A Vision for Research Cyberinfrastructure at UCI

Draft 4.0, February 2016

Table of Contents

1. Executive Summary
2. Our Vision for UCI Research CyberInfrastructure
3. Research Data Storage
    - 3.1. Vision
    - 3.2. Current
    - 3.3. Competitive Risk
    - 3.4. Recommendations
        - 3.4.1. Cost
        - 3.4.2. Delay
4. Curation
    - 4.1. Vision
    - 4.2. Current
    - 4.3. Competitive Risk
    - 4.4. Recommendations
        - 4.4.1. Cost
        - 4.4.2. Delay
5. Research Computation
    - 5.1. Vision
    - 5.2. Current
    - 5.3. Competitive Risk
    - 5.4. Recommendations
        - 5.4.1. Cost
        - 5.4.2. Delay
6. RCI Working Environment
    - 6.1. Vision
    - 6.2. Current
    - 6.3. Competitive Risk
    - 6.4. Recommendations
        - 6.4.1. Cost
        - 6.4.2. Delay
7. RCI Staffing Needs
    - 7.1. Vision
    - 7.2. Current
    - 7.3. Competitive Risk
    - 7.4. Recommendations
        - 7.4.1. Cost
        - 7.4.2. Delay
8. Education and Training
    - 8.1. Vision
    - 8.2. Current
    - 8.3. Competitive Risk
    - 8.4. Recommendations
        - 8.4.1. Cost
        - 8.4.2. Delay
9. Organizational Considerations
    - 9.1. Vision
    - 9.2. Current
    - 9.3. Competitive Risk
    - 9.4. Recommendations
        - 9.4.1. Cost
        - 9.4.2. Delay
10. Budgetary Requirements and Funding
    - 10.1. Vision
    - 10.2. Current
    - 10.3. Competitive Risk
    - 10.4. Recommendations
        - 10.4.1. Cost
        - 10.4.2. Delay
    - 10.5. Charging Models

1. Executive Summary

Theory & experimentation have been supplemented over the last deacde by modeling and data. All those foundational pillars of science require a correspondingly robust Research CyberInfrastructure (RCI). RCI capability at UCI is well below that of other R1 universities and campus research & scholarship is already impaired by the lack of investment in Storage, Computation, & Staff. Without substantial new investment in this area we will increasingly fall behind and lose the ability to even sustain current levels of research, much less accelerate and expand it.

To support the competitiveness of UCI researchers in the evolving cyber environment, our vision is to:

- Change how RCI is coordinated, funded, and delivered, by placing the responsibility for its support, etc under a separate organization, the UCI RCI Center (UCI-RCIC). [I personally hate this name and all names containing CyberInfrastructure, but...] In order to concentrate attention and responsiveness in this area, we propose to place RCI direction under the control of its end users. That is, to break the current Reseach Computing Support out of OIT and establish it under the direction of a supervisory committee of interested faculty who will set direction, staffing, act as coPIs on grant applications, and provide feedback on suggestions for increased performance from the staff.

The RCIC would also coordinate with other like-minded units on campus (the Data Science Initiative, the UCI/SDSC Computation program(?)) that need RCI and support for research and instruction, since requirements for both are growing rapidly. Center & shared equipment grants would also be coordinated via the collaboration between the RCIC and other such units. Expanded Description

- Initiate construction of a scalable Petabyte storage system that can be accessed by all researchers that can be leveraged to provide the multiple types of storage and data sharing that assist most research endeavors. This includes centralized active file storage & backup, easier sharing of even large data sets, secure web distribution of data, file syncing if necessary, and tiered data archiving locally and to cloud archives.
- Provide a mechanism to upgrading and renewing the CPUs of the UCI's compute clusters. (expand?)
- Provide a baseline or *birthright* level of storage, connectivity, and computing for faculty in all disciplines. If the funding requested is provided, within a year, we can provide >1TB of robust, secure, central storage, 1 Gigabit/ second network connectivity, and 100,000 hours/yr of 64bit compute hours using Open Source Software (OSS) to each faculty member who requests it. These allocations should increase over time and could be augmented to support research projects through grant funding. Expanded Description
- Establish a widely available & scalable Research Desktop Computing Environment (RDCE) to facilitate computational & data science research and teaching. This environment would include access to shared software (both proprietary and OSS), high performance computing resources, visualization tools, tools for data sharing and collaboration, assisted access to external UC and national facilities, and appropriate cloud resources. While the RDCE would be more secure than traditional desktop computing, more secure computational and storage environments would be provided for compliance with Data Use Agreements, and other information security frameworks (e.g. HIPAA/FISMA) and Data Sharing policies (e.g. for Genomic Data Sharing). Expanded Description
- Hire staff to support the RCI and provide much more assistance to researchers to fully leverage the hardware and software . None of the projects under discussion will advance without staff expertise, which is not cheap, but with UCI at the bottom of RCI staff by most measures, this is critical. Career staff would maintain, upgrade, and expand RCI operation, train students, other staff, & faculty in current computational techniques, document processes, provide catalytic programming services, assist with grant prep, work with existing staff in other units to provide statistical, analytical, and advanced computing needs, and assist in maintaining compliance with federal requirements for data robustness, backup, retention, re-use, archiving, sharing, and security. Expanded Description

Executing this vision will speed the ramp-up of research programs, increase productivity by offloading in-lab administration & support, provide a much higher baseline of RCI services in all Schools, and offer much-increased security and access to tools for all researchers.

2. Our Vision for UCI Research CyberInfrastructure

Information Technology (IT) and the ability of researchers to exploit it, plays a critical and evolving role in University research, teaching, and other scholarly pursuits. The goal of the RCI Workgroup is to

provide recommendations that significantly advance and accelerate UCI research as widely and as economically as possible. Most key recommendations can be effected within months of the requested funding.

Existing RCI staff and facilities have provided a large amount of quality service to campus researchers since 2013 when the recommendations from the Faculty Assessment of the State of Research Computing (FASRC) cited above were issued. OIT, with the support of faculty and the Office of Research, has received two NSF grants to enhance RCI. The first funded UCI LightPath a 10 Gb/s Science DMZ restricted to research data. *LightPath* now connects the two largest campus compute clusters with labs in seven additional buildings. The second grant will fund a Cyberinfrastructure Engineer for two years. The UCI Libraries launched the Digital Scholarship Services unit to support data curation and promote Open Access to data produced by UCI researchers. The 2 largest compute clusters, underfunded and aging as they are (HPC and GreenPlanet) have been used to produce 10s of papers in multiple domains.

However, in spite of these isolated successes, the RCI at this university remains a distinct weakness, to the extent that some researchers still rely on RCI at other institutions. Specialized computational and storage resources especially are notably underfunded at UCI, with some facilities such as HIPAA/FISMA-secure research computing facilities completely absent.

RCI impacts every aspect of research and scholarship and must be addressed as a campus priority. Besides computation *per se*, it includes networking, storage, data management, and support services required by all disciplines. As well, since requirements continue to expand, there is a critical need for long-term RCI planning as well, one reason for establishing and empowering the above-mentioned RCI Center.

3. Research Data Storage

3.1. Vision

Researchers should be able to interact with large data sets as easily as they interact with email and desktop documents. Tools to compose, share, backup & archive, forward, edit, analyze, and visualize multi-TB datasets should be available to all faculty. A requirement that underlies all those aims is the secure & reliable physical storage required to contain that data.

3.2. Current

DM

I get where this section is going, but I think it needs a strong statement of why it matters at the very beginning of the section. Otherwise, we're forcing people to read four paragraphs of relatively technical background before we tell them, "look, we have a problem". I suggest we say there is a problem first and then explain why there is one and what it is.

Much research at UCI generates or uses vast amounts of data; researchers are largely left on their own to manage it and to prevent catastrophic loss of sometimes critical data. This situation is inefficient, unsustainable, exposes UCI to liability, and is thus highly risky for research, legal, and fiduciary reasons.

All research organizations are seeing data storage requirements increase dramatically as more devices produce higher resolution digital data. Without access to robust, scalable, *medium to high* performance storage, modern research just does not work. The various types of storage, metrics by which they are distinguished, and rationale are LINKTO[discussed here in more detail], but universal access to storage for recording, writing, analysis, backup, archiving, dispersal, and sharing are the *de facto* **papyrus** of this age. The on-campus availability of the data is not enough. The data must be available globally to those who have valid need for it, and in many cases, secured against unauthorized access for reasons of privacy, sensitivity, intellectual property, or other legal prohibition.

Such storage systems also require automatic backup, since data loss can unexpectedly abort a project with substantial fiscal loss as well as incurring long-term penalties from funding agencies.

While some of this storage can be outsourced to Cloud providers, much of the storage a research university requires is not amenable to remote Clouds. Much research storage must be *medium to high* performance, from streaming reads and writes as required in video editing and bioinformatics, to small high Input-Output (IO) operations per sec, as with relational databases. These characteristics require a local LINKTO[Campus Storage Pool], which can be leveraged to provide much of the storage described above by providing specialized, highly cached *IO nodes* communicating in parallel to the storage pool.

Such *IO nodes* could provide desktop file services, web services, file-syncing and sharing, archival services, and some kinds of backup LINKTO[see Diagram 1].

The minimum standard for a useful Campus Storage Pool is one with: - Large capacity, scalable to multi-Petabyte size. UCLA's CASS is an example of such a Campus Storage Pool, tho an expensive one. - Low latency, high-bandwidth access to Compute Clusters and other analytical engines. - Backed and mirrored up to multiple locations (including off-campus) - Physically secured with appropriate authentication/authorization fences to enable secure file sharing and collaboration among project teams internationally. - Accessible via a range of protocols; As an example, LINKTO[Purdue's Data Depot] is available to Windows and Macs as a network drive on campus, and accessible by SCP/SFTP/Globus from anywhere.

The Campus Storage Pool would be available to all faculty as a baseline, no-cost service. Additional storage needs would be funded through a cost recharge model where the administration would support the cost of the storage server chassis and the disk-equivalents would be bought by researchers.

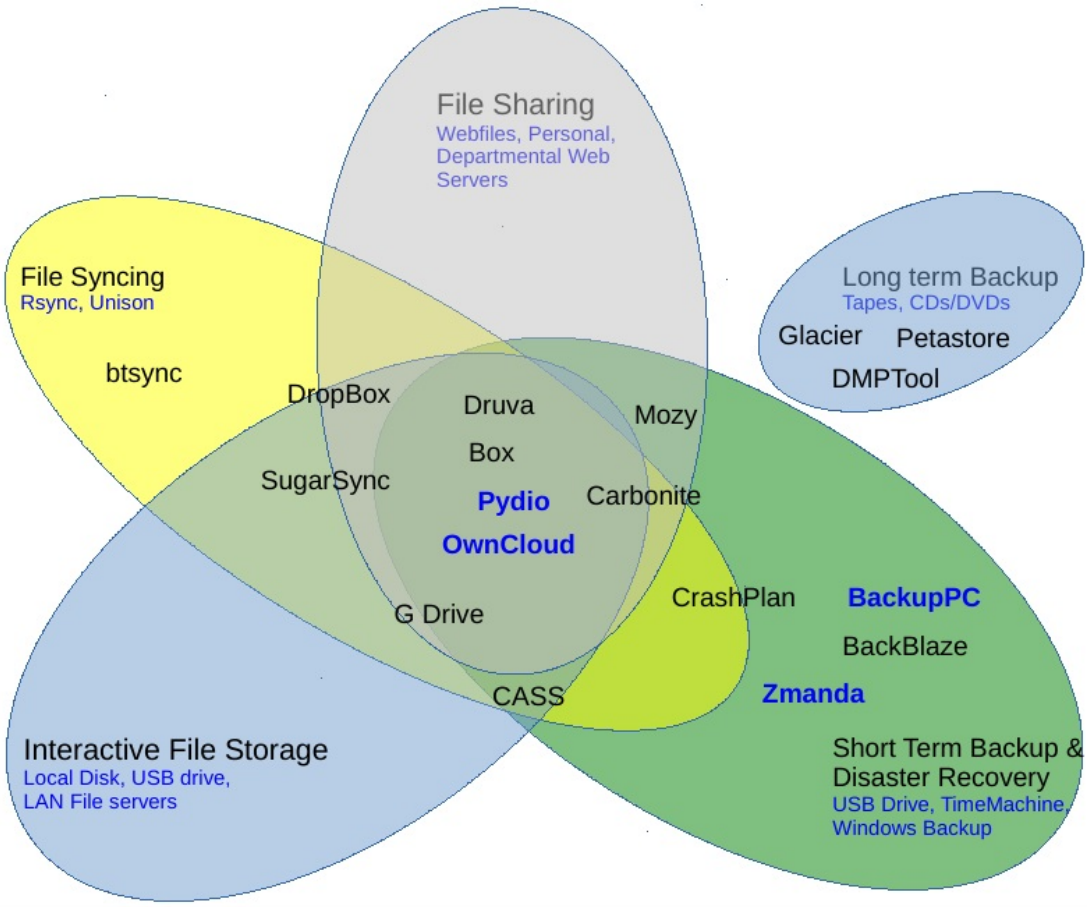

Figure 1.

Overlapping service requirements by Application and Service (Cloud or local). Labels at the outer lobe of each ellipse denote the generic service and some local examples. Black labels show commercial/cloud services. Bold blue names are Open Source services that show promise.

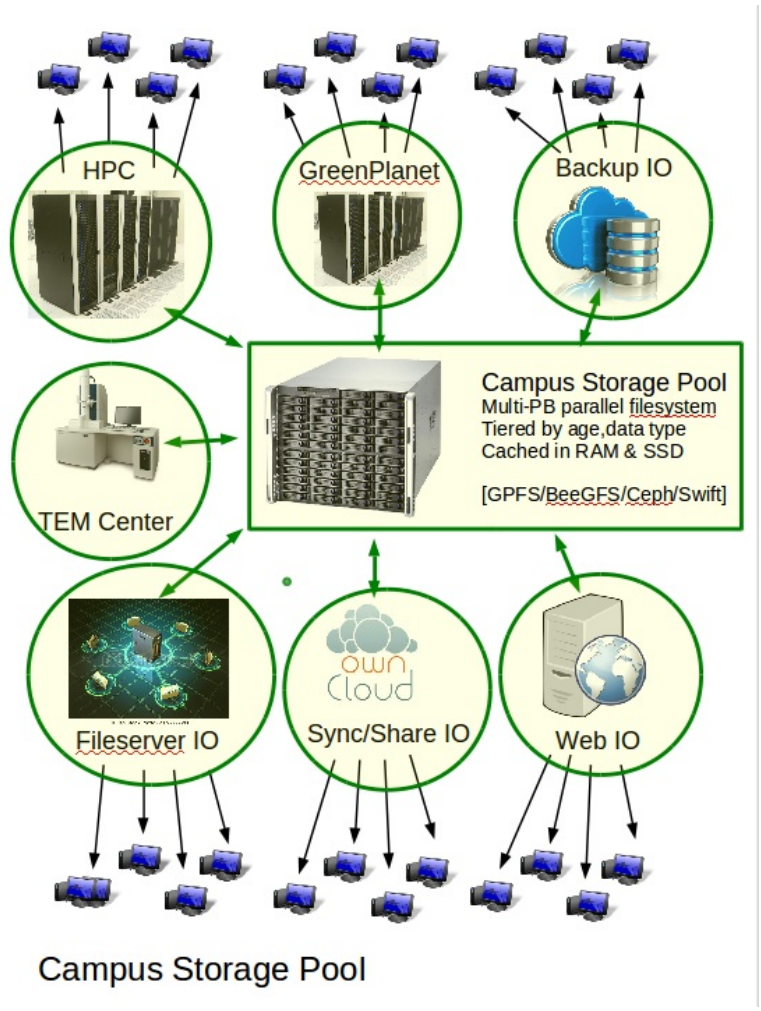

Figure 2.

Functional Diagram of Campus Storage Pool with the PB sized central pool feeding high performance links and Input/Output (IO) nodes via parallel connections (in green). Macs & PCc would connect via those IO nodes for file storage, Web access, backup, and sync/share operations using commodity ethernet connections (in black).

3.3. Competitive Risk

There are 3 risks of not implementing this. First is the risk of not providing what is increasingly considered to be a university *birthright* - part of the basic research infrastructure. This results in non-competitive grant applications and inability to compete for attractive hires. The second is the financial risk of not providing backup of research data. Currently considered the very poor cousin of administrative data, research data is the data that actually *brings in* money, altho the *dollar density per byte* is much lower most of the time. The 3rd risk is the fiduciary risk of not protecting data that must be shielded for intellectual property, legal, or security reasons.

3.4. Recommendations

We advise that this is Priority #1. We recommend the immediate funding of a baseline 500TB Campus Storage Pool and matching backup system, based on a review of the technical details LINKTO[described in detail here].

3.4.1. Cost

Based on the technical details mentioned above, this system would cost on the order of $300,000 for a raw storage of ~1PB & networking hardware, and another $200,000 for additional software, support, documentation & faculty assistance, networking support, and software integration, for a total amount of about $500,000 (!!this is largely made up on the spot!!).

3.4.2. Delay

The delay largely depends on the delay in funding but also the delay in review. OIT has already started an internal review for aspects of the backup system, but the technical review for the Campus Storage Pool has not begun. We estimate the time to review, spin up a test platform, and agree on such a plan is 3 months, given the equivalent of 2 FTEs for that period. The delay from that point until implementation is 4 months, depending on the delay in signing agreements with vendors.

4. Curation

4.1. Vision4.2. Current4.3. Competitive Risk4.4. Recommendations

4.4.1. Cost

4.4.2. Delay

UCI researchers are increasingly being asked to manage and share their research data in order to comply with funding agency requirements1 and system-wide Open Access policies2. Grant agencies direct researchers to document plans for disseminating and preserving their work. They must also demonstrate implementation and show that publications and data are openly available for use and reproducibility. Failure to do so risks the ability to retain funding and win subsequent grant applications. The NIH, for example, recently announced they will suspend funding for awardees who do not document deposit of papers into PubMed Central.3 So new services, tools, and staff support are needed to insure and grow grant funding and to save UCI researchers valuable time. Examples: ICTS is working with the Libraries' to develop procedures to identify appropriate versions of publications for deposit, verify citation information, and document required persistent link identifiers. Additional support staff is required to scale campus training programs on data management plans and open access distribution as funder mandates increase. Professional staff are also needed to perform enabling activities to keep data usable when large discipline-specific repositories aren't available or suitable.4 This is especially germane for Arts and Humanities data. The Libraries' administer a number of digital repositories and metadata services which preserve content and make it discoverable.5 Repositories and metadata alone won't fulfill a vision of cyber-research. There are new modes of inquiry and research where scholars (individually and collaboratively) engage with online libraries/archives using methods like augmented editions, data mining, visualization, deep textual analysis, concept network analysis, mapping, etc. Examples: Librarian expertise is integral to creating scholarship tools based on digital collections such as the current NEH funded project creating linked data and visualization tools for analyzing digital representations of artists' books. The wait time grows to fulfill project requests such as: designing infrastructure and digitizing content for an international online Critical Theory Archive; integrating crowdsourced translation tools for Ming dynasty organizational names within the China Biographical Database; and assisting a doctoral student procuring social science and humanities data (local crime statistics, transcripts of Dragnet radio shows) to map perceived and actual crime locations over time and create novel publishing mechanisms to interact with the model.

Campus support for data and digital asset management is distributed, loosely coordinated, and not staffed to the level of peer institutions.6 We advise

- Immediate funding for a Library Data Curation Specialist to support funder compliance, manage collections of campus produced data, work with Office of Research to implement data management training program, and promote open access
- Designing a campus space to bring together and highlight data and digital content management tools and services now distributed across the Libraries', OIT, Humanities Commons, and other units.
- In the longer term, funding Digital Humanities Librarian and additional programmer analysts to develop scholarship enabling tools over digital collections

1 - A growing number of government and private funding bodies have requirements. For example, DOE, NSF, NIH, NEH, Gates Foundation, Howard Hughes Medical Institute, MacArthur Foundation, Gordon and Betty Moore Foundation mandate preservation and sharing of results

2 - UC Open Access Policies http://osc.universityofcalifornia.edu/open-access-policy/index.html

3 - "For non-competing continuation grants with a start date of July 2013 or beyond NIH will delay processing of an award if publications arising from it are not in compliance with the NIH public access policy" http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-042.html

4 - Such as a small set of microscopy images used to devise new methods for evaluating cytoskeletal orientation deposited in UCI DASH data sharing repository http://bit.ly/10340OR

5 - Calisphere, DASH data sharing, eScholarship, EZID, Merritt are provided centrally by UCOP/CDL. UCI contributes to the development, governance, and promotion of CDL hosted repositories and tools while managing local configurations and work flow. UCI fully administers the UCISpace repository.

5. Research Computation

5.1. Vision5.2. Current5.3. Competitive Risk5.4. Recommendations

5.4.1. Cost

5.4.2. Delay

Research computation provides the computational power used by our research, and campus support for this area is far behind that of peer institutions. Computation is essential to collecting, processing and analyzing increasingly large and complex data with appropriate speed. Broadly defined, RCI includes computational resources that support these endeavors from smartphones and tablets, laptops and desktops, shared servers and clusters as well as national and international facilities. Scientific computing needs are growing exponentially, and High Performance Computing (HPC) resources are required to address this.

High Performance Computing

Computation is an integral aspect of modern science. Most research groups at UCI rely on computation in some form. Capture and analysis of dynamic systems is computationally limited. These systems range from interactions of atomic particles and molecules, to cellular and organismic living systems, to artistic performances, to urban environments, to geophysical dynamics and astronomical events. Modeling of such dynamic systems is even more computationally intensive. Similarly analyzing and understanding interactions and causality at the level of molecular dynamics, social and legal networks, health, business, economics, and creative arts is computationally intensive.

The convergence of data-driven experimental science coupled with high-throughput technologies and computer-driven simulations has created a huge need for computing power which is currently not met at UCI. Despite the growing demand for scientific computation, UCI has only two major computational facilities, the Physical Sciences Greenplanet and campus HPC clusters. Both of these facilities are operating with aging equipment and largely outdated (?) commodity hardware. UCI presently faces a large and rapidly growing shortfall of at least several hundredseveral 100 teraFLOPS in computing capability. This shortfall is limiting the ability of UCI faculty to perform their research and to compete for extramural funds. Many competing institutions are much better equipped; for example, Purdue, which is of comparable size to UCI, has a Conte community compute cluster providing 943 teraFLOPS. This is more than ten times the capability of Greenplanet. This need was highlighted in the recent Research CyberInfrastructure Vision Symposium at which new faculty described their surprise at UCIâs limited computational resources. In fact, many are continuing to rely on resources at their previous institutions (University of Washington, University of North Carolina, UC Berkeley etc.) through professional contacts. This is not only detrimental to our reputation, but creates potential security risks.

SDF

As I recall, the reliance on non-UCI resources was more in terms of storage than computational cycles and this is in keeping with the comment about security risks. Thus, we might want to move the last 3 sentences to the section on storage rather than bundle it with compute cycles.

Simulations on tens of thousands of cores are becoming the new de facto standard for computing-enabled and data-driven science. Science at this scale simply is not possible at UCI right now. Single-investigator funded node purchases help maintain the status quo, but their volume is too small to shift UCIâs competitive position.

The maintenance of basic research facilities such as buildings, lab space, or shared research facilities

including the Laboratory of Electron and X-Ray Instrumentation (LEXI), Transgenic Mouse Facility, Greenhouse, Optical Biology Core Facility and Genomics High-Thoughput Facility are essential to UCIâs success. Computational facilities should be considered a comparable and essential aspect of UCIâs basic research facilities, and maintained accordingly. Adequate computational hardware is not only just as important, but where lacking, can limit the impact of these other research resources.

xxIn addition, the effective lifetime of computational hardware is only about three years. A sustainable funding model is required for periodic upgrades of hardware and software. xx

Technical Expertise and Software

Any utility derived from hardware requires expertise and software that allow researchers to exploit it. Enabling servers to work together effectively is a complex process, requiring broad system administration expertise. In addition to high speed network connections and parallel file systems, special scheduling software is necessary to guarantee that jobs from different researchers are assigned appropriately and efficiently to combinations of CPU cores and servers. The expertise for developing and debugging parallel applications is even more specialized.

The Linux Operating System is widely regarded as the most important environment for scientific computing. Because of the low cost and high scalability of Linux-based clusters, the ease of writing software for the command-line (vs graphical applications), and the speed of Open Source Software (OSS) propagation, the overwhelming majority of research software is written first for Linux. It is necessary to train and retain Linux-savvy system administrators and programmers who are able to set up and maintain Linux- based software, servers, large filesystems, and configure environments.

High Performance Computing Vision

SBS

This strikes me as a section that might suggest some goal for the increase in computation? Some metric of increase. Otherwise it just sounds like we need a little more and sustainable and we can keep up. We said Purdue was 10X GP.

DM

I agree. How can this be projected to grow? Doubling in computer power every N years? Financial investment to track with total research outlays? Some sort of rule of thumb?

A major investment in support of maintenance and expansion of higher performance compute clusters is needed. Annual funding must also be identified for the appropriate level of support staffing and to enable the computational infrastructure to remain current. This does not require wholescale renewal each year, but it should provide basic capabilities for all researchers and a framework that can be augmented by grant funding.

A baseline compute hour allocation on Linux compute clusters should be made available to all researchers, with additional computation required by specific projects addressed through grant funding and other mechanisms. Additional capacity must be allocated for educational use to facilitate teaching on modern parallel computing and related techniques.

An increasing amount of social, medical research (as opposed to specifically patient) and physical data have special security requirements to satisfy federal funding agencies. In order to remain competitive, campus investment is also needed to support demonstrable establishment of HIPAA/FISMA secure cluster resources.

6. RCI Working Environment

6.1. Vision6.2. Current6.3. Competitive Risk6.4. Recommendations

6.4.1. Cost

6.4.2. Delay

DM

In keeping with my suggestions above, might want a single sentence summary to motivate this section before giving background.

The working environment that researchers use varies considerably across and within disciplines. In almost every case, the most immediate aspect is the operating system, its user interface and application software present on personal computer systems and, increasingly, on various other âpersonal devices.â While the majority of faculty find themselves preferring Apple OS X or Microsoft Windows, as discussed above, Linux is the system of choice for high performance computing and for those who are using specialized or open-source software (including software they develop). Campus RCI must include support for all of these working environments.

DM

I still don't totally get the point of this paragraph/section. Is it just that we have to support faculty using different OSs? Or is it that we need to provide the software faculty need? I get that people use some particular computing environment, but I think we might need to get to the aspect of this that requires campus support more quickly (or at least give an example of it) to avoid losing people. Right now I still read this first section down to the underlined "...Research remote..." text as basically saying, "People use lots of different stuff for research. We need to help them collaborate, share, and visualize data."

In addition to operating systems with the user interfaces which researchers most easily interact, working environments include applications and data sources (licensed at a cost, open source, and even locally developed) which may be broadly used or specific to particular disciplines. Whether such components are part of a centrally directed campus-wide cyberinfrastructure or are best viewed in terms of locally directed components, all should be coordinated and integrated at a campus level.

One class of applications and services that deserve special mention are those that enable collaboration, both on campus and with colleagues worldwide. While there are significant variations in collaboration practices and preferences across and within disciplines, a very important aspect of cyberinfrastructure is facilitating collaboration from interpersonal interactions, to data sharing, to information dissemination.

Visualization software and facilities are additional key requirements in RCI working environments.

SBS

Good if Crista Lopes could look at this particular para-she has commented software for interactions is key.

The Research Remote Virtual Desktop

Maintaining research computing environments with the requisite software requires significant administration and upkeep. One method that has proven effective here and elsewhere is the provision of standardized âvirtual desktopâ environments using Remote Desktop protocols.

SBS

Here we say that we have and are using effectively; just below here we say we will establish. Could we have an example of how this happens currently? I am only aware of admin software unless we mean like Galaxy or HPC software? How would the new system improve or differ?

This is an efficient mechanism that provides an exportable display from a large server or cluster that can be brought up on any device, anywhere. It centralizes storage for convenience, cost, backup, security, and reduces administration costs. This approach can be used for providing applications for native Windows, MacOS and Linux. It also allows sharing of research software licensing across a large set of users who make occasional use of a particular title to lessen overall campus costs for software that is not being used constantly.

We propose to make a new Research Virtual Desktop service available to the campus to facilitate access to well-maintained software in an integrated environment.

Access to External Resources

UCI hosts significant components of a robust cyberinfrastructure: the campus wired and wireless network, the UCI-LightPath high-speed science network, systems housed in the OIT Data Center (including high performance compute clusters), resources in various academic units, and staff in OIT, the UCI Libraries and various academic units who provide RCI services.

The network also provides access to external facilities, both those dedicated to academic work, such as the San Diego Supercomputer Center (SDSC) and commercial providers. Commercial services include Amazon and Microsoftâs Azure for computational needs and Google Apps and Microsoft Office 365, for cloud communication, collaboration, and storage services. This last group of resources form an increasingly essential component of research cyberinfrastructure capacity to enable and to facilitate scholarly productivity and research efforts.

In addition to network connectivity, providing access to off-campus resources includes addressing contractual agreements, security measures, and access permission (authentication and authorization). UCIâs participation in Internet2âs federated identity management confederation (InCommon) allows UCInetID credentials to be used to access external resources. These range from the world-wide wifi access provided by Eduroam to InCommonâs âResearch & Scholarshipâ service providers (e.g., the GENI Experimenter Portal, the Gravitational-wave Candidate Event Database).

The list of relevant computational and information resources available via the network varies considerably by discipline and is continually changing. Providing expertise in selecting external resources is as important as the resources themselves.

7. RCI Staffing Needs

7.1. Vision7.2. Current7.3. Competitive Risk7.4. Recommendations

7.4.1. Cost

7.4.2. Delay

âPeopleâ are the part of RCI that make it all work and enable it to be effectively leveraged by researchers. IT professionals are required to support, maintain and enhance RCI services. Researchers also require access to research computing specialists to help implement effective IT solutions to research problems and assist with research workflow.

UCIâs current level of RCI staffing is substantially below that of comparable peer institutions. We have approximately 3 FTE assigned to supporting the Greenplanet and HPC clusters across Physical Sciences and OIT and a similar number assigned to supporting researchers in other ways. Purdue and Indiana University have 20 or more staff assigned to these tasks. UCLA has a RCI team of 21 FTE, 11 of which focus on compute clusters and other high performance services. UCB has 21 individuals (15 FTE) on their Research IT team. Current Physical Sciences and OIT RCI staff support is insufficient to cover the areas for which they are completely or partially responsible. Including computing architecture and operation, research data storage and transfer and programming support, and graphical information systems.

Additionally, tThere are 3 FTE to support RCI within the Digital Scholarship Services unit of the UCI Libraries. As an exemplar, Purdue has a dedicated data curation center and 7 FTE supporting data curation work: 4 data specialists, a software developer, a data curator, and an administrator.

Current UCI Libraries staffing support is insufficient to cover the areas for which they are responsible which include: promoting researcher compliance with funder data usage requirements and best practices in data curation; training in data curation; promoting Open Access repositories; development of tools to describe data domain ontologies; development of robust institional repositories and exhibits, particularly with humanities applications; and negotiating software licensing from external sources.

RCI staffing Vision

```
RCI Staffing
```

Compute Cluster Support: Additional system administrators/engineers are needed to help maintain and enhance existing campus computational clusters including complex networking, queuing and job control, runtime check-pointing, security and other tasks that are currently not being done or are significantly delayed. Additional need is driven by the large increase in both sophisticated and naÃ¯ve users and

application diversification (requiring software installation, user training/assistance, code optimization and development).

Storage System Development and Support: Additional system administration/ engineering staffing will be required to develop, maintain, and then assist users in adopting the proposed research storage service (see Research Data Storage).

Research Computing Support: The workgroup additionally recommends adding staff to provide for: installation and networking of workstations for advanced instrumentation; user training and programming to assist users with open-source Linux based software; user training and assistance with statistical packages R, SAS, and MatLab; assistance in identification of appropriate commercial software; development of visualization tools for research results, and assisting investigators with establishing appropriate levels of security to be in compliance with funding agency requirements.

Discipline-Oriented Specialist Staffing

Library Sciences Specialist: The library identified a need for an additional data science librarian with vision and leadership skills to grow library-related data management consulting services and administer data repository systems. This position would coordinate activities with OIT and Office of Research staff supporting data management and liaise with staff within the schools. The individual filling this position would should have expertise in research funder requirements for data preservation, project management, functional requirements specification and application development, metadata, and digital preservation. In addition, the position requires experience administering the common open source repository systems for the management, discovery, and access to UCI produced data.

Discipline-oriented âRCI S specialistsâ: Full RCI support requires professional level specialists with computer science, but also disciplinary backgrounds (e.g. math, chemistry, biology, social sciences, engineering, humanities and the arts) working in partnership with research teams. These staff positions might be jointly sponsored by RCI and the Schools. They could include accomplished part-time graduate or even undergraduate appointees. We refer to these as RCI specialists. Examples of important skills in addition to discipline domain specific knowledge stipulated above would be: High performance programming skills and techniques (OpenMP, MPI, GPGPUs, etc.); high performance networking, data transfer and storage; statistical and mathematical computing tool utilization; scientific visualization; Linux and open source software; database design, construction and utilization; and website development.

We expect to increase RCI support service staffing for both mission critical RCI services and RCI specialists incrementally over the next several years based on annual assessment of unmet needs and effectiveness of current services.

Accessibility and awareness of RCI Resources

The increasing complexity and planned enhancement of computational resources requires systematic outreach so that the full potential of the RCI investment is realized. As discussed below, establishment of the RCIC Research Cyber-Infrastructure Center (RCIC) is one aspect of this. We describe here implementation at the level of RICI RCIC staff outreach to the UCI community. An RCI staff specialist will be tasked with coordination of outreach to the campus community regarding RICI RCIC facilities and resources.

Outreach and training coordinator: One or more of the RCI specialists familiar with research computing software, resources, and techniques will provide outreach to schools and distributed research staff and be responsible for maintaining a RCI Program Website that will contain a directory and links/descriptions to relevant services and resources; federal guidelines and templates to facilitate compliance with data use policies; links/descriptions to shared software and means to request assistance with new acquisitions; comprehensive calendar for RCI related workshops and events; and current campus RCI site map. Such a specialist would also develop a recommended minimum training program for incoming faculty, students and staff to ensure baseline awareness of resources and best practices.

The team: All staff at UCI who support RCI, whether they are in the central RCI unit, other units such as the UCI Libraries, a school, research group, or ORU/Institute would be considered part of an extended UCI RCI support team. The RCIC coordinator would facilitate this approach, which would be further strengthened through joint central/school assignments where appropriate.

8. Education and Training

8.1. Vision8.2. Current8.3. Competitive Risk8.4. Recommendations

8.4.1. Cost

8.4.2. Delay

The need for computational and data analysis skills is increasing rapidly and impacting almost every research discipline. There are very strong statistical and computer sciences departments at UCI in which students systematically learn from experts and a thoughtful curriculum. However, no solution is currently in place for incoming students in disciplines that have not been traditionally computationally oriented, or for postdocs, faculty and others who have limited time for formal classes, but going forward require working knowledge in these areas.

Vision

Training facilities: Currently many/mostall classroom buildings have wireless access that can continue to update. However, there is also a need for specialized training in HPC computing that is not currently met. Some number of classrooms should be available to enable actual teaching in the HPC environment for state-of-the art education. but some require upgrades. Regular and Computer Classrooms are in great demand and bringing the new Instructional building online is an important goal. A subset of computer classrooms should be equipped with visualization and scientific software to facilitate RCI training.

Education and training for base level proficiency in use of computational tools is a necessary aspect of preparedness to function in the modern world. At UCI this training should be integrated into the undergraduate and graduate curriculum. The university should instituteestablish a proficiency requirement for a minimum skill set in computation and use of computational tools for undergraduates at the earliest possible time. Formal graduate or undergraduate courses related to interdisciplinary computation (ie not all computation/data courses for majors) will be listed on the RCIC website. These could include as examples, graduate courses in statistics for biologists from several departments and in bioinformatics from ICS and from Epidemiology.

The RCI working group specifically addressed training in the context of faculty, undergraduate, graduate, postdoctoral and visiting scholar researchers. Education and training targeted to this group is typically either remedial and intensive or targeted to specific and relatively immediate needs. Researchers find it difficult to accommodate a formal class schedule and expect to do significant amounts of learning independently. Research training acknowledges this and will focus on three approaches: 1) individual or small group targeted training initiated by investigators or prompted by RCI computing staff and domain specialists in response to perceived need; 2) relatively short and intensive workshops (e.g. day or week-long) for groups of ten to thirty combining lecture and hands-on use of tools with rolling on-line sign up to allow instruction as soon as critical number of students is reached; and 3) on-line training. The RCIC would not be responsible for each of these modes of training but would ensure that the RCIC website lists the available modules and has links to sign-ups and schedules.

RCI training will include:

Parallel programming, data analysis, and visualization techniques. Data asset management Computing skills: basic hardware and software knowledge, Linux operating system, shell commands, simple scripting, basic programming, web tools, etc (Software Carpentry courses provide a framework). Also, leveraging Macintosh and Windows environments for scholarly and scientific applications. Scientific and scholarly software applications including Matlab, Mathematica, R, SAS, ArcGIS, Python [include additional cross-disciplines application examples]. Collaboration and sharing: Available tools, how they can be leveraged in collaborative research, and policies regarding security, confidentiality, open access and ownership of intellectual property. Procedural training: everything from how one accesses various components of cyberinfrastructure to the licensing or purchasing of software or devices. Such procedural training needs to be closely integrated with educational outreach and information dissemination about existing capabilities. Collaborative training with specialized groups including ICTS and Department of Statistics in statistical software, and Genomics High-Throughput Facility (GHTF) in genomics open-source software.

Examples of currently available workshops that would be advertised and scheduled are statistics workshops with ICTS, Big Data with Data Sciences Intitiative and BioLinux through GHTF. Online courses in HPC operations currently exemplify the online training mode. Possible outcomes for researchers, in addition to enhanced skills, could include certification as occurs for Data Science Initiative Courses or access to more specialized commercial software in the case of GHTF.

RCI resources must be made readily available for use in education, both for education on RCI tools and techniques, and as a platform for other subject matter. This includes providing access to compute clusters and RCI working environments for classroom use, and equipping instructional labs with visualization and other RCI capabilities.

9. Organizational Considerations

9.1. Vision9.2. Current9.3. Competitive Risk9.4. Recommendations

9.4.1. Cost

9.4.2. Delay

The Office of Information Technology currently supports RCI through its Research Computing Services group, with assistance from OIT data center and operations staff. The UCI Libraries provides RCI services through its new Digital Scholarship Services unit and other mechanisms. UCI Health provides support to researchers through secure access to clinical data and staff support. Some schools, notably Physical Sciences, have support team dedicated to RCI support. In addition, there are numerous staff involved in RCI as components of research units and other entities.

Any organizational model for RCI services must be flexible and able to leverage expertise across the university (and beyond). As there will be numerous funding and priority tradeoffs that will evolve over time, effective faculty governance is also imperative.

We propose to establish the UCI Research Cyberinfrastructure Center (RCIC) to manage and coordinate campus RCI. The RCIC would have a full-time staff director reporting to the Chief Information Officer with a âdotted lineâ to the Vice Chancellor for Research. The Office of Information Technology would house the RCIC administratively. A faculty panel, including representatives from schools, UCI Libraries, research units, and institutes, would provide oversight and prioritize investments . The RCIC would directly manage a subset of campus RCI and facilitate coordination and access to RCI resources within UCI, UC, and beyond.

UCI staff who support RCI, whether they are in a central RCI unit, other units such as the UCI Libraries, a school, research group, or research unit would be considered part of an extended UCI RCI support team. The RCIC Director would provide leadership and coordination to enable and leverage this approach. Team integration would be further strengthened through joint central/school assignments where appropriate.

It is critical that RCI staff be closely aligned with the faculty and research groups they support. Whereas it will not be possible to co-locate RCI staff with research groups in all cases, we should do so as much as possible. Where it is not practical, other mechanisms should be employed to make RCI staff function as collegial members of extended research teams.

10. Budgetary Requirements and Funding

10.1. Vision10.2. Current10.3. Competitive Risk10.4. Recommendations

10.4.1. Cost

10.4.2. Delay

OITâs current annual RCI budget is approximately $700k, covering the costs of 3.8 FTE staff, hardware maintenance, and other operational expenses. Additional funding will be required to establish and maintain baseline services that will be made available to faculty across disciplines. These services will form the foundation for UCIâs RCI; guaranteeing access to all faculty, facilitating collaboration and data safety and providing resources to that lead to funded projects.

Planning Budget Scenario More planning and prioritization review is required to flesh out an augmented RCI budget, but we present the following scenario to facilitate discussion. This represents an additional investment of approximately $2.3m yearly and would fund the following (all salaries are approximate and include benefits at 50%):

Research Cyberinfrastructure Director ($180k) Campus-Wide Storage System: 1 FTE storage programmer/administrator ($130k) Storage hardware ($200k) Compute Cluster Support: 2 FTE system administrators ($250k) Hardware refresh ($500k) Research Cyberinfrastructure Specialists: 2 FTE research computing staff ($250k) 1 FTE Data Curation Specialist ($130k) Scientific/scholarly software licensing ($100k) Research Information Security Compliance Engineer ($150k) Networking: UCI-Lightpath connectivity in additional campus buildings or UCInet enhancement in support of research ($400k)

10.5. Charging Models

In addition to providing core funding to build and maintain UCIâs Research Cyberinfrastructure

foundation, a critical goal for the first year of UCIâs RCI program is to develop recharge models to fully leverage grant funding. Fee structures would fund access to cluster computing cycles, storage, and staff services above baseline allocations.

Last updated 2016-02-13 13:36:42 PST