

**WINCHESTERSYSTEMS®**

*Storage Without Complexity*

## Technology Update White Paper

### “Enterprise RAID 6”

©2006 Winchester Systems Inc. – 101 Billerica Ave., Bldg. 5 – Billerica, MA 01862 - 800-325-3700

[www.winsys.com](http://www.winsys.com)

## Technology Update White Paper

### “Enterprise RAID 6”

#### **Why Enterprise RAID 6?**

Winchester Systems has developed Enterprise RAID 6 to provide superior data protection for enterprise data storage in response to key storage industry trends and customer requirements.

Simply, Enterprise RAID 6 is a disk array that uses enterprise class disk drives, whether Fibre Channel, SAS or SATA, protects against two drive failures and ensures the successful completion of rebuilds by providing RAID 5 protection after a single disk drive failure. Thus, Enterprise 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, Enterprise RAID 6 is powerful, and yet not 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 the lack of improvement in the unrecoverable bit error rate of enterprise class disk drives.

##### *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 500 GB in just 25 years - a 100,000 *times* increase. Disk drive capacities are up 500 *times* from 1 GB to 500 GB in just the last 10 years.

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 300 GB to 500 GB 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 not changed in the past ten years. Given the dramatically higher recording densities of today’s disk drives, keeping this bit error rate constant should be viewed as a major accomplishment of the disk drive manufacturers. Nonetheless, the disk drive capacities are up 500 times. This means that 500 times the error rate is now expected compared to when RAID 5 was introduced. Soon, drives will contain 1 TB and the increase in storage per device will be 1,000-fold with commensurately 1,000 times the error rate. Any time a design is stressed 1,000 times the error rate of the original purpose – it is high 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, the rebuild fails and stops.

#### *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 500-fold increase in storage capacity per drive. Today, with multi-terabyte 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  $10^{15}$  unrecoverable error rates are considered to be enterprise class drives and those with 1 bit in  $10^{14}$  error rates are typically secondary storage comprised of low cost SATA disks which are enhanced desktop class drives.

#### *Probability of Errors*

As the data in Figure 1 clearly shows, the probability of a rebuild failure in a 2 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 and is a pretty scary 16%.

If a data center has 20 TB of total enterprise class storage and rebuilds each array once in its lifetime, there is also a 16% chance of a permanent data loss. If a total of 125 TB of rebuilds occurs over a three year period in a data center, it is a virtual guarantee that a permanent data loss will occur – even with the best drives and full RAID 5 protection. With SATA class drives 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 drives on the market used by all the storage manufacturers.

**Figure 1 – Probability of Permanent Data Loss**

Drive Type	Bit Error Rate	Byte Error Rate	Prob. Data Loss (2 TB)	Prob. Data Loss (20 TB)	Number of Data Losses (125 TB)
Secondary	1 in $10^{14}$	12.5TB	16%	160%	10
Enterprise	1 in $10^{15}$	125 TB	1.60%	16%	1

### *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.

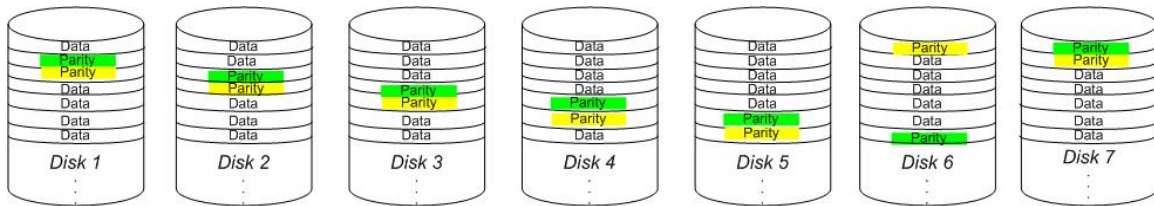
It is true that the probability of a rebuild failure is smaller for a single array it is not smaller for the storage installation as a whole. Further, smaller arrays have many disadvantages. They require more parity drives and waste storage, limit volume sizes and thus risk “disk full” messages and numerous 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 extra drives are used anyway. 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 longer than comparable RAID 5 arrays as will be demonstrated in a later section.

### **Hobson’s Choice**

Today’s storage choices typically involve enterprise class arrays with only RAID 1, 5 or 10 or desktop class arrays that offer RAID 6 – a difficult choice. Now there is a solution with the best of both worlds – Enterprise RAID 6 – that offers enterprise class disk drives and RAID 6 data protection

#### ***Enterprise drives***

$10^{**15}$  unrecoverable bit error rate  
1,200,000 hour MTBF  
100% duty cycle  
5-year drive warranty

#### ***RAID 6 protection***

500 to 30,000 *times* MTDL  
Protects against 2<sup>nd</sup> drive failure  
Increases rebuild success rate  
RAID 5 protected after drive failure

While typical commercial arrays that offer enterprise class drives and RAID 5 comprise the bulk of the installed base of equipment – and has served the IT industry well – it is time to move to the next level of protection. When presented with a choice of enterprise class drives and RAID 6 data protection – at lower cost – the choice should be simple – deploy RAID 6 when using high capacity disk drives.

### **Enterprise Class Disk Drives**

Enterprise class drives are typically Fibre Channel or SCSI disk drives that offer 1 bit in  $10^{**15}$  unrecoverable bit error rates, 1.2 million hour MTBF and 100% duty cycle.

Today, there are also new alternatives including enterprise class drives with SATA interfaces. There are two forms of these drives as outlined in Figure 3. One type is denoted as E-SATA-150 with 10,000 rpm, which originated as an enterprise class SCSI drive now has a SATA interface. This design provides the ability to provide 20 years of SCSI enterprise class reliability at SATA prices – a good combination. The other type is denoted as E-SATA-400 with 7,200 rpm, which is high quality SATA drive built to enterprise standards – again providing enterprise class reliability at SATA prices.

**Figure 3 – Enterprise Class Disk Drives**

Enterprise Disk Drive	RPM	MTBF (hours)	Error Rate	Duty Cycle	Factory Warranty	Avg Seek	Rotational Latency
E-SATA 150	10K	1.2M	10**15	100%	5 Yrs	4.6 ms	3.0 ms
FC 146	10K	1.2M	10**15	100%	5 Yrs	4.6 ms	3.0 ms
SCSI 146	10K	1.2M	10**15	100%	5 Yrs	4.6 ms	3.0 ms
E-SATA 400	7.2K	1.2M	10**15	100%	5 Yrs	8.7 ms	4.2 ms
FC 300	10K	1.2M	10**15	100%	5 Yrs	4.6 ms	3.0 ms
SCSI 300	10K	1.2M	10**15	100%	5 Yrs	4.6 ms	3.0 ms

### Cost Savings

Enterprise RAID 6 disk arrays using E-SATA-150 drives typically cost about 20% less than comparable RAID 5 arrays using Fibre Channel disk drives. These enterprise class drives are designed to deliver SCSI reliability in the drive and take advantage of high volume SATA pricing. Similarly Enterprise RAID 6 arrays using E-SATA-400 drives can cost up to 40% less than comparable RAID 5 Fibre Channel arrays. In this case, the drives are slightly lower performance with 7,200 rpm. However, the overall system performance is still extremely high and will satisfy all but the most demanding high performance applications with ease at a tremendous cost reduction.

### 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 gets worse with increasing array sizes.

*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 4 shows the MTDL figures 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 a fraction of a single year.*

Figure 4 – Mean Time to Data Loss

Array Capacity	RAID 5 MTDL (months)	RAID 6 MTDL (months)	Reliability Increase Ratio
1 TB	100	3,000,000	30,000x
2 TB	50	500,000	10,000x
5 TB	10	20,000	2,000x
10 TB	7	7,000	1,000x
20 TB	2	1,000	500x

Source: Intel Corp.

SATA 10\*\*14 disk drives

### Unbeatable Combination

When hardware powered RAID 6 is combined with enterprise-class disk drives with 1 in  $10^{**15}$  unrecoverable bit error rates, 1.2 million hour MTBF and rated for 100% duty cycle, the combination is an unbeatably 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.

### Enterprise RAID 6 Performance Breakthrough

Most 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.

Winchester Systems designed RAID 6 for enterprise class performance and uses dual hardware ASICs, one for each of the two independent parity calculations involved in the dual parity algorithms in RAID 6. Using this high-speed hardware, Enterprise RAID 6 demonstrates minimal performance drop from RAID 5. Figure 5 demonstrates the performance of FlashDisk SA-3400 series products with sixteen 400 GB, 7,200 rpm disk drives. Notice that the sequential write performance of RAID 6 (35,089 IOPS) is within 3% of the comparable RAID 5 performance (36,513). RAID 6 random writes have a higher degradation of about 30%, however, the absolute performance is still excellent compared to many other RAID 5 arrays, and will have I/O capacity to spare is the bulk of today's applications. The data for the 150 GB, 10,000 rpm disk drives show similar results. Thus, performance is still high and data protection is vastly improved by orders of magnitude.

**Figure 5 – E-SATA-400 Performance**

Measure	RAID	Seq Read	Seq Write	RND Read	RND Write
IOPS	R1	50,008	35,337	1,240	1,443
IOPS	R5	50,931	36,513	1,277	584
IOPS	R6	49,447	35,089	1,266	406
MB/s	R1	343	266	145	97
MB/s	R5	420	371	164	68
MB/s	R6	412	280	163	67

**Figure 6 – E-SATA-150 Performance**

Measure	RAID	Seq Read	Seq Write	RND Read	RND Write
IOPS	R1	45,668	34,383	3,137	2,134
IOPS	R5	44,180	31,586	2,284	481
IOPS	R6	42,673	31,559	2,387	1,247
MB/s	R1	446	264	219	124
MB/s	R5	428	367	167	59
MB/s	R6	415	260	163	73

*Results above are for FlashDisk SA-3400 models with single controller and 16 disk drives.  
Dual controller modes generally offer twice the data rates.*

### **Rebuild Times**

The question of rebuild times naturally arises when using the dual distributed parity algorithm of RAID 6. Yes, rebuild times are longer. For example, it takes 2 hours and 37 minutes to rebuild a RAID 6 array after one drive failure compared to 1 hour and 16 minutes for a comparable RAID 5 array. However, the RAID 6 rebuild is completely protected during those critical hours against a second drive failure and more importantly against the possibility of a single bad block ruining the rebuild entirely and causing permanent data loss. The extra hour invested during rebuild is an extremely trivial price to pay for this valuable protection. You will also note that the RAID 6 rebuild after two drive failure is 5 hours and 5 minutes – twice as long as with a single rebuild in process. However, everyone with two drives failures would gladly let the RAID array process data for 5 hours to automatically recover the data rather than experience the certain and permanent data loss that would be incurred in RAID5. Easy choice – wouldn't you agree?

**Figure 7 – Rebuild Times**

Configuration	(16) 150 GB	(24) 150 GB	(16) 400 GB
RAID 1	0:41	0:40	3:05
RAID 5	1:16	2:06	3:07
RAID 6 (1)	2:37	3:52	3:15
RAID 6 (2)	5:05	7:50	6:06

**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. Most other data would be best served on RAID 6 arrays for its added reliability and vastly superior data protection. Finally, less critical data would be stored on RAID 5 arrays, especially copies of data including snapshot and replicated data, on-line archives where tape backups exist, temporary files, output files, reports or other reproducible data.

**Physical Arrays**

Disk arrays from Winchester Systems offer the ultimate flexibility with 12, 16 and 24 drive enclosures and support for simultaneous use of RAID 1, 5, 6 and 10. Thus a single 24-drive enclosure can be set up to support multiple physical arrays. For example, a 24-drives enclosure could be set as follows:

- RAID 10 with 8 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 4 disks for snapshot data for adequate protection at low cost

**Phase In Plan**

After reviewing the benefits of RAID 6, it makes sense to determine how to phase in the new higher reliability technology. Most people cannot easily replace all their old equipment at one time. Thus it makes sense to install Enterprise RAID 6 for critical applications and redeploy or trade in older RAID 5 equipment. RAID 5 equipment can be demoted to less critical applications, disk-to-disk backup or remote replication site use or may be traded in to the original vendor for credit or sold on the third party market.

**Summary**

Enterprise RAID 6 offers a unique combination of enterprise class disk drives with 1 bit in  $10^{15}$  bit error rate, 1.2 million hour MTBF and 100% duty cycle and RAID 6 dual parity protection. It increases MTDL by 500 to 30,000 times thus slashes risk of data loss by 2 to 4 orders of magnitude. When combined with dual hardware ASICs it delivers enterprise level I/O performance and is faster than most commercial RAID 5 arrays and far less expensive than typical RAID 1 arrays. 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**

Enterprise 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 eighth 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.

For More Information:

### **Winchester Systems Inc.**

10 Billerica Ave., Bldg. 5  
Billerica, MA 01862

**800-325-3700**

**[www.winsys.com](http://www.winsys.com)**

## 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 extra drive for single distributed parity
- 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 extra drives for dual distributed parity
- Second parity drive replaces the “passive hot spare” with an “active hot spare” at no extra cost
- Protects against loss of *any* 2 disk drives
- RAID 5 protected after 1 drive failure
- Rebuild in background

#### ***Enterprise RAID 6 Overview – All of the above RAID 6, PLUS***

- Utilizes enterprise class disk drives featuring
  - 1 in  $10^{**15}$  bit error rate
  - 1.2 million hour MTBF (Mean Time Between Failure)
  - 100% duty cycle rating
- Redundant hardware ASICs for high-speed parity calculations
- Runs RAID 6 at near RAID 5 speeds
- Increases MTDL (Mean Time to Data Loss) by 500 to 30,000 times

## 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:

$$P(7) = 6/36 = 1/6$$

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

$$P(7|7) = P(7) * P(7) = 1/6 * 1/6 = 1/36$$

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 is 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 4.

## Appendix C

### Vastly Superior Data Protection

#### 24-Bay Enterprise RAID 6

- High speed RAID 6 with redundant hardware ASIC parity chips
- 24 Enterprise drives
- 770 MB/second read
- 460 MB/second write
- 87,216 IOPS read
- 33,462 IOPS write
- Single or dual active redundant controllers
- (2) U320 SCSI or (8) 2 Gb Fibre Channel ports
- 2 GB cache RAM



**FlashDisk SA-4500**

## Appendix D

### 16-Bay Enterprise RAID 6

- High speed RAID 6 with redundant hardware ASIC parity chips
- 16 Enterprise drives
- 510 MB/second read
- 350 MB/second write
- Single controller
- 4 Gb Fibre Channel
- (2) U320 SCSI or (2) 4 Gb Fibre Channel ports
- 2 GB cache RAM



**FlashDisk SA-3000**