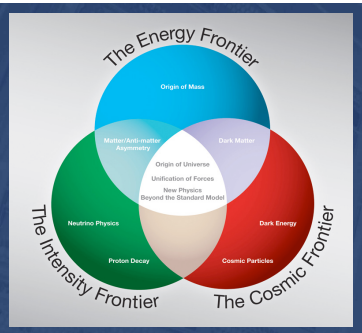
US Shared Cyber-Infrastructure

Ruth Pordes
Associate Head, Computing Division
Fermilab,
Executive Director Open Science Grid

(& member of CMS LHC experiment and advisory to NEES earthquake engineering simulation CI.)



an intro..Fermilab

Fermilab is the U.S. High Energy/Particle Physics User Facility.

>2300 users from around the world. Runs the Tevatron. <~2000 staff - ~1000 physicists, engineers and computing professionals. Last year ~50 students got doctorates based on work at Fermilab & ~1,000 college students in onsite programs.

Funded by the DOE through the Fermilab Research Alliance (FRA). Some NSF funding small projects, R&D.

Worldwide collaborations, support includes data processing and analysis development for HEP.

Early contributors for Grids in HEP in DOE and NSF leading to sustained infrastructure since 2003.

Grid2003 -> OSG

From *symmetry breaking*

Interesting effect at the Tevatron hints at new physics

Fermilab's Wilson Hall in the shape of a t, the symbol for the top quark.

Scientists are finding signs of new physics through the study of a particle Fermilab physicists discovered at the Tevatron, the top quark.

When top quarks and their anti-particles, anti-top quarks, are created in particle collisions at the Tevatron, detectors note the direction in which they fly. Theory predicts that the particles will favor one direction slightly over the other, traveling that way about 5 percent of the time more.

However, in studies by the DZero collaboration and the CDF collaboration, the particles seemed to be picky 15 percent of the time. Top quarks went forward and anti-top quarks went backward. This month, the CDF collaboration announced results with an even larger asymmetry.

an intro..Open Science Grid

Distributed high throughput services that integrate distributed, reliable and shared resources to support computation at all scales. Supports serial to "whole node" jobs.

US contribution to the World Wide LHC Computing Grid. Supported by both DOE and NSF. Continuous production operations, Service Level Agreements written for OSG and major stakeholders for most operations services (and operate as "Information Technology Infrastructure Library (ITIL)-like" processes").

18 research Virtual Organizations, 8 are Campus/Regional based and are multi-disciplinary. Usage 50% LHC, 30% other physics (Tevatron, LIGO, 20% other sciences (molecular biology, mathematics, earthquake simulations)

95 Lab and University resources.

Includes ~40 LHC Tier-3 sites.

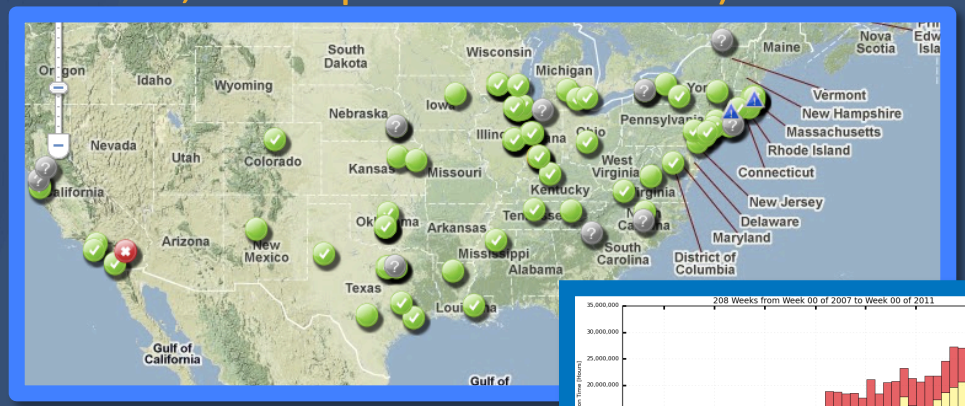
A few non-US sites.

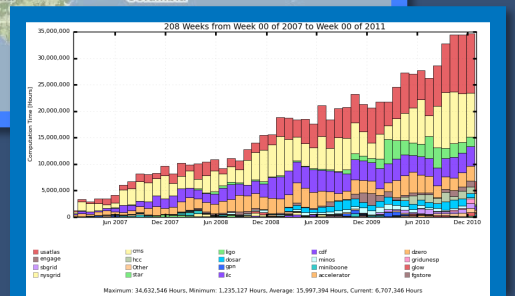
Per Day: ½ million jobs

1 million CPU hours;

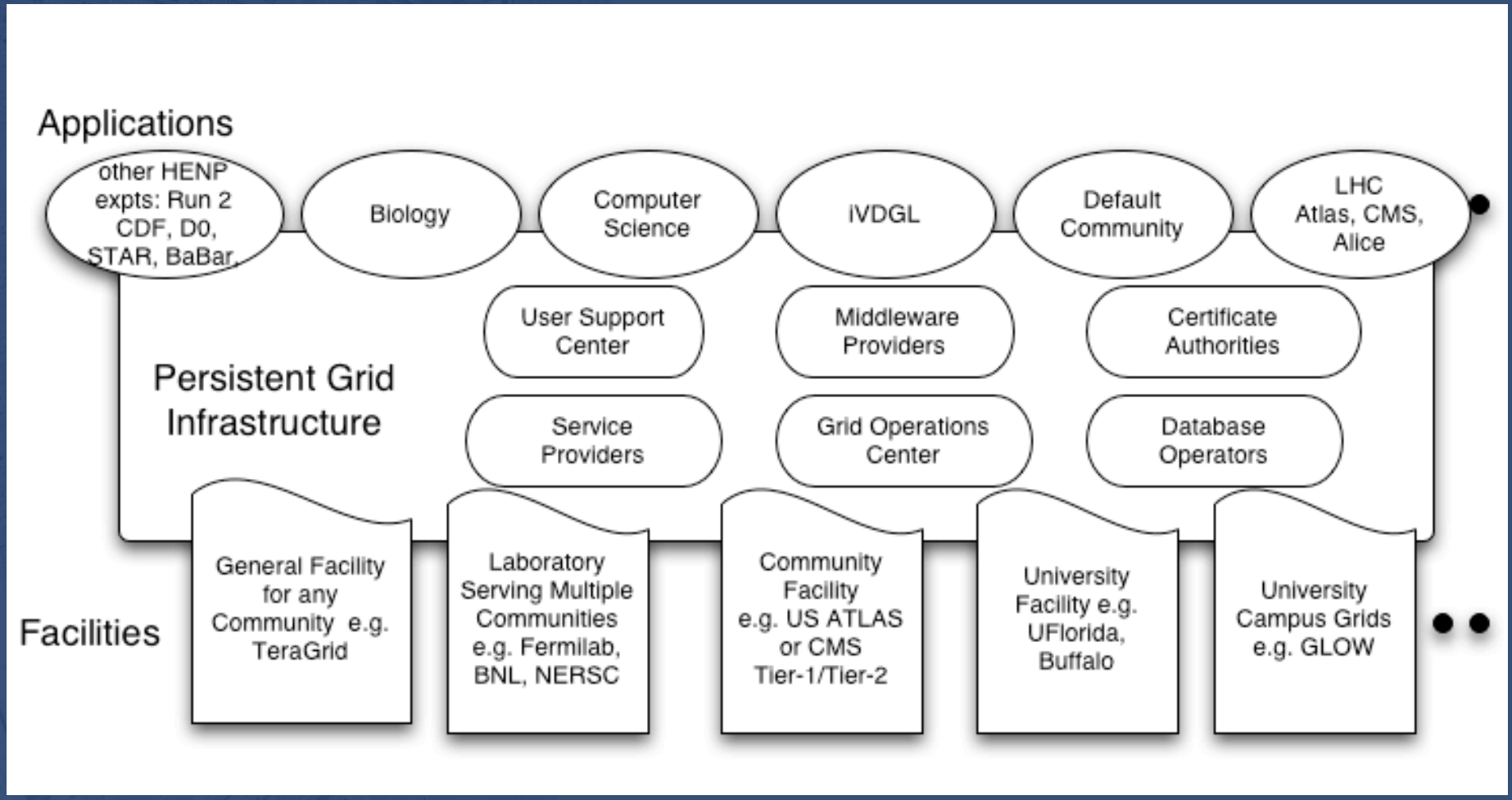
1 million files transferred;

4 Petabyte data transferred

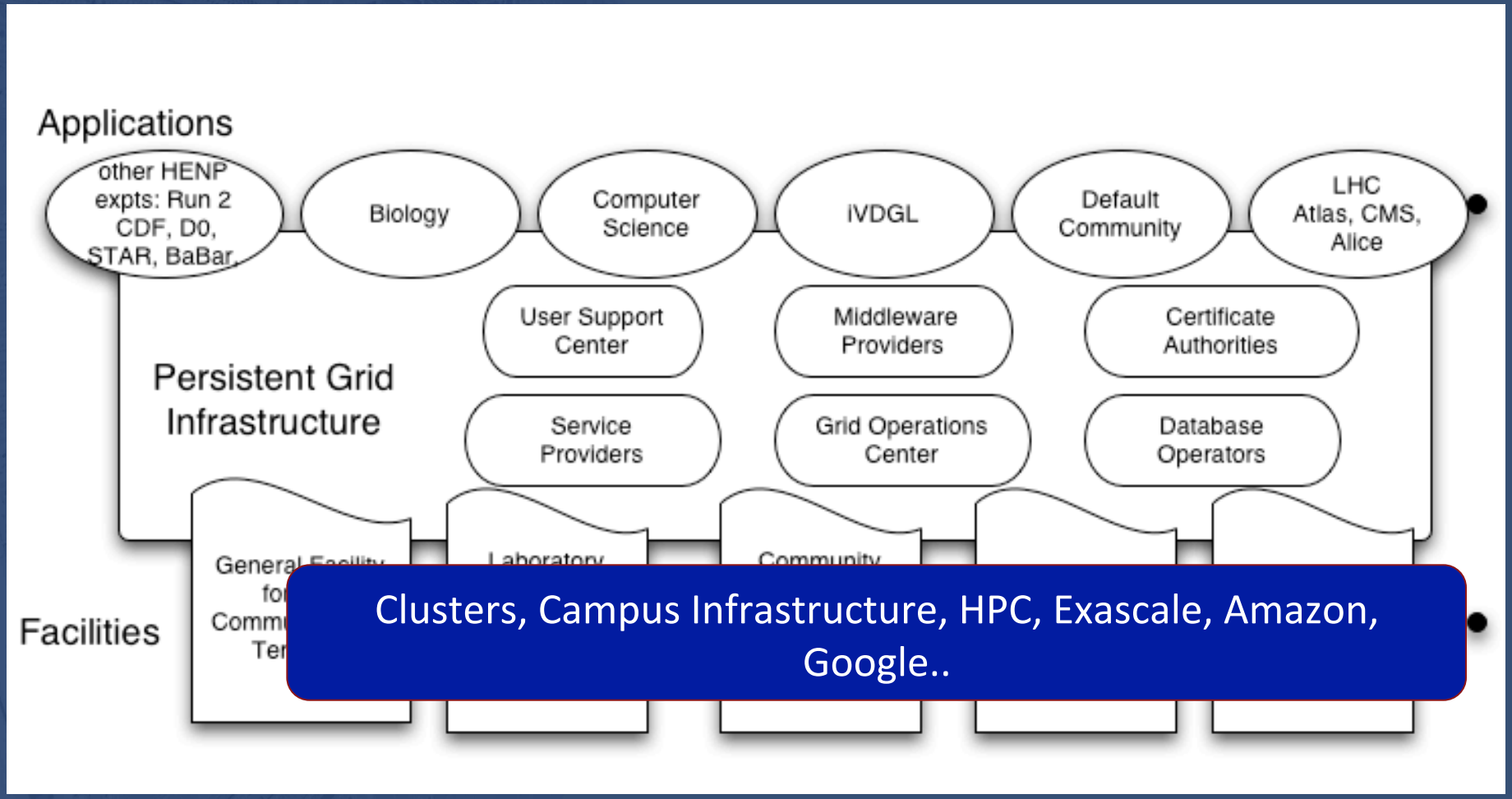




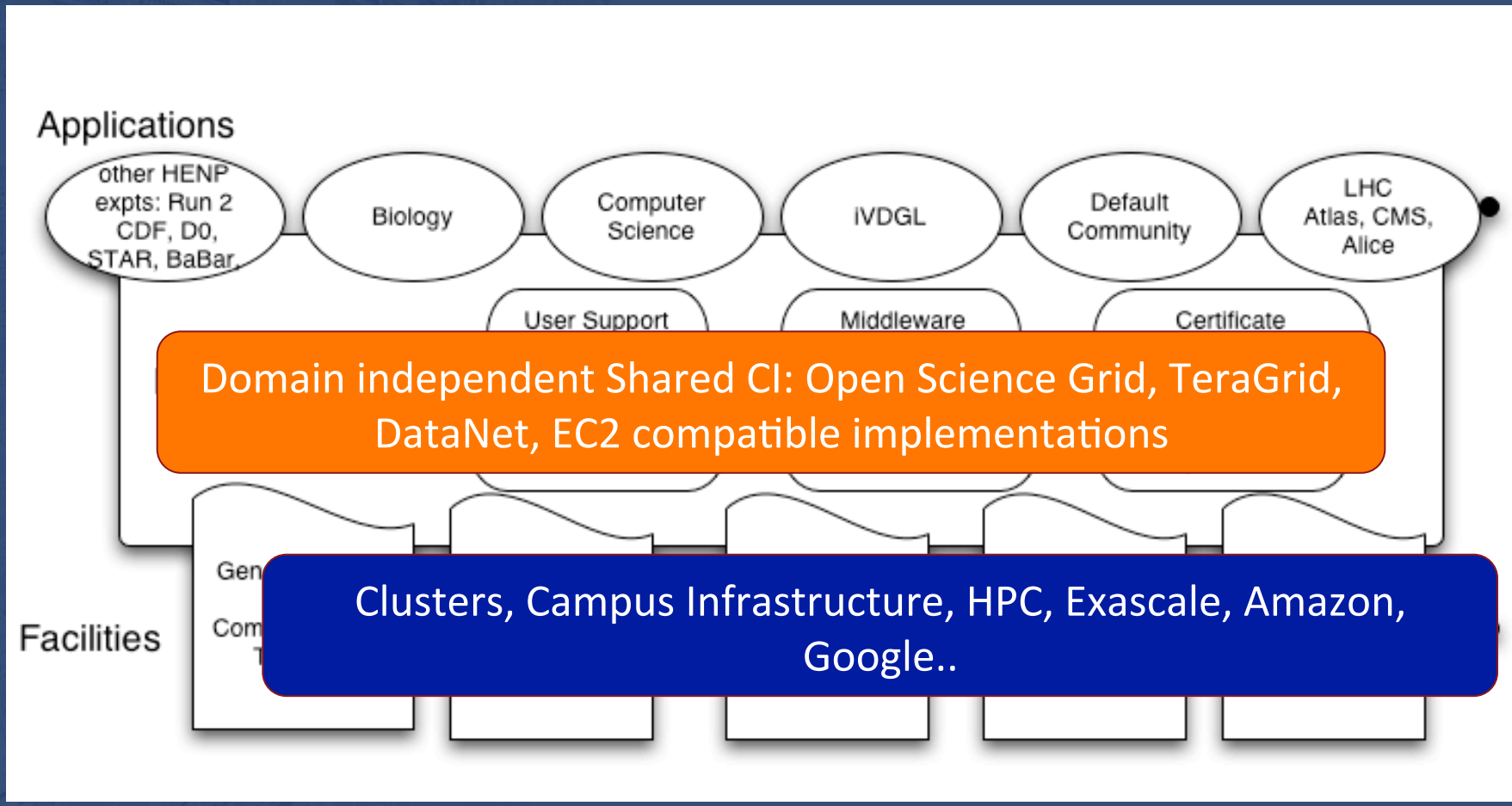
US Shared CI – circa 2005
Open Science Grid



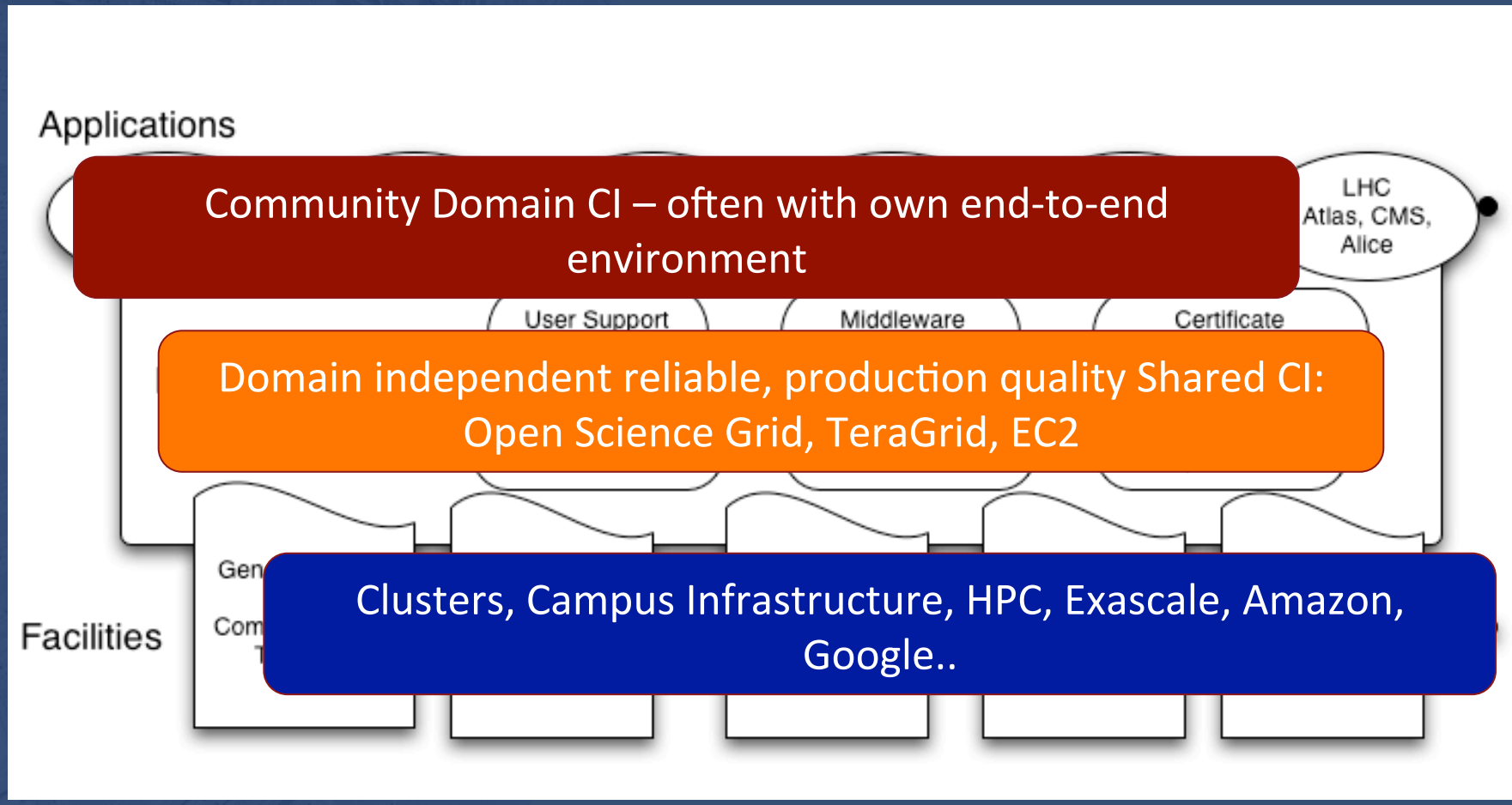
US Shared CI -2011 Extension and Expansion



US Shared CI -2011
Multi-Science Domain CIs

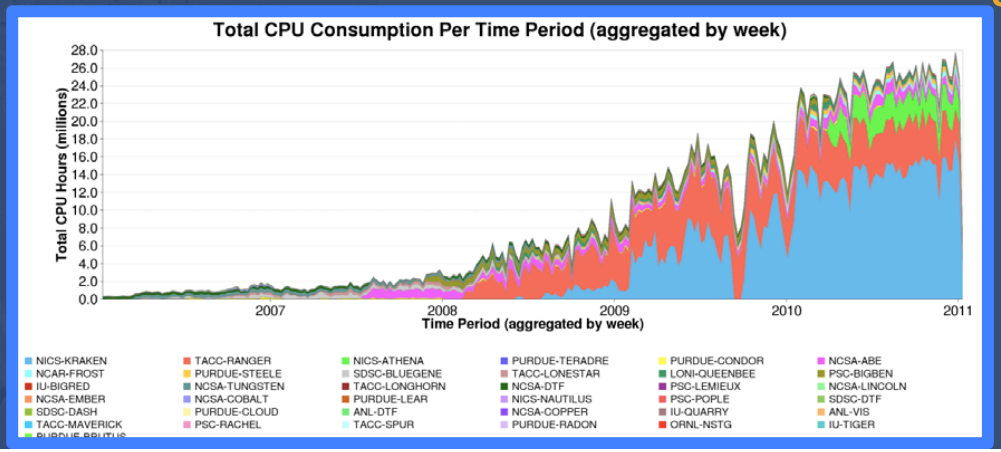


More Large Domain Communities have own Shared CI middle tier



Intro to.. TeraGrid

- 9 HPC Resource institutions (IU, LONI, NCSA, NCAR, ORNL, PSC, Purdue, SDSC, TACC). Supports jobs up to >1000 core.
- ~3.5 million cpuhours/day
- >30 science gateway community portals.
- 2010Q4: ~319 institutions with users per quarter. (~50 institutions responsible for 80% of the total).
- shared file systems on/across most resources.
- ssh and GRAM interfaces supported for job submission.
- GridFTP for data transfer and iRODS for data management.



TG and OSG

TeraGrid -> XD

- 5 year Xsede proposal being presented to National Science Board in May. Tentative start of project July 1 2011. Transition planning in full swing.
- OSG is an XD Service Provider with a seat on the Service Provider Board. Small XD effort proposed towards OSG users: Campus Champions, User Support.

Cross-infrastructure usage and joint activities continuing to expand:

- Usage of both OSG and TG resources: e.g. Earthquake simulations, molecular biology.
- ExTENCI project to jointly harden and deploy LUSTRE-WAN, Virtual Machines/Science Clouds.
- Plan to understand how to better accommodate both allocation methodologies.

OSG -> OSG'

- Current 5 year project finishes in 10/211.
- 1 year DOE HEP extension from 3/11 to 3/12.
- 5 year NSF proposal for ~40% of full need.
- DOE ASCR SciDAC-3 proposal for s/w research deployment institute 5/2/11

Sustain and evolve services – in collaboration with VOs and, for some, CERN

- Data Center Quality Infrastructure
- Ubiquitous access to data "anywhere, anytime"
- Improved integration with and cross-facility support for muti-core, clouds, advanced networks, virtualization, preservation
- Scale number of jobs/users/cpu

Goal to make additional contributions to

- Campus infrastructures,
- Distributed High Throughput Computing (DHTC) Technologies,
- Software usability and sustainability.

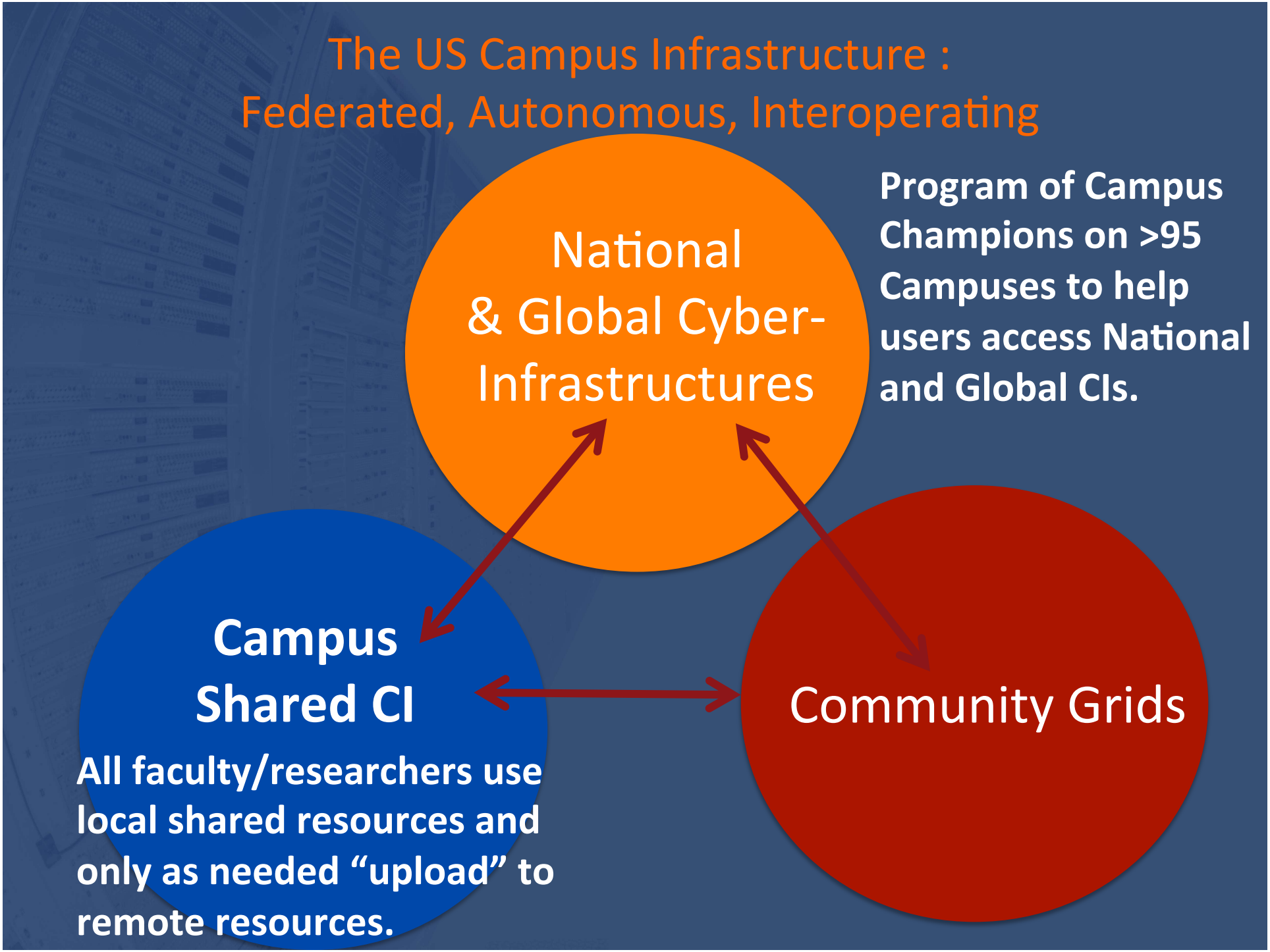
Interoperation

- Long history of interoperation with EDG, EGEE I-III
- Continue interoperation with EGI operations, software and policy services
- Continue interoperability of softwar services with EMI (gLite & ARC) releases.
- Increase usability across OSG and XD.

Future challenges as OSG and TG evolve:

- TG/XD software services to include Unicore, Genesis II, ?Nimbus?.
- OSG VDT to provide access to clouds, support inter- and intra-campus shared CI, leverage EPEL & Fedora for s/w packaging and distribution.
- Data interchange.

Future challenges as OSG and EGI evolve?


The US Campus Infrastructure :
Federated, Autonomous, Interoperating
National
& Global Cyber-
Infrastructures
Program of Campus Champions on >95 Campuses to help users access National and Global CIs.
Campus
Shared CI
All faculty/researchers use local shared resources and only as needed "upload" to remote resources.
Community Grids

OSG services on the Campus

- Support for local campus identity management services (no need grid certificates.)
- An integrated software package that does not require root privileges, easily installed by a "lightly trained" campus researcher.
- Coordinated education, training and documentation.
- A campus job submission point capable of routing jobs to multiple batch schedulers.
- An OSG job submission point that can route jobs directly to on-campus resources.
- Prototype implementation on 6 campuses uses Condor glidein and flocking capabilities.

Executive Summary

Prominent researchers in many fields have written about the disciplinary imperatives behind investment in cyberinfrastructure for research and education. In "A Research Cyberinfrastructure Strategy for the CIC," we provide a CIO perspective on how to respond to these imperatives, including ideas about how to monitor the changing environment and continuously adjust to it.

Our goal should be to enable scholarship at the cutting edge of every discipline, while getting as much value as possible from every dollar spent on cyberinfrastructure. The CIC campuses are very richly endowed with cyberinfrastructure resources but can be even more effective by ***adopting good practices*** that support greater coordination at every level:

- Plan.
- Share (at the highest level possible).
- Design funding models that promote scholarship and stewardship.
- Rely on user governance.
- Conduct cyberinfrastructure impact analysis.

CIO input to the Provosts

Over the long run, these practices should help our institutions produce more scholarship per dollar invested. In the near term, the perspectives underlying these recommended practices allow us to identify six high-priority enhancements of campus cyberinfrastructure that will be necessary if we are to meet the present needs of researchers and the expectations of funding agencies. ***We should all currently be investing in:***

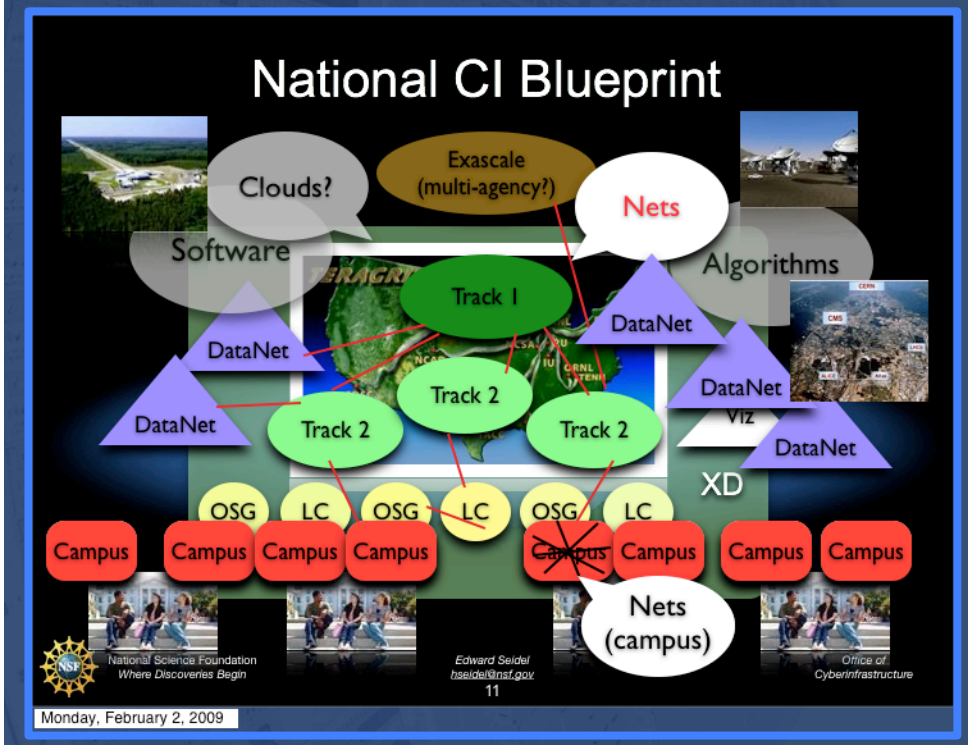
- Preparing for "federated identity management" and other enabling technologies for virtual organizations.
- Maintaining state-of-the-art communication networks.
- Providing institutional stewardship of research data.
- Consolidating computing resources while maintaining diverse architectures.
- Expanding cyberinfrastructure support to all disciplines.
- Exploring cloud computing to manage long-term financial commitments.

The CIC's well-established pattern of IT-related collaboration puts our institutions in a position to contribute more together than the sum of our individual contributions. We recommend leveraging this strength to amplify the value of every dollar of cyberinfrastructure investment along with the following ***Provostial actions:***

- Build interest among other Provosts in federated identity management.
- Review recruiting and retention practices that affect campus cyberinfrastructure.
- Continue supporting competition as a consortium.
- Influence federal spending rules.

>300 CI "ready" Campuses in the US

Computing Resources



Tiered approach with continuous renewal of Track1 and Track2 systems.

Petascale machines on TeraGrid/XD.

Agreements to support testing usage of commercial clouds.

University Clusters for specific research goals and local sharing.

Requirement for open data and data retention beyond the end of the project from all new projects.

Datanet systems, digital libraries to provide multi-science domain data archiving and curation.

Computing Resources

Leadership class machines at Argonne (ANL) and Oakridge National Labs.

Multi-purpose HPC at Lawrence Berkeley National Lab (NERSC).

Single purpose resources at Brookhaven, Fermilab, SLAC, JLAB, NREL Labs.

Investigations of GPUs and application specific hardware.

R&D for Terapath networking.

Short term Scientific Cloud evaluation at ANL and NERSC. Looking at DOE wide agreement for Amazon costs.

Major focus on Research to stand up Exascale computing - address performance growth of multiple 10 Petascale computers - in ~2014

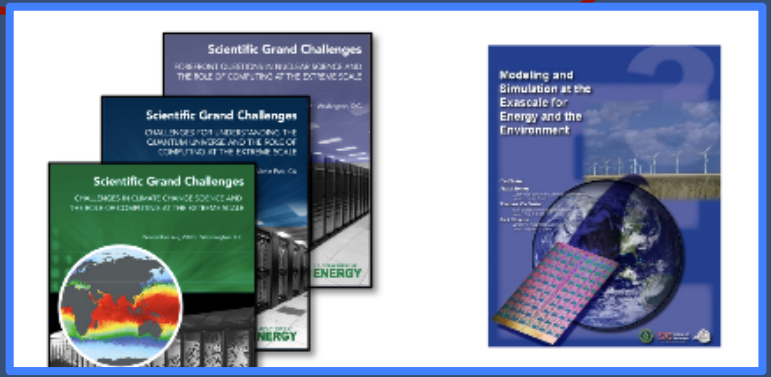
Software

Co- design partnerships between application, applied maths, computer science towards Exascale.

International cooperation for Exascale software (exascale.org):

Processors, compilers, tools, data and workflow management.

Application adaptation, new algorithms and methods, performance and fault tolerance

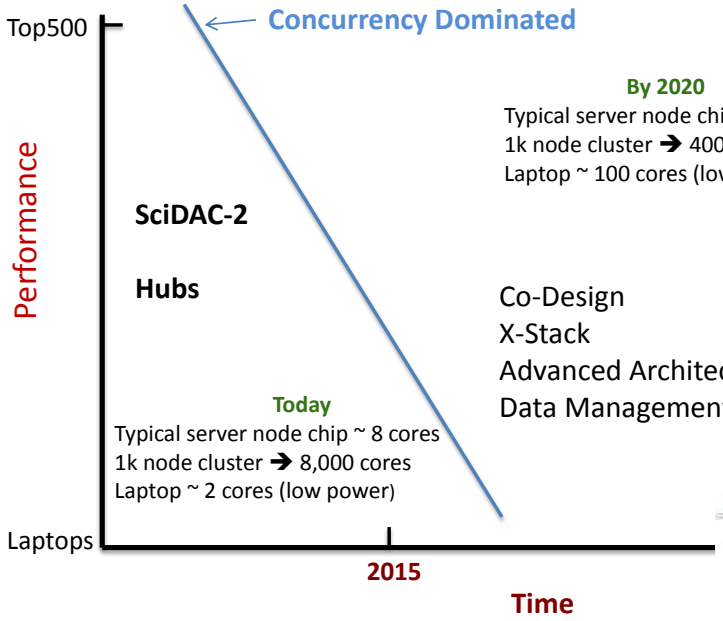


Integration of 100 Gigabit and then 1000 Gigabit network capabilities into drivers, OS, middleware, applications over next few years.

Issues of sustaining current software stacks relied on by the scientific community.

Federated data archives, curation and access

Phase Change in Computing



U.S. DEPARTMENT OF ENERGY | Office of Science

Ensure performance enhanced by the revolution in number of cores

Challenges to Exascale

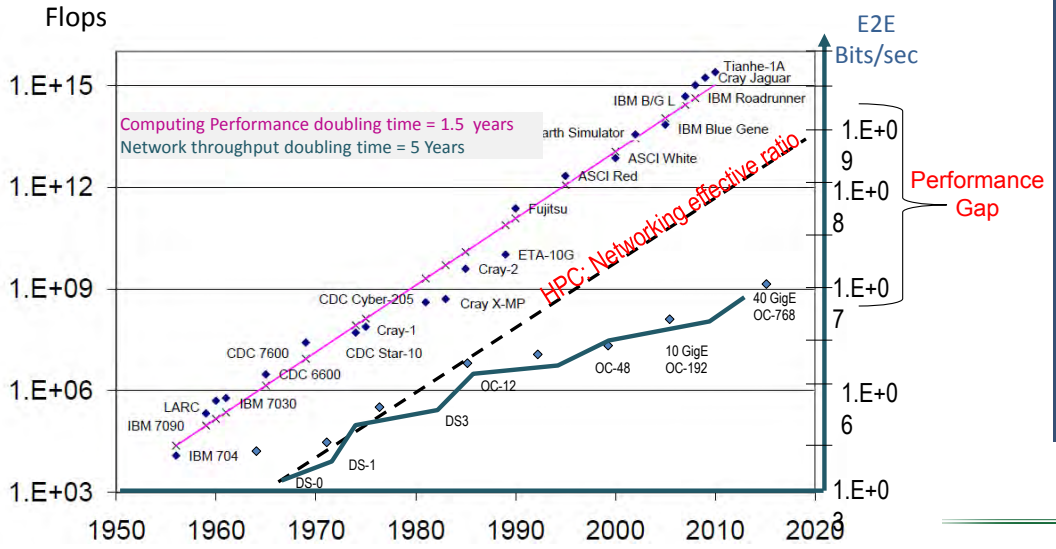
Performance Growth

1) **System power** is the primary constraint
2) **Memory** bandwidth and capacity are not keeping pace
3) **Concurrency** (1000x today)
4) **Processor** architecture is an open question
5) **Programming model** heroic compilers will not hide this
6) **Algorithms** need to minimize data movement, not flops
7) **I/O bandwidth** unlikely to keep pace with machine speed
8) **Reliability and resiliency** will be critical at this scale
9) **Bisection bandwidth** limited by cost and energy

Unlike the last 20 years most of these (1-7) are equally important across scales, e.g., 100 10-PF machines

Data-Intensive Science:

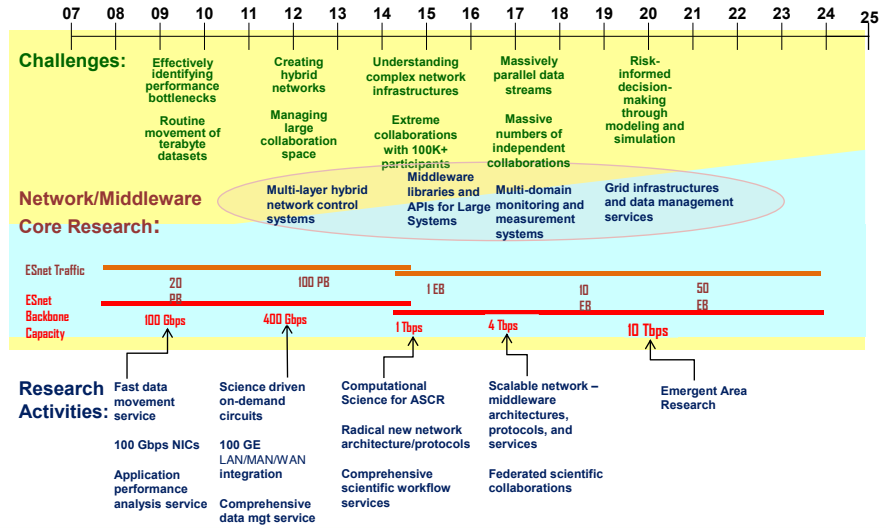
HPC performance to : Networking Performance Gap



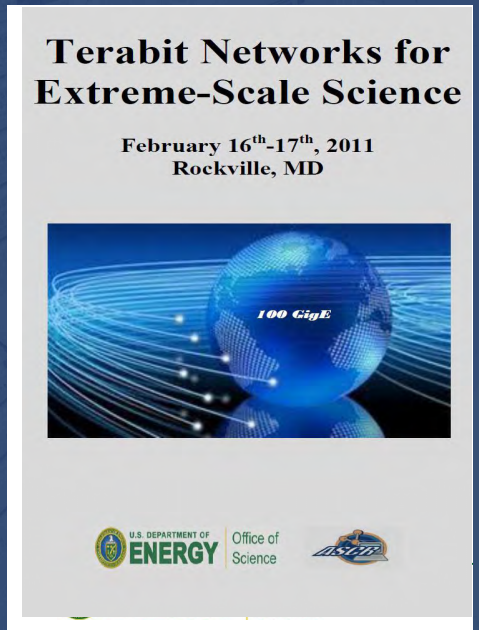
U.S. DEPARTMENT OF ENERGY | Office of Science

Ensuring advanced networks are available and usable.

Themes and Portfolio Directions



U.S. DEPARTMENT OF ENERGY | Office of Science



Cloud Resources

Early adopters of Commercial Clouds from our field

- STAR Nuclear Physics, Belle-2, Harvard Medical School use of Amazon
- US ATLAS & US CMS testing of job submission infrastructure with EC2 interfaces.
- Investigation of DOE wide agreement with Amazon
- NSF support for testing of Microsoft azure

February 4, 2010

Microsoft Corp. and the National Science Foundation (NSF) today announced an agreement that will offer individual researchers and research groups selected through NSF's merit review process free access to advanced cloud computing resources. By extending the capabilities of powerful, easy-to-use PC applications via Microsoft cloud services, the program is designed to help broaden research capabilities, foster collaborative research communities, and accelerate scientific discovery. Projects will be awarded and managed by NSF. More details about funding opportunities are available at http://www.nsf.gov/dir/index.jsp?org=CISE.

Microsoft will provide cloud computing research projects identified by NSF with access to Windows Azure for a three-year period, along with a support team to help researchers quickly integrate cloud technology into their research. Windows Azure provides on-

Scientific Clouds – Nimbus, Eucalyptus, Open Nebula, OpenStack

- Specific installations: DOE Magellan at ANL and NERSC; FutureGrid XD testbed
- Overlaid on existing facilities: As part of ExTENCI project, CMS use of WISPY@Purdue/Teragrid; FermiCloud/OSG

Standards

- Being developed by Open Grid Forum.
- National Institute of Standards actively participating and evaluating clouds for e-government.

NIST Cloud Computing Collaboration Site

About | Reference Architecture | SAJACC | Security | Standards Roadmap | Business Use Cases

Useful Documents for Cloud Adopters

NIST publications, work-in-process, and working group products that may be helpful to US government agencies and others in making decisions regarding the effective and secure implementation of cloud computing. To actively participate on this wiki collaboration site and NIST Cloud Computing mailing list, please register through NIST Cloud Computing website.

NIST Cloud Computing Program Strategic Efforts

Business Use Case Working Group

- Twiki: Business Use Cases
- [Work-In-Progress] DRAFT Cloud Computing Business Use Case Template
- [Work-In-Progress] FGDC GeoSpatial GeoCloud Business Use Case
- [Work-In-Progress] USAID Virtual Desktop Infrastructure (VDI) Business Use Case
- [Work-In-Progress] USAID Applications (Email, productivity, collaboration) Business Use Case
- [Work-In-Progress] STIDS Incident Response Business Use Case

Communities: Bio Energy Centers

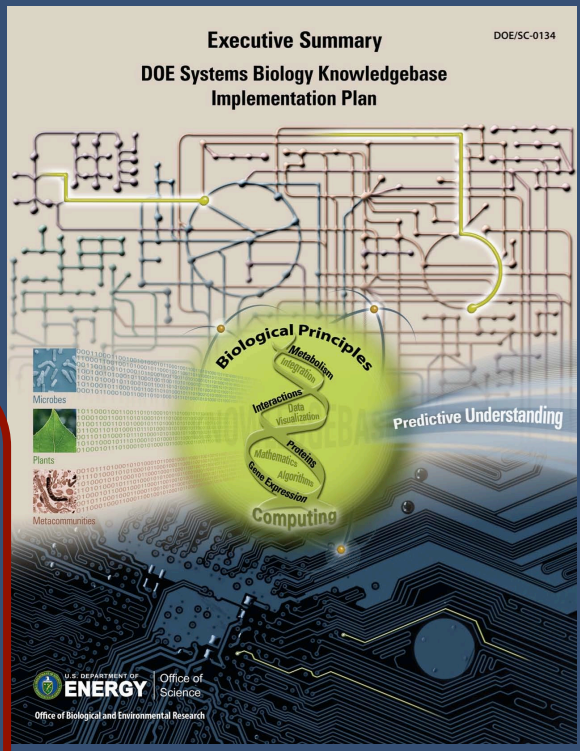
Development of next-generation bioenergy crops

Discovery and design of enzymes and microbes with novel biomass-degrading capabilities

Development of transformational microbe-mediated strategies for biofuel production

The three centers are

- BioEnergy Science Center (BESC) led by Oak Ridge national Lab
- Great Lakes Bioenergy Research Center (GLBRC)
- Joint BioEnergy Institute, led by LBNL

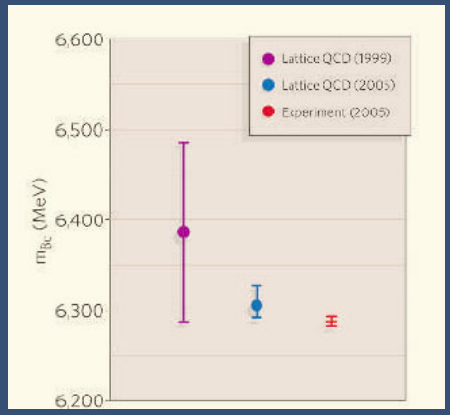


Systems Biology Knowledge Base Initiative: Implementation Plan from the Community. initial projects in proposal stage.

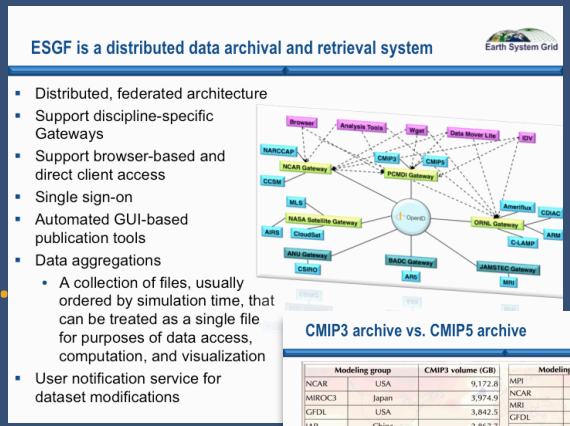
"provide the research community with a comprehensive cyberinfrastructure to advance systems biology research over the next several years. It articulates the scope and plans necessary to begin the Kbase effort and outlines a strategy for Kbase support of key research objectives in the microbial, plant, and metacommunity sciences"

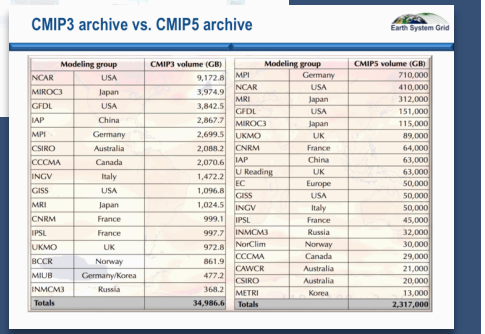
Other Communities

- US Lattice Quantum Chromodynamics Consortium.
- Particle and nuclear physics. US LHC, RHIC, Neutrino
- Astrophysics – experiment and computational.
- Earth System Grid (ESGF): Globally Federated Databases and data access.
- Energy, Chemistry..



Llattice QCD calculation of Allison *et al.*[1] (central point) for the mass of the B_c meson compared with an earlier calculation[7] (left) and the experimental measurement of Acosta *et al.*[2] (right).





Modeling group		CMIP3 volume (GB)
NCAR	USA	9,172.8
MIROC3	Japan	3,974.9
GFDL	USA	3,842.5
IAP	China	2,867.7
MPI	Germany	2,699.5
CSIRO	Australia	2,088.2
CCCMA	Canada	2,070.6
INGV	Italy	1,472.2
GISS	USA	1,096.8
MRI	Japan	1,024.5
CNRM	France	999.1
IPSL	France	997.7
UKMO	UK	972.8
BCCR	Norway	861.9
MIUB	Germany/Korea	477.2
INMCM3	Russia	368.2
Totals		34,986.6

Modeling group		CMIP5 volume (GB)
MPI	Germany	710,000
NCAR	USA	410,000
MRI	Japan	312,000
GFDL	USA	151,000
MIROC3	Japan	115,000
UKMO	UK	89,000
CNRM	France	64,000
IAP	China	63,000
U Reading	UK	63,000
EC	Europe	50,000
GISS	USA	50,000
INGV	Italy	50,000
IPSL	France	45,000
INMCM3	Russia	32,000
NorClim	Norway	30,000
CCCMA	Canada	29,000
CAWCR	Australia	21,000
CSIRO	Australia	20,000
METRI	Korea	13,000
Totals		2,317,000

NIH focused Projects
(I know much less about this)

Large scale computing still mainly depends on existing/otherwise provided infrastructures.

Sharing makes significant progress while same challenges:

Biomedical Informatics Research Network (BIRN) – **consolidated focus** on " data-sharing infrastructure, software tools, strategies and advisory services – all from a single source.."

Cancer Biomedical Informatics Grid (CABig) – **address core mission** (2011 review report)

Protein Data Bank –**sustained for 40 years** to date ..

How is the current BIRN model different from its previous one?

We have refocused BIRN's offerings in significant ways. Most importantly, BIRN is now open to all biomedical research groups, regardless of specialty. We also see our mission as supporting efficient data transfer, in observance of Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy and security guidelines – ***not*** as a creator of, or repository for, user data.

As a result, BIRN is no longer provides hardware, and does not offer or maintain servers (previously called "racks") for storing user information. Nor do we use a system in which participants' computers function as data "nodes" on a unified computing grid.

The WG would like to stress that, going forward, the creation of an infrastructure for data collection, management, analysis, and dissemination remains a critical and only partially addressed problem. It is thus critical that the WG's findings about the caBIG® program's progress and traction does not diminish NCI's enthusiasm for and commitment to supporting this critical area of development. Specifically, we recommend that caBIG® return to its original mission and premises and that NCI focus separately on informatics tools for clinical and basic research components. The former should become more driven by the requirements of the organizations that run clinical studies. The latter should be better integrated with NCI's existing

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the **wwPDB**, the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

CI included in all programs.

e.g. large Facilities manual

These resources, many of which are now in development, are collectively known as "cyberinfrastructure."

II. Planning and Managing for the MREFC

INTRODUCTION

NSF investments provide state-of-the art infrastructure for research and education, such as laboratory and field instrumentation and equipment, multi-user research facilities, distributed instrumentation networks and arrays, and mobile research platforms. In addition, investment is increasing in highly sophisticated information technology (IT)-based infrastructure, including distributed sensor networks, extensive data-storage and transmission capabilities, advanced computing resources, and Internet-based distributed user facilities.[7]

Addressing need for sustainability and evolution

The operations stage of large facilities is typically more than 10 years, with significant variation within NSF's portfolio.[50] To stay at the research frontier, upgrades and renewal of equipment are usually required. In the case of an observatory, this may include new instruments and cameras. For a sensor network, it may include the deployment of additional sensors or renewal of cyberinfrastructure. At an accelerator facility, the upgrades may take the form of higher energy or luminosity or new detectors. In general, these upgrades and renewals will be funded from R&RA funds, often from a portion of the operating funds designed for such purposes. Funding

Science Community CI – a sample from large to small (not in order):

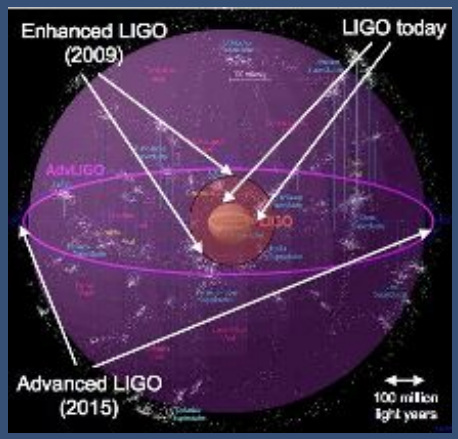
- Large Synoptic Survey Telescope – LSST
- US LHC
- Advanced Large Interferometer Gravitational Wave Observatory – Advanced LIGO.
- Ocean Observatories Initiative
- Dark Energy Survey
- iPLANT

Mission and Vision
The iPlant Collaborative (iPlant) will foster the development of a diverse, multi-disciplinary community of scientists, teachers, and students, and a cyberinfrastructure that facilitates significant advances in the understanding of plant science by this community through the application of computational thinking and approaches to Grand Challenge problems in plant biology and computer science.

- Network for Earthquake Engineering Simulation - NEEScom
- National Ecological Observatory Network - NEON
- Great Lakes Ecological Observatory Network.
- Incorporated Research institutions for Seismology (IRIS)
- Social Sciences

  …
- Coming?:
  - Combustion

Advanced LIGO



Revolution in sensitivity via the start of data taking of Advanced LIGO, Advanced VIRGO, Large-scale Cryogenic Gravitational wave Telescope, and possibly the Australian Interferometric Gravitational Observatory.

The search for gravitational waves transitioning from the quest for "first observation" to a standard tool of astronomy.

30 year roadmap document : "an inter-continental scale network made up of advanced interferometers in the US, Europe, Asia, and Australia would provide an all-sky array that could detect, decode and point to the sky-position of gravitational wave sources in the audio bandwidth where many of the most interesting sources are located".

Computation needs, rapid-response computations will increase vastly

Network for Earthquake Engineering Simulation

CI Goals

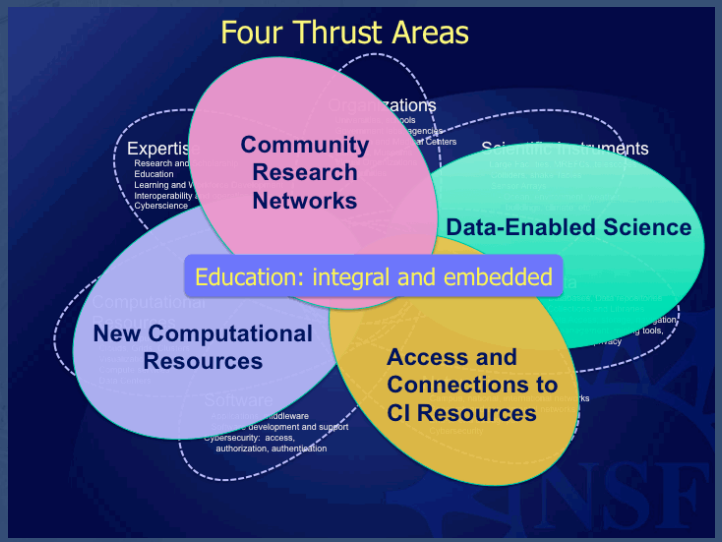
- Characterize and curate measurement data from 1-1000 sensors, lasting seconds to days, up to 20,000 files per project to date from 14 shaketable and other collection sites..
- Data Repository – NEES Project Warehouse, 6 TB in 10/10; 10 TB in 2/11.
- Tools to search, display, analyze, manipulate data - data ranges from photos and movies to xls spreadsheets and text documents.
- Computational simulation – including hybrid simulations where shake table measurements and computational simulations "cooperate".
- Extensive education, outreach and community building.

NSF Planning for the Future – CF21

As a logical next step, it is imperative that NSF develop a strategic long term vision of what is being called a *Cyberinfrastructure Framework for 21st Century Science and Engineering (CF21).* This vision will need to address concepts and capabilities such as,

1. High end computational, data, visualization and sensor-based systems and the associated user support needed for transformative science; HPC systems and services of all types and flavors, networking, interoperable data systems and mining, including a focus on sustainability and extensibility.
2. Activities that link cyberinfrastructure framework into campuses (including government and business) and programs that provide the widely dispersed, most broadly based activities and resources; grids, cloud computing, loosely coupled campus services, federated ID management and hybrid networks involving wireless and social networks.
3. Major national and international research facilities and collaborations including large-scale NSF collaborative facilities and projects, cyberinfrastructure developed across NSF and other government agencies, including international partners.
4. A comprehensive plan for education and outreach in computational science to support learning and workforce development for 21st century science and engineering.

Overarching goals applied to new projects and programs.

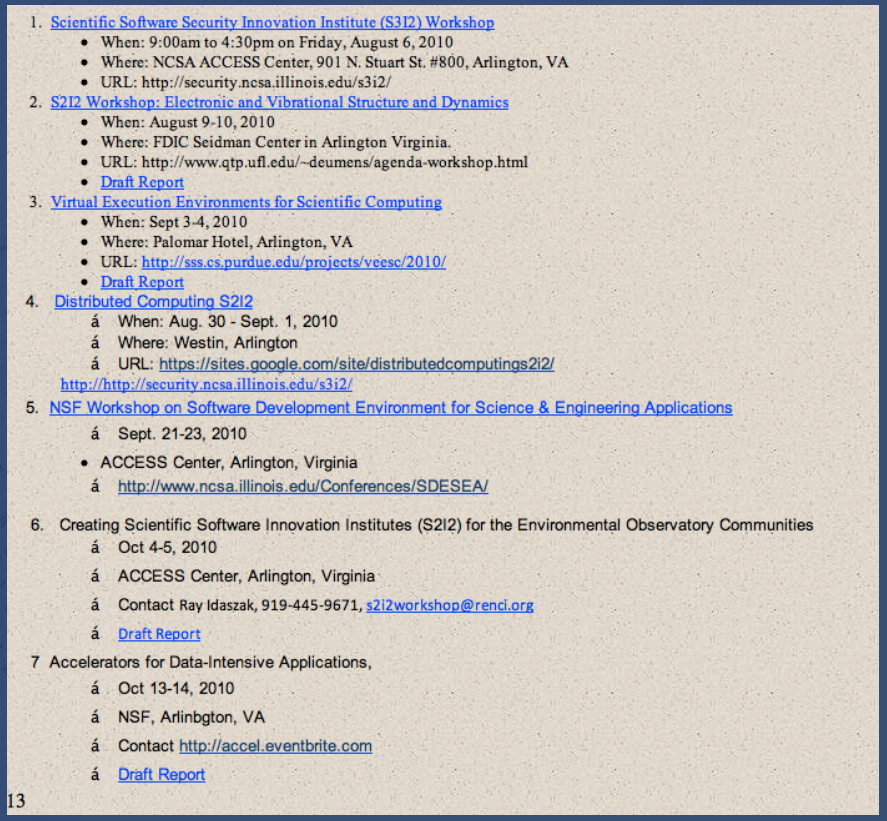


Researchers, scientists, students, educators experience ubiquitous transparent presence from where-ever they are.

NSF Planning for the Future – Software Sustainability

The SI2 program envisions that the Scientific Software Innovation Institutes would provide the expertise, processes, resources and implementation mechanism to transform computational science and engineering innovations and community software into a robust and sustained software infrastructure for enabling science and engineering. They would provide pathways for community involvement, enable software elements developed within the community to be transitioned to conform to community software frameworks standards, and make then accessible, usable and extendible by the community. Furthermore, education, and workforce and career development would be an integral part of an S2I2. S2I2 awards are expected bring together multidisciplinary teams of domain scientists and engineers, computer scientists and software engineers, technologists and educators.

Series of sponsored workshops from the community covering a spectrum of topics: some domain specific some broad.



1. Scientific Software Security Innovation Institute (S3I2) Workshop
   - When: 9:00am to 4:30pm on Friday, August 6, 2010
   - Where: NCSA ACCESS Center, 901 N. Stuart St. #800, Arlington, VA
   - URL: http://security.ncsa.illinois.edu/s3i2/
2. S2I2 Workshop: Electronic and Vibrational Structure and Dynamics
   - When: August 9-10, 2010
   - Where: FDIC Seidman Center in Arlington Virginia.
   - URL: http://www.qtp.ufl.edu/~deumens/agenda-workshop.html
   - Draft Report
3. Virtual Execution Environments for Scientific Computing
   - When: Sept 3-4, 2010
   - Where: Palomar Hotel, Arlington, VA
   - URL: http://sss.cs.purdue.edu/projects/veesc/2010/
   - Draft Report
4. Distributed Computing S2I2
   - When: Aug. 30 - Sept. 1, 2010
   - Where: Westin, Arlington
   - URL: https://sites.google.com/site/distributedcomputings2i2/ http://http://security.ncsa.illinois.edu/s3i2/
5. NSF Workshop on Software Development Environment for Science & Engineering Applications
   - Sept. 21-23, 2010
   - ACCESS Center, Arlington, Virginia
   - http://www.ncsa.illinois.edu/Conferences/SDESEA/
6. Creating Scientific Software Innovation Institutes (S2I2) for the Environmental Observatory Communities
   - Oct 4-5, 2010
   - ACCESS Center, Arlington, Virginia
   - Contact Ray Idaszak, 919-445-9671, s2i2workshop@renci.org
   - Draft Report
7. Accelerators for Data-Intensive Applications,
   - Oct 13-14, 2010
   - NSF, Arlinbgton, VA
   - Contact http://accel.eventbrite.com
   - Draft Report

13

Software Infrastructure Workshop

Report from the Workshops on Distributed Computing, Multidisciplinary Science, and the NSF's Scientific Softwar Innovation Institutes Program

Miron Livny, Ian Foster, Ruth Pordes, Scott Koranda, JP Navarro

October 2010

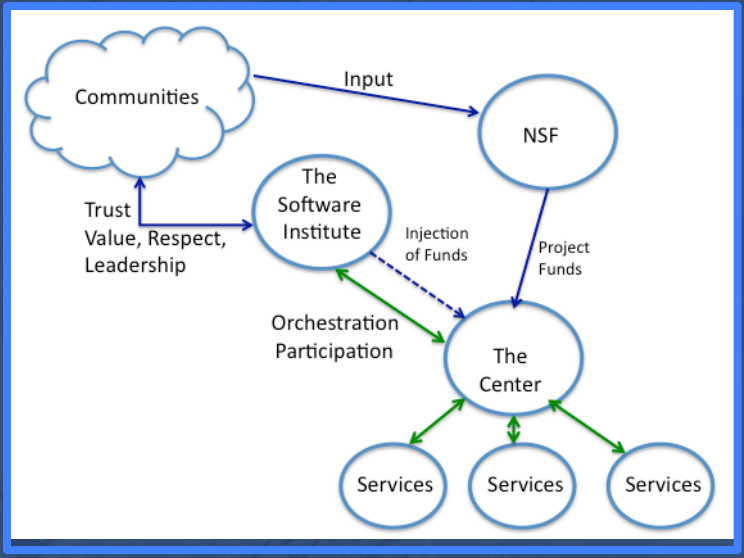
Calls for cross-domain, cross-program, cross agency architecture, validation and auditing, software technologies forum.

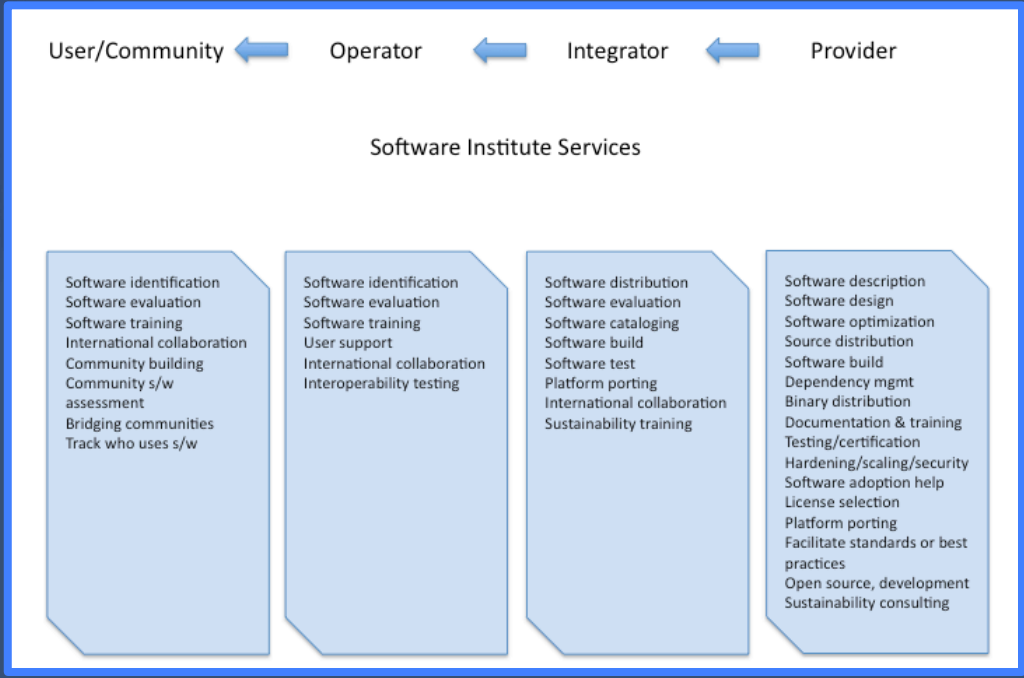
3 Outcome – Recommendation 1

The most significant outcome of the workshops was the vision (and key attributes) for:

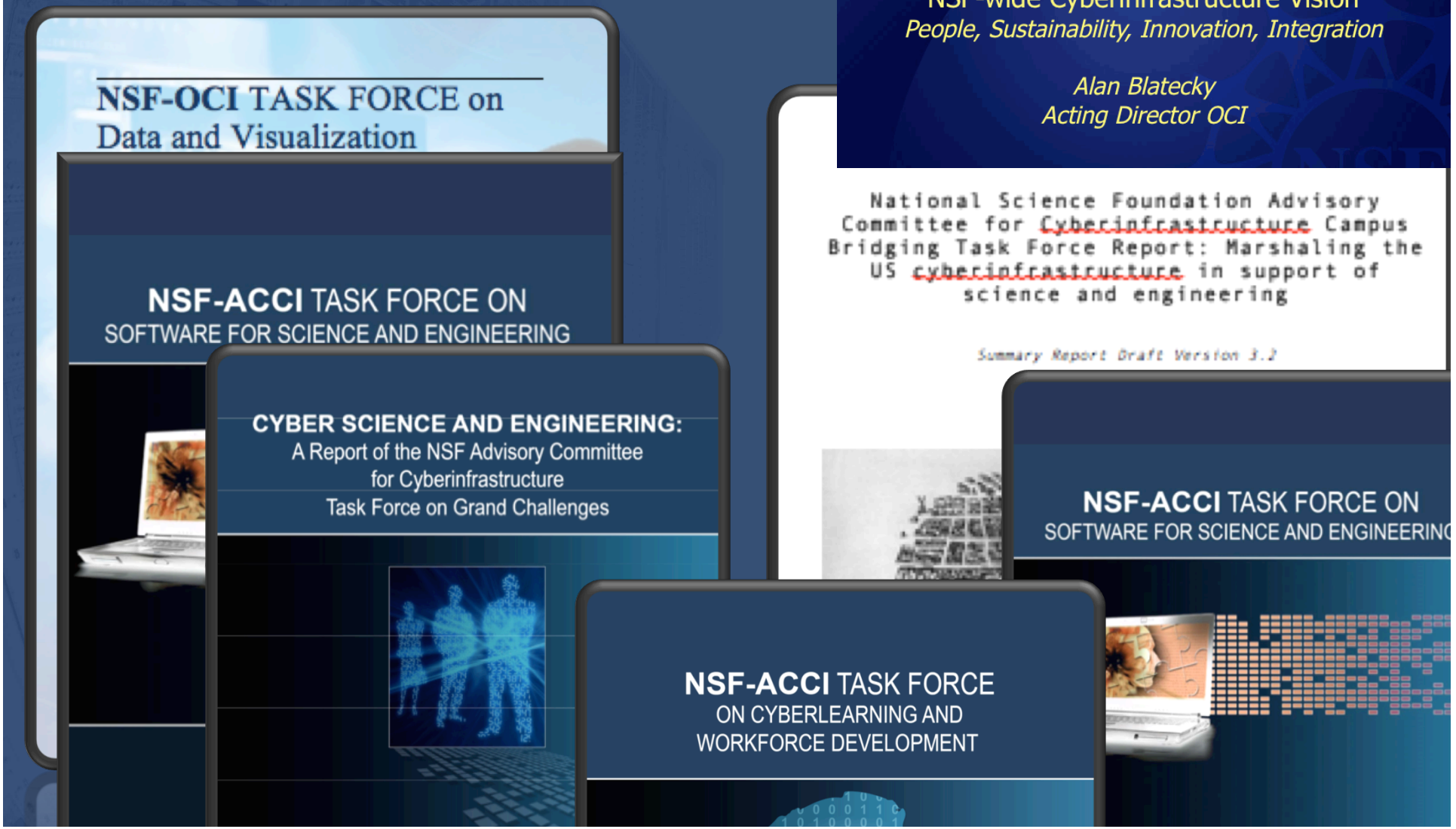
> "***A US Software Infrastructure Institute that provides a national center of excellence for community based software architecture, design and production; expertise and services in support of software life cycle practices; marketing, documentation and networking services; and transformative workforce development activities.***"

The measure of success of such an institute should be the cost effectiveness (as measured in scholarly work) of our software infrastructure. This will be accomplished by a thriving and innovative software infrastructure community anchored by the Software Infrastructure Institute. The Institute will play a unique role by addressing organizational and life cycle elements not covered by infrastructure implementation and deployment projects that are driven by scientific objectives or technological trends.





NSF Planning for the Future: ACCI Task Forces- reports reviewed by NSF April 1 2011



Grand Challenges (Draft)

- An interagency group in CDS&E should be established to address national goals and priorities and to ensure coordination of efforts
- Support of diverse HPC activities (hardware, methods, algorithms) should remain a high priority. University researchers need open access to these resources at all levels
- The development of robust, reliable and useable software at all levels needs to supported by NSF and recognized as an important component of the research portfolio of NSF
- Support CI for data and visualization
- Learn how to create grand challenge communities and VOs (and do it!)

Campus Bridging (Draft)

- Study successful campus CI implementations to document and disseminate the best practices for strategies, governance, financial models and deployment
- Establish a blueprint and roadmap for national CI, including
  - Standard Authentication (InCommon)
  - MRI awards at campus level
  - National Data infrastructure, including national networking backbone
- Campuses should
  - Develop a Cyberinfrastructure master plan with the goal of identifying and planning for the changing research infrastructure needs of faculty and researchers
  - Work toward a goal of providing their educators and researchers access to a seamless Cyberinfrastructure which supports and accelerates research and education

Grand Challenges (Draft)

- An interagency group in CDS&E should be established to address national goals and priorities and to ensure coordination of efforts
- Support of diverse HPC activities (hardware, methods, algorithms) should remain a high priority. University researchers need open access to these resources at all levels
- The development of robust, reliable and useable software at all levels needs to supported by NSF and recognized as an important component of the research portfolio of NSF
- Support CI for data and visualization
- Learn how to create grand challenge communities and VOs (and do it!)

o document and disseminate the best practices for strategies, governance, financial models and deployment

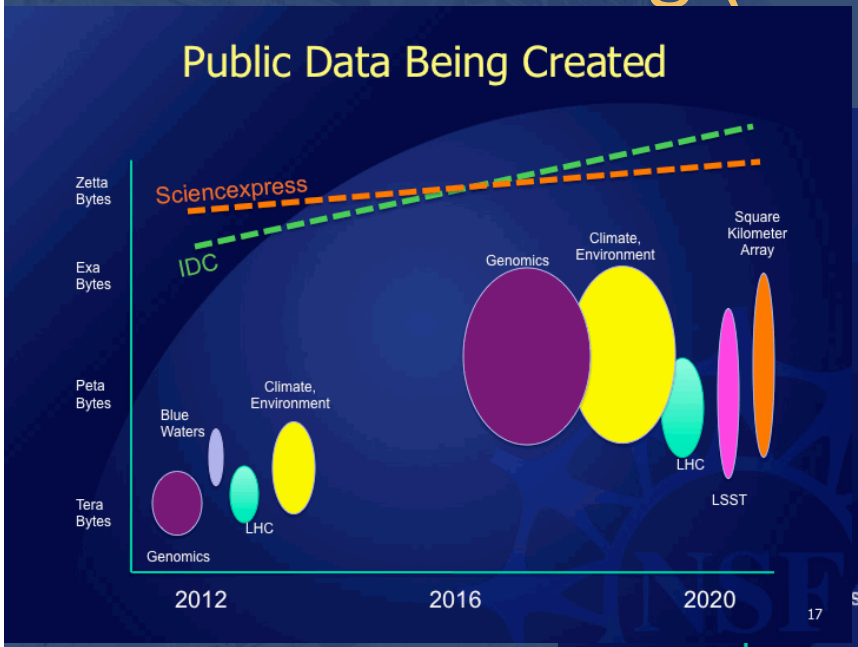
- Establish a blueprint and roadmap for national CI, including
  - Standard Authentication (InCommon)
  - MRI awards at campus level
  - National Data infrastructure, including national networking backbone

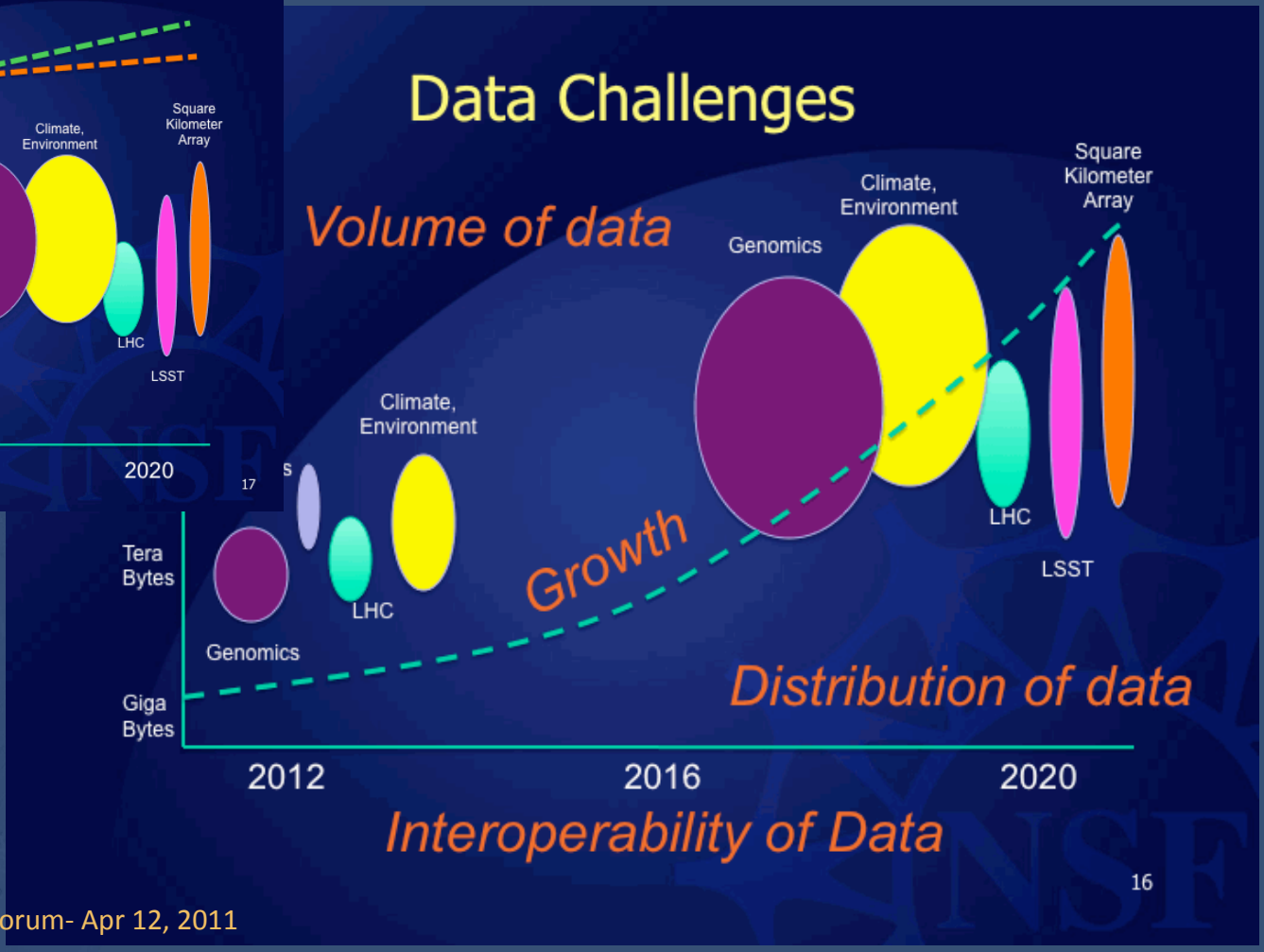
- Campuses should
  - Develop a Cyberinfrastructure master plan with the goal of identifying and planning for the changing research infrastructure needs of faculty and researchers
  - Work toward a goal of providing their educators and researchers access to a seamless Cyberinfrastructure which supports and accelerates research and education

More parochially from Campus Bridging Draft …

Tactical Recommendation to the NSF #1:The NSF should fund the TeraGrid eXtreme Digital program, as currently called for in existing solicitations, and should continue to fund and invest in the Open Science Grid.

NSF Continuing and Increasing focus on Data – archive, access, curation, provenance, sharing (From Alan Blatecky)





Datanet

- Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline.

- Initially 2 projects:
  - Dataone – focus earth observational data from all disciplines, all places, at all scales led by the University of New Mexico.
  - The Data Conservancy, A Digital Research and Curation Virtual Observatory lead bu Johns Hopkins University.

- Digital libraries increasingly contributing: California digital library, North Carolina, Georgia, Purdue, UIUC…

March 21, 2011

The National Science Board's (NSB) Committee on Strategy and Budget (CSB), Task Force on Data Policies will host a two-day expert panel discussion on March 28 and 29, a the National Science Foundation (NSF) headquarters in Arlington, Va.

Experts from across the United States and from the United Kingdom and Germany will participate in this workshop discussion. The full meeting agenda is online.

Members of the media and the public are invited to the meeting. Highlights include:

MONDAY, MARCH 28, 2011

8:20a.m. - 10 a.m. Session I: The Vision of Data-Intensive Science

Guiding questions: What are some of the defining characteristics of data-intensive science? What are the goals for enabling re-use and re-purposing of data? What new opportunities and new types of science have yet to be realized? These questions build upon the vision for a new NSF-wide program in computational and data-intensive science.

10:15 a.m. - 12 p.m. Session II: Reproducibility, First Steps and Guiding Principles

Guiding questions: Reproducibility starts to scope the problem and drives all sorts of related issues (curation, cost, etc.). What does this mean for types of discovery that need data sharing? What are the implications for data publishing and data citation? Would complete data release include the original, "raw" data; cleaned-up, publication-ready data, along with the methods for clean-up; publication-ready data with the meta-data necessary to reproduce any interpretations of the data; raw data with software to make it usable to others; data organized in a way that is interoperable to some standard; etc.?

12:30 p.m. Lunch Presentation: *High Performance Cyberinfrastructure is Needed to Enable Data-Intensive Science and Engineering*

Dr. Larry Smarr, Harry E. Gruber Professor, Department of Computer Science and Engineering, University Of California, San Diego Jacobs School of Engineering; and Director, California Institute for Telecommunications and Information Technology

1 p.m. - 3 p.m. Session III: Exemplars, Lessons Learned