

**As hard-drive capacities continue to outpace their throughput, the time has come for a new level of RAID.**

BY ADAM LEVENTHAL

# Triple-Parity RAID and Beyond

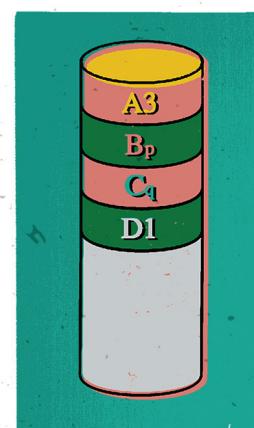
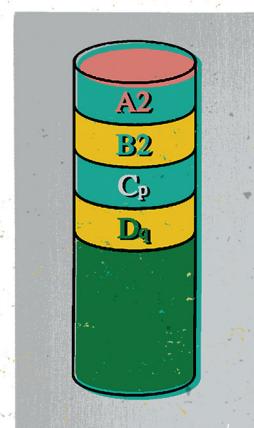
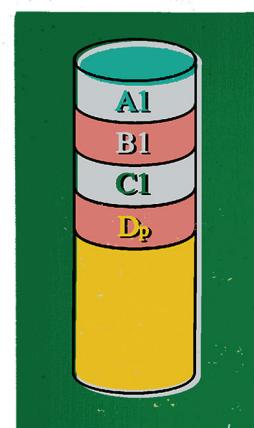
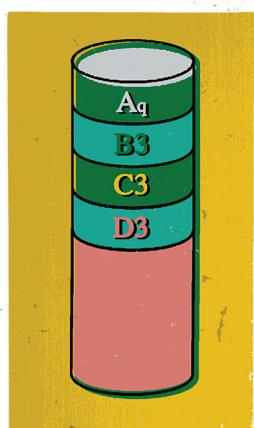
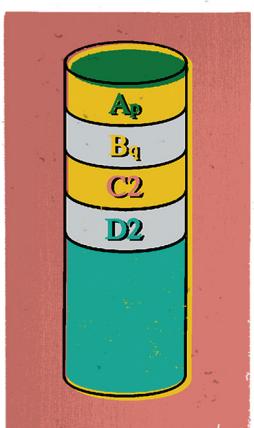
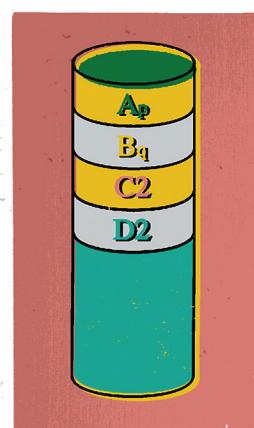
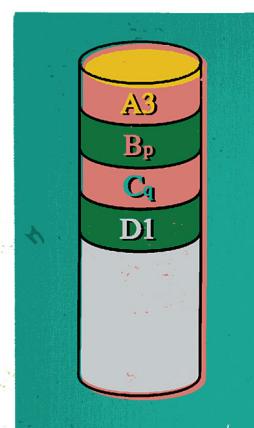
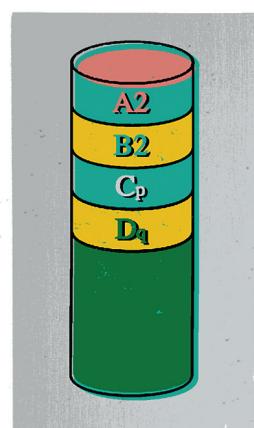
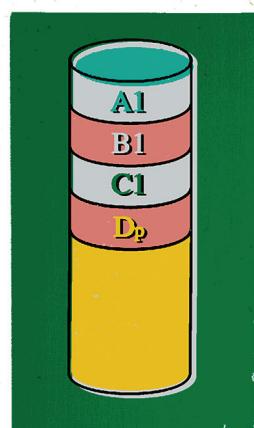
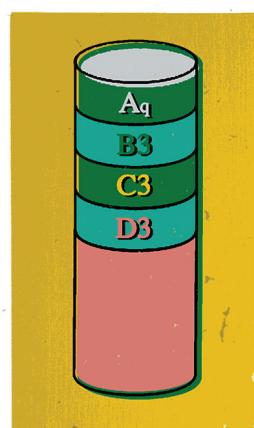
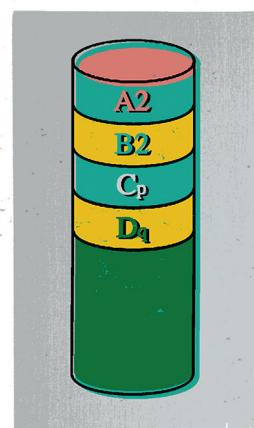
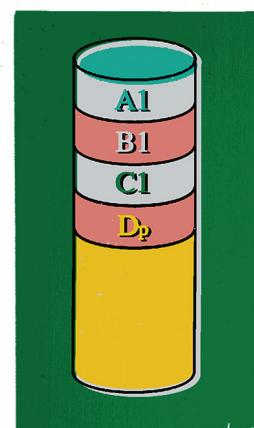
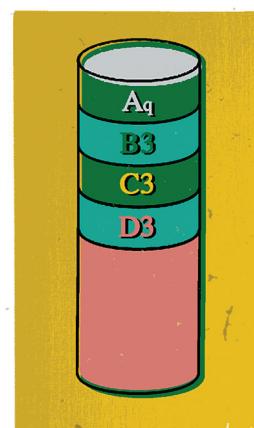
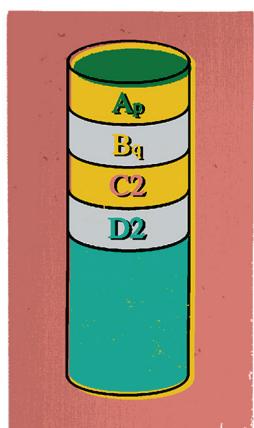
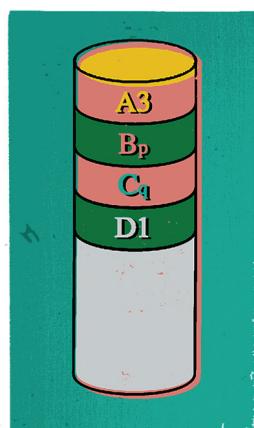
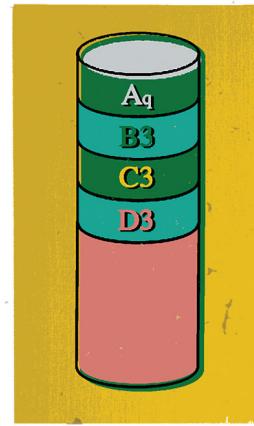
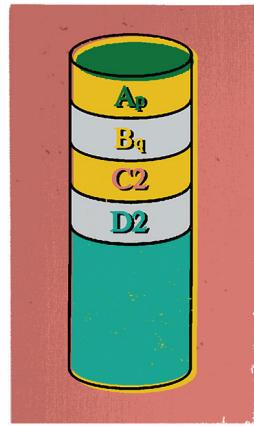
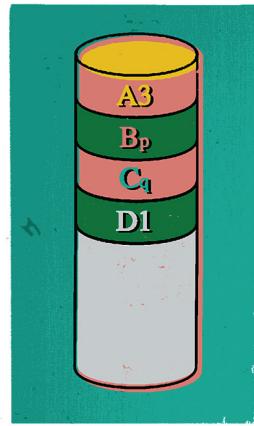
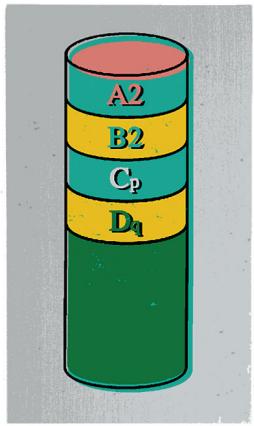
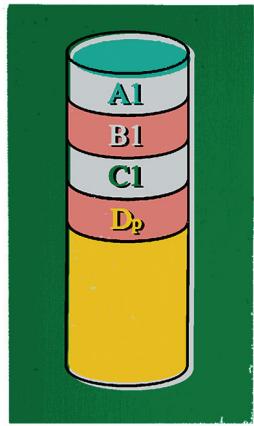
HOW MUCH LONGER will current RAID techniques persevere? The RAID levels were codified in the late 1980s; double-parity RAID, known as RAID-6, is the current standard for high-availability, space-efficient storage. The incredible growth of hard-drive capacities, however, could impose serious limitations on the reliability even of RAID-6 systems. Recent trends in hard drives show that triple-parity RAID must soon become pervasive. In 2005, *Scientific American* reported on Kryder's Law,<sup>11</sup> which predicts that hard-drive density will double annually. While the rate of doubling has not quite maintained that pace, it has been close.

Problems for RAID, hard-disk throughput has failed to match that exponential rate of growth. Today repairing a high-density disk drive in a RAID group can easily take more than four hours, and the problem is getting significantly more pronounced

as hard-drive capacities continue to outpace their throughput. As the time required for rebuilding a disk increases, so does the likelihood of data loss. The ability of hard-drive vendors to maintain reliability while pushing to higher capacities has already been called into question in these pages.<sup>5</sup> Perhaps even more ominously, in a few years, reconstruction will take so long as to effectively strip away a level of redundancy. What follows is an examination of RAID, the rate of capacity growth in the hard-drive industry, and the need for triple-parity RAID as a response to diminishing reliability.

The first systems that would come to be known as RAID were developed in the mid-1980s. David Patterson, Garth Gibson, and Randy Katz of the University of California, Berkeley, classified those systems into five distinct categories under the umbrella of RAID (redundant arrays of inexpensive disks).<sup>9</sup> In their 1988 paper, RAID played David to the Goliath of SLED (single large expensive disks). The two represented fundamentally different strategies for how to approach the future of computer storage. While SLED offered specialized performance and reliability—at a price—RAID sought to assemble reliable, high-performing storage from cheap parts, reflecting a broader trend in the computing industry. The economics of commodity components are unstoppable.

Patterson et al. were seemingly prescient in their conclusion: "With advantages in cost-performance, reliability, power consumption, and modular growth, we expect RAIDs to replace SLEDs in future I/O systems."<sup>9</sup> However, their characterization of RAID as "a disk array made from personal computer disks" was a bit too specific and a bit too hopeful. While RAID is certainly used with those inexpensive, high-volume disks, RAID in its de facto incarnation today combines its algorithmic reliability and performance improvements with disks that are themselves often designed for performance and reliability, and therefore



remain expensive. This evolution is reflected in the subtle but important mutation of the meaning of the I in RAID from *inexpensive* to *independent* that took place in the mid-1990s (indeed, it was those same SLED manufacturers that instigated this shift to apply the new research to their existing products).

In 1993, Gibson, Katz, and Patterson, along with Peter Chen, Edward Lee, completed a taxonomy of RAID levels that remain unamended to date.<sup>3</sup>

Of the seven RAID levels described, only four are commonly used:

► **RAID-0.** Data is striped across devices for maximal write performance. It is an outlier among the other RAID levels as it provides no actual data protection.

► **RAID-1.** Disks are organized into mirrored pairs and data is duplicated on both halves of the mirror. This is typically the highest-performing RAID level, but at the expense of lower usable capacity. (The term *RAID-10* or

*RAID-1+0* is used to refer to a RAID configuration in which mirrored pairs are striped, and *RAID-01* or *RAID-0+1* refer to striped configurations that are then mirrored. The terms are of decreasing relevance since striping over RAID groups is now more or less assumed.)

► **RAID-5.** A group of  $N+1$  disks is maintained such that the loss of any one disk would not result in data loss. This is achieved by writing a parity block,  $P$ , for each logical row of  $N$  disk blocks. The location of this parity is distributed, rotating between disks so that all disks contribute equally to the delivered system performance. Typically  $P$  is computed simply as the bitwise XOR of the other blocks in the row.

► **RAID-6.** This is like RAID-5, but employs two parity blocks,  $P$  and  $Q$ , for each logical row of  $N+2$  disk blocks. There are several RAID-6 implementations such as IBM's EVENODD,<sup>2</sup> NetApp's Row-Diagonal Parity,<sup>4</sup> or more generic Reed-Solomon encodings.<sup>10</sup> (Chen et al. refer to RAID-6 as  $P+Q$  redundancy, which some have taken to imply  $P$  data disks with an arbitrary number of parity disks,  $Q$ . In fact, RAID-6 refers exclusively to double-parity RAID;  $P$  and  $Q$  are the two parity blocks.) For completeness, it's worth noting the other less prevalent RAID levels:

► **RAID-2.** Data is protected by memory-style ECC (error correcting codes). The number of parity disks required is proportional to the log of the number of data disks; this makes RAID-2 relatively inflexible and less efficient than RAID-5 or RAID-6 while also delivering lower performance and reliability.

► **RAID-3.** As with RAID-5, protection is provided against the failure of any disk in a group of  $N+1$ , but blocks are carved up and spread across the disks—bitwise parity as opposed to the block parity of RAID-5. Further, parity resides on a single disk rather than being distributed between all disks. RAID-3 systems are significantly less efficient than with RAID-5 for small read requests; to read a block all disks must be accessed; thus the capacity for read operations is more readily exhausted.

► **RAID-4.** This is merely RAID-5, but with a dedicated parity disk rather

Figure 1. Comparison of RAID-5 and RAID-6 reliability.<sup>1</sup>

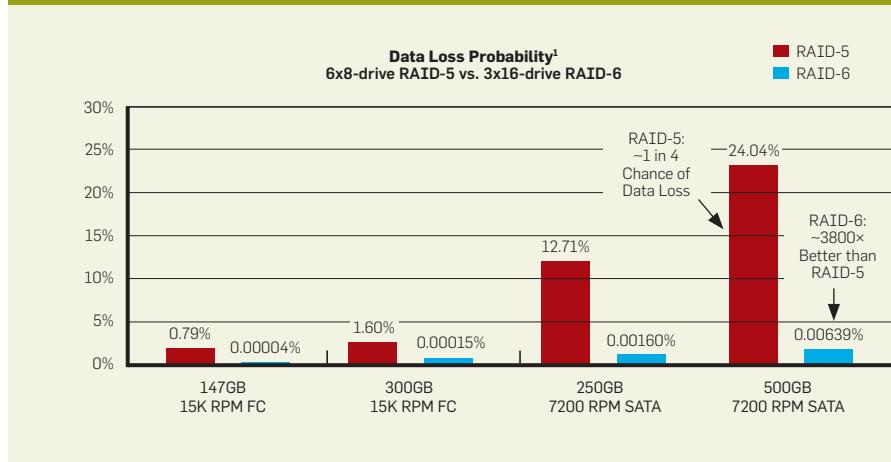


Figure 2. Historical Capacity/Throughput of 7200 RPM SATA HDDs.

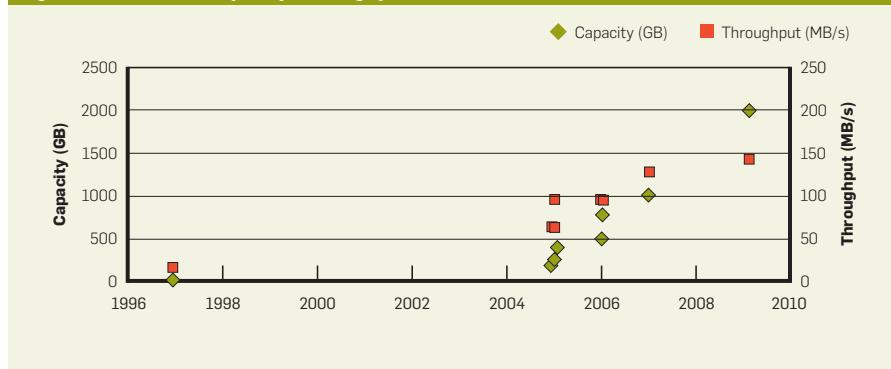


Figure 3. Historical Capacity/Throughput of 10K RPM FC HDDs.



than having parity distributed among all disks. Since fewer disks participate in reads (the dedicated parity disk is not read except in the case of a failure), RAID-4 is strictly less efficient than RAID-5.

RAID-6, double-parity RAID, was not described in Patterson, Gibson, and Katz's original 1988 paper<sup>9</sup> but was added in 1993 in response to the observation that as disk arrays grow, so too do the chances of a double failure. Further, in the event of a failure under any redundancy scheme, data on all drives within that redundancy group must be successfully read in order for the data that had been on the failed drive to be reconstructed. A read failure during a rebuild would result in data loss. As Chen et al. state:

"The primary ramification of an uncorrectable bit error is felt when a disk fails and the contents of the failed disk must be reconstructed by reading data from the nonfailed disks. For example, the reconstruction of a failed disk in a 100GB disk array requires the successful reading of approximately 200 million sectors of information. A bit error rate of one in  $10^{14}$  bits implies that one 512-byte sector in 24 billion sectors cannot be correctly read. Thus, if we assume the probability of reading sectors is independent of each other, the probability of reading all 200 million sectors successfully is approximately

$$(1 - 1/(2.4 \times 10^{10}))^{(2.0 \times 10^8)} = 99.2\%.$$

This means that on average, 0.8% of disk failures would result in data loss due to an uncorrectable bit error."<sup>3</sup>

Since that observation, bit error rates have improved by about two orders of magnitude while disk capacity has increased by slightly more than two orders of magnitude, doubling about every two years and nearly following Kryder's law. Today, a RAID group with 10TB (nearly 20 billion sectors) is commonplace, and typical bit error rate stands at one in  $10^{16}$  bits:

$$(1 - 1/(2.4 \times 10^{12}))^{(2.0 \times 10^{10})} = 99.2\%.$$

While bit error rates have nearly kept pace with the growth in disk capacity, throughput has not been given its due consideration when determining RAID reliability.

As motivation for its RAID-6 solution, NetApp published a small comparison of RAID-5 and -6 with equal capacities (7+1 for RAID-5 and 14+2 for RAID-6) and hard drives of varying quality and capacity.<sup>1</sup> Note that despite having an additional parity disk, RAID-6 need not reduce the total capacity of the system.<sup>7</sup> Typically the RAID stripe width—the number of disks within a single RAID group—for RAID-6 is double that of a RAID-5 equivalent; thus, the number of data disks remains the same. The NetApp comparison is not specific about the bit error rates of the devices tested, the reliability of the drives themselves, or the length of the period over which the probability of data loss is calculated; therefore, we did not attempt to reproduce these specific results. The important point to observe in Figure 1 is the stark measured difference in the probability of data loss between RAID-5 and RAID-6.

When examining the reliability of a RAID solution, typical considerations

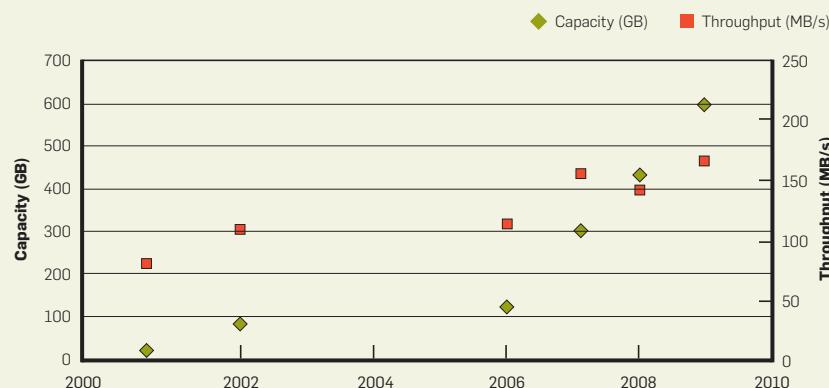
range from the reliability of the component drives to the time for a human administrator to replace failed drives. The throughput of drives has not been a central focus despite being critical for RAID reconstruction, because throughput has been more than adequate. While factors such as the bit error rate have kept pace with capacity, throughput has lagged behind, forcing a new examination of RAID reliability.

### Capacity vs. Throughput

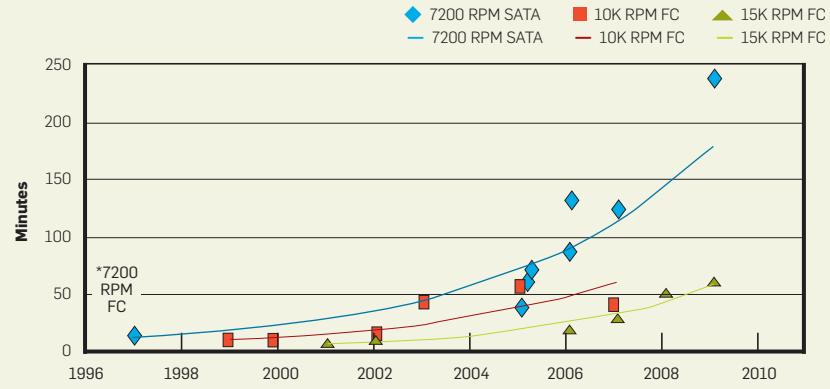
Capacity has increased steadily and significantly, and the bit error rate has improved at nearly the same pace. Hard-drive throughput, however, has lagged behind significantly. Using vendor-supplied hard-drive data sheets, we've been able to examine the relationship between hard-drive capacity and throughput for the past 10 years. Figures 2–4 show samples for various hard-drive protocols and rotational speeds.

This data presents a powerful con-

**Figure 4. Historical Capacity/Throughput of 15K RPM FC HDDs.**



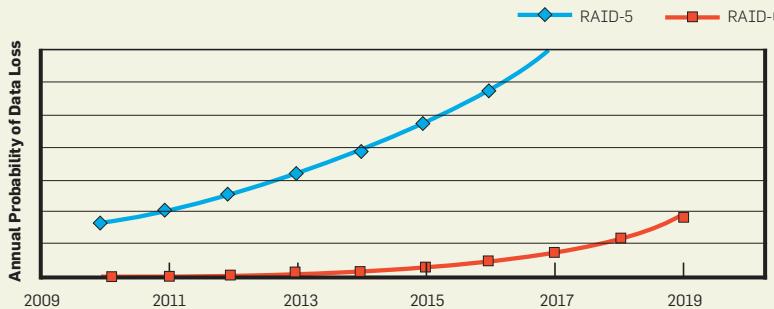
**Figure 5. Minimum time required to populate HDDs through the years.**



clusion about the relative rates of capacity and throughput growth for hard drives of all types—there's obviously no exponential law governing hard-drive throughput. By dividing capacity by throughput, we can compute the amount of time required to fully scan or populate a drive. It is this duration that dictates how long a RAID group is operating without full parity protection. Figure 5 shows the duration such an operation would take for the various drive types over the years.

When RAID systems were developed in the 1980s and 1990s, reconstruction times were measured in minutes. The trend for the past 10 years is quite clear regardless of the drive speed or its market segment: the time to perform a RAID reconstruction is increasing exponentially as capacity far outstrips throughput. At the extreme, rebuilding a fully populated 2TB 7200-RPM SATA disk—today's capacity champ—after a failure would take four hours operating at the theoretical optimal throughput. It is rare to achieve those data rates in practice; in the context of a heavily used system the full bandwidth can't be dedicated exclusively to RAID repair without adversely affecting performance. If

Figure 6. Projected relative reliability of single- and double-parity RAID.



one assumes that only 10%–50% of the total system throughput is available for reconstruction, the minutes-long RAID rebuild times of the 1990s balloon to multiple hours or days in practice. RAID systems operate in this degraded state for far longer than they once did and as a consequence are at higher risk for data loss.

Latent data on hard drives can acquire defects over time—a process blithely referred to as bit rot. To mitigate this, RAID systems typically perform background scrubbing in which data is read, verified, and corrected as needed to eradicate correctable failures before they become uncorrectable.<sup>5</sup> The phenomenon of scrub-

bing data necessarily impacts system performance, but the time required for a full scrub is a significant component of the reliability of the total system. A natural tension results between how priorities are assigned to scrubbing versus other system activity. As throughput is dwarfed by capacity, either the percentage of resources dedicated to scrubbing must increase, or the time for a complete scrub must increase. With the trends noted previously, storage pools will easily take weeks or months for a full scrub regardless of how high a priority scrubbing is given, further reducing the reliability of the total system as it becomes more likely that RAID reconstructions will encounter latent data corruption.

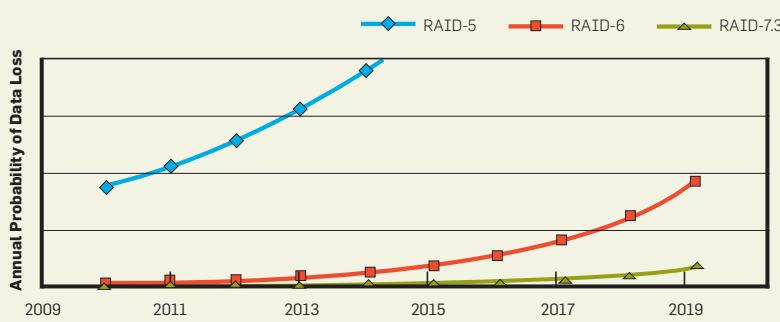
Given the growing disparity between the capacity growth of hard drives and improvements to their performance, the long-term prospects of RAID-6 must be reconsidered. The time to repair a failed drive is increasing, and at the same time the lengthening duration of a scrub means that errors are more likely to be encountered during the repair. In Figure 6, we have chosen reasonable values for the bit error rate and annual failure rate, and a relatively modest rate of capacity growth (doubling every three years). This is meant to approximate the behavior of low-cost, high-density, 7200-RPM drives. Different values would change the precise position of the curves, but not their relative shapes.

RAID-5 reached a threshold 15 years ago at which it no longer provided adequate protection. The answer then was RAID-6. Today RAID-6 is quickly approaching that same threshold. In about 10 years, RAID-6 will provide only the level of protection that we get from

## A Classification for Triple-Parity RAID

None of the existing RAID classifications apply for triple-parity RAID. One option would be to extend the existing RAID-6 definition, but this could be confusing, as many RAID-6 systems exist today. The next obvious choice is RAID-7, but rather than applying the designation merely to RAID with triple-parity protection, RAID-7 should be a catch-all for any RAID technique that can be extended to an arbitrary number of parity disks. Specific techniques or deployments that fix the number of parity disks at N should use the RAID-7.N nomenclature with RAID-7.3 referring to triple-parity RAID, and RAID-5 and RAID-6 effectively as the degenerate forms RAID-7.1 and RAID-7.2, respectively.

Figure 7. Projected relative reliability of single-, double-, and triple-parity RAID.



RAID-5 today. It is again time to create a new RAID level to accommodate the realities of disk reliability, capacity, and throughput merely to maintain that same level of data protection.

### Triple-Parity RAID

With RAID-6 increasingly unable to meet reliability requirements, there is an impending but not yet urgent need for triple-parity RAID. The addition of another level of parity mitigates increasing RAID rebuild times and occurrences of latent data errors. As shown in Figure 7, triple-parity RAID will address the shortcomings of RAID-6 for years (see the accompanying sidebar “A Classification for Triple-Parity RAID”). The reliability is largely independent of the specific implementation of triple-parity RAID; a general Reed-Solomon method suffices for our analysis.

A recurring theme in computer science is that algorithms can be specialized for small fixed values, but are then generalized to scale to an arbitrary value. A common belief in the computer industry had been that double-parity RAID was effectively that generalization, that it provided all the data reliability that would ever be needed. RAID-6 is inadequate, leading to the need for triple-parity RAID, but that, too, if current trends persist, will become insufficient. Not only is there a need for triple-parity RAID, but there's also a need for efficient algorithms that truly address the general case of RAID with an arbitrary number of parity devices.

Beyond RAID-5 and -6, what are the implications for RAID-1, simple two-way mirroring? RAID-1 can be viewed as a degenerate form of RAID-5, so even if bit error rates improve at the same rate as hard-drive capacities, the time to repair for RAID-1 could become debilitating. How secure would an administrator be running without redundancy for a week-long scrub? For the same reasons that make triple-parity RAID necessary where RAID-6 had sufficed, three-way mirroring will displace two-way mirroring for applications that require both high performance and strong data reliability. Indeed, four-way mirroring may not be far off, since even three-way mirroring is effectively a degenerate, but more

reliable, form of RAID-6, and will be susceptible to the same failings.

### Implications for RAID

While triple-parity RAID will be necessary, the steady penetration of flash solid-state storage could have a significant effect on the fate of disk drives. At one extreme, some have predicted the relegation of disk to a tape-like backup role as flash becomes cheap and reliable enough to act as a replacement for disk.<sup>6</sup> In that scenario, RAID is still necessary as even solid-state devices suffer catastrophic and partial failures, but the specific capacities, error rates, and throughputs for such devices could mean that triple-parity RAID is not required. Unfortunately, too little is known about the properties of devices that might flourish, and that scenario is too far in the future to obviate the need for triple-parity RAID.

At another extreme, the integration of flash into the storage hierarchy<sup>8</sup> could address high-performance needs though solid-state caching and buffering, thus decoupling system performance from that of the component hard drives. This could hasten current trends as hard-drive manufacturers would be able to increase capacity even more quickly, unhindered by performance requirements, while likely slowing the rate of throughput increases. Further, divorced from performance, RAID stripes could grow very wide to optimize for absolute capacity; this would reduce the reliability further with the same amount of parity protecting more data. In this scenario, the need for triple-parity RAID would be made all the more urgent by accelerating current trends.

If Kryder's Law continues to hold, the burden of correctness will increasingly shift from the hard-drive manufacturers to the RAID systems that integrate them. Today, RAID reconstruction times factor more into reliability calculations than ever before, and their contribution will increasingly dominate. Triple-parity RAID will soon be critical to provide sufficient reliability even in the face of exponential growth.

### Acknowledgments

Many thanks to Dominic Kay for gath-

ering the historical hard-drive data, and to Matt Ahrens, Daniel Leventhal, and Beverly Hodgson for their helpful reviews. C

### Related articles on queue.acm.org

#### Flash Storage Today

Adam Leventhal

<http://queue.acm.org/detail.cfm?id=1413262>

#### Hard Disk Drives: The Good, the Bad and the Ugly

Jon Elerath

<http://queue.acm.org/detail.cfm?id=1317403>

#### You Don't Know Jack about Disks

Dave Anderson

<http://queue.acm.org/detail.cfm?id=864058>

### References

1. Berriman, E., Feresten, P., and Kung, S. NetApp RAID-DP: Dual-parity Raid-6 protection without compromise; <http://www.mochadata.com/download/NetApp-raid-dp.pdf> (2006).
2. Blaum, M., Brady, J., Bruck, J., and Menon, J. EVENODD: An optimal scheme for tolerating double disk failures in RAID architectures. In *Proceedings of the International Symposium on Computer Architecture* (1994), 245–254; <http://portal.acm.org/citation.cfm?id=191995.192033>.
3. Chen, P., Lee, E., Patterson, D., Gibson, G., and Katz, R. RAID: High-performance, reliable secondary storage. Technical Report CSD 93-778 (1993); <http://portal.acm.org/citation.cfm?id=893811>.
4. Corbett, P., English, B., Goel, A., Grcanac, T., Kleiman, S., Leong, J., and Sankar, S. Row-diagonal parity for double disk failure correction. In *Proceedings of the 3rd Usenix Conference on File and Storage Technologies* (2004), 1–14; <http://portal.acm.org/citation.cfm?id=1096673.1096677>.
5. Elerath, J. Hard-disk drives: The good, the bad, and the ugly. *Commun. ACM* 52, 6 (June 2009), 38–45; <http://portal.acm.org/citation.cfm?id=1516046.1516059>.
6. Gray, J. and Fitzgerald, B. Flash Disk Opportunity for Server-Applications. Microsoft Research; <http://research.microsoft.com/en-us/um/people/gray/papers/FlashDiskPublic.doc> (2007).
7. Hitz, D. 2006. Why “Double Protecting RAID” (RAID-DP) doesn't waste extra disk space; [http://blogs.netapp.com/dave/2006/05/why\\_double\\_prot.html](http://blogs.netapp.com/dave/2006/05/why_double_prot.html) (2006).
8. Leventhal, A. Flash storage memory. *Commun. ACM* 51, 7 (July 2008), 47–51; <http://portal.acm.org/citation.cfm?id=1364782>.
9. Patterson, D., Gibson, G., and Katz, R. A case for redundant arrays of inexpensive disks (RAID). In *Proceedings of ACM SIGMOD International Conference on Management of Data* (1988), 109–116; <http://portal.acm.org/citation.cfm?id=50214>.
10. Plank, J. A tutorial on Reed-Solomon coding for fault-tolerance in RAID-like systems. Technical Report UT-CS-96-332; <http://portal.acm.org/citation.cfm?id=898928> (1996).
11. Walter, C. Kryder's Law. *Scientific American* (Aug. 2005); <http://www.scientificamerican.com/article.cfm?id=kryders-law>.

**Adam Leventhal** is a senior staff engineer and flash architect for Sun's Fishworks advanced product development team responsible for the Sun Storage 7000 series. He is one of the three authors of DTrace, for which he and his colleagues were named one of *InfoWorld's* Innovators of 2005 and won top honors from the 2006 *Wall Street Journal's* Innovation Awards.