2010

Computing & Communications Services

2010 / 09 / 28

Kent Percival, M.Sc., P.Eng.

Expanding CCS Storage Services

CCS now provides storage to individuals and departments. This investigation considers issues of expanding the range of service to meet additional user requirements.

While users ask for more low cost storage, the underlying issues of data protection, accessibility, findability, management, and cost control are key service requirements. Based on experience at other universities and consultant recommendations,

- It is not possible to provide "raw-disk" enterprise storage service at the price point of commodity workstation drives, without institutional subsidization.

- Enterprise storage services can provide great value-add for users, and for organizational units, in managing the growing volume of digital content; content that increasingly needs to be secured but accessed by others; content that has value beyond its period of active use.

CCS should *not* attempt provide low cost raw disk storage but focus on establishing a broader suite of information management applications and storage services focusing on efficient data management, and value-added productivity tools for major content management applications.

A data management platform would optimize costs by directing data to appropriate, cost effective, types of storage. Information lifecycle management would also move data to alternate tiers as the data ages, ultimately securely destroying obsolete data or moving it to a permanent archive to control storage growth. Separating storage technology from storage services permits the introduction of newer more cost efficient technologies as they become available, including cloud storage.

Contents

Expanding CCS Storage Services

Investigation of gaps in storage strategy

Executive Summary

Request

While CCS is offering a storage service to campus on new storage infrastructure, there seems to be an opportunity for a second tier, lower cost, service to cover other storage needs of departments and individuals.

Conclusions

1. **Digital Storage cost management is a growing issue** across the IT industry and a major concern in many higher education institutions. Rapid growth of digital information continues in most units of higher education. Data volume is expected to continue to grow at > 50% per year. Increasing the largest proportion of stored information is ready-only or inactive data that is more challenging to manage (Appendix 1). Managing this content can achieve significant storage consolidation to reduce volume and hence cost.

2. **Requirements of the management of digital information are rapidly exceeding the capability of individual users or smaller departments.** The critical areas include security, access control, findability, data loss prevention, and long retention of valuable read-only information. Users are finding it a significant burden to manage their own data.

3. While significant integrated storage vendor platforms are still leading, **the IT community is evolving to a model that distinguishes higher level user-oriented storage services from the backend infrastructure issues.** A software information/data management platform links this services delivery to backend infrastructure management.

4. Cost of locally-attached "raw storage" drives will continue to drop. **It will not possible to provide enterprise storage at comparable prices** without central subsidization.

5. Even if qualified by an SLA, **user expectations of enterprise storage solutions grow as the user becomes less concerned their local technology and their data volume continues to grow.** A user-managed storage service is most risky when their data volume exceeds their capability and tools to manage a space with many different types of data.

6. **With an information management platform more choices are available in the storage infrastructure.** Low cost MAID (Massive Array of Idle Disks)" SATA drive technology provides the lowest unit cost if the data is managed and protected by the data management platform software. Savings from purchasing lower cost storage technology can be offset by increased data management effort.

7. **Current Cloud storage services are still expensive** (bandwidth charges) and add another layer of complexity for the user. **A new generation of Cloud services are expected to integrate at the infrastructure level**, permitting Cloud storage to be used where cost effective, transparently to the users of data management applications/services.

Proposed CCS Strategy

Investment in automated information/data management will improve institutional business productivity by providing better data management services to the community. An information/data management platform will permit deployment of lower cost storage and reduce staff effort on manual data administration.

Summary of Recommendations

Tactical

1. **... that CCS not implement a low cost, unmanaged file volume storage service.**
2. **... that CCS develop a suite of extended storage service offering with several SLA profiles and incorporating metadata-based data management policies.**
3. **... that CCS plan to implement a lower cost tier of storage in the data centre, fronted by data management applications for read-only data objects.**

Strategic

4. **... that CCS invest in data management software to automate management of data objects according to the SLA profiles across different tiers of storage hardware.**
5. **... that CCS collaborate with the Library Research Enterprise & Scholarly Communication group to coordinate data management services for researchers.**
6. **... that CCS develop a technology architecture "vision" that supports coordinated tactical plans across platforms, services, and projects.**

Investigation - Expanding CCS Storage Services

Request

While CCS is offering a storage service to campus on new storage infrastructure, there seems to be an opportunity for a second tier, lower cost, service to cover other storage needs of departments and individuals. Question asked include

- Is this perceived need common?
- If it is, how is the storage industry reacting?
- What solutions are being offered or developed to meet these needs?
- Where does this tier of storage fit in the overall future of storage?
- Where does cloud storage fit in?

Investigation

This report is based on an investigation of the industry direction and how some higher education institutions are looking at enterprise storage.

- review of EDUCAUSE online papers and other industry content from publishers (Network World) and whitepapers from several leader vendors.
- searches of EDUCAUSE constituent group email archives provide information about other institutions talking about storage issues. Several leads were followed up with personal contact.
- telephone consultations with Gartner consultants provided an overview of enterprise storage issues and then a more detailed conversation on low cost storage equipment.

Local requirements were identified by

- discussion with several CCS staff,
- discussion with some library staff regarding their views on research information management.
- review of my earlier notes on research data needs (2008-9) from consultations with Assoc. Deans, Research (ADRs) and principle investigators (PIs)

Conclusions

One upon a time, users managed their own data on a limited budget. They knew what data they had and how to find it. As hard drives become full, they deleted older, obsolete data, or purchased more storage with spare funds. When others wanted to use their data, the user provided a static copy of parts of their data to others they trusted. The user periodically backed up their critical data. If critical data went missing the individual apologised and handled the workaround. The individual user was custodian, data manager, gatekeeper, and risk manager.

For some users, managing large amounts of data is becoming too complex.

When the data became valuable to the academic/admin unit or to a research team, shared systems were devised and data was entrusted to systems administrators who become the technical custodian, data manager, gatekeeper, and risk manager.

Some units find that their local data is being used in other critical applications in the University requiring much more effort to manage access and security.

Can't make price point

Users will continue to find "personal" storage, either workstation drives, or networked (CFS), a convenient and cost effective solution for their data as long as the data volume and/or usage requirements are not onerous. Enterprise "raw-disk" storage solutions cannot compete on a cost unless strategically subsidized.

- Low cost "raw-disk" MAID storage solutions exist. On a per-byte basis, for small volume, they will never be competitive workstation drives. Low cost comes without the fault tolerance and manageability of SAN/NAS-type storage so data loss risk levels are much higher.
- Workstation storage uses shorter life drives and the storage device refresh cost is hidden within workstation refresh. An enterprise system will use 10-year drives and will be refreshed every 5 years as a capital cost of central IT.

Controlling Duplicated and Obsolete Data

As data volume grows, data managers can lose track of the value of older un-changing data objects. Considering all institutional data including user-managed storage, it is estimated that 80% (Appendix 1) of stored data is obsolete and likely unknown. Many objects are copied to storage devices of others, revised, and backed up again, resulting in lots of duplicated storage.

If the volume of this "inactive" class of data objects can be reduced, overall storage costs can be significantly reduced. Organizations with strong central IT control have demonstrated that data management of this "inactive" class can reduce organizational storage costs by 40+ %.

Growing Expectations

More importantly, enterprise storage solutions generate higher client service expectations, even if defined by a service level agreement. When a central IT unit provides storage for others, they become, by default, a partner custodian, data manager, gatekeeper, and risk manager.

Value-Added Storage Services

CCS' opportunity is to target services to those individuals and departments that no longer have the interest or capacity to manage their rapidly growing digital content and data.

The strategic opportunity for enterprise storage is to provide value-added services to campus users and units who have challenges managing their data. These value added services should be flexible to

- target particular classes of information (e.g. storage and cataloguing of large numbers of digital images.
- capture appropriate metadata to support findability, manageability, and access control.
- provide secure access control, managed by the user, to other consumers (both users and applications) of the content.

Storage Service Platforms

Delivery of these differentiated services can be accomplished with appropriate service technologies in front of mass storage.

- **Content Management System** These systems capture metadata for managing the data and providing access control.
- **Database Management** Organization and access to information to some classes of information may be best handled in a database structure which capture not only the raw data but associated metadata.

- **Information Lifecycle Management** This process, manual or automated, tracks data object value, based on metadata, moving objects to appropriate an class of storage device and destroying (note security aspect) obsolete objects.
- **Digital Archive** Distinct from disaster backups, the archive is a specific service for users who require long preservation but infrequent access. The archive service includes guarantees of long term accessibility based on permitted and maintained storage formats.

Data Asset Management

One value of a data management platform is the ability to report on the data asset profile and its changing nature. A good system would provide a dashboard view of the storage assets to support service development and infrastructure resource planning. (Currently CCS does not have a good picture of the data assets.)

Storage System Flexibility

Technology management

Separating storage services and storage devices is a key strategy in managing enterprise storage. This permits selection, deployment and replacement of storage equipment without affecting the user. Data objects can be moved transparently to appropriate storage based on class or object metadata. Data can be migrated to new equipment.

Data Loss Prevention (DLP)

Statistics show that up to 80% of stored data objects have content that does not change. Managing enterprise storage with a set of value-added differentiated services permits classification significantly reduces DLP volume and frequencies.

The Digital Archive service also reduces DLP loads by ensuring there is a separate protected copy on reliable storage. DLP backups of “read-only” data repositories can be much less frequent and disaster recovery times can be longer.

A Digital Archive service significantly impacts DLP. Providing users the opportunity to move older data objects to the archive reduces the volume of static objects to be backed up. New DLP practices are required for the archive store.

Isolation of the Storage Infrastructure

With a good front end Content and Life Cycle Data Management platform in place, data can be adequately targeted to different storage systems, permitting a backend collection of *independent* storage devices with different characteristics.

This storage virtualization approach has merit in the longer term allowing individual storage hardware systems to be replaced independently of each other and of the data management platform. Data migration occurs more transparently. Differential growth of classes of storage can be accommodated. New technologies such as solid state drives (SSDs) can be implemented for specific high availability applications.

Thin provisioning (automatic allocation as needed) will be common place soon on higher performance storage.

Virtualization and thin provisioning require confident views of data classes, requirements, and anticipated growth. This requires a good data management platform that provides good reporting including a decision analysis dashboard to permit agile updating of data management policies.

Proposed CCS Strategy

Investment in automated information management will improve institutional business productivity by providing better information management services to the community. An information management platform will permit deployment of lower cost storage and reduce staff effort on manual data administration.

Savings from purchasing lower cost storage technology are offset by increased data management effort. Manual data management effort is a significantly increasing cost overwhelming device costs and creating substantial lost staff opportunity cost. Investment in automated data management will not only provide more capacity in CCS but will permit better data management and data storage services to the community.

It is proposed that the CCS Storage Services strategy separate storage infrastructure technology issues from storage service definitions by implementing data service platform technology. This strategy is a form of two sided *virtualization*. From the consumer side virtual information management and storage services are virtually delivered by an integrated enterprise information management system. On the infrastructure side, enterprise storage is a virtual common platform utilizing a variety of storage devices with different cost and performance attributes.

This strategy would include information management applications (such as image management with search and edit capability) as well as offer content specific storage services with different SLA profiles and features. The data would be stored on different types of storage infrastructure selected to meet different cost / performance characteristics. The data management system would automate policies directing data from the various service levels to the appropriate storage and provide dashboard-type reporting on use and content-type mix for further service and infrastructure planning.

CCS can begin to develop this strategy by defining and deploying very targeted to meet the needs of smaller classes of users, potentially using the Drupal content management platform or by licensing a specific technology of interest to a group of users (e.g. image management). Wider marketing of data base services may attract interest is structured storage of data now managed with flat files.

The Library, as well as discipline specific groups, is delivering data management for researchers. For some information management services, especially for researchers, a CCS partnership with the Library may result in information management platforms that achieve these value-added services that are disconnected from the storage platform.

Recommendations

Tactical

1. **... that CCS not implement a low cost, unmanaged file volume storage service.**
   Such a service would result in a service that grows unbounded and which CCS increasingly assumes management risk for.

2. **... that CCS develop a suite of extended storage service offering with several SLA profiles and incorporating metadata-based data management policies.**
   A permanent archive service should be one of the first service groups.

3. **... that CCS plan to implement a lower cost tier of storage in the data centre, fronted by data management applications for read-only data objects.**
   Fronted by data management applications for read-only data objects, this lower cost storage would have lower access performance as well as lower-frequency DLP backups

Strategic

4. **... that CCS invest in data management software to automate management of data objects according to the SLA profiles across different tiers of storage hardware.**
   Recognizing the opportunities of separating front end services and backend infrastructure is a key strategic step. To implement such a strategy, an appropriate data management platform is required to link them. This platform will consist of several key components, including, CMS/DMS, database, ILM, repository, and archive management systems.

   Further investigation is required to understand this layer. It may be appropriate that a project team lead this phase by hearing from leading vendors, developing a requirements list and piloting some of the technology.

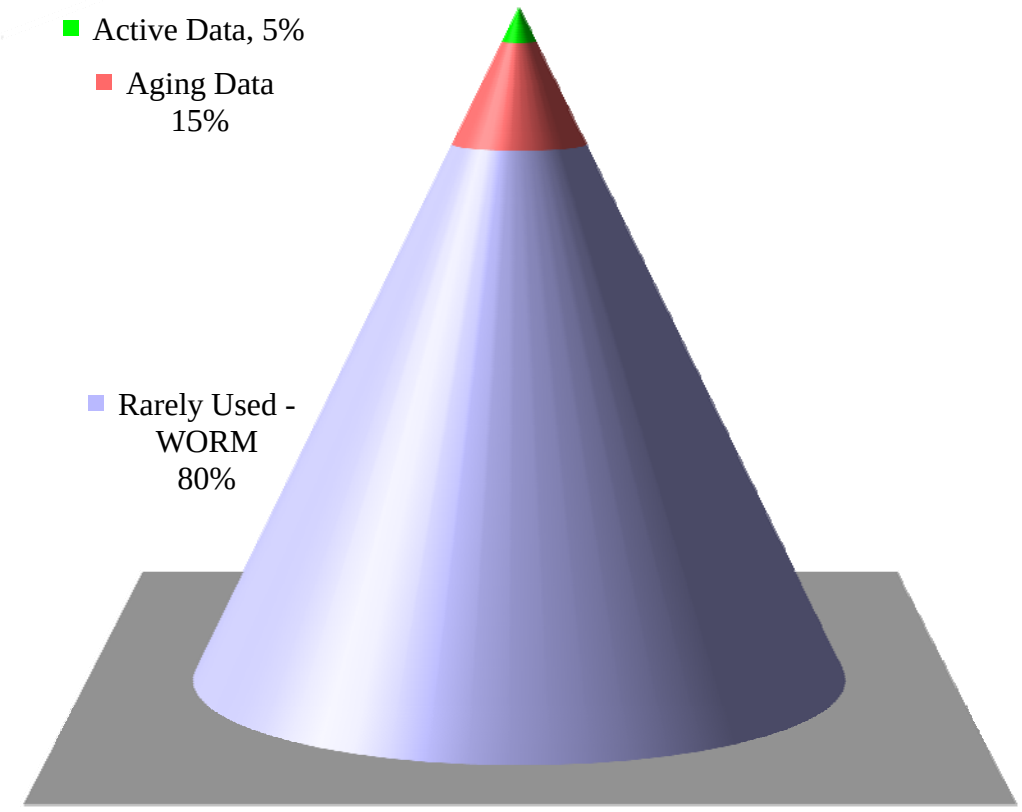
5. **... that CCS collaborate with the Library Research Enterprise & Scholarly Communication group to coordinate data management services for researchers.**
   The volume of such data on campus is already large and expected to grow significantly. As funding organizations consider strategic funding for data storage, they are also looking at the value of original data to published reports and papers. In some disciplines, "publishing" the raw data in structured form (with descriptive metadata) will become a priority. This data and papers can also be referenced from LMS environments to enhance undergraduate learning … in fact the DRC (and the UoG BRDC) is a prime example of a CCS/Library service for researchers.

6. **... that CCS develop a technology architecture "vision" that supports coordinated tactical plans across the storage platforms services, and infrastructure and associated projects.**
   Projects, such as server virtualization, hosted virtual desktops, high performance computing, and cloud(SAS) application deployments will need to support applications accessing data in various storage services and may be optimized with the an appropriate server/storage architecture.

Appendices

Appendix 1 Industry observation of Data Classification across Enterprise

From a University of Minnesota paper (2006) describing their research concerning a new storage infrastructure. Similar data results are mirrored in Gartner and Burton reports of industry sectors including higher education, health, and research institutes.



Appendix 2 Impact of Green Storage Technologies (future considerations)