A NEW APPROACH TO DISK-BASED MASS STORAGE SYSTEMS

Aloke Guha
COPAN Systems, Inc.
2605 Trade Center Drive, Suite D
Longmont, CO 80503-4605
Tel: +1-303-827-2520, Fax: +1-303-827-2504
e-mail: aloke.guha@copansys.com

Abstract

We present a new approach to create large-scale cost-effective mass storage systems using high-capacity disks. The architecture is optimized across multiple dimensions to support streaming and large-block data access. Because I/O requests in these applications are to a small fraction of data, we power-cycle drives as needed to create a very high-density storage system. Data protection is provided through a new variant of RAID, termed power-managed RAID™, which generates parity without requiring all drives in the RAID set to be powered on. To increase bandwidth, the system employs several concurrency mechanisms including load balancing I/O streams across many parallel RAID controllers. A number of optimizations are made to the interconnection architecture and caching schemes that lowers the cost of storage to that of typical automated tape library systems while exploiting the performance and reliability advantages of disk systems. Recent results at COPAN Systems have proven the viability of this new storage category and product.

1. Background

Growing reference information, regulatory data and rich media archives [1] are providing new impetus for mass storage systems. Unlike transaction applications, these systems access a small fraction of the data infrequently or on a scheduled basis. One such application is backup and restore that is characterized by large-block streaming writes and infrequent reads. With their need for large-scale storage, these applications are very sensitive to cost. Historically, they have been addressed by automated tape libraries. While concerns with performance, lack of data protection, and the reliability of tapes [2] have always favored disk systems, their limited capacity and high cost have prevented them from replacing tape. The advent of high-capacity SATA drives and their cost approaching that of tape motivates us to revisit this design issue.

Existing approaches to high-capacity storage have been incremental, typically substituting Fibre Channel (FC) disks with lower cost ATA disks in standard RAID arrays. This does not result in increased storage density or the cost per unit storage of tape libraries which usually have a 3X advantage over disk systems. Therefore, a fundamentally different approach is required.

We believe the best approach is to use an application and workload-driven architecture to provide the performance and reliability of disks at the scale and cost of tape.

2. A New Approach: Application Specific Storage System

The specific needs of archival storage applications help define the architectural constraints. These include:

- I/O requests to large block with sequential or predictable access

- Performance metrics based on data rates (Mbytes/sec) and not on I/Os per sec (IOPs)
- Time to first byte should be in ms to seconds, desired for mission-critical systems

Given the above, we can eschew many common but complex features of traditional disk storage architectures:

1. *No need for large primary RAM cache*. Since the storage is not used for high-transaction I/O, there is little need for large shared caches, except for reads.
2. *No need for high-speed switched interconnection from host to disks*. With more tolerance to latency than transactional systems, access from these applications do not benefit from non-blocking switched connectivity.
3. *No need to access all the data all the time*. With access to a small fraction of the data (e.g., <5% in tape systems), a majority of disks could be taken off-line if the mean latency is bounded. In addition, most writes can be done sequentially.
4. *No need to limit capacity based on interconnection bandwidth.* Since the capacity to data rate ratio is high, the ratio of total drive bandwidth to interconnect bandwidth can be higher than traditional disk systems.
5. *Ensure data availability*. Given the scale of data they maintain, archival storage must have high reliability, data protection and availability.

COPAN Systems' architecture is based on power management of a large number of SATA drives. The basic concept was first used in the MAID (massive array of idle disks) project [3] that examined tradeoffs in disk power consumption and performance. The results showed that a MAID with cache can effectively support most reads from a large database archive.

To meet enterprise class storage needs using a power-managed disk design, we had to satisfy multiple criteria. These included data protection, scalable capacity, high bandwidth, storage and data manageability, small footprint, and a cost equal to or better than tape storage costs. This introduces a number of design guidelines:

1. *Provide parity based redundancy*: Tradeoff redundancy with effective cost of storage. RAID 1 provides 100% redundancy, but it also doubles the cost per unit storage.
2. *Ensure data protection and performance with drives power-cycled*. Optimize tradeoffs between power cycling, storage density, performance, redundancy and cost. Keeping 30%-50% of drives online implies 10X more data online than tape systems.
3. *Ensure that I/O rate and parity protection are maintained, even when disks are in transition during the power-cycling of individual drives*.
4. *Provide ways to scale total bandwidth of the storage system*

An implicit design objective was the optimal packing and configuration of the drives. The architecture that works best for a given volume turns out to be a three-level interconnect. If the total number of drives in the system is N, then N is decomposed into a 3-tuple, such that

$$N = s.t.d$$

where s is the number of storage enclosures or shelves, t is the number of "stick", or column of disks, in the each shelf unit, and d is the number of disk drives in each stick in a shelf (Figure 1). All I/O requests to SCSI disk volumes arrive over FC links at the system controller which maps the logical volume to physical volumes on shelves connected by FC.

If physical constraints of the rack can be satisfied, N can be chosen to support very large storage capacities in a single system. The packaging of drives must also provide adequate heat dissipation so that the disks operate at or below the specified ambient temperature.

We provide brief descriptions of the data protection, performance and reliability of the system.

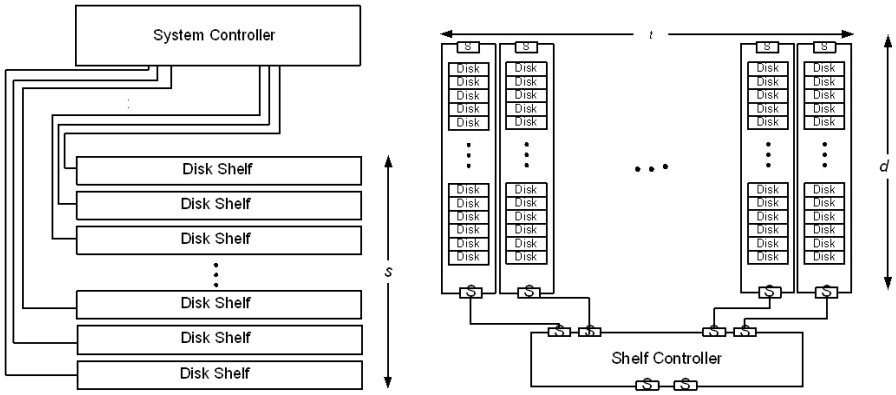

Figure 1. Three-tiered storage: decomposition into shelf, stick and disk

2.1. Efficient Data Protection: Power-Managed RAID™

Data protection using parity can be provided by power-cycling either full RAID sets or individual drives within the RAID set. COPAN Systems supports both approaches. However, power-cycling individual drives within the RAID set, termed power-managed RAID™ (PM-RAID™), has the advantage of keeping more RAID volumes on-line (assuming nominally 1 volume per RAID set) and reducing power swings.

In PM-RAID™, all data is written sequentially from drive to drive and the parity drive is fixed. At a minimum, only one data drive and the parity drive are powered on. When writing initially to drive D0, the data on the parity drive P will also be D0 if the parity drive was initialized to zeros. When the writes exceed the capacity of D0, the second drive D1 is powered up and data is written to it, while drive D0 is powered down. At this point, the parity drive P will contain the XOR of the earlier data written to D0 and the new data written to D1. Similarly, after the writes to the third drive, the parity drive will contain the XOR of the data from drives D0, D1 and D2. Thus, after writes to all data drives in the RAID set, the parity drive will indeed be the parity for all drives in the RAID set. On the failure of a data drive, all drives in the RAID set will be powered on to reconstruct the failed data drive as in a normal RAID reconstruct.

Figure 2 shows data written sequentially using a stripe size s of 1. If more bandwidth is desired, a larger stripe size s, $1 \leq s \leq n$ in an n+1 RAID, should be used. The most power-efficient and least bandwidth case is $s = 1$ with only 2 drives powered on, while the maximum bandwidth case is $s = n$ when all n+1 drives are powered on. Note that the $s = n$ case corresponds to traditional RAID 4 [4]. PM-RAID™ is therefore the most generalized version of RAID 4 under power constraints.

Various configurations of the number of on-line volumes and associated bandwidth are possible with PM-RAID™. For example, if there are 1000 data drives in the system that are set up as 4+1 RAID sets with a budget to power only 25% of all drives, the maximum number of volumes that can be powered is 125, or 50% of all volumes or 2X the drive fraction

powered on. Each volume can be accessed sequentially at a data rate possible from one disk drive.

PM-RAID™ utilizes a few always-on mirrored drives that maintain information on volume metadata, drive health, and read and write caches. Unlike previous approaches on disk on disk caching [5] used for performance, write caching is used to ensure that parity generation and I/O data rates can be sustained during power transitions of the drives. In general, including the spare and the metadata drives, less than 30% of all drives can be kept online while maintaining high bandwidth given sufficiently large number of disks in the system.

To ease storage management, the drive power management is transparent to the end user or application. The stripe width of the PM-RAID™ as well as any striping of the user's volumes across RAID sets in the shelves is managed by the system and not the user.

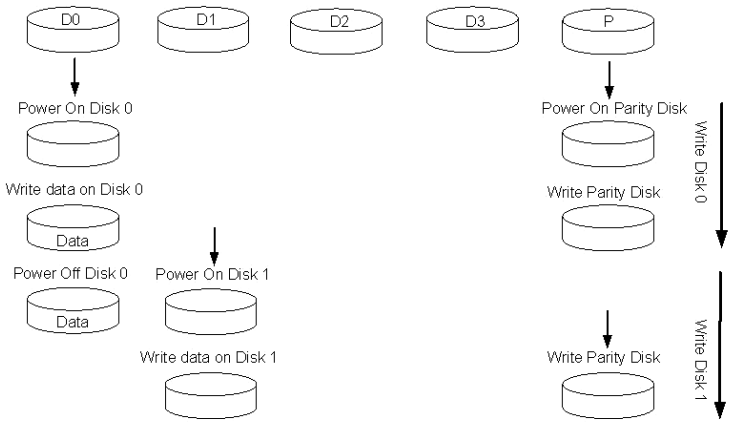

Figure 2. Writes in a power-managed RAID™

2.2. Increasing Performance: Concurrent I/O

Power management limits the number of drives powered on at any time. This limits the total I/O per storage shelf. To increase aggregate bandwidth, we use a number of concurrency techniques.

First, because the system controller has independent FC connections to storage shelves (Figure 1) and each storage shelf contains its own RAID controller, write and read bandwidth is increased *s*-fold with *s* shelves. Each shelf RAID controller can stripe data across up to *t* sticks.

Second, since each volume on the shelf is minimally comprised of 1 active data drive, the bandwidth can also be increased by striping (RAID-0) logical volumes from the system controller across volumes in the shelves.

Third, more I/O concurrency between the shelf controller and the SATA drives is provided in the stick using a custom SATA data and command router. This router provides concurrent access and command queuing of I/O to all *d* drives in a stick. This enables high-bandwidth I/O operations across a large number of drives in the shelf.

Finally, we use disk caches to cost-effectively ameliorate latency effects of disk spin-up, if there is no a priori information on access patterns. When write requests can be scheduled, as in a backup, a currently powered-down disk can be scheduled to power up with sufficient lead

time before the end of file is reached on the current drive. When the write requests cannot be scheduled, the data can be redirected to a spinning drive cache and later staged to the true disk target. Read disk caches also store data from all data drives. A read request to a powered-off drive is directed to the read cache, while the target drive is powered up. In the case of a read cache miss, our current measurements indicate that the read penalty is well below 10 seconds.

We also note that for streaming applications that access large blocks of data, the read latency of a few seconds on a cache miss is usually dwarfed by data transfer time.

2.3. Increasing Data Reliability: Effect of Power Management

Using high-capacity ATA drives and high-density drive configuration to create PB-sized storage systems raises the importance of data protection and data availability. We address the issue in many ways.

First, power-cycling the drives has a direct impact of increasing disk life and therefore system reliability. With 1000 drives and a drive MTBF of 400,000 hrs, the expected first failure of a drive is 18 days. Such low MTTF implies frequent drive swap-outs that is not be acceptable for most data centers. With PM-RAID™, if drives are powered with an average duty cycle of 25%, the net effect is to extend the effective drive MTTF to 4X the nominal value, or 1.6M hours, greater than typical SCSI or Fibre Channel.

Second, we closely monitor the total power cycles, also known as contact start-stops (CSSs) since all drives have a specified maximum CSS. We therefore ensure during operations that all drives are all well below the CSS threshold specified by the drive vendor.

Third, to accommodate drive failures that can put a large amount of data at risk, we use a unique proactive replacement mechanism. By continuously monitoring drive attributes, embedded controller intelligence determines whether a drive susceptible to failure should be replaced. Proactive drive replacement enables recovering data from the drive before failures, allowing the system to protect data even as failing drives are replaced.

3. Results

At the time of writing, we have demonstrated end-to-end I/O from the user to the drives. We have proved two important concepts. First, PM-RAID™ and the interconnect architecture perform as expected, with linear scaling of bandwidth with increasing I/O load. With the current design, 1000 MB/s or more bandwidth is possible in a single system. Second, the implementation shows that close-packing of drives in the system kept environmental attributes within specifications so data reliability is assured to be better than traditional disk systems. More detailed results will be presented at the conference.

REFERENCES

[1] Fred Moore, Are You Ready for MAID Technology? July 8, 2003, Computer Technology News, http://www.wwpi.com/lead_stories/070803_3.asp

[2] Bob Cramer, 'It's the restore, stupid!" April 23, ComputerWorld, http://www.computerworld.com/hardwaretopics/storage/story/0,10801,78483,00.html

[3] Dennis Colarelli, Dirk Grunwald et al, The Case for Massive Arrays of Idle Disks (MAID), Usenix Conference on File and Storage Technologies, Jan. 2002, Monterey CA.

[4] David A. Patterson, G. Gibson, and Randy H. Katz, "A Case for Redundant Arrays of Inexpensive Disks (RAID)," SIGMOD 88, p. 109-116, June 1988.

[5] Y. Hu and Q. Yang, DCD -- Disk Caching Disk: A New Approach for Boosting I/O Performance. Proc. of the 23rd ISCA95, Philadelphia, PA, 1995.