Building a Massive, Distributed Storage Infrastructure at Indiana University

Anurag Shankar, Gerry Bernbom
University Information Technology Services
Indiana University
2711 East Tenth Street
Bloomington IN 47408
ashankar, bernbom@Indiana.Edu
tel: +1-812-855-9255
fax: +1-812-855-8299

Abstract

Anticipating an onslaught of data in research, administrative, and academic computing, Indiana University (IU) undertook in 1998 the ambitious task of architecting a massive, distributed storage infrastructure to meet its long-term storage needs. The task, now nearly complete, has resulted in the institution of the High Performance Storage System (HPSS), a hierarchical storage management (HSM) system, augmented by the Distributed Computing Environment Distributed File System (DCE DFS) acting both as a file system front end to HPSS and as a common file system (CFS) for IU campuses. Using gateways, IU's distributed storage system today currently offers a user on its eight geographically distributed campuses a capacity for securely storing and accessing nearly 200 Terabytes of data from any networked (Windows, Mac, or Unix/Linux) desktop equipped with a web browser.

HSM systems such as HPSS have traditionally been used by high-end users at large research labs (for example Los Alamos, Livermore, Sandia, Brookhaven National Labs in the U.S. and at CERN in Europe), at supercomputer centers (for example the San Diego Supercomputer Center), and at government agencies such as NASA. IU's installation is unique in two respects. It is the first production HPSS that is geographically distributed over a wide area network (WAN). Second, we have made available a high-end storage system in an academic setting not only to traditional, high-performance research users (for example astronomers, physicists, chemists, etc.), but also more generally (to users in economics, fine arts, apparel design, music, libraries, life sciences, etc.).

1 Re-centralization of Storage

Why build a centralized data storage system when typical personal computer hard disks today offer tens of gigabytes of storage at a very low (acquisition) cost? While it is certainly true that the availability of cheap, abundant personal storage capacity in the early nineties started a trend toward de-centralization of storage (from a highly centralized mainframe era), this trend is slowing. Researchers on university campuses a decade ago found to their delight that, for the most part, they were able to acquire (through grants) the resources necessary to store their data locally, on personal workstations or on servers in their offices or in departments. However, their initial enthusiasm soon dissipated when the high, after-purchase cost and effort of ensuring the integrity, protection, and long-term storage of data became apparent. As hard disk drive sizes have swelled to gigabytes and then to tens of gigabytes, backups have become increasingly costly, even painful. Also, enterprise-wide, the need for protecting institutional intellectual assets (in the form of research and other data created by users) has grown progressively stronger over the past decade, forcing many institutions to reconsider centralizing data storage.

2 Infrastructure Choices

The design of a storage infrastructure ultimately depends on a number of factors, chief among which are a) the amount of data to be stored, b) user data access patterns, and c) the available budget. With disk prices continuing their free fall, the storage industry seems to have agreed on storage area networks (SANs) to provide redundantly configured disk-based storage. However, SANs or alternatives that utilize spinning disks alone are simply not cost effective in building petabyte class data repositories at the present time. This leaves us with tapes and with HSM technology. The largest data repositories in the world are thus built using HSM systems. The tape to disk price ratio per megabyte of storage (especially at the high end) still favors tape over disk.

In a traditional HSM system, data bits are migrated seamlessly (from a user's perspective) from finely tuned, fast but (relatively) small disk caches (ours is a TB) to massive tape libraries (again, ours offers 200TB) when unused for a period of time. Metadata resides on disk forever (and is backed up carefully and redundantly). The user pays a price for having easy access to terabytes of data in the form of tens of seconds to possibly minutes-long delay in retrieving data bits that have migrated to tape. However, this appears to be acceptable for the majority of academic users who are happy to have access to massive storage capacities normally outside the scope of their individual or departmental budgets.

Armed with this information, we began looking for a HSM solution that provided a) long-term vendor viability, b) excellent hardware and software support, c) scalable performance, d) ease of access (preferably via a file system), and e) the ability to distribute software and hardware components geographically. At the conclusion of our request for proposal (RFP) process, only one contender remained, namely the High Performance Storage System. The HPSS[1] is the result of a fruitful collaboration between a number of government labs, academia, and IBM. It is not a vended solution in the usual sense; one pays instead a membership fee to join the HPSS collaboration. Each member gains access to the source code and is free to modify it within the mechanisms provided by the collaboration. Excellent software support is also included. Another attractive feature of HPSS is its ability to present a file system interface to data stored on tapes via DCE DFS[2], a distributed, scalable and secure file system.

At the high end, campus projects needing massive data storage at IU included candidates such as next generation high-energy physics experiments, with the potential to generate petabytes of data each year. With possible analysis times extending to decades, protection against software and hardware obsolescence is paramount. We felt that HPSS fit these needs and our environment well, by giving us long-term control over our destiny. [HPSS is also the HSM system of choice at some of the world's largest data repositories (for example SDSC, Brookhaven National Labs, CERN, etc.).]

Finally, while a tape-based system is ideal for archiving large files (tapes perform best when streaming), many campus users needed persistent, disk-based storage as well. In the past, this need was met (though inadequately) by the Novell Netware file system. By 1999 however, the future of Novell itself was in question and the existing Novell infrastructure was in urgent need of repair or replacement. With DCE DFS software already installed for HPSS purposes, it was natural to use it in lieu of Novell. DCE DFS is one of the most highly scalable and distributed file systems in use currently in the industry, to deliver high-end, secure file service. However, since DCE DFS clients are available only for a number of Unix flavors and for Windows NT4, it was clear that appropriate gateway servers would be needed to extend DCE DFS to the pervasive base of Windows and Mac desktops and servers on campus.

3 Building IU's Distributed Storage System

Our service design included campus users (using their personal workstations or departmental servers or our supercomputers) who either required massive, archival storage and/or who needed traditional, disk-based storage. A major design goal for us was also to provide storage ubiquitously, either via the web or via a file system front-end. Though these methods do not provide the highest performance, they were targeted for a non-savvy computer user due to the simplicity of use.

The majority of our users were located on two of IU's eight campuses, namely IU Bloomington (IUB) and IU-Purdue University at Indianapolis (IUPUI), located around fifty miles apart in central/south-central Indiana. Since the intercampus bandwidth (45Mbps) was insufficient to move massive amounts of data between campuses, we decided to experiment distributing IU's HPSS hardware and software across the two campuses. While the metadata engine remained at IUB, the actual user data was to reside where the user was located physically, either at IUB or at IUPUI. The idea was to use the intercampus link efficiently, to carry metadata traffic only. Users were to access their data over their local LAN at each campus via third party transfers. Extensive tests in partnership with IBM validated the idea and the experiment transformed into the first production instance of a remote HPSS mover at IUPUI in late 2000.

The file system front-end to HPSS is configured via "migrating" DFS servers. Data placed into HPSS via DFS arrives first in the DFS server disk caches, and later migrates to HPSS disk caches via a bi-directional DFS-HPSS link. The migrating DFS is thus a dedicated, external subsystem to HPSS. Static (i.e. non-migrating) DFS was also configured using separate DFS servers, with no link to HPSS, to provide the "Common File System" (CFS) service to the masses (directly, via DFS clients, and via SMB, Appleshare IP, and web gateways). Security for both HPSS and for CFS is provided through DCE (based on Kerberos 5).

We configured our core HPSS on a dedicated IBM RS/6000 SP located at IUB. This allows the eleven PowerPC "Silver" thin and wide SP nodes (which run core HPSS servers, disk/tape movers and migrating DFS servers) to communicate over the IBM SP switch at 130MB/s. Our supercomputer (another IBM SP) users are able to transfer data to/from HPSS using an ASCEND router at better than 100MB/s. A terabyte of IBM's serial storage array (SSA) disk attached to the eleven nodes forms the HPSS and migrating DFS disk caches. We use IBM's Magstar (3590E) tape drives in an IBM 3494 tape library and Storage Technology Corporation's 9840 "Eagle" tape drives in a STK 9310 tape library to store HPSS data at IUB. Remote HPSS disk and tape movers and a DFS server are configured on an IBM H70 server at IUPUI in Indianapolis. We have roughly 1TB of UltraSCSI RAID5 disk configured on the H70 as HPSS and migrating DFS disk caches. A number of IBM's Magstar drives inside an IBM 3494 tape library are SCSI-attached to the H70 at IUPUI.

HPSS is accessed in a high-performance mode via especially written Unix clients or more easily via FTP, DFS or via the web. We currently have around a thousand users distributed across various IU campuses, with roughly 55TB of data stored in HPSS.

IU's non-migrating DFS (or CFS) runs at IUB on several IBM's low-end B50 servers with IBM's UltraSCSI RAID5 arrays. Five Sun E220R servers run Samba[3], Netatalk[4], and Apache-SSL[5] servers which allow Win9x, Mac, and Linux users to access DFS from any networked desktop. The gateways are accessed by users as a single, round-robin DNS name. User authentication occurs securely (via modifications to Samba, Netatalk, and Apache server code[6]) directly against DCE. This allows no name space information to be maintained on the gateways, thus helping load balance and scale the service up as appropriate, without user impact. The non-migrating DFS servers and the gateways together form our CFS environment which is available to all campus users, either as a mapped drive under Windows, an an Appleshare IP volume on Macs, via smbmount or a native DFS client under Unix (or smbfs under Linux), and via the web. We are serving roughly 25,000 CFS customers currently with 250GB of data stored and backed up regularly.

4 Future

We are currently working in partnership with IBM to investigate developing an interface between IBM's high-performance general parallel file system (GPFS) and HPSS. This could enable high-speed, parallel, file system based data transfers between Linux clusters and HPSS (these clusters are currently served largely via low-performance NFS). We are also expanding the HPSS infrastructure at our Indianapolis campus (to nearly 400TB capacity) to support life sciences research and will start tests soon thereafter to mirror all HPSS data in real time across i-light[7], a newly installed high-speed optical fiber infrastructure between IUB and IUPUI. This should provide us with better protection against a disaster at either site. Finally, CFS is being extended to the IUPUI campus and will replace the local Novell infrastructure during 2002.

5 Conclusions

Indiana University is one of the few academic institutions to successfully anticipate and to build an ambitious infrastructure to provide massive data storage to its users. Using HPSS, a highly scalable and distributed hierarchical storage management system, along with DCE DFS and SMB, AppleShare IP and web gateways, a campus user at IU can store and access terabytes of data from their desktops. We have also found that it *is* possible to implement and to offer a high-end storage system to the masses, with significant cost savings over the long run.

We are happy to share our knowledge and experiences with anyone who is interested.[8]

References

[1] Information about the High Performance Storage System (HPSS) is available at the website http://www.clearlake.ibm.com/hpss/.

[2] IBM’s DCE website: http://www.ibm.com/software/network/dce/. IBM’s Transarc Labs DCE/DFS website: http://www.transarc.ibm.com/Product/.

[3] Samba project website: http://www.samba.org/.

[4] Netatalk project website: http://www.umich.edu/~rsug/netatalk/.

[5] Apache project website: http://www.apache.org/.

[6] Paul Henson’s mods for Samba/Netatalk/Apache are available at the URL http://www.intranet.csupomona.edu/~henson/www/projects/.

[7] Indiana’s high-speed research network website: http://www.i-light.iupui.edu/.

[8] Information about IU’s distributed storage services is available at the URL http://storage.iu.edu/. Our distributed storage services group website address is http://www.indiana.edu/~dssg/.