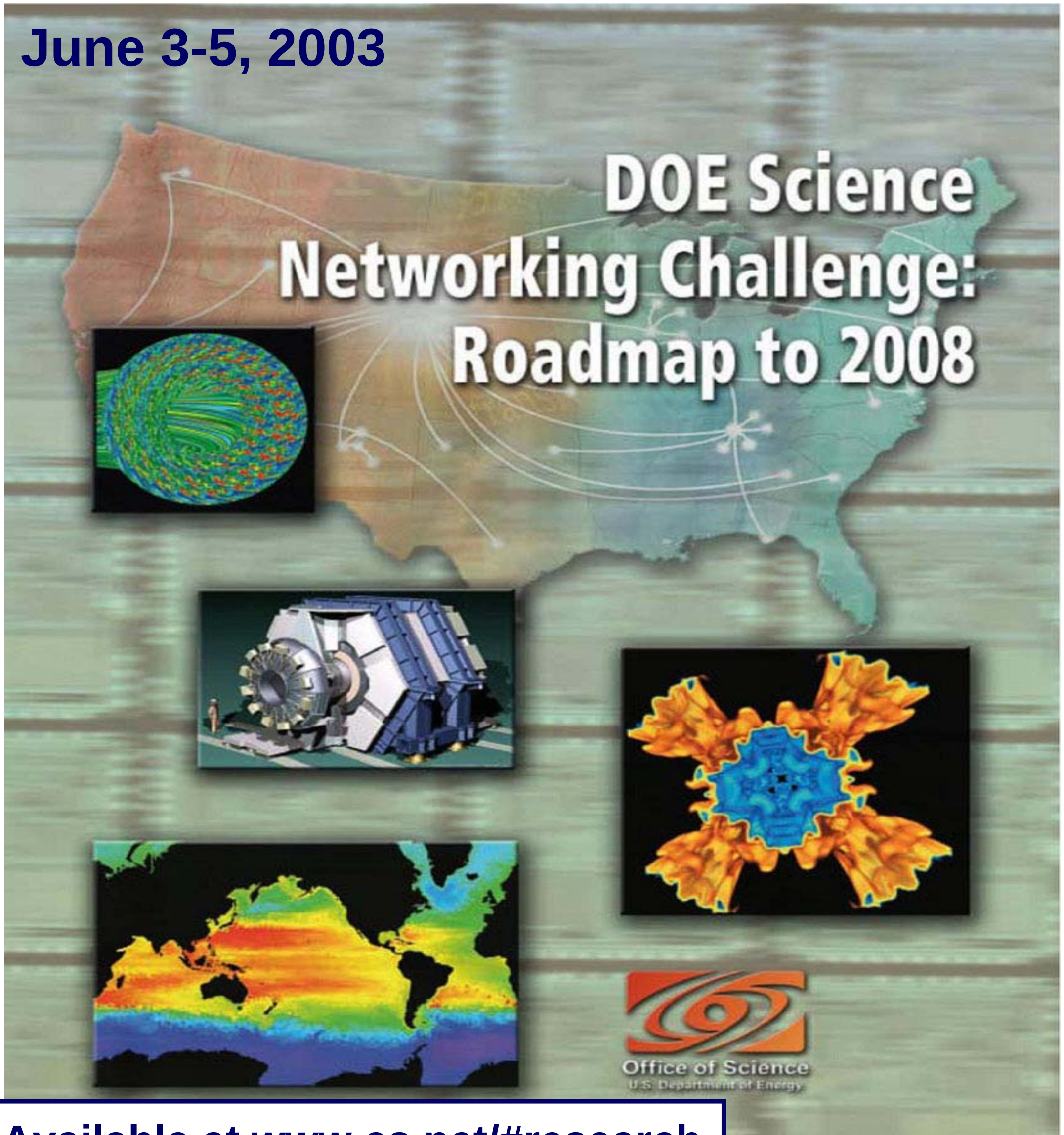
Bill Johnston and Mike Wilde
Rick Stevens
Ian Foster and Dennis Gannon
Linda Winkler and Brian Tierney
Sandy Merola and Charlie Catlett

Evolving Quantitative Science Requirements for Networks

Science Areas	Today *End2End* Throughput	5 years End2End Throughput	5-10 Years End2End Throughput	Remarks
High Energy Physics	0.5 Gb/s	100 Gb/s	1000 Gb/s	High bulk throughput
Climate (Data & Computation)	0.5 Gb/s	160-200 Gb/s	N x 1000 Gb/s	High bulk throughput
SNS NanoScience	Not yet started	1 Gb/s	1000 Gb/s + QoS for Control Channel	Remote control and time critical throughput
Fusion Energy	0.066 Gb/s (500 MB/s burst)	0.198 Gb/s (500MB/ 20 sec. burst)	N x 1000 Gb/s	Time critical throughput
Astrophysics	0.013 Gb/s (1 TByte/week)	N*N multicast	1000 Gb/s	Computational steering and collaborations
Genomics Data & Computation	0.091 Gb/s (1 TBy/day)	100s of users	1000 Gb/s + QoS for Control Channel	High throughput and steering

New Strategic Directions to Address the Network Needs of DOE Science

June 3-5, 2003



Available at www.es.net/#research

Focus on what was needed to achieve the science driven network requirements of the previous workshop

High Impact Network Roadmap

- **Year 3: Switched Lambdas Between MANs at 10 Gbps**
- **Year 4: Nat'l E2E Testbeds with Full 20 Gbps Use**
- **Year 5 (2008): Start Native 40 Gbps Services**

Organized by the ESSC

Workshop Chair
Roy Whitney, JLAB
Report Editors
Roy Whitney, JLAB
Larry Price, ANL

GNEW 2004

Grid and Network Workshop at CERN March 15-16, 2004

CONCLUDING STATEMENT

"Following the 1st International Grid Networking Workshop (GNEW2004) that was held at CERN and co-organized by CERN/DataTAG, DANTE, ESnet, Internet2 & TERENA, there is a wide consensus that hybrid network services capable of offering both packet- and circuit/lambda-switching as well as highly advanced performance measurements and a new generation of distributed system software, will be required in order to support emerging data intensive Grid applications, Such as High Energy Physics, Astrophysics, Climate and Supernova modeling, Genomics and Proteomics, requiring 10-100 Gbps and up over wide areas."

Transition beginning now to optical, multi-wavelength Community owned or leased fiber networks for R&E

National Lambda Rail (NLR)

15808 Terminal, Regen or OADM site

Fiber route

NLR

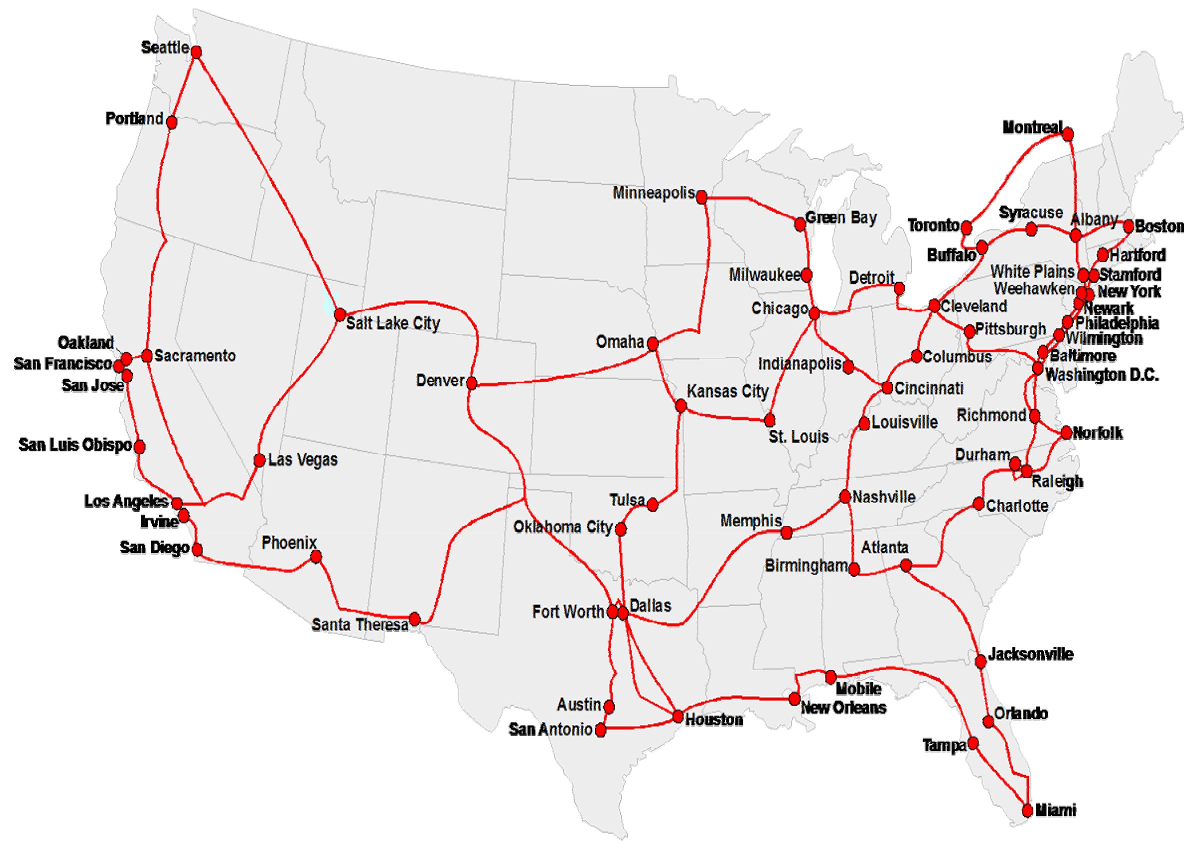
- Coming Up Now; Initially 4 10G Wavelengths
- Full Footprint by ~End 04
- Internet2 HOPI Initiative (w/HEP)
- To 40 10G Waves in Future

Complemented by Dark Fiber Initiatives in 18 US States

18 State Dark Fiber Initiatives In the U.S. (As of 3/04)

California (CALREN),
Colorado (FRGP/BRAN)
Connecticut Educ. Network,
Florida Lambda Rail,
Indiana (I-LIGHT),
Illinois (I-WIRE),
Md./DC/No. Virginia (MAX),
Michigan,
Minnesota,
NY + New England (NEREN),
N. Carolina (NC LambdaRail),
Ohio (Third Frontier Net)
Oregon,
Rhode Island (OSHEAN),
SURA Crossroads (SE U.S.),
Texas,
Utah,
Wisconsin

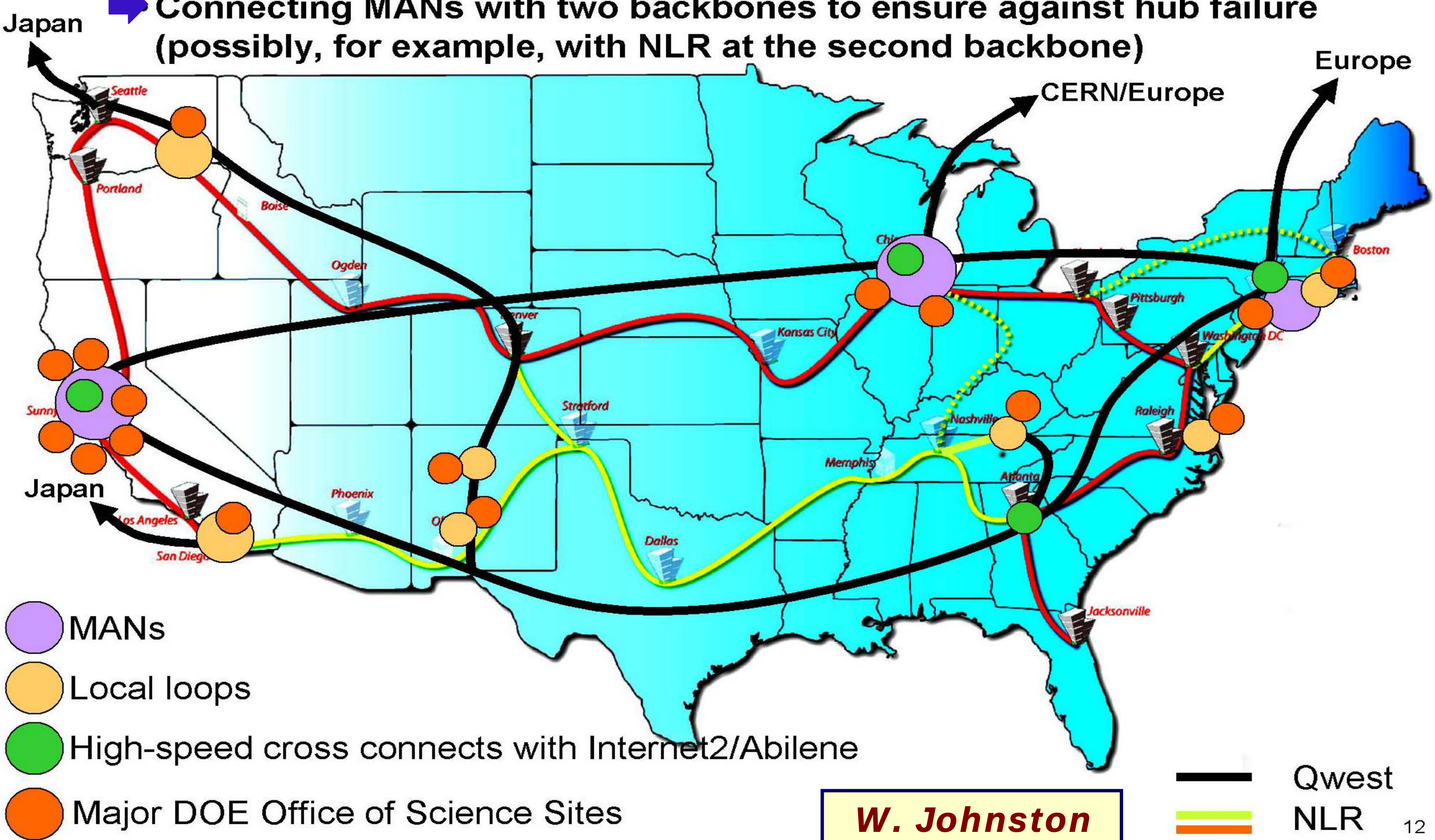
The Move to Dark Fiber is Spreading



FiberCO

Long-Term ESnet Connectivity Goal

- MANs for scalable bandwidth and redundant site access to backbone
- Connecting MANs with two backbones to ensure against hub failure (possibly, for example, with NLR at the second backbone)



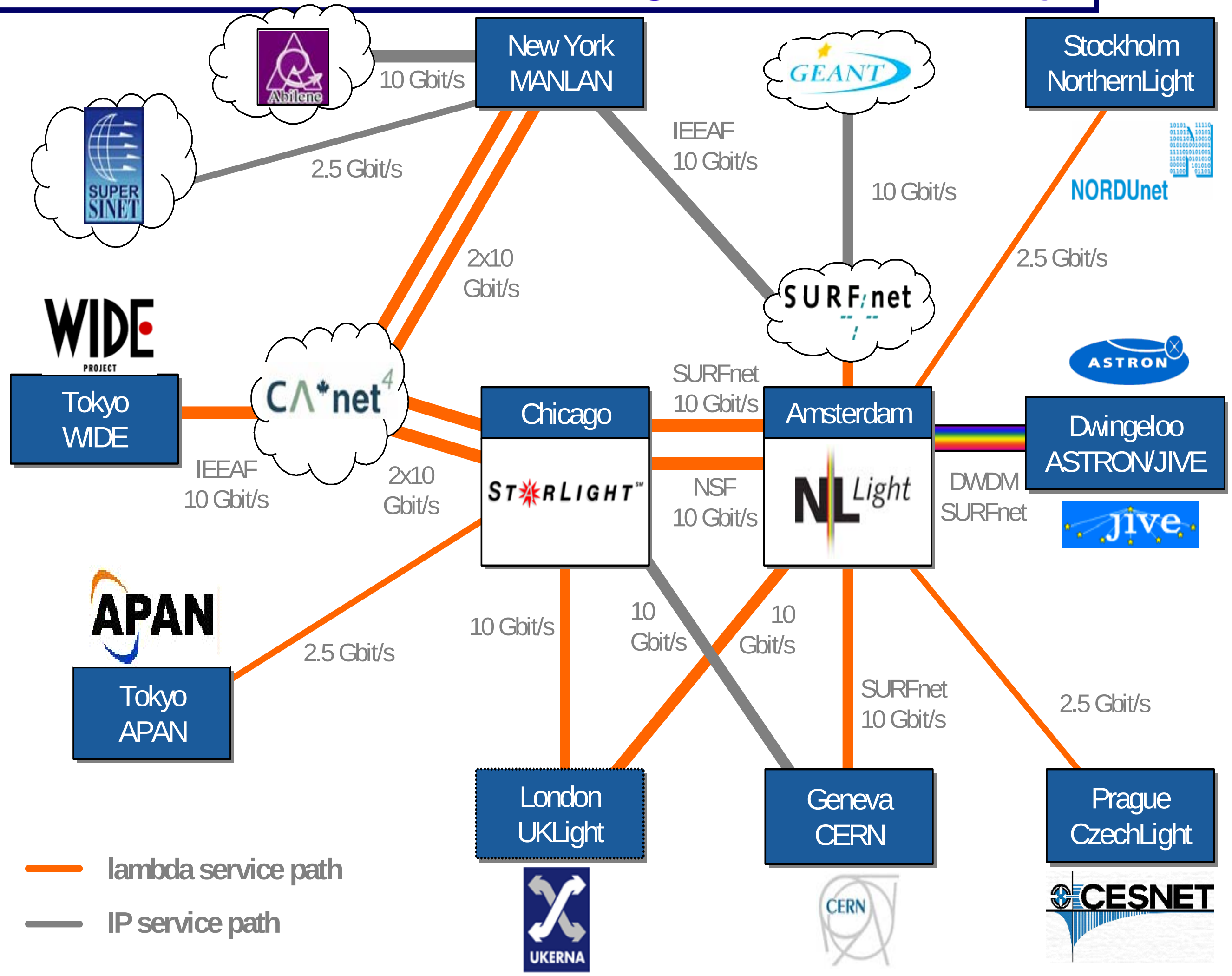
W. Johnston

ICFA SCIC

GLIF: Global Lambda Integrated Facility

"GLIF is a World Scale Lambda based Lab for Application & Middleware development, where Grid applications ride on dynamically configured networks based on optical wavelengths ...

GLIF will use the Lambda network to support data transport for the most demanding e-Science applications, concurrent with normal Internet paths for the remaining traffic mix."



10 Gbps Wavelengths For R&E Network Development Are Proliferating, Across Continents and Oceans

PROGRESS in SE Europe (Sk, Pl, Cz, Hu, ...)

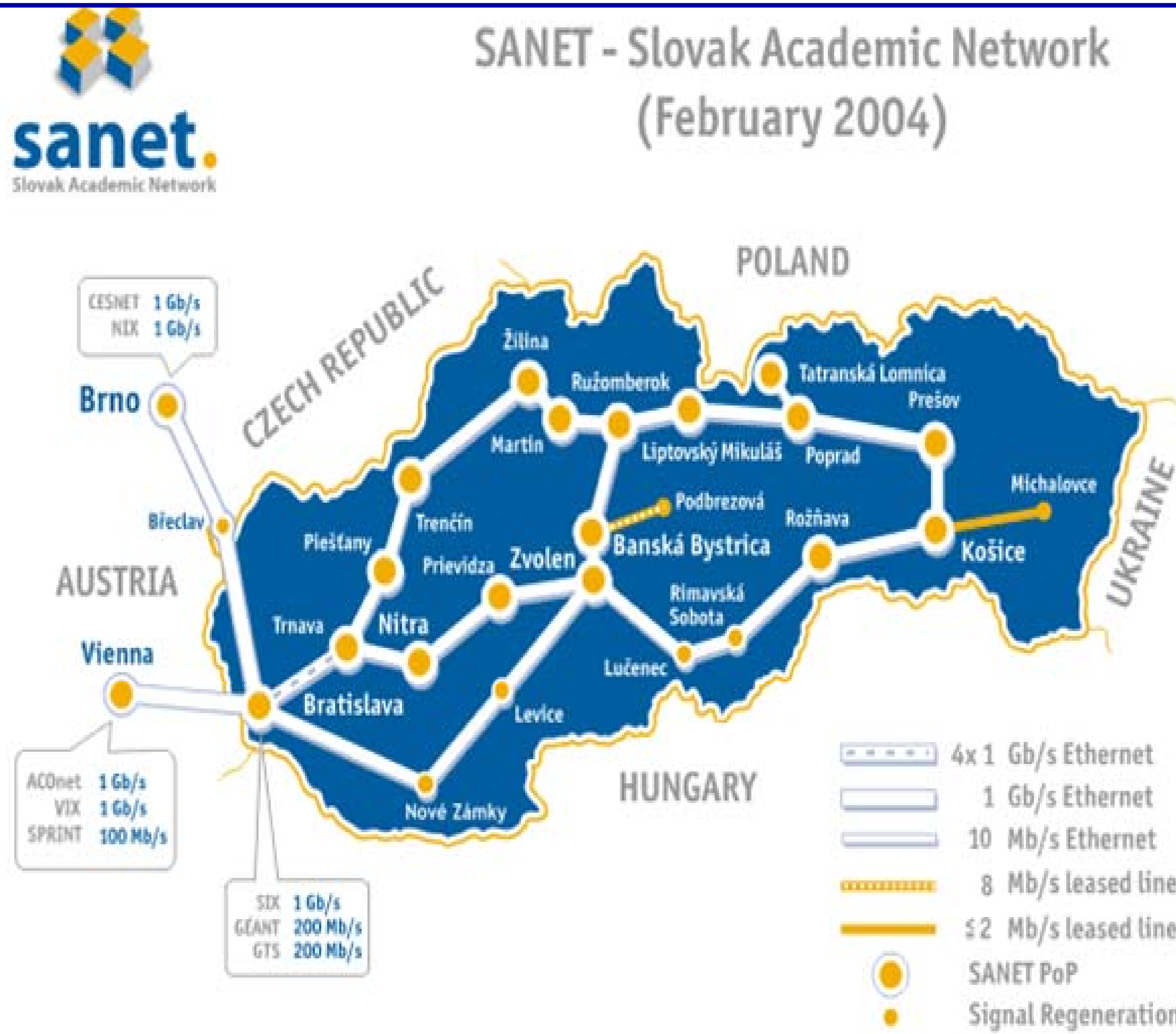
1660 km of Dark Fiber CWDM Links, up to 112 km.

1 to 4 Gbps (GbE)

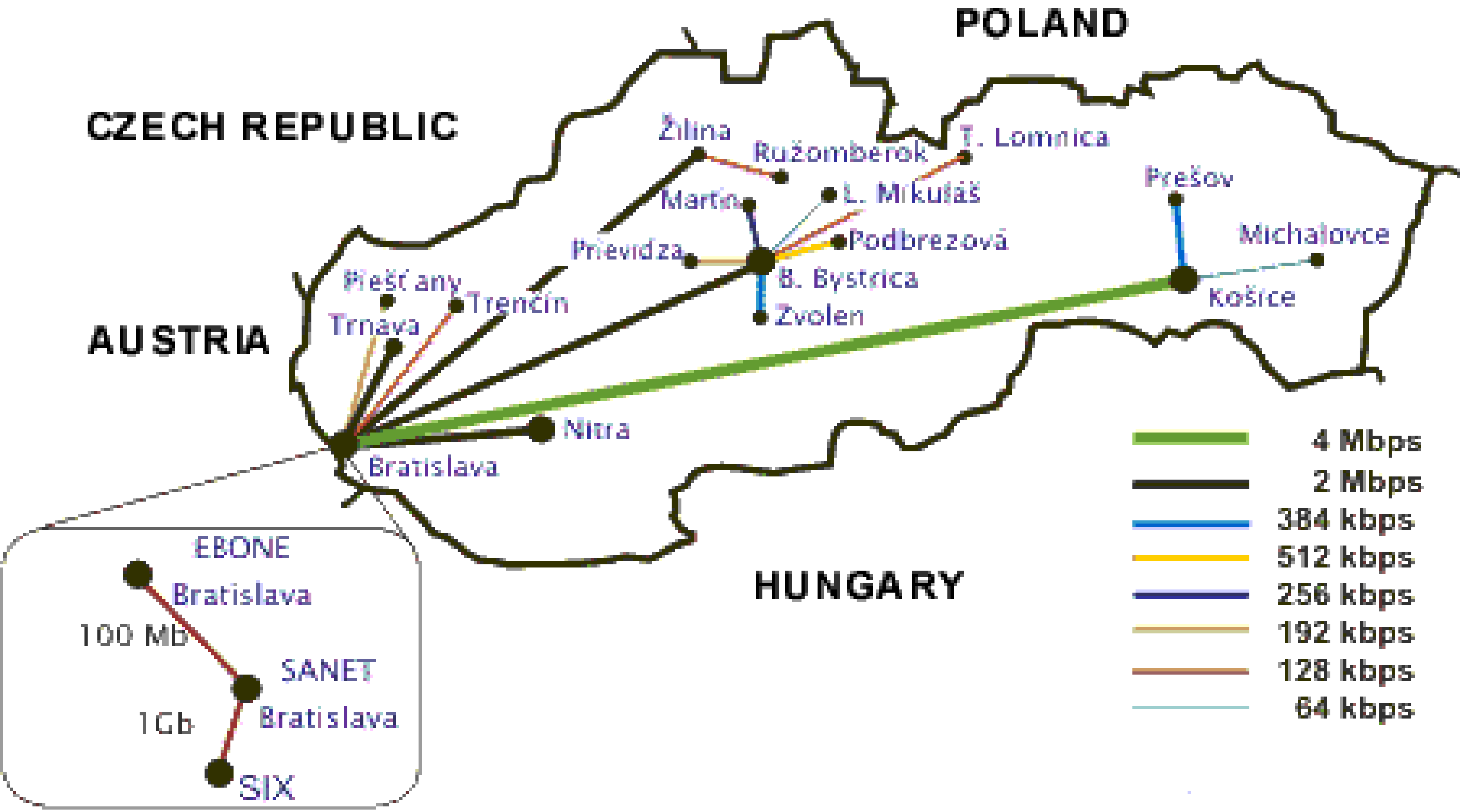
August 2002: First NREN in Europe to establish Int'l GbE Dark Fiber Link, to Austria

April 2003 to Czech Republic.

Planning 10 Gbps Backbone; dark fiber link to Poland this year.



SANET - Slovak Akademic Data Network
(January 2002)

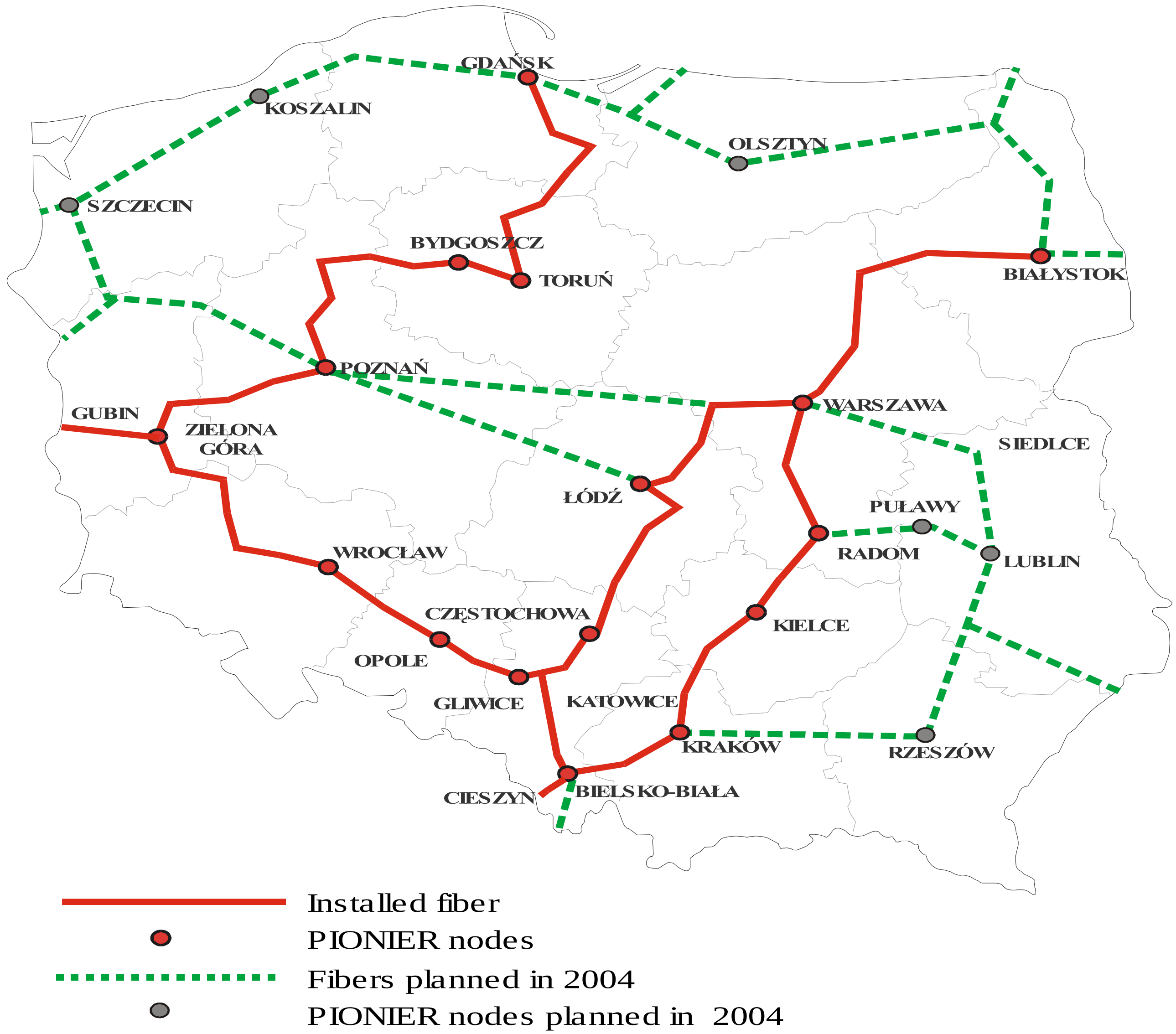


Dark Fiber in Eastern Europe
Poland: *PIONIER* Network

2650 km Fiber Connecting 16 MANs; 5200 km and 21 MANs by 2005

Support

- Computational Grids Domain-Specific Grids
- Digital Libraries
- Interactive TV
- Add'l Fibers for e-Regional Initiatives



HENP Data Grids, and Now Services-Oriented Grids

- The original Computational and Data Grid concepts are largely stateless, open systems
  - ➔ Analogous to the Web
- The classical Grid architecture had a number of implicit assumptions
  - ➔ The ability to locate and schedule suitable resources, within a tolerably short time (i.e. resource richness)
  - ➔ Short transactions with relatively simple failure modes
- HENP Grids are *Data Intensive & Resource-Constrained*
  - ➔ Resource usage governed by local and global policies
  - ➔ Long transactions; some long queues
- HENP ➔Stateful, End-to-end Monitored and Tracked Paradigm
  - ➔ Adopted in OGSA [Now WS Resource Framework]

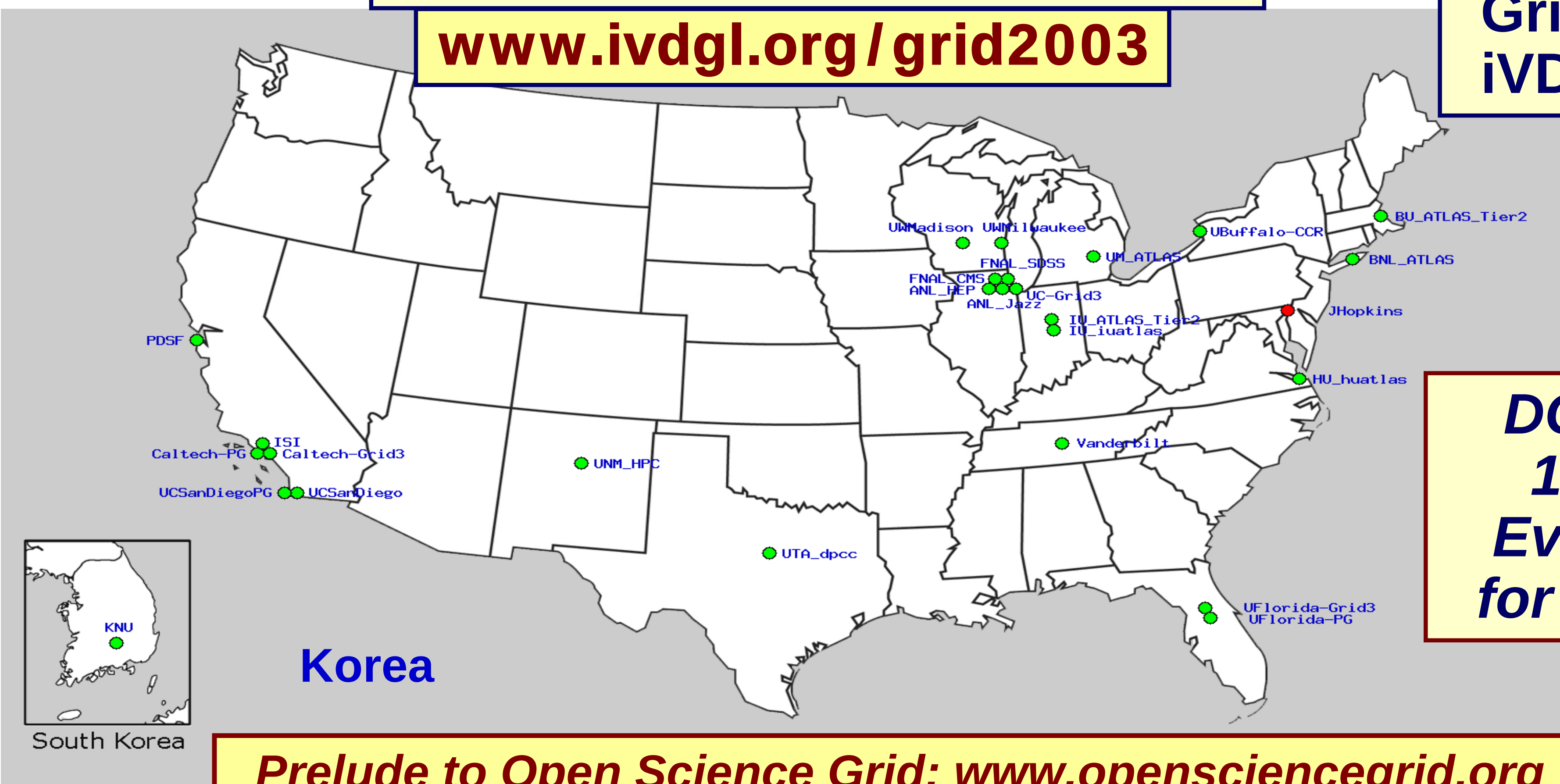
➔ *Analysis:* 1000s of users competing for resources at dozens of sites: complex scheduling, management, security

Grid2003: An Operational Production Grid, Since October 2003

- **27 sites (U.S., Korea)**
- **2300-2800 CPUs**
- **700-1100 Concurrent Jobs**

www.ivdgl.org/grid2003

Trillium:
PPDG
GriPhyN
iVDGL



DC04: 14M Events for CMS

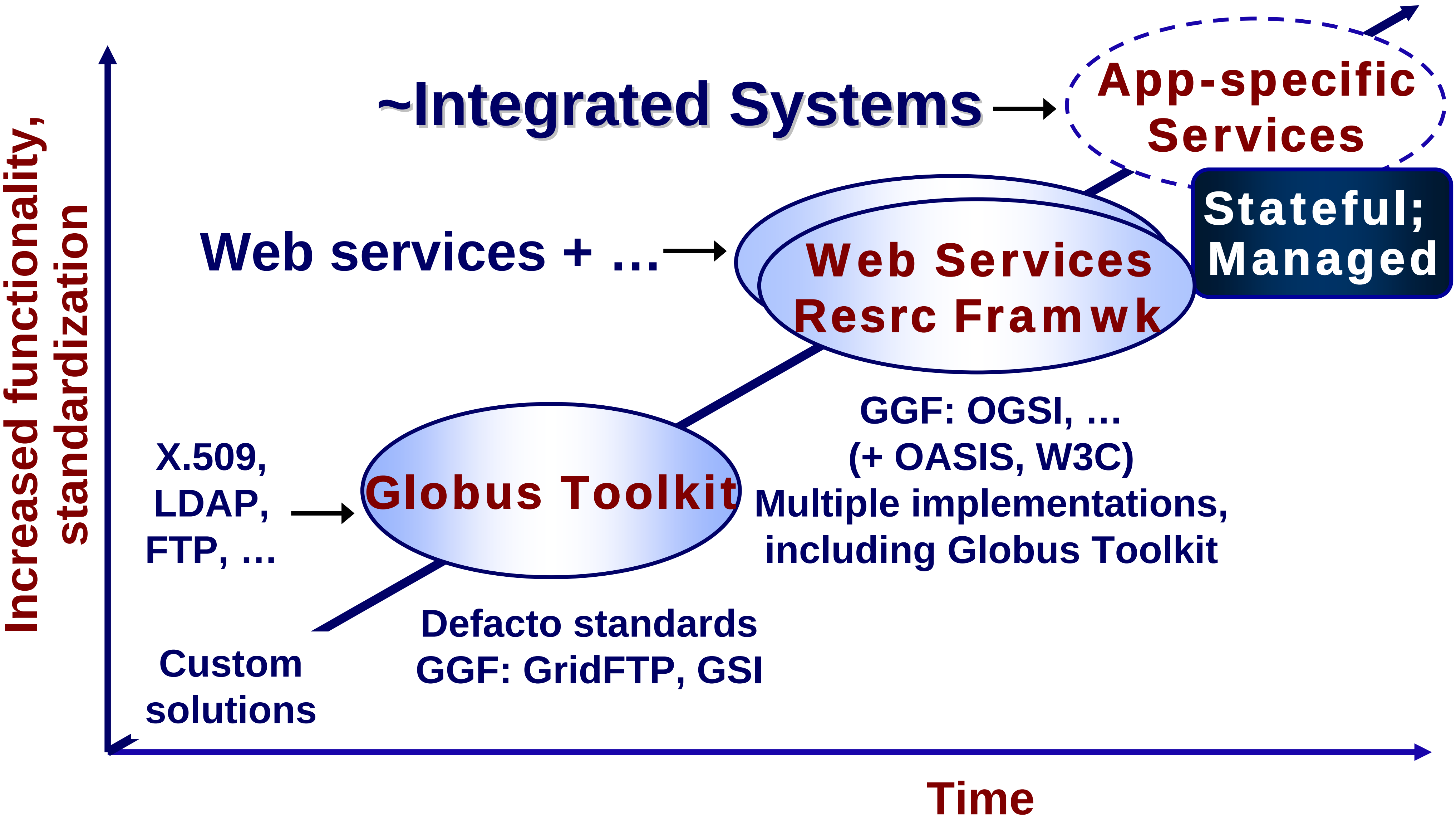
Prelude to Open Science Grid: www.opensciencegrid.org

The Open Science Grid
http://www.opensciengrid.org

The *Open Science Grid* will

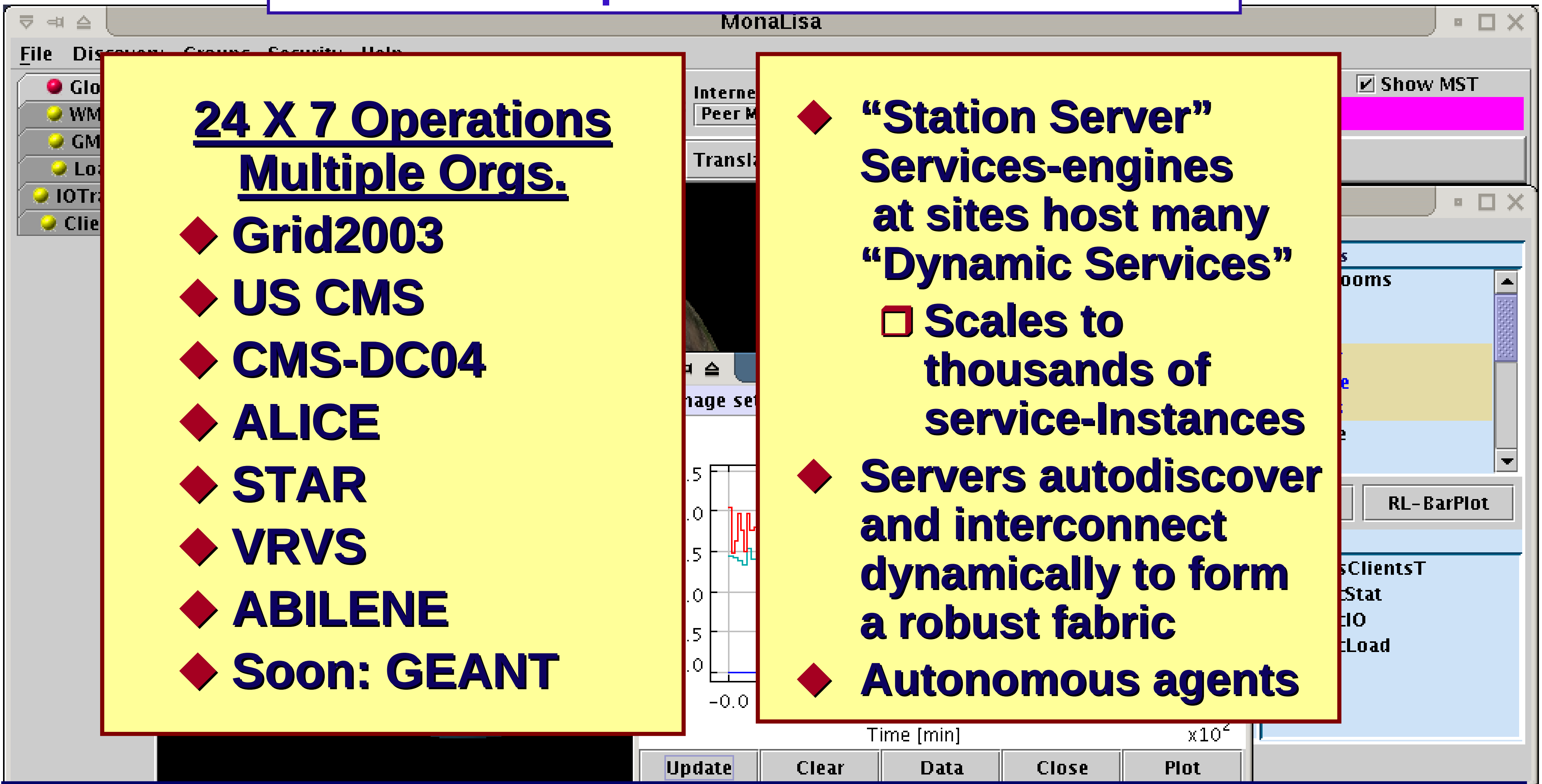
- **Build on the experience of Grid2003, as a persistent, production-quality Grid of national and international scope**
- **Ensure that the U.S. plays a leading role in defining and operating the global grid infrastructure needed for large-scale collaborative and international scientific research.**
- **Combine computing resources at several DOE labs and at dozens of universities to effectively become a single national computing infrastructure for science, the Open Science Grid.**
- **Provide opportunities for educators and students to participate in building and exploiting this grid infrastructure and opportunities for developing and training a scientific and technical workforce. This has the potential to transform the integration of education and research at all levels.**

The Move to OGSA and then Managed Integrated Systems



Managing Global Systems: Dynamic Scalable Services Architecture

MonALISA: http://monalisa.cacr.caltech.edu



24 X 7 Operations Multiple Orgs.

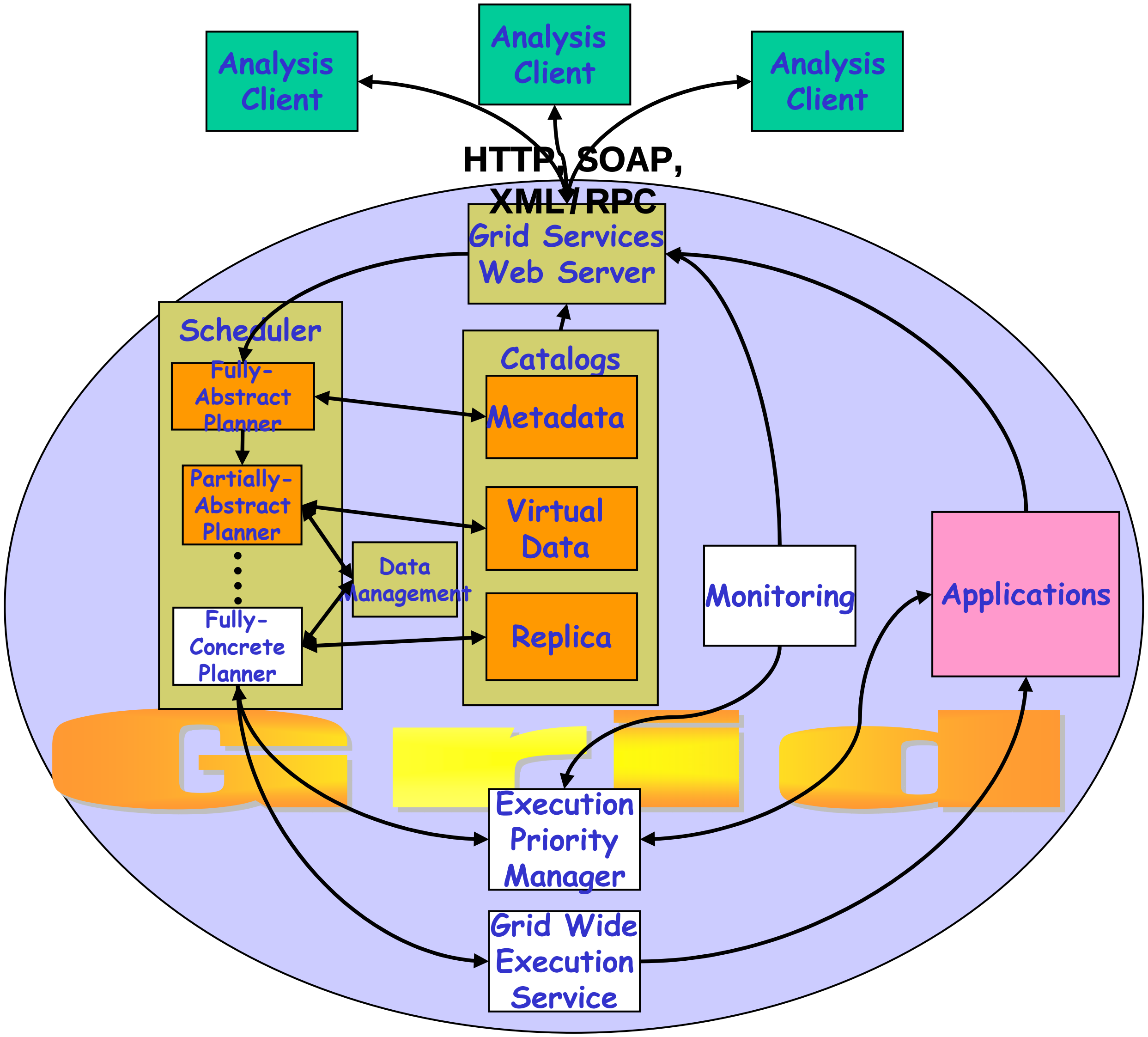
- Grid2003
- US CMS
- CMS-DC04
- ALICE
- STAR
- VRVS
- ABILENE
- Soon: GEANT

- "Station Server" Services-engines at sites host many "Dynamic Services"
  - Scales to thousands of service-Instances
- Servers autodiscover and interconnect dynamically to form a robust fabric
- Autonomous agents

+ CLARENS: Web Services Fabric and Portal Architecture

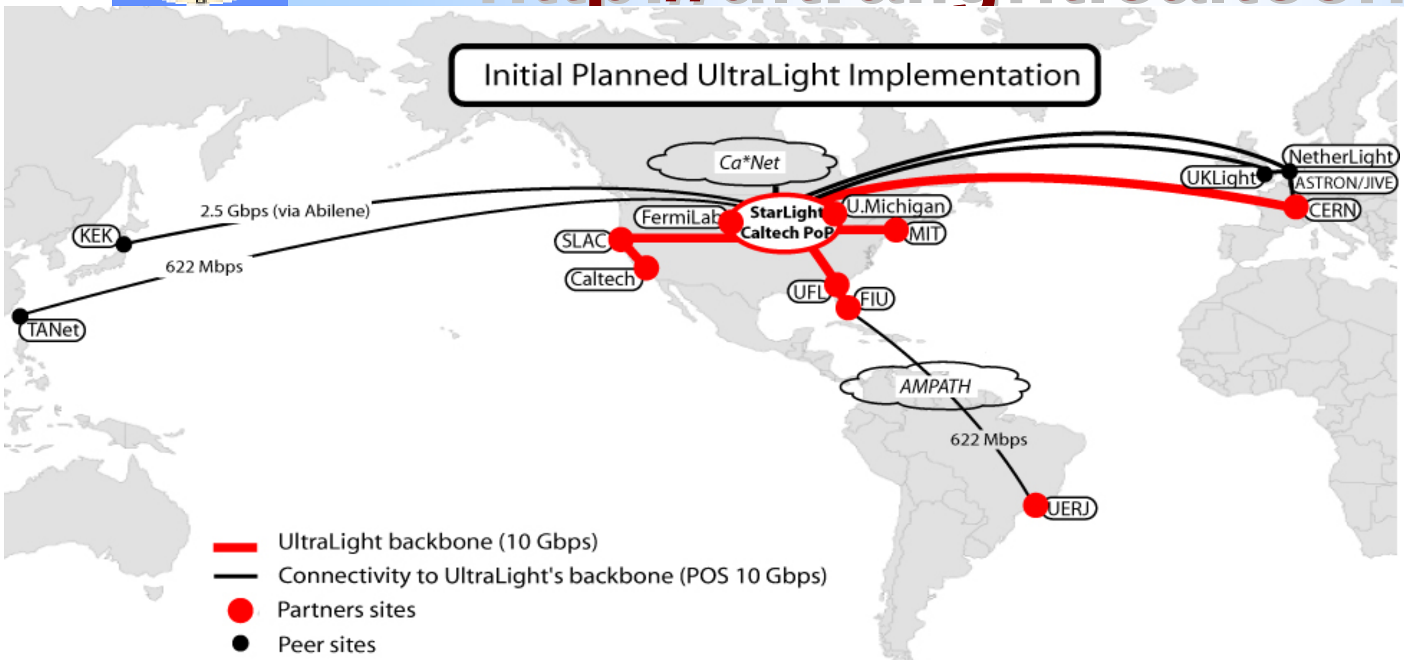
Grid Analysis Environment

CLARENS: Web Services Architecture



- **Analysis Clients talk standard protocols to the *CLARENS "Grid Services Web Server"***
- **The secure Clarens portal may hide the complexity with a simple Web Services API**
- **Key features: *Global* Scheduler, Catalogs, Monitoring, and Grid-wide Execution service**
- ***Clarens servers form a Global Peer to peer Network***

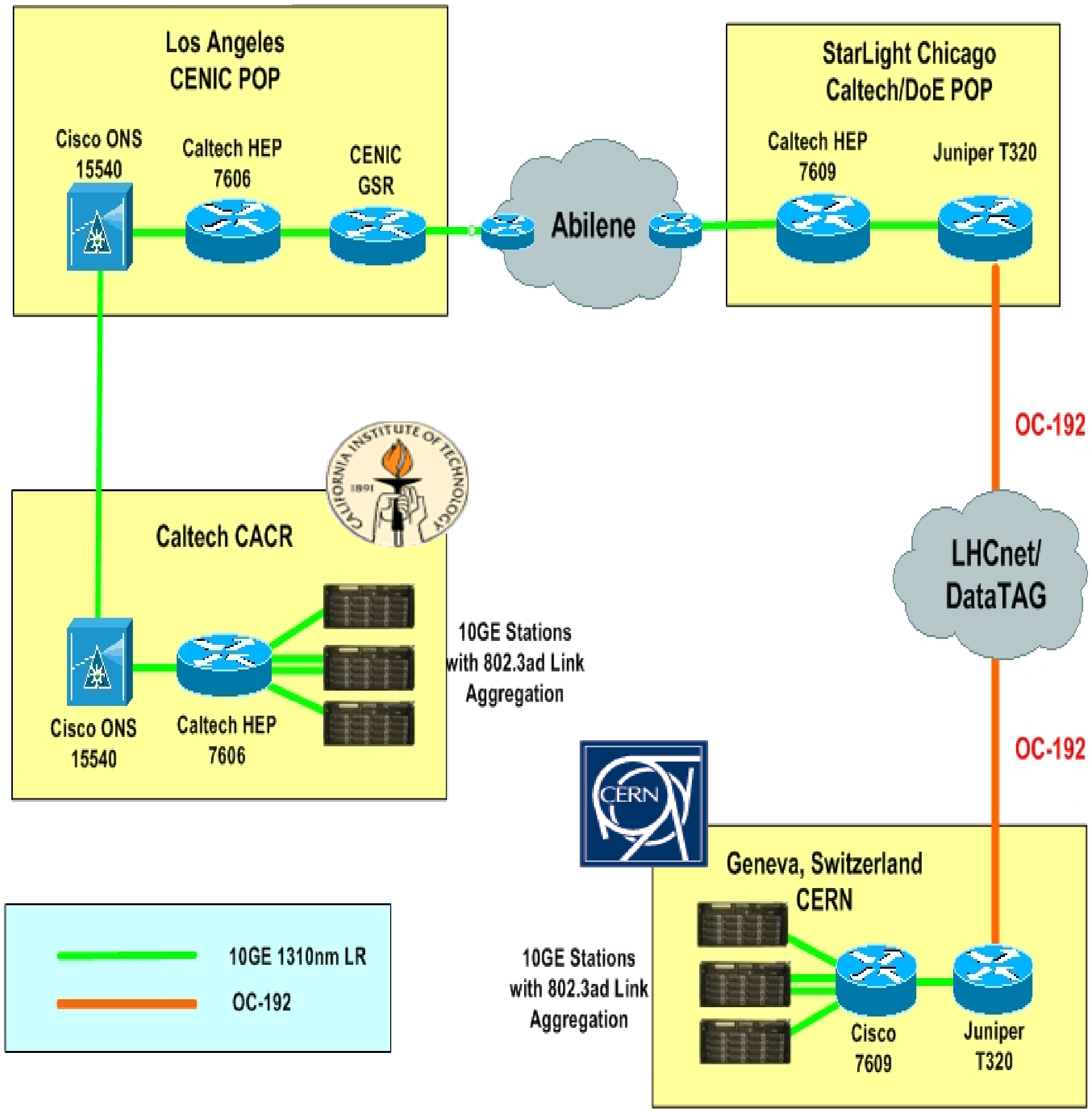
UltraLight Collaboration: http://ultralight.caltech.edu



- Caltech, UF, FIU, UMich, SLAC,FNAL, MIT/Haystack, CERN, UERJ(Rio), NLR, CENIC,UCAID, Translight, UKLight, Netherlight, UvA, UCLondon, KEK, Taiwan
- Cisco, Level(3)

- **Integrated hybrid experimental network, leveraging Transatlantic R&D network partnerships; packet-switched + dynamic optical paths**
  - **10 GbE across US and the Atlantic: NLR, DataTAG, TransLight, NetherLight, UKLight, etc.; Extensions to Japan, Taiwan, Brazil**
- **End-to-end monitoring; Realtime tracking and optimization; Dynamic bandwidth provisioning**
- ***Agent-based services spanning all layers of the system,*** **from the optical cross-connects to the applications.**

Next: 1+ GByte/sec Disk-to-Disk Trials



- **Caltech / Microsoft / AMD / NewiSys /S2IO Collaboration**
- **Low Cost ATA RAID Arrays**
- **10G Link Aggregation (802.3ad): 20 Gbps Input**
- ***Entree to:***
  - ***GB/sec Production Data Services***
  - ***Hybrid N X 10G Nets***

Target: Early May

SCIC Report 2004
Main Conclusion: On the Digital Divide

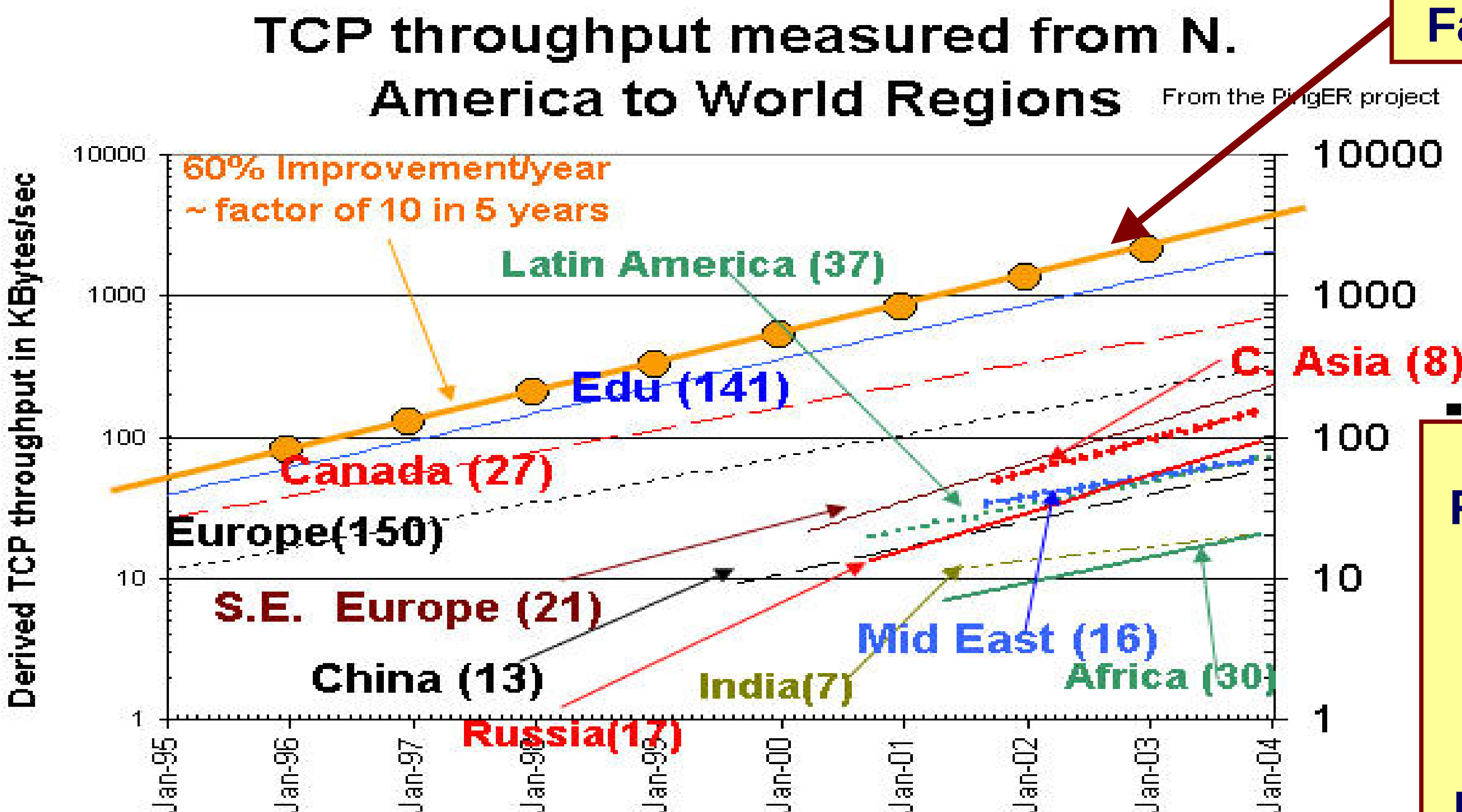
- ***As the pace of network advances continues to accelerate, the gap between the economically "favored" regions and the rest of the world is in danger of widening.***
- ***We must therefore work to Close the Digital Divide***
  - ***To make Scientists from All World Regions Full Partners in Their Experiments; and in the Process of Discovery***
  - ***This is essential for the health of our global experimental collaborations, our plans for future projects, and our field.***

SCIC Monitoring WG (L. Cottrell, SLAC): Throughput Improvements 1995-2004

IEPM
Internet End-to-end Performance Monitoring

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

60% annual improvement
Factor ~100/10 yr



Some Regions ~5-10 Years Behind

SE Europe, Russia, Central Asia May be Catching Up (Slowly);

India Ever-Farther Behind

Progress: but Digital Divide is Mostly Maintained

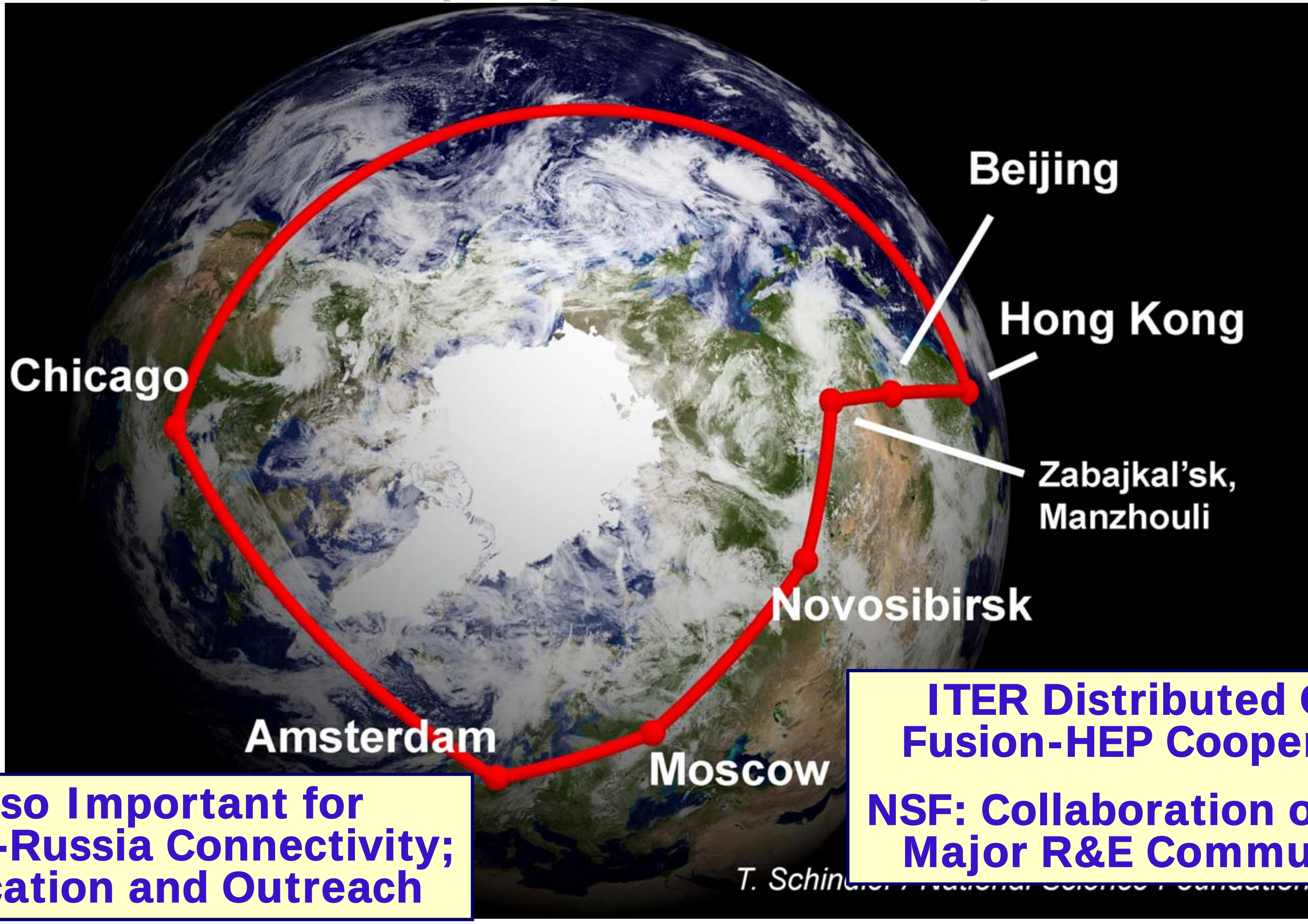
(1) Matthis et al., Computer Communication Review 27(3), July 1997

Work on the Digital Divide: Several Perspectives

- Work on Policies and/or Pricing: pk, in, br, cn, SE Europe, …
  - Find Ways to work with vendors, NRENs, and/or Gov'ts
  - Exploit Model Cases: e.g. Poland, Slovakia, Czech Republic
- Inter-Regional Projects
  - *GLORIAD,* Russia-China-US Optical Ring
  - South America: CHEPREO (US-Brazil); EU @LIS Project
  - Virtual SILK Highway Project (DESY): FSU satellite links
- Workshops and Tutorials/Training Sessions
  - For Example: Digital Divide and HEPGrid Workshop, UERJ Rio, February 2004
- Help with Modernizing the Infrastructure
  - Design, Commissioning, Development
  - Tools for Effective Use: Monitoring, Collaboration
- Participate in Standards Development; Open Tools
  - Advanced TCP stacks; Grid toolkits, systems

GLORIAD: Global Optical Ring (US-Russia-China)

“Little Gloriad” (OC3) Launched January 12; to OC192



Also Important for Intra-Russia Connectivity; Education and Outreach

ITER Distributed Ops.; Fusion-HEP Cooperation

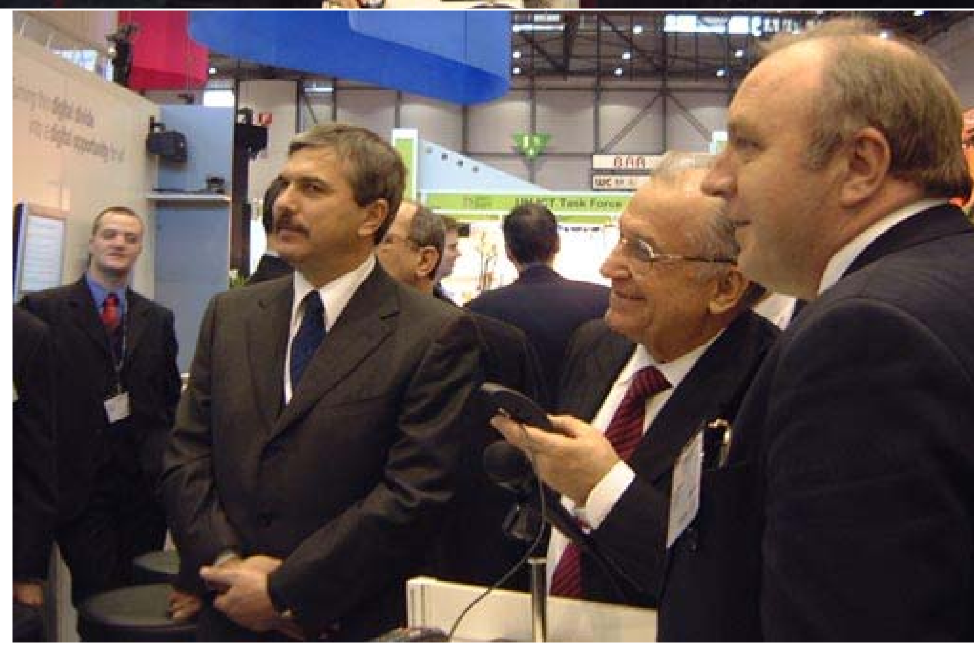
NSF: Collaboration of Three Major R&E Communities

World Summit on the Information Society (WSIS): Geneva 12/2003 and Tunis in 2005

- ***The UN General Assembly adopted a resolution in 2001 endorsing the organization of the World Summit on the Information Society (WSIS), under UN Secretary-General Kofi Annan, with the ITU and host governments leading its preparation.***
- ***GOAL: To Create an Information Society:***
  "Tokyo Declaration" of January 2003:
  "… One in which highly developed ICT networks, equitable and ubiquitous access to information, appropriate content in accessible formats and effective communication can help people achieve their potential"
- ***Kofi Annan Challenged the Scientific Community to Help (3/03)***
- ***CERN and ICFA SCIC have been quite active in the WSIS in Geneva (12/2003)***

Role of Science in the Information Society. Palexpo, Geneva 12/2003

- **CERN SIS Forum, and**
- **CERN/Caltech Online Stand**
- **<u>Visitors:</u>**
  - **Kofi Annan, UN Sec'y General**
  - **John H. Marburger, Science Adviser to US President**
  - **Ion Iliescu, President of Romania; and Dan Nica, Minister of ICT**
  - **Jean-Paul Hubert, Ambassador of Canada in Switzerland**
  - **Carlo Lamprecht, Pres. of Economic Dept. of Canton de Geneva**
  - **...**



Role of Sciences in Information Society. Palexpo, Geneva 12/2003

- **Demos at the CERN/Caltech RSIS Online Stand**
  - Advanced network and Grid-enabled analysis
  - Monitoring very large scale Grid farms with MonALISA
  - World Scale multisite multi-protocol videoconference with VRVS (Europe-US-Asia-South America)
  - Distance diagnosis and surgery using Robots with "haptic" feedback (Geneva-Canada)
  - Music Grid: live performances with bands at St. John's, Canada and the Music Conservatory of Geneva on stage

VRVS 30k hosts
103 Countries
2-3X Growth/Year

Networks and Grids for HENP and Global Science

- **Network backbones and major links used by HENP and other fields of data intensive science are advancing rapidly**
  - **To the 10 Gbps range in < 2 years; much faster than Moore's Law**
- **We are learning to use long distance 10 Gbps networks effectively**
  - **2003-2004 Developments: to 6+ Gbps flows over 11 kkm**
- **Removing Regional, Last Mile, Local Bottlenecks and Compromises in Network Quality are now On *the Critical Path***
- **A transition to community-owned and operated R&E networks is beginning (us, ca, nl, pl, cz, sk …) or considered (de, ro, …)**
- ***US and Global Science Network Roadmaps: A National Priority***
  - **A Factor ~1000 Improvement Per Decade**
  - **Hybrid Optical Networks; and Beyond**
- ***We Must Work to Close to Digital Divide***
  - ***Allowing Scientists and Students from All World Regions to Take Part in Discoveries at the Frontiers of Science***