RESEARCH CYBERINFRASTRUCTURE VISION WORKGROUP

December 4, 2015
1-2:30 pm, Science Library 6110a

Research Computing at UC Berkeley

David Greenbaum
Director, Research Information Technologies
Information Services & Technology

Developing RCI Vision Document

- Review tentative components
  - Introduction
  - Perspective on criticality of RCI
  - RCI requirements: immediate, and longer term
    - Staff services
    - Network connectivity (pervasive and advanced)
    - Storage (pervasive and advanced)
    - Data management/sharing/curation
    - Computation (pervasive and advanced)
    - Working environment (software/database licenses, tools, etc.)
    - Cyber-facilities and access to external resources
  - RCI requirement perspective/priorities across disciplines
  - Education and training
  - Funding models/considerations
  - Organizational considerations
  - Draft budget – Phase 1 (immediate) and Phase 2 (longer term)
- Request for workgroup members to provide RCI vision input from the perspective of their discipline and/or school

Discuss Initial Component Drafts

- Storage (Harry Mangalam)
- Curation (Laura Smart)

RCI Symposium Update (Allen Schiano)

- Identifying panelists and other participants

UC RCI Vision Draft

- Next Generation Research and the University of California: Planning for the Future of UC's Cyberinfrastructure; A report on the UC VCR-CIO 2015 Summit

---

Cyberinfrastructure consists of computational systems, data and information management, advanced instruments, visualization environments, and people, all linked together by software and advanced networks to improve scholarly productivity and enable knowledge breakthroughs and discoveries not otherwise possible. [EDUCAUSE, 2009]

The goal of the workgroup is to build on the 2013 Faculty Assessment of the State of Research Computing effort to develop a "Research Cyberinfrastructure Vision" that significantly advances UCI research and scholarship capabilities.

Committee Organization:

- Co-Chairs: Suzanne Sandmeyer, Dana Roode
- "Executive Committee:" Suzanne Sandmeyer, Dana Roode, Padhraic Smyth, Filipp Furche, Ali Mortazavi, Lorelei Tanji, Laura Smart, Allen Schiano
- Administrative support: Nyma Cain

Core Workgroup:

- Suzanne Sandmeyer, Biological Chemistry
- Padhraic Smyth, Computer Science / Data Science Initiative
- Ali Mortazavi, Developmental & Cell Biology
- David Mobley, Pharmaceutical Sciences
- Filipp Furche, Chemistry
- James S. Bullock, Physics & Astronomy
- David Theo Goldberg, Comparative Literature and Anthropology / Humanities Research Institute
- Said Elghobashi, Mechanical & Aerospace Engineering
- Jasper Vrugt, Civil & Environmental Engineering
- Aparna Chandramowlishwaran, Electrical Engineering & Computer Science
- Bryan Sykes, Criminology, Law and Society
- Crista Lopes, Informatics
- John Crawford, Dance
- Lorelei Tanji, University Librarian
- Laura Smart, E-Research & Digital Scholarship Services Librarian
- Dana Roode, Chief Information Officer & Associate Vice Chancellor
- Allen Schiano, OIT Research Computing
- Harry Mangalam, OIT Research Computing

Additional Interested Parties:

- Elizabeth Martin, Psychology and Social Behavior
- Susan Turner, Department of Criminology, Law and Society
- Doug Tobias, Chemistry
- Keith Moore, Earth System Science
- Ioan Andricioaei, Chemistry
- Michael Franz, Computer Science
- Babak Shahbaba, Statistics
- Soroosh Sorooshian, Civil & Environmental Engineering
- Feng Liu, Mechanical & Aerospace Engineering
- Jesse Colin Jackson, Art

Research Computing Infrastructure Requirements: immediate, and longer term

Harry Mangalam, Research Computing Support, OIT
version 1.0; November 16, 2015

Table of Contents

This is a strawman version intended to provoke discussion. It is not the final version.

1. Storage.

This is a summary of a more extensive description of the problem, but that document provides no explicit recommendations. This one does.

2. The Problem

As machines of all kinds become more digital they produce more digital data, and the techniques to digest and analyze this data require additional storage as well, often 10X the raw data. Especially in research, this leads to problems in handling that research for people who are experts in their fields, but not at Information Technology. Therefore it is up to the organizations that do understand this to provide solutions that provide the optimum balance of speed, flexibility, cost, scalability, reliability, and especially, easy of use.

3. The Recommendation

UCI should immediately provide a local storage system similar to Perdue's DataDepot of 120TB (2 half-populated storage servers + metadata server + Input/Output node (IO node)), where local researchers can rent local storage for:

- active files

- short term backup
- web distribution of files
- sync & share with collaborators

OIT should buy the required chassis hardware with the actual disks being paid for by storage rental (at a slight premium to cover the next chassis and maint costs). The cost for raw disk is currently ~ $50/TB for disks of decent reliability.

This storage can be accessed in a variety of ways via the caching IO nodes that provide the actual services, so that the users rarely access the fileservers directly.

The filesystem should probably be a commercial appliance, but the IO nodes can be run by OIT to provide the services required by researchers. The IO node(s) should provide CIFS & NFS access, web services.

Optimally, the same storage would support spinning disks for bulk storage, SSDs for complex access, encrypted volumes for data needing extra security, parallel access for high-speed RW, and support for SMB/CIFS, Network File System, WebDAV, and other protocols as needed.

Users who require the very fast parallel access to the filesystem can purchase the specialized clients to do so. Such clients could be the compute clusters on campus or specialized machines at the microscope facility.

4. The Rationale

From both internal and external surveys, by far the most critical resource that faculty desire is storage in various forms. Storage is needed for actively used files, short-term backup, and long-term archives. This storage needs to be *medium to high performance* on many metrics, from streaming reads & writes (RW) as in bioinformatics & video editing, to small high-jitter RW operations with rapidly changing offsets (relational databases, access to zillions of tiny files (ZOTfiles).

Researchers particularly need storage for terms appropriate for their publishing cycles which is typically 1-2 years in the life sciences, longer in some social sciences. They also need backup for these files, since their loss can often terminate the project with the fiscal loss to the campus often in the range of $10K to multiple $M. The administrative contribution to this system is provided by the overhead of successful grants, themselves considerably assisted by the availability of this storage.

Research now often involves access to these files both on campus and off, by lab members and external collaborators. Some of this data is restricted by intellectual property agreements or privacy concerns and therefore requires special protections of permission, local firewalls, and even on-disk encryption.

Commercial cloud storage can provide some of these resources, and therefore offload the cost of technical oversight. Services like Amazon Glacier or Univ Oklahoma's Petastore are hard to beat for archiving data, which would otherwise require a ~$100K investment in hardware. For unrestricted data that has been published and no longer has to be accessed quickly, cloud archiving is the recommended solutions.

However, locality is attractive for data that has critical latency, bandwidth, security, and

legal/ownership requirements. Even peering with another UC campus (such as UCLA's CASS) cannot provide the kind of resources that many local users require. For example, it cannot provide untunneled SMB/CIFS file services to local users due to the insecurity of that protocol and the encryption adds even more latency to the connection. Additionally, any operation that requires frequent communication between client and server will take longer with the longer packet round trip times due the increased number of network hops. While opening a single small file for editing is hardly more noticeable from CASS, unpacking a 100MB archive from a client at UCI to CASS at UCLA takes about 700X longer than the local operation.

The latest version of this document can be found here.

Version 1.0; November 16
Last updated 2015-11-16 11:01:43 PST

The rationale

Research problems are increasingly interdisciplinary and complex. A fundamental way to advance research is to openly share data, which can facilitate greater collaboration and reproducibility of results that ultimately leads to solutions and new knowledge that benefits society. The 2013 Public Access to Federally Funded Research Memo from the White House Office of Science and Technology Policy directed most grant funding agencies to develop policy requirements for public access to research articles and data from grants awarded. Grant funders require data management plans as a condition of making awards. The goal is to ensure public access to publically funded research and to move towards greater interoperability and re-use of data. Some publishers (Nature, Science, Cell) require that data associated with articles be openly available.

Research data is increasingly seen as a valuable asset for the campus and takes many forms (numeric, textual, images, videos, etc.). Ownership and management of research data is intrinsically tied to research funding [UCI Strategic Plan Pillar 1]. And it can be a source of distinction that brings prestige to a university and impacts its rankings [UCI Strategic Plan Pillar 3]. New ways of using data can be game-changing. For example, pooling and mining health data is being looked at as a possible alternative to lengthy clinical trials in terms of accelerating medical breakthroughs. These new modes of using research data require a new approach, new methodologies, new training for the next generation of researchers.

The problem

Curation is the active and ongoing management of content. It is not enough to back up data with sufficient disk space and replication for disaster recovery. Bit level preservation does not guarantee data will be re-usable. Digital content is difficult to maintain for several reasons including hardware obsolescence, software obsolescence, media degradation, proliferation of file formats, obsolescence of encoding and file formatting schemes, intellectual property concerns, and funding. Preservation for re-use requires the keeping of context as well as content. That means including sufficient metadata for discovery and access, ensuring enabling technology like software is well documented and runnable (i.e. can migrate to new operating systems or can be emulated), clearly indicating how data is licensed for reuse, and dedicating funds for sustainability beyond server space. Humans need to monitor content periodically to ensure data has not been corrupted either physically (i.e. bit fixity checks) or figuratively (i.e. security has not been breached).

Bottom line, researchers are focused on doing their research and too busy to manage their data. So unless there is an easy mechanism embedded into their research process, the incentive to curate and preserve research data is low.

The vision

All UCI researchers understand and utilize best practices in data curation and preservation, incorporating data management techniques throughout the research lifecycle. Data support services are seamlessly integrated within researcher work flows and help secure research

funding. UCI created research data is openly shared and demonstrably illustrate our capacity to improve lives.

The recommendations:

Examine the campus research workflow holistically to ensure that there are the tools and services to support scholars so that they can focus their time on doing their research and that there are mechanisms that facilitate the data management, sharing and discovery, and preservation of campus research data that will support collaboration and interdisciplinarity.

Strengthen existing research data management consulting services within OIT, the Libraries, and the Office of Research. Actively promote the full suite of data services which covers all stages of the research cycle:

- management planning – guidance in writing plans, choosing a repository to host your data set, obtaining persistent URLs, choosing the appropriate license
- selection and use – obtaining and organizing data, utilizing data science techniques mining and visualizing, supporting collaborative use of data with easily accessible yet secure network storage, publishing and citing data.
- storage and retrieval – selecting and applying metadata standards, semantic services and ontologies, reformatting to preserve-able file formats, search engine optimization

Provide a robust training program promoting data literacy skills

Increase use of current repository services for data. Grow data collections in UCI Dash. Promote reuse of UCI data. Track usage and impact and report on it. Share our success stories.

Contribute to study and use of emerging financial models for data preservation. Support UC participation in the Digital Preservation Network.

Track and internally share UCI data management plans from successful grant applications.

Consider a campus-wide implementation of ORCID (unique identifiers for researchers) which would minimize name disambiguation and streamline the workflows from manuscript to grant submission and beyond.

Potential strategies for implementation

- Integrate currently distributed data support services by creating a campus coordinating body to share knowledge, develop training programs, target outreach, and increase usage and impact of UCI research data. Include Office of Research, Libraries, OIT, Data Science Initiative,
- Create interdisciplinary data curation pilot projects representing UCI research priorities to gauge technical requirements and support required
- Fund additional FTE within data curation services to meet documented demand
- Cultivate data curation experts within academic programs to provide peer support and liaise with data support services
- Secure funding for post-doctoral research in domain informatics.

- Incorporate data curation training within the data science initiative
- Design incentives to promote best practices in data management, rewarding exemplary researchers implementing data curation within all stages of the research process
- Utilize collaborative work space to expand use of data mining and visualization
- Support work flow tools which aid the data creation to preservation pipeline
- Research and document the relationship between data curation practices and extramural funding
- Assist researchers with data audits, identifying digital data at-risk, and funding the preservation re-formatting of data with high potential for re-use and impact.

Appendix A: Foundational Cyberinfrastructures, Services, and Facilities,

All campuses should make these once-forward-looking, now-basic infrastructures, services, and facilities available and accessible to their researchers. This may be done by leveraging federated or intercampus services, by looking to cloud providers and external vendors, or with local campus solutions as is most efficient, effective and applicable.

It is recommended that the ITLC's UC-Wide Research IT Group be called upon to elaborate, extend, and specify these foundational (also called "birthright") infrastructures, services, and facilities, revisiting expectations on a regular basis (perhaps triennial, as three years is often considered the lifespan of modern computing equipment).

Colocation

Although the trend is towards virtual machines in a cloud environment, not all research needs are well suited to distant locations, nor is the current pricing and performance efficient for all applications (notably high-frequency I/O, big-data, and high-performance computing applications). Centralized locations for server storage preserve precious lab space and allow server owners to take an important first step towards the cloud — adjusting to having their servers and instruments in separate locations.

High-Performance Computing

The surge in data set size and the improved speed of data collection has caused a requirement for high-performance compute cycles across all disciplines.

High-Speed Networking

Campuses should have high-speed connections to the research Internet (in 2015, 10G should be considered a minimum requirement) and should provide a reasonable minimum speed to the research wall jack (in 2015, 1G should be considered). Where possible, campuses should be able to provide a higher speed connection to data-intensive labs and facilities.

Working Storage and Backup

A basic allocation of working storage is critical to the functioning of nearly all research labs. Practical backup services protect not only the researcher but also critical UC intellectual property. Storage options should take into account the needs

occasioned by compliance issues such as patient data, personally-identifiable information, export control etc.

Data Curation

Central facilities where valuable research data sets can be curated, preserved, searched and reused must be developed. Data citation is growing in importance to both the individual and the institutional academic reputation, and new data sharing and archiving requirements at federal agencies mean that efficiency is best realized where these facilities are centrally provided.

Information Security

All researchers should have access to appropriate tools, services, guidelines, and compliance-sensitive best practices to assist in the protection of critical research data and associated high-risk assets. These may include tools to interrogate data stores for the existence of protected forms of data, scanning services designed to uncover vulnerabilities without interfering with sensitive experiments, and guidance on safe use of cloud and other remote-storage technologies.