Survey of Campus Research Storage Needs, etc.

Sustainability Workshop

October 2, 2013

Richard Moore
Elements of UCSD’s Integrated Research CyberInfrastructure (RCI) Program

- Data Center Colocation
- Networking
- Research Computing
- Centralized Storage
- Data Curation
- Technical Expertise
Campus Survey of Researchers’ Data Requirements

• Conducted survey of a broad sample of ~50 representative PIs to understand technical and cost requirements

• An additional motive was to increase awareness of the RCI program

• Identify common needs, and define sustainable RCI business model with strong adoption

• Develop centralized, production storage services
## PI Interview Responses: Where is Your Data Coming From?

### Table 1. Data Sources and Relative Distribution

<table>
<thead>
<tr>
<th>Data Source</th>
<th>%</th>
<th>Representative Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencers</td>
<td>28</td>
<td>Biology</td>
</tr>
<tr>
<td>Software applications</td>
<td>28</td>
<td>Biology, Physics</td>
</tr>
<tr>
<td>Field sensors/instruments</td>
<td>20</td>
<td>Marine Biology, etc.</td>
</tr>
<tr>
<td>Audio visual equipments</td>
<td>10</td>
<td>Arts</td>
</tr>
<tr>
<td>Mass spectrometers</td>
<td>8</td>
<td>Biology</td>
</tr>
<tr>
<td>Tomographic instruments</td>
<td>8</td>
<td>Biology, medicine</td>
</tr>
<tr>
<td>External data repositories</td>
<td>8</td>
<td>Biology</td>
</tr>
<tr>
<td>LHC particle dectors</td>
<td>3</td>
<td>Physics</td>
</tr>
<tr>
<td>Archeological studies</td>
<td>3</td>
<td>Humanities</td>
</tr>
<tr>
<td>Curation</td>
<td>3</td>
<td>Sociology</td>
</tr>
</tbody>
</table>

- Indicates use cases for storage and connectivity requirements
- Data sources:
  - ~50% campus instruments
  - ~30% simulations (XSEDE, campus, lab systems)
  - ~20% field instruments
  - ~15% other external sources
- %’s reflect PIs, not data volume

Numbers reflect percentages of PIs surveyed that utilize each solution; Individual PIs use multiple solutions, so %’s add up to >100%.
How do You Handle Data Storage/Backup?

Table 2. Data Storage Devices and Services Utilized

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
<th>Primary purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network attached storage (NAS) devices</td>
<td>73</td>
<td>Standard performance network filesystem</td>
</tr>
<tr>
<td>USB Drives</td>
<td>70</td>
<td>Storage and backup</td>
</tr>
<tr>
<td>Local server hard disk drives</td>
<td>65</td>
<td>Storage and backup</td>
</tr>
<tr>
<td>Dropbox</td>
<td>33</td>
<td>Data sharing</td>
</tr>
<tr>
<td>SDSC Project Storage</td>
<td>13</td>
<td>Standard performance network filesystem</td>
</tr>
<tr>
<td>XSEDE Lustre Filesystem</td>
<td>10</td>
<td>Parallel filesystem</td>
</tr>
<tr>
<td>Google Drive</td>
<td>10</td>
<td>Storage and sharing</td>
</tr>
<tr>
<td>Amazon S3</td>
<td>8</td>
<td>Storage and sharing</td>
</tr>
<tr>
<td>SDSC Cloud Storage</td>
<td>8</td>
<td>Storage and sharing</td>
</tr>
<tr>
<td>Tape library</td>
<td>5</td>
<td>Storage and backup</td>
</tr>
<tr>
<td>Small Area Network Storage Array</td>
<td>3</td>
<td>Databases</td>
</tr>
<tr>
<td>CD/DVD</td>
<td>3</td>
<td>Storage and backup</td>
</tr>
<tr>
<td>Hadoop Filesystem</td>
<td>3</td>
<td>Replication and Map Reduce</td>
</tr>
<tr>
<td>iRODS</td>
<td>3</td>
<td>Metadata driven storage and sharing</td>
</tr>
</tbody>
</table>

• **Storage Devices**
  • Network accessible storage (NAS), USB and server local drives dominate
  • Use of Dropbox for sharing
  • Others use Google Drive, Hadoop, XSEDE, SDSC co-location

• **Backup modes**
  • Replicated copies in two NAS
  • A copy in the NAS,
  • A copy in local hard drive (laptop/workstation),
  • And a copy in a USB drive
  • Maybe a copy in email/Dropbox

• **Problems:**
  • Out of sync
  • Lost track of its location
  • Lost version control
  • High cost of recovery

Numbers reflect percentages of PIs surveyed that utilize each solution; Individual PIs use multiple solutions, so %’s add up to >100%.
For PIs interviewed, current needs 1-1000TB
Increasing in future
Perceptions of permanent storage interesting – none for some, intermediate for many, large for a few
Metadata and retention requirements

Do you need metadata annotation capability?

Metadata Annotation Needs

- Yes, 23%
- Not sure, 5%
- Not yet, 33%
- No, 40%

How long do you need to retain your data?

Research Data Life Time

- 1 year: 3%
- 2 years: 3%
- 3 years: 3%
- 5 years: 20%
- Permanently: 30%
- Duration of Project: 63%
- Other: 3%
Table 4. Top 10 requirements for campus cyberinfrastructure

<table>
<thead>
<tr>
<th>Type</th>
<th>%</th>
<th>Comments</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better CI with minimal direct cost</td>
<td>91</td>
<td>Least burden on research budget</td>
<td>Cost</td>
</tr>
<tr>
<td>Network Attached Storage</td>
<td>73</td>
<td>Shared POSIX compliant filesystem</td>
<td>Sharing</td>
</tr>
<tr>
<td>Data replication as backup</td>
<td>66</td>
<td>Keep a second copy somewhere safe</td>
<td>Recovery</td>
</tr>
<tr>
<td>Dropbox- or Google Drive-like service</td>
<td>43</td>
<td>Ease of access and worry free backup</td>
<td>Ease of use</td>
</tr>
<tr>
<td>10G network connection</td>
<td>38</td>
<td>High speed network bandwidth</td>
<td>Network bandwidth</td>
</tr>
<tr>
<td>Minimal cost beyond hardware cost</td>
<td>24</td>
<td>Little operating cost</td>
<td>Cost</td>
</tr>
<tr>
<td>Shared technical expertise</td>
<td>20</td>
<td>Infrastructure, software and application consulting</td>
<td>Expertise</td>
</tr>
<tr>
<td>Distributed multisite replication</td>
<td>18</td>
<td>Geographical safety</td>
<td>Recovery</td>
</tr>
<tr>
<td>Desktop backup</td>
<td>18</td>
<td>Routine research data safety</td>
<td>Backup</td>
</tr>
<tr>
<td>Compliant and secure storage for sensitive data</td>
<td>16</td>
<td>Personal and clinical data safety</td>
<td>Security</td>
</tr>
<tr>
<td>Tiered storage plans</td>
<td>16</td>
<td>Data retention and automatic removal</td>
<td>Cost</td>
</tr>
</tbody>
</table>

**Top Requirements for Campus Cyberinfrastructure**

- Cost effectiveness tops list
- Ease of use follows
- “Cost is King, Ease of Use Follows”
- Reliable, NFS/CIFS storage most common platform
- Many responses relate to data durability – backups/copies/tiered storage
- High-speed networking enhances quality of service
- “Compliant” environment (storage/computing)
- Tiered storage options is desirable
Research Computing (in production now)

- RCI is evolving SDSC’s Triton system to the “Triton Shared Computing Cluster” (TSCC)
- Condo model: Researchers purchase compute nodes which are operated as part of shared cluster for 3-4 years
  - PI buys hardware & modest ops fee
  - Lower ops cost than local PI cluster; larger-scale resource; professionally-managed
- Hotel: Purchase time by the core-hour; shared queue
Data Curation – in pilot (production FY13-14)

• Completing a two-year pilot phase
  • How do lab personnel work with librarians to curate their data?
  • How much work is required to curate data and what are options?
  • What is a sustainable business model for curation within RCI project?

• Five representative programs across UCSD selected as pilots
  • The Brain Observatory (Annese)
  • Open Topography (Baru)
  • Levantine Archaeology Laboratory (Levy)
  • SIO Geological Collections (Norris)
  • Laboratory for Computational Astrophysics (Wagner)

• Using existing tools whenever possible
  • Storage at SDSC, campus high-speed networking, Digital Asset Management System (DAMS) at UCSD Libraries, Chronopolis digital preservation network

• Also, develop Data Management Plan tools and provide training

• Anticipate production curation services in FY13-14
Some Comments and Lessons Learned

• Campus multi-year budget commitments make a difference to adoption – obvious but …
• In-person interactions very important to adoption
• Wish we’d hired an expert in conducting survey
• Comment yesterday re campus requiring that PIs put skin in the game – not only $, but litmus test
  • However, makes it hard to plan and prepare for 3-5 years out
• ‘Economies of scale’ leverage varies for different services (e.g. colocation -> data curation)
• UC systemwide pilot project (may also apply to some regional collaborations) - getting one person to say yes is a lot easier than N people