A Vision For
Research CyberInfrastructure at UCI

Version 4.7e - 3/14/2016

Executive Summary

UCI has put forward a visionary strategic plan that recognizes the critical role that cyberinfrastructure plays in achieving campus goals. Research Cyberinfrastructure (RCI) at UCI is currently below that of other R1 universities. Substantial new investment in storage, computation, and staffing is needed to sustain and accelerate research productivity in support of the strategic plan.

Existing RCI staff and facilities have provided quality service to campus researchers since 2013 when the recommendations from the Faculty Assessment of the State of Research Computing (FASRC) were issued. The Office of Information Technology (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 dedicated to the exchange of 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 two largest compute clusters (HPC and GreenPlanet), although underfunded and aging, have been used to produce many publications in multiple domains.

In spite of these successes, RCI remains a distinct weakness at UCI, to the extent that some researchers rely on services at their previous institutions. Specialized computational and storage resources especially are underfunded at UCI, with some facilities such as HIPAA/FISMA-secure research computing facilities not readily available on campus.

Theory and experimentation have been supplemented over the last decade by modeling and data. Those four foundational pillars of science require correspondingly robust RCI, implemented in a way that is fully integrated into the research workflow. The goal of the RCI Vision Workgroup is to provide recommendations that significantly advance and accelerate UCI research as broadly and 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 and management, and support services required by all disciplines. These requirements continue to expand and 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 distinct new organization, the UCI RCI Center (RCIC). OIT's current Research Computing Support group will form the core of the new RCIC. A faculty panel will be established to provide overall direction, set service priorities, serve as co-PIs on grant applications, and oversee an annual process to assess service effectiveness. The RCIC will coordinate with related units on campus that need RCI and support for research and instruction. Center and shared equipment grants will also be coordinated through collaboration among RCIC and other units. OIT will support the RCIC administratively and provide supplementary IT services. [Page 3]

- Hire staff to support research IT services and provide assistance to researchers to fully leverage the hardware, software, and services. None of the projects under discussion will advance without staff

expertise, which is a critical aspect of RCI. Career staff will maintain, upgrade, and expand services; train students, staff, and faculty in current computational techniques; develop standard operating procedures, provide programming, assist with grant preparation, work with staff in other units to support advanced computing needs, and assist in maintaining compliance with federal requirements for data management. [Page 3]

- Initiate construction of a scalable Petabyte storage system that can be accessed by all researchers and leveraged to provide the multiple types of storage and data sharing that assist research endeavors. This includes centralized, active file storage and backup, ready sharing of large data sets, secure web distribution of data, file syncing, and tiered data archiving locally and in cloud archives. [Page 5]

- 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 and to use the resulting ease of access to encourage cross-domain collaborations among researchers at UCI and elsewhere. [Page 6]

- Initiate the upgrade and renewal of the UCI's compute clusters to bring UCI into parity with similar R1 institutions. [Page 7]

- Provide a baseline level of storage, connectivity, and computing for faculty in all disciplines. At least 1 Terabyte of robust, secure, central storage, 1 Gigabit/second network connectivity, and 100,000 compute hours annually will be made available to each faculty member upon request. These allocations will increase over time and be augmented by research grant funding.

- Establish a widely available and scalable Research Desktop Computing Environment (RDCE) to facilitate computational and data science research and teaching. This environment will include access to shared software, high performance computing resources, visualization tools, tools for data sharing and collaboration, assisted access to external UC and national facilities, and appropriate cloud resources. The RDCE will be more secure than traditional desktop computing, and additionally secure computational and storage environments will 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). [Page 9]

Executing this vision will accelerate development of research programs, with increased productivity, reducing redundancy among units, increasing the overall level of support expertise, and increasing RCI tools and security for all researchers.

Putting Researchers in Charge - the RCI Center.

Vision

Researchers know best when they need a particular tool or function, so it is essential that they guide how RCI is provided. A small, rotating committee of faculty with a strong interest in RCI will consult with the larger campus research community and set direction for RCIC staffing and use of funds to optimize service delivery. RCI services will thereby be appropriately applied, and expertise leveraged across the university and beyond.

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 teams dedicated to RCI support. In addition, there are numerous staff involved in RCI as parts of research units and other entities.

Recommendations

We propose to establish the UCI Research CyberInfrastructure Center (RCIC) to manage and coordinate campus RCI. The RCIC will have a full-time Staff Director reporting to the Chief Information Officer with a dotted line report to the Vice Chancellor for Research. OIT will house the RCIC administratively. A faculty panel including representatives from schools, UCI Libraries, research units, and institutes, will provide oversight and prioritize investments. The panel will oversee an annual process to assess service effectiveness. The RCIC will directly manage a large subset of campus RCI and facilitate access to RCI resources within UCI, UC, and beyond through contacts at the San Diego Supercomputer Center, the NSF-funded XSEDE organization, and others. The RCIC Director will facilitate a flexible and inclusive approach to RCI staffing and expertise, as explained in the next section.

Competitive Risk

The RCIC is proposed to facilitate decision-making and implementations to support research computing. If it is not created and funded to appropriate levels, the risk is that RCI support will proceed at an unacceptable level, with collateral impacts on research, publication rates, recruitment, and retention.

RCI Staffing

Vision

People are the part of RCI that enable it to be effective. Research Computing Support requires in-depth knowledge of CPUs, file systems and formats, schedulers, the Linux operation system, networks, compilers, libraries, provisioning systems, data-flow, and applications. Often staff must know the domain context and be competent to document domain-specific procedures.

Sufficient staffing is needed to maintain services as well as engage with researchers beyond the current level of simply answering their most immediate questions. Domain specialists in the areas of highest RCI use (e.g. bioinformatics, engineering, physics) with training in the most popular applications (e.g. MATLAB, R, PyLab) would be an enormous help to computational researchers, the numbers of whom are increasing steadily. The HPC cluster alone has almost 2,000 registered users of whom about 500 use the cluster every month.

Current

OIT and Physical Sciences have approximately three FTEs assigned to supporting the GreenPlanet and HPC clusters and a similar number for supporting researchers in other ways including research data storage and transfer, application support, and programming. This is substantially below that of comparable peer institutions (UCB, UCLA, Purdue), and others based on a variety of metrics. Purdue and Indiana University have 20 or more staff assigned to these tasks. UCLA has a RCI team of 21 FTEs, 11 of which focus on compute clusters and other high performance services. UCB has 15 FTEs on their Research IT team. *(See Appendix C)*

In the specialized area of Digital Scholarship Services there are three FTEs to support RCI associated with the UCI Libraries, compared to an exemplar like Purdue which has a dedicated Data Curation Center and seven FTEs supporting data curation work. UCI's three FTEs are insufficient to cover the areas for which they are responsible which include: promoting researcher compliance with funder data management requirements; providing training in best practices for data curation; promoting Open Access of UCI data and scholarly works; developing access and retrieval systems; building robust institutional repositories and online exhibits, particularly with humanities applications; and negotiating licenses for software and externally sourced data.

Recommendations

We recommend increasing RCI 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. Specifically, we need to add system administration staff to assist with:

- **Cluster support**: filesystems, backup, maintenance, and upgrades.
- **Installation and upgrades:** to existing software packages (over 1,000 now on HPC)
- **User training** in:
    - the Linux operating system, cluster computing, optimal data handling techniques
    - Bash, Perl, Python, Jupyter, R, and MATLAB, and best practices
    - BigData-compatible analytic and visualization tools
- **Data security:** assisting investigators with establishing appropriate levels of security to be in compliance with funding agency requirements.

Beyond the support required to maintain research computing services, the following *Discipline-Oriented Specialists* are also required:

- **Domain Specialists:** Full RCI support requires domain specialists in math, chemistry, biology, social sciences, engineering, humanities and the arts, with complementary technical skills, working in partnership with research teams. These staff positions, could be jointly sponsored by the RCIC and the Schools and include part-time graduate or even undergraduate appointees.
- **Library Sciences Specialist:** The library identified a need for an additional data science librarian who will have expertise in funding agency requirements for data preservation, project management, functional requirements specification and application development, metadata, and digital preservation. This position will coordinate activities with OIT and Office of Research staff supporting data management and liaise with school staff.
- **RCI Outreach and 'Concierge':** The increasing complexity of RCI requires systematic outreach so that the full potential of RCI investment is realized. The RCIC team will provide outreach to the campus community regarding available services. The RCI Director will be ultimately responsible for assuring guidance on *any and all aspects of RCI*. As such, this person must be familiar with most aspects of RCI on campus and know the principals well. The RCIC will also develop a training program for incoming faculty, students and staff to ensure baseline awareness of resources and best practices, and maintain a website with information about RCI services and contacts.

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, will be coordinated through an extended UCI RCI support team. Staff will continue to be directed by the entities through which they are funded, but participate in expertise sharing where it is mutually beneficial to project and campus goals. The RCIC Director will facilitate this approach, which will 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.

Competitive Risk

There are two risks. The most obvious is that with the current minimal RCI staff, researchers receive only a minimal level of support, resulting in ineffective use of resources and impacts to research productivity. The second is that understaffing results in insufficient maintenance, and thus, even endangers existing infrastructure. In short, investments in RCI without matching FTEs would limit their positive impact.

Research Data Storage

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 and archive, forward, edit, analyze, and visualize multi-terabyte datasets should be available to all faculty. A requirement that underlies all those aims is the secure and reliable physical storage required to contain that data.

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 and not conducive to collaboration; it exposes UCI to liability and is 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 Campus Storage Pool Technical Diagram *(Appendix D, Figure 2)*. Universal access to storage for recording, writing, analysis, archiving and sharing are the *de facto* **papyrus** of this age. The on-campus availability of 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, intellectual property, or other legal prohibition. Such storage systems 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 this. 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 rates, as with relational databases. These characteristics require a local **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.

Recommendations

We recommend the immediate implementation of a baseline Campus Storage Pool and matching backup system. 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.
- Low latency, high-bandwidth access to Compute Clusters and other analytical engines.
- Important data backed-up and mirrored to multiple locations (including off-campus).

- Physical security 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.
- Accessibility 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 secure protocols (such as SCP/SFTP/Globus) from anywhere.

The Campus Storage Pool will be available to all faculty as a baseline, no-cost service. Additional storage needs will be funded through a cost recharge model. Figure 1 in Appendix D represents the various requirements of an academic storage system and shows how the Campus Storage Pool will implement these requirements in software, which will mostly be run on the IO nodes that provide specific services. See also the Campus Storage Pool Technical Diagram (*Appendix D, Figure 2*).

Competitive Risk

There are three risks of not implementing this. First is the risk of not providing what is increasingly considered to be a universal requirement of basic research infrastructure. This results in non-competitive grant applications and inability to compete for attractive hires. The second is the financial risk of losing data that does not have backup. The third risk is the fiduciary risk of not protecting data that must be shielded for intellectual property, legal, or security reasons.

Curation

Vision

Large amounts of data demand new ways of thinking about management. Federal and other agency guidelines are requiring strict compliance with good practices. Adoption of support for current data management tools and practices will enable sharing, increase security, decrease time-to-publication, and increase UCI's competitiveness for funding and visibility in the public and commercial sectors.

Current

UCI researchers are increasingly asked to manage and share their research data to comply with funding agency requirements such as those of NSF, NIH, and UC Open Access Policies. Granting agencies direct researchers to document plans and demonstrate implementation for disseminating and preserving their work. The UCI Libraries are working with the Institute for Clinical and Translational Science to develop procedures to identify appropriate publication formats for deposit, citation verification, and document required persistent link identifiers. 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 crowd-sourced 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. For example while UCI Digital Scholarship Services has three FTEs, Purdue has nine, and the University of Oregon has ten.

Recommendations

The key missing element for data curation and management is staffing. Staff are required to assist faculty directly, develop the tools and workflows required for data management, and to conduct training programs on data management plans and open access distribution as funder mandates increase. Data storage as described above is required for curation and management, 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 programs, 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.

Competitive Risk

Failure to enhance data management services 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 (see NIH public access policy).

Insufficient staff for maintenance of repositories endangers the longevity of important projects, especially in the Arts and Humanities. The Libraries share administration of a number of digital repositories and metadata services which preserve content and make it discoverable, such as Calisphere, DASH data sharing, eScholarship, EZID, Merritt (hosted centrally by UCOP/CDL). UCI fully administers the UCISpace repository. However, these resources are insufficient to sustain service in the face of greatly increased development of these types of repositories.

Research Computation

Vision

Computation is an integral aspect of modern science and requires CPU cores, the more the better. Medicine and biology, domains that until recently used little computation, are now the HPC cluster's largest consumers of CPU cycles. Other domains that previously had almost no large-scale computational requirements (Arts, Social Sciences) are using massive social media databases to study trends and relationships in every conceivable arena. Analyses of this scope, whether social and legal networks, molecular dynamics, healthcare, business intelligence, economics, or the increasingly creative arts, is computationally intensive.

Given the evolving role that computation is playing, it is important to provide a baseline amount of access across all disciplines. 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, we see mobile devices providing interfaces, connecting to large central compute resources with the power required to carry out the complex analysis. These heterogeneous mobile devices will increasingly 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 multimedia editing), those that require dedicated hardware for 3D visualization, and the like, are not well-served by compute clusters.

A competitive research environment requires enough resources such that compute jobs do not languish for days in wait queues, that computationally intense jobs can 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.

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 do not 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.

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-Throughput 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 recommended. Annual funding must also be identified 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. It is essential that sufficient cluster capacity (thousands of cores) be made available for training students of all disciplines in the use of parallel computing.

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

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. 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 *(Appendix B*). 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.

RCI Working Environment

Vision

RCI should strive to provide as much functionality as possible, as unobtrusively as possible, using interfaces with which researchers are comfortable. However, there are areas where the native interface either does not exist or cannot be scaled economically; in those cases, the RCI environment should be driven by long-term functionality, with instruction provided to allow effective use by researchers. Transparent functionality is driven by appropriate integration of hardware and software tools. This will be facilitated in the RCI project. Software tools facilitate individualizing the environment. Visualization software and facilities are additional key requirements in RCI working environments.

Many researchers require access to proprietary software. The RCIC will 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, the small number of beneficiaries would effectively share 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 central RCI 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 will help provide the infrastructure to do this.

In a wide variety of situations, Open Source Software (OSS) should be actively encouraged. OSS makes services both economical and scalable. It reduces legal exposure and minimizes human support needs for licensing. While most OSS is available for Linux, it exists in the Macintosh and Windows environments as well. Analytical software tends to appear first on Linux as OSS and (sometimes years) later is wrapped into proprietary form for Macintosh and Windows, so the ability to use this software in its command line form on Linux confers an additional months-to-years time advantage.

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, the RCI project will collaboration from interpersonal interactions, to data sharing, to information dissemination.

The Research Desktop Computing Environment

Maintaining research computing environments with the requisite software requires significant administration and upkeep. One method that has proven effective 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, Macintosh 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. Activating that icon starts a Remote Desktop application that presents another Desktop as an application window. In that window, all the research applications required will be presented as further icons or in the familiar nested menu system. The applications started on that Desktop will execute on the CPUs on the cluster and will 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.

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 control. UCI's participation in Internet2's identity management confederation (InCommon) allows UCInetID credentials to be used to access external resources.

UCI has left the provision of RCI working environments largely in the hands of the individual researchers and schools. Some schools (Physical Sciences) have internal staff to support research computing, but most have only have IT staff to address routine desktop support. Research software is often funded via a “pass the hat” approach and is not consistently available across campus.

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.

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, secured environment.

Competitive Risk

Like other resources, if applications and the instruction in those applications is not supported, UCI researchers will not have access to 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 and from being locked into a proprietary systems.

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 and economical as possible by bundling the interface into the virtual desktops.

Education and Training

Vision

The need for computational and data analysis skills is increasing rapidly and impacting almost every research discipline. This is where the RCIC staff can perform what can honestly be called a *transformational service.* UCI will become a leader in *data- and compute-savvy graduates,* with RCI training as outlined in *Recommendations* below.

Current

The UCI Statistics and Computer Sciences departments are very strong, but no training strategy is currently in place for incoming students in traditionally non-computational areas or for postdocs, faculty and others who have limited time for formal classes. Incoming students, graduate as well as undergrad, generally have minimal computational and data analysis skills. The university must address this deficit, which requires instructors, time, and classrooms. There are some good introductory and advanced classes being taught already outside of official channels, mostly by student instructors and staff. Faculty are currently not incentivized and in fact, are actively discouraged from taking on roles in interdisciplinary (non-major) training courses. The natural alternative to address this deficiency is an organization like the RCIC. RCIC staff can teach the introductory classes and provide teaching assistance to other courses with substantial computational depth. Examples of the currently available workshops are **Statistics** workshops with *ICTS*, **Introduction to Linux**, and **Big Data** with the *Data Sciences Initiative* and **BioInformatics** through *GHTF*.

Recommendations

RCI resources must be made available for teaching, both for education on RCI tools/techniques, and as an instructional 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. Training topics should include the following:

- Linux OS, basic bash commands, utilities, bash programming, internet tools
- Cluster computing; use of the scheduler, debugging, batch scripts, filesystems, data movement
- Data handling, including modern data formats, regular expressions, caching effects , I/O bottlenecks, parallel operations, how to move data effectively, encryption, compression, and checksums
- Data analysis, and visualization techniques, using interpreted languages such as Perl, Python, R, and Julia, as well as proprietary applications such as Matlab, Mathematica, SAS, ArcGIS
- Finding, Installing, compiling, debugging Open Source Software
- Data asset management
- Tools for collaboration and sharing: how they can be leveraged in collaborative research, and policies regarding security, confidentiality, open access and ownership of intellectual property
- 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

Generally, these classes can be taught in small, on-demand groups or in high-intensity day- or week-long sessions, with course content available at all times online for reference or for those whose schedules are not open to dedicate to available class schedules.

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. Furthermore, new standards of robustness in research require competency in data management and statistics. Failure to train our students puts them and us at risk of research non-compliance. Graduates that do not have minimal computational skills will be unemployable in the future marketplace.

Budgetary Requirements and Funding

Vision

More planning and prioritization review is required to flesh out an augmented RCI budget, but we present a four year plan to guide discussions in Appendix E. We propose that a new annual campus allocation of $2.1M be provided to establish the Research Cyberinfrastructure Center. This allocation will be augmented by OIT, grant, and recharge funding totalling $950k. Grant and recharge funding will increase over subsequent years to allow further RCI evolution and growth.

These combined funding sources will allow us to do the following:

1. Implement a campus storage system with backup and appropriate levels of security;
2. Augment current computer cluster system administration, libraries data curation, and RCI specialist staffing;
3. Hire a full-time director for the Research Cyberinfrastructure Center; and
4. Provide an annual funding source to refresh a subset of cluster hardware, provide additional software licensing, and enhance research networking.

Addressing these priorities will enhance current RCI and lead to establishing baseline services for 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. Metrics will be collected to inform an annual process overseen by the RCIC faculty panel to assess service delivery and identify changes required to ensure overall RCI effectiveness.

Current

OIT's current annual RCI budget is approximately $750k, covering the costs of 3.75 FTE staff, hardware maintenance, and other operational expenses. Funding in the Libraries and schools supports additional RCI services.

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 will fund access to cluster computing cycles, storage, and staff services above baseline allocations and provide the basis of future RCI growth.

Appendix A

UCI Research Cyberinfrastructure Vision Workgroup

In September 2015, Provost Enrique Lavernia charged the UCI Research Cyberinfrastructure Vision Workgroup as follows:

> *Simulation and modeling are now the "third pillar of science" after theory and experimentation, with "data" becoming a fourth pillar. Computational and data-enabled science is key to modern research and scholarship.*
>
> *The goal of the workgroup is to build on the 2013 Faculty Assessment of the State of Research Computing effort to develop a "Research Cyberinfrastructure Vision" that significantly advances UCI research and scholarship capabilities.*

Workgroup Participants:

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

** Co-Chairs

Appendix B

UCI RCI Symposium - January 27, 2016

The RCI Vision Workgroup and the UCI Data Science Initiative co-sponsored a one-day symposium on the state of Research Cyberinfrastructure. The event was held at the CalIT2 Auditorium where approximately 120 UCI faculty, students, and staff heard presentations by faculty across disciplines regarding the critical role RCI plays in research (see below). Feedback from the symposium has contributed to the creation and validation of the UCI RCI Vision Document.

The goal of the symposium was to hear from faculty on current RCI application and problems they encounter as well as their thoughts about future plans. UCLA and UC Berkeley RCI Service Directors were also on hand to present about RCI at those campuses. A number of UCI staff also presented about current campus RCI services.

Given that RCI affects all areas of campus and many faculty were willing to present and attend the conference, faculty presentations and subsequent discussions were divided into four sections focusing on schools that make similar use of RCI:

- Section 1 – Biological Science and Medicine;
- Section 2 – Humanities and Arts;
- Section 3 – Social Science, Social Ecology, Law, Business, and Education; and
- Section 4 – Engineering, Physical Sciences, and Information and Computer Science.

Each section started with six to seven presentations followed up by a discussion panel led by two section faculty. The section leads invited faculty to present that were representative of other faculty in their section. At the end of the day the faculty leads discussed what they had heard and wished to share from their sections.

To assist with the Vision Workgroup effort and to provide a symposium proceedings document, a video recording of the event was made by UCI Media Services. Presentation materials that were displayed at the symposium will also be made into a compendium that will be available through the Workroup website (http://sites.uci.edu/rci). This appendix provides a brief overview of the comments that were shared and discussed by symposium attendees.

Symposium Takeaways

RCI is critical to the research efforts of a large and diverse set of UCI faculty. All speakers outlined how their research makes critical use of several different areas of RCI. RCI is not limited to high performance computing efforts in a small subset of schools.

The use of data in UCI research efforts is the underlying need that ties RCI efforts across all domains. Whether data are *big* or not is not as crucial as the fact that the management, transfer, analysis, generation, and acquisition of data are crucial. UCI's current cyber infrastructure environment to do all these tasks is not adequate for increasing research needs. A particularly strong concern is the lack of robust data management and storage capabilities. Growing compliance constraints from funding agencies on data owners to provide proper data management and curation is putting grant awards at risk.

The use of sensitive or confidential data in research crosses many fields. The ability to store and work with such data goes well beyond medical research. Research efforts are hampered by the inability of faculty to use secure data storage and computational resources.

Computational needs for research require resources at the campus level. While campus computing clusters are marginally meeting that need, the aging infrastructure and reliance on pure grant funding for resources puts future research needs at risk of not being met. The growth of computational needs in many areas such as Biological Science and Medicine is putting additional demands on cluster resources and support.

Many faculty are continuing to make use of RCI resources available to them at other institutions due to previous graduate student, postdoc or faculty appointments. This access is required to carry out their research as UCI does not have comparable services available to them. These include data storage, software, user support, and access to datasets. Several faculty reported that they expected these tools when they came to UCI but were surprised that they did not exist. The lack of RCI resources presents a significant recruitment issue for future faculty hires as well as a retention issue for current faculty.

In addition to the lack of RCI resources in several areas (computation, data storage, high speed networking, and software), there is a strong need for human support resources in many areas. There is a strong need for application specialists with significant academic and IT credentials that can assist faculty as colleagues in research efforts.

Symposium Speakers

Opening Remarks

- Jim Hicks, Interim Vice Chancellor of Research
- Dana Roode, CIO & Associate Vice Chancellor, IT (Workgroup Co-Chair)
- Suzanne Sandmeyer, Biological Chemistry (Workgroup Co-Chair)
- Allen Schiano, OIT Research Computing (Symposium Emcee)

UC Guests

- David Greenbaum, Director, Research Information Technologies, UC, Berkeley
- Bill Labate, Director, IDRE Research and Technology Group, UC Los Angeles

UCI Faculty – Biological Sciences and Medicine

- Ali Mortazavi, Developmental and Cell Biology
- Karen Edwards, Epidemiology
- Robert Spitale, Pharmaceutical Science
- J. J. Emerson, Ecology and Evolutionary Biology
- Xiaohui Xie, Computer Science
- Lisa Dahm, Clinical Informatics, UCIMC

UCI Faculty - Humanities and Arts

- David Theo Goldberg, Comparative Literature
- John Crawford, Dance
- Julia Lupton, English
- Tim Elfenbein, Informatics
- Antoinette LaFarge, Studio Art
- Jesse Colin Jackson, Studio Art
- Robert Fuller, East Asian Languages

UCI Faculty - Social Sciences, Social Ecology, Business, Law and Education

- Bryan Sykes, Criminology
- John Hipp, Criminology
- Jade Jenkins, Education
- David Min, Law
- Elizabeth Martin, Psychology
- Tim Bruckner, Public Health
- David Brownstone, Economics

UCI Faculty – Physical Sciences, Engineering, Information and Computer Science

- Said Elsghobashi, Mechanical Engineering
- Filipp Furche, Chemistry
- Daniel Whiteson, Physics and Astronomy
- Michael Carey, Computer Science
- Aparna Chandramowlishwaran, Electrical Engineering
- Jim Bullock, Physics and Astronomy
- Steve Barwick, Physics and Astronomy
- David Mobley, Pharmaceutical Sciences
- Doug Tobias, Chemistry

UCI RCI Service Providers

- Joseph Farran, HPC Services, OIT
- Laura Smart, Data Management and Curation, UCI Libraries
- Tony Soeller, GIS Services, OIT
- Babak Shahbaba, Center for Statistical Consulting
- Francisco Lopez, System Administration Services, OIT
- Jessica Yu, Network Services, OIT
- Patty Furukawa, School IT Directors

Appendix C

RCI at other Universities

A review of RCI staffing, services and organization at other universities reveals a wide range of staffing and organization scenarios. It is difficult to provide an "apples to apples" comparison among them since RCI units do not provide a consistent set of services. In some cases, IT staff outside the core RCI group provide system administration and other services. However, cluster computing, shared storage, data management, and advance RCI consulting are the most consistently provided services.

This appendix provides staffing benchmark data from the Coalition for Academic Scientific Computing along with information about RCI at several other universities.

Coalition for Academic Scientific Computing Benchmark

The Coalition for Academic Scientific Computing (CASC, http://casc.org/), is an educational organization with 85 member institutions dedicated to advocating the use of advanced computing technology to accelerate scientific discovery. In 2013 CASC published a paper entitled "Academic Computing Management Benchmarks" with information gathered by surveying its members (*http://ssrn.com/abstract=2313089*). This provided some perspective on RCI staffing levels and funding that are summarized here.

Benchmark results were based on completed survey responses from 48 CASC members. Responding institutions were broken down into categories based on the size of the High Performance Computing they provided:

- 4 "Extra Large" institutions with between 90,000 and 400,000 compute cores
- 17 "Medium to Large" institutions with 10,000 to 50,000 cores
- 27 "Small" institutions, with 576 to 9,000 cores

UCI has a total of about 12k compute cores between the GreenPlanet and HPC compute clusters, putting us at the bottom of the "Medium to Large" category above.

Number of Cores per Cluster FTE support

The survey showed a median of 867 cores per FTE for the "Small" CASC members, and a median of 2100 cores per FTE for "Medium to Large" CASC members. Given UCI's 16k cores, this suggests between 7.6 and 18.4 FTE would be needed to support our compute clusters. UCI currently has 3 FTE doing this work across OIT and Physical Sciences.

Other Staffing

- Data Management: "Small" CASC members had between 0 and 3 FTE; "Medium to Large" had between 1 and 4 FTE.
- Advanced consulting and other support: 2 FTE was the median for "Small" CASC members, and 4 FTE was the median for "Medium to Large" CAC members.

Funding Sources

The survey indicated that “Medium to Large” members received 24% of their funding from federal sources, with the remainder from university, state and other sources. “Small” CASC members received 20% of their funding from federal sources.

Reporting Relationships

Of the combined group of “Small” and “Medium to Large” CASC members, 22 RCI Directors report to the VC of Research, and 24 report to the CIO.

UC Berkeley

Until 2 years ago, UC Berkeley provided only minimal central research IT services, relying on individual units and the neighboring Lawrence Berkeley National Laboratory for this. Larry Conrad became Chief Information officer in early 2013, and later that year faculty petitioned the Vice Chancellor of Research for enhanced research IT services including local shared compute clusters. The result was an initiative funded by the CIO, Vice Chancellor for Research, and Chancellor that led to the establishment of their current program managed by the UCB Research IT group.

UC Berkeley’s Research IT Group

- Comprises 20 people (15 FTE)
- Reports to the CIO in partnership with the Vice Chancellor for Research
- Approximate annual budget of $3.5 - $4m
- Provides a shared compute cluster (including 200,000 compute hour allowance for all faculty PIs), a range consulting services, and data management support
- http://research-it.berkeley.edu/

UC Los Angeles

UCLA provides broad research IT services through the Institute for Data Research and Education (IDRE) Research Technology Group (RTG). The RTG:

- Resides in the Office of Information Technology (dotted line to the VCR)
- Annual budget of approximately $4.3M; largely core funded, 10% from grants
- 22.25 FTE, more than half of which are PhD research scientists:
    - Geographical Information Systems & Visualization – 2.5 FTE
    - HPC Research & Projects – 4.5 FTE
    - HPC Systems & Networking – 7 FTE
    - Statistical Consulting – 4 FTE, 1 FTE students
    - Senior Staff – 3 FTE
- https://idre.ucla.edu/people/rtg

Key Services and Initiatives

- Computational cluster services (14k nodes)
- High Performance Computing/Storage Services (4 Petabytes)
- High Performance Precision Medicine (HPPM)
- HIPAA computation and storage infrastructure

- Overhead waiver for research services
- Data Management In partnership with the Library

Other Universities

- The University of Florida has a team of 16 research computing staff: https://www.rc.ufl.edu/contact/people/
- The University of Colorado at Boulder has a team of 10 staff and an academic director: https://www.rc.colorado.edu/about/staff
- Purdue has a large research computing program with over 50 staff: https://www.rcac.purdue.edu/about/staff/

Appendix D
Supporting Diagrams

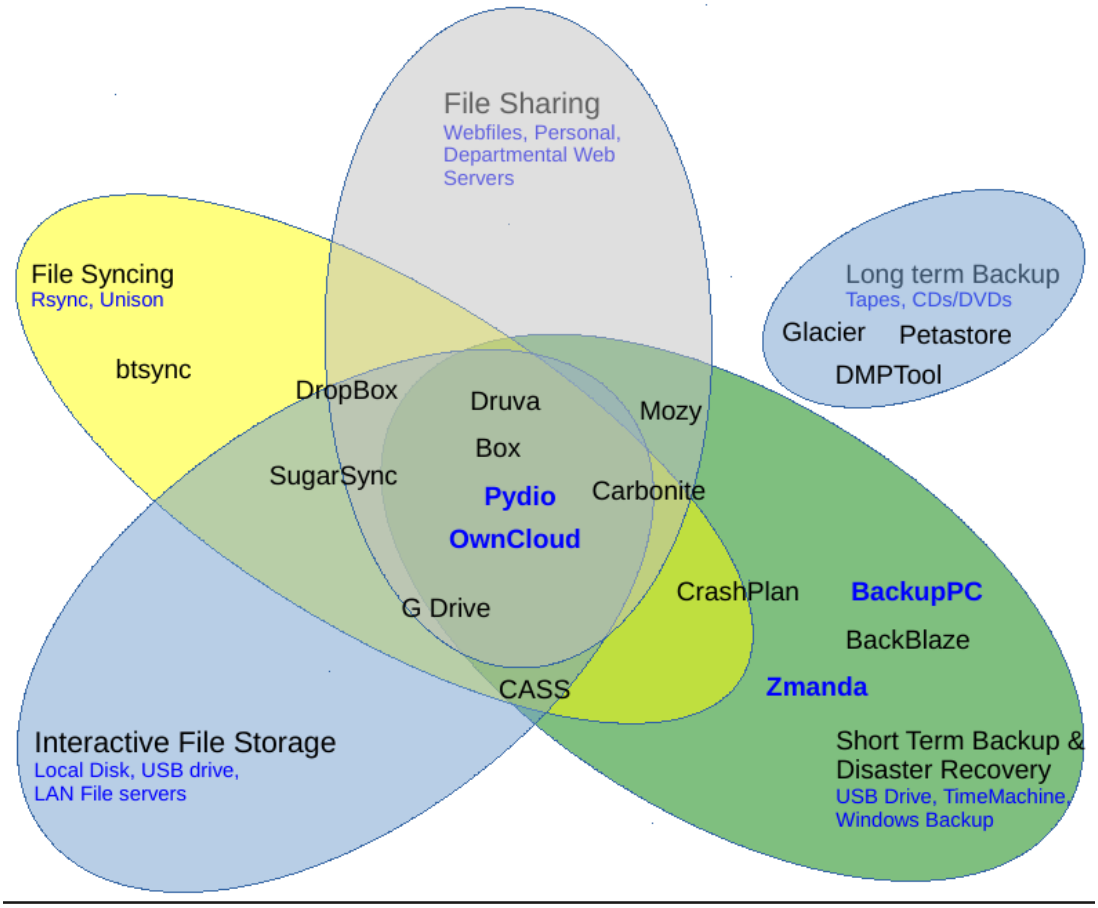

Figure 1

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.

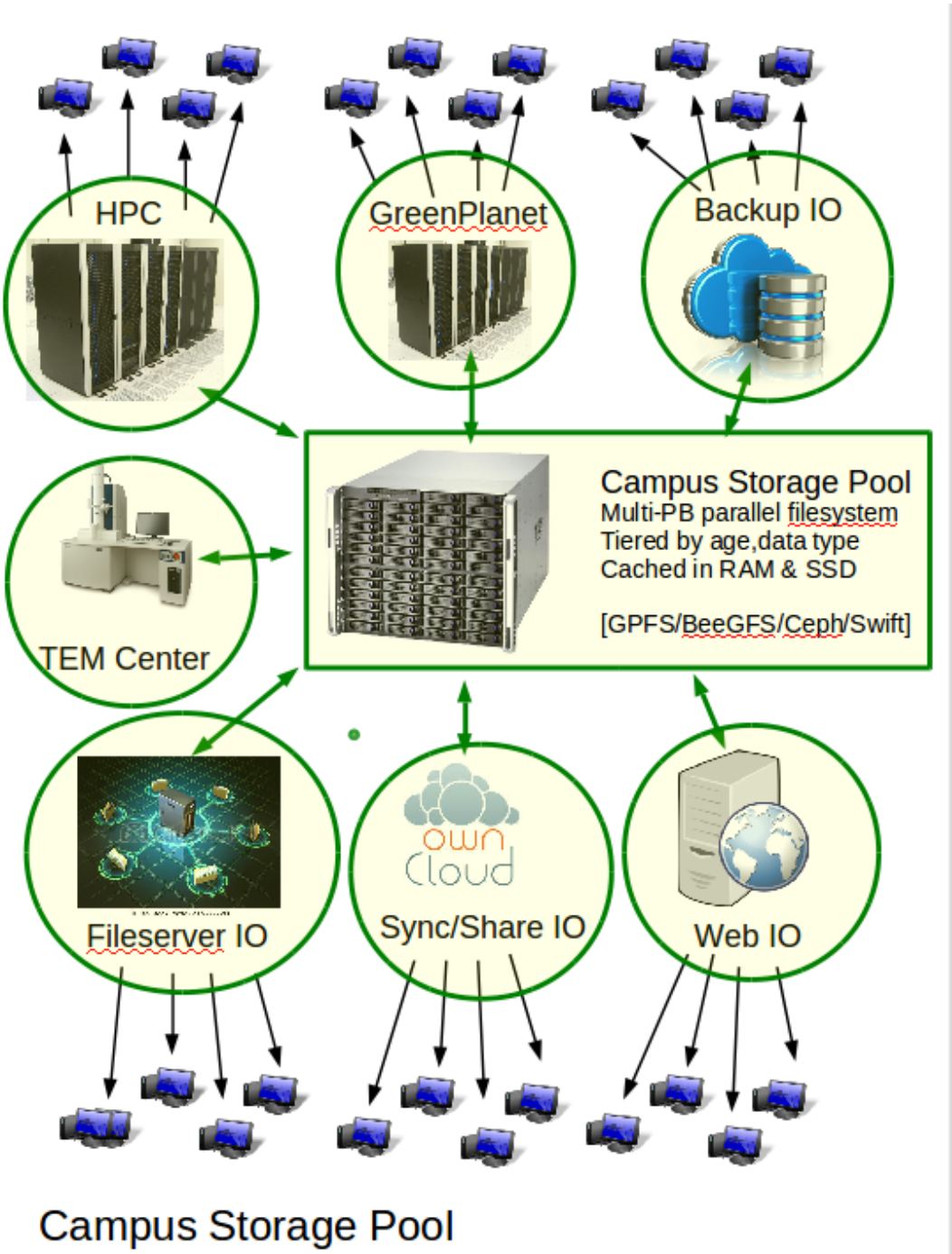

Figure 2

Campus Storage Pool Technical Diagram

Appendix E
Draft RCI Center Budget

	2016/17		2017/18	2018/19	2019/20
	Annual	Partial-Year	Annual	Annual	Annual
Salary & Benefits (New)					
RCIC Director	190,000	142,500	190,000	195,700	201,571
Cluster Admin	130,000	107,900	130,000	133,900	137,917
Cluster Admin	130,000	65,000	130,000	133,900	137,917
Storage Admin	130,000	107,900	130,000	133,900	137,917
RCI Specialist	130,000	97,500	130,000	133,900	137,917
CI Engineer	130,000	130,000	130,000	133,900	137,917
Library Data Specialist	130,000	97,500	130,000	133,900	137,917
Additional RCI Staff TBD	130,000	32,500	130,000	133,900	137,917
Additional RCI Staff TBD			65,000	130,000	130,000
Additional RCI Staff TBD				65,000	130,000
Salary & Benefits (Existing)					
Manager (.25 FTE)		54,000	55,620	57,289	59,007
Cluster Architect		166,000	170,980	176,109	181,393
RCI/HPC Specialist		153,000	157,590	162,318	167,187
GIS Specialist		153,000	157,590	162,318	167,187
Programming Support (.5 FTE)		77,000	79,310	81,689	84,140
Administrative (.5 FTE)		45,000	46,350	47,741	49,173
Total Salary and Benefits		*1,428,800*	*1,832,440*	*2,015,463*	*2,135,077*
Other Expense					
Storage Hardware		500,000	400,000	400,000	400,000
Cluster refresh		600,000	600,000	600,000	600,000
Research Networking		400,000	400,000	400,000	400,000
Software Licensing		100,000	100,000	100,000	100,000
Supplies & Misc		50,000	50,000	50,000	50,000
Total Other Expenses		*1,650,000*	*1,550,000*	*1,550,000*	*1,550,000*
Total Expense		**3,078,800**	**3,382,440**	**3,565,463**	**3,685,077**
Funding Sources					
From OIT		750,000	750,000	750,000	750,000
Recharge/Grant		200,000	400,000	500,000	600,000
Central		2,128,800	2,232,440	2,315,463	2,335,077