Next-Generation Networking for Science

ASCAC Presentation

March 23, 2011

Program Managers

Richard Carlson <richard.carlson@science.doe.gov>

Thomas Ndousse <thomas.ndousse-fetter@science.doe.gov>

Presentation Outline

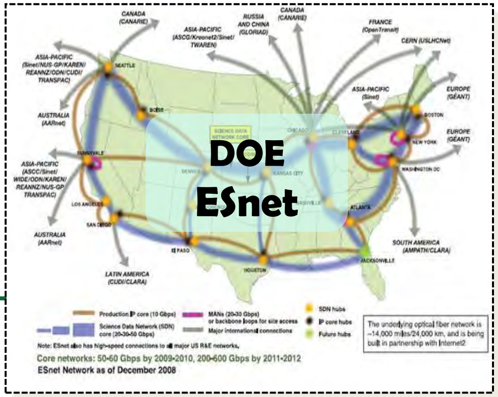
- Program Mission
- Program Elements
- Science Drivers
- Budget
- Previous Accomplishments
- Current Portfolio
- Program Highlights
- Future Directions
- Conclusions

Next-generation Networks for Science

Mission:

The Goal of the program is to research, develop, test and deploy advanced network technologies critical in addressing networking capabilities unique to DOE's science mission. The program's portfolio consists of two main elements:

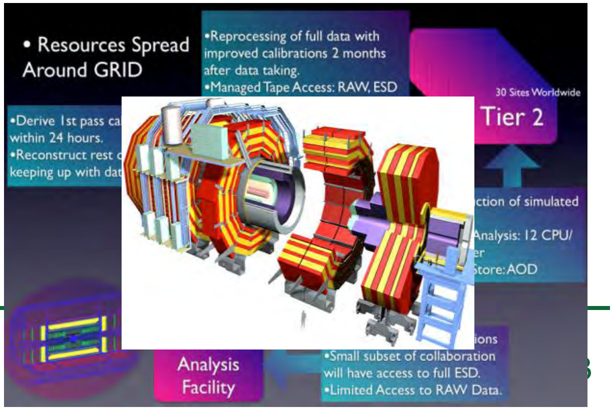
High-Performance Networks
High-Performance Middleware


DOE
ESnet
CANADA (CANARIE)
ASIA-PACIFIC (ASCG/Kreonet2/Sinet/TWAREN)
RUSSIA AND CHINA (GLORIAD)
CANADA (CANARIE)
FRANCE (OpenTransit)
CERN (USLHCNet)
ASIA-PACIFIC (Sinet/NUS-GP/KAREN/REANNZ/ODN/CUDI/TRANSPAC)
ASIA-PACIFIC (Sinet)
EUROPE (GÉANT)
AUSTRALIA (AARnet)
EUROPE (GÉANT)
ASIA-PACIFIC (ASCC/Sinet/WIDE/ODN/KAREN/REANNZ/NUS-GP/TRANSPAC)
AUSTRALIA (AARnet)
SOUTH AMERICA (AMPATH/CLARA)
LATIN AMERICA (CUDI/CLARA)
SEATTLE
BOISE
SUNNYVALE
LOS ANGELES
SAN DIEGO
EL PASO
HOUSTON
DENVER
KANSAS CITY
CHICAGO
CLEVELAND
NASHVILLE
ATLANTA
JACKSONVILLE
BOSTON
NEW YORK
WASHINGTON DC
SCIENCE DATA NETWORK CORE
Production IP core (10 Gbps)
Science Data Network (SDN) core (20-30-50 Gbps)
MANs (20-30 Gbps) or backbone loops for site access
Major international connections
SDN hubs
IP core hubs
Future hubs
The underlying optical fiber network is ~14,000 miles/24,000 km, and is being built in partnership with Internet2

Note: ESnet also has high-speed connections to all major US R&E networks,

Core networks: 50-60 Gbps by 2009-2010, 200-600 Gbps by 2011-2012

ESnet Network as of December 2008



- Resources Spread Around GRID

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD

30 Sites Worldwide

Tier 2

•Derive 1st pass ca
within 24 hours.
•Reconstruct rest o
keeping up with dat

ction of simulated

Analysis: 12 CPU/
er
Store:AOD

Analysis Facility

ions
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

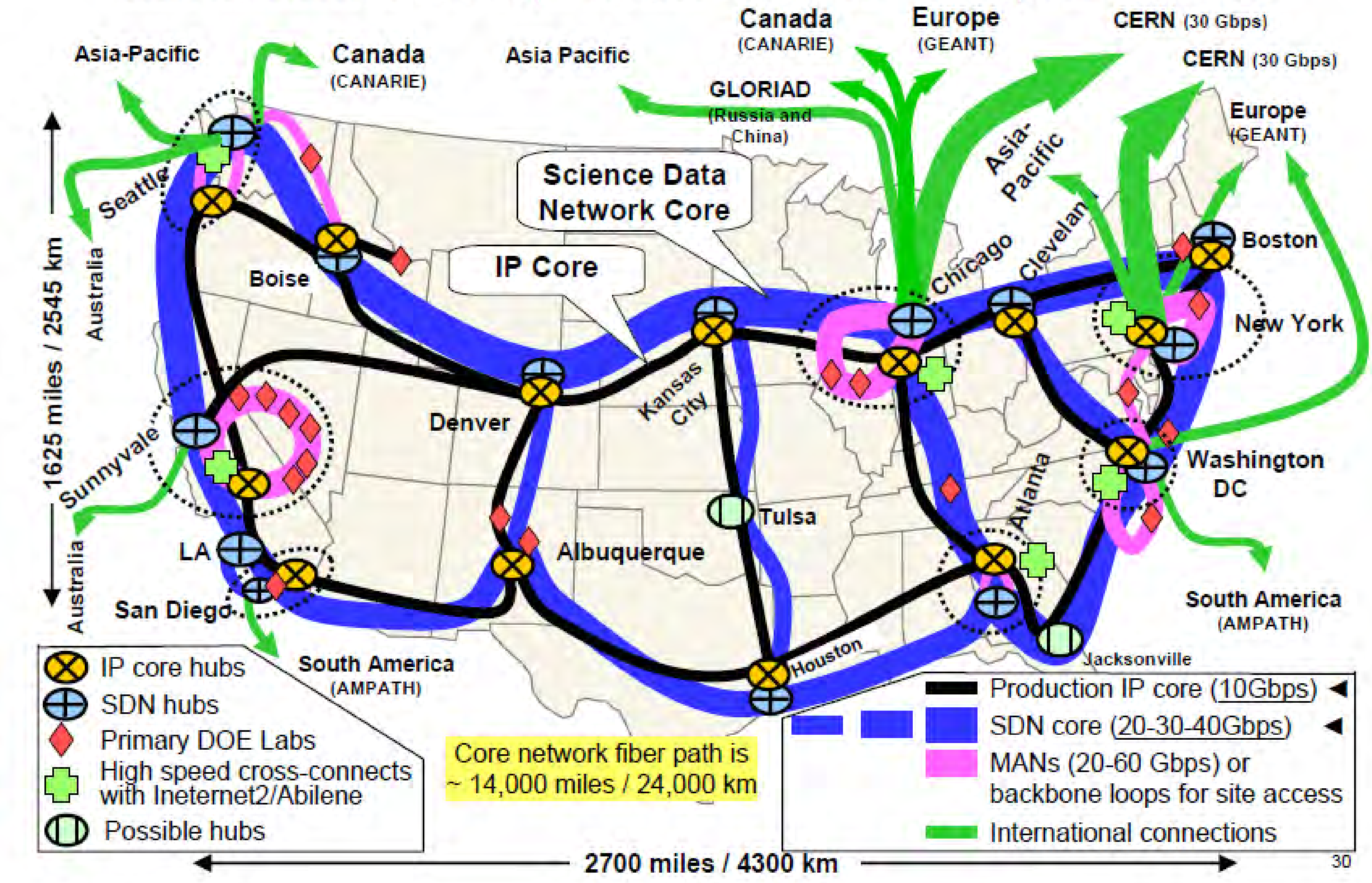
Program Elements

- **High-Performance Networks – Research and development of advanced technologies which include technologies for rapid provisioning of hybrid packet/circuit-switched networks, ultra high-speed transport protocols, high-speed data distribution tools and services, secure and scalable technologies for bandwidth and circuits reservation and scheduling, secure and scalable tools and services for monitoring and managing of federated networks.**

- **High-Performance Middleware – research and development to support distributed high-end science applications and related distributed scientific research activities. These include advanced middleware to enable large-scale scientific collaborations; Secure and scalable software stacks to manage and distribute massive science data, software and services to seamlessly integrated science workflows to experiments and network infrastructures; cyber security systems and services to enable large-scale national and international scientific collaborations.**

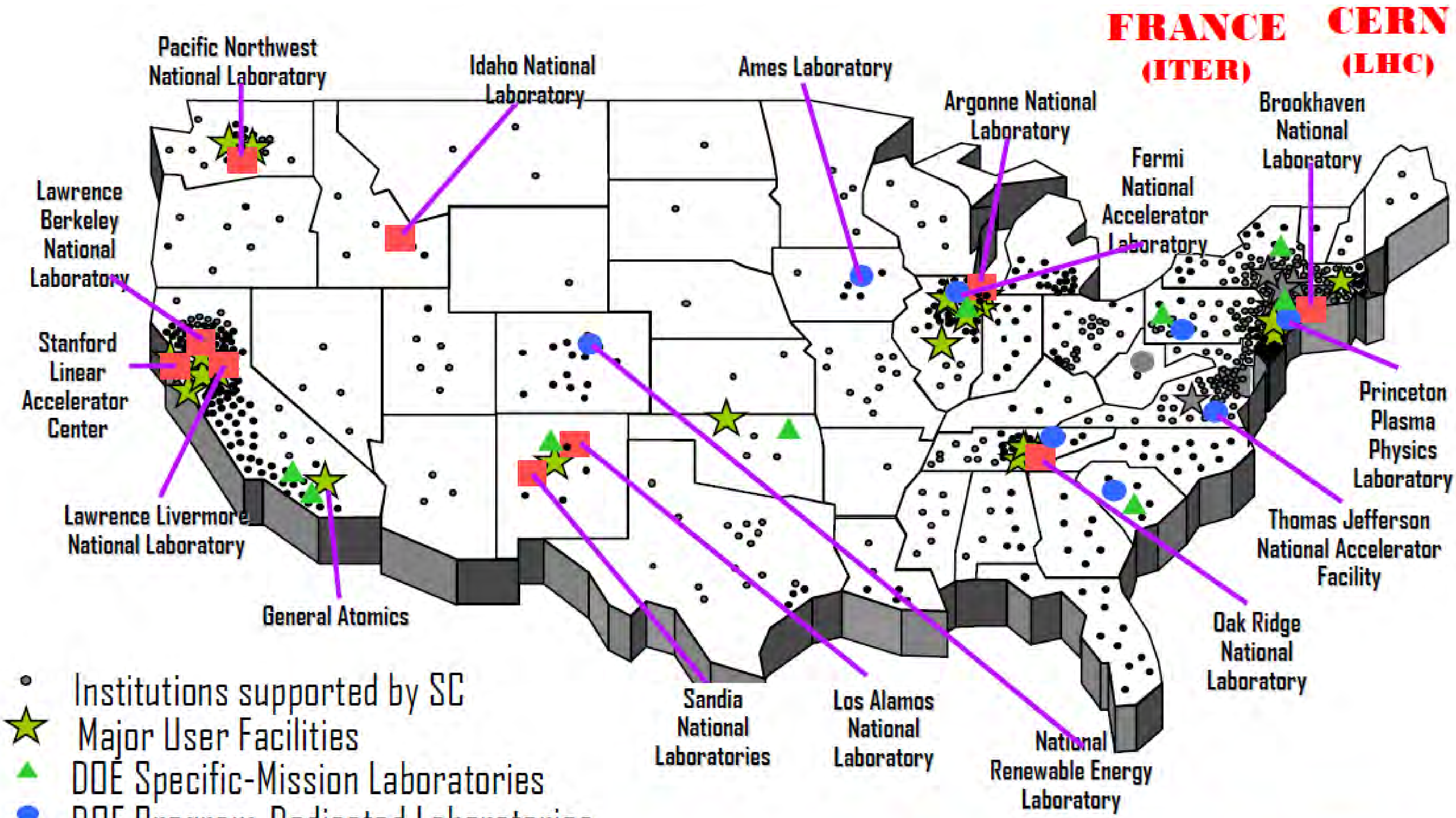
DOE’s ESnet Capabilities Projection

Core networks: 40-50 Gbps in 2009-2010, 160-400 Gbps in 2011-2012



Core network fiber path is ~ 14,000 miles / 24,000 km

DOE Distributed Science Complex

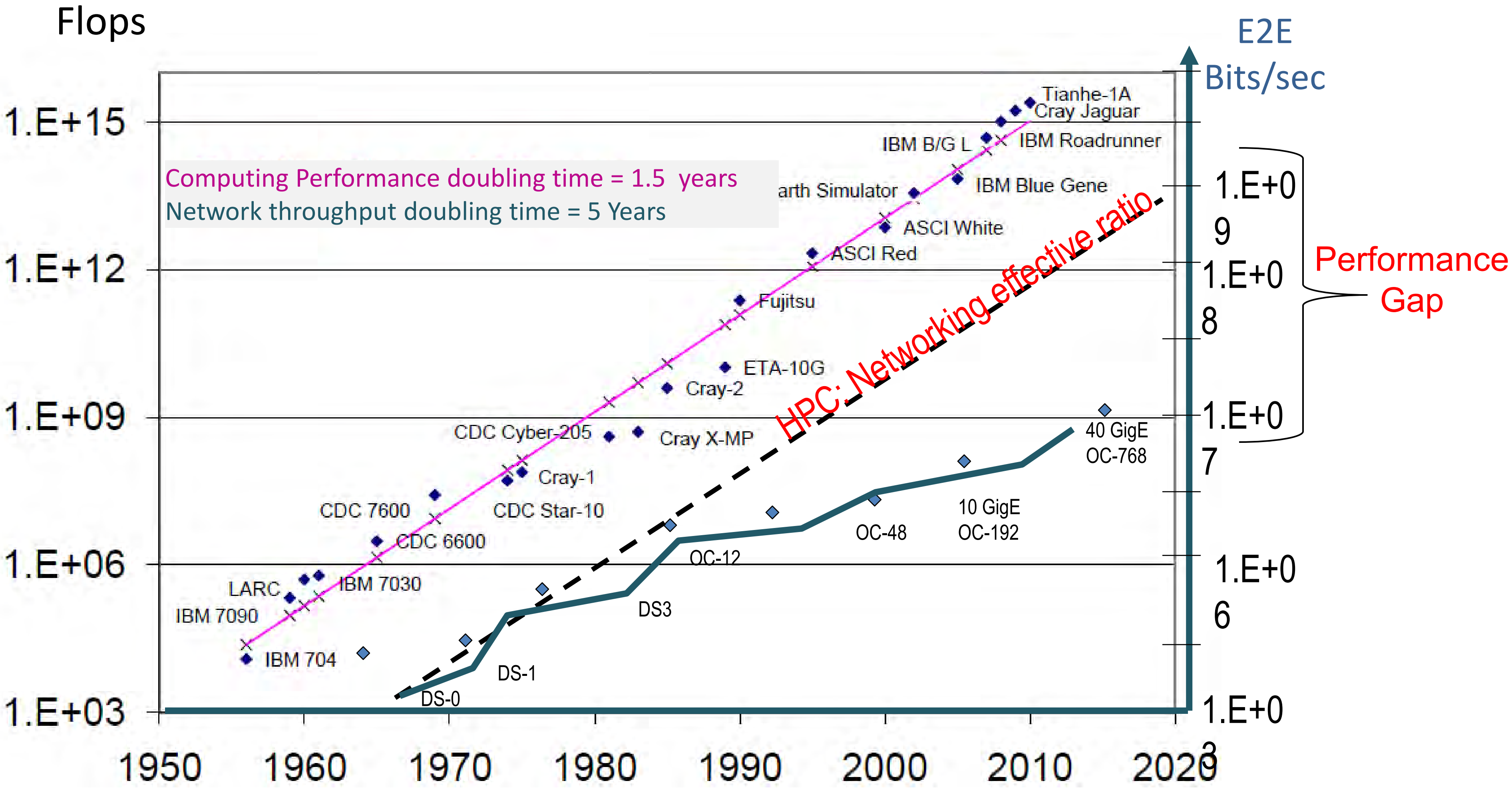

FRANCE
(ITER)
CERN
(LHC)
Pacific Northwest National Laboratory
Idaho National Laboratory
Ames Laboratory
Argonne National Laboratory
Fermi National Accelerator Laboratory
Brookhaven National Laboratory
Lawrence Berkeley National Laboratory
Stanford Linear Accelerator Center
Lawrence Livermore National Laboratory
General Atomics
Princeton Plasma Physics Laboratory
Thomas Jefferson National Accelerator Facility
Oak Ridge National Laboratory
Sandia National Laboratories
Los Alamos National Laboratory
National Renewable Energy Laboratory

- ● Institutions supported by SC
- ★ Major User Facilities
- ▲ DOE Specific-Mission Laboratories
- ● DOE Program-Dedicated Laboratories
- 🟥 DOE Multiprogram Laboratories

Large-scale Distributed Scientific Collaborations

Data-Intensive Science:

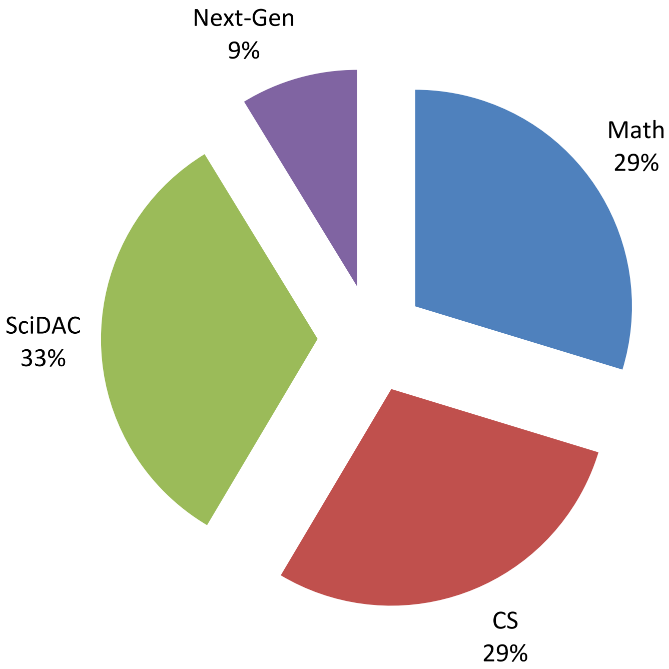
HPC performance to : Networking Performance Gap



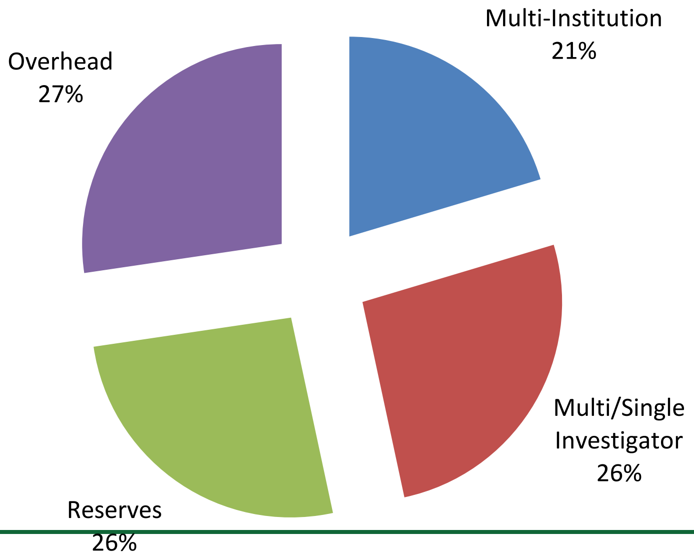
U.S. DEPARTMENT OF ENERGY | Office of Science

Next-Gen Funding Profile

FY11 - $165M



FY11 - $14.3M



Historical Perspective of Networking R&D at DOE

- 1989 – Van Jacobson's congestion control algorithm at LBL helps avert an imminent Internet congestion collapse.
- 2003 -The search for efficient scientific collaboration environment services at ANL leads to grid computing (Globus & GridFTP).
- 2005 – ORNL prototypes a working model of on-demand and dynamic circuit services over Ultra-Science Network testbed.
- 2007 – ESnet extends on-demand bandwidth services to provide production grade services to scientists.
- 2010 – DOE/ESnet initiates the deployment of a 100 Gbps end-to-end demonstration network prototype.

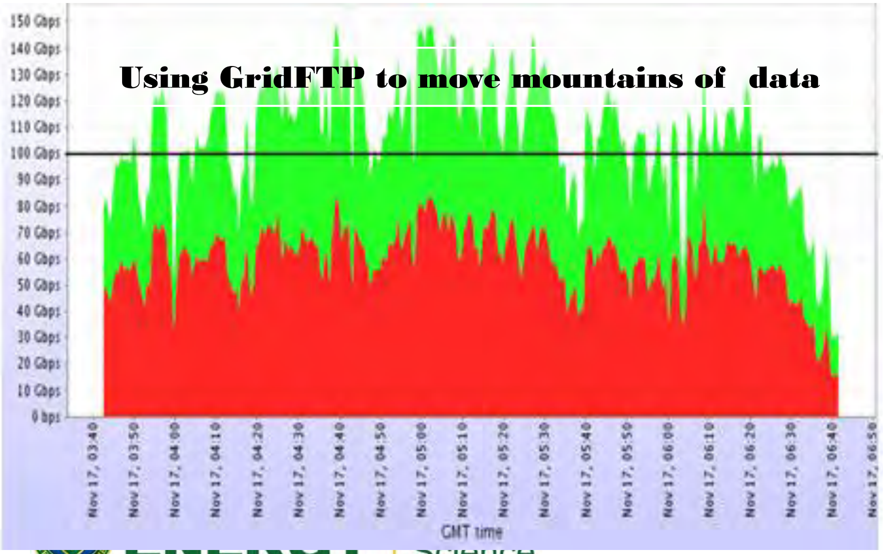
Next-Generation Research Contributions to Science and the Internet

GridFTP

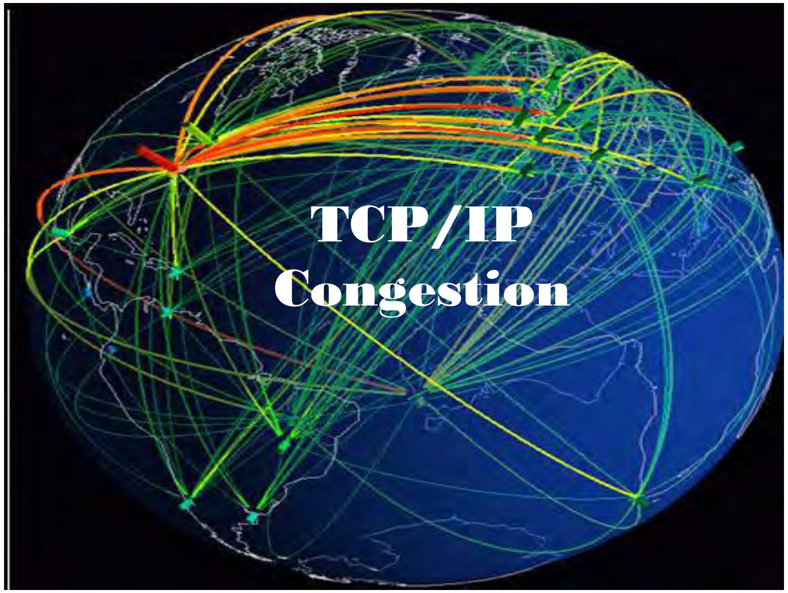
2003 - The GridFTP data transfer protocol developed by ASCR network researchers delivers 100x throughput over traditional Internet –based FTP in data transfer.

GridFTP is now the de-facto data transfer protocol used in scientific and commercial grid computing world-wide for data movement.

GridFTP is the primary data transfer protocol used to distribute the massive data generated by the LHC experiment and the global climate modeling communities.



Using GridFTP to move mountains of data



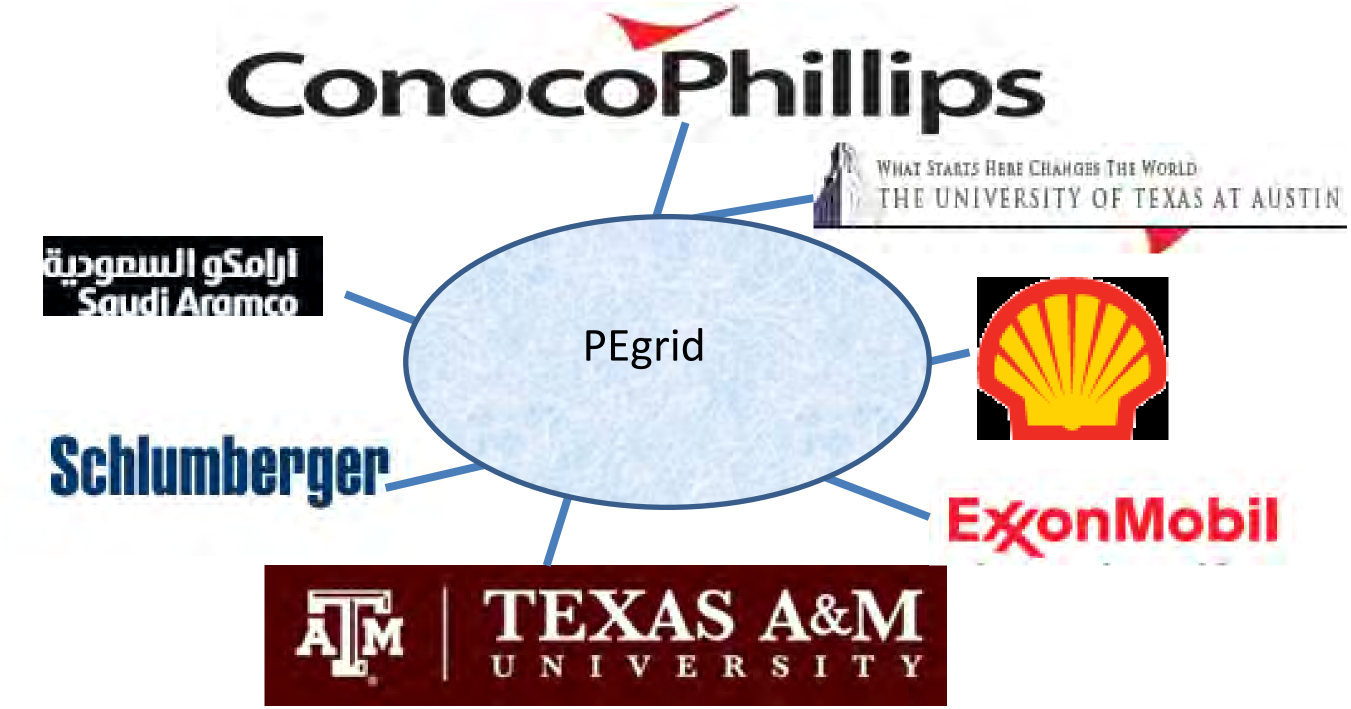
TCP/IP Congestion

TCP/IP Congestion Control

1989 - Van Jacobson, working at Lawrence Berkeley Nation Laboratory, developed the algorithm that solved the congestion problem of TCP protocol used in over 90% of Internet hosts today.

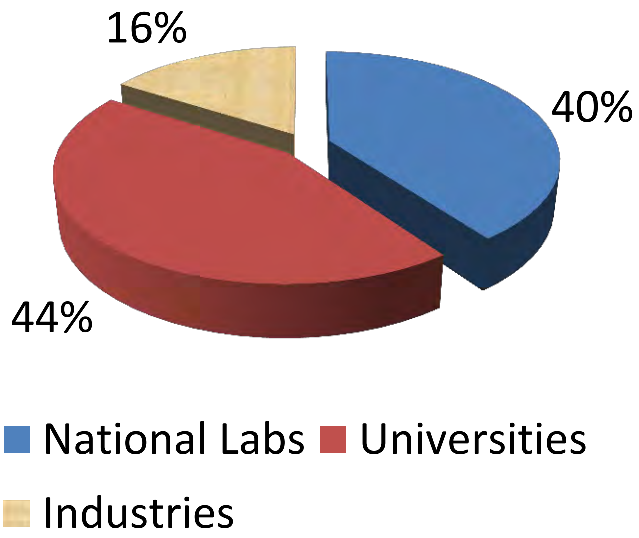
This algorithm is credited for enabling the Internet to expand in size and support increasing speed demands. The algorithm helped the Internet survive a major traffic surge (1988-89) without collapsing.

Next-Generation Networking for Sustained Large-Scale, Global Science Leadership



ASCR funded Globus and OSG middleware provides the fundamental building blocks for many grid infrastructure projects. PEgrid, a $45M collaboration of Texas universities, software companies, and oil companies leveraged these services to create an environment where students can be trained to use multi-dimensional supercomputer simulation models.

Current Portfolio



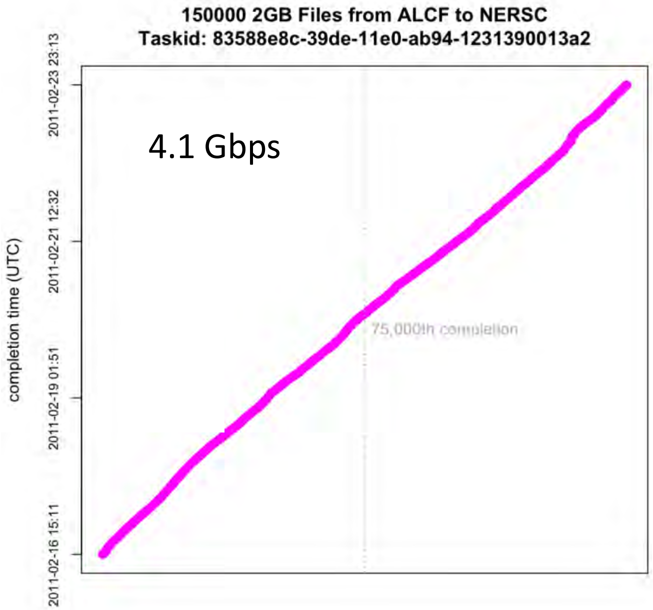
All Awards made through open solicitations

- **Affiliation of our Researchers**
  - National Lab – 39%
  - University – 42% in
  - Industry – 19%
- **Portfolio distribution**
  - Long-term R&D – 20%
  - Short – 60%
  - Testbed activities 20%
- **Project size**
  - 3 large multi-institution
  - 15 single investigator

Portfolio Breakdown

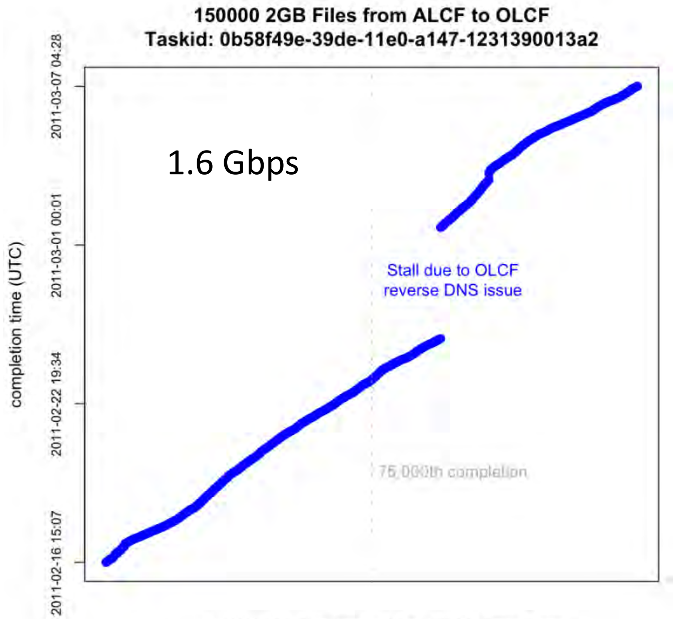
- **Data Movement**
  - Globus On-line
- **Advanced Network Provisioning**
  - OSCARS, Terapaths, Lambda Station
- **perfSONAR based Network Performance Monitoring**
  - IP based network infrastructure
  - Dynamic circuit infrastructure
- **Large-scale Scientific Collaboration**
  - Earth System Grid
  - Open Science Grid
- **Advanced Network Concepts**
  - Hybrid optical networking
  - Network virtualization

CEDPS - SaaS Data Management



150000 2GB Files from ALCF to NERSC
Taskid: 83588e8c-39de-11e0-ab94-1231390013a2

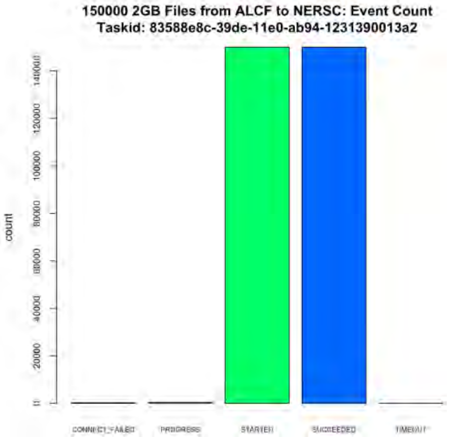
4.1 Gbps



150000 2GB Files from ALCF to OLCF
Taskid: 0b58f49e-39de-11e0-a147-1231390013a2

1.6 Gbps

Stall due to OLCF reverse DNS issue

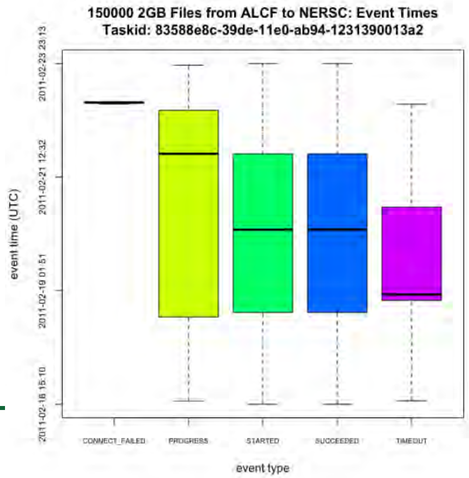


150000 2GB Files from ALCF to NERSC: Event Count
Taskid: 83588e8c-39de-11e0-ab94-1231390013a2

Moving 322 TeraBytes of data from ANL to each remote site

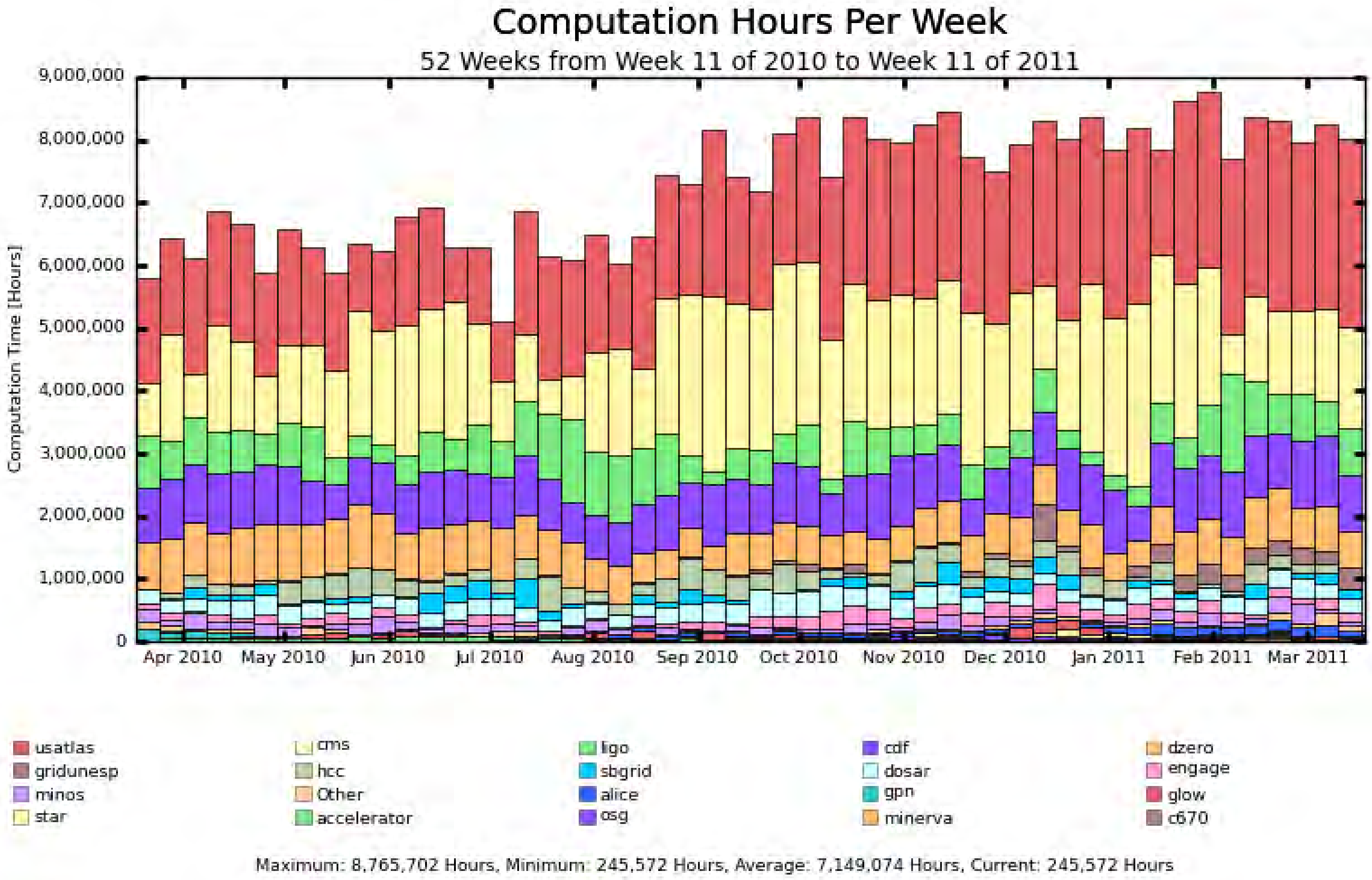
7.34 Days to NERSC

18.56 Days to ORNL



150000 2GB Files from ALCF to NERSC: Event Times
Taskid: 83588e8c-39de-11e0-ab94-1231390013a2

Open Science Grid



The Open Science Grid (OSG) promotes scientific discovery and collaboration in data-intensive research by providing a set of services that supports computation at multiple scales. Strong growth by several international physics communities highlights how OSG resources are used to convert raw instrument data into meaningful information that leads to scientific discoveries.

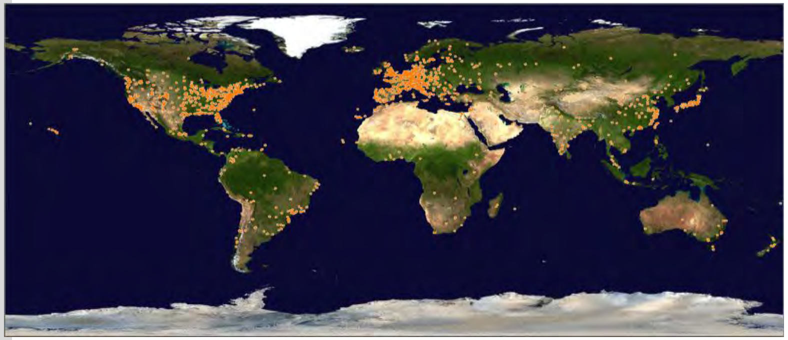
Earth System Grid

SciDAC
Scientific Discovery through Advanced Computing

ESG serves climate data to the world: *"The past successes of ESG enabled the CMIP3 to be wildly successful and a large contributor towards the success of the IPCC 4th Assessment Report. It is hard for me to imagine an effort being more successful given where the field was at that time. The Earth System Grid is a crucial component towards the success of the data serving for CMIP5 and AR5".* **Ron Stouffer, NOAA Scientist & IPCC Author**

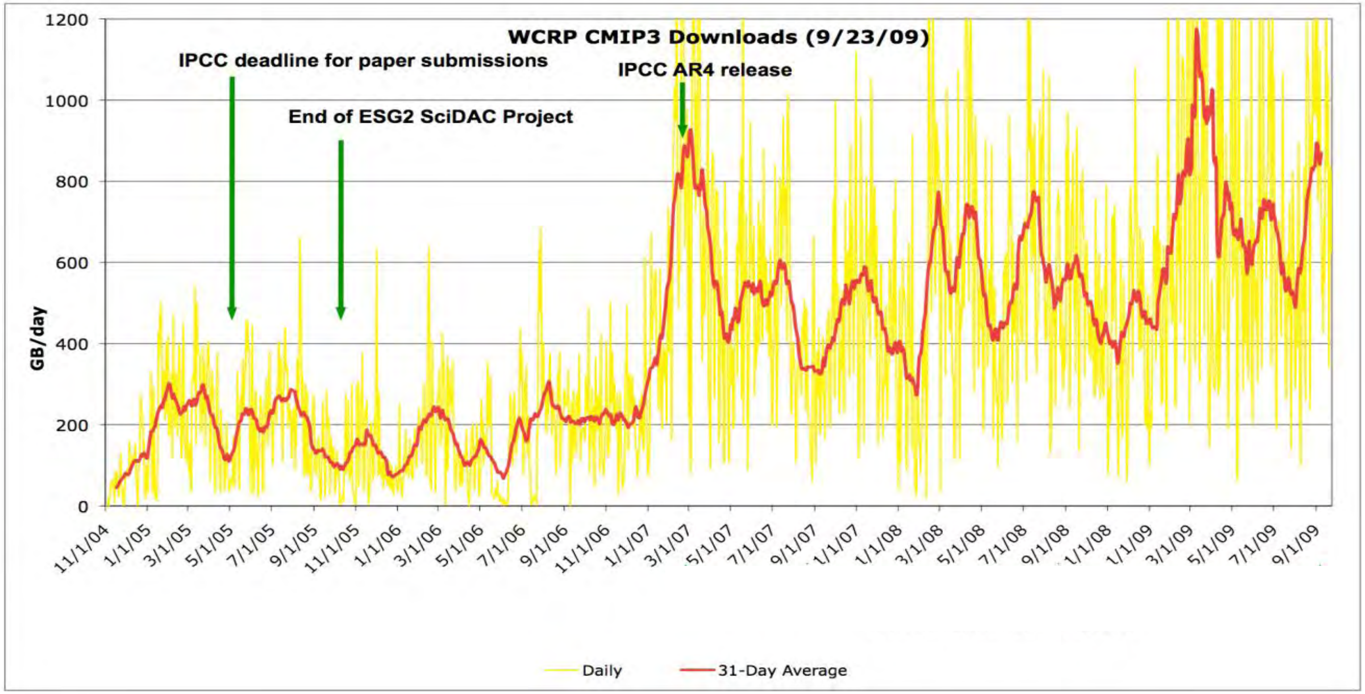
" A major advance of this 4th assessment of climate change projections compared with the 3rd Assessment Report is the large number of simulations available from a broader range of models. Taken together with additional information from observations, these provided a quantitative basis for estimating likelihoods fro many aspects of future climate change. "

p.12 IPCC 4th Assessment Report



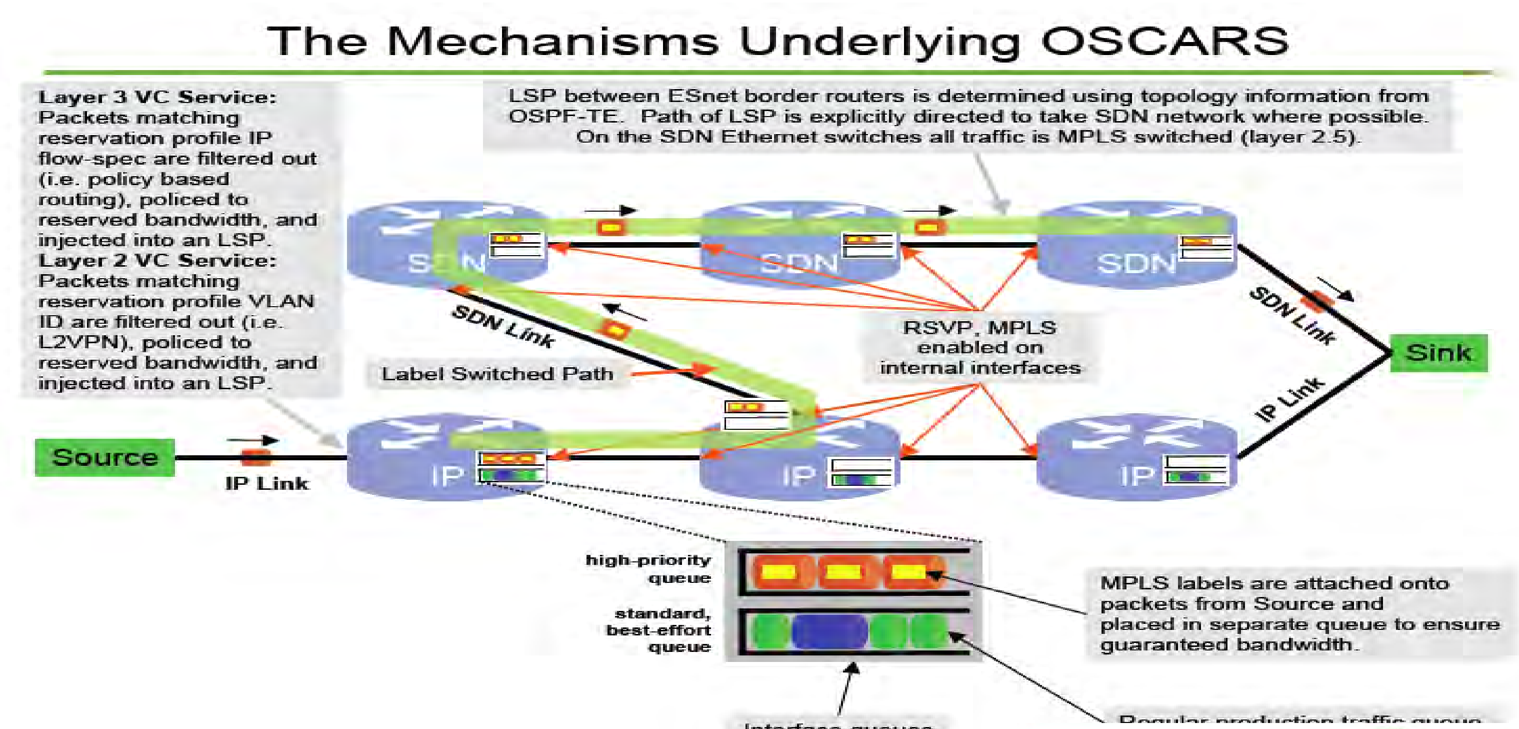
The Nobel Peace Prize 2007

"*For leadership …. which led to a new era in climate system analysis and understanding.*"
Award to PCMDI by American Meteorological Society, Jan. 2010



The Earth System Grid… is an essential component to tapping into the wealth of information about climate and climate models embedded within the CMIP archives."
Gavin Schmidt, NASA Scientist

E2E On-Demand Bandwidth and Circuits



The Mechanisms Underlying OSCARS

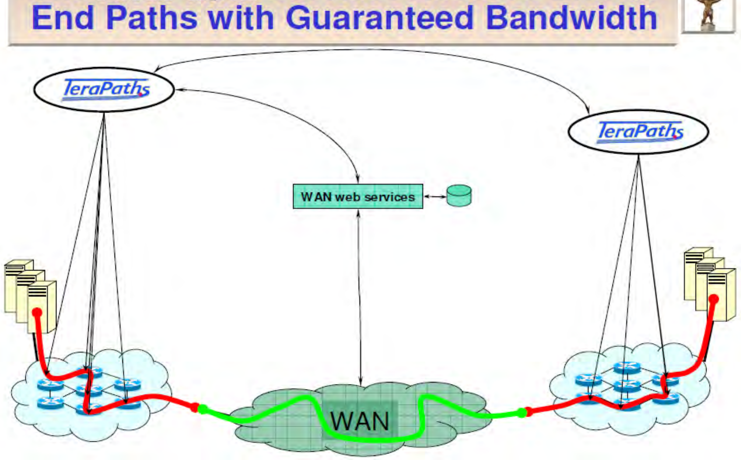
Layer 3 VC Service: Packets matching reservation profile IP flow-spec are filtered out (i.e. policy based routing), policed to reserved bandwidth, and injected into an LSP.

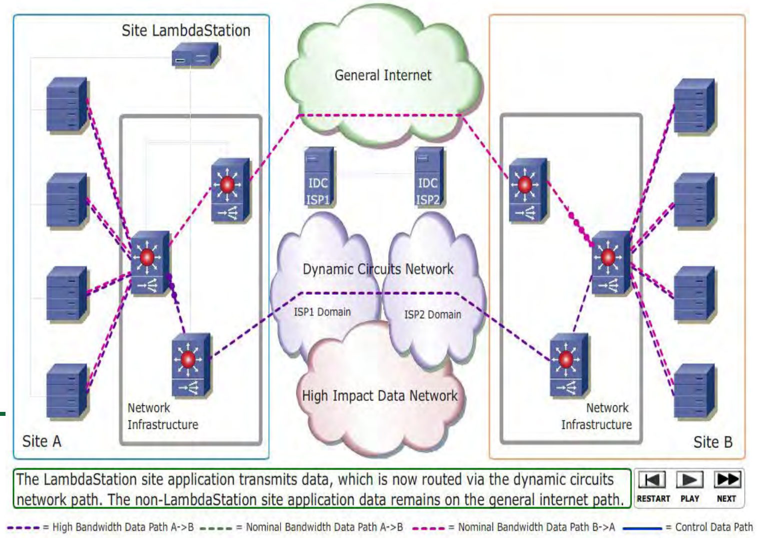
Layer 2 VC Service: Packets matching reservation profile VLAN ID are filtered out (i.e. L2VPN), policed to reserved bandwidth, and injected into an LSP.

LSP between ESnet border routers is determined using topology information from OSPF-TE. Path of LSP is explicitly directed to take SDN network where possible. On the SDN Ethernet switches all traffic is MPLS switched (layer 2.5).

MPLS labels are attached onto packets from Source and placed in separate queue to ensure guaranteed bandwidth.

The *TeraPaths* Service: Reserve End-to-End Paths with Guaranteed Bandwidth

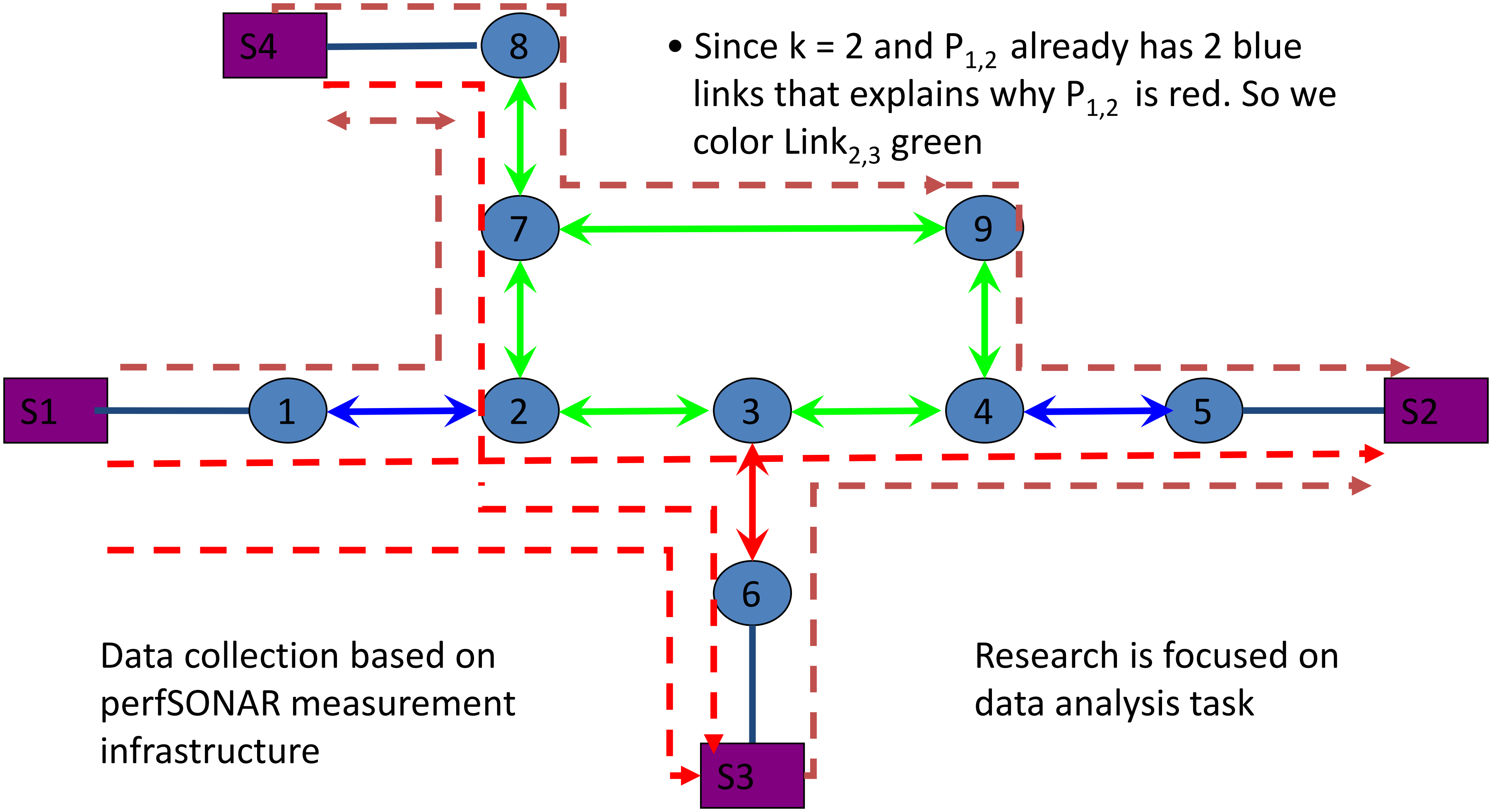




The LambdaStation site application transmits data, which is now routed via the dynamic circuits network path. The non-LambdaStation site application data remains on the general internet path.

---- = High Bandwidth Data Path A->B ---- = Nominal Bandwidth Data Path A->B ---- = Nominal Bandwidth Data Path B->A —— = Control Data Path

Advanced Diagnostics and Analysis



- Since k = 2 and $P_{1,2}$ already has 2 blue links that explains why $P_{1,2}$ is red. So we color $\text{Link}_{2,3}$ green

Data collection based on perfSONAR measurement infrastructure

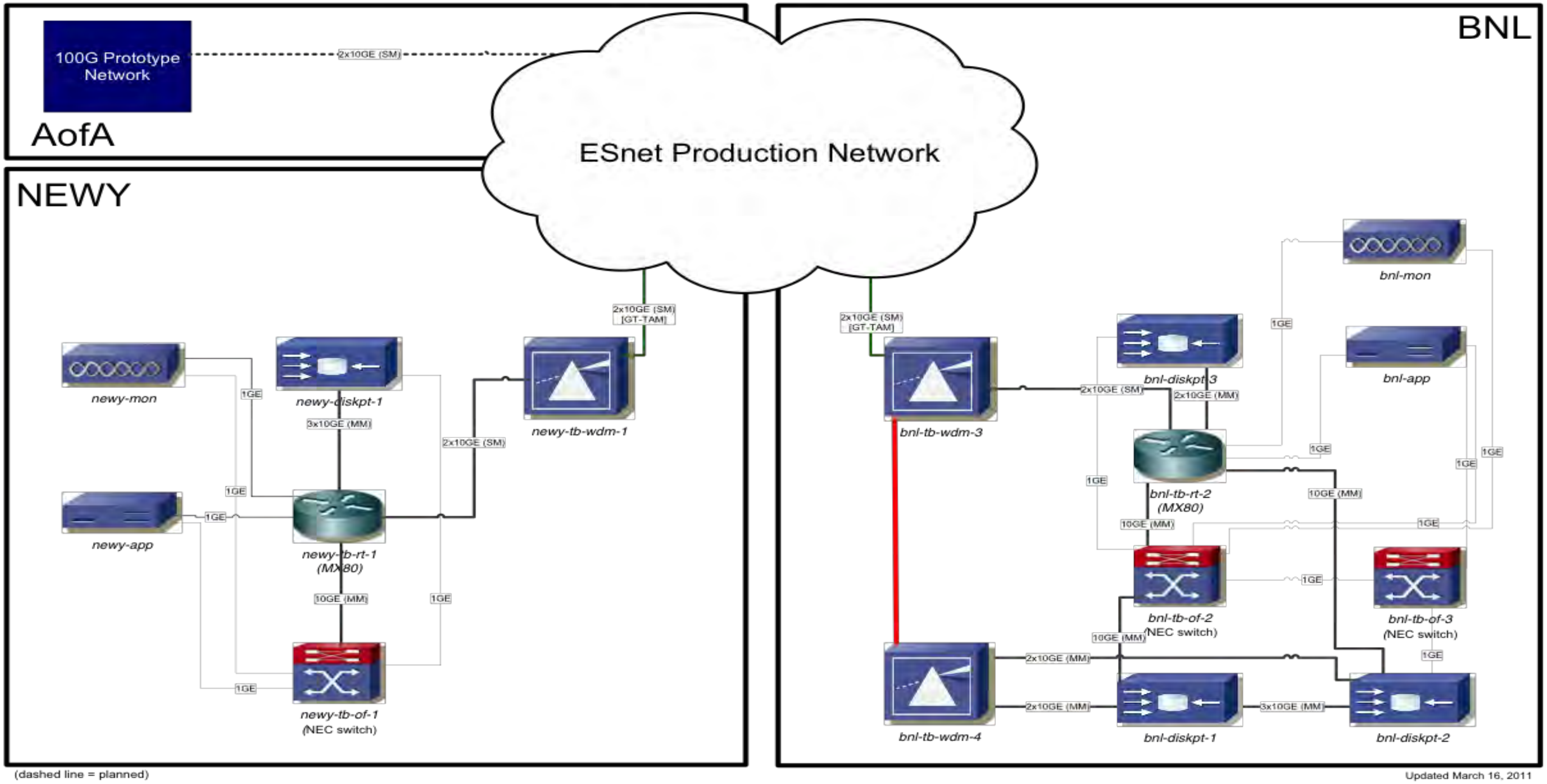
Research is focused on data analysis task

ARRA Activities

- **Advanced Network Initiative - Nation-wide 100 Gbps network demonstration prototype**
  - FTP100
  - 100Gbps Network interface
  - 100 Gbps experimental network facility
  - Scaling ESG and OSG to 100 Gbps
- **Magellan Cloud computing – Exploring the capabilities of cloud computing for scientific application**
  - NERSC
  - ANL

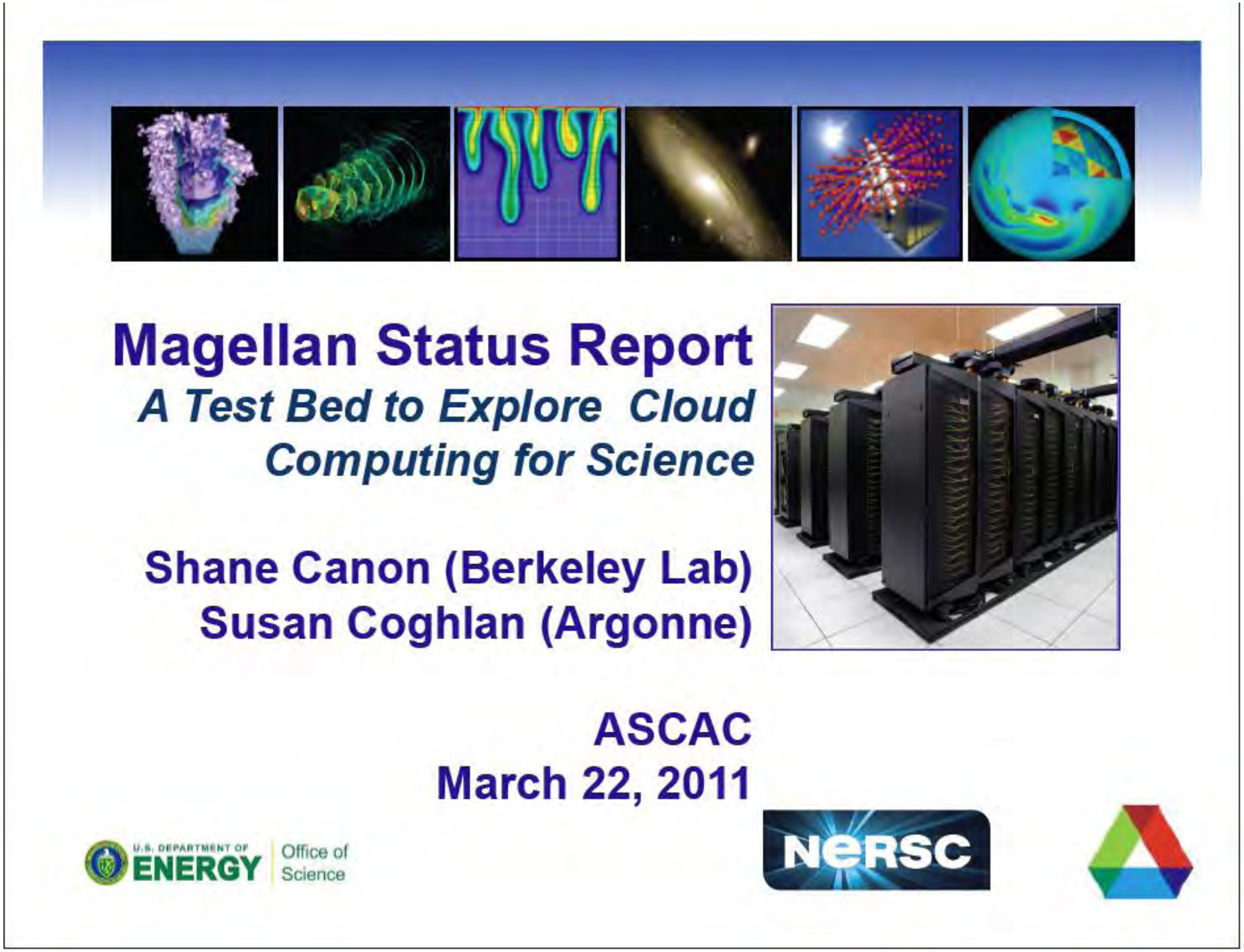
DOE ANI Testbed

Long Island MAN (LIMAN) Testbed Architecture



Notes:
-"App Host": can be used for researcher application, control plane control software, etc. Can support up to 8 simultaneous VMs
-"I/O Testers" are capable of 15 G disk-to-disk or 35G memory-to-memory
-Other infrastructure not shown: VPN Server, file server (NFS, webdav, svn, etc.)

ANI Funded Magellan Project



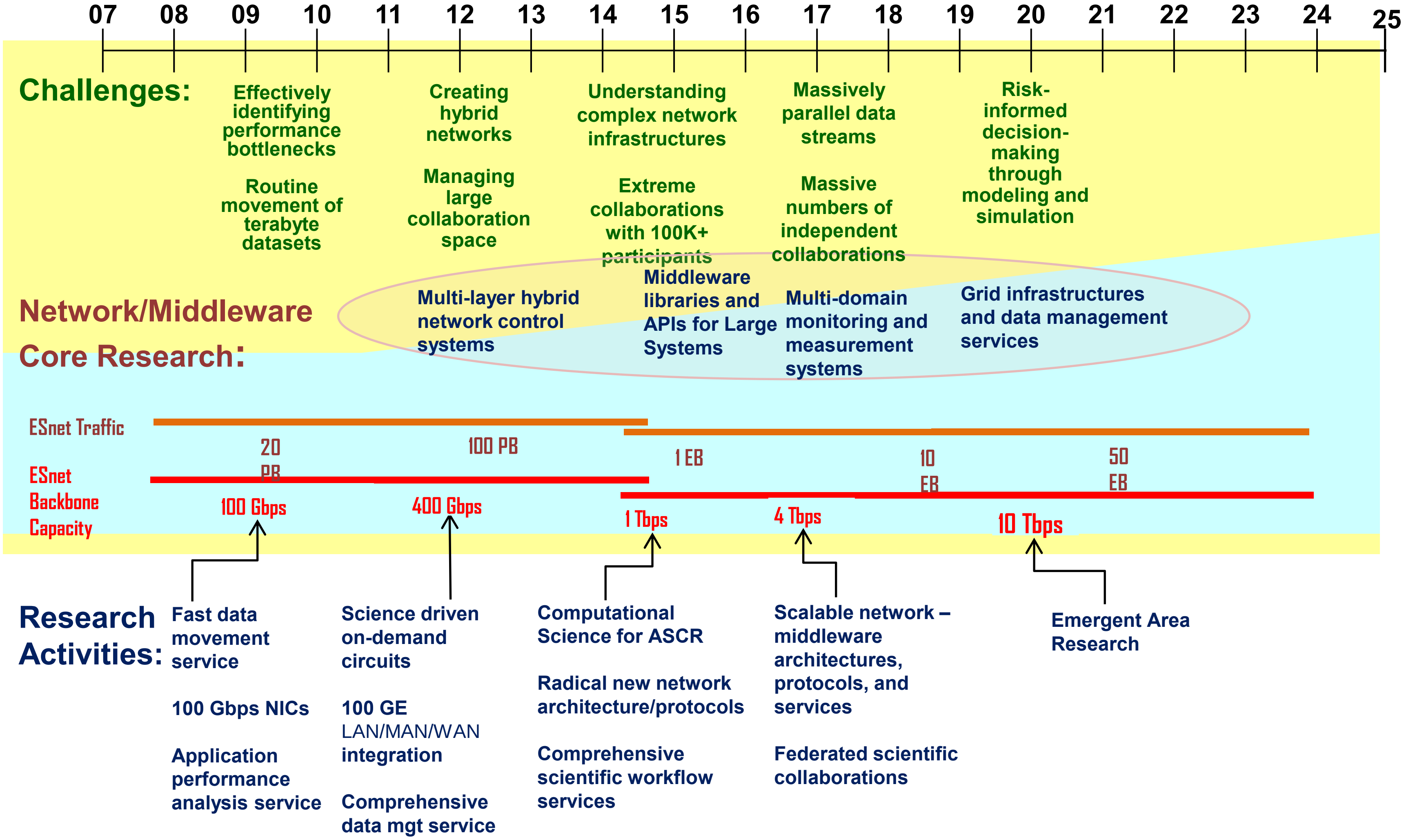
Magellan Status Report

A Test Bed to Explore Cloud Computing for Science

Shane Canon (Berkeley Lab)
Susan Coghlan (Argonne)

ASCAC
March 22, 2011

Themes and Portfolio Directions



Challenges:

Effectively identifying performance bottlenecks

Routine movement of terabyte datasets

Creating hybrid networks

Managing large collaboration space

Understanding complex network infrastructures

Extreme collaborations with 100K+ participants

Massively parallel data streams

Massive numbers of independent collaborations

Risk-informed decision-making through modeling and simulation

Network/Middleware Core Research:

Multi-layer hybrid network control systems

Middleware libraries and APIs for Large Systems

Multi-domain monitoring and measurement systems

Grid infrastructures and data management services

ESnet Traffic

20 PB

100 PB

1 EB

10 EB

50 EB

ESnet Backbone Capacity

100 Gbps

400 Gbps

1 Tbps

4 Tbps

10 Tbps

Research Activities:

Fast data movement service

100 Gbps NICs

Application performance analysis service

Science driven on-demand circuits

100 GE LAN/MAN/WAN integration

Comprehensive data mgt service

Computational Science for ASCR

Radical new network architecture/protocols

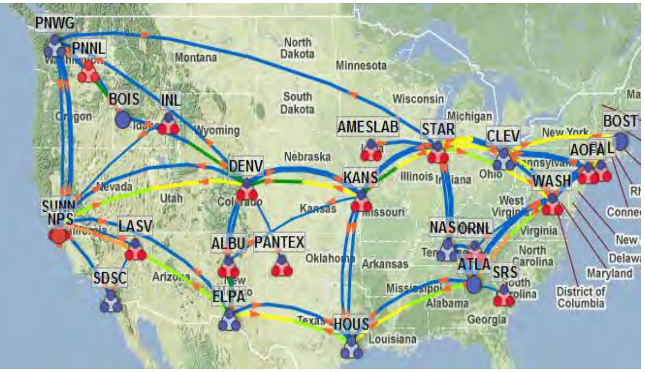
Comprehensive scientific workflow services

Scalable network – middleware architectures, protocols, and services

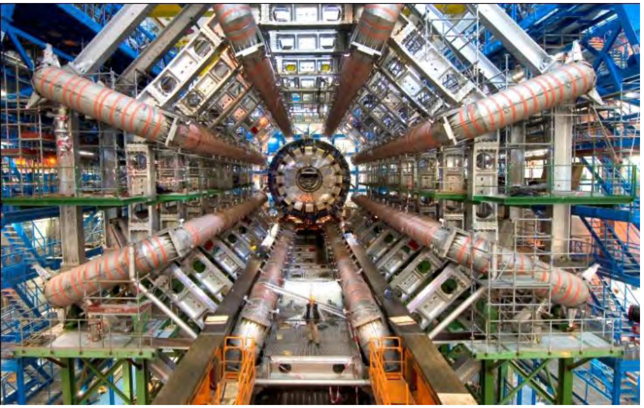
Federated scientific collaborations

Emergent Area Research

Challenges for Next-Gen Program



- **Develop a fundamental understanding about how DOE scientists use networks and how those networks behave**



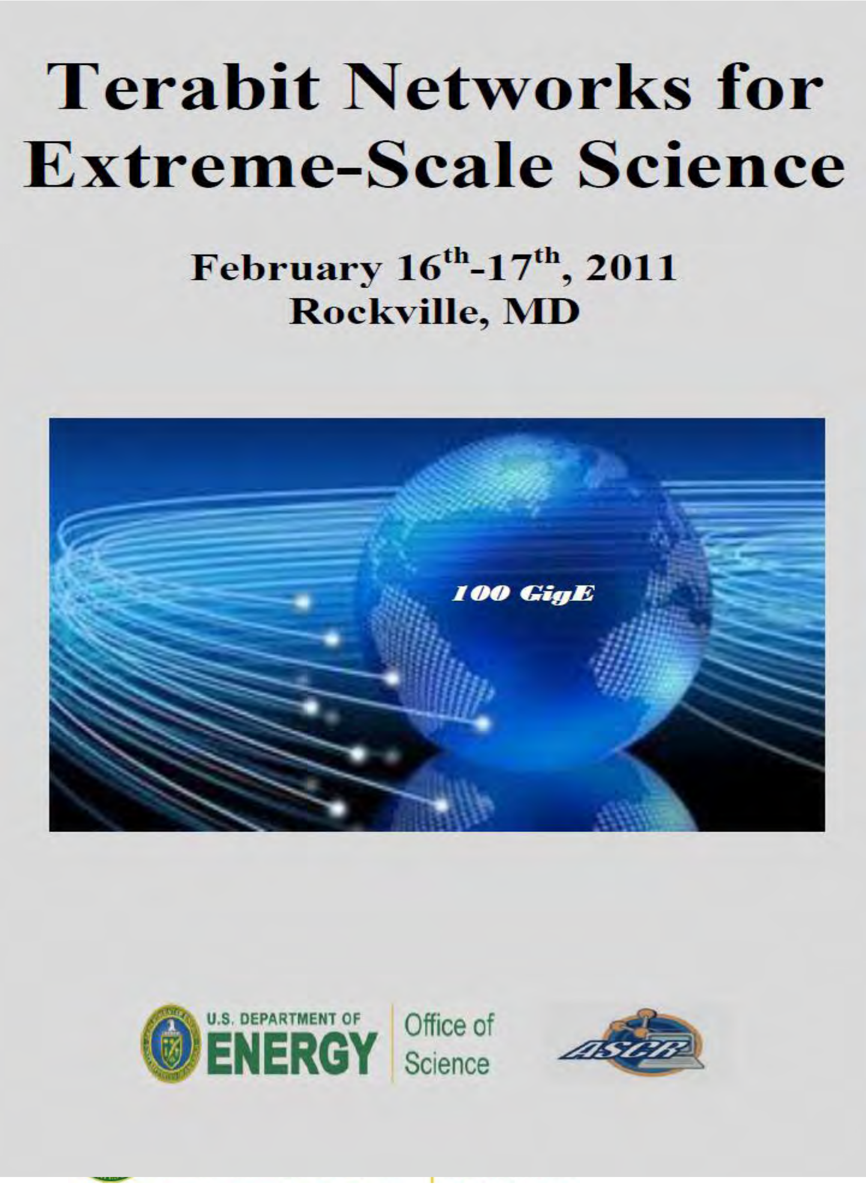
- **Provide scientists with advanced technologies that simplify access to experimental facilities, Supercomputers and scientific data**

- **Provide dynamic and hybrid networking capabilities to support diverse types of high-end science applications at scale.**

Future Planning Activities

- **A series of targeted workshop to identify significant research challenges in networks and collaborations**
  - Terabit Network Workshop Feb 16-17, 2011
- **Solicitations will be release as funding becomes available (tentatively looking at)**
  - FY11 – Focus on Terabit Network issues
  - FY12 – Focus on Data issues in DOE complex

Terabit Networks Workshop



Terabit Networks for Extreme-Scale Science

February 16th-17th, 2011
Rockville, MD

U.S. DEPARTMENT OF ENERGY | Office of Science

ASCR

Workshop Objective

- To identify the major research challenges in developing, deploying, and operating federated terabit networks to support extreme-scale science activities

- Participants from industry, academia, and nation laboratories (50 attendees)

- Major technical areas of discussions included
  - Scaling network architectures and protocols by several orders of magnitude over today's network performance.
  - Terabit LANs, host systems, and storage
  - Advanced traffic engineering tools and services

Terabit Network Workshop

- **3 Breakout Sessions**
  - Advanced User Level Network-Aware Services
  - Terabits Backbone Networking Challenges
  - Terabits End Systems (LAN's, Storage, File, and Host Systems)
  - Exascale Security Considerations (virtual session only)
- **1.5 days of in-depth discussions by each breakout group**
  - Numerous challenges identified by each group
  - Report to be issued shortly

- **https://indico.bnl.gov/conferenceTimeTable.py?confId=319**

FY10 FOA: High-Capacity Optical Networking and Deeply Integrated Middleware Services

FOA Topics:

1. Hybrid packet/circuit-switched networks
2. Multi-Layer Multi-Domain Dynamic Provisioning
3. 100 GE System-Level Network Components and Services
4. Multi-Layer Multi-Domain Network Measurement and Monitoring

FOA Areas:

a) High-Capacity Optical Networks

b) Deeply Integrated Middleware

Submissions/Budget

Total Submission - 31

- **High-capacity Networks: 16**
- **Deeply Integrated Middleware: 15**

Budget

- Total request: $31,274.45
- Recommended Budget:

  a) $900k/yr - (6 Labs)

  b) $690k/yr - (2 Universities)

Summary

- Research conducted by this program has enabled ESnet to deliver outstanding performance to DOE scientists (winning 2 industry awards)
- On-demand bandwidth (OSCARS) service is used daily by ESnet, Internet2, and other R&E networks
- Large globally distributed science communities (OSG, ESG) directly benefit from ASCR research
- ARRA investment will have direct impact on current and future network infrastructure

Conclusions

- Focus is on E2E performance in large distributed scientific computing environments
- Integrated approach combining Network, Middleware, and Collaboratory activities
- Interact with ESnet operations staff to bridge the "Valley of Death" issue
- Work with community to identify new challenges and fundamental research topics that will align program with ASCR's strategic objectives and DOE's science mission

- **Supplemental slides**

Portfolio Breakdown

- **Large Multi-institution Collaborations**
  - Center for Enabling Distributed Petascale Science
  - Earth Systems Grid
  - Open Science Grid
- **Multi-Investigator projects**
  - Network Weather and Performance Services eCenter
  - Virtualized Network Control
  - Integrating Storage Management with Dynamic Network Provisioning for Automated Data Transfers
  - End Site Control Plane Subsystem (ESCS)
  - Collaborative DOE Enterprise Network Monitoring Deployment

Portfolio Breakdown

- **Single Investigator projects**
  - Sampling Approaches for Multi-Domain Internet Performance Measurement Infrastructures to Better Serve Network Control and Management
  - Towards a Scalable and Adaptive Support Platform for Large-Scale Distributed E-Sciences in High-Performance Network Environments
  - Towards a Scalable and Adaptive Application Support Platform for Large-Scale Distributed E-Sciences in High Performance Network Environments
  - End Site Control Plane Subsystem (ESCPS)
  - Resource Organization in Hybrid Core Networks with 10 G Systems
  - Data Network Weather Service Reporting
  - Dynamic Provisioning for Terascale Science Applications Using Hybrid Circuit/Packet Technologies and 100G Transmission Systems
  - Dynamic Optimized Advanced Scheduling of Bandwith Demands for Large-Scale Science applications
  - Assured Resource Sharing in Ad-Hoc Collaboration
  - Orchestrating Distributed Resource Ensembles for Petascale Science
  - Detection, Localization and Diagnosis of Performance Problems Using perfSONAR
  - COMMON: Coordinated Multi-Layer Multi-Domain Optical Network Framework for Large-Scale Science Applications
  - VNOD: Virtual Network On-Demand for Distributed Scientific Applications

American Reinvestment and Recovery Act

- **Magellan**
  - Scientific Cloud Computing Research
- **ANI Projects**
  - Esnet upgrade and backbone infrastructure
  - Testbed infrastructure
  - Climate 100: Scaling the Earth System Grid to 100 Gbps Networks
  - End-System Network Interface Controller for 100 Gb/s Wide Area Networks
  - An Advanced Network and Distributed Storage Laboratory for Data Intensive Science
  - 100G FTP: An Ultra-High Speed Data Transfer Service Over Next Generation 100 Gigabit Per Second Network