The NSF Cyberinfrastructure Vision: Act Locally and Engage Globally

Daniel E. Atkins
Director,
U.S. National Science Foundation
Office of Cyberinfrastructure
datkins@nsf.gov

http://egee-technical.web.cern.ch/egee-technical/conferences/EGEE06/index.html
What I know about the Oklahoma Supercomputing Context

- Regional leadership and facilitation of this well-established and valued supercomputing symposium series.
- Well known for computational science education and training.
- Culture of collaboration between academia, industry, and government.
- International leadership in modeling/predicting extreme weather events.
Act Locally; Engage Globally

- “engage globally” broadly defined to include geography, inter-disciplinary, inter-institutional, integrated research and education.
- Supercomputing as a major resource (often the centerpiece) of broader distributed knowledge communities supported by CI.
- Working in global reach virtual organizations.
- Supercomputing in the age of virtualization; relaxed constraints of distance and time.
The University of Michigan Upper Atmospheric Research Collaboratory (UARC)

The Initial Facility at Sondrestrom, Greenland
UARC Interface (circa 1988)

Real-time instruments

computational models

dynamic work rooms

team chat

Archival data

Journals
Evolved into a Network of Instruments (one global instrument)
UARC Patterns of Communication

Pattern of Communication, UARC Campaign, April 9, 1997

1998 Smithsonian Science Award
Vignettes from UARC/SPARC

- Shared, tele-instruments & expertise.
- Rapid response, opportunistic campaigns.
- Multi-eyes, complementary expertise.
- Isolated instruments became a global instrument chain.
- Cross-mentoring/training.
- New & earlier opportunities/exposure for grad students.
- Enhanced participation. Legitimate peripheral participation.
- Support for authentic, inquiry-based learning at UG and pre-college level.
- Distributed workshops for post-campaign data analysis.
- Session re-play for delayed participation.
- Data-theory closure.
- A “living specification” to stretch vision of possibilities.
Nomenclature

ACLS-Mellon Study: Cyberinfrastructure for Humanities
http://www.acls.org/cyberinfrastructure/

HASTAC Project
http://www.hastac.org/

e-research

e-science

e-infrastructure

cyberinfrastructure (CI)

Cyber science

research & learning

science

engineering

rapid response

humanities

world universities

CI-enhanced | enabled

e = electronic | enhanced | enabled
Dualities

- **cyberinfrastructure** enables **research & development**
  
  *Cl is both an object and means for R&D*

- **collaboration** enables **CI environments**
  
  *Multi-stakeholder collaboration required to create, provision, and apply CI; CI supports collaborations across time and distance (geographic, disciplinary, institutional)*

- **learning | education** enables **CI environments**
  
  *Learning and workforce development initiatives required to create and use CI; CI enables/enhances learning/education*
Instances of Virtual Organizations (VOs)

Alternate Names for Instances of VOs:
- Co-laboratory
- Collaboratory
- Grid (community)
- Network
- Portal
- Gateway
- Hub
- Virtual Research Environment (VRE)
- Cyberinfrastructure Collaborative
- Other?
“a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology, and pulled by the expanding complexity, scope, and scale of today’s challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive “cyberinfrastructure” on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy.”

http://www.nsf.gov/od/oci/reports/toc.jsp

NSF Blue Ribbon Advisory Panel on Cyberinfrastructure

Daniel E. Atkins, Chair
University of Michigan

Kelvin K. Droegemeier
University of Oklahoma

Stuart I. Feldman
IBM

Hector Garcia-Molina
Stanford University

Michael L. Klein
University of Pennsylvania

David G. Messerschmitt
University of California at Berkeley

Paul Messina
California Institute of Technology

Jeremiah P. Ostriker
Princeton University

Margaret H. Wright
New York University
Vision and Activities Based on Broad and Diverse Community Engagement

Advances in components of CI-systems for S&E R&E

NSF internal working groups

30+ disciplinary workshops on CI vision & impact

CI Council, Directorate/Office CI Activities, OCI, ACCI

NSB & Community Input

Complex, multi-scale, multidisciplinary S&E research challenges

High Performance Computing

Data, Data Analysis & Visualization

Virtual Organizations

Learning & Workforce Development

Vision Framework

• All directorates and offices support cyberinfrastructure.
• Science-driven partnerships between creation, provisioning and use of CI
• Supports integrated research and education and broadened access and participation.
Some Science Drivers

- Inherent **complexity and multi-scale** nature of today's frontier science challenges.

- Requirement for **multi**-disciplinary, **multi**-investigator, **multi**-institutional approach (often international).

- High **data intensity** from simulations, digital instruments, sensor nets, observatories.

- Increased value of data and demand for data **curation & preservation** of access.

- Exploiting infrastructure **sharing** to achieve better stewardship of research funding.

- Strategic need for **engaging more students** in high quality, authentic science and engineering education.
Biosystems Need for Multiscale Computational Modeling

The Complexity of Biosystems

Time Scale (seconds)

10^{-15} 10^{-12} 10^{-9} 10^{-6} 10^{-3} 10^{0} 10^{3} 10^{6} 10^{9}

Number Scale (over size scale from Angstroms to Km)

Atoms Biopolymers Organisms

Scale (over size scale from Angstroms to Km)

Organisms 10^6

Enzyme Mechanisms

Ab initio Quantum Chemistry

First Principles Molecular Dynamics

Molecular Dynamics

Electrostatic continuum models

Finite element models

Organ function

Cell signalling

DNA replication

Protein Folding

Homology-based Protein

Discrete Automata models

Ecosystems and Epidemiology

Evolutionary Processes

Finite element models

Electrostatic continuum models

Ab initio Quantum Chemistry

First Principles Molecular Dynamics

Molecular Dynamics

Empirical force field Molecular Dynamics

The Complexity of Biosystems

From J. Collins, NSF/BIO AD

D. E. Atkins
Achieving the NSF CI (e-science) Vision requires synergy between 3 types of activities.

**Transformative Application** - to enhance discovery & learning

**Provisioning** - Creation, deployment and operation of advanced CI

**R&D** to enhance technical and social effectiveness of future CI environments

**Borromean Ring**: The three rings taken together are inseparable, but remove any one ring and the other two fall apart. See [www.liv.ac.uk/~spmr02/rings/](http://www.liv.ac.uk/~spmr02/rings/)
### NSF CI FY07 Budget Request

Total of $600M in CI Funding with $182M in OCI

#### Cyberinfrastructure Funding

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological Sciences</td>
<td>$77.00</td>
<td>$84.00</td>
<td>$90.50</td>
<td>$6.50</td>
<td>7.7%</td>
</tr>
<tr>
<td>Computer and Information Science and Engineering</td>
<td>45.32</td>
<td>63.00</td>
<td>68.00</td>
<td>5.00</td>
<td>7.9%</td>
</tr>
<tr>
<td>Engineering</td>
<td>52.00</td>
<td>52.00</td>
<td>54.00</td>
<td>2.00</td>
<td>3.8%</td>
</tr>
<tr>
<td>Geosciences</td>
<td>71.35</td>
<td>71.35</td>
<td>75.00</td>
<td>3.65</td>
<td>5.1%</td>
</tr>
<tr>
<td>Mathematical and Physical Sciences</td>
<td>56.52</td>
<td>59.30</td>
<td>63.56</td>
<td>4.26</td>
<td>7.2%</td>
</tr>
<tr>
<td>Social, Behavioral and Economic Sciences</td>
<td>20.39</td>
<td>20.54</td>
<td>20.54</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Office of Cyberinfrastructure</td>
<td>123.28</td>
<td>127.12</td>
<td>182.42</td>
<td>55.30</td>
<td>43.5%</td>
</tr>
<tr>
<td>Office of International Science and Engineering</td>
<td>0.22</td>
<td>1.00</td>
<td>1.05</td>
<td>0.05</td>
<td>5.0%</td>
</tr>
<tr>
<td>Office of Polar Programs</td>
<td>25.38</td>
<td>26.24</td>
<td>26.24</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Subtotal, Research and Related Activities</td>
<td>471.47</td>
<td>504.55</td>
<td>581.31</td>
<td>76.76</td>
<td>15.2%</td>
</tr>
<tr>
<td>Education and Human Resources</td>
<td>20.27</td>
<td>15.02</td>
<td>15.52</td>
<td>0.50</td>
<td>3.3%</td>
</tr>
<tr>
<td>Total, Cyberinfrastructure Funding</td>
<td>$491.74</td>
<td>$519.57</td>
<td>$596.83</td>
<td>$77.26</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

Totals may not add due to rounding.
Several Active Solicitations Posted
Seeking more program officers.


D. E. Atkins
High Performance Computing

increasingly important tool for understanding

Track 1: One solicitation funded over 4 years: $200M acquisition + additional O&M cost.

Track 2: Four solicitations over 4 years: $30M/yr acquisition + additional O&M cost. First track 1 approved 8-07

Satellite tobacco mosaic virus, P. Freddolino et al.

Aldehyde dehydrogenase, T. Wymore and S. Brown

I. Shipsey

John Q Public

S.-Y. Kim, M. Lodge, C. Taber.
Press Release 06-137
National Science Foundation Awards Texas Advanced Computing Center $59 Million for High-Performance Computing

University and industry consortium to deploy powerful general-purpose computing system

Scientists will use the TACC computer to simulate the 10 million atoms in this bacterial organelle.

Credit and Larger Version

- Challenges: **increased scale, heterogeneity, and re-use value** of digital scientific information and data. Inadequate digital preservation strategy of long-lived data.
- Taking initial steps to **catalyze the development** of a federated, global system of science and engineering data collections that is open, extensible, evolvable, (and appropriately curated and long-lived.)
- Complemented by a **new generation of tools** and services to facilitate data mining, integration, analysis, visualization essential to transforming data into knowledge.
- NSF Leadership for OSTP/Interagency Working Group on Digital Data
Distributed virtual organizations are **based upon CI** that provides flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources.

**Resources and services** include HPC, data/information management, sensor-nets/observatories, linked through global networking and middleware, and accessed by people through web portals and workflow environments.

Increasing numbers of **virtual organizations are required** by S&E research and education communities. Referred to by many names, e.g. *collaboratory, co-laboratory, grid, gateway, portal, hub, ....*

**Challenges** being address include tools for more rapid building and ease of use, interoperability/middleware, high performance, end-to-end networking, and dynamic reconfiguration, social issues, assessment of impact, and economic and technical sustainability.
Virtual Organizations offer additional modes of interaction between People, Information, and Facilities.

**Time**

- **Same** (synchronous)
  - **ST-SP**: Physical mtgs, Print-on-paper books, journals, Physical labs, studios, shops
  - **ST-DP**: AV conference, Web search, Online instruments

- **Different** (asynchronous)
  - **DT-SP**: Shared notebook, Library reserves, Time-shared physical labs, ...
  - **DT-DP**: Email, Knowbots, Autonomous observatories

**Geographic Place**

- **Same**
  - **ST-SP**
  - **ST-DP**

- **Different**
  - **DT-SP**
  - **DT-DP**

**Symbols**

- **P**: people
- **I**: information
- **F**: facilities, instruments

**Notes**

- Time: Same (synchronous), Different (asynchronous)
- Geographic Place: Same, Different
Virtual Organizations offer additional modes of interaction between People, Information, and Facilities.

- **ST-SP**: Physical mtgs, Print-on-paper books, journals, Physical labs, studios, shops.
- **DT-SP**: Shared notebook, Library reserves, Time-shared physical labs, ...
- **ST-DP**: AV conference, Web search, Online instruments.
- **DT-DP**: Email, Knowbots, Autonomous observatories.

**P**: people, **I**: information, **F**: facilities, instruments.
The number of nation-scale projects is growing rapidly!
Need a socio-technical approach: Realizing the potential of e- | cyber science to support effective VOs requires attention to institutional infrastructure

- Need more than good ICT systems and tools for individuals and organizations. (Technology determinism alone is not enough).

- No less important is the institutional contexts (i.e. norms of practice and rules) to facilitate collaboration within science and technical research communities.

- The institutional and organizational environment of e-science encompasses a wide and diverse array of interrelated social, economic, and legal factors that
  - create incentives for, and constraints upon individual and collective action; and
  - thereby shape the production, utilization, consumption, and governance of e-science capabilities and products.

Adapted from Paul David, see www.oii.ox.ac.uk/resources/publications/RR2.pdf
VO-substrate: International R&E Networking
TeraGrid is a facility that integrates computational, information, and analysis resources at the San Diego Supercomputer Center, the Texas Advanced Computing Center, the University of Chicago / Argonne National Laboratory, the National Center for Supercomputing Applications, Purdue University, Indiana University, Oak Ridge National Laboratory, the Pittsburgh Supercomputing Center, and the National Center for Atmospheric Research.
TeraGrid PI’s By Institution as of May 2006

Blue: 10 or more PI’s
Red: 5-9 PI’s
Yellow: 2-4 PI’s
Green: 1 PI

TeraGrid PI’s

Charlie Catlett (cec@uchicago.edu)
TeraGrid Science Gateways Initiative: Community Interface to Grids

- Common Web Portal or application interfaces (database access, computation, workflow, etc).
- “Back-End” use of TeraGrid computation, information management, visualization, or other services.
- Standard approaches so that science gateways may readily access resources in any cooperating Grid without technical modification.
Let’s look at a few real example Grid Science Gateways

(about a dozen ... many more exist!)

These example slides courtesy of D. Gannon
NEESGrid

Realtime access to earthquake Shake table experiments at remote sites.
The Biomedical Informatics Research Network (BIRN) Portal provides BIRN members with a single sign on web portal to access data grid files, computation grid resources, and a variety of collaboration tools to facilitate the scientific needs of BIRN researchers. Non-BIRN participants may access the portal through a guest registration.
Geological Information Grid Portal

GEONgrid Portal

Select a Subject to Show Resources

- Biological oceanography
- Chemical oceanography
- Cryology
- Ecology
- Education
- Environmental science
- Forestry
- Geochemistry
- Geologic time
- Geology
- Geophysics
- Human geography
- Hydrology
- Mineralogy or petrology
- Natural hazards
- Other
- Palaeontology
- Physical geography
- Physical oceanography
- Soil science
- Structural geology
- Technology

Resources in Geology

- Title: Arizona Geology Map
- Format: shapefile
- Dataset id: GEON-25db3db-e71c-11d9-b225-ab22ed7501c0
- Spatial Coverage: North: 37 East: -109.04 South: 31.33 West: -114.82
- Temporal Coverage: any
- Description: This is a geology map of Arizona in USA.
- Semantic Annotations: see details
Mesoscale Meteorology

NSF LEAD project - making the tools that are needed to make accurate predictions of tornados and hurricanes.
- Data exploration and Grid workflow
The LEAD Vision: Adaptive Cyberinfrastructure

DYNAMIC OBSERVATIONS

Analysis/Assimilation
- Quality Control
- Retrieval of Unobserved Quantities
- Creation of Gridded Fields

Prediction/Detection
- PCs to Teraflop Systems

Product Generation, Display, Dissemination

Models and Algorithms Driving Sensors

The CS challenge: Build cyberinfrastructure services that provide adaptability, scalability, availability, usability, and real-time response.

From D. Gannon
Renci Bio Portal

Providing access to biotechnology tools running on a back-end Grid.

- leverage state-wide investment in bioinformatics
- undergraduate & graduate education, faculty research
- another portal soon: national evolutionary synthesis center
Nanohub - nanotechnology
Welcome to the Crystallography Portal

The Purdue Chemistry Crystallography Center

Disable your browser's cache to get the live stream!

These values are updated approx. every 60 sec.

<table>
<thead>
<tr>
<th>Instrument Enclosure Temp. &amp; Humidity:</th>
<th>23.4 C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rel. Humid. 43.1 %</td>
</tr>
<tr>
<td>Chill Water In:</td>
<td>16.4 C</td>
</tr>
<tr>
<td>Chill Water Out:</td>
<td>19.3 C</td>
</tr>
<tr>
<td>Generator Relay Voltage:</td>
<td>3.42</td>
</tr>
<tr>
<td>All previous voltages</td>
<td></td>
</tr>
<tr>
<td>X-ray Generator is:</td>
<td>OFF</td>
</tr>
</tbody>
</table>

Times in UTC

2005-09-24 16:35:59
2005-09-24 16:36:25
2005-09-24 16:36:48

Local date/time: 2005-09-24 11:36:54

Streaming video from the crystal microscope on the Nonius diffractometer

Streaming video from the Nonius Kappa CCD detector (Under development!!)

Total Number of .jpg: 10
Frame: s0100010.jpg
All available .jpg images
Browse the 20 latest .jpg images

Data from Nonius Kappa CCD detector

LabJack U12
Welcome to the ServoGrid Portal

SERVO Grid
Solid Earth Research Virtual Observatory Grid

- QuakeSim home page
- Old GEM, General Earthquake Modeling Web Site
- SLIDE Distributed File System for NASA Computational Technology Project
- Report from the Earth Science Enterprise Computational Technology Requirements Workshop April 30-May 1 2002 where SERVO concept first introduced
- Discover the Grid at the Grid Forum or at this collection of papers
- Other collected papers and presentations on SERVOgrid and related topics are available from the Community Grids Lab publications page.

QuakeSim Web Portal  User Manual  Support  Report Bugs  QuakeSim Web Site

Participating Institutions:
IU CGL | NASA JPL | UC Davis | UC Irvine | USC
Welcome to the GeneGrid Prototype - Release 0.6

This is the GeneGrid Test Bed release 0.6 managed by the Belfast e-Science Centre, utilising resources in BeSC, Queen's University of Belfast, Melbourne University, BT and the San Diego Super Computing Centre.

Users are limited to selected staff of both commercial partners - Fusion Antibodies, Amtec Medical - and the Belfast e-Science Centre. To obtain a user account, please contact the appropriate representative - P.V. Jithesh (BeSC), Mark McCurley (Fusion) or Dr. Shane McKee (Amtec). Authorized users will be provided with a username and password by BeSC.

All users are requested to subscribe to the GeneGrid mailing list and to use it for directing queries etc. Mail GeneGrid, and place the word “subscribe” (without the quotes) in the message body.

For more on the GeneGrid project, please click here.

Important Note: Current GeneGrid Users please continue to use the Release 0.5 available here.
myGrid is a collection of services and components that allows the high level integration of biological applications. The architecture provides the infrastructure necessary, in a web service environment, e-science workbench that actively supports the scientific lifecycle. Each component or service contributes to a system that allows the e-scientist to perform complex in-silico experiments across distributed bioinformatics resources.
Learning supported by CI. (cyber-enabled learning).

Workforce development to create and use CI for S&E research and education.

Broadened participation: Exploit the new opportunities that cyberinfrastructure brings for ... people who, because of physical capabilities, location, or history, have been excluded from the frontiers of scientific and engineering research and education.

Explore CI support for integrated research and education.
To be a genuinely competitive knowledge economy, Europe must be better

- in producing knowledge through research
- in diffusing it through education
- in applying it through innovation
e-Infrastructures in FP7 - strategy - Virtual Organizations (VO)

Bringing the best brains together
Sharing the best scientific resources

Weather Forecast VO
Biomedics VO
Astrophysics VO

Scientific Data
Sharing Scientific Resources
Communication Network

Producing the best science
Collaborating Infrastructures

Potential for linking ~80 countries by 2008
Some Existing & Potential Interactions

International Science Projects, e.g. ATLAS, CMS

- EGEE
  - gLite
  - Experience with large, production, international Grid operation
  - Other?

Funding & science collaboration

staff R&D interactions
use of components
shared development

U.S. Investments

- U.S. part of international science/engineering research projects.
- Open Science Grid (OSG)
- TeraGrid & Science Gateways
- Grid Interoperabilty Now (GIN)
- GLOBUS
- Condor Technologies
- Virtual Data Toolkit (VDT)
- NMI Build and Test
- Shibolleth
- GridShib
- Other?

Other National/Regional Grid Projects

International Science Projects, e.g. ATLAS, CMS
New Opportunities

“Cyberinfrastructure-enhanced knowledge communities offer the potential for enabling a new wave of global-scale collaboration across multiple disciplines, geography, and institutions. It could empower a revolution in what science explores, how it is done, and who participates.

Realizing this potential will, however, also required a new wave of commitment to collaboration between the complex array of stakeholders necessary to create, deploy, sustain, and apply cyberinfrastructure in transformative ways.

Cyberinfrastructure both enables and requires a new wave of collaboration.”
Questions & Discussion