Technology Update

White Paper

“High Speed RAID 6”

Powered by

Custom ASIC Parity Chips
“High Speed RAID 6”

 Powered by Custom ASIC Parity Chips

Why High Speed RAID 6?
Winchester Systems has developed “High Speed RAID 6” to provide superior data protection for both primary and secondary enterprise data storage in response to key storage industry trends and customer requirements.

Simply, “High Speed RAID 6” is a disk array that uses today’s “enterprise class” Fibre Channel, SAS or SATA disk drives, protects against two drive failures and ensures the successful completion of rebuilds by providing RAID 5 protection after a single disk drive failure – and does so at high speed. Thus, “High Speed RAID 6” offers vastly superior protection from permanent data loss – increasing the MTDL (mean time to data loss) by 2 to 4 orders of magnitude. While simple in concept, “High Speed RAID 6” is not yet available from most major commercial storage vendors.

Storage Industry Trends
Two trends are driving the data storage industry inexorably towards RAID 6. One is the well-known and remarkable increase in drive capacities and the second is that the small improvements to the unrecoverable bit error rate of enterprise class disk drives have fallen seriously behind increases in capacity. This leaves users exposed to data loss without a more reliable technology like RAID 6.

Capacity
The data storage industry has followed Moore’s law for semiconductor memory. That is, capacity doubles every 12 to 18 months. This has held true for the past three decades or more in hard disk storage. Disk drive capacities have gone from 5 MB to 2 TB in just 25 years - a stunning 400,000 times increase.

The industry response was to design disk arrays with redundancy, notably RAID 5, which uses parity information to rebuild an array if one drive fails. RAID 5 was designed when disk drives were under 1 GB each. Today, disk drives are often 600 GB to 2 TB and thus design goals and assumptions made a decade ago are in need of review.

Bit Error Rates
The unrecoverable bit error rates of enterprise class disk drives have changed only once in the past dozen years by one order of magnitude. Drives improved from 10**15 to 10**16 on primary storage Fibre Channel and SAS drives and from 10**14 to 10**15 for SATA secondary storage. However, drive capacities have increased by over three orders of magnitude. Given the dramatically higher recording densities of today’s disk drives, improving the bit error rate by even this much should be viewed as a major accomplishment by the disk drive manufacturers. Nonetheless, the disk drive capacities are up 2,000 times from 1 GB to 2 TB in just the past 12 years. With 2,000 times more capacity but only 10 times more reliability, means that 200 times the error rate (2,000/10) is now expected compared to when RAID 5 was introduced. Any time a design is stressed 200 times the error rate of the original purpose – it is time to review the assumptions and determine if they are still valid.

Permanent Data Loss
Most every IT manager knows that RAID 5 protects against a single drive failure. They also know that the array can be rebuilt from the remaining drives. To perform a successful rebuild, every sector on all
the other drives must be readable – or permanent data loss occurs. If even one sector is bad, data is lost, perhaps in a critical sector, like a directory or key file. Many disk arrays just fail the rebuild and stop, losing all data.

Challenged Assumptions
When drives were small, it was reasonable to assume that a rebuild would be successful. What many IT managers do not realize is the effect of the 2,000-fold increase in storage capacity per drive. Today, with 10 and 20 TB disk arrays, the probability of a complete and successful rebuild diminishes proportionally with the size of the array – to the point of that there are now potentially unacceptable risks of data loss during RAID 5 rebuilds. Remember, a successful rebuild requires that every single sector of every remaining disk read correctly. A single bad sector ruins the rebuild and causes data loss. Figure 1 shows the probability of data loss using various classes of disk drives. Drives with 1 bit in $10^{16}$ unrecoverable error rates are considered to be primary storage and those with 1 bit in $10^{15}$ error rates are considered secondary storage comprised of low cost SATA disks.

Probability of Errors
As the data in Figure 1 clearly shows, the probability of a rebuild failure in a 20 TB array is a measurable number of 1.6%. This probability was a tiny fraction of one percent when drives were smaller. Using SATA technology, the error rate is 10 times higher at a pretty scary 16%.

If a data center has 200 TB of total primary storage and rebuilds each array once in its lifetime, there is also a 16% chance of permanent data loss. If a total of 1,250 TB of rebuilds occurs over a three year period in a data center, it is a virtual guarantee that permanent data loss will occur – even with the best drives and full RAID 5 protection. With secondary SATA storage there will be 10 permanent data losses in the same time.

Notice, that these calculations are independent of the vendor, RAID controller and parity mechanism. It is just a fact determined by the unrecoverable error rate of the highest reliability enterprise class drives on the market used by all the storage manufacturers.

Figure 1 – Probability of Permanent Data Loss

<table>
<thead>
<tr>
<th>Drive Type</th>
<th>Bit Error Rate</th>
<th>Byte Error Rate</th>
<th>Prob. Data Loss (20 TB)</th>
<th>Prob. Data Loss (200 TB)</th>
<th>Number of Data Losses (1,250 TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary</td>
<td>1 in $10^{15}$</td>
<td>125 TB</td>
<td>16%</td>
<td>160%</td>
<td>10</td>
</tr>
<tr>
<td>Primary</td>
<td>1 in $10^{16}$</td>
<td>1,250 TB</td>
<td>1.60%</td>
<td>16%</td>
<td>1</td>
</tr>
</tbody>
</table>

RAID 5 Limitations
Thus, everyone needs to understand the limitations of the drives and the RAID 5 protection algorithms that have served the industry so well to date. Note also that the calculations in Figure 1 take the manufacturer’s specifications at face value. Drive manufacturers determine these error rates by running tests of hundreds or thousands of drives for short periods under ideal conditions to simulate field experience and then extrapolate the results from short duration measurements. I/O intensive applications wear drive mechanisms and increase error rates. Older drives can surely expect higher, perhaps significantly higher, error rates than manufacturer’s published rates.
Finally, RAID 5 arrays have always been subject to human error. If someone removes the wrong drive in a RAID 5 array after a failure – all the data will be permanently lost.

Something needs to be done – especially as new drives will have higher capacity and place data at ever-higher risks.

**Interim solutions**
There are, of course, some workarounds that can be done, but these are merely delaying the tactics that do not make good use of resources. RAID 5 arrays can minimize the chance of failure on a particular rebuild by configuring smaller arrays.

It is true that the probability of a rebuild failure is smaller for a single array it is not smaller for the storage installation as a whole. Further, smaller arrays have many disadvantages. They require more parity drives and waste storage, limit volume sizes and thus risk “disk full” messages and numerous volume expansions and require more management intervention. Smaller arrays deliver reduced performance since a smaller number of disks reduces the stripe size and limits the parallel reading and writing that is provided by accessing many disks at once.

The industry trends discussed are stressing the design limits of RAID 5 and putting enterprise data at risk. It is time to address the unseen part of the iceberg in reliable data storage. A better solution is needed.

**RAID 6 Solution**
Arrays with RAID 6 protection require two drives for distributed parity. This is one more drive than RAID 5 but many fewer than RAID 10. (Refer to Appendix A for a brief review of RAID modes.)

Figure 2 depicts an example a RAID 6 array with seven disk drives and shows how the two independent sets of parity information are distributed across all seven disks. If one disk drive fails, the array survives and is still RAID 5 protected! That is the distinct advantage of RAID 6. If a second drive fails, the data is still available but is no longer RAID protected until a rebuild is completed.

**Figure 2 – RAID 6 Dual Distributed Parity**

While RAID 6 protection requires two parity disks rather than the one disk required by RAID 5, most RAID 5 users deploy a “hot-spare” drive, which is inactive until needed, so two drives are also used. The hot spare in the RAID 5 array can be thought of as an “inactive” parity drive while the second extra drive in the RAID 6 array can be thought of as an “active” parity drive. Thus, two drives are utilized to protect the data in each case but RAID 6 provides vastly superior data protection with no added disk drives.

The hot spare is still an option but not nearly as necessary in RAID 6 since the array is still RAID 5 protected after the first drive failure. The dreaded RAID 5 “window of vulnerability” is closed in RAID 6.
after a single drive failure. RAID 6 thus provides MTDL that is 2 to 4 orders of magnitude better than comparable RAID 5 arrays as will be demonstrated in a later section.

**Mean Time to Data Loss**

RAID 6 is a dual distributed parity mechanism that permits two disk drives to fail in an array and still be able to recover and rebuild data from the remaining disk drives. This is a huge advantage over RAID 5 – especially in the MTDL. While the probability of a second drive failure during a rebuild is still low, the probability of a bad sector is still relatively high and increases with array size.

_After a drive failure in RAID 6, the array is still running with RAID 5 protection._ This enables the surviving RAID 5 array to automatically correct the expected bad block errors during rebuild. Figure 3 shows the MTDL results prepared by Intel Corp. that demonstrates the overwhelming advantage of data protection offered by RAID 6 versus RAID 5.

**RAID 6 provides 500 to 30,000 times the MTDL as a comparable RAID 5 array.** Thus RAID 6 provides MTDL measured in thousands of months or hundreds of years instead of as little as year or two.

**Figure 3 – Mean Time to Data Loss**

<table>
<thead>
<tr>
<th>Array Capacity</th>
<th>RAID 5 MTDL (months)</th>
<th>RAID 6 MTDL (months)</th>
<th>Reliability Increase Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 TB</td>
<td>1000</td>
<td>30,000,000</td>
<td>30,000x</td>
</tr>
<tr>
<td>2 TB</td>
<td>500</td>
<td>5,000,000</td>
<td>10,000x</td>
</tr>
<tr>
<td>5 TB</td>
<td>100</td>
<td>200,000</td>
<td>2,000x</td>
</tr>
<tr>
<td>10 TB</td>
<td>70</td>
<td>70,000</td>
<td>1,000x</td>
</tr>
<tr>
<td>20 TB</td>
<td>20</td>
<td>10,000</td>
<td>500x</td>
</tr>
</tbody>
</table>

Source: Intel Corp. SATA 10^15 disk drives

**Unbeatable Combination**

When hardware powered RAID 6 is combined with today’s Fibre Channel, SAS or SATA disk drives the combination is an exceptionally reliable data storage device. When the RAID 6 parity calculations are performed in dual custom ASICs in the RAID controller, then high performance can also be expected. The dual high speed ASICs permit a properly designed RAID 6 system to be faster than typical commercial RAID 5 arrays.

**“High Speed RAID 6” Performance Breakthrough**

Many commercial RAID vendors have not embraced RAID 6 simply because implementation on their current family of RAID controllers would result in a product that would be too slow to be accepted by the bulk of the IT community. Existing designs perform RAID parity calculations in software or use just a single ASIC. Thus RAID 6 is often associated with a large performance impact known as a write penalty – often a 65-80% performance reduction – a good reason to avoid using RAID 6.

However, hardware powered by custom high speed ASICs for RAID 5 and RAID 6 parity calculations can deliver performance in RAID 6 comparable to RAID 5 performance as shown in Figure 4 below for FlashDisk by Winchester Systems. Thus, FlashDisk allows using the clearly valuable dual parity protection of RAID 6 while delivering the performance of RAID 5 – a true performance breakthrough.
Figure 4 – FlashDisk RAID 6 vs. RAID 5 Performance

<table>
<thead>
<tr>
<th>SAS/SATA Array with Eight 8 Gb Fibre Channel Host Ports</th>
<th>Single Controller</th>
<th>Dual Controllers</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAID 5 IOPS (R/W)</td>
<td>73K / 57K</td>
<td>117K / 57K</td>
</tr>
<tr>
<td>RAID 6 IOPS (R/W)</td>
<td>72K / 56K</td>
<td>114K / 56K</td>
</tr>
<tr>
<td>RAID 5 Throughput MB/sec. (R/W)</td>
<td>1,319 / 846</td>
<td>1,893 / 824</td>
</tr>
<tr>
<td>RAID 6 Throughput MB/sec. (R/W)</td>
<td>1,265 / 729</td>
<td>1,848 / 813</td>
</tr>
</tbody>
</table>

Rebuild Times
The question of rebuild times naturally arises when using the dual distributed parity algorithm of RAID 6. For single drive failures, the RAID 5 and RAID 6 rebuild times are comparable at about 2 hours and 20 minutes as shown in Figure 5. For two drive failures, the RAID 5 array loses all the data. The RAID 6 array protects and restores all the data. Rebuild of the RAID 6 array with two drives failures takes less than twice as long as the rebuild with one drive failure at 4 hours and 34 minutes.

During the RAID 6 rebuild, data is protected during those critical hours against a second drive failure, removal of the “wrong” bad drive and more importantly against the possibility of a single bad block ruining the rebuild entirely and causing permanent data loss. The extra two hours invested during rebuild is a small to pay for this valuable protection.

Figure 5 – Rebuild Times

<table>
<thead>
<tr>
<th>Configuration</th>
<th>(16) 600 GB</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAID 1</td>
<td>2:18</td>
</tr>
<tr>
<td>RAID 5</td>
<td>2:20</td>
</tr>
<tr>
<td>RAID 6 (1)</td>
<td>2:26</td>
</tr>
<tr>
<td>RAID 6 (2)</td>
<td>4:34</td>
</tr>
</tbody>
</table>

Best Practices
Products using custom ASICs for parity calculations deliver relatively high speed in every RAID mode. However, some RAID modes are inherently faster than others due to their architecture. Thus, RAID 10 striped mirror drives are still recommended for the small amount of data, typically critical databases, with very high performance requirements where the cost of the extra disk drives is warranted – especially with a high percentage of write activity. Most other data would be best served on RAID 6 arrays for its added reliability and vastly superior data protection.

Physical Arrays
Disk arrays from Winchester Systems offer the ultimate flexibility with expandable 12 and 16 drives base units each capable of supporting an additional six expansion shelves. The SX-2300 Series supports 84 disk drives and 168 TB of storage and the SX-3400 Series supports 112 disks and 224 TB of storage. Arrays can be built with up to 24 drives per array. Both products support simultaneous use of RAID 1, 5, 6 and 10. Thus a system with three shelves can be set up to support multiple physical arrays:

- RAID 10 with 12 disks for a critical database for performance
- RAID 6 with 12 disks for the bulk of the critical data for maximum protection
- RAID 5 with 12 disks for snapshot data for adequate protection at low cost
Summary
“High Speed RAID 6”, powered by custom high speed ASICs for parity calculations, increases MTDL by 500 to 30,000 times thus slashes risk of data loss by 2 to 4 orders of magnitude and avoids the typical performance degradation associated with most RAID 6 implementations. The FlashDisk line of primary and secondary storage with “High Speed RAID 6” delivers performance that is faster than most commercial RAID 5 arrays and far less expensive. Finally, RAID 6 eliminates the window of vulnerability for a second drive failure during a rebuild and virtually eliminates the possibility that a single bad sector discovered during a critical rebuild can cause a permanent data loss.

Conclusion
“High Speed RAID 6” disk arrays from Winchester Systems provide a unique and unmatched combination of reliability, performance and affordability in a truly open storage product that is in its tenth generation of development.

These products are available now and are simple to install, manage and service and provided by a company that has been delivering data storage products for demanding commercial, industrial and government applications since 1981.
Appendix A

RAID Mode Definitions

RAID 1 & 10 Overview
- Copy each drive to a mirror drive
- Stripe mirror drives for performance
- Doubles drive costs, power and space
- Withstand failure of one drive per pair
- Withstand failure of up to half of the drives
- Data loss if 2 drives in same pair fail

RAID 5 Overview
- One drive for single distributed parity
- Usually a second drive as a “passive hot spare” for rebuilds
- Reduces cost of protection by almost half
- Protects against loss of 1 disk drive
- No protection after 1 drive failure
- Rebuild in background
- Hardware ASIC for parity calculations

RAID 6 Overview
- Two drives for dual distributed parity
- Second parity drive replaces the RAID 5 “passive hot spare” with an “active parity disk”
- Protects against loss of any 2 disk drives
- RAID 5 protected after 1 drive failure
- Rebuild in background
Appendix B

Probabilities and MTDL Explained

MTDL calculations basically represent a compound probability. It is the probability that a second problem will occur after the first problem has occurred. The formula for compound probabilities is simply the product of the two probabilities:

\[ P(A|B) = P(A) * P(B) \]

For example, in a dice game, there are 6 ways to roll 7 out of 36 possible rolls so:

\[ P(7) = \frac{6}{36} = \frac{1}{6} \]

The probability of rolling 7 twice in a row is simply:

\[ P(7|7) = P(7) * P(7) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36} \]

When the probability of an event is small, the probability of two of them in a row is exceedingly small since the probability is multiplied by itself and is thus squared. For example, for the small probabilities of 10%, 1% and 0.1%, the compound probabilities are exceedingly small:

| P(A)   | P(A|A)   | Ratio |
|--------|---------|-------|
| 0.10%  | 0.0001% | 1000x |
| 1%     | 0.01%   | 100x  |
| 10%    | 1%      | 10x   |

This is why the MTDL in a RAID 6 array is so incredibly superior to RAID 5. The second distributed parity mechanism actually does not merely double the margin of safety, it basically squares it – resulting in the 2 to 4 orders of magnitudes higher MTDL that is depicted in Figure 3 in the white paper above.