A Vision for Research Cyberinfrastructure at UCI

Draft 4.2, February 15, 2016; Mon Feb 15 12:22:44

Table of Contents

1. Vision & Executive Summary

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.

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.

Theory & experimentation have been supplemented over the last decade by modeling and data. Those four foundational pillars of science require a correspondingly robust Research CyberInfrastructure (RCI), implemented in a way that allows researchers to exploit it in a way that naturally fits with the way the scientist works. The goal of the RCI Workgroup is to provide recommendations that significantly advance and accelerate UCI research as widely and as economically as possible.

Because RCI impacts every aspect of research and scholarship, it must be addressed as a campus priority. Besides computation *per se*, it includes networking, storage, data curation & 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 as immediate support.

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 this under a separate organization, the UCI RCI Center (RCIC). 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 group 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 and 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. Expanded Description
- Initiate the upgrade and renewal of the UCI's compute clusters to bring UCI into parity with similar R1 institutions. Expanded Description
- 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.
- 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
- Provide increased support for research Data Management and Curation, not only because funding agencies demand it, but also to increase the re-use of data created at UCI and to use the resulting ease of access to encourage cross-domain collaborations among researchers at UCI and elsewhere. Expanded Description
- Hire staff to support the RCI and provide much more assistance to researchers to fully leverage the hardware, software, and services. 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, curation, 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 data security and access to tools for all researchers.

2. Putting Researchers in Charge - the RCI Center.

2.1. Vision

Researchers tend to know best when they need a novel tool or function, so our recommendation is to have *them* provide the guidance for how RCI is provided. A small, rotating committee of faculty who have a strong interest in how RCI is provided will consult with the larger campus research community and drive the direction of how to use the staff and budget of the RCIC to pursue better support. This should cut current feedback process of hierarchical complaint loops which have historically taken multiple years to result in any changes.

2.2. Current

The Office of Information Technology (OIT) currently supports RCI through its Research Computing Support 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.

2.3. Competitive Risk

The RCIC is proposed to speed decisions and implementations to support research computing. If it is not created and funded to appropriate levels, the risk is that support for RCI will proceed at the same unacceptable rate as it has before, with follow-on impacts on research, publication rates, recruitment, and retention.

2.4. Recommendations

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 dotted line 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.

2.4.1. Cost

See Budget.

2.4.2. Implementation

Should the RCIC be approved, it can start to operate quickly with an interim Director. Since it is a virtual organization, there are no requirements for new physical infrastructure, although it would be favorable in the future to bring all its staff together in a fairly close setting so that discussions and decisions could be made face to face.

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

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 discussed in the Technical Diagram Legend, 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 Campus Storage Pool (see Technical Diagram), 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.

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. This storage should minimally match the security requirements for HIPAA/FISMA and other federally mandated access.
- Accessible via a range of protocols; As an example, Purdue's Data Depot is available to Windows and Macs as a network drive on campus, and accessible by SCP/SFTP/Globus from anywhere.

† *CASS* could even act as a cross-campus Storage Pool but packet latency to it is poor and bandwidth to it, while decent for word-processing is unacceptably slow for HPC and *many-file* operations. CASS also does not allow the kind of file access used by Macs and Windows, nor does it allow shell access to set up other services.

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.

The following diagrams the various requirements of an academic storage system and shows how the

Campus Storage Pool would implement these requirements in software, which would mostly be run on the IO nodes that provide specific services. See also the Campus Storage Pool Technical Diagram.

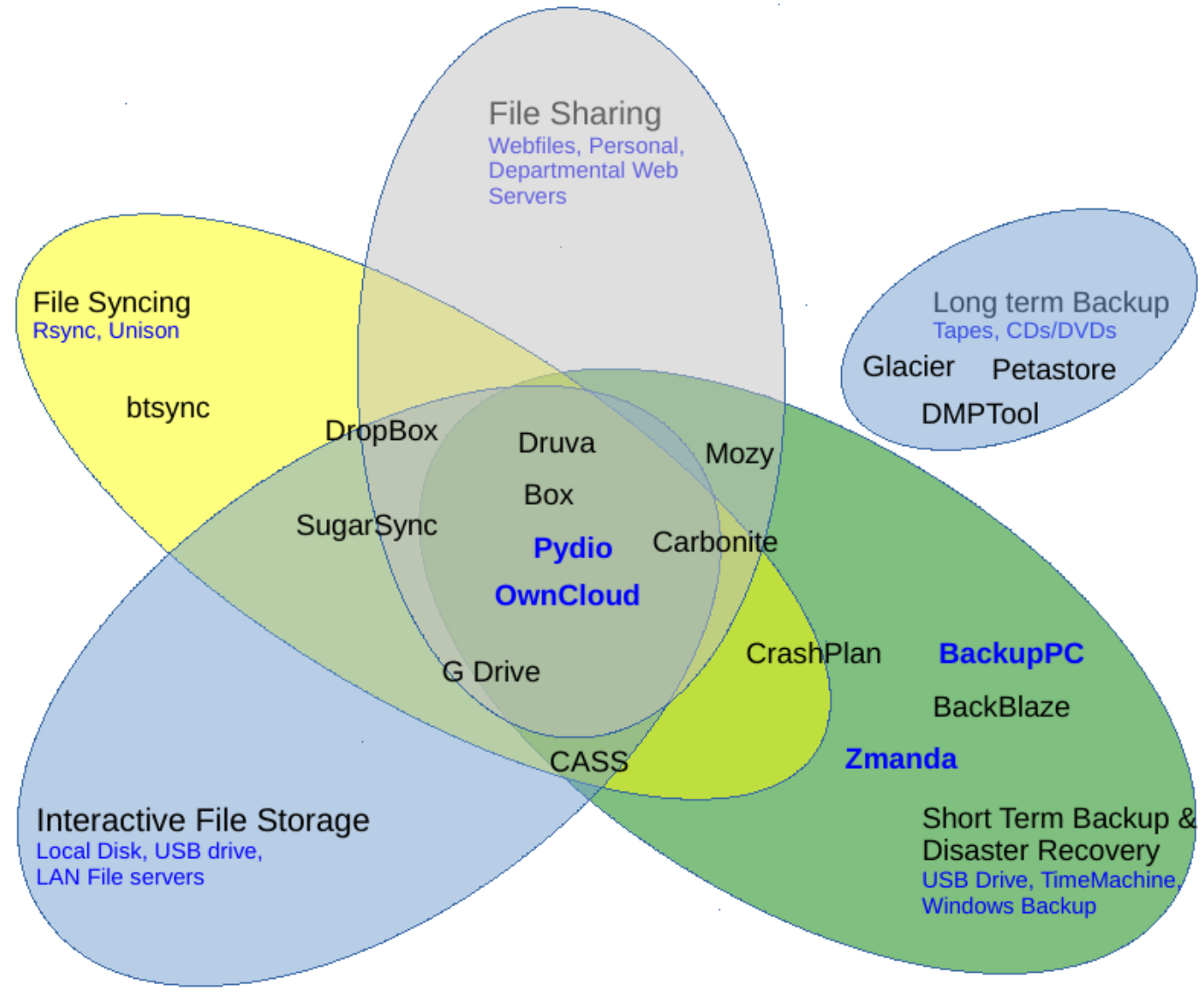

Figure 1.

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 for comparison. Bold blue names are Open Source services that could be implemented locally to replace the commercial service.

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 similar to the smaller system 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.

3.4.2. Implementation

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. Vision

In the same way that Researchers need web browsing or analytical tools such as Firefox or MATLAB, data management, archiving, and re-use requirements from funding and publishing organizations demand a similar set of tools and support. These tools not only decrease time-to-publication, but increase UCI's visibility re: data availability and competitiveness for new grants.

4.2. Current

UCI researchers are increasingly being asked to manage and share their research data in order to comply with funding agency requirements Ref 1 and system-wide Open Access policies Ref 2. Grant agencies direct researchers to document plans and demonstrate implementation for disseminating and preserving their work.

Example 1.

The UCI Library is working with ICTS to develop procedures to identify appropriate publications formats for deposit, verify citation information, and document required persistent link identifiers.

Example 2.

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 the increasing load of data and digital asset management is currently distributed, loosely coordinated, and not staffed to the level of peer institutions. Ref 6.

4.3. Competitive Risk

Failure to provide such implementation plans risks the loss of current funding and inability to win subsequent grants. For example, the NIH recently announced they will suspend funding for awardees who do not document deposit of papers into PubMed Central. Ref 3.

Professional staff are also needed to perform enabling activities to keep data usable when large discipline-specific repositories aren't available or suitable. Ref 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. Ref 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, and mapping.

4.4. Recommendations

The key missing part that addresses data curation and management is staffing. Staff are required to assist faculty directly, develop the tools and workflows required for data management, and to manage campus training programs on data management plans and open access distribution as funder mandates increase. In the area of providing data storage for the curation and management, there is considerable overlap with the Storage Section above, but much of the functionality for data curation is completely separate.

We recommend:

- 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.

4.4.1. Cost

Laura: Can you estimate a cost for each of the things you request above?

4.4.2. Implementation

Laura: Can you address how long you think it would take to

References

1 - A growing number of government and private funding bodies have Data Management 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. These plans are often part of the competitive process and lack of verifiable Data Management support will decrease the fundability of the applicant.

2 - UC Open Access Policies

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"

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.

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.

6 - UCI Digital Scholarship Services has 3FTE. Purdue has 9 FTE. U. Oregon has 10 FTE

5. Research Computation

5.1. Vision

As noted in the Executive Summary, computation is an integral aspect of modern science. and computation requires CPU cores, the more the better. Medicine and biology, domains that recently used little computation, are now the HPC cluster's largest consumers of CPU cycles, and other domains that previously had almost no large-scale computational requirements (Arts, Social Sciences) are using the massive social media databases to study trends and relationships in every conceivable arena. Analyses of

this scope, whether social & legal networks, molecular dynamics, healthcare, business intelligence, economics, or the increasingly creative arts is computationally intensive.

Faculty are very supportive of the idea of the *baseline* RCI, and with that computational requirement, CPU requirements increase that much more. In addition, the notion of a Virtual Research Computing Desktop - see below backed by the power of a large computational resource adds further requirements to the number of cores that must be available to provide reasonable response time.

Increasingly, as has already started in the web and commercial projects, we see more (semi-mobile) devices providing the interfaces, connecting to large central compute resources providing the analytic power required to run these increasingly large compute-bound jobs. These heterogeneous mobile devices will increasingly be bought by the end users but via standard protocols, be used to access web-based or native graphical user interfaces running on powerful multi-processor back ends in **secure** environments, allowing the analysis of restricted and/or proprietary data.

While centralized resources like a compute cluster are an attractive mechanism to be able to address multiple requirements for raw processing, there are a large number of needs that are not well-served by them. Efforts that require real-time processing, those that require specialized data pipes (as for multi-media editing), those that require dedicated hardware for 3D visualization, etc are not well-served by an over-emphasis on clusters.

The optimal research environment would provide enough resources so that compute jobs would not languish for days in wait queues, that computationally intense jobs could run to completion instead of having to be diced into smaller ones, and that the infrastructure is renewed regularly (with additional resources provided by Investigators as needed) to maintain the state of the art.

5.2. Current

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 hardware. HPC currently has a theoretical speed of about a 0.55 TeraFLOPS (TeraFLOPS = 1 Trillion Floating Point Operations / Second; a modern desktop CPU ~10 Billion FLOPS) and GreenPlanet is smaller. UCI presently faces a large shortfall of at least several **hundred** 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 the Conte Community Compute Cluster providing an aggregate 943 TeraFLOPS (includes 2 Phi accelerators per node). This is more than 1700 times the speed of HPC.

In terms of non-cluster resources, some of the problems reported in the humanities are addressable by improvements in the Research Data Storage section, but there are also specific needs that don't map well into compute clusters. These fall mostly into the areas of multimedia work, real-time processing of input data (a la the Internet of Things), and improvements in Networking for large-scale collaboration.

5.3. Competitive Risk

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 UCI's competitive position. In the past, local compute resources were used as testbeds for optimizing codes for very large analytical runs which were then moved to National Supercomputer Centers, but while this is still happening, the scale of all computation is increasing to the point where that model is also being overwhelmed.

This need was highlighted in the recent *Research CyberInfrastructure Vision Symposium* at which new faculty described their surprise at UCI's limited computational resources, especially the complete lack of **secure** computing facilities where HIPAA/FISMA and other forms of restricted data can be analyzed. In fact, many are continuing to rely on resources at their previous institutions (University of Washington, University of North Carolina, UC Berkeley) through professional contacts. This is not only detrimental to our reputation, but creates potential security risks, and is closely related to the above stanza on Research Data Storage.

The obvious risk in allowing UCI's computational resources to wither is that our computational scientists will no longer be competitive nationally, that incoming grant dollars will have to be shared to those institutions that have the facilities, and that research involving restricted data simply cannot be performed easily at UCI.

5.4. Recommendations

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 UCI's success. Computational facilities should be considered a comparable and essential aspect of UCI's 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.

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 whole scale 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.

5.4.1. Cost

See Budget.

5.4.2. Implementation

Additions to clusters can be completed on the order of 2 months. Providing secure compute facilities will require a careful review of the current Data Center and advice from UCI's security team as to how to provide more security without making access so difficult as to make it unusable.

6. RCI Working Environment

6.1. Vision

Researchers use (and feel quite strongly) about the interfaces to their computing devices. RCI should strive to provide as much functionality as possible, as unobtrusively as possible. However, there are areas where the native interface either does not exist or can't be scaled economically and in those cases, the RCI environment should be driven by long-term functionality, with instruction to bring researchers to effective use of them.

Besides lacking the underlying hardware to accomplish a task, the other major impediment to RCI is the software. Many researchers, for reasons of history, dependence on certain libraries, familiarity with interfaces, or functionality will require access to proprietary software. When this is the case, UCI should strive to provide that software at the lowest price via bulk or network licensing. Where the demands of the work require proprietary tools, and they cannot be economically licensed, it is not unfair to require the small number of beneficiaries to contribute to the cost.

Working environments also include 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. Again, a secure Campus Storage Pool goes a long way in providing the infrastructure to do this.

In many other situations, Open Source Software (OSS) should be strongly considered, and where appropriate, promoted. This makes services much more scalable, both via lack of legal exposure as well as reduction of human support for the licensing. While most OSS is available for Linux, it exists in almost equitable amounts for Mac OSX and in a surprising amount for Windows. This is not only because it is free, but because analytical software tends to appear first on Linux as OSS and (sometimes years) later is wrapped into proprietary form for Mac OSX and Windows, so the ability to use this software in its commandline form on Linux confers a months-to-years time advantage, as well as the financial one.

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 RCI is facilitating collaboration from interpersonal interactions, to data sharing, to information dissemination.

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

6.2. The Research Desktop Computing Environment

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 *virtual desktop* environments using Remote Desktop protocols.

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, Mac OSX 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.

The implementation mechanism is as simple as placing an icon on the Desktop of a personal computer regardless of OS. Activating that icon starts a Remote Desktop application that presents another Desktop as another application window. In that application window, all the research applications required would be presented as further icons or in the familiar nested menu system. The applications started on that Desktop would execute on the CPUs on the cluster and would have access to both interactive and batch sessions. These Desktops are long-lived - they can be closed and then re-activated at another location, doing exactly what was being done previously. The Research Desktops have access to the same or similar facilities and collaboration tools as the native desktop.

6.3. Current

UCI does host significant components of a robust RCI: the campus wireless and wired network with good connectivity to CENIC and Internet2, the LightPath high-speed science network, systems (including compute clusters) housed in the OIT Data Center, staff in OIT and in the UCI Libraries.

In addition to network connectivity, providing access to off-campus resources includes addressing contractual agreements, security measures, and access permission (authentication and authorization). UCI's participation in Internet2's 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 InCommon's *Research & Scholarship* service providers (e.g., the GENI Experimenter Portal, the Gravitational-wave Candidate Event Database).

However, in terms of addressing RCI requirements, UCI has left this largely in the hands of the individual Schools. Some schools (Physical Sciences) have internal staff to address and assist with research-related problems, but many have only *Computer Support Coordinators* to address Desktop-level and *MS Office*-level issues.

One of the persistent problems faced by faculty in the latter group is that they don't know who to ask for advice with their research computing problems. While the Research Computing Support group has been active for almost a decade, this is still an issue that needs to be addressed.

The Data Sciences Initiative, via their Short Courses outreach program has been a significant driver to educate the UCI research community in the use of various techniques and especially in the use of OSS for data analysis. These courses are often over-subscribed and the feedback is extremely positive.

6.4. Competitive Risk

Like other resources, if the applications and especially the instruction in those applications is not supported, UCI researchers will not have (or will not be able to use) some of the very powerful tools that they can exploit. One way of looking at this is also in the preparation of future scientists who can be taught how to use these powerful OSS tools and therefore be freed for the rest of their lives from licensing costs, or spend those funds on licensing, which locks the students into a proprietary system which will require funding every license period.

Similarly, since we cannot ignore some of the very powerful proprietary tools, we can make access to them and to the OSS ones as simple as possible by bundling the interface into the virtual desktops.

6.5. Recommendations

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.

6.5.1. Cost

The software cost for implementing the Linux version is zero since it is OSS. The hardware costs are fairly low since we can run much of this software on older servers that we have in excess. Licensing costs for the proprietary software are dependent on whatever deals can be made with the vendors. However, the effect of providing better and better OSS tools and making them more easily available to students will result in more pressure on proprietary vendors to decrease prices, as we have seen with the cost of OSs going to near zero.

Windows- based remote desktops will cost more, thought how much more depends on the agreements with Microsoft as well as the actual software vendors.

The main cost of implementation is the human cost of setting it up and documenting it to the level that it becomes easy to use. After that, the human costs are for setting up the software on the back-end cluster which have to be done anyway.

6.5.2. Implementation

There will be some delay in testing and selecting the best Desktop software to use in, but since this project is actually ongoing in RCS, we can move forward fairly quickly and test versions of it could be ready for use from the HPC cluster within 2 months with widespread availability in 6 months, based on availability of login servers.

7. RCI Staffing

7.1. Vision

People are the part of RCI that enable it to be effective and *Research Computing Support* is different than generic *Computer Support* in that not only are they asked to debug strictly computational problems, but the problems requires an in-depth knowledge of how CPUs, file systems an formats, schedulers, the Linux OS, the various cache levels, utilities, networks, compilers, libraries,provisioning systems, data-flow, and of course the applications all work… and/or don't. Added to this are the vagaries and context of the backing science. Add the requirement for being able to clearly document complex procedures, and the necessity of teaching these techniques, and you may understand why well-qualified people are hard to find.

Much RCI support is catalytic - a little knowledge can make an insurmountable problem disappear, but a good part of it is ongoing and quite demanding, such as programming support, investigating new techniques for improving and maintaining RCI, and especially *answering questions from researchers*.

Optimally, there would be enough RCI staff to maintain the RCI as well as engage much more with researchers, which is currently not possible beyond directly answering their most immediate questions. Domain specialists in the areas of highest RCI use such as bioinformatics, engineering, physics, with particular training in the most popular applications such as MATLAB, R, PyLab, etc would be an enormous help to computational researchers, the numbers of whom are increasing steadily (HPC alone has almost 2000 registered users of whom about 500 use the cluster every month).

7.2. Current

UCI's current level of RCI staffing is substantially below that of comparable peer institutions (UCB, UCLA, Purdue), and others based on a variety of metrics. We have approximately 3 FTEs 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 cluster operation, research data storage and transfer, application and programming, the Data Center Information System, and Graphical Information Systems.

Additionally, there are 3 FTEs 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 FTEs 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 institutional repositories and exhibits, particularly with humanities applications; and negotiating software licensing from external sources.

7.3. Competitive Risk

There are 2 vectors of risk. The most obvious is that with so few RCI support people, it is difficult for researchers to obtain more than glancing assistance from any of them. The second is that the actual infrastructure is at risk since there are no spare cycles to do proactive work on the actual RCI. This shows that providing more hardware or proprietary software without matching FTEs will result in little improvement in usable RCI.

7.4. Recommendations

We recommend adding staff to assist with:

- sysadmin support for cluster, filesystem, and backup maintenance and upgrades.
- installation and upgrades to existing software (>1000 packages and versions now on HPC)
- installation and networking of workstations for advanced instrumentation;
- user training to introduce users
    - to the Linux OS, cluster computing, optimal data handling techniques
    - to bash, Perl, Python, Jupyter, R, & MATLAB
    - to Open Source visualization tools
- installation of, documentation about, and training with other Open Source tools
- answering researcher questions about techniques, errors, obtaining and using data sources, etc.
- assisting investigators with establishing appropriate levels of security to be in compliance with funding agency requirements.

Beyond these general necessities, we strong advocate for adding the following *Discipline-Oriented Specialists*

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.

RCI Specialists: 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.

Outreach Specialist / Concierge: The increasing complexity of RCI requires systematic outreach so the full potential of the RCI investment is realized. An RCI staff specialist will be tasked with coordination of outreach to the campus community regarding RICI RCIC facilities and resources and will be the main point of contact for anyone wanting guidance on all aspects of RCI. As such, this person must be familiar with most aspects of RCI on campus and know the principals well. This person or another person familiar with 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 Director would facilitate this approach, which would be further strengthened through joint central/school assignments where appropriate.

7.4.1. Cost

See Budget

7.4.2. Implementation

Depending on the administrative level of this unit, it could be in place fairly rapidly or be delayed for a decade.

HJM I have no experience to estimate this at all

8. Education and Training

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.

8.1. Vision

Incoming students, undergrad as well as graduate, have execreble computer skills if any. At most, they have been trained on *MS Office* which is close to useless for modern data analysis and visualization. Since the university cannot ban them for not learning about computers, we must teach them, and that requires both instructors, time, and classrooms. The classrooms are fairly easy to find; instructors with time are not.

This is where the RCIC staff can perform what can honestly be termed a transformational service; to transform computer-naive students into those who have a chance of dealing with modern data. Such training will require learning new applications on their personal devices, but mostly learning Linux and how to use it for data analysis.

What we aspire to is the graduation of a group of data-savvy students; they know about modern data formats, cleansing, regular expressions, transformation, effects of caching, IO problems, parallel operations (if not programming), how to move data effectively, encryption, compression, checksums, and of course, analysis, statistics, and visualization of large data sets.

RCI training will include:

- the Linux OS, basic bash commands, utilities, bash programming, internet tools, etc (Software Carpentry courses provide a framework). Also, leveraging Macintosh and Windows environments for scholarly and scientific applications using the MacOSX terminal and the Cygwin environment on Windows.
- Cluster computing; use of the scheduler, debugging, batch scripts, filesystems, data movement
- data analysis, and visualization techniques, basic intepreted languages such as Perl, Python, Julia
- installing, compiling, debugging Open Source Software.
- Data asset management
- Scientific and scholarly software applications including Matlab, Mathematica, R, SAS, ArcGIS, Perl, Python
- 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: 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.
- Using cloud services for academic analysis.

- Collaborative training with specialized groups, for example using:
  - statistical software with ICTS / Department of Statistics
  - genomics open-source software with the Genomics High-Throughput Facility

8.2. Current

Whether this proficiency is demanded as a requirement of a program or just made available on an *ad hoc* basis, these introductory courses are critical to launch students on the right path to competency. Generally, with a few such short-courses under their belts, students can Google their way to proficiency and even expertise. For those students entering a numerically based program, it is a good idea to make such courses compulsory, since without them, later courses will be a nightmare of catch-up and backtracking.

There are some good introductory and advanced classes being taught already outside of official channels, mostly by student instructors and since faculty have little incentive to contribute to this effort, the natural alternative is an organization like the RCIC. RCIC staff can teach the introductory classes and act as dedicated sysadmins adjuncts and TAs to other courses with substantial computational depth.

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. We acknowledge this and 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 all of these modes of training but would ensure that the RCIC website lists the available modules and has links to sign-ups and schedules.

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.

8.3. Competitive Risk

The glaringly apparent risk is that our students are not able to do research using the most basic of modern numerical tools. Excel is a useful tool but not for assembling genomes. Nor is it capable of doing social network analysis on Facebook logs. For those kinds of approaches, we need students with experience on Linux using modern stream-processing tools.

8.4. Recommendations

RCI resources must be made available for teaching, both for education on RCI tools/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.

8.4.1. Cost

Overwhelmingly, the cost here is human. The current RCS personnel are teaching multiple courses, but this is essentially a labor of love and is not scalable. This works in the reverse direction as well; with campus RCI being used for instruction, instructional monies can be used to support RCI. See Budget

8.4.2. Implementation

Since a few of the courses are already being taught in cooperation with the Data Science Initiative, the spin-up time will be short for those courses, but new courses are extremely labor intensive and will probably take months for new hires to develop or learn them.

9. Budgetary Requirements and Funding

9.1. Vision

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%):

9.2. Current

OIT's 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 UCI's RCI; guaranteeing access to all faculty, facilitating collaboration and data safety and providing resources to that lead to funded projects.

9.3. Competitive Risk9.4. Recommendations

In addition to providing core funding to build and maintain UCI's RCI foundation, a critical goal for the first year of UCI's 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.

9.4.1. Cost

```
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)
                                                 Total
($2,290k)
```

9.4.2. Implementation

??