Disk Storage, Heal Thyself

No-maintenance RAID arrays will deliver more efficiency and higher performance

TODAY'S RAID ARRAYS don't look very different from those of the 1990s: Most of the world's data is on very similar systems, with banks of hot-swappable drives and annual service contract costs of 15% to 20% of the purchase price. But that may be about to change.

Recently, a couple of pioneering vendors have introduced disk arrays that will require no maintenance over their three- to five-year lifetimes, and come with warranties that include on-site repairs for the same time span.

If these self-healing devices can live up to their promise of no maintenance, this technology could become a standard feature of disk arrays—and one we'll wonder how we ever did without. For now, organizations with some tolerance for risk could save themselves a bundle by adopting this technology sooner rather than later.

RETHINKING RAID

Xiotech's Intelligent Storage Element (ISE), used in its Emprise 5000 and 7000 drive arrays, and Atrato's Sealed Array of Independent Drives (SAID), the key component of its Velocity 1000 array, combine revamped mechanical design, advanced RAID technology, built-in spares, drive scrubbing, and, most significantly, drive rehabilitation to achieve high performance and low maintenance costs, potentially saving customers thousands of dollars per year.

The most visible difference between these systems and a typical midrange disk array is that they don't have front-accessible hot-swappable drives. The typical arrangement of 12 to 16 3.5-inch drives in a 3U package reduces mean time to repair and allows failed drives to be hot-swapped but also limits airflow. Because all the drives are mounted facing the same way, this setup allows multiple drives' rotational vibrations to reinforce one another, causing data errors and premature failure.

Rather than building RAID sets from whole drives, both systems break data into chunks, then create a logical RAID drive by distributing the data, parity, and 10% to 15% spare space across all physical drives. So a 4+1 RAID-5 set will have one parity chunk for every four data chunks, but the data will be spread across all the drives in the system.

Spreading RAID sets across all those spindles isn't a huge breakthrough: Hewlett-Packard's EVA and Xiotech's Magnitude, among others, have been doing it for years. This kind of data distribution has several advantages, though. First is the performance boost from having large reads and writes dispersed across 20 or more drives with independent positioners active at the same time. By spreading the spare space around, these drives are all working to deliver data.

SPEEDIER REBUILDS

The real innovation is what these arrays do when a drive fails. When a typical RAID controller encounters any drive error bigger than a single bad sector that it can remap to another location on the same drive, it marks the entire drive as bad and stops using it. If a spare drive is available, the controller starts rebuilding the RAID set.

When a self-healing array sees a drive error, it starts the rebuild process but also sends the drive that generated the error to rehab. First, it cycles the power, just to see if there was a firmware glitch—CPUs and firmware on drives occasionally hang. Then, it starts to run diagnostics to determine exactly what's wrong. The array then works to rehabilitate the drive by low-level formatting. If it still finds a head that's bad, it can return the rest of the drive's space to service.

This level of rehab requires close cooperation between the array and drive manufacturers to ensure that the controller knows the logical block addressing or SCSI block to a particular head, if nothing else.

It also means that self-healing

The Evolution Of RAID

1988	1989	1991	1994	1998	2002	2008
"A Case For Redundant Arrays Of Inexpensive Disks (RAID)" is published	Compaq SystemPro, the first RAID controller for LAN servers, arrives	EMC introduces Symmetrix RAID for the Mainframe	ANSI Fibre Channel Standard opens the SAN era	Xiotech's Magnitude offering virtualizes RAID	RAID-6 (double parity) goes mainstream	Self-healing arrays from Xiotech and Atrato come to market

arrays won't trust a failed drive just because it powers up and says it's OK. They put a suspect drive through its paces with reads, writes, and long and short seeks, then bring it up to operating temperature and run some diagnostics before releasing it from rehab and using it to store data again.

Duty-cycle management is another way to prevent drive failures. Self-healing arrays will throttle back requests to a drive if the controller sees that the drive temperature is too high.

NO ADMIN REQUIRED

Atrato is pitching the Velocity 1000 to video-on-demand, IPTV, and video-surveillance markets, where storage arrays may be installed in branch security offices, cable system distribution sites, and other locations that don't have professional storage administrators to swap drives and may not even be easily accessible to a dispatched service technician.

The Velocity 1000 is based on the SAID module. Each SAID contains 160 2.5-inch laptop-style 360-GB drives, providing 47.5 TB of usable space, taking into account both the overhead of parity information for RAID-5 and the allocation of spare space. The drives are mounted in counter-rotating pairs arranged so air can flow over all sides.

At $2 per gigabyte, the Velocity 1000 is priced competitively with low-end serial ATA arrays but offers many times their performance. A Velocity 1000 system has one or more controllers, with 4-Gbps Fibre Channel access ports and SAID cabinets. The controller is connected to the SAIDs via multiple SAS connections. A basic one-controller/one-SAID system costs about $100,000, but video-on-demand providers won't want just one. Getting this kind of capacity and performance from a conventional array would cost several times as much.

The Velocity 1000 supports RAID-1, RAID-5, and triple mirror virtual volumes. A low duty cycle option will shut down duplicate or triplicate drives in shifts to reduce the workload of each drive by as much as 60%. The Velocity 1000 also performs background disk scrubbing and uses the diagnostic data it collects to duplicate the data from the drives most likely to fail into the spare space on other

THE LOWDOWN

» THE PROMISE New-technology disk arrays rethink RAID to eliminate drive swapping and other ongoing maintenance, obviating expensive maintenance plans through built-in spares and other self-healing features, including drive rehabilitation. These arrays distribute data and spare space across all their drives, boosting performance as well.

» THE PLAYERS Atrato, Xiotech, Hewlett-Packard. Both Atrato and Xiotech have put their money where their mouths are, providing three- to five-year warranties, including on-site repairs. We expect others to follow suit.

» THE PROSPECTS While we don't yet know for sure if these arrays can run nonstop for five years without preventative maintenance, there's every reason to believe they will. This means IT groups, especially in organizations with remote locations, could save themselves a bundle by adopting this technology sooner rather than later. These arrays also offer several times the performance for the buck compared with more conventional systems.

drives to reduce the amount of data that needs to be rewritten during a rebuild.

Xiotech acquired ISE, the technology in its Emprise 5000, from Seagate Technology. The ISE is a 3U module that mounts 20 3.5-inch or 40 2.5-inch drives on two aluminum frames, called DataPacs, in back-to-back pairs, so the drives' rotational vibrations cancel out and air can flow over the entire system. Each ISE also includes dual active/active controllers with 1 GB of battery-backed-up cache and 4-Gbps Fibre Channel interfaces.

The basic building block of an Emprise storage infrastructure is the 1.1-, 2.4-, or 8-TB ISE. As you add capacity, you're also adding more RAID controllers and more cache memory. On a more conventional array, in contrast, adding more drive shelves for new applications would mean sharing I/O channels and cache.

Xiotech's Emprise 7000 adds a controller running the same Xiotech ICON Manager GUI as its Magnitude 3D array line to two or more Emprise 5000 ISEs and can share storage with the Magnitude 3D in virtual volumes while adding replication to the feature set. An Emprise 7000 system can manage as many as 64 ISEs for as much as a petabyte of storage.

Both systems post some sizzling performance: Atrato says a single SAID Velocity 1000 can support 3,600 simultaneous video streams or 11,000 I/Os per second. The Emprise 5000 is the current record holder in price/performance on the SPC-1 and SPC-2 benchmarks: A single ISE churned out 5,892 I/Os per second for $20,800, for a net price/performance figure of $3.52 per I/Os per second, less than 25% of the cost of conventional arrays.

Critics have pointed out that the failure avoidance and self-healing techniques used in these arrays can result in performance variations as drives are throttled back and/or disabled. This results in fewer positioners being active to serve data. However, given the high performance these systems provide, we don't see small variations in performance as an issue for the vast majority of users. And with multiterabyte drives on the horizon, IT departments need to look at RAID with fresh eyes.

—HOWARD MARKS (hmarks@nwc.com)