K. Keeton and E. Anderson. "A backup appliance composed of high-capacity disk drives," HP Laboratories SSP Technical Memo HPL-SSP-2001-3, April 2001. A shorter version of this paper appears in *Proc. of the Eighth Workshop on Hot Topics in Operating Systems (HotOS-VIII)*, p. 171, Elmau, Germany, May 2001.

A Backup Appliance Composed of High-capacity Disk Drives

Kimberly Keeton and Eric Anderson, Hewlett-Packard Labs
{kkeeton,anderse}@hpl.hp.com

Abstract

Disk drives are now available with capacity and price per capacity comparable to (and in some cases better than) nearline tape systems. Given disks' superior performance, density and maintainability characteristics, it seems likely that disks will soon overtake tapes as the backup medium of choice. In this paper, we outline the potential advantages of a backup system composed of high-capacity disk drives and describe what implications such a system would have for backup software.

1 Introduction

The world has an enormous rate of data creation, and this rate is growing over time. A recent study by Lyman and Varian at UC Berkeley's School of Information Management and Systems estimates that the world creates between 1 and 2 exabytes (10^{18} bytes) of data per year, the equivalent of 250 MB per person for each man, woman and child on earth [18]. They estimate that information production will increase by roughly 50% per year.

Further corroborating evidence is provided by Dr. Greg Papadopoulos, the Chief Technical Officer (CTO) of Sun Microsystems, who estimates that the I/O capacity and associated processing demands for decision support (DSS) databases double every 6 to 12 months [22], a rate supported by the large commercial systems summarized in the Very Large Database (VLDB) surveys performed by Winter, et al. [29] [30].

Regardless of the exact growth rates, the lessons are clear: data storage demands are huge and they are growing over time. An important challenge of this growth is the requirement of backing up this data for archival purposes and for recovery from various errors and failures.

Backups are done for multiple reasons [2] [4] [20]:

1. protection against user or software errors
2. protection against independent media failure (e.g., disk or other component failure)
3. protection against correlated media failure (e.g., site failure or disaster)
4. long-term storage of archival data, and
5. bulk movement of data between weakly connected sites.

All backup solutions are based on some form of data redundancy, with the precise solution dependent on the backup goal. User and application software errors are most often corrected through the use of online file system snapshots or the creation of backup tapes. Independent media failures are generally addressed through the use of hardware redundancy, such as RAID techniques. Correlated media failures must be solved through the use of additional uncorrelated media. For instance, sites ship locally-created backup tapes off-site or duplicate data over a wide-area network. Finally, archival storage uses easy-to-read-formats on very stable storage media.

Although existing solutions are mostly adequate for backup of today's systems, several challenges remain that can be improved by disk-based backup.

First, the poor random access performance of tape media means that partial restores of a data set take a long time, and that different users can't simultaneously share a single tape for restores. Disks have random positioning performance over 1000x better than tapes do.

Second, the expense of tape drives, and the need to verify a tape with a different drive because of head drift leads some administrators to skip testing their backup tapes. Disks have better sequential performance, so the time to test the backup is reduced. Furthermore, each disk has its own read/write head, eliminating the concern about head drift.

Third, some industries have legal requirements to store data for a long time period, which requires a very stable storage medium and a data format that will be decipherable in several decades. Tapes have had many formats over the recent decades and tape drives can have calibration problems. Conversely, disks hide the physical format, and have had few interfaces. Similarly with the integrated read/write head, disks have fewer calibration problems.

Finally, disk bandwidth and capacity are starting to outstrip tape bandwidth and capacity, leading to solutions that require multiplexing of disks and tapes. This problem will only be exacerbated as the performance of tapes continues to get worse compared to the performance of disks.

The characteristics of disk-based backup have implications for the creation of backup software. Backup software protects data by maintaining a read-only copy that cannot be inadvertently corrupted, and by providing an alternate (and possibly simpler) software path to that of a snapshotting file system. A backup system that keeps a fraction of its disks online may be able to approximate the performance of an online snapshot, allowing greater simultaneous sharing, while still maintaining the data protection properties of a backup. Key challenges lie in designing backup software for optimizing reliability and data integrity, scheduling the resources of the backup appliance, and developing APIs for giving users and applications more control over how backups are performed.

The remainder of the paper is organized as follows. In Section 2, we present data to support the observation

that disks are starting to win the capacity and price/GB race against tape. Section 3 describes the numerous performance, density and maintainability advantages that disks possess over tapes. We describe the implications that these disk characteristics present for backup software in Section 4. We review related work in Section 5 and conclude in Section 6.

2 Technology Trends

Although we traditionally think of magnetic tape as possessing higher capacity and lower price/GB than magnetic disk, technology trends demonstrate that these tides are turning.

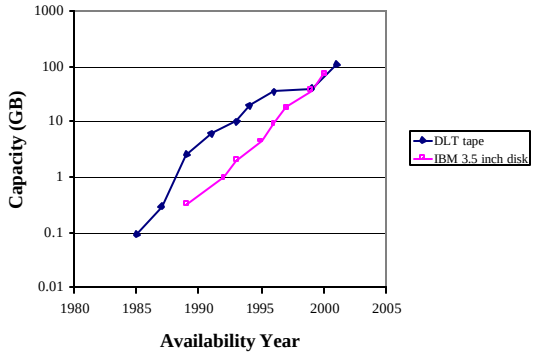

Figure 1: Capacity trends for magnetic tape and magnetic disks [8] [24].

Figure 1 shows that DLT (digital linear tape) tape's capacity lead over IBM 3.5" magnetic disks has shrunk over the last 15 years. In fact, with the introduction of Seagate's 180 GB Barracuda (scheduled to ship in the first quarter of 2001), disk has taken the lead in the capacity race [26].

Surprisingly, high-capacity disk drives provide comparable price/GB to tape media. Table 1 compares the media price per GB of disks and tapes, as well as the system price per GB, including an enclosure for disks, or a tape autoloader and tape drive for tapes. Disk media price is within 3X of tape media price, and disks cost less per GB than tapes, once the drive and tapes supported by the autoloader are taken into account.

Gray and Shenoy have made similar observations regarding the relative price between disks and tapes [6]. They report prices of $33/GB for server-class packaged disks and about $10/GB for a tape system, including a single drive and 15 tape cartridges.

Given the large capacity of disks relative to tape and the comparable price per capacity, it is time to rethink tape's role as the backup medium of choice.

3 Potential Benefits of Disk-Based Backup

Disks confer numerous benefits over tape, which may make them a more attractive alternative for use in backup systems. In this section, we enumerate the performance, density, legacy, maintainability, and lifetime benefits of disks and their implications for backup systems.

3.1 Performance

Disks offer a nearly 7X speedup over tape for sequential accesses; their average positioning time is over three orders of magnitude better than tape. These disk performance advantages affect the design of a backup appliance. First, due to their superior sequential bandwidth, disks will be better for creating and fully restoring backup volumes. This performance also allows for easier verification and more efficient data scrubbing operations. Furthermore, disk bandwidth scales more cheaply, as each disk adds bandwidth whereas only expensive tape drives improve tape system bandwidth.

Second, the superior random positioning performance of disk implies that disk will be better at partial restorations, which require the backup system to search for a particular file or directory. Data can be read directly from an online backup disk nearly instantaneously, allowing the backup system to provide performance similar to a file system snapshot stored in the traditional online disk subsystem. An additional benefit of disks' random performance is that a disk-based backup system is more shareable, in that it could easily satisfy restore requests from multiple users.

3.2 Density

Disk-based systems may provide potential packaging density advantages, which will be attractive for environments where space is at a premium, such as large ISPs.

If we design the appliance so that only a small fraction of the drives are powered on simultaneously, power and cooling requirements are reduced, and we can package the drives more densely. Keeping only a few drives on at a time is similar to how tape systems keep only a few tapes loaded. The delay for starting a drive is less than 10 seconds, which is comparable to tape load time [17].

The dimensions of a low-profile drive are 25.4 mm (H) x 101.6 mm (W) x 146.1 mm (D). In the 44.0 mm x 482.6 mm x 715.0 mm volume of one standard rack unit (commonly abbreviated as 1U), we could fit a grid four disks wide by four disks deep, assuming that all power supplies, cabling, etc. are placed "above" the disks in the package.

Using 60 GB Maxtor drives, our 1U backup appliance could hold 960 GB of disk storage, yielding a packaging density of 63.2 TB/m^3. In contrast, the ADIC FastStor22 tape autoloader holds 22 40 GB tape drives in a space 241.0 mm x 589.0 mm x 190.0 mm, yielding a packaging density of 32.6 TB/m^3. Thus, a disk-based backup appliance could provide roughly 2X more capacity per volumetric unit than a tape-based system.

3.3 Support for Legacy Devices

Disk-based backup systems may handle legacy archival devices better than tape-based backup systems. The

Enclosure Product	Enclosure Capacity	Enclosure Price (US$)	Media Capacity	Media Price (US$)	Media Price/GB	System Price/ GB
HP SC10 JBOD disk tray (assume 10 - 20 LP disks)	600 GB (min) to 1200 GB (max)	$4989.00	60 GB x 20	$4500.95	$3.75/GB	$15.82/GB to $7.91/GB
ADIC FastStor22 DLT8000 auto-loader (1 drive, 22 tape capacity)	880 GB	$9494.95	40 GB x 7	$466.95	$1.67/GB	$12.46/GB
AIT2 tape autoloader (1 drive, 19 tape capacity)	950 GB	$8816.95	50 GB x 10	$1133.95	$2.27/GB	$22.17/GB
IBM 3581 Ultrium tape autoloader (1 drive, 7 cartridges)	700 GB	$12178.00	100 GB x 1	$131.95	$1.32/GB	$18.72/GB

Table 1: Price per capacity for online magnetic storage [3] [12]. We show native data capacities for the various media; additional capacity may be available through the use of various compression techniques. Disk media data is provided for the Maxtor DiamondMax EIDE UDMA 3.5 low profile disk [17]. We calculate system price per GB as the price of the disk enclosure or tape autoloader plus the minimum number of media packages to fill the device. For example, the price/GB for the ADIC FastStor22 includes the price of 4-7 paks of DLT tapes.

magnetic tape market has been unable to agree on a single standard over the years, resulting in a proliferation of formats, such as digital audio tape (DAT), digital linear tape (DLT), advanced intelligent tape (AIT), 8mm Mammoth tape, and linear tape open (LTO). In order to restore data from a tape, an administrator must find a tape drive of the correct format to read the tape's contents. In contrast, the disk industry has far fewer standard interfaces, namely SCSI and (E)IDE/ATA. Although these standards have evolved over the years to increase the available bandwidth, in general, backwards compatibility has been retained. In addition, disks include their own read/write heads, eliminating the need to search for a separate drive to retrieve data.

3.4 Maintainability

Disk-based systems require less maintenance to continue operating properly than tape-based systems. Tape drives need to be periodically cleaned with a special cleaning cartridge in order to work properly. Tape drives also need to be periodically serviced [20] to make sure that head drift doesn't render a tape written on one drive unreadable on all others. In contrast, disk drives are enclosed media, which don't require cleaning, and each disk includes its own read/write head, so head drift is not an issue. Like tapes, disks should be periodically read to ensure the disk mechanism still works properly.

3.5 Lifetime

Empirical evidence suggests that disks could have a shelf life as long as, and possibly longer than, the shelf life of tapes, implying that they may ultimately be better archival media. Although most DLT tape product literature states that DLT tapes have a shelf life of at least 30 years [21], system administration experts generally advise re-recording data that is stored longer than three years, to ensure that the data remains readable [20]. Disks generally come with warranties for three years for desktop (E)IDE/ATA disks to five years for enterprise-class SCSI disks. According to experts in the disk drive industry, the operational lifetime of a disk drive, which is determined by the number of power on hours, could conceivably be as long as ten years [1]. Powering the disk off, with its head parked away from the disk surface, might provide such longevity. Disk drives with even longer lifetimes and the appropriate non-operating characteristics could be designed for use as archival removable media, but so far the market has not demanded a disk drive optimized for such purposes.

4 Implications for Backup Software

As described in the previous section, disks provide many potential hardware-related benefits for the development of a backup appliance. In this section, we describe the implications of disk-based backup for the development of backup software.

4.1 Design for Reliability

As described in Section 1, one solution to data loss due to user errors and application software bugs is using versioning, snapshots or checkpoints in the file system [9] [10] [11] [19] [23] [25]. Online copies provide fast access in the event of inadvertent data deletion or corruption.

Unfortunately, snapshots or versions do not protect against kernel or filesystem bugs. Filesystems are complex, and online snapshots and versioning increase the complexity. Backup software, conversely, does not support updates to data, so it can treat the backup media as write-once, lowering the chance of accidental corruption. Furthermore, backup software provides a separate code path for archiving data, which avoids correlated system software failures.

The backup system's focus on reliability means that we can design it to trade off reliability for performance. One approach is to increase the checks applied by the

backup software. If each software layer checks the results of computations it has requested from the other layers, the software can more easily diagnose when a software error occurs and assign accountability. For example, a component could immediately re-read data after writing it to verify that it has been written correctly.

As another example, if the backup software computes a checksum or hash for each data block written to a backup archive, it is easy to verify whether the data has become corrupted. As mentioned in Section 3.4, a backup disk's contents should be periodically read to ensure that the disk mechanism still works properly. During this scan, checksums and hashes could be recomputed, thus verifying that the data remains uncorrupted.

As a final example, if the backup system uses a redundant representation for the data, the software could simulate failures of the hardware and check that the data is correctly re-constructed before considering the write to be successful.

4.2 Design for Sharability

Fast random disk seeks make it possible to multiplex a backup appliance among multiple users, permitting fair sharing of user-initiated backup and restore. To bound the amount of time spent seeking to 1%, a disk-based system with a .01s seek time can schedule a particular user for 1 second. Conversely, a tape-based system with a 60s seek time must spend 6000s on each user. A disk-based system can therefore share the appliance between users, as a 1s quanta is acceptable for restore operations. The ability to schedule and share the backup appliance is especially valuable for xSPs, who may want to provide a backup "service" for the clients they host.

A disk-based backup system can also achieve the "instant restore" behavior of an online shapshot while retaining the data protection benefits of a separate backup system. The directory structure is restored immediately while the actual data is restored on-demand (and in the background), as is done in hierarchical storage management systems [27].

A disk-based backup system has to schedule on-demand requests, background restores, and periodic data scrubbing. Recall from Section 3.2 that we only allow a small number of drives in each 1U slice to be active at a time. Thus, in addition to scheduling different types of requests, the backup system must schedule the order in which drives are made active.

However, we have the advantage over many scheduling systems that all of the operations required for a restore are generally known in advance, which makes the scheduling task easier. Furthermore, if we have multiple mirrors of a particular file for further redundancy in the backup system, we can choose which of the copies to restore, providing more flexibility in the scheduling.

We therefore envision an interface where the administrator controls the fraction of the backup utility assigned to each user (a la proportional share scheduling [28]), and the system is responsible for scheduling the order of the users and the operations for each user efficiently (for example by earliest deadline first).

4.3 Design for Longevity

A final opportunity for backup software is to automatically convert data formats commonly used today into formats that will be easy to read many years in the future. For instance, the backup system could extract the text from a Microsoft Word document, or generate a bitmap of each page. This service is most useful for important documents, whose contents must be preserved for the long-term. The backup system could define an API for letting users specify which files should be translated, and for what source and destination formats. It could even automatically translate recognized formats when the archival backups are created.

5 Related Work

Although several researchers have made the observation that cheap disk storage could change the way we think about protecting data, we are aware of no work that explores this concept in depth. The closest work is Gray and Shenoy's paper on rules of thumb in data engineering [6]. The authors observe that disks are replacing tapes as backup devices, because disk prices are approaching nearline tape prices. They derive the useful rule of thumb that nearline tape:online disk:DRAM storage price ratios are approximately 1:3:300. We observed the same tape:disk ratio in Section 2 for media prices, but found that the price of packaged disks is comparable to the price of nearline tape systems.

Santry, et al., have proposed that large cheap disks should be leveraged to provide a file system that retains all important versions of a file [25]. Their Elephant file system manages data using file-grain user-specified retention policies, an approach that contrasts with checkpointing file systems, such as Plan-9 [23], AFS [11], and WAFL [10], where file system-wide policies govern periodically generated snapshots of entire file systems. Many of these systems also provide backup that is integrated with file system snapshotting [7] [14] [15] [23]. All of these techniques apply to a disk-based backup system, as well. The key goal that must be preserved is the distinction between the read-only backup copy and the live file system to provide data integrity.

Hierarchical storage managers, such as Tivoli [27] and IBM's ADSM [13], manage the migration of data between magnetic disk, magnetic tape, and optical disk drives. Upon access to files stored on slow media, the HSM system restores the file system structure immediately, and data is restored as files are accessed. As described in Section 4.2, a disk-based backup system can leverage this technique to achieve the instant restore behavior of online snapshots.

Chervenak, et al., provide a thorough survey of backup

techniques, including device-based vs. file-based backup schemes, full vs. incremental backups, optional data compression, and techniques for backing up a live system [4]. The authors then classify several research and commercial backup systems according to these parameters.

6 Conclusions

An astounding amount of data is created each year, and it is not clear that current tape-based backup technology can keep up with this data production rate. We observe that tapes no longer hold their tremendous capacity and price per capacity advantages over disks: disks are now available with capacity and price per GB comparable to that of tape. We believe that these technology trends will result in a shift towards using disks for backup.

Disks confer tremendous hardware-related benefits for a backup system. We described the advantages of disk performance, packaging density, support for legacy devices, maintainability, and potential lifetime benefits relative to tapes.

The characteristics of disk-based backup have implications for the creation of backup software. Backup software protects data by maintaining a read-only copy that cannot be inadvertently corrupted, and by providing an alternate (and possibly simpler) software path to that of a snapshotting file system. A backup system that keeps a fraction of its disks online may be able to approximate the performance of an online snapshot, allowing greater simultaneous sharing, while still maintaining the data protection properties of a backup. Key challenges lie in designing backup software for optimizing reliability and data integrity, scheduling the resources of the backup appliance, and developing APIs for giving users and applications more control over how backups are performed.

7 References

[1] D. Anderson, Seagate Technology, personal communication, January 2001.

[2] E. Anderson and D. Patterson. "A retrospective on twelve years of LISA proceedings." *Proc. of LISA XIII*, pages 93 - 105, November 1999.

[3] Buy.com. Various price quotes from `www.buy.com`.

[4] A. Chervenak, V. Vellanki and Z. Kurmas. "Protecting file systems: a survey of backup techniques," *Proc. of Mass Storage Symp.*, 1998.

[5] J. Douceur and W. Bolosky. "A large-scale study of file system contents," *Proc. of SIGMETRICS '99*, pages 59 - 69, May 1999.

[6] J. Gray and P. Shenoy. "Rules of thumb in data engineering," *Proc. of Intl. Conf. on Data Engineering*, February 2000.

[7] R. Green, A. Baird and C. Davies. "Designing a fast, online backup system for a log-structured file system," *Digital Technical Journal*, October 1996.

[8] E. Grochowski. "IBM leadership in disk storage technology," `www.storage.ibm.com/storage/technolo/grochows/grocho01.htm`.

[9] R. Hagmann. "Reimplementing the Cedar file system using logging and group commit," *Proc. of Symp. on Operating Systems Principles*, pages 155 - 162, November 1987.

[10] D. Hitz, J. Lau and M. Malcolm. "File system design for an NFS file server appliance," *Proc. of Winter USENIX Technical Conf.*, pages 235 - 246, January 1994.

[11] J. Howard, M. Kazar, S. Menees, D. Nichols, M. Satyanarayanan, R. Sidebotham and M. West. "Scale and performance in a distributed file system," *ACM Transactions on Computer Systems*, 6(1):51-81, February 1988.

[12] IBM 3581 Ultrium tape autoloader. `www.storage.ibm.com/hardsoft/tape/3581/prod_data/g225-6851.html`.

[13] "IBM ADSTAR distributed storage manager (ADSM) - distributed data recovery white paper," `www.storage.ibm.com/storage/software/adsm/adwh-ddr.htm`.

[14] J. Johnson and W. Laing. "Overview of the Spiralog file system," *Digital Technical Journal*, 8(2):5-14, 1996.

[15] S. Lammert. "The AFS 3.0 backup system," USENIX *Proceedings of the 4th Conf. on Large Installation System Administration*, pages 143 - 147, October 1990.

[16] E. Lee and C. Thekkath. "Petal: distributed virtual disks," *Proc. of Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS-VII)*, pages 84 - 92, October 1996.

[17] Maxtor DiamondMax. `www.maxtor.com/products/DiamondMax/DiamondMax/QuickSpecs/42072.txt`.

[18] P. Lyman, H. Varian, J. Dunn, A. Strygin, and K. Swearingen. "How much information?" `www.sims.berkeley.edu/how-much-information/`.

[19] K. McCoy. *VMS file system internals.* Digital Press, 1990.

[20] E. Nemeth, G. Snyder, S. Seebass and T. Hein. *UNIX system administration handbook.* Prentice Hall, Inc., Upper Saddle River, New Jersey, 1995.

[21] Nstor Technologies. "DLT tape technology frequently asked questions," `www.nstor.com/support/faqs/dlt/answers.html`.

[22] G. Papadopoulos. "Moore's Law ain't good enough," keynote speech at Hot Chips X, August 1998.

[23] R. Pike, D. Presotto, K. Thompson and H. Trickey. "Plan 9 from Bell Labs," *Proc. of United Kingdom UNIX Systems User Group (UKUUG)*, pages 1 - 9, July 1990.

[24] Quantum Corporation, *The DLTtape handbook*, `www.dlttape.com/technology/handbook/`.

[25] D. Santry, M. Feeley, N. Hutchinson, A. Veitch, R. Carton and J. Ofir. "Deciding when to forget in the Elephant file system," *Proc. of the 17th Symp. on Operating Systems Principles (SOSP-17)*, pages 110 - 123, December 1999.

[26] Seagate Technology, "Seagate delivers disk drive with world's highest capacity," `www.seagate.com`.

[27] Tivoli space manager. `http://www.tivoli.com/products/index/space_mgr/`.

[28] C. Waldspurger and W. Weihl. "Lottery scheduling: flexible proportional-share resource management," *Proc. of the First USENIX Symp. on Operating Systems Design and Implementation (OSDI)*, pages 1-11, November 1994.

[29] R. Winter and K. Auerbach. "Giants walk the earth: the 1997 VLDB survey," *Database Programming and Design*, volume 10, number 9, pages S2 - S9+, September 1997.

[30] R. Winter and K. Auerbach. "The big time: the 1998 VLDB survey," *Database Programming and Design*, volume 11, number 8, August 1998.