Most managers and business owners would agree that a company’s most valuable asset is its information. In fact, losing even a few bytes of information can have catastrophic results. For instance, online transaction processing databases typically log thousands to millions of transactions a day, but critical portions of these transactions, such as the account balance field, comprise only a few bytes of data. Likewise, a single corrupted block of data can render an entire file unreadable. Protecting data is crucial to a company’s survival in today’s competitive business environment.

System administrators may feel that because they store their data on a redundant disk array and maintain a well-designed tape-backup regimen, their data is adequately protected. However, undetected data corruption can occur between backup periods—backing up corrupted data yields corrupted data when restored. Scenarios that can put data at risk include:

- Controller failure while data is in cache
- Power outage of extended duration with data in cache
- Power outage or controller failure during a write operation
- Errors reading data from disk
- Latent disk errors

Dell|EMC storage systems protect against data corruption in all of these scenarios. These systems offer a comprehensive set of data integrity features that protect data before, during, and after it is stored.

Data Integrity: The Dell|EMC Distinction

By Michael H. Darden

The Dell|EMC partnership has expanded the existing line of Dell™ PowerVault™ storage servers with the new Dell|EMC storage systems. This article describes how the Dell|EMC products protect data before, during, and after it is stored.

Protecting data before it is written

Because disks are mechanical devices, they are relatively slow by nature. Under certain conditions, host servers can realize enormous performance gains by caching—reading and writing directly to RAM on the storage system—rather than waiting for data to be written to disk. In particular, write-back caching is popular; when a host writes data to the storage system, the storage system sends an acknowledgment back to the host after the write is received in cache but before it is written to disk. Dell|EMC products offer multiple gigabytes of cache, and thus provide electronic-speed access to several billion bytes of data.

However, cache memory is volatile: if power is removed, its contents are lost. If a system uses write-back caching and it unexpectedly loses power, multiple gigabytes of data can potentially be lost. Additionally, because the cache is usually physically located on the storage processor board, cached data can be lost if the storage processor fails. Caching introduces multiple vulnerabilities into the storage system, but Dell|EMC systems greatly reduce the risk of cache data loss by using cache mirroring and cache vaulting.

The technique of cache mirroring

Cache mirroring protects against cache data loss that occurs when a storage processor fails. As a prerequisite to cache mirroring, the storage system must have redundant storage processors with failover capability. Half of the write cache on each storage processor is dedicated to holding a copy of its partner’s write cache. If one controller fails, the other controller assumes ownership of the failed processor’s storage and continues where it left off.

A complex operation, cache mirroring involves transferring large amounts of rapidly changing data between controllers. Many storage systems are designed so that the cache data path between controllers is the same as the data path (for example, SCSI bus or Fibre Channel loop) to the disk drives. In these designs, cache mirror traffic can have a
Battery backup has several disadvantages and hidden costs. If a storage processor fails during the mirroring operation, will the controller have the correct data in cache? What happens if the communication path between the processors has problems?

Dell EMC storage systems effectively protect against the problems introduced by cache mirroring. Through Dell EMC Control Center Navisphere® management software, these systems provide a high level of control over the amount of cache allocated to the mirror. They also include a dedicated Cache Mirroring Interface (CMI) bus in hardware, which separates the mirrored-cache data path from the main data path to minimize the impact of mirroring on overall storage system performance. In addition, should a Dell EMC storage system encounter a problem with the CMI path or a storage processor, it will automatically flush the cache to disk and disable write-back caching until the problem is corrected.

**The problem with battery backup**

Battery backup is a popular method of protecting the cache against unexpected power loss. Internal or external batteries keep power applied to the cache memory chips if the external power source (the wall power) is removed. Because memory chips consume minimal power, small, low-capacity batteries can be used to maintain the cache, sometimes resulting in nominal cost savings. However, battery backup has several disadvantages and hidden costs.

**Extended power outage.** Most battery backup systems have a cache hold time (the period during which they can sustain cache contents) of up to 72 hours, or three days. If the power outage lasts longer than the cache hold time, the cache contents are lost. Examples of power outages that have lasted more than three days include those caused by Hurricane Andrew (south Florida and Louisiana, 1992), flooding from Tropical Storm Allison (Houston, 2001), and the destruction surrounding the World Trade Center (New York, 2001).

**Difficulty determining cache hold time.** All batteries, even those from the same manufacturing lot, vary in capacity. The capacity of each battery in higher quality battery-backup systems can be tested, but testing often involves pulling the battery offline for several hours. In addition, capacity changes with time and temperature, and it changes based on the battery’s history. Because precise battery capacity is difficult to determine at any given point in time, system administrators will never know exactly how much time remains on the cache batteries during an unexpected power outage; the time will likely be different, even for identical systems.

**Battery management.** Cache batteries do not last forever, and many need to be periodically reconditioned to maintain their rated capacity. Most need to be replaced every two years, and during that time the battery must be constantly monitored to ensure it has enough capacity to maintain the cache for the expected time period.

The types of batteries commonly used for backup—nickel cadmium (NiCad), nickel metal hydride (NiMH), and lithium ion—have periodic conditioning or replacement requirements. These requirements can lead to asset tracking and maintenance problems, particularly if an administrator is responsible for multiple storage systems or if the equipment is redeployed to a different system administrator. Tracking the age and condition of every cache battery can be difficult, even with proactive notices from the manufacturer.

**The benefits of cache vaulting**

Dell EMC products employ a simpler, more reliable cache preservation scheme called cache vaulting. In this scheme, a Dell EMC standby power supply (SPS) keeps the entire storage system powered until all cache data is stored safely on disk, after which an orderly system shutdown is executed. This process occurs within a matter of minutes and indefinitely protects the data, regardless of the power outage duration.

The Dell EMC SPS does not suffer the management and maintenance problems of other battery systems because it uses sealed lead-acid batteries. This proven battery technology sacrifices size and weight for extremely long shelf life and zero reconditioning requirements—perhaps not a good trade-off for notebook computers, but perfect for cache vaulting. Cache vaulting and the Dell EMC SPS are standard features of every Dell EMC product.

**Protecting data while it is being written**

Most disk storage systems provide some level of protection against failed disks, one of the most common being RAID-5. The RAID-5 level does not require a complete copy of data for redundancy. Instead, the equivalent of one disk from the RAID-5 disk group contains parity information. When a disk fails, parity information and data from the remaining operational disks are used to reconstruct the data from the failed disk.

The RAID-5 write process (see Figure 1) first reads old values from the target location and from parity. It calculates the new parity and then writes the new data and adjusted parity to disk. Data and parity may be written at different times, but under normal circumstances the data and parity remain consistent.

When the host requests a read from a failed disk, the storage processor must reconstruct the data by reading data and parity from operational disks and then reconstructing the missing data. If the parity is consistent, the data is correctly reconstructed (see Figure 2).

One challenge in maintaining parity is to make sure that the data and the parity remain consistent under all conditions. In other words, at every instant the parity must accurately represent the complete set of data it protects. Parity consistency is critical: if a disk fails, inconsistent parity will cause the missing data to be incorrectly
reconstructed, thus producing corrupted data (see Figure 3). Worse yet, this condition can go undetected.

Unfortunately, data and corresponding parity are not always simultaneously updated. Depending on the individual disk load, parity can be updated before or after the user data. Parity can become inconsistent if a problem occurs in the storage system—power failure to a storage processor, hardware failure in the storage processor, or power failure to a disk enclosure—after the data is written but before its parity is updated (see Figure 4). If the storage system cannot preserve the status of a write in progress, this condition can go undetected.

All RAID-5 storage systems allow data reconstruction from failed disks, but not all do a good job of assuring consistent parity. Dell | EMC storage systems have patented safeguards for maintaining consistent parity.

**Nonvolatile write-in-progress records**

The first line of defense against inconsistent parity is nonvolatile random access memory (NV-RAM). NV-RAM is persistent; unlike RAM used for caching, NV-RAM holds its contents even when power is removed. With each write operation, Dell | EMC storage systems record in NV-RAM the disk group identifier, the starting sector into which the data is to be written, and the number of sectors to be written. This write-in-progress record remains until the write operation (including the parity update) is complete. Should the power unexpectedly fail, the storage system can quickly discern what write operations were in progress and correct the parity accordingly.

Because NV-RAM is physically located on the storage processor board, it can fail if a storage processor fails. However, Dell | EMC systems protect against inconsistency caused by failed controllers by using advanced data formatting at the sector level.

**Sector-level protection**

Safeguards against data corruption are embedded in every sector of Dell | EMC data. When disks are formatted, they are divided into a series of data blocks or sectors. The standard size for a data sector is 512 bytes, but Dell | EMC products are formatted into 520 byte sectors: 512 bytes of user data plus 8 bytes of validation data. Part of the validation area

---

**Figure 1. RAID-5 write sequence**

**Figure 2. RAID-5 reconstruction using consistent parity**

**Figure 3. RAID-5 reconstruction using inconsistent parity**

**Figure 4. RAID-5 write sequence resulting in inconsistent parity**
contains a time stamp, a write stamp, and a shed stamp that are used to ensure parity consistency if the controller fails.  

**Time stamp and write stamp.** The time stamp is a random but unique number generated by the storage processor. Every time a major stripe update occurs (such as during a write command that causes all sectors in the stripe to be updated), a new time stamp is written into the validation area of all disks in the group.  

The write stamp is a series of bits, one for each disk in the disk group. Every time a write occurs in the sector, the state of the write stamp is changed not only on the data sector, but also on its corresponding parity sector. Thus, if the write operation has completed, the write stamp bit on the data sector should match the write stamp bit on the parity sector. All write stamps are written to zero at every major stripe update—the same time a new time stamp is written.  

Together, the time stamp and the write stamp create a type of version number for data and parity. After completed operations, the version numbers across data sectors and parity sectors will match. Every time a Dell | EMC storage processor opens a new disk group (for example, on power up, after being replaced, or after taking the load of a failed partner), the storage processor checks the time stamp and write stamp for each stripe. If it finds a mismatch, it recalculates the parity.  

Time stamps and write stamps work well to guard against inconsistent parity caused by failed controllers. However, if a problem is found, parity must be recalculated, which implies that all drives must be operational. If a disk drive and a controller simultaneously fail, parity would be impossible to recalculate because the complete set of original data is no longer available. Dell | EMC products address this issue through parity shedding and a shed stamp.  

**Shed stamp.** A patented algorithm, parity shedding is applied immediately upon failure of a disk drive. It uses parity to reconstruct the data from the failed drive and then overwrites the parity with the reconstructed data. Overwriting parity may seem risky, but once a drive fails, the only reason to have parity is to reconstruct the failed data. Reconstructing in advance and writing over the parity will not adversely affect system availability and in fact can significantly increase performance while the array is in a degraded state. After parity is shed, a shed stamp is written in the validation area of the parity sector to differentiate between true parity and reconstructed data.  

**Protecting data after it is written**  
Data can become corrupted as it is read back to the host. Furthermore, latent disk errors can corrupt data while it is sitting on the disk doing nothing. Dell | EMC systems use checksums and data scrubbing to protect against these problems.  

**Checksums**  
The validation area of the 520 byte sector contains a checksum for the entire data sector. When data is read from the disk, the controller calculates a new checksum and compares it to the checksum read with the data sector. If the controller finds a problem, it rejects the data and reads from disk again. The checksum verifies that the data stored is the data retrieved.  

**Data scrubbing**  
Data scrubbing, a proactive data-protection feature typically found only on higher end storage systems, is available on all Dell | EMC products through a firmware feature called SNiFFER. The SNiFFER process continuously reads blocks of data and checks for read errors reported by the disk drives. These errors can be either disk-recovered errors (the disk drive encountered errors but managed to correct them) or medium errors (the disk drive was not able to recover the data). SNiFFER responds to a disk-recovered error by proactively relocating the data to another area on disk before the data becomes unreadable. If SNiFFER encounters a medium error, it will reconstruct the data using redundant information from the lost data’s RAID group. As an added bonus, the process of finding and correcting drive errors results in the logical consistency of all data and parity information within RAID groups on the storage system.  

SNiFFER works concurrently with host I/O requests and does not interfere with ongoing operations. It can run as a continuous, low-priority background task that cycles through the complete system every few days, or it can run as a once-through process at higher priority that completes in a matter of hours.  

**Dell | EMC: Delivering data integrity**  
To protect data, RAID and backups alone are not enough. Administrators must guard against a multitude of scenarios that can potentially corrupt data: cache vulnerabilities, system failures during write operations, erroneous reads, and latent disk errors. Dell | EMC storage systems provide comprehensive, effective, production-proven data protection in these scenarios.  

Michael H. Darden (michael_darden@dell.com) is a product manager for Dell | EMC storage systems at Dell. His responsibilities include developing strategic roadmaps for Dell™ enterprise storage systems and defining next-generation product features for Dell | EMC products. Before joining the Dell Storage Systems Group in 2000, he spent seven years in various engineering and technical marketing positions in the high-tech industry. Michael holds a B.S. in Electrical Engineering, an M.S. in Manufacturing Systems Engineering, and an M.B.A. from the University of Texas at Austin.