ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Campus Data Storage Services Task Force
Final Report

March 9, 2012

Submitted by:

Mike Grady, Office of the CIO ; Task Force co-chair
Beth Sandore Namachchivaya, University Library; Task Force co-chair
Jason Alt, NCSA
Jack Brighton, College of Media, Center for Multimedia Excellence (CME)
Michelle Butler, NCSA
Mike Corn, Office of the CIO
Dan Davidson, Institute for Genomic Biology
Jennifer Eardley, Division of Biomedical Science, OVCR
Michael Edwards, College of LAS
David Gerstenecker, College of ACES
Gabe Gibson, College of LAS
Howard Guenther, Office of the Vice Chancellor for Research (OVCR)
Tom Habing, University Library
Maggie Helms, Division of Biomedical Science
Josh Henry, College of ACES
Alice Jones, AITS/UA
Joanne Kaczmarek, University Archives
Jackie Kern, Facilities & Services
Charley Kline, CITES
Carol Livingstone, Division of Management Information
Carol Malmgren, Office of the Registrar
Glenda Morgan, Office of the CIO
Frank Penrose, College of Engineering
Sarah Shreeves, University Library
Jason Strutz, University Library
Chuck Wallbaum, School of Chemical Sciences
Kristopher Williams, Materials Research Laboratory

Table of Contents

Executive Summary and Recommendations

The Task Force on Data Storage Services was charged with several responsibilities: surveying the local and peer institution practices around data storage and services, identifying unmet needs, and recommending both solutions and a strategy for the Urbana campus approach to data storage and related services, to support the campus' educational, research, and administrative mission. The Task Force began its work in August 2011 and submitted a final report with recommendations on March 9 2012.

From August 2011 through January 2012 the Task Force had the opportunity to survey more than 50 campus units, representing research, educational, administrative, and auxiliary functions on the Urbana campus. Over 80% of the academic units responded to the survey, sharing in-depth information about their use of storage and related services. Further, the Task Force established several working groups that focused on different aspects of the storage and storage services challenge, including research, workplace productivity, instructional, and institutional assets, sensitive data, security and privacy, and architecture. These working groups developed use cases that provide actual scenarios on campus for storage and service needs. Almost thirty individuals from units across campus contributed to carrying out the needs assessment and formulating the recommendations in this report.

In this process it has become clear that storage and related services are the new "baseline" technical requirement. It is a given that a professionally operated and consistent base level of computer networking is a core requirement for a major research university to remain competitive in today's educational and research landscape. The work of the Task Force has confirmed that storage services has reached that same level of criticality for universities and their mission, and an appropriate base level of storage and related services should now become a given. Further, evidence that the Task Force has gathered suggests that the Urbana campus could achieve substantial savings (both financial and personnel) in offering centrally-managed storage and related services. By minimizing the number of storage services, the campus could effectively decrease the amount of staff needed to manage storage. This strategy, while it may not work for all, could effectively free up edge IT professionals to support more specific data management needs at the unit level.

The Task Force recommends that the Urbana campus adopt the following storage services strategy, composed of seven key themes. The storage strategy will require ongoing action in a number of areas:

Strategy: Data Storage Services

I. Articulate Leadership and Establish Campus Storage Management Governance
Establish a campus governance structure that guides planning, policy, and operations around storage services. The Executive Governance Committee for Data Center Consolidation group offers a model and an opportunity for transforming into a Data Center Shared Services group.

II. Share Storage Resources
Enable federated storage among campus units: There are over seven petabytes of storage supported collectively across 50+ units on campus. Implement an architecture that enables sharing of storage among units, and allows for effective and efficient access to and transport of data. Enabling disparate pools of storage to be used as a more cohesive whole could quickly provide better usability of current storage, and potentially less need for each unit to over-provision storage to meet unexpected new needs. With the cost of a given amount of storage decreasing each year (a conservative estimate is 15% a year), buying storage before it is needed costs more.

III. Provide Centrally Managed Storage
Develop an architecture that supports an amount of centrally-managed, “common good” storage that enables provision of access to storage at both file system and more abstracted levels. Incorporate capabilities both from a central service, and among major data center nodes on campus. This builds on the advantages of federated storage, minimizing the number of distinct storage services across the campus, reducing FTE needed to manage storage, while also providing a consistent base upon which to offer value-added storage services.

IV. Provide Storage Management Services
Develop and offer centrally critical storage-related services, including backup, replication, and sensitive data management, as fundamental services available to all units for institutional, administrative, and research data. Provide services with tools for accessibility and ease of use across campus units. These services can be local or outsourced, depending on what the campus / university already does efficiently and effectively, and what is available in the "industry" (academia or commercial) to support these services.

V. Incorporate Cloud Services
"Bridge" to cloud services that support elastic, diverse, and evolving needs, both short- and long-term, as economics, scale and expertise warrant. Multiple approaches to, and types of, cloud services should be considered, including collaboration with other academic and research institutions, use of commodity cloud storage services, and participation in consortial activities like Internet2 and the CIC.

VI. Provide Storage Management Best Practices and Policies
Specify central solutions with tools for ease of use and best practices and policies for all storage environments, whether central or edge, for institutional (administrative, instruction, individual) as well as research storage needs and services. Establish an educational, communication, marketing and support effort around storage choices for campus users and units, and on effective data management practices and applicable policies, including sensitive data. Work in conjunction with other groups/units/roles, such as the campus Data Stewardship Committee, the Office of the Vice Chancellor for Research, and the Chief Information Security Officer, in the development of best practices and policies for data management.

VII. "Evergreen" Approach

Storage services and strategies are not a "one and done" effort. Effective long-term storage services will require consistent and constant refresh, investment, and support and maintenance. This includes an ongoing focus on a rapidly changing landscape of storage technologies, service options, user requirements, regulatory obligations, and technology opportunities and economics.

Specific Recommendations and time frame

The Task Force makes the following recommendations that address the strategies identified above. These recommendations are further described and explained in Section VI of this report. Note that the "Who" column represents key units and groups responsible for seeing that recommendation gets acted upon, not all of the units and groups that will be involved.

#	Recommendation	Start Date	End Date or Duration	Strategy theme addressed	Recommendation aligns with needs in these areas:	Who (see key below)
1.	Define a lead role in the Office of the CIO	March 2012	Open / continuing	I.	All (Architecture, Institutional, Research, Sensitive data)	CIO
2.	Establish governance of campus storage services within a Data Center Shared Services (DCSS) committee	April 2012	Continuing	I.	All areas as above	CIO, IT Council, EGCDCC
3.	Implement Box storage service for collaboration and file syncing across devices	In progress	August 2012 and continuing	IV., V.	Institutional, Research	Box project team
4.	Create sensitive data service pilot proposal	March 2012	June 2012	III., IV., V., VI.	Research, Sensitive data	CIO, OVCR, DBS, CISO, Library
5.	Establish policy and solutions for sensitive research data management	June 2012	December 2012	III., IV., V., VI.	Research, Sensitive data	CIO, OVCR, CDSC, CISO
6.	Develop best practices and policies, and a recommended solution set, for data management	In progress	Continuing	VI.	All	CIO, CDSC, DCSS, Library
7.	Establish communication, education, marketing , and training efforts	March 2012	Continuing	VI., VII.	All	CIO, CDSC, DCSS
8.	Implement a pilot of federated storage	March 2012	December 2012	II.	All	DCSS, CITES and several major units on campus

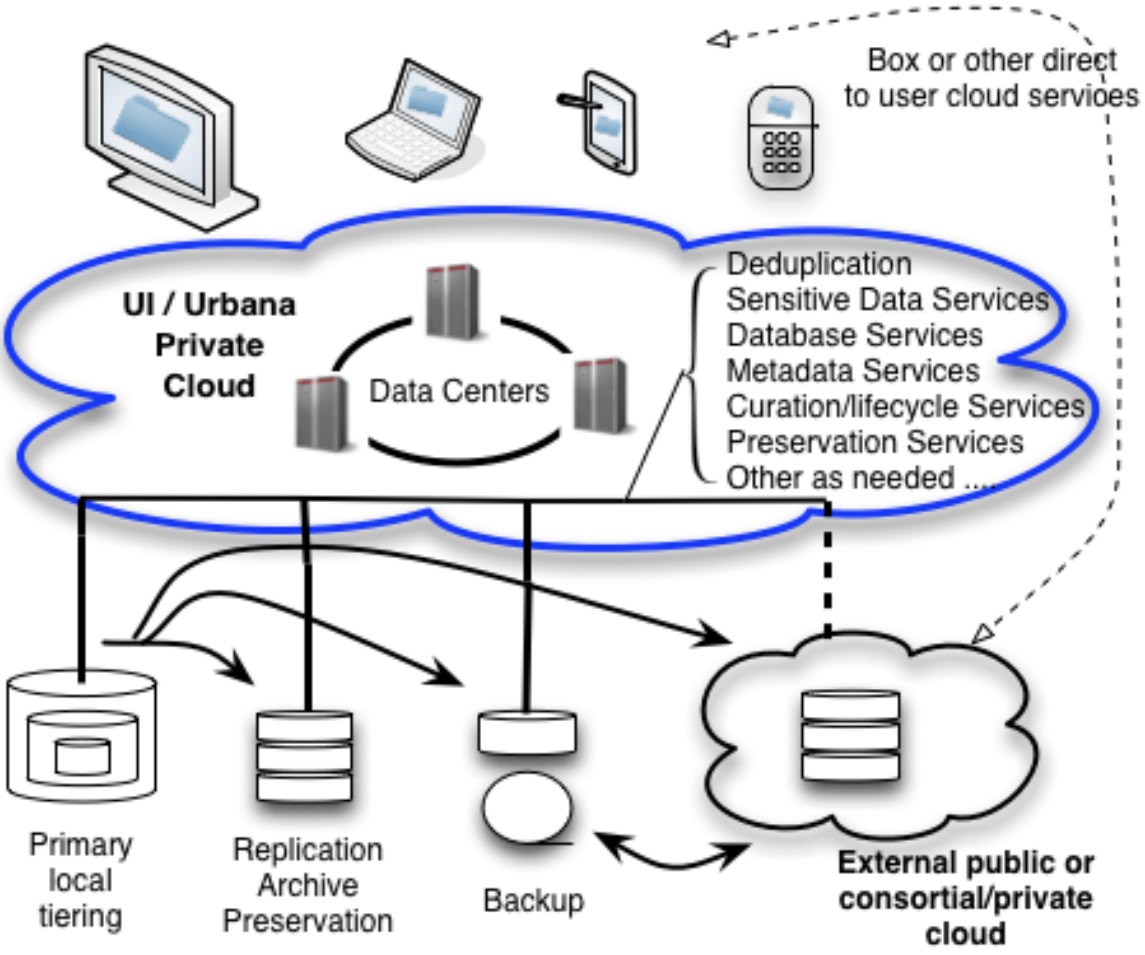
9.	Implement a pilot of central tiered storage for units	May 2012	December 2012	III., IV.	Architecture, Institutional, Research	DCSS, AITS, CITES
10.	Create large research data storage service proposal	April 2012	June 2012	III., IV.	Architecture, Research	DCSS, OVCR, NCSA, CIO, Library
11.	Establish cloud technical & contracting expertise	April 2012	Continuing	IV., V., VII.	All	CIO, DCSS
12.	Identify subsidized backup, replication & archiving services	August 2012	December 2012	VI.	All	DCSS
13.	Develop central tiered storage capability to address unit and "common good" needs	Fall 2012	1 – 2 years	III., IV.	Architecture, Institutional, Research	DCSS, CITES, AITS
14.	Establish a new Data Security Compliance role	2013	1 year	IV., VI., VII.	Sensitive data	CIO, OVCR, CISO
15.	Define and create value-added storage services such as database management, curation, etc.	2012	~ 2 years, continuing after that	IV., VI., VII.	All	DCSS, CDSC

Acronym key: CDSC: Campus Data Stewardship Committee; CISO: Chief Information Security Officer; DBS: Division of Biomedical Sciences; DCSS: Data Center Shared Services; EGCDCC: Executive Governance Committee for Data Center Consolidation; OVCR: Office of the Vice Chancellor for Research

The Task Force views this work as an evolutionary, not a precipitous set of activities. Because of the campus' decentralized approach to storage and their related services, some units are adequately provisioned for the next several years. By contrast, some units do not currently have substantial storage capabilities, and others may never achieve the required types of storage and services necessary to support the numerous types of work carried out in the unit. For a successful centralized, "common good" storage model, buy-in ought to be established incrementally on campus, through opportunities with units, over time.

The Task Force believes that the risk to the University is too great to ***not*** have a forward-looking strategy and an effective set of storage services. The investigation of the Task Force has led us to conclude that there are tangible risks and costs, both financial and legal, that are related to putting off needed storage planning and implementation. These include unintentional exposure of personal information and other sensitive (e.g. HIPAA) data, or failing to provide effective management of research data. Well-stewarded data is at the heart of maximizing our research, teaching, learning, and public outreach impact. It supports reproducible research, reusable data enabling new research, and informed administrative decision making and assessment. It is a necessity in order to satisfy records retention rules, research funding agency expectations, and business continuity and disaster recovery needs. Now is the time to move forward with a cohesive storage strategy and set of services that not only minimizes risk and maximizes value, but also effectively provides our campus a competitive advantage in addressing its mission. Our vision for the data storage services future that can yield these advantages is encapsulated in the following diagram:

Future Storage Services Vision for the University



* Note: Variation of a diagram from a presentation shared with us by Dell

I. Introduction and Task Force charge

The Data Storage Services Task Force was appointed by Paula Hixson, Interim Chief Information Officer for the Urbana campus at the end of July 2011 with a charge letter included in Appendix A. The primary goal of the Task Force, as stated in that letter, was to "develop a comprehensive strategic plan for addressing the central data storage needs of this campus, including specific operational ideas for implementation". The Task Force was asked to carry out its work expeditiously, delivering an interim report by the beginning of October 2011, with an initial due date of the end of 2011 for a full report. In October it was determined that the Task Force work required additional time, and the final report due date was extended to March 9, 2012. The membership, process, and activities of the Task Force are identified and described in Appendixes B and C of this report. The elements of the charge are summarized and grouped in categories below:

Overall:

- Develop a strategic plan (include strategy, tactic and operational recommendations) to support the central data storage needs of the campus, including specific operational ideas for implementation;
- Recommend a cohesive set of storage services that efficiently and effectively meets the data, storage, backup, and data management needs of faculty, staff, and students.

Process:

- Study what other peer institutions are doing in this same area. Evaluate whether any of the models being pursued elsewhere would work well here, and if so, whether it would be possible or wise for us to simply adapt/adopt that model;
- Survey the campus landscape and identify existing data storage;
- Consult with other campus committees dealing with related data services issues, including the Executive Governance Committee for Data Center Consolidation, the campus Data Stewardship Committee, the new campus community cluster effort, the Media Commons project, and the Center for Media Excellence, the IT Professional community, and the IT Council.

Needs assessment included identification of storage needs across a varied community:

- Common needs;
- Unmet needs on this campus (unmet needs articulated by end-users, as well as those that are currently identified only by the expert community, but not yet recognized by the end-user community);
- Different constituencies (faculty, staff, and students) as well as institutional and unit needs;
- Researchers, including funding agency requirements for data stewardship and rapid growth of the volume of research data;
- Value-added services (such as database design or hosting, data archiving, data

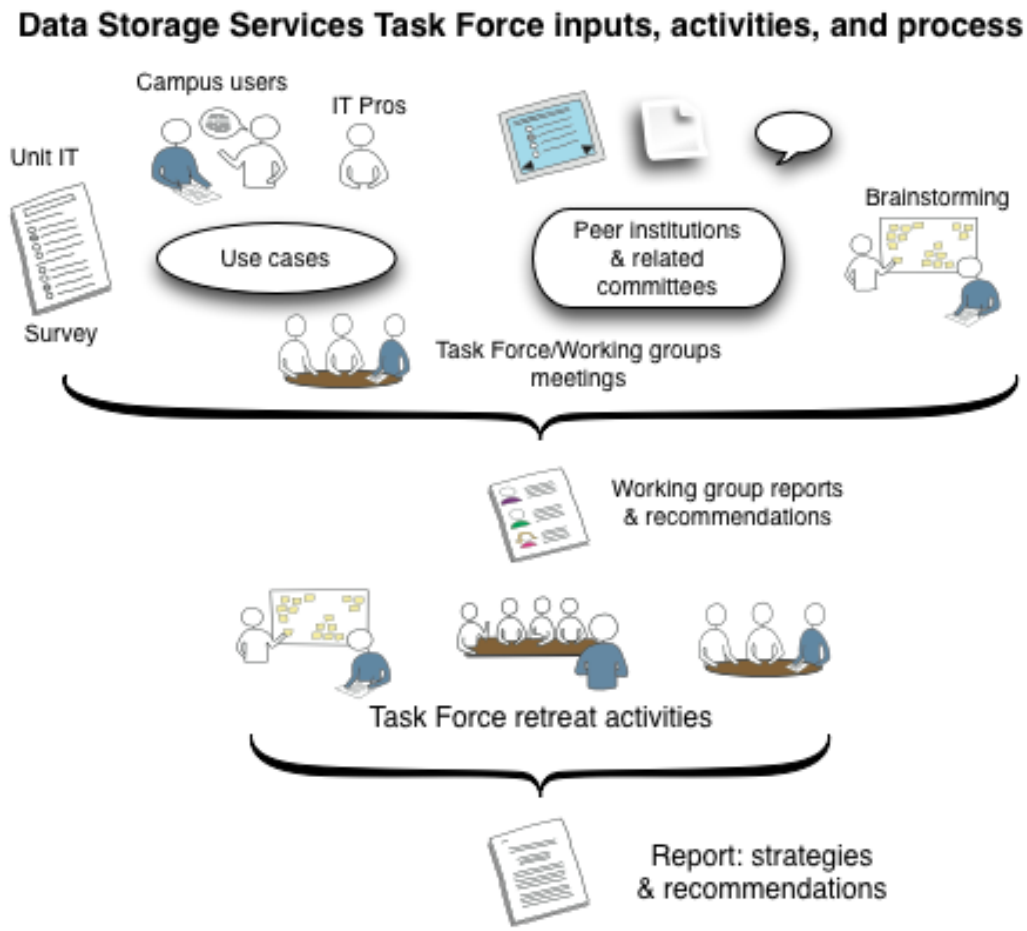
curation, data backup, etc.) to whatever physical storage solutions are recommended.

Partnerships:

- Identify potential collaborations and partnerships that have strong potential to help the Urbana campus address its needs;
- Articulate the role of cloud storage in the campus storage strategy;
- Reach out to colleagues at UIC and UIS and invite them to participate.

The co-chairs of the Task Force did contact the CIOs at UIC and UIS to see if either or both campuses were interested in participating actively in the Task Force. Both UIC and UIS expressed interest in this effort, but indicated they were not in a position to participate at the time. Some limited input has been received from the UIC campus. The Task Force does feel confident that much of this work may have application for the UIC and UIS campuses, and that many of the proposed strategies and recommendations have the potential to be extended to both UIC and UIS.

The following graphic summarizes the Task Force's work; a more detailed account of how the Task Force executed its charge is in Appendix C.



II. Review of peer institution storage strategies and services

In its charge letter, the Task Force was asked to review what types of storage models peer institutions offer, and to analyze the suitability of those models to meet the Urbana campus storage needs. The Task Force identified a variety of peer institutions with informative materials around their current and proposed storage services and strategies, with resources from Stanford University, the University of Texas at Austin, Iowa State University, and several peer CIC institutions (Indiana, Northwestern) being particularly useful. Additional useful resources were identified at the University of California, Berkeley, the University of Virginia, and the University of Iowa. (See Appendix D: "Storage at Peer Institutions" for links to a number of peer institution resources.)

The Task Force has also had the timely opportunity to leverage the conversations and the work of a CIC-wide Data Storage Working Group (focused on research data), which was formed in January 2011 to address issues related to research cyberinfrastructure. Three members of our Task Force (Grady, Guenther, Namachchivaya) represent Illinois on that CIC-wide group, which has provided key additional insight as to storage services, plans and strategies at our peer CIC institutions.

While we did not identify an existing storage model that addressed all aspects of storage services that the charge articulated, the Task Force did note that a number of institutions are planning services and strategies that are similar to those under investigation for the Urbana campus. The work of the Task Force to identify storage service needs across the campus, aligns closely with the services, plans and strategies of these peer institutions. Key points worth highlighting that are consistent across all or a number of these peers:

- Tiered storage offerings are critical (i.e., storage that supports file system, block, and other types of needs of varying speeds and qualities);
- "Common good" storage is provided by many institutions, including file services layered over such. 77% of CIC institutions are providing such today;
- A "common good" centrally provided storage service focused on research data is a growing trend, with such services now having a "leading edge" of 50 GB of disk and 1 TB or more of archive storage;
- Backup and archiving services are critical, but survey feedback and followup conversations suggest that the campus needs more than one backup strategy and related services, depending on the content that is being backed up or archived;
- Storage services that enable collaboration with colleagues outside the institution are important;
- Central storage pools, versus storage scattered across the institution, provide economic, scalable, and secure, advantages that address data lifecycle needs. This storage must provide services at several levels, including file system access (e.g. CIFS, NFS) and block-level access;

- Cloud storage needs to be a part of the storage strategy going forward, and can be the right choice in some circumstances today;
- Defining and supporting data retention lifecycles is critical;
- Dealing effectively with sensitive data is critical;
- Storage must be easy to use or it won't be used.

See Appendix D for links to resources at the above-mentioned institutions, along with a summary of focus group input at the University of Texas at Austin on storage service needs from an end user perspective.

III. Baseline survey findings

A campus-wide survey was designed to gather and report information about several aspects of storage across a variety of campus units. The purposes of the survey were to gather baseline information on current storage practices and services on campus as well as utilized external services, to identify current unmet needs, to identify future anticipated needs, and to identify the potential for more effective ways to support unit-level data storage services. The survey asked questions that elicited responses in the following categories:

- Amount of storage, access methods, and related services currently in use
- Storage services and practice
- Funding and fee model
- Storage strategy, unmet needs, and value-added services

There were 43 survey responses returned (more than 80% of those distributed), on behalf of most of the colleges, instructional units, research centers, institutes, administrative units, and auxiliary units within the campus. Several of these responses represented multiple units, and depending on exactly how one counts, the responses represented more than 50 units. The responses were rolled up into 34 units (there were multiple departmental responses within two colleges) for broad analysis. (Appendix E lists all the units responding.) We estimate that the survey responses represent over 80% of faculty, staff and students on campus, based on an examination of DMI numbers for the responding units. The twenty-one questions on the survey yielded over 45 data points for analysis, some of which were closed-end responses, and some of which were open-end text responses. The survey revealed the following baseline information about the current use of storage on campus:

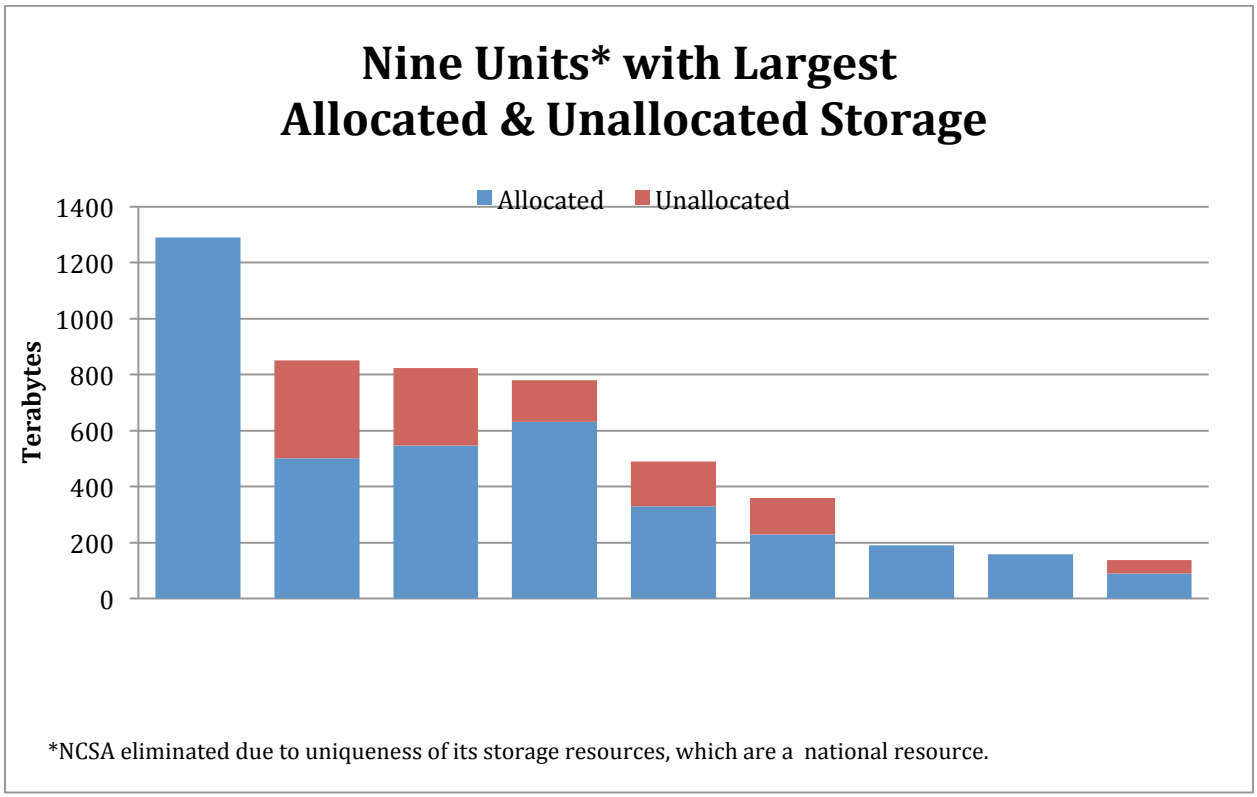
- Quotas: few units reported that they utilize storage quotas for any user group (fewer than 10 units use specific quota guidelines);
- Standard allocations: a number of units provide standard allocations for staff, faculty, and graduate students, ranging from one gigabyte (1 GB) on the low end to more than a terabyte (> 1 TB) on the high end (fewer than 3 units reported that they provide more than a terabyte of storage for individuals or groups)/
- Storage access methods: Windows-based storage access (CIFS) appears to be supported by more units than other storage access methods; this suggests that many units support a Windows-based storage infrastructure;
- Staff: Most units allocate some staff to the management of storage and access—ranging from a minimum of .25 FTE to a maximum of 3 FTE;
- Amount of storage: roughly seven petabytes (7 PB) of storage are supported in aggregate by the units that responded to the survey. Of that total, 4.8 PB of storage are currently in use, and 2.4 PB is designated as "unallocated." [1] This suggests that about 34% of the total storage infrastructure that is maintained across these campus units

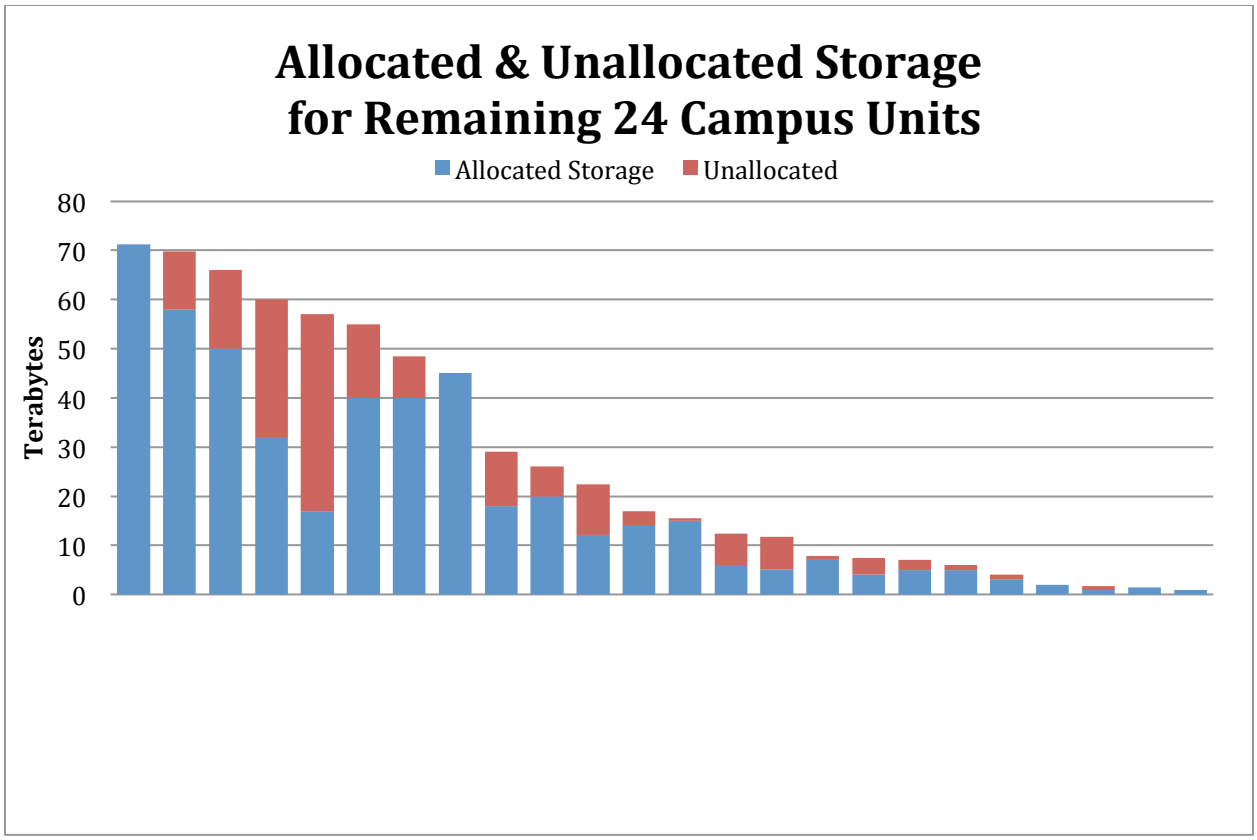
[1] These figures do not include the NCSA storage, which is mostly a national resource.

may not be actively allocated. There may be a number of reasons for this, including anticipated use in the near future;

- Ten units support over 90% of the storage that is currently in use by the units reporting in the survey. These same ten units support over 95% of the unallocated storage currently managed across the units reporting. The top ten storage supporters (and users) include a mix of central units like CITES, AITS, and the University Library, and colleges or research units that have a significant investment in research, such as Engineering, LAS, ACES, NCSA, IGB, and Beckman.

The following two charts show the reported allocated and unallocated storage for each of the 34 units used for analysis, except for NCSA, where most of the storage is a national resource, not a campus resource. The first chart shows the numbers for the top 9 units, and the second chart has the other 24 units – these were split this way because of the large difference in the quantity of storage managed between these sets of units. The unit names aren't listed because they aren't important to seeing the general pattern, and because we indicated that we'd keep survey responses confidential.





The top priorities and unmet needs for storage services that the survey responses highlight are:

- Increasing need for storage capacity, and in many cases, the rate of need for growth is accelerating;
- Affordable backup services, archiving, and the ability to recover data in a timely manner for business continuity and disaster recovery are critical and frequently unmet (or at least not well-met) needs;
- Easy ways to effectively share and collaborate on data are needed, both within the campus and with users and groups external to the campus. The need for "Dropbox-like" services was often noted, both for sharing and synchronizing data;
- Remote accessibility to data (including from mobile devices) is needed and not well-met today;
- "Cloud storage" was frequently mentioned, not so much as being used today, but as something many units are speculating may soon be part of the solution to their storage needs.

The most frequently mentioned obstacles to meeting the storage service needs of units and the individuals they serve today are:

- Funding/cost—lack of recurring funds;
- Staffing/expertise—insufficient staff or lack of necessary staff expertise;
- Lack of necessary services/features/ease of use;
- Education about what is available, and how to appropriately use it.

The survey responses support the idea that units are quite willing to use storage services that are affordable (from their perspective), easy to use, and meet the needs of the unit and the individuals they support, regardless of who provides those services. Some units are finding that some central services such as backups, file services (often in conjunction with virtual servers), and SAN services meet at least some of their needs today. Many more units indicated that they do not take advantage of current central services because the services do not meet their needs at an affordable price, and/or they do not provide the needed services in an easy-to-use and effective enough form.

One key point highlighted in the survey responses was that campus units purchase significantly more storage than they need due to several arbitrary budgetary limitations:

- Expecting the need for storage to increase, and wanting to be ready in case of a sudden significant new need (generally related to research);
- Uncertain/infrequent unit funding available for storage, which results in units "over-spending" to ensure that they meet some future, unspecified need.
- In most units, the implementation, migration, and management of storage is considered a component of the job (<50%); Most IT professionals seek to minimize the number of times they install, and configure storage, to reduce overall time devoted to storage management.

Given those factors, "overprovisioning" storage is commonplace on the Urbana campus. As noted earlier in this report, the cost of storage per terabyte is decreasing at an average rate of 15% per year (a conservative estimate from industry reports and some data shared by NCSA). Storage purchased today that sits unused for several years costs significantly more per terabyte than purchasing storage "just in time," although it does guarantee the certainty that a unit has some type of storage available. To minimize cost, units would ideally follow a "thin provisioning" strategy, obtaining storage as needed. A shared pool of unallocated storage that was available on a reliable, short turnaround basis to meet sudden significant new needs, could accommodate a high percentage of storage needs for campus units. It would also enable a cost-efficient purchase and the thin provisioning strategy.

The Task Force recommendation that the campus invest in a shared pool of unallocated storage is a key driver for several of its recommendations. The recommendation that the campus invest in federating existing storage resources across large data center nodes is a "quicker win" to enable immediate sharing of existing pools of storage across units. The added advantage of federating storage is that it allows easy access to data where and when a user needs it, and it reduces the number of times a user is required to copy files from one storage resource to

another. One example that promises to become widespread is that of a researcher who generates a large data set within IGB, but needs to perform computation on that data using the Campus Cluster[2]. If the data can be accessed and delivered quickly across the network, the researcher can avoid the need to duplicate the data by copying to the Campus Cluster storage.

The longer-term strategy is to centralize storage resources, which will naturally allow for a shared pool of unallocated storage. That also would provide a single point (or at least fewer points) where cloud storage resources (whether commercial or in partnerships with other academic institutions and organizations) can be "linked in" and provide an easy way to quickly expand storage capacity.

Another key point revealed by the survey is the high amount of effort devoted to storage management across campus, regardless of the amount of storage to be managed by an individual unit. The availability of central storage options can reduce unit time spent on managing basic storage, and enable units to shift that effort to the support of higher-value activities, such as assisting researchers with data management planning. Evidence from recent studies, including a recent CIC storage staffing survey, strongly suggest that the FTE necessary to manage storage is more closely tied to the number of distinct storage services managed, and not as closely tied to the amount of storage managed. A reduction in the total number of distinct storage services across the campus has the potential to significantly reduce the total effort required to manage storage across the campus.

There are a number of additional survey findings noted by the Task Force in Appendix F.

Finally, based on the evaluation of the survey and its responses, the Task Force recommends that a follow up survey be administered on a regular basis, similar to the annual Campus Computer Room Inventory, to gauge storage needs as well as the impact of any changes that are implemented as a result of the Task Force recommendations. Potential areas for deeper exploration are the types of storage technologies deployed and storage devoted to CIFS service delivery (because of the great potential for functional and cost efficiency in pooling this type of storage).

[2] https://campuscluster.illinois.edu/

IV. Needs Based on Analysis of Topical Working Groups

A. Analysis, method, and use cases

In the late fall of 2011 the Task Force created four working groups to examine storage and service needs related to key areas called out in the Task Force's charge: research, institutional (including instruction and administrative), and two overarching topics—sensitive data, and storage architecture. A specific charge was written for each working group, and the working groups were chaired by members of the core Task Force. Membership included a mix of IT professionals, faculty, administrators, and librarians. The charge for each working group, and more information and artifacts produced by the group, can be found on the Data Storage Services Task Force wiki space, under the "Working Groups" heading.

Each of the working groups had informal conversations with a number of individuals and groups that are engaged in activities that require the use of storage and related services. In order to document the types of work and research practices that require storage and related services, three of the four working groups—research, institutional, and sensitive data—developed a set of use cases. The use cases are located on the Task Force wiki space[3], under the "Use Cases" heading. The working groups integrated their analysis of the use cases as they formulated their respective baseline perspectives on the storage and related service needs to support different areas of activity in the campus community. The issues discussed by the sensitive data working group overlapped with both the institutional and the research working groups. The Task Force took care to involve representatives from the sensitive data working group in the conversations and recommendations of both the institutional and the research working groups. The architecture working group's work depended heavily on the recommendations of the three other working groups.

The activities of the core Task Force and the Working Groups culminated in a mini-retreat, held at the end of January 2012. At the retreat, each working group made a presentation that articulated storage needs for research, administration, instruction, sensitive data, and architecture. Over 25 participants developed recommendations for storage and related services based on the reports of the working groups, which the Task Force believes accurately represent needs of a substantial number of individuals, units, and groups on campus.

B. Research and Storage as a part of the Research Lifecycle

The challenges presented by the campus research enterprise illustrate the broad scope and dimensions presented in establishing an effective data storage program. Indeed,

[3] https://wiki.cites.uiuc.edu/wiki/display/DSST

the very strengths that distinguish the Urbana campus research programs extend the requirements and conditions of data storage provisions and services.

The Research Storage Working Group examined a wide range of issues associated with data storage. A total of 15 use cases were developed and discussed in terms of types and formats of research data, diversity of storage needs across disciplines, expectations of researchers, and federal agency requirements.

The Working Group utilized a broad array of inputs to provide background and context to their analysis. Included were direct interviews with researchers, knowledge of storage operations within their own units, and related initiatives currently under way on the Urbana campus and in peer institutions, as well as external groups, such as the CIC, Association of Research Libraries, and the GSLIS CIRSS Center. The Working Group developed a list of primary storage needs and services that are important to effectively supporting research on campus, and that list can be found in Appendix G.

The Working Group also reviewed a summary prepared by Bill Mischo of Data Management Plans (DMPs) submitted from the Urbana campus to the NSF since January 2011. This overview was useful for compiling the present data storage practices and provisions currently implemented by the Urbana campus researchers. A summary of information gleaned from this analysis of DMPs is included in Appendix H: Summary of NSF data management plans prepared by researchers on the Urbana campus, January – November 2011.

The current storage solutions are not optimal, but researchers are very resourceful in utilizing storage options and capacity to meet their needs. Compliance, in particular with federal regulations around privacy and confidentiality, is an issue in some cases. It's clear that more storage space is a general concern and the current solutions aren't meeting those requirements. Many other needs for storage services are not effectively addressed at even the most basic levels. Overall needs articulated by the group include:

- Reliable, long-term (5 years +) managed storage for high capacity datasets;
- Ease of use—enable researchers to easily move data from HPC and other computing analysis environments (like the campus research computing cluster) to easily-identified storage environments that are suited to the data needs
- Ability to create and manage pools of storage across and among campus units now;
- Collaboration tools to enable sharing of research data among individuals/groups
- Storage-related services:
    - Tiered storage to support a variety of research needs, including sensitive data (HIPAA, classified data, etc.), high and low availability;
    - Backup, replication, and archiving capabilities that meet the needs of the research;

- Research data consulting services, including support for the development of research data management plans in conjunction with grant and other data collection projects, databases, curation, migration, and preservation;
- Data archiving
- Access to the data by other researchers

The optimized approaches will require cooperation and coordination across multiple administrative and academic units. While the storage infrastructure is important, the issues surrounding long -term preservation and the curation of datasets necessitates further development of policies, training, and education. Resources will be an issue and development of an effective business model is essential. User buy-in poses a huge challenge, but an important area that will be greatly facilitated by regular communication at all levels.

C. Workplace Productivity, Instruction, and Institutional Assets

The Working Group on Workplace productivity, instruction, and institutional assets considered diverse storage and services needs across several audiences within the University, excluding research. The needs of faculty, staff, and students were considered by looking at administrative support, workplace productivity, learning and instruction, and other institutional assets where storage is a primary concern.

Administrative support functions include those related to human resources, finance, and student systems; as well as college and unit level support of myriad users for daily operational business. This encompasses storage of raw data, enterprise systems data, operational and managerial reports, local databases and reports, communication services (both internal and external; web accessible and mobile delivery methods; etc.), support of marketing services, and myriad sharing and backup services.

Because the scope of this working group's charge included a number of critical and distinct activities (instruction, communication, administration, financial, cultural), the working group identified a common, cross-cutting set of functions as a basis for analyzing the different use cases and needs. The most common functions that have an impact on how and where data is stored include:

- Creation (workplace productivity tools, data entry methods, etc.)
- Management (individual) (drafts, working docs, etc.)
- Validation (check sums, access log files, system audits, etc.)
- Aggregation and analysis (software tools, batch processing, etc.)
- Display and re-use (web content, publications, proprietary applications, etc.)
- Transmission (e-mail attachments, web forms, ftp, etc.)

In addition to varying roles and workplace environments mentioned above, the function and type of data, as well as the period of period of time over which data may be needed or required to be retained can vary greatly. All of these factors impact the data storage needs.

The working group identified the following core storage needs that cut across the areas of instruction, administration, and institutional stewardship:

- Ability for units and groups to create and manage pools of storage across and among campus units (federated storage);
- Centrally-managed storage pools that support file-system storage, with services built around them that facilitate easily finding and using files, and support institutional security requirements, ease of access from remote locations. This should address the common needs across many units for functions such as document storage that do not require high performance disk. This storage should support administrative, business, instructional, and other related institutional needs.
- Central storage pools that support administrative and other needs for storing, accessing, and managing data stored in databases--virtual machine style data. Users of this type of data require services beyond simple storage such as data warehouses, databases, and virtualization.
- Campus-wide pool of multimedia storage, requiring varying styles of storage speeds to support media in current use, as well as historical media that is in the "archive" or "preservation" state. Fast access to this storage is required for multi-media professionals to manipulate and edit current use files that are current. The appropriate version of these files would be on storage with services for streaming of that media to the public or students through the LMS.
- Digital preservation storage at both the file system and the preservation system levels, for administrative, institutional, and educational assets.

D. Sensitive Data, Security and Privacy

The Sensitive Data Working Group's examination revealed that there is more activity involving the use of sensitive data on campus than initially anticipated. The scope of sensitive data encompasses virtually all areas of campus activity, including research, administrative, and instruction. That scope has expanded considerably when one considers: FERPA protected data (i.e. student data), classified projects, export control issues, genomic data, and other sensitive human subjects research studies. While the campus research community does engage in "clinical" studies, there is a widely-held perception that "clinical" data doesn't exist on campus. However, activities focused on human subjects permeate our research environment, and as a widely recognized center of supercomputing activities, our reputation only serves to attract increased bioinformatics activity. More to the point, while much of this research may fall out of

the strictly “clinical” realm, it may often include highly sensitive information on individuals.

Recommendations of this Working Group include:

- Launch a pilot project providing secure storage for select human subject researchers with highly sensitive data as an immediate interim-only solution. Simultaneously, research, propose and develop the facilities, be it local or outsourced or cloud-sourced, to support all campus researchers working with highly sensitive data, human subjects or otherwise. Develop a long-range plan with a flexible funding model and policies for this facility to ensure institutional compliance as a hybrid entity;
- Establish a new role on campus of Data Security Compliance with dual reports to the OVCR and the Chief Information Security Officer. This role would be responsible for:
    - Tracking and interpreting the relevant regulations around the security of sensitive data,
    - Disseminating that information to the parties that need to be aware of the regulations,
    - Improving the compliance of departmental IT professionals with security breach protocols and reporting, and
    - Developing non-intrusive processes for monitoring of human subjects studies, clinical and otherwise. (Are records being stored properly? Are data access protocols being followed? Are data access audit trails, sometimes called “accounting,” being reviewed? Are people being consented properly? Is correct version of research protocol and Informed Consent being used at all times after the first approved Protocol Amendment?);
- Provide a centrally-managed sensitive data storage option for administrative data.

E. Storage Architecture, Technology, Delivery, and Cost Models

The Storage Architecture Working Group's plans depended largely on the recommendations of the other working groups. The Urbana campus does not have a robust central storage infrastructure, owing to the fact that storage has been developed by individual units. Findings from the survey and the use cases suggest that the campus ought to invest in the infrastructure to enable the federation of storage pools that are currently supported by academic and research units. The goal here is to allow for more flexible management of existing storage investments.

The review of peer institutions reveals that most of Illinois' peers support robust central storage facilities for administrative, instructional, and research needs. The benefits to centralized storage are considerable, including cost efficiencies in volume purchases, ability to deploy flexibly, and the potential for fewer staff dedicated to managing centralized storage. Although Illinois' peers have constructed and provisioned substantial central investments, the timing of Illinois' entrée into the storage arena and the availability of cloud storage may allow our institution to develop an architecture that is somewhat different from its peers.

At the minimum, however, it is clear that there is a strong need for the provisioning of several kinds of centralized storage—file-system, easily accessible storage for administrative, instruction, and research use; block storage to support storing large files such as media; and preservation storage. And a centralized, consistent storage service would provide a simpler base upon which to layer value-added services such as database hosting and management, metadata and curation services, automated lifecycle management, preservation services, etc.

While recent industry figures suggest that most data stored are never accessed again, many of us do not know when we store information in a file whether or not we will need it again. Data are stored based on the assumption that they may be needed in the future. Until recently storage has been viewed as a separate and distinct component of the scope of resources required to support overall productivity. Now storage is viewed as an integral component of the productivity picture—it is a critical component of the lifecycle of data.

Technology advances of the past several years support virtualization of software and hardware. Further, the use of networking extends access to geographically dispersed infrastructure. What the Task Force found is that it is redundant at a low level for campus units to replicate numerous instances of the same kind of storage. Many units remain at a disadvantaged position because they must support significant storage startup costs in order to offer basic storage services. Each unit that supports storage must continuously invest in the equipment, software, and personnel to refresh and manage storage.

A document describing an internal storage architecture for the campus, and the various storage technologies and options that the architecture could encompass and provide, and the tradeoffs involved in choosing from amongst those various options, is included in Appendix I: Storage Architecture.

V. Opportunities for collaboration

The Task Force was also specifically charged with identifying potential collaborations and partnerships that have strong potential to help the Urbana campus address its storage service needs. The collaboration and partnership opportunities exist both within the Urbana campus, across the University as a whole, and with a growing number of potential external partners: institutions, consortiums, non-profit and commercial organizations, etc.

The first thing that is important to highlight is that the recommendations of this Task Force align with, and are complementary to, the work of the following campus committees and efforts:

- Data Center Consolidation: centralizing storage services with a focus on a substantial shift of file services, shared storage pools (e.g. SAN), and backup and archiving services into the consolidated data centers while using virtual servers where possible will accelerate the consolidation of data center space on campus.
- Governance: another key opportunity is to consider transforming the Executive Governance Committee for Data Center Consolidation into a broader "shared services" governance body that also deals with storage, server services, etc. This is one of the Task Force recommendations.
- Campus Data Stewardship Committee: the recommendations involving research data storage services can have a profound impact on providing an effective infrastructure and services for research data management, support, and stewardship, and thus directly align with a primary goal of the campus Data Stewardship Committee and its Illinois Research Data Initiative. Additionally, the University Library is doing active work around data management and stewardship, and should be a partner in activities around developing best practices for data management, lifecycle, preservation, etc.
- Campus Cluster initiative: the recommendation around exploring a large research dataset service that complements the Campus Cluster effort directly aligns with that effort.
- E-Science and Cyberinfrastructure: Illinois' participation in the ARL E-Science Institute effort, and the campus Cyberinfrastructure Master Planning activities.
- Center for Multimedia Excellence (CME): the recommendations of the Task Force to establish an effective set of storage services which can be used to store and manage multimedia digital objects aligns with a primary campus need identified by CME.

There are also a large number of potential collaborations and partnerships to be explored with external partners. An incomplete list of some of the possibilities that we can begin to leverage today, continue to explore, and/or establish ongoing tracking of are:

- Joint solutions with the UIC & UIS campuses. Several initial touchstones are around HIPAA and other sensitive data storage services, and research data storage related to

the Campus Cluster effort (which is now in the process of specifying how researchers at the other UI campuses can participate in the Campus Cluster).

- Box pilot effort that has already commenced at the University of Illinois, as part of the broader Internet NET+ pilot Box service offering.
- Other Internet2 NET+ services as the portfolio of such expand.
- Peer institutions: there are possibilities to be explored both within the broader scope of the CIC-wide Data Storage Working Group, and potentially with specific partner CIC institutions such as Indiana University. (Indiana has already been asked by another institution to provide a pricing model for providing a "bit level" storage service.)
- Various repository efforts around the nation today. The University of Minnesota Digital Library has provided a report they commissioned to compare the pricing and services of a number of repository services that could be used for preservation, including UC3/Merritt, HathiTrust, MetaArchive, etc. The CIC Data Storage Working Group is also compiling a list of repository services and cloud storage services specifically targeted at researchers and research data.
- Federal agency funded efforts such as the set of NSF DataNet projects and the data repository services and models those projects are creating.
- DuraSpace.org and its DuraCloud service.
- XSEDE and its campus bridging initiatives such as GFFS.

These are just some of the possible partnerships, services, or collaborations that can be explored and pursued as we go forward in executing on our storage strategies. The above possibilities span the gamut from "raw storage" (with various cloud vendors being available in the future through the Internet2 NET+ services, and/or partnerships with specific institutions or consortiums) to specialized research data repositories, from initial point of data creation to long lasting archives, and from commercial offerings to non-profit organizations and universities. In order to fully understand and take advantage of these opportunities as appropriate, the campus storage services governance group will want to work with the Office of the CIO, the University Library, the campus Data Stewardship Committee and others to ensure the campus continues to closely follow, participate in, and assess the economic and service advantages these various service offerings and/or partnerships can provide.

VI. Recommendations, and related challenges and opportunities

The Task Force, based on all of the inputs and work described above, has formulated a storage strategy composed of the seven key themes described in the Executive Summary at the beginning of this report. Based on those themes, the Working Group reports, and the broad input the group has synthesized, the Task Force makes the following fifteen (15) specific recommendations that we believe will address all of the storage services strategic themes. These recommendations are presented in summary form in the Executive Summary, but are presented here with a fuller description. Some of these recommendations can be achieved relatively quickly, and others will take several years to fully implement. And several will have an initial deliverable, but be an ongoing process.

In this section we highlight the challenges that need to be addressed in order to achieve the targeted benefits that a cohesive strategy and specific actions are intended to accomplish. The Task Force has targeted specific opportunities that could hasten the broad acceptance of this storage services strategy and adoption of best practices and centrally managed storage services.

A. Recommendations

1. **Define a lead role in the Office of the CIO**.
   Articulate the program lead role in the Office of the CIO that is responsible for spearheading efforts around campus-wide data storage and services, and align this role with the efforts of a transformed Executive Governance Committee for Data Center Consolidation (the new Data Center Shared Services (DCSS) Group). There needs to a "program officer" with the responsibility to see that this entire portfolio of recommendations and the pilots, groups, efforts, etc. that should get launched indeed do get successfully started and make effective progress.

2. **Establish governance of campus storage services within a Data Center Shared Services (DCSS) committee.**
   Consider transforming the Executive Governance Committee for Data Center Consolidation into a "Data Center Shared Services (DCSS) Group", with storage services as part of its overall planning and operational portfolio, and articulate how this relates to the campus IT governance structure. Effective governance that is clearly seen to have broad campus representation and scope will be a key to ensuring both that storage services meet the needs of campus users and units, and that those services are broadly accepted and utilized.

3. **Implement Box storage service for collaboration and file syncing across devices**.
   Implement the Box collaboration/file storage accounts for the Urbana campus. There is a project that has already commenced to provide a University-wide service that provides a Box enterprise account for all faculty, staff and students. Box is an

"enterprise-grade" service that is similar to other commercial services like Dropbox, and provides an easy way to share files with collaborators and to synchronize files between various devices. Box provides institutional (and end user) advantages in having rich and flexible access control, encryption capabilities, and a growing number of integrations to an ecosystem of related services. Box is one of the first services being offered through the new Internet2 NET+ program[4] referred to in Section V (collaboration opportunities) of this report. Box addresses two needs expressed by all types of users and units – 1) an effective way to share files with collaborators whether across or beyond (e.g. at other institutions) the campus; and 2) a way to access and manage files across a variety of devices. Box does not, however, meet the requirement for "common good" storage – the kind of storage that supports the administrative work of units and staff within those units, faculty in their research, education, and service roles, and students for their instructional and research needs. It is not an appropriate general backup solution for desktops, laptops and other devices.

4. **Create sensitive data service pilot proposal.**
   Create a proposal for a pilot project that provides secure storage for select human subject researchers with highly sensitive data as an immediate and temporary solution, and present to campus administration for funding. The Sensitive Data Working Group along with the Research Working Group identified a critical storage service need – a secure storage service for the data collected, generated and analyzed by human subject researchers. An *ad-hoc* group has just formed under the umbrella of the Campus Data Stewardship Committee to address secure research data storage and management issues. The group includes representatives of multiple units (most identified in the Recommendations Table), to scope out and create a proposal for a pilot sensitive data service that will involve several researchers currently doing human subject research. This pilot will address an urgent current need and provide knowledge and experience that will be an input to the following Recommendation 5.

5. **Establish policy and solutions for sensitive research data management.**
   Set up a campus-level group that formulates policy related to the storage and management of sensitive research data, and identifies best practices and solutions for implementation (joint between the Office of the Vice Chancellor for Research and the Office of Security and Privacy). The pilot project described in Recommendation 4 can address an immediate need, but a focused and sustained effort by the campus is required to determine the policy needs, practices and solutions that must be put in place to effectively support the needs of researchers who work with sensitive data. The longer-term work ought to commence before the pilot project is finished. This group should work with the broader storage service best practices and policies to incorporate sensitive data policies, practices, and solutions into the long-term solutions, and to create and ensure there are workshops and other educational efforts focused on the management of sensitive data.

[4] http://www.internet2.edu/netplus/

Note that in using the term "sensitive data", we are specifically referring to data where there are regulatory and legal obligations concerning the handling and access to the data. Many researchers will consider their data to be "sensitive", not wanting other than a select few to be able to access that data, until at least after their research results drawn from that data are published (or at least are in the "pipeline" for publishing). Most of the rest of these recommendations apply to the storage services, policies and best practices that are appropriate for all research data not covered by regulatory and legal obligations.

6. **Develop best practices and policies, and a recommended solution for data management.**
   Develop best practices and policies, and recommended solutions, for storage and data management for both non-sensitive research data and data associated with administrative support, workplace productivity, learning and instruction, and other institutional assets, leveraging efforts already underway by the campus Data Stewardship effort and the University Library, and peer efforts such as within the CIC Data Storage Working Group. This should include focused work on defining appropriate data lifecycles and retention periods. Because issues related to research data can be fundamentally different than those related to institutional data due to issues related to publication, openness, et cetera, this may be two complementary efforts. Sensitive data handling will still be a concern in this effort, as there are also regulatory and legal obligations around some types of institutional data, such as data covered under FERPA, personally identifying information (PII), etc. There will potentially be some overlap in this effort with the sensitive research data effort, and these three efforts should collaborate and combine results as appropriate.

7. **Establish communication, education, marketing , and training efforts.**
   Launch communication, education, marketing, and training efforts on storage service options and best practices as they are developed, including the creation of "ask the expert" data management consulting services for newly developed needs and support of users. This may include a web site that contains the options for storage services, including a matrix to help users make sense of what is available, includes best practices and policies for data management (particularly for research data), and pointers to resources for assistance and help. This effort should establish training for IT professionals on data management issues that can be updated and held regularly, as well as training for researchers and graduate students on research data management issues such as that offered at MIT: http://libraries.mit.edu/guides/subjects/data-management/. This training might be integrated into the RAMP training (http://www.ospra.illinois.edu/RAMP.html) offered by OSPRA or as a stand-alone workshop offered regularly. In addition, regular communications about good data management practices should be broadcast on campus to ensure awareness and to reach new members of the community. These efforts, particularly in the area of

research data, should be coordinated with those already underway by the campus Data Stewardship effort and the University Library.

8. **Implement a pilot of federated storage.**
   Implement a pilot of federated storage between at least three of the largest storage providers on campus today, and implement the architecture for future enhancement. The goal is to enable more effective management of decentralized storage pools, and to enable easier transport of data. This pilot can identify any network upgrades or capabilities that are needed to take full advantage of federating storage now, while at the same time ensuring that the network is not a barrier to getting full value from a central tiered storage service.

9. **Implement a pilot of central tiered storage for units.**
   Implement a pilot effort that will provide central tiered storage to several campus units; this can serve as one pilot of central tiered storage, and help inform the effort to build a broader central tiered storage service as in Recommendation 14 below. AITS has some storage capacity that is available because of an upgrade to new storage, and this could serve to host the data from several units that are otherwise considering new purchases of unit-specific storage. This pilot could potentially include multimedia storage.

10. **Create large research data storage service proposal.**
    Create a proposal for storage services (replication/archive pool) that support large research datasets, with representatives from units including NCSA, IGB, Beckman, CITES, the University Library, and the OVCR. It may be appropriate to have UIC representation. This service should complement the Campus High Performance Computing Cluster effort. Ask that group to define and recommend such a service and potential funding models.

11. **Establish cloud technical & contracting expertise.**
    Establish central IT technical and contracting expertise to scope and simplify cloud storage options for the campus (Internet2, commodity, CIC, other consortia and groups). With cloud service options growing rapidly, procurement requires expertise that ought to be available centrally so that units can take advantage of it rather than develop the expertise locally. Establishing a specific role for tracking cloud options, helping identify which are of potential value to the campus, and helping shepherd through the process of arranging contracts for such could provide significant value to the campus. It would avoid duplication of effort, so each unit (or even each researcher) doesn't need to replicate this expertise (and time), and greatly increase the chances of identifying cloud opportunities that could deliver great value to the campus.

12. **Identify subsidized backup, replication & archiving services.**
    Form a "replication and archiving" working group, perhaps with one or more focus groups, to articulate the variety of campus backup needs and to analyze how the current backup services can meet these economically, what slight variations could meet

more needs, and to identify any needs that require different solutions. While there is one type of backup service (TSM) available, the survey suggests that it does not meet all of the backup needs on campus. Both CITES and AITS have strong backup technologies today, and the CITES service is about to have a significant price decrease implemented, lowering the cost of using the service. But, even with such a price decrease, the backup technologies in use do not economically and efficiently cover all of the backup needs on campus. And, while it can be effective for disaster recovery, it is not as well-suited to supporting business continuity needs. Having more than one copy of data that can be accessed as quickly as is needed is too important to leave to chance and to the vagaries of budgets and a landscape of disparate and uncoordinated replication, backup and archiving solutions. These services are so critical that the Task Force recommends that they be subsidized so that they are within financial and technical reach of any campus unit that needs them.

13. **Develop central tiered storage capability to address unit and "common good" needs.**
Develop central tiered storage capability that includes file-system as well as block storage, based on requirements determined from both the survey findings and broader conversations. As an initial service, establish a pilot of "common good" storage for faculty and staff, focusing first on research, and then expanding this to cover a base of storage for all campus users. As identified in the section on what our peer institutions are doing, providing "common good" storage for campus users is a norm, and some of the leading institutions are in particular going further and providing a significant amount of "common good" storage for researchers.

A central tiered storage solution can deliver the greatest cost/benefit return to the campus, and its units and individuals. As described in other parts of this report, the total FTE to manage storage is reduced and there can be significantly less storage purchased years before it is needed. It also provides a consistent storage service upon which to layer value-added services, as described in Recommendation 15 below. Central tiered storage can meet a need that has already been identified by the Center for Multimedia Excellence (CME) for effectively managed and persistent multimedia storage, potentially provide a "competitive advantage" to campus researchers, etc. Some part of this central tiered service may reside in the cloud. But it will take time to shift the campus in this direction. Establishing this service will require one-time investment, and a steady continuing investment. While it was not within the scope of the Task Force to develop a specific cost model, it is clear from peer institution experience that the Urbana campus could begin to realize reasonable savings within a five-year period. That savings would come to individual units because they would no longer be required to support significant storage infrastructure.

14. **Establish a new Data Security Compliance role.**
Establish a new role on campus of Data Security Compliance with dual reports to the Office of the Vice Chancellor for Research (OVCR) and the Chief Information Security Officer. This role is described in some detail in Section IV. B. of this report, the section

describing the findings and recommendations of the Sensitive Data, Security and Privacy Working Group.

15. **Define and create value-added storage services such as database management, curation, etc.**
    These efforts should include exploring the creation of campus-wide value-added storage services such as:

    - Deduplication
    - Database hosting and management
    - Metadata services
    - Data curation services
    - Data lifecycle management services
    - Preservation services
    - Archiving services

    Deduplication can save on the amount of storage needed. Both the Research and Institutional Working Groups identified database hosting and management services as a need, and peer institutions (e.g. Indiana University) already provide this, including such a service targeted at researchers. And preservation services have already been identified as important for multimedia and at least some research data. We need to refine our knowledge and gather more input past this report itself -- including further/deeper analysis of the Urbana campus Data Management Plans, follow-on storage surveys, faculty/researcher survey on data and its management, continued work on sensitive data management, etc. These steps can leverage efforts already underway in the campus Data Stewardship Steering Group and the University Library. We know the above services are important today, but we don't know the full need, the total cost of providing such services, and effective ways of funding and resourcing all of these services. And it will be much easier and less costly to layer these services over a few centrally managed storage services than the current multitude of disparate storage services. So moving forward on all of the other recommendations will provide an enabling step for accomplishing this last recommendation.

B. Challenges and Opportunities

It is critical to the University's competitive edge in research and education that the campus adopt the storage services strategy, and recommendations, presented in the Task Force Executive Summary. Moving carefully but purposefully toward central support for capacity and management of storage services provides everyone with access to better quality storage and related services, as well as real cost savings.

The Task Force is not recommending that all data storage be centrally mandated. Some units with unique requirements may need to maintain storage infrastructure. However, a

substantial percentage of storage needs could be addressed more effectively if they were managed centrally, with sufficient staff and services to support the specific needs of units, groups, and individuals. The Task Force underscores the critical need for individualized services around centralized storage, so that units and individuals are able to accomplish their work without significant delays due to an inflexible central configuration.

The Task Force recognizes that there are challenges that will need to be overcome in order to implement this storage service strategy. The biggest challenge is "buy-in", where the key factors are ease of use, control, and cost.

There is widespread agreement that unless storage services are easy to use, they won't be used. Access to what you need, when you need it and where you need it, is paramount. We need to provide storage services that fit seamlessly into the user's daily work activities. Users need file services that are as simple to use as accessing another folder on their desktop/laptop/other devices. They need storage that fits into the research workflows and cyberenvironments in which they are doing their research. If the campus creates widely available storage services with these properties, adoption will be much easier.

Faculty, staff and students on the Urbana campus are currently accustomed to highly decentralized and heterogeneous storage services. They are accustomed to working with unit level IT staff when they have specific storage needs, or doing it on their own. This suggests a greater sense of control over storage resources and data management. By providing flexible central storage options with appropriate storage services, training, and consulting, the campus can ensure that individuals and units can better manage their data than current local options allow. In this scenario, unit IT resources will be free to focus on higher-level data management issues.

There is common misconception that storage is cheap. This is fueled by what the Task Force refers to as the “Best Buy” syndrome—the widely-held belief that almost anyone can buy an external 1TB hard drive for $100 or less from a big box store, and thereby solve their storage problems. Reliable, well-managed storage services that support research, education, and institutional memory require up-front and ongoing investment. The Urbana campus can ensure the lowest cost possible by developing some level of tiered storage services. Tiered storage keeps data stored on the lowest quality/performing storage that is appropriate to the access patterns for that data, while ensuring that sufficient copies exist elsewhere so that a failure does not cause data loss.

The University faces risk of data exposure (in research and institutional data) with the current highly decentralized storage environment. This can be resolved by developing centralized storage options. Although there is an overall institutional investment, the long-term benefits outweigh the costs. Some basic services such as backup of critical data, ought to be fully subsidized, while others, like database hosting and curation, ought to be offered as cost recovery. Costs to the unit can be mitigated by subsidizing services, and in the case of research grants, making it easy to include any appropriate costs in the grant

budget. Stewarding our data reliably is critical to the academic, administrative, and service activities of the campus, just as the campus network is a critical underpinning today. Storage services need to be viewed in a similar manner. Easily affordable backup, replication, and archiving services should be part of the infrastructure.

A final mitigation to cost is establishing deduplication services and data lifecycle management, automated as much as feasible, migrating data to cheaper storage when it is less likely to be needed, and deleting the data when it is no longer needed. That helps ensure you are storing no more than you need to. Of course, data lifecycle management relies on having first established best practices and guidelines around data retention, so that is also an important effort to undertake.

Well-stewarded data is at the heart of maximizing the impact of the University's research, teaching, learning, and public outreach. It supports reproducible research, reusable data enabling new research, and informed administrative decision making and assessment. It is a necessity in order to satisfy records retention rules, research funding agency expectations, and business continuity and disaster recovery needs. Now is the time to move forward with a cohesive storage strategy and set of services that not only minimizes risk and maximizes value, but also effectively provides our campus a competitive advantage in addressing its mission.

Appendices

Appendix A: Data Storage Services Task Force Charge Letter

July 25, 2011

To: Mike Grady, Executive Program Officer for Cyberinfrastructure
Beth Sandore Namachchivaya, Associate University Librarian for Information Technology Policy & Planning

From: Paul Hixson, Interim CIO

Re: Charge Letter for Data Storage Services Taskforce

Mike and Beth, I am writing to ask each of you to serve as co-chairs of a new taskforce to develop a comprehensive strategic plan for addressing the central data storage needs of this campus, including specific operational ideas for implementation. As one of your first steps as co-chairs, I would ask you to select a diverse group of colleagues from across the campus to serve with you on this taskforce, keeping in mind the need to have the group be large enough to gather various informed viewpoints and small enough to be productive.

The work of this taskforce will be critical in helping the campus establish a cohesive set of storage services that efficiently and effectively meets the data stewardship, storage, backup, and data management needs of faculty, staff, and students. It will be very important for the work of this new taskforce to complement and, where appropriate, coordinate with the ongoing work of the Data Center Consolidation Committee, the campus Data Stewardship steering group, the new campus community cluster effort, the Media Commons project, and the Center for Media Excellence.

In carrying out this assignment, the Data Storage Services Taskforce should focus on the needs of all data users on this campus (faculty, staff, and students) as well as institutional and unit needs. It is expected that some of the needs that will be identified through this process may overlap with – or complement – other areas of need, while others will present a more unique need case. The taskforce is encouraged to consider which needs might best lend themselves to be treated collectively as "common needs" (and therefore possibly be addressed in a comprehensive, global manner as "common goods") versus those areas that would best be dealt with more narrowly. The taskforce should undertake the following:

- Begin by conducting a census of all of the existing data storage solutions that are currently being used by faculty, staff, and students on this campus. Special attention will need to be given to the unique needs of researchers, including the growing

requirements of funding agencies for researchers to meet greater data stewardship responsibilities, as well as the rapid growth of the volume of research data needing to be stored.

- Identify all current data storage needs that are not currently being met on this campus (separate those that are recognized as unmet needs by end-users from those that are currently identified only by the expert community, but not yet recognized by the end-user community)
- Evaluate the need for providing value-added services (such as database design or hosting, data archiving, data curation, data backup, etc.) to whatever physical storage solutions are recommended.
- Consult with all on-going campus committees dealing with related data services issues (such as those mentioned above) plus any others your taskforce is aware of, to insure that the deliberations and recommendations of this taskforce represent a comprehensive and thorough examination of all current data storage service needs.
- Study what other peer institutions are doing in this same area. Evaluate whether any of the models being pursued elsewhere would work well here, and if so, whether it would be possible or wise for us to simply adapt/adopt that model.
- Consider the question of whether we should be partnering with other peers (including the CIC) in pursuing a joint collaborative model for meeting our campus needs.
- Consider the role that cloud storage should play, now and in the future, in meeting campus long-term storage needs. Even if the committee determines that it would not be wise to utilize cloud storage at the present time, the committee is encouraged to design current service offerings in such a way that they could accommodate cloud storage in the future.
- In developing data storage recommendations for this campus, the taskforce is encouraged to think both strategically and operationally. And, thus, in any taskforce proposals that are developed, some consideration should be given to how both strategic and operational oversight/guidance of any service will be staffed and maintained.

Finally, although the charge to this taskforce has been written with a focus on addressing the needs of the Urbana-Champaign campus, it is recognized that this project actually has a high potential to be of both interest and benefit to our sister campuses in Chicago and Springfield. Therefore, early on, we encourage you to reach out to colleagues at UIC and UIS and invite them to participate, if they are interested, in the work of this group. If the work of this taskforce could be expanded to meet the needs of the larger University without slowing down the efforts of addressing the needs of this campus, that would be a good thing.

In order to meet the needs of the campus in a timely manner, I need to ask your taskforce to adhere to a fairly aggressive timeline. I would appreciate receiving the committee's final report by January 1, 2012 and an intermediate progress report by October 1, 2011. The taskforce should understand that I intend to share the interim October report with Executive CIO Michael Hites and with the CIO¿s at both UIC and UIS so that the IT Governance groups on our sister campuses can be kept informed of the work of this committee and your tentative findings can inform their planning processes for the coming year. Finally, the taskforce should also

understand that your final report in January will probably go to the future UIUC governance committee on its way to me, the Provost, Chancellor, and Executive CIO.

Please let me know as soon as possible if you will be able to accept this important assignment.

c: Richard Wheeler
Michael Hites
Ravi Iyer
Cynthia Lindstrom
Robert Goldstein
Farouk Eslahi
CIO Council members
IT Governance Committee members (UIUC)

Appendix B: Data Storage Services Task Force Membership

Core Task Force

- Mike Grady, Office of the CIO ; Task Force co-chair
- Beth Sandore Namachchivaya, University Library; Task Force co-chair
- Jason Alt, NCSA
- Michelle Butler, NCSA
- Mike Corn, Office of the CIO
- Jennifer Eardley, Division of Biomedical Science, OVCR
- David Gerstenecker, College of ACES
- Gabe Gibson, College of LAS
- Howard Guenther, Office of the Vice Chancellor for Research (OVCR)
- Alice Jones, AITS/UA
- Jackie Kern, Facilities & Services
- Charley Kline, CITES
- Carol Malmgren, Office of the Registrar

Sensitive data, security and privacy Working Group

- Mike Corn, Office of the CIO
- Jennifer Eardley, Division of Biomedical Science, OVCR
- Maggie Helms, Division of Biomedical Science

Storage architecture, technology, delivery, and cost models Working Group

- Michelle Butler, NCSA (consulting & review)
- Michael Edwards, College of LAS
- Alice Jones, AITS/UA; co-chair
- Charley Kline, CITES; co-chair
- Frank Penrose, College of Engineering

Storage as a part of the Research lifecycle Working Group

- Michelle Butler, NCSA, co-chair
- Dan Davidson, Institute for Genomic Biology
- Gabe Gibson, College of LAS; co-chair
- Mike Grady, Office of the CIO (ex officio)
- Howard Guenther, OVCR; co-chair
- Maggie Helms, Division of Biomedical Science
- Josh Henry, College of ACES
- Sarah Shreeves, University Library
- Chuck Wallbaum, School of Chemical Sciences

Workplace productivity, instruction, and institutional assets Working Group

- Jack Brighton, College of Media, Center for Multimedia Excellence (CME)
- David Gerstenecker, College of ACES; co-chair
- Tom Habing, University Library
- Joanne Kaczmarek, University Archives
- Jackie Kern, Facilities & Services; co-chair
- Carol Livingstone, Division of Management Information
- Carol Malmgren, Office of the Registrar; co-chair
- Glenda Morgan, Office of the CIO
- Kristopher Williams, Materials Research Lab

The Task Force additionally met with, consulted with, and had help from a number of other campus groups and individuals. These included:

- Randy Cetin, Office of the CIO, Executive Governance Committee for Data Center Consolidation (EGCDCC)
- Lauren Garry, Center for Advanced Design, Research, and Exploration (CADRE) at the UIC campus
- Bill Goodman, College of AHS; EGCDCC
- Jamie McGowan, University Library (survey analysis)
- Bill Mischo, University Library (Urbana DMP analysis)
- Mary Schlembach, University Library (Urbana DMP analysis)
- John Towns, NCSA
- Center for Multimedia Excellence (CME)
- Data Stewardship Committee
- Executive Governance Committee for Data Center Consolidation (EGCDCC)
- Urbana campus IT Pro community
- CIC Data Storage Working Group

Appendix C: Data Storage Services Task Force Process and Activities

The Interim CIO for UIUC (Paul Hixson) created the campus Data Storage Services Task Force at the end of July 2011 with a charge letter included as Appendix A. Our primary goal, as stated in that letter, was to "develop a comprehensive strategic plan for addressing the central data storage needs of this campus, including specific operational ideas for implementation". The Task Force was asked to carry out its work expeditiously, delivering an interim report by the beginning of October 2011, with an initial due date of the end of 2011 for a full report.

We recognized that the most important elements to succeeding were inviting and ensuring broad input, being open in our work, and focusing on identifying a broad range of key use cases from which we could derive a set of requirements and needs that would drive our storage strategy recommendations. At a greater level of detail, all of the following steps of our process were critical to successfully delivering on our charge:

- Get a fast start by identifying willing individuals that reflected a diversity of functional and organizational perspectives within the campus.
- See if the other UI campuses were interested and able to participate in the Task Force. Both UIC and UIS expressed interest in our work, but couldn't commit to fully participating given the compact timeline the Task Force was charged to follow. Some minimal input has been received from key individuals at UIC.
- Start with a smaller core group that could work quickly to fully scope out the Task Force process and the key steps that were needed to ensure effective and comprehensive input from the campus community on the current storage landscape and emerging storage needs. Besides the co-chairs, a core group of ten individuals (listed in Y) was identified and invited to participate, and all agreed.
- Meet weekly until early October, in order to be able to deliver an interim report with some substance by then.
- Conduct our work, to the greatest extent possible, in an open manner, and invite the full campus IT community to provide input at any time, with a particular focus on collecting use cases illustrating the range of storage uses and needs. Establishing a CITES wiki space (https://wiki.cites.uiuc.edu/wiki/display/DSST/), open to the entire campus community, was one of the first steps that was taken. That was followed by inviting IT Pros to an initial Caffeine Break where the Task Force charge was discussed, and all were invited to provide their input on the wiki or to any member of the Task Force.
- Begin by conducting a campus storage census: create a storage survey instrument, identify the key individuals to whom to administer the survey, collect the results, and analyze the survey responses. The survey process and results are discussed in Section V of this report.
- Identify and study our peer institutions' storage services and strategies (called out in our charge), to evaluate what we could learn from and/or leverage from their storage service models. The next section of the report summarizes what we've learned from that effort.
- Begin to develop the range of storage use cases that were needed in order to understand the storage service needs of the faculty, staff, students and units on campus. Besides the

Task Force members themselves developing use cases, there was a an additional Caffeine Break focused on storage service use cases held in September.

- Conduct a brainstorming session to generate an initial set of anticipated storage priorities and likely recommendations that would match those priorities, to serve to help plan the rest of our work and as talking points with various complementary campus committees and groups (e.g. Executive Governance Committee for Data Center Consolidation, Data Stewardship Committee).

The above steps were carried out in August and September, leading to an interim report submitted to Paul Hixson on Oct. 5, 2011. The Task Force discussed this interim report, and its proposed next steps, with Paul Hixson on Oct. 10, 2011. It was agreed that the important next steps to enable the Task Force to successfully complete its work would be:

- Establish a set of working groups that would expand the membership of the Task Force and allow for a deeper focus in the functional areas that were identified as critical to explore more fully:
    - Storage architecture, technology, delivery, and cost models (5 members)
    - Storage as a part of the Research lifecycle (9 members)
    - Workplace productivity, instruction, and institutional assets (9 members)
    - Sensitive data, security and privacy (3 members)

  In total, the Task Force membership expanded to 26 individuals (see full list above in Appendix B). The core charge of each working group was to construct use cases articulating the storage services needs within their designated functional area/campus community, and from those recommend storage service strategies that best serve those needs. The working groups were led by core Task Force members and did excellent work, and each submitted a report to the full Task Force summarizing their work and recommendations by late January.

- Work with related campus committees, groups, and efforts to identify work they've done that could be leveraged by the Task Force, and opportunities for complementary strategies. These included the Executive Governance Committee for Data Center Consolidation, the Data Stewardship Committee, the University Box pilot team and the broader Internet2 NET+ effort, and further interaction with campus IT Pros through a presentation and discussion at the IT Pro Forum in November.
- Continue to gather and consider the work of peer institutions, particularly drawing on the CIC-wide effort to look at research data storage services and strategies at CIC institutions.
- Extend the timeframe for the Task Force to complete its work and submit its full report to the beginning of March 2012.
- Have a mini-retreat in late January 2012 where each working group would present on its work and recommendations, and through breakout groups and full group discussion, arrive at a final set of proposed recommendations to go into the final report.
- With help from various Task Force members, agree on a report template and draft the final report and present that draft to Paul Hixson and Michael Hites on Feb. 23, 2012. After

incorporating their feedback, finalize the report by the end of February 2012.

Finally, we want to highlight that the Task Force has benefited greatly from the willingness of everyone participating to give of their time and expertise to help inform, guide, lead and contribute to the work. To the extent that this report succeeds in identifying the campus storage service needs and an effective set of strategies to best meet those needs, that is fully due in part to the hard work of all of the members of the Task Force, and more broadly to all of the IT Pros, units, and campus community members that contributed.

Appendix D: Storage at Peer Institutions

The following resources describing storage services, strategies and planning at peer institutions and consortia in which Illinois participates were particularly useful to the work of the Task Force. Many of these helped to either spark ideas for strategies and recommendations, and/or substantiate that our analysis of the campus storage census and use cases gathered aligned with what a number of our peer institutions are similarly determining.

Committee on Institutional Cooperation (CIC) Data Storage Working Group: The CIC has established a group to focus on data storage services (http://www.cic.net/db/memDisp1.asp?id=348), in order to learn from each other and identify potential areas of collaboration to be explored. One recommendation that the CIC Data Storage working group has already made is that all institutions should consider a "common good" storage service provided to researchers for their research data. There are several exemplary "common good" research data storage specific services at CIC institutions. Indiana and Northwestern have been identified as "leading edge" for this, and other CIC institutions are now establishing a similar service at the 50GB level or more (e.g. University of Chicago).

- **Indiana University's** Scholarly Data Archive and Research File System
- **Northwestern University's** Vault research storage services

Common Solutions Group (CSG): presentations on storage services and plans at several CSG member institutions (http://www.stonesoup.org/meetings/1005/work2.pres/):

- Overall survey results: http://www.stonesoup.org/meetings/1005/work2.pres/CSG-StorageSurvey-20100512.htm
- **University of California, Berkeley**: http://www.stonesoup.org/meetings/1005/work2.pres/waggener.pdf
- **University of Iowa**: http://www.stonesoup.org/meetings/1005/work2.pres/shafer.htm
- **University of Virginia**: http://www.stonesoup.org/meetings/1005/work2.pres/jokl.pdf

Iowa State University: Iowa State has a fairly simple straightforward model with several distinct services and tiers of storage defined. Follow the Resources links from the right column off the main page: http://www.cio.iastate.edu/projects/storage/

Stanford University's Storage Strategy Documents. Each is structured with an overview/current state, a vision, goals, roadmap and measures of successes. These are interesting documents, focused on storage technologies.

- Data Archive and Backup
- Data Storage Management
- Networked Storage
- Cloud Storage

University of Texas at Austin's Central Storage Project:
"The Central Storage project is a two-phase initiative that will enable ITS to cost-effectively expand and enhance centralized data storage services offerings for campus. Phase I was completed in summer 2010 and addressed the immediate needs to sustain current campus storage environments. Phase II is currently under way, and focuses on developing a storage roadmap that allows for optimal allocation of ITS funds and resources to meet campus storage needs over the next 3 to 5 years".

- Analysis of Data Storage and Data Protection Options Offered by Peer Institutions and 3rd Party Vendors (pricing comparisons and indication of what services are available): http://www.utexas.edu/its/central-storage/governance/Data%20storage%20options%20at%20peer%20institutions%20v3.pdf

- UT/Austin conducted focus group sessions on storage needs, and their Executive Summary from that activity yielded the following key themes for data storage (http://www.utexas.edu/its/central-storage/governance/Executive%20Summary%20for%20Focus%20Group%20Data%20Analysis.pdf):

  "According to frequency of mention by focus group participants, the most important themes for data storage on campus are:
  1. *User education/Help with choosing a solution*
  2. *Sharing documents with users both within UT-Austin and external to the university*
  3. *Easy to provision and use the service*
  4. *Support for the data retention lifecycle*
  5. *Backups (unspecified - desktop and/or server)*
  6. *Encryption/Cat 1 Data/Data security*
  7. *Ability to store and share large files*

  It is important to note that based on the user feedback, the solution for the number one theme (User education / Help with choosing a solution) may involve simplifying the choices customers have and automating the provisioning process (theme #3) as much as possible rather than providing additional documentation or training classes."

Appendix E: Units represented by the storage survey responses

The 43 storage survey responses included five from units within the College of ACES and five within the College of LAS. The College of ACES also provided a summary response on behalf of the college. The five responses within LAS were rolled up into one composite response. This yielded 34 total units that were used for broad analysis. The units responding, and the set of units used for analysis, are listed here:

Colleges & Instructional Units:

College of Agricultural, Consumer and Environmental Sciences (ACES)
- Agricultural and Biological Engineering
- Agriculture & Consumer Economics
- Animal Sciences
- Extension
- HCD, FSHN, NutrSci, AgEd, CCRS

College of Business
College of Education
College of Engineering
College of Fine and Applied Arts (FAA)
Graduate College
Graduate School of Library and Information Science (GSLIS)
College of Liberal Arts and Sciences (LAS)
- Astronomy
- ATLAS
- Atmospheric Sciences
- Biology-related schools and programs
- School of Chemical Sciences

College of Law
College of Medicine
School of Labor and Employment Relations (LER)
School of Social Work

Other Academic Units:

Online and Continuing Education
University Library

Research Centers & Institutes:

Beckman Institute
Center for Advanced Study (CAS)
Fire Service Institute
Institute for Genomic Biology (IGB)
Illinois Natural History Survey
Illinois State Geological Survey
Illinois State Water Survey

National Center for Supercomputing Applications (NCSA)
W.M. Keck Center (Bioinformatics Unit)

Administrative/service units:
AITS/UA
CITES
Division of Management Information (DMI)
Facilities & Services
McKinley Health Center
Public Safety
Swanlund IT Service Center (24 offices/programs/etc. covered)

Auxiliaries & Affiliated Agencies:
Campus Recreation
Division of Intercollegiate Athletics (DIA)
Housing
UI Foundation

Appendix F: Additional information to note from the storage survey responses

At a greater level of detail, the following are a number of useful takeaways from the storage survey responses that the Task Force was able to glean:

- Not surprisingly, the need for storage capacity keeps growing, and in some cases, very rapidly. In particular, research data needs and multimedia (video, images, etc.) were frequently noted;
- High interest in using storage services that are easy to use, cost-effective, and meet unit needs;
- Virtual servers (VMs) and the storage services associated with them were often mentioned. Centrally-hosted VMs appear to offer one key leverage point at which to centralize storage pools;
- That current central backup services do get used and meet some needs, but there are many backup needs for which the current central service is not a good match. Expense was the most cited reason, but that was often linked to the current backup model not providing the necessary levels of discrimination between what is backed up and how;
- Administrative storage appears to have a more consistent backup strategy than research storage ("admin" units have a higher percentage of backup storage to mainline storage, in general). [In particular, ratio of backup to primary storage seems particularly low in three of the biggest players -- Engineering, IGB, and Beckman. Is that because the nature of a lot of the data on primary storage doesn't need backup, an artifact of the survey, or just too expensive to do so with technologies/funding they have today? (Also in some of the smaller storage-managing units like FAA.) This is one area where an additional survey/further work to understand what and how things are being backed up today would be useful.]
- Less than half (44%) of the respondents indicated internal practices around the management of sensitive data, and 14% indicated they do not store sensitive data, which seems unlikely. Access control lists appear to be the basic strategy today for restricting/managing access to sensitive data. Data encryption appears to both have limited current deployment and limited plans to deploy such in the near future;
- Technologies that are rapidly growing in adoption in the storage industry -- snapshots, replication, de-duplication, encryption -- not only appear to have very limited use currently on campus, but also very few units indicating explicit plans to do so. The Task Force would speculate that a likely reason for that would be the lack of resources (expertise and/or funds) to explore and deploy such, versus a lack of interest/need. In fact, a number of responses indicated interest in exploring these technologies and/or having such available;
- Limited value-added services are being provided today, such as database hosting/management, data modeling, metadata, curation, preservation, etc. But a need for them was highlighted by a number of units, both in the research and administrative data arenas;
- Other value-added storage services noted in several responses are document

management and digital asset management;
- Recognition that the campus needs better defined data lifecycles/retention periods for a range of data;
- The need for archiving services was frequently noted.

Note that units were asked to return the surveys within a relatively tight timeframe, and encouraged to provide their best estimates and “ballpark figures”. In analyzing the responses, the Task Force has learned a number of valuable things about how to reshape the survey to gather more specific information in some areas that will be useful for future planning and executing on a number of the recommendations.

Appendix G: Research Working Group: Primary Storage Needs and Services

The primary storage issues, requirements and services identified for research and the research lifecycle are:

- Common good allocation: This seems to be almost a necessity for any basic data storage program. The most likely scenario is a fairly basic level of storage provided at little or no cost to researchers, students, and staff members.

- Additional/expanded memory tiered storage options and services: From most of the use cases considered, it was apparent that researcher requirements will demand very large storage capacities and user services beyond any common good allocation. The presumption is this additional capacity will be provided in the form of fee-based models.

- Data systems security: The complex regulatory environment of protected and confidential information, HIPPA requirements, intellectual property, export controlled, etc. will make this aspect paramount in any storage provisions and capabilities.

- Ability to replicate research studies: This aspect is a primary intent of the federal regulations and is presumed to be fundamental for federal disclosures.

- Longitudinal aspects: A large segment of research projects will require datasets that can be extended over long periods of time and possibly accessed for applications much different from the original objectives of the research.

- Data retention provisions: Many key questions need to be addressed regarding what type of data will be retained, in what format, and for what duration.

- Centralized repositories: A key factor in a campus-wide program is the extent to which centralized approaches will be utilized, considering factors of efficiency, cost, and service levels.

- Access provisions, including cluster storage: A fundamental concern for any level of external outreach, especially research project collaborators.

- Digital conversion of data: Important consideration, but determined to be outside the scope of this task force and/or working group.

- Authentication, authorization, and accounting: Cornerstones that need to be included for any data storage approaches.

- Rapid data uploading and downloading: Critical benchmarks for service levels and adoption by the research community.

- Disaster recovery and backup provisions: Fundamental questions, but need to resolve the mechanisms for providing these services on a long-term operational basis.

- Centralized database hosting: Indicated to be a very strong element of architecture and technology design.

Appendix H: Research Working Group -- Summary of NSF data management plans prepared by UIUC researchers, January – November 2011

The Research Working Group reviewed a summary prepared by Bill Mischo (University Library/Grainger) of Data Management Plans submitted to the NSF since January 2011. This overview was useful for compiling the present data storage practices and provisions currently implemented by UIUC researchers. The results of this analysis included:

- 341 proposals with DMPs (updates and supplements not included)
- 43 proposals used Grainger Library template and mention assistance from Grainger in their proposal (12.61%)
- 57 proposals identified IDEALS as a location where data will be deposited (includes the 43 from above) (16.72%)
- 52 proposals used the single sentence "See GPG Chapter II.C.2.j for guidance on contents" for their DMP (15.25%)

The current storage solutions are not optimal, but researchers are very resourceful in utilizing storage options and capacity to meet their needs. Compliance is an issue in some cases. It's clear that more storage space is a general concern and the current solutions aren't meeting those requirements. Many other needs do not seem to be effectively addressed at even the most basic levels. The current storage solutions include, in order of most frequent to least:

- RAIDs
- work or lab computers
- research group servers
- external hard drives
- no storage format provided (2 cases)
- outside repository
- "Unique storage solutions"
- NCSA data test bed
- custom built processor
- research group cluster

Appendix I: Storage Architecture

The universe of discourse when discussing storage solutions is much too large to consider as a whole when developing a vision for a campus storage architecture. As a first step in design methodology, we considered as out of scope:

Systems which directly attach (via SATA, SAS, or USB) their storage devices (in other words, storage which is completely private to the system using it)
Cloud storage
Federated storage, as a special case of cloud storage

It's important to understand that things declared as out of scope for a storage *architecture* are certainly not out of scope for the overall storage *solution.* For instance, we envision that cloud storage services will play a very important role for portable personal storage as well as for accommodating special needs such as HIPAA compliance. However, since such solutions are closed and/or connect to the rest of the overall storage solution only at the very highest levels of file copying, there is little room to discuss integration with the on-campus storage architecture.

For an on-campus solution, we limited our explorations to a basic service stack imparted by Service-Oriented-Architecture (SOA) design. In particular, we operate according to the following foundation:

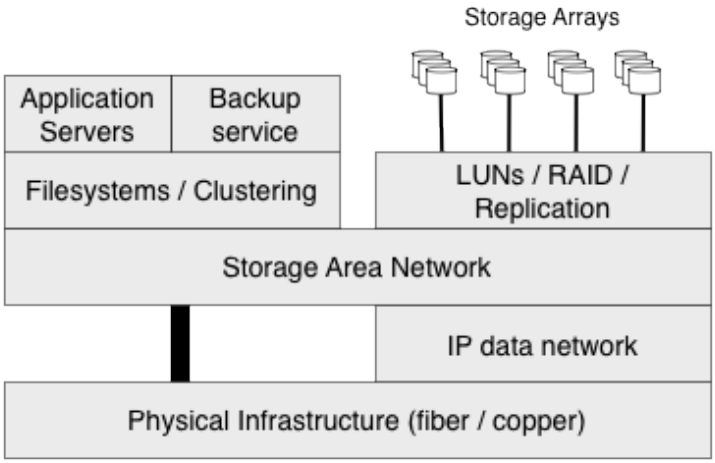


Fig. 1: Storage Service Architectural Stack

To understand the layers of this architectural stack, it helps to begin at the storage itself. Storage is organized into arrays of disks, which are divided into partitions (segments of disks). Partitions may stand alone, but more often are organized into a RAID (Redundant Array of Independent Disk). Either way, the result is a LUN or Logical Unit.

Logical Units, RAID, and Replication

A Logical Unit represents the most fundamental service offering of a storage architecture. It appears to the system accessing it as an unformatted physical disk, to be used in any way the host system requires. Depending on the type of physical disks involved and the RAID level in use, a Logical Unit will provide a certain level of performance, and a certain level of reliability. For example:

RAID-0 keeps only one copy of data, "striped" across multiple disks. This is efficient in terms of utilization since all disks directly contain usable data. Performance is high since writes and reads can be distributed across all disks simultaneously. However, reliability is quite low, since the failure of any one disk in the set results in the loss of *all* data stored in the set.
RAID-1 keeps two exact copies of all data written, on at least two separate partitions. This is inefficient in terms of utilization since it requires at least twice as much raw storage as provided, and write performance is low. But reliability is high due to the multiple copies, so that an entire physical disk can fail with no loss of data. Operation is also quite simple.
RAID-5 uses one extra "parity partition" to store recovery information in the event of the loss of one other partition in the set. This is efficient in terms of utilization since only one parity partition is required for several storage partitions. Reliability is moderately high; any one physical disk can fail without loss of data, but with a large number of disks the probability that two will simultaneously fail must be considered. Performance is low, however, due to the need to continuously rewrite the parity data, and RAID-5 solutions require dedicated hardware.
RAID-6 is similar to RAID-5 but utilizes multiple parity partitions to be even more proofed against failure. Since rebuild time on a very large RAID filesystem can be large, there can be a significant window of vulnerability should another disk fail during rebuild, and RAID-6 protects against that. Performance characteristics are similar to that of a RAID-5 set, and again, dedicated hardware is required.

There are many combinations of disk technology, disk performance, and RAID design. One of the first tasks to be made in a central storage architecture is to determine a list of storage tiers, which provide a range of options from inexpensive to expensive, from low to high performance, and from low to high reliability, that can adequately span the set of campus needs.

Some storage applications may require, as a part of their business continuity plans, that data be replicated among several locations, either in another building on campus or even in a distant location on another campus. Some storage arrays can be connected across a wide-area network and will maintain exact replicates of data between them. This has great value in terms of protection against disasters, but is expensive, and performance is usually low.

Note that some applications, particularly database systems, can replicate their data between locations at that level, rather than having the storage systems do it transparently. This often is preferable to storage-level replication as performance sacrifices are lower, but obviously such solutions are highly dependent on the particular application, and are not in scope for the storage architecture. They do, however, present a use case for an application server being able to access storage located some distance away.

Storage Area Network

For each application making use of the campus storage system, then, a Logical Unit of the appropriate tier is created, and presented to the application's system across a Storage Area Network or SAN.

A SAN can be built either directly on dedicated physical fiber and copper infrastructure available on campus, or by transporting SAN data across the existing campus data network, mingling with other data transport applications. The former solution is more expensive and requires dedicated resources but provides much higher performance; the latter is inexpensive, requires relatively little dedicated hardware, but performs less well and also is subject to bottlenecks and congestion interference from other data traffic.

Because a SAN is expensive either in terms of actual cost or occupied network bandwidth, we state that the desired configuration consist of both the storage infrastructure, and the systems hosting the applications using that storage, being located in shared data center space. This minimizes SAN buildout and network bandwidth required, and also keeps equipment housed in managed environments where space, power, and cooling can be controlled.

There will certainly be use cases for the SAN to span distances across campus, and even between campuses. Two good examples are:

Existing storage solutions which are not yet at the end of their life cycle. They should be accommodated as a least-effort way to maximize use of current deployments.
Large amounts of storage which are accessed infrequently, such as archival repositories. High performance is typically not required, and the amount of storage traffic generated is relatively low. The storage traffic can be carried across the existing campus data network at low cost and with low impact.

The former requires a physical build of a campus-wide SAN between CITES telecommunication nodes and into buildings on an as-needed basis. When the campus data network was upgraded to gigabit speeds, much of the multimode fiber plant was abandoned, and this can often be used to reach into buildings.

Filesystems and Clustering

Once a Logical Unit is presented to an application server, it must be formatted as a filesystem, which is a way of organizing the raw data space on the Logical Unit into directories and files, with ownership and access permissions. There are several kinds of filesystem organizations available, depending on operating system, security requirements, metadata requirements, performance issues, and other factors.

Commonly, a filesystem design assumes that a single filesystem driver on a single application server is manipulating the data on the Logical Unit. In other words, a Logical Unit belongs to exactly one application server, which formats it as a filesystem for its own exclusive use. This necessarily means that any other storage services, such as backup, archive, and metadata creation, must be performed by that same application server. (The backup service itself can still

be centralized, but the backup functions would need to be distributed among the application servers.)

If it is desired that backup services be performed centrally, the implication is that a separate system providing backup service *also* access the data on the Logical Unit *at the same time as* the application server. This requires a special kind of filesystem organization called a clustered filesystem, the design of which is much more complex and the management of which is more involved. Any architectural requirement for central backup, archive, metadata, or any other service which needs access to formatted filesystems being used by application systems creates an additional requirement for a clustered filesystem deployment, which greatly changes the way in which application servers access the storage provided. An additional layer in the architectural stack is created, and application systems do not directly access Logical Units but instead mount network disks via CIFS or NFS protocols. Functionality is greatly increased, but so is complexity of the architecture.

Backup and Archive

CITES currently operates a "traditional" backup system based on the IBM Tivoli Storage Management product. This has been a successful service and recent activity-based costing exercises have zeroed in on an extremely competitive price for the service. Having undergone major software and hardware refreshes in FY12, the service is at the start of a life cycle and can operate and scale in the strategic timeframe.

The TSM service also provides a lesser-known archive function which can accept files into archives in a client-driven manner and keep multiple on-site and off-site copies. This archive function is also at the core of the hierarchical storage architecture described below.

However, the world of data backup and archive is rapidly changing, and the University will need to remain keenly involved in developments in these areas, and try to strike a mark where new technologies will intersect our strategic goals in the three-to-five year timeframe.

We recommend that, fairly soon, we undertake a study of next-generation storage management architectures and technologies. The current model of "store your data, and back it up for safekeeping" may be becoming deprecated, and any of the managed storage and archive solutions on the horizon will be a significant paradigm shift from the current model, and thus will address a completely different set of requirements and modes of operation. It will take at least two years to develop a new storage architecture that leverages new technology. In the end, however, it may well be worth it to pursue a more integrated storage management model which leverages hierarchical storage, automated tier management, automated replication, and use of cloud storage services.

Hierarchical Storage Solutions

Hierarchical storage management (HSM) refers to the management of stored files according to a central policy. Rather than being statically located on the filesystem they are written to, files can be moved from there to less expensive (but lower performance) storage, de-duplicated and combined with other copies of the same file owned by other users, copied to near-line or off-

line media such as tape, or copied out to a commercial cloud storage service. The goal is to provide a seamless experience for the user, who is presented with an essentially limitless amount of online storage which is implemented by moving unused or very large files to more and more cost-efficient media.

CITES is currently exploring an HSM solution based on the archive function of TSM, combined with a GPFS clustered filesystem. Results have been encouraging, but the complexity of these solutions should not be underestimated, and they occupy considerable staff resources. Because HSM relies on the GPFS filesystem, the service cannot be offered as a part of a Logical Unit service, but only as a mounted filesystem service.

Service Offering Points

Establishing a manageable set of service offerings will be key to a successful service. Offering too few services will cause us to miss important use cases, while offering too many results in an architecture which is difficult to manage, rigid, and confusing to the end user.

The current set of services essentially consists of Tier-3 LUNs presented via iSCSI, the TSM backup and archive service, and a handful of other "one-off" services. GPFS and HSM have not been formalized as services, and this is an area that requires focus since there are many opportunities to be realized there.

Ancillary services such as metadata, curation, and federated access also require further study and probably need to be included in the requirements study for a future storage architecture.

Cost

No enterprise-grade, managed storage solution can hope to reach the cost-per-terabyte of a consumer-grade, standalone hard disk drive. This leads to the so-called "Best Buy Syndrome" in which customers seeking economical storage solutions are led to inexpensive but unreliable and unmanaged solutions involving external USB drives.

While it is hopeless to try to reach a cost per terabyte rate that matches commercial hard drives, it is important to offer a set of tiered storage solutions that cover a sufficient range from the affordable to the high-performance. Clustered filesystems and hierarchical storage solutions can help drive costs down since not all data needs to reside on local spinning disks.

A survey of central storage offerings from other service providers in higher education reveals that our backup service costs and Tier-3 storage (via iSCSI LUNs) costs are in the same ballpark as our peers, which is encouraging news. We must continue to pursue efficiencies in scale and service provisioning, however, to keep costs low.

Summary

The current set of storage solutions being offered are robust and mature, but offer only a small subset of the necessary service offerings needed to meet the many requirements discovered by the Data Storage Services Taskforce. At the same time, new, paradigm-shifting storage

solutions are starting to become available which will blur the lines between storage, backup, replication, and hierarchical storage management.

Soon CITES and the campus will need to seriously explore new technologies and develop a technology roadmap to guide storage development into the future. In the mean time, existing technologies and services need to be aggressively maintained and expanded to meet current needs and to bridge the gap toward future solutions.