September 2010

A Better RAID Strategy for High Capacity Drives in Mainframe Storage

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Disclaimer 1
Introduction 2
RAID 6 Data Integrity Now and Then.......... 2
What Causes This Change 4
Time to Reshuffle RAID 6 5
  Increasing the Number of Parities from Two to Three 5
  Limiting Drive Capacities 6
  Global Hot Spare Drive and Others 6
Conclusion 7

Introduction

RAID6 has a long, proven history with Mainframe enterprise storage customers. However, two trends render RAID 6 less capable of meeting Mainframe enterprise storage systems' reliability requirements today. First, there have been significant increases in Hard Disk Drive (HDD) capacities. Second, the integration of SATA and SAS HDDs, currently many of which are evolved from consumer class HDDs, expose flaws in RAID 6 that require a different data protection strategy. This paper demonstrates that RAID 6 is less capable to prevent data loss with today's large capacity SATA and SAS HDDs. It will also share Oracle's designs to mitigate this risk. The integrity of customers' data is in jeopardy if these large capacity SATA/SAS HDDs are simply packaged into a RAID 6 group without additional mitigation.

RAID 6 Data Integrity Now and Then

Table 1 shows the specifications for the different types of drives this study covered. 146 GB Fibre Channel (FC) drives are popular in Mainframe enterprise storage systems. Other common FC HDDs have a capacity of 300GB or 400GB. While there is a reduction in RAID 6's Mean Time To Data Loss (MTTDL) from FC 146 GB to FC 400GB, it is not significant, so only FC 146 GB drives are used as an example. Table 1 also shows the specifications for the largely accepted high capacity drives in Mainframe enterprise storage systems.

Note in Table 1:

- The HDD vendors specify Mean Time Between Failure (MTBF) for each type of drives. Many independent researchers and storage vendors' field data, including Oracle's, found a lower MTBF than these specifications. In addition, there is still a debate about whether SATA or SAS HDD's MTBF are really less than FC HDDs. Because the specifications are so debatable, and to ensure our conclusion is widely applicable, a MTBF of 1,000,000 hours is used for all three types of drives.
- HDD rebuild times are normalized to protect proprietary information. Exact rebuild time can be varied by many factors, but we believe the increasing trend in Table 1 still applies.

TABLE 1. HDD AND RAID 6 CONFIGURATION

Parameter	Value			Notes
	FC 146GB	SATA 1TB	SAS 2TB	
RAID 6	12 drives	12 drives	12 drives	RAID 6: 11 total drives including 2 rotational parities, plus 1 additional global hot spare drive.
HDD Capacity	146 GB	1 TB	2TB	
MTBF	1,000,000	1,000,000	1,000,000	HDD MTBF (hours)
P2p	2	2	2	Ratio of time to rebuild 2 parity vs. rebuild 1 parity
Rebuild	1x	3.7x	7.7x	HDD Rebuild time (hours)
Response	24	24	24	Field response time to replace the failed drive (hours)

UBER	10^{16}	10^{15}	10^{15}	Uncorrectable Bit Error Rate

Table 1 includes Uncorrectable Bit Error Rate (UBER), which is an important reliability metric. As some reference, UBER for LTO tape is 10^{16} and Oracle T10K tape is 10^{19}. More discussion of tape UBER is beyond the scope of this paper.

Based on the above configurations, the MTTDL for different RAID 6 groups can be calculated using a Markov chain[1]. The results are shown in Figure 1. To effectively display different MTTDLs, a large range of values shown in Figure 1, a logarithmic scale presentation of the same data is given in Figure 2.

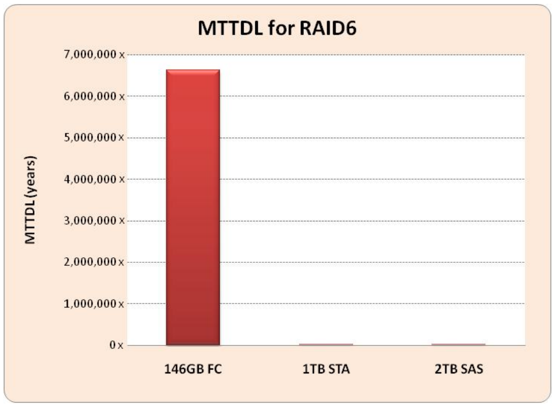

FIGURE 1. MTTDL FOR DIFFERENT TYPE OF HDD

While the high capacity drives enjoy a reduced cost, they also have dramatically lower reliability. In fact, one RAID 6 group based on 1TB SATA drives will have 1,000 times lower data integrity than the RAID 6 group with 146GB FC drives. And that's just one RAID group of drives. In reality, Mainframe enterprise storage systems usually employ hundreds of HDDs in multiple RAID 6 groups so the probability of data loss is even higher, especially when you factor in software-caused data loss errors. We believe that a RAID 6 strategy is not adequate in today's environment. It will only leave customers bitter after data loss failures. Clearly, a new data protection paradigm is necessary.

[1] A modeling technique for reliability analysis, where a system is described by its possible states and the possible transitions between these states.

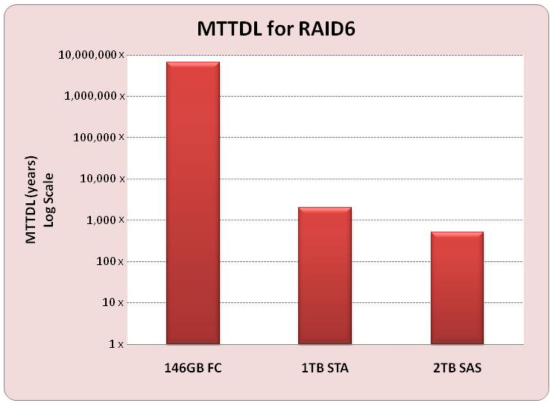

FIGURE 2. MTTDL FOR DIFFERENT TYPE OF HDD (LOG SCALE)

What Causes This Change

There are two main reasons that the MTTDL is so much lower in today's SAS and SATA drives:

- The vast increase in HDD capacities
- The UBER during reads on the SAS and SATA drives

The large capacity of new HDDs means that the same data protection algorithm needs to cope with much more data. For example, a 1 TB drive has 6.8 times more capacity than its 146 GB counterpart, so the risk of data loss is also 6.8 times higher assuming everything else is equal. Unfortunately, the risk increases even more because the larger drives also require longer rebuild times as shown in Table 1.

Another significant change is the new drives' UBER drops from 10^{16} to 10^{15} because of their roots in the consumer class market. This drop may seem insignificant, but its impact becomes compounded in the MTTDL formula. In a RAID 6 group (11 total drives with 2 parities) when a rebuild is necessary, all eight alive HDDs and one parity drive would need to be read successfully for the rebuild of the failed drive to be successful as well[2]. If one bit error returned during the read process, the rebuild would be deemed a failure by the RAID controller. In a RAID 6 implementation, the controller would then issue another rebuild, pulling in the second parity. In either case, the probability of a rebuild failure caused by the lower UBER is:

$1-(1-11015)9\times8\times1012=1-0.9358=6.94\%$

In contrast, the rebuild failure probability for a RAID 6 group consisting of the same number of 146 GB FC drives is only 0.117%. This is a huge degradation from the FC drives to the SAS and SATA drives.

[2] For illustration purposes only. The actual data in RAID 6 are striped, and two parity blocks are distributed across all 11 member drives.

There are other factors that can also impact MTTDL, but our sensitivity study reveals that they are relatively insignificant compared to the capacity and UBER issues. Other factors include high I/O throughput drives having a higher probability of discovering an UBER error in the same backup window, different number of drives per RAID group, different rebuild times, and varied drive scrubbing times.

Time to Reshuffle RAID 6

There are several potential alternatives to mitigate the MTTDL issues that result from the traditional RAID 6 strategy. Possible mitigations include:

- Assigning fewer data drives in RAID 6
- Increasing the number of parities from two to three
- Limiting drive capacities
- Implementing one or more global hot spare drives
- Shortening rebuild times
- Shortening disk scrubbing times

Allocating fewer drives per RAID 6 group can be an easy technical solution. Reducing the number of drives in a group from 11 to 10 increases the MTTDL by 39%. However, the benefits may defeat the purpose of having high capacity drives in the first place.

Increasing the Number of Parities from Two to Three

To optimize the drives' capacity while minimizing the MTTDL risk, a different RAID strategy can be implemented. The Oracle Zettabyte File System (ZFS) increases the number of parities from two to three in a group. This new RAID strategy is called RAID Z3. The comparison between the traditional RAID 6 group described previously that has eleven total drives, with two rotating parities, and a global hot spare with a RAID Z3 group that has twelve total drives, with three rotating parities, and a global hot spare is shown in Figure 3. The drive used in Figure 3 is a 2 TB drive. The MTTDL increases by a whopping 516% with RAID Z3. RAID Z3 can now meet the Mainframe enterprise storage customer's data integrity requirements.

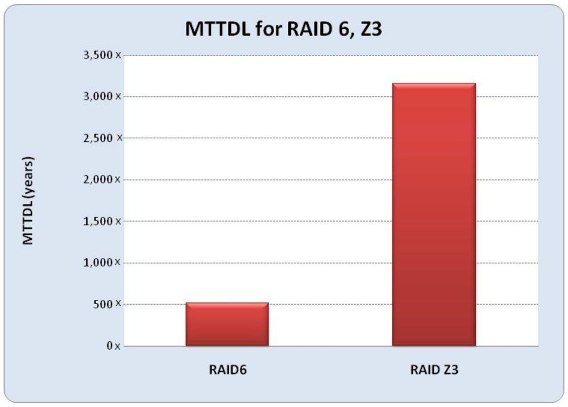

FIGURE 3. MTTDL FOR RAID 6, RAID Z3

Limiting Drive Capacities

At first, the idea of limiting a drive's capacity would seem counterproductive. However, slightly shaving the total capacity does have benefits as shown with the Oracle Virtual Library Extension (VLE) product. Based on customer usage profiles, VLE presents 85% of the HDD's physical capacity to the customer. This does not impact the customer's normal backup operation, but improves the RAID Z3 data reliability by 62%. Other reliability benefits are also observed from this change. It's believed that drives might be more prone to reliability issues when their capacity reaches a very high mark because of especially busy drive movements and garbage collection activities. Limiting a drive's capacity to only 85% gives us a chance to bypass this potential hot spot.

Implementing a Global Hot Spare Drive and Other Possible Mitigations

Once a failed drive is detected, a hot spare drive will allow the RAID controller to rebuild the failed drive automatically without requiring field service personnel to physically replace the failed drive. Allowing hot spare drives to be globally available to any RAID group is another important reliability feature. If hot spare drives are only dedicated to one specific RAID group, the overall RAID 6 reliability is inferior to the global sharing.

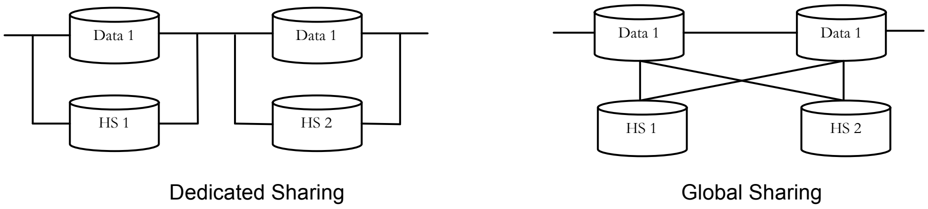

FIGURE 4. HOT SPARE DRIVE SHARING STRATEGY

To illustrate the effect of global sharing, consider four drives in a simple configuration (ignoring RAID implementation) with two data drives, and two hot spare drives. The two potential configurations can have dedicated or globally shared spares as described in Figure 4.

The MTTDL for a configuration with globally shared spares is 18% higher than with dedicated sharing. The Oracle VLE always employs a global hot spare design to reinforce the RAID Z3 reliability.

Shortening rebuild times, and faster disk scrubbing are other possible alternatives to increase a RAID group's data integrity. While these alternatives sound intriguing, our study shows the changes in these parameters do not significantly improve MTTDL.

Oracle VLE applies RAID Z3 protection, limits drive capacity (2 TB raw) to present only 85% to the end user, and assigns hot spare drives globally to all RAID groups. All these features enable VLE to easily meet the Mainframe enterprise storage customer's data integrity requirement. The improved MTTDL is shown in Figure 5. To be consistent, RAID 6 and RAID Z3 in Figure 5 utilizes the same configuration as in Figure 3, while RAID Z3 (85%) confines the drive capacity to be used to only 85%. VLE has slightly different RAID Z3 configuration, which has eleven (not twelve) total drives, with three rotating parities, and a global hot spare.

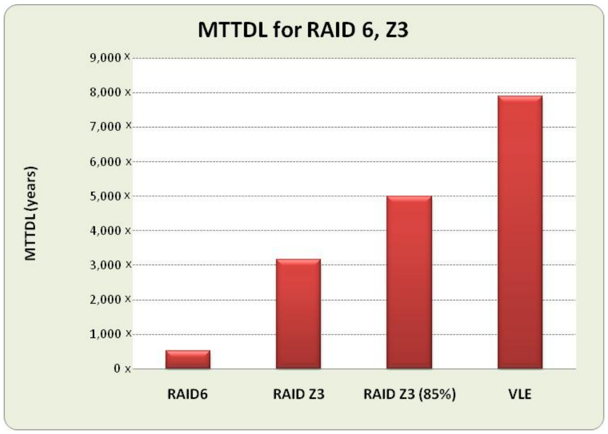

FIGURE 5. MTTDL FOR VLE V.S. **OTHER COFIGURATIONS**

Conclusion

Our study demonstrates that adopting large capacity SATA and SAS drives into Mainframe enterprise storage system requires a more reliable system design to tolerate HDD and rebuild failures. RAID 6 is incapable of meeting Mainframe enterprise customer's data integrity requirements, with large capacity SAS or SATA drives. A combination of RAID Z3 and other options is an excellent alternative. Without a fault tolerant system design, simply packaging new wine of large capacity HDDs into the old RAID 6 bottle will leave a sour taste in Mainframe enterprise customers' mouths with the high probability of data loss. For a smoother experience with today's drives, a modern RAID Z3 implementation is required.

ORACLE®

Oracle is committed to developing practices and products that help protect the environment

A Better RAID Strategy for High Capacity Drives
in Mainframe Storage
September 2010

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

SOFTWARE. HARDWARE. COMPLETE.