Developing Sustainable Data Services in Cyberinfrastructure for Higher Education

Requirements and Lessons Learned

Wilfred W. Li, Richard L. Moore, Matthew Kullberg, Brian Battistuz, Steve Meier, Ronald Joyce, Richard P. Wagner
San Diego Supercomputer Center, UC San Diego
La Jolla, CA 92093, USA
(wilfred, rlm, mck, battistuz,rjoyce,rpwagner, smeier)@sdsc.edu

Tad Reynales, Qian Liu
Qualcomm Institute
UC San Diego
La Jolla, CA 92093, USA
(reynales, qianliu)@ucsd.edu

Abstract— The University of California, San Diego (UC San Diego) Research Cyberinfrastructure (RCI) program provides long-term quality services in centralized storage, colocation, computing, data curation, networking and technical expertise. To help define the data storage needs and set priorities, the RCI data services (RCIDS) team conducted a series of interviews with faculty and senior staff members between September 2012 and February 2013. A total of 50 groups from 29 separate departments and organized research units (ORUs) participated in the interviews, representing more than 600 UC San Diego researchers. From human genomic sequences, marine natural products, to cosmological simulations, their diverse datasets are shared with hundreds of thousands of users worldwide. The top 10 requirements on data services and the top 5 existing challenges and risks as reported by UC San Diego researchers have been identified. Based upon these requirements, the RCIDS team recommends a *Network Attached Storage* (NAS) data service to be first deployed with a sustainable business model. Additional services will be developed through further discussion with the research community and in view of emerging cloud computing technologies. An extensive discussion is provided on the implementation plan, cloud-based data services, and the lessons learned in building sustainable e-science infrastructure for higher education research.

Keywords—Higher education; research cyberinfrastructure; data cloud; sustainable data services; network-attached storage.

I. INTRODUCTION

Cyberinfrastructure [1] is increasingly recognized as essential for the state of the art universities in the 21st century because it enables "faster, better, and different scientific capabilities" [2]. Institutional development of high performance computing environment has been correlated to increases in National Science Foundation (NSF) funding and research competitiveness [3]. The investment in cyberinfrastructure is no longer a question of "why and when" but "what and how". Data is the cornerstone of successful research laboratories and reputable universities in a data-driven knowledge economy. The proliferation of data from sensor networks, large scale mapping and surveys, scientific simulations, and user contributed content in social networks is constantly redefining the perception and understanding of "big data." Big data is characterized by the volume, variety and velocity of data at an unprecedented scale [4]. Building a sustainable infrastructure to support the data life cycle and expedite the retrieval of information and genesis of knowledge is absolutely necessary for any university to lead higher education.

In April 2009, the UC San Diego RCI design team (RCIDT) published a report titled "Blueprint for the Digital University" [1]. The campus-wide cyberinfrastructure is designed to be environmentally conscious through the adoption of green technology and economically sound through cost-effective services. The goals are 1) to provide UC San Diego researchers with competitive advantages when they conduct research and apply for grants, 2) to enhance the quality of student education at UC San Diego, 3) to preserve the University's data and intellectual property, 4) to reduce redundant campus infrastructure, and 5) to conserve the global environment through green technology. The six core elements are centralized storage, colocation, data curation, research computing, research network, and technical expertise. Examples include the deployment of UC-wide computing resources such as the Triton Resource [2] and the new Triton Shared Compute Cluster (TSCC) [3].

In the area of centralized data storage, the rapid growth of big data sciences has outpaced existing support mechanism to meet the demands. To stay current with the investigators' research advances and data growth, the RCIDS team has conducted a series of in-person interviews with UC San Diego faculty investigators and information technology (IT) staff members, guided with a questionnaire [4] focused on research data services. Ultimately, a long-term sustainable business

This work is supported in part by the UC San Diego RCI program and the NSF award to PRAGMA OCI-1234983.

[1] http://rci.ucsd.edu/_files/Blueprint.pdf
[2] http://tritonresource.sdsc.edu/
[3] http://rci.ucsd.edu/computing/index.html
[4] http://tinyurl.com/omfchcq

model is to be developed with the active participation of UC San Diego researchers to build an infrastructure that meets their needs to create, share and discover.

The rest of this paper will first describe related works in this area, followed by the demographics of the interview participants, and the results including common data flows, usage patterns, requirements, risks, and challenges. Then the implementation plan is presented with respect to the business model, the rationale and recommendation of an initial NAS data service, and an exploratory view on cloud based data storage solutions in the near future. Finally, the conclusion summarizes the key findings that provide insight to the practical implementation of sustainable campus cyberinfrastructure for e-science in higher education.

II. RELATED WORKS

A. Data management and sharing requirements

There has been a gradual recognition that tax dollars funded research could be made more valuable by increasing public access, long-term preservation and re-use. Researchers need to prepare for increasing expectations from funding agencies, as directed by the White House Office of Science and Technology Policy (OSTP) guidance in early 2013 [5]. Funding agencies already impose various guidelines on how long the data may be kept and shared through requirements of data management plans (DMP) by the NSF and data sharing plans (DSP) by the National Institutes of Health (NIH). The Royal Society published a report [6] titled "Science as an open enterprise" in June 2012 outlining ten recommendations that may lead to a new paradigm for scientific research. As the number two recommendation, "universities and research institutes should play a major role in supporting an open data culture… ." Darby *et al* have conducted a series of interviews and workshops to identify drivers, barriers, and enablers to the process and context of data sharing and reuse [5]. In particular, they have identified the need for a persistent and sustainable preservation infrastructure through a combination of journal, learned society, funding agency, and institutional archives. Xuan *et al* have proposed architectural design principles for infrastructure to support data integration and curation in higher educational institutions [6].

B. Other institutional efforts

There have been increasing efforts at different institutions around the world to develop sustainable models for research computing and data management services. Frequently these efforts are characterized by campus-wide activities with the participation of the faculty, staff, information technology experts in research computing, and librarians for data curation. For example, University of Colorado Boulder published a comprehensive report on their recommendations in research data management [7] with detailed analyses on the funding models from peer institutions. Purdue University has established a PURR data repository using the HubZero platform [7] with a very active user community. In the data storage area, the RCIDS team has collected raw data storage pricing information from about 50 Universities that offer storage space. The numbers range from $80/terabyte (TB)/year for a single copy of data, or $250/TB/year for replicated data, to about $1000/TB/year. In the data curation area, curated data storage is often offered in the range of $1000~$2000/TB/year with long-term preservation up to 10 years or more. It should be noted that these prices vary greatly depending on the number of copies, the quality of hardware, the value of added services, and the level of university subsidies.

Univerisity of Michigan (UM) offers a comprehensive set of cyberinfrastructure services with a useful "storage solution summary chart" for the different options [8]. UM is among the first to negotiate contracts with Box and Google to offer cloud based storage solutions for university use. These commercial solutions offer friendly user interface and ease of access from mobile devices.

C. Shared research cyberinfrastructure activities

As the price of commodity hardware for computing and storage drops, more and more researchers find themselves mired in the business of managing laboratory needs in computing and storage on a routine basis. In recent years, the condo computing model has gained popularity. The principal investigators (PIs) purchase the computing hardware through their grants and their universities fund the professional management to various degrees. However, the condo storage model has not gained much traction. The complexity of providing condo storage due to the persistent nature of data and funding agency requirements has hampered its implementation and adoption. Only a handful of universities offer services for PI-purchased storage hardware. For example, Purdue University and Clemson University have programs in which the universities pay the full cost of operating the hardware. UCLA is currently rolling out its own condo storage plan in 2013 allowing UCLA researchers to purchase shared storage space.

III. METHOD

A. Demographics of interviewees

The selection of the participants is based on known knowledge of data storage needs and representative of similar investigators on campus in their respective areas of research. The participants represent 29 different departments and organized research units (ORUs) on campus, including researchers from the fields of biology, physics, sociology, marine studies, arts and humanities, medicine and pharmacy. The total number contacted for interviews was 80 with 50 completed interviews at a 63% response rate. This is about 4% of the UC San Diego ladder rank faculty members. According to an estimate based upon the average group size, the total number of personnel covered under these interviews is about 600 including faculty, staff, researchers, and students. Similarly, the estimated number of *external* users who may benefit from the released research data from these laboratories is about 300,000. The major funding sources for these researchers are NIH, NSF, and private foundations.

[5] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[6] http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

[7] http://hdl.handle.net/10971/1398

[8] http://www.itcs.umich.edu/itcsdocs/r1468/

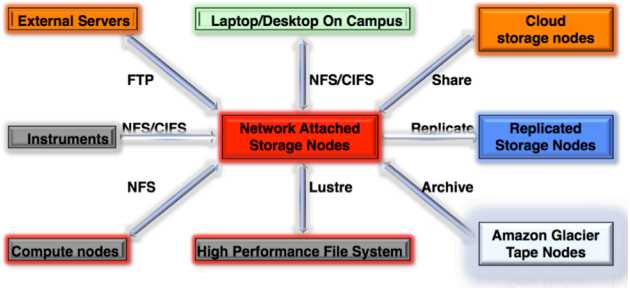

Fig. 1. Simplified mashup of common data flow patterns. Researchers use a subset of these components usually depending on actual requirements.

B. *Interviews*

The interviews typically last one hour. The questionnaire was provided ahead of time and the responses were entered by the interviewers based upon the replies from the interviewees. This provides the opportunity to clarify any confusion that may exist due to misinterpretations of technical jargons. Detailed notes were taken to capture any issues not mentioned in the questionnaire but deemed important by the PI's. Any response entries not covered by the multiple choices provided are entered as exceptions in a text field category of "Other." During the compilation phase, these responses were expanded into new categories or combined with existing categories when appropriate.

IV. RESULTS

A. *Common data flow patterns*

Researchers from the interviews often use a subset of the illustrated components in their routine data flows (Fig. 1). Data generated from software applications or instruments such as sequencers are sent to the primary storage nodes, which may be processed and analyzed using cluster compute nodes, or a high performance file system in TSCC or XSEDE. Usually data is replicated to a secondary storage node through snapshots, which also serve as a limited backup mechanism when the data turnover rate is low. With high data turnover rate, the only data retrievable is from a snapshot up to 24 hours earlier with a daily snapshot policy. The primary storage is considered "hot," as the data is always accessible from laptops or workstations using the Network File System (NFS) or Common Internet File System (CIFS) protocol. Amazon Glacier is a long-term archival solution for data rarely needed. Cloud storage nodes are often used for sharing data through a private cloud provider such as the San Diego Supercomputer Center (SDSC) Cloud Storage [9] or commercial providers such as Dropbox, Google Drive, or SkyDrive.

Table 1. Data Sources and Relative Distribution

Data Source	*%*	**Representative Fields**
Sequencers	28	Biology
Software applications	28	Biology, Physics
Field sensors/instruments	20	Marine Biology, Astronomy
Audio visual equipments	10	Arts
Mass spectrometers	8	Biology
Tomographic instruments	8	Biology, Medicine
External data repositories	8	Biology
LHC particle dectors	3	Physics
Archelogical studies	3	Humanities
Curation	3	Sociology

Table 2. Data Storage Devices and Services Utilized

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

The data sources for Fig. 1 are listed in Table 1, and the type of storage devices and services utilized are shown in Table 2. Topping the data sources are sequencers and software applications for simulations in computational chemistry, biology and physics. The genomic sequencing effort has been generating data faster than Moore's law with the cost of sequencing an entire human genome lowering to about $1000. Soon it might be cheaper to resequence a genome than to store the data for the long term.

The majority of researchers (73%) use NAS, 70% use USB drives, and 65% use local hard drives on workstations to store their data. Of these, USB drives and local hard drives are considered economical storage options, and both may be taken offline for long-term archival. Up to 43 % of the participants use Dropbox (33%) or Google drive (10%) to manage their small data sharing (up to tens of GB) needs.

The storage density continue to double approximately every 13 months, according to the Kryder's Law [8], faster than the 18 months cycle for transistor density in the Moore's Law. In addition, plug'n play NAS devices such as Drobo drives or Synology NAS servers have seen increased adoption

[9] http://cloud.sdsc.edu

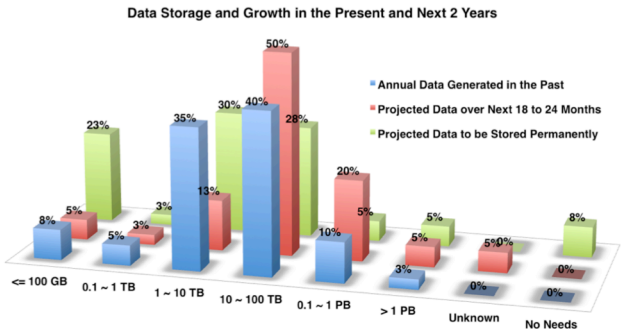

Fig. 2. Growth of data storage neeeds in the present and the next two years.

as a local NAS option for small research groups that do not have strong performance requirements.

B. Growth of data storage needs

The current, projected and permanent amount of data storage over the next two years are illustrated in Fig. 2. The researchers are primarily dealing with data in the 1 to 100 TB range (75%), though more users will be dealing with data in the 10 to 1 PB range (70%) in the next two years. However, the amount of data to be kept permanently are under 100 TB for most users and under 100 GB for 23% of users.

Spending on equipment shows that the majority of PI's spend less than $25K per year on inventorial storage equipment (Fig. 3). This indicates that most researchers are dealing with data in the 10~40 TB range, using an estimated raw storage hardware cost of about $500/TB. Those who spent money in the past 18 months also tend to spend less in the next 18 months because of the storage hardware tend to last 3 to 5 years on average.

C. Data life cyle characteristics

When asked about how long they intend to keep their data, the researchers mostly only intend to keep the data for the duration of projects (Fig. 4). The majority also recognize the need for metadata annotation and data curation (data not

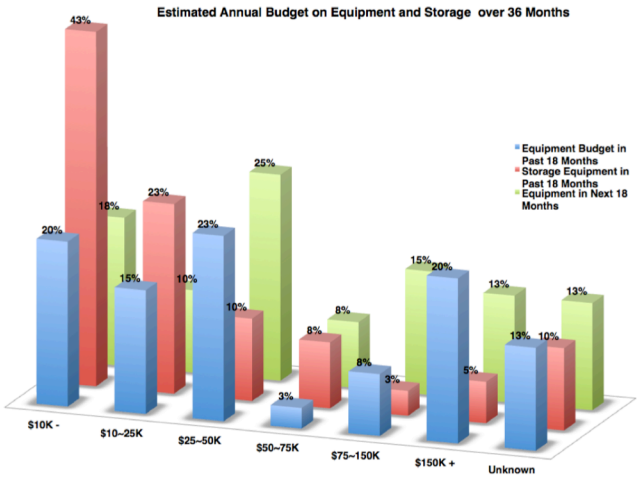

Fig. 3. Annual spending on inventorial equipment which are over $5000 in value over the past and future 18 months. Red columns are for storage equipment only in the past 18 months.

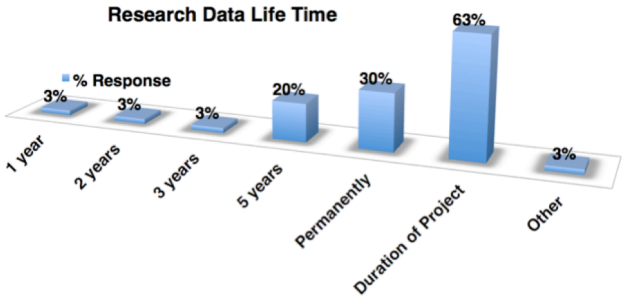

Fig. 4. Research data life time as reported.

shown). About one quarter of the respondents have metadata annotation or data curation needs, and about one third of the respondents will add metatdata when better tools are available. Yet 40% of the respondents have no foreseeable metadata or curation needs, and five percent are unsure of their metadata needs at the moment. This distribution may change as funding agencies develop more detailed requirements and funding models for data curation, archival, and preservation.

D. Risks and Challenges

The top 5 risks and challenges facing the researchers are shown in Table 3. The number one risk is the sustainability of the campus research cyberinfrastructure program. The termination of the RCI program poses a serious problem if they have already invested time and grant money into the research cyberinfrastructure themselves. This concern outranks the constant increasing demand of data storage needs (65%) and lack of back plans and dedicated support staff to keep data safe (50%). Another top concern is that the cost may be higher down the line in a "bait and switch" style (53%). These concerns stress the need to develop a long-term strategic plan for campus research cyberinfrastructure.

E. Requirements

The top 10 requirements from the interviews are shown in Table 4. Of these, the top categories are cost, sharing, ease of use, backup and recovery, and network bandwidth. These requirements are in agreement with the proposed minimal criteria for cloud based data services [6]. However, there is one important distinction. The majority of research groups are using NAS as the principal platform for data storage, as illustrated in Fig. 1. Cloud based storage solutions such as SDSC Cloud and Amazon S3 only received adoption by a small percentage of users. On the other hand, the dropbox- or Google Drive-like services are preferred by up to 43% of

Table 3. Top 5 risks and challenges today and in the near future

Risks and Challenges	***%***	**Category**
Campus may cease funding for RCI	85	Adoption barrier
Constantly increasing storage demands	65	Distraction from research
Bait and switch with increased cost later	53	Poor business practice
Poor backup plan	50	Lack of expertise
No dedicated support staff	50	Distraction from research

Table 4. Top 10 requirements for campus cyberinfrastructure			
Type	***%***	**Comments**	**Category**
Better CI with inimal direct cost	91	Least burden on research budget	Cost
Network Attached Storage	73	Shared POSIX compliant filesystem	Sharing
Data replication as backup	66	Keep a second copy somewhere safe	Recovery
Dropbox- or Google Drive-like service	43	Ease of access and worry free backup	Ease of use
10G network connection	38	High speed network bandwidth	Network bandwidth
Minimal cost beyond hardware cost	24	Little operating cost	Cost
Shared technical expertise	20	Infrastructure, software and application consulting	Expertise
Distributed multisite replication	18	Geographical safety	Recovery
Desktop backup	18	Routine research data safety	Backup
Compliant and secure storage for sensitive data	16	Personal and clinical data safety	Security
Tiered storage plans	16	Data retention and automatic removal	Cost

researchers. Users will choose different platforms depending on actual requirements and usage scenarios. While it is tempting to develop a single solution for all scenarios, it may be more practical and economical to have a number of options to choose from.

V. IMPLEMENTATION PLAN

Most PI's are interested in using grant funding to purchase hardware storage and have them professionally managed to keep their data safe. Thus, a sustainable business model is needed for condominium style shared storage solutions where the PI's purchase full or partial condo storage based upon their needs.

The UC San Diego Cyberinfrastructure Planning and Operation Committee (CIPOC) report in 2010 outlined a guiding principle in the development of a business model for RCI Centralized Storage. The PIs pay "79% of the start-up costs, 59% of the annual costs for the initial years, and 68% of the steady-state annual costs" and UC San Diego provides the rest of the cost through energy savings and indirect cost recovery. The PIs pay approximately $100/TB/year on hardware only for a replicated data storage service without including the costs for networking, colocation, and labor. As a rule of thumb, doubling the cost of the hardware provides a rough estimate of the total cost of ownership (TCO) for replicated data storage over the useful lifetime of the equipment. Thus, a ballpark figure is $200/TB/year minimum cost if the required economy of scale is achieved. Universities that pay the full cost of operating the hardware provide roughly a 50% subsidy to the TCO.

A. Business model considerations

The current business model under development builds upon a fixed investment that covers the startup, optimization, and operation of the core business processes. The economy of scale allows all participating PIs to share the incremental cost and lowers the average cost of operating individual storage nodes. The PIs will enjoy professional management of their critical research data at the most economical price and collectively bear the incrementally growing labor requirements. The viability of this business model may be influenced by the following factors:

1) Cost-effectiveness. Most users would perform cost comparisons before they make purchase decisions. The storage pricing has to be low without affecting the quality of service to provide the most cost-effective solution. It is noteworthy that the published rates by universities are often for data curation, designed for high quality data in small amounts to be kept permanently. Therefore, a service description and the advertised cost should be examined together.
2) The subsidy from a university reduces the cost for the PIs and the number of participating PIs builds up the economy of scale. Commitments from both sides are required to generate a recharge rate that is cost-effective to the PIs and sustainable to the university. Of the services following the condo storage model, Notre Dame University offers $250/TB/year without backup; University of Michigan offers the same rate with backup; UCLA offers volume discounted prices starting from $236/TB/year with backup and single copy is available for half of the costs.
3) Service longevity. The PIs need assurance that the service will be there at the approved recharge rate. It may take up to a year for researchers to get grants funded using the published recharge rates. Researchers should not have to fear increased costs when they get their grant awards or lose access to their data when they have funding gaps. The budgetary threat must be removed in the strategic planning process for a university to ensure the economy of scale necessary for long-term viability of research cyberinfrastructure programs.
4) Market demand. The interview participants are selected because they likely have major data management requirements. The actual market size for a university as a whole may not be extrapolated directly from the interview data. The biggest users may have application-specific requirements and smaller users may be happy with USB or local NAS drives. Therefore, the main target users may be those dealing with tens to hundreds of TB of data (Fig. 2). A total amount on the order of 3~5 petabytes (PB) may be required to achieve the economy of scale necessary and maximize cost-effectiveness.
5) Actual adoption rate. The adoption rate is influenced by all the other three factors above, and is the indicator of a successful business model. Advertising at different levels

may help increase the community awareness of research data services to encourage adoption.

6) Service scalability. The number of users and complex support requests from an individual group will increase the cost of user support. A missing campus-wide UID/GID system at a university should be resolved with a long-term solution with consideration of federated identities with other universities. The business processes must be streamlined with greater automation to reduce the operating cost of adding and supporting individual users and groups. Short-term users increase the operating cost and reduce the service scalability unless the business process is highly automated. Complex support requests may be resolved through limited support of shared technical expertise and a recharge mechanism.

7) Integrated services. Due to the diversity of requirements by the researchers (Table 4), there may be budgetary constraints on the number of services provided. Further discussion on the specific requirements is necessary to determine whether they can be met with existing campus-based recharge services, as opposed to developing new services. Moreover, the most cost-effective solutions are often developed with the complete data life cycle [10] in mind, i.e., from data collection, processing, distribution, discovery, analysis, to repurposing, and archiving. Coordinated efforts using different technologies are required to develop integrated solutions from the moment of data creation to archival.

B. *Recommendation*

The NAS data service is central to the research data workflow of many laboratories (Fig. 1) and utilized by 73% of the interview participants (Table 2). Even though the 50 groups of researchers represent less than 5% of the UC San Diego faculty members, they generated more than two PBs of data over the past year, and expect to add new data at that rate over the next two years.

A NAS data service from RCI addresses the following risks and challenges:

- Constantly increasing storage demand (65%)
- Lack of or incomplete data management and backup plan (53%)
- No dedicated support staff (50%)

SDSC has offered a NAS data service called Project Storage [11] over the past 18 months. It currently comprises 8 nodes, each with 88TB usable space and fully replicated. More than 20 research groups have taken advantage of the service. The adoption of Project Storage is gradually increasing with good performance and stability from the underlying architecture (Fig. 5). The Project Storage offers a hotel model as well, which allows users to pay a higher rate without having to purchase any hardware. The existing Project Storage customers are converted into the RCI NAS data service to enjoy a reduced rate when the service goes into production.

[10] http://www.ddialliance.org/what
[11] http://project.sdsc.edu

The RCI NAS data service enables the flexible provisioning of professionally managed storage space on demand with full replication as the backup option and support of disaster recovery. The service also guarantees campus-wide accessibility for data storage and sharing. The NAS data service is priced favorably to encourage sufficient adoption to achieve the economy of scale. There is minimal operating cost over hardware and the bait and switch practice is avoided.

C. *Cloud computing*

In the era of big data, the demand for a high integrity, centralized, and durable storage solution is getting stronger and stronger. According to the Gartner, Inc., "big data is forecast to drive $34 billion of IT spending" in 2013 [12]. From the perspective of data availability and protection, storage is the most critical component of an IT infrastructure. Managing the design, deployment, growth, consolidation, backup, recovery, and archival solutions and making informed and strategic purchasing decisions are constant challenges facing both educational and industrial institutions.

The RCI NAS data service is a storage solution that provides reliability and standard performance at a modest cost. It is widely used by small- to medium-sized research labs in the university environment. This platform complements the high performance parallel file systems such as Lustre, GPFS (global parallel file system), and PVFS (parallel virtual file system), which may scale to thousands of nodes and tens of thousands of cores. However, these high performance solutions are much more expensive and are usually associated with high performance computing (HPC) systems. As illustrated in Fig. 1, researchers need to pick the best data storage platform based upon their application and performance requirements.

Cloud computing, as originally described in the 2009 RCIDT report, combines storage and computing resources into "a set of managed infrastructure operated by third-party providers", with emerging technology that enables on-demand provisioning of machines (IaaS, infrastructure as a service), execution environments (PaaS, platform as a service), applications (SaaS, software as a Service), and network

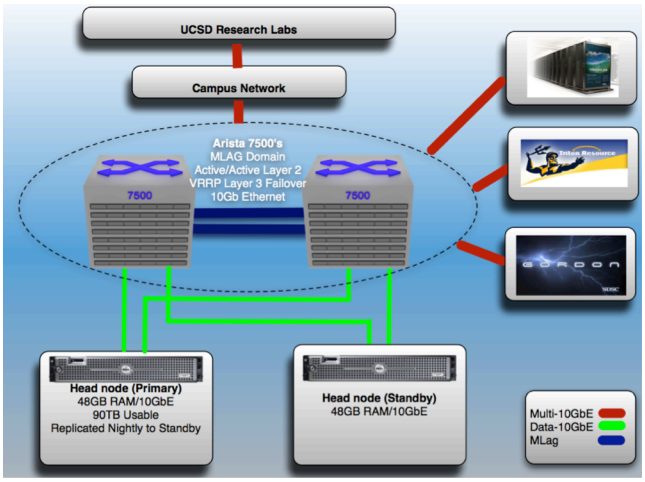

Fig. 5. The architecture diagram of the RCI NAS data service.

[12] http://www.gartner.com/newsroom/id/2200815

resources (NaaS, network as a service). "More than 30% of organizations will move at least one enterprise data center workload to the cloud in 2013. Cloud services will be deployed even more broadly for private storage, including disaster recovery, backup, and archiving," according to Jay Kidd from NetApp [13]. Amazon EC2 (elastic computing cloud), Amazon S3, and Google Cloud Platform are leading cloud computing providers. Many popular services such as Dropbox and Google Drive leverage the cloud computing platforms to provide SaaS.

There is strong competition between cloud storage providers such as Amazon or Rackspace and high capacity storage solution providers such as Dell, Data Direct Networks (DDN), or EMC Corporation. Cloud based storage provides a reasonable cost of entry but lacks the performance, ease of access, and cost effectiveness desirable in the university research environment currently. On-site enterprise solutions such as Dell Compellent and Oracle SAM-QFS provide robust, high performance, and tiered storage but have high purchase prices and maintenance fees. Many of these providers such as Rackspace and Dell also offer private cloud solutions to organizations that want full control of their cloud environment behind their own firewalls. While they are expensive to implement at universities, commercial vendors often offer academic discounts to increase adoption. For example, Cornell Univeristy has partnered with DDN to offer storage solutions [14] that include endpoints for Globus Online [15].

D. Data cloud services

SDSC Cloud Storage was unveiled with more than 2 PB of storage capacity in 2011 as the largest academic private cloud storage environment to date. The recharge rates of SDSC Cloud Storage have been lower than SDSC Project Storage because of the OpenStack Swift object-based storage platform [16]. However, most researchers require POSIX file systems such as the NAS data service that support traditional applications. While it is currently possible to use additional hardware enabled configurations to access SDSC Cloud Storage, the price-performance ratio has not reached a level required by the mass market.

GlusterFS and cephFS are two open source cluster filesystems that provide easy scale-out up to the PB range using commodity hardware. Both are POSIX compliant and offer a single namespace and native support for S3 and Swift APIs. They are excellent choices as processing workspace in genomics applications. However, they are still maturing and extensive testing and customization may be required for production use currently. Commercial support is provided by RedHat and Inktank respectively.

Many UC San Diego departments have built smaller IaaS private cloud services to provision virtual servers and virtual storage to minimize the TCO. UCLA is rolling out cloud storage solutions that utilize IaaS using virtual frontends to enable automatic data replication as soon as data is written to one node [17]. This provides the benefit of load balancing and "hot replication," with little downtime for disaster recovery.

Through a recent NSF award, researchers at SDSC and Calit2 have begun to build a new "big data freeway," a state of the art research network named PRISM@UCSD [18], which will provide 10G and 40G connections between participating research laboratories. CENIC[19] is contributing to establishing 100G connections to UC San Diego that will provide even bigger pipes for data sharing. This may enable commercial cloud access from the campus using virtual private cloud. Together, campus-based solutions will continue to offer the best performance in the "big data freeway". The emergence of software defined network (SDN) will make NaaS a reality and change how secure and compliant storage is implemented.

In short, advances in cloud computing and network technology will change how data services are implemented in the future.

VI. DISCUSSION

Many researchers use the free data storage provided by Dropbox, Google Drive, or other commercial services despite privacy concerns. While Dropbox and Google Drive may be sufficient for small data sharing and backup, they become more expensive when the amount of data increases beyond the low TB range. UM has provided a leading example in using its collective bargaining power to negotiate favorable licensing and privacy terms to contract these services to commercial companies. However, these commercial services may suffer from degraded synchronization performance over wireless network for data sizes beyond tens of MBs. They also have limitations on maximum file size support.

A campus based Dropbox-like solution may offer better network performance for data transfer and synchronization. For example, researchers may use the NAS data service coupled with commercial software such as GoodSync or CrashPlan for data synchronization and backup. SDSC Cloud Storage also provides several interfaces such as a web interface, a CyberDuck GUI application, and a command line client for data storage. All the data stored in SDSC Cloud Storage may be shared easily through its web interface. Globus Online has been a successful example for transferring big data. There are also ongoing research efforts to build biomedical research data cloud services using the Duckling [20]Collaboration Library and Opal [21] services (Dong *et al*, unpublished data). Researchers may choose the most economical option depending on data sizes, sharing needs, production readiness, and performance requirements.

Like Dropbox or Google Drive, researchers also expect research data services to be accessible, affordable, reliable, and sustainable. For example, Google Drive content is searchable for easy data retrieval. Metadata and provenance tools may be provided to enable research intelligence and data mining.

[13] http://searchstorage.techtarget.com/Top-Storage-Trends-for-2013-NetApp-for-SearchStorage
[14] http://www.globusworld.org/files/2013/12-Lifka-Cornell_University_Center_for_Advanced_Computing_a_Sustainable_Business_Model_for_Advanced_Research_Computing-Keynote.pdf
[15] http://www.globusonline.org
[16] http://www.openstack.org/software/openstack-storage/
[17] http://registration.cenic.org/cenic2013slides/UCLACloudStorage.pptx
[18] http://bits.blogs.nytimes.com/2013/03/20/a-big-data-freeway-for-scientists/
[19] http://www.cenic.org
[20] http://duckling.sourceforge.net
[21] http://sourceforge.net/projects/opaltoolkit/

These types of tools are under exploration in the RCI data curation element [22]. DuraSpace [23] hosts several open source projects in for digital data management that may offer turn-key solutions, e.g., DSpace or flexible and customizable solutions, e.g., Fedora Commons. DSpace has more than one thousand active instances in various universities worldwide.

In a recent talk on "Big Process for Big Data [24]", I. Foster suggested the scenario of "Dropbox for science" where research cyberinfrastructure is delivered in a "frictionless, affordable and sustainable" fashion and eventually provide a new model for cloud computing, aka, "science as a service." The enthusiam for Dropbox- and Google Drive-like services from the interviews is reflective of the desire for cyberinfrastructure that can transform the way scientific research is conducted in the digital age.

VII. CONCLUSION

The RCI NAS data service addresses the major current requirements for the interview participants and may be integrated into customized solutions for different schools, departments and ORUs. However, cloud computing and storage technologies are offering many evolving solutions that may change how research data services are delivered. The strong global initiatives to conduct scientific endeavors as open enterprises will lead to new paradigms for science. Future efforts in data services need to address the data life cycle integratively. Strong institutional commitment to support research cyberinfrastructure is a key requirement in the sustainability of the cyberinfrastructure for higher education

ACKNOWLEDGMENT

We would like to thank Drs. P. Arzberger, C. Bagwell, R. Hawkins, P. Papadopoulos for comments and suggestions.

REFERENCES

[1] Atkins, D., et al.: 'Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrasturcuture', 2003

[2] Hey, T., and Trefethen, A.E.: 'Cyberinfrastructure for e-Science', Science, 2005, 308, (5723), pp. 817-821

[3] Apon, A., et al.: 'High Performance Computing Instrumentation and Research Productivity in U.S. Universities', J Information Tech Impact, 2010, 10, (2), pp. 87-98

[4] Mattmann, C.A.: 'Computing: A vision for data science', Nature, 2013, 493, (7433), pp. 473-475

[5] Darby, R., et al.: 'Enabling Scientific Data Sharing and Re-use'. Proc. 8th IEEE International Conference on e-Science, Chicago, 8-12, Oct 2012.

[6] Xuan, P., et al.: 'An Infrastructure to Support Data Integration and Curation for Higher Educational Research'. Proc. 8th IEEE International Conference on e-Science, Chicago, 8-12, Oct 2012.

[7] McLennan, M.: 'Managing data within the HUBzero platform', Omics : a journal of integrative biology, 2011, 15, (4), pp. 247-249

[8] Walter, C.: 'Kryder's law', Sci Am, 2005, 293, (2), pp. 32-33.

[22] http://libraries.ucsd.edu/about/digital-library/index.html
[23] http://www.duraspace.org
[24] http://www.slideshare.net/ianfoster/pnnl-may-2013