Cyberinfrastructure

David Minor
UC San Diego Libraries
San Diego Supercomputer Center

January 4, 2012

Cyberinfrastructure:

• History

• Definitions

• Examples
History

mid-1990s:
  – High performance computing becoming more focused on distributed resources
  – Internet boom

Foundational questions:
  – How do we bring together distributed resources?
  – What impact will it have on science?
National Science Foundation

2003: “Revolutionizing Science and Engineering Through Cyberinfrastructure”

• Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure

• The “Atkins Report”

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

The Panel’s overarching finding is that 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.
Recommendation

NSF should establish and lead a large-scale cyberinfrastructure program that is:

– Interagency
– Internationally coordinated
– Creates, deploys, and applies cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education

Office of Cyberinfrastructure (OCI)

• Funds the creation of tools and services
• Coordinates research programs
• Funds national supercomputing efforts
• Guides national discourse

Questions?

Definitions
Definition One – The Stuff

Cyberinfrastructure is comprised of the hardware and staffing needed to create and support environments for data acquisition and management, as well as the necessary tools, computing and information processing services.

Definition Two – The Plan

For scientists and researchers, cyberinfrastructure includes the technology and programs needed to efficiently connect laboratories, data, computers, and people. This is done to create new forms of science.
Computers, storage and networks

Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Computers, storage and networks
Computers, storage and networks

Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Enabling hardware, algorithms, software, communications, institutions, and personnel.
... uh, and E-Science

In most of Europe and much of the rest of the world, what we are talking about is called “E-Science” or “eScience.”

– Operationally and philosophically the same
– The name and concept driven by the United Kingdom in the previous decade

Questions?

ARL DLF
Examples

TeraGrid / XSEDE

• TeraGrid 2004 – 2011
  – 2.5 petaflops of computing capability
  – More than 50 petabytes of online and archival data storage
  – More than 100 domain-specific databases

• XSEDE
  – 2011-2016
  https://www.xsede.org
nanoHUB

- Computational nanotechnology research
  - Simulation programs
  - Science gateways, networks and applications
  - Teaching and learning materials
    http://nanohub.org

DataNet program

- “Sustainable Digital Data Preservation and Access Network Partners”
  - 2008-present
  - ~$45 million dollars
  - Several large programs with many partners
    http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141
Network for Earthquake Engineering Simulation (NEES)

http://nees.org

OpenTopography

http://www.opentopography.org
HathiTrust

http://www.hathitrust.org

Penn State University

• Digital Library Technologies, a division of Information Technology Services
  – Defining service levels for archival storage and repository services for the university

• Digital Library Technologies and University Libraries
  – Publishing and curation services program in support of digital stewardship at the institution
  – Working with other organizations around the country to create flexible, shareable tools and services

http://www.dlt.its.psu.edu/
Cornell - DataStaR

• Investigating methods for collaborating and sharing data as well as promoting high-quality metadata creation.

http://datastar.mannlib.cornell.edu/

Johns Hopkins University

Data Management Service (DMS)
JHU Data Management Service (DMS) represents the culmination of two years of research, design, development and implementation of Data Conservancy.

- Service launched in July 2011
- DC instance launched in October 2011
- Important, essential foundations in place
- There remains work to be done

JHU Data Management Services

Two sets of services provided:

- Consultative data management planning services to support NSF proposals
  - *direct funded by schools with PIs seeking NSF funding*
- Post award data management planning and deposit in the JHU Data Archive (DC instance)
  - *funded through charge back on grant*
  - *Defined time limits, volume of data deposited per project, unencumbered data only for now*
Purdue University

Purdue University Research Repository (PURR)

• Collaboration of Purdue Libraries, Office of the Vice President for Research, and Information Technology at Purdue (ITaP)
• Currently hosting 44 projects and 127 registered users (6 grant-funded projects)
• 4 campus workshops to raise awareness
• 34% of NSF proposals submitted from Purdue since January 17 include PURR as a component of its data management plan
• Still under development; budget and development plan submitted in November
• For more information, visit PURR: http://research.hub.purdue.edu
Subject librarians and data service specialists advise and collaborate with investigators to write effective data management plans. Based on the department affiliation of the project creator, the appropriate subject librarian is associated with the project upon its creation. A project receives a default allocation of private file storage (500M for 3 years) and a set of collaborative tools (e.g., wiki, calendar, blog, messaging).

The project creator can register a grant award to receive a larger allocation of storage (100G for life of project). Data staged in the project space can be described and previewed before being submitted for publication and/or archiving. Submissions are reviewed and approved by the subject librarian and data service specialist. After this point, datasets are assigned DOIs and enter curation. Curated datasets remain published and/or archived for 3 years by default or 10 years for grant-supported projects. When this initial commitment expires, datasets are remanded to the library where they are managed for the long-term as a part of the library’s collections (including selection and appraisal and deselection).

Indiana University
IU Libraries and Cyberinfrastructure
Factors driving consolidation of cyberinfrastructure

- Economies of scale in management
  - Air conditioning
  - Electricity supply
  - Staff
- Security
- VM hosting – intelligent infrastructure
- Economies of vendor partnerships and large purchases
- Redundant backups (2 locations) and our own network
- Some places are best not consolidated (Astro, some of the clusters in Chemistry)

IU Bloomington Data Center
http://dcops.iu.edu/

Campus Scholarly Infrastructure

Domain Specific Discovery & Innovation, Teaching & Learning

Shared Cyberinfrastructure

Shared Cyberinfrastructure

Necessary Infrastructure Leveraged

Genomics
  - Innovation
  - Visualization
  - Models
  - Metadata
  - Curation
  - Computation, Storage
  - Networks

Anthropology
  - Innovation
  - Publication
  - Visualization
  - Searching, Retrieving
  - Curation
  - Storage
  - Networks

Arts
  - Innovation
  - Visualization
  - Computation
  - Models
  - Metadata
  - Curation
  - Retrieval & Analysis
  - Storage
  - Networks

Physics
  - Innovation
  - Visualization
  - Computation
  - Models
  - Metadata
  - Curation
  - Distributed Storage
  - Networks
  - Primary Storage

Discipline Research Stacks...

© Brad Wheeler, Indiana University, Creative Commons Attribution License 2.5
IU Libraries and Cyberinfrastructure

- IU Libraries are involved in a variety of collaborative projects that utilize shared cyberinfrastructure services with a variety of partners on the IU campus. These include areas of OVPIT (Research Technologies, Learning Technologies and Enterprise Software) and areas within the Pervasive Technology Institute Research Centers, the Office of the Vice President for Research, and the College of Arts and Sciences.
  - Shared storage infrastructure for digital collections, institutional repository, publishing, data curation, and digital humanities projects.
  - Shared computational infrastructure for use with the HathiTrust Research Center, data curation, and other data transport and transformation needs.
  - Shared large-area file system for mass-data transfers and redundancy –
    - Shared data MOU with TACC
  - Shared virtual infrastructure within the IU Cloud – used for most library technologies, digital libraries and other library application services at IU.
- These resources when utilized for enterprise-wide solutions are part of the IU Empowering People Plan of abundant unmetered resources.

UC San Diego

Research Cyberinfrastructure (RCI)
RCI elements

- High-Performance Computing
- Data Center Colocation
- Storage
- Data curation
- Networking and other services

High-performance computing

- Triton Resource: a cost–effective and accessible high-performance computing system primarily for UC San Diego and UC researchers

- The Triton Affiliates and Partners Program (TAPP): high performance cluster computing time.
  http://www.sdsc.edu/us/tapp
Data center colocation

- Standard rack provided with ISO-Base seismic protection, aisle containment, and 2x30A power distribution

- 10+ Gb networking fabric connectivity both throughout SDSC aggregation fabric and into CENIC

- 24/7 operations staff providing facility oversight and emergency "remote hands" hardware assistance

http://rci.ucsd.edu/services/colocation.html

---

Storage

<table>
<thead>
<tr>
<th>Storage Type</th>
<th>Cost per Terabyte-Year</th>
<th>Availability</th>
<th>Application Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel File System</td>
<td>Free while running on an SDSC HPC machine. Medium-term parking space available by special arrangement with Project Storage purchased in an equal quantity.</td>
<td>99.5%</td>
<td>Up to 100 GB/s</td>
</tr>
<tr>
<td>Project Storage</td>
<td>• Standard Availability, Single-Site Durability - $600&lt;br&gt;• High Availability, Multiple-Site Durability - $900</td>
<td>• 99.5% &lt;br&gt;• 99.95%</td>
<td>• Up to 1 GB/s &lt;br&gt;• Up to 1 GB/s</td>
</tr>
<tr>
<td>Cloud Storage</td>
<td>• Single-Site Durability - $390&lt;br&gt;• Triple Copy - $650</td>
<td>• 99.5% &lt;br&gt;• 99.5%</td>
<td>• Up to 100 MB/s &lt;br&gt;• Up to 100 MB/s</td>
</tr>
</tbody>
</table>
Networking and other services

• Web & Database Hosting
• Oracle Database Hosting
• 10GigE research network throughout campus

http://rci.ucsd.edu/services/other-services.html

Data curation

• Starting with in two year pilot phase

• Using existing tools whenever possible
  – Storage at SDSC
  – Digital Asset Management System at UCSD Libraries
  – Campus high-speed networking
  – Chronopolis digital preservation network

http://rci.ucsd.edu/services/data-curation.html
Data management plans

• Resources and contacts available to UCSD researchers
• Examples from submitted proposals
• Guidance, tips and recommendations for DMP preparation

http://rci.ucsd.edu/dmp/index.html

Cyberinfrastructure is …
Computers, storage and networks

Software programs, services, instruments, data, information, knowledge, and social practices applicable to specific projects, disciplines, and communities of practice.

Enabling hardware, algorithms, software, communications, institutions, and personnel.

National

Community

Institutional
Questions!

Cliff Lynch

Director of the Coalition for Networked Information