

Spotlight: Storage

Section III, goal 2c, "Provide a shared pool of storage that many collaborative tools can use," defines targets to enhance its shared data storage infrastructure in support of collaboration. This appendix provides more in-depth discussion regarding those targets.

Like identity management, shared storage is one of the facilities that helps make online collaboration possible. Shared storage underlies the collaborative "spaces" in which people work together online, in real-time or asynchronously.

However, storage is frequently a key constraint, due both to price and availability. The cost of storage is often a significant component of the cost-recovery prices that must be charged for recharge-based collaborative tools, which in turn limits the use of such tools. Those costs also often make it prohibitive to make collaborative tools ubiquitous: available for use by the entire campus population. Limitations on available storage, such as maximum quotas available for email, file sharing, and collaborative workspaces, are also an important constraint. Storage limits have in some cases prevented departments from transitioning services to central campus offerings, or have led campus users toward the use of consumer- or business-oriented collaborative services from outside providers, which frequently offer larger amounts of storage.

There are several areas in which the campus's data storage infrastructure can be enhanced to better support online collaboration:

- 1. Seek to reduce the impact of storage costs and capacity limits on collaboration**

IST and departmental IT organizations have made great strides in reducing the cost of shared storage: storage which is shared among multiple hosts. However, these costs need to be further reduced in order to help remove the impact of shared storage as a cost and capacity bottleneck for collaborative tools, and to make these tools more available to broader campus audiences.

Offering shared storage, regardless of cost considerations, in itself offers many technical challenges. Finding ways to further reduce the costs of this storage will also be challenging, as it may require investigating new organizational and technical approaches and even taking some occasional risks. Among some possible approaches:

- a. Seek economies of scale.
 - Pool resources at the campus or UC-wide level.
 - Consider whether, and if so, when to outsource some storage, and possibly also associated services, to vendor-run data centers with well-honed system and storage administration practices, access to inexpensive power.

(See "Spotlight: Legal and Policy" for some concerns related to that approach.)

- b. Reduce the quantity of data that must be stored, particularly on high-cost disk.

- Investigate de-duplication and compression technologies, particularly for lower-value data and its backups. [1]
- Identify less valuable or active data, and move it to lower-cost storage

The campus Data Center already offers tiered storage, with lower costs for lower performance storage. There may also be additional ways – through both automated techniques and institutional data handling practices – to identify less valuable and/or less active data, and to migrate that data to lower-cost storage.

Such data can be relegated to less expensive, lower-performance disk storage, as well as be an earlier candidate for offline storage or outright deletion.

- c. Monitor industry trends and work with storage partners.

- Continue to monitor trends in enterprise-ready storage technologies.

An example of one significant recent trend is the advent of iSCSI as a lower-cost alternative to Fibre Channel, which has already helped generate a new IST service offering.
(<http://istpub.berkeley.edu:4201/bcc/Spring2008/1176.html>)

- Continue to partner with storage and services vendors to review potential solutions.

- d. Explore low-cost emerging technologies.

Some emerging storage technologies offer considerable promise, particularly for lower-performance, lower-feature set, higher-latency storage. Certain of these technologies may become sufficiently mature, over time, to be ready for specialized campus uses or for enterprise use, and it may be productive for the campus to periodically monitor this space.

The following are examples of the broad range of emerging storage technologies:

- Storage over IP (SOIP), ATA over Ethernet (AoE) and other lower-cost technologies for providing network-accessible storage or SANs, using commodity disks. [2]
- Emerging services from "cloud" providers, such as Amazon and Google, which provide pay-as-you-go access to storage, compute time, and other infrastructure services at these providers' data centers.
- Peer-to-peer architectures or other distributed storage techniques for redundantly distributing storage across the campus, to take

advantage of temporarily unused storage within high performance clusters, and locally attached storage on servers and workstations.

2. Provide a baseline level of storage for use within collaborative contexts.

The campus, both centrally and locally – through its colleges, schools and departments, and its research, service and administrative units – provides storage through a widely varying, patchwork set of technologies. The quality, availability, and features of these services varies widely across campus, with some units and many individuals lacking access altogether.

Because of the wide variability in access to storage across the campus, a baseline amount of storage should be made available at no cost to any collaborating community. As a goal, the lack of storage availability – at least at some modest, baseline level – should never be an impediment toward initiating collaborative activities.

As well – although implementing this may be challenging – this baseline storage should ideally be accessible from any number of collaborative tools. Stated another way, this storage should be automatically and transparently made available when a collaborating group uses any of their collaboration tools, up to the limits of their allocated space. This may also require mechanisms for reclaiming unused or inactive storage, perhaps accompanied by methods for migrating inactive data to paid storage or to other providers' storage offerings.

3. Make additional storage readily purchasable.

Any individual or collaborating group should have the opportunity to purchase as much storage for collaborative purposes as they may need.

This storage may either be used for general purposes, or within the context of a particular collaborative tool or tool set they are using, such as email or a collaborative worksite tool.

4. Provide a set of services layered on top of storage.

Value-added services should be implemented in conjunction with storage. If these services are provided as a service layer offering, many collaborative tools can rely on their existence and build upon them, rather than each tool having to implement them individually.

Some examples of these services that have been identified as needs by members of the campus community, some of which are also foundational to collaboration tools, include:

- a. Desktop-mountable share points.
- b. Granular permissions.
- c. Secure storage of sensitive and restricted data.
- d. Sharing files beyond the institution.
(Granting non-UC Berkeley affiliates the ability to read and write selected

documents.)

- e. Web-accessible files.
- f. 'Permanent' URIs and digital object identifiers.
(This service is further described in footnote [3], below.)
- g. Versioning of files.
- h. Backup and restore.

These services may either be integral features of the storage technologies used, or may be implemented at a higher layer. Examples of the latter include storage or network appliances that implement these services; storage software or file systems that incorporate value-added services; or collaborative workspace tools, such as SharePoint or Sakai.

5. Offer mechanisms for transferring very large files.

Important new areas of research, using techniques such as environmental sensor networks, medical and aerial imaging, high definition audio and video documentation, and manipulation and analysis of large datasets, produce enormous data files that must somehow be shared with colleagues and peers. In other contexts, too, both academic and administrative, there are needs to transfer very large files.

The campus should provide inexpensive storage from which large individual files and more generally, large amounts of data can be offered, for short periods of time, to collaborating partners. This service can incorporate various methods to discourage its use for purposes that violate university policies.

The campus should also provide high-speed data pipelines and other mechanisms for transferring those files. This need may already be partly met – at least for modestly large datasets and particular sets of collaborating partners – through the CALREN-2 network. In other cases, physically swapping large hard drives is still the only currently feasible option available, and the campus can examine whether there may be opportunities to help streamline this process.[4]

Finally, there may be potential for these services to be productively addressed at a scale beyond the campus, such as at the UC-wide level.

6. Archive and preserve data with long-term value.

Collaborating communities, such as researchers working together on problems, generate large amounts of data. While much of that data may ultimately prove to be relatively ephemeral, some data retains value for many years or generations, either as primary source materials (e.g. data sets, unique digital images or video), or in a historical context.

Any data that has long-term value, in a scholarly context or otherwise, requires both archiving and long-term preservation, much as books and journals are preserved in traditional libraries for current and future generations. Increasingly, scholarly production includes digital content that may need to be preserved in the form of digital data, such as data sets and multimedia, and does not lend itself to

conversion to traditional, longer-lived media.

Unfortunately, digital data has a high propensity to degrade, due to the impermanence of magnetic and optical media. For instance, the Digital Preservation Coalition cites "generic figures" for the lifetime of CD-ROMs from a minimum of as little as 3 months to a maximum of 30 years, based on environmental conditions.

Data formats also change over time. Documents written in a word processing application under an older operating system in the 1990s may not be readily usable today, without having an old computer set up to replicate the original environment in which those documents were written. Even documents written in the earliest versions of application programs may not be readable by the current versions of those same applications.

If data of lasting value is to be effectively stored over time, this may require periodically migrating it to newer media and data formats, and in some instances converting it to presumably longer-lived, standards-based formats. As with any migrations or conversions, however, there some risk of data loss or alteration. The Media Vault Project, an IST pilot project being conducted with multiple campus units as partners, is underway to better understand these issues and to lay the groundwork for future campus work in this area.

[1] Stephen J. Bigelow, "What is data deduplication," *SearchStorage.com*, March 22, 2007, http://searchstorage.techtarget.com/sDefinition/0,,sid5_gci1248105,00.html; Robert L. Scheier, "Eight Ways to Reduce Your Data Center Storage Costs," *Computerworld*, April 9, 2007, <http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=286057>

[2] Chris Mellor, "Storage over IP: Will SOIP be as successful as VOIP?", *Techworld*, May 11, 2006, www.techworld.com/storage/features/index.cfm?featureid=2508; Scott Lowe, "ATA Over Ethernet: Worth Considering?", *TechRepublic*, August 17, 2006, http://articles.techrepublic.com.com/5100-10878_11-6106721.html; Paul Virijevich, "Reduce network storage cost, complexity with ATA over Ethernet," *Linux.com*, July 3, 2006, <http://www.linux.com/feature/55334>

[3] In scholarly communications such as journal articles, as well as in online contexts, faculty, researchers, and students increasingly cite hyperlinks that point to underlying data sets or other supporting materials, whether created by themselves or others. However, those links are often ephemeral, failing when hosting services and the online contents they make available change in any number of major or minor ways. The campus Library, and the library and information sciences communities more generally, may have some solutions to offer in this area, such as as 'permanent' Uniform Resource Identifiers (URIs) and digital object identifiers (DOIs). These solutions may require the creation or sourcing of infrastructure on campus, at the UC-wide level, or outside the university. Ideally, these solutions can be overlaid on top of shared storage and made transparent, so that any item stored there – or in specific locations within that store – would automatically receive

DRAFT

permanent URIs and possibly also DOIs. This would be particularly appropriate for scholarly documents and materials referenced from those documents.

[4] Darren Waters, "Google helps terabyte data swaps," BBC News, March 7, 2007, <http://news.bbc.co.uk/1/hi/technology/6425975.stm>