Sun ORACLE | STORAGETEK

January 2013

A Better RAID Strategy for High Capacity Drives in Mainframe Storage

ORACLE

Introduction

RAID6 has a long, proven history with Mainframe enterprise storage customers. However, two trends render RAID6 less capable of meeting Mainframe enterprise storage systems' reliability requirements today. First, there have been significant increases in Hard Disk Drive (HDD) capacities, with a single disk drive having more than 20x the capacity of previous drives. Second, the integration of SATA and SAS HDDs, currently many of which are evolved from consumer class HDDs, expose flaws in RAID6 that require a different data protection strategy. This paper demonstrates that RAID6 is less capable of preventing 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 in a RAID6 group without additional mitigation.

RAID6 Data Integrity Now and Then

Table 1 shows the specifications for the different types of drives this study covered. 146 and 450 GB Fibre Channel (FC) drives are still in use in Mainframe enterprise storage systems. Table 1 also shows the specifications for the largely accepted high capacity SAS and SATA drives in Mainframe enterprise storage systems.

TABLE 1. HDD AND RAID6 CONFIGURATION

Parameter	Value				Notes
	FC 146 GB	FC 450GB	SATA 1TB	SAS 3TB	
RAID6	12 Drives	12 Drives	12 Drives	12 Drives	RAID6: 11 total drives including 2 rotational parities, plus 1 additional global hot spare drive.
HDD Capacity	146 GB	450 GB	1 TB	3TB	
MTBF	1,000,000	1,000,000	1,000,000	1,000,000	HDD MTBF (hours)
Multiple Rebuild Factor	1.7	1.7	1.7	1.7	Ratio of time to rebuild 2 parity vs. rebuild 1 parity
Rebuild	1x	1.5x	2x	6x	HDD Rebuild time (hours)
Response	24	24	24	24	Field response time to replace the failed drive (hours).
UBER	10^{16}	10^{16}	10^{15}	10^{15}	Uncorrectable Bit Error Rate

Note in Table 1:

- The HDD vendors specify Mean Time Between Failure (MTBF) for each type of drive. 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 vary due to many factors, but we believe the increasing trend in Table 1 still applies.
- Table 1 includes Uncorrectable Bit Error Rate (UBER), which is an important reliability metric. As a point of reference, UBER for LTO tape is 10^{16} and Oracle T10K tape is 10^{19}. UBER frequency is higher with SAS and SATA drives now being used in enterprise class mainframe environments.

UBER, in conjunction with the other metrics in Table 1, were used to calculate the Mean Time To Data Loss (MTTDL) for various size drives. Figure 1, below, shows MTTDL calculated using Markov Chains for RAID6 configurations of eleven drives per RAID group plus one spare drive[1].

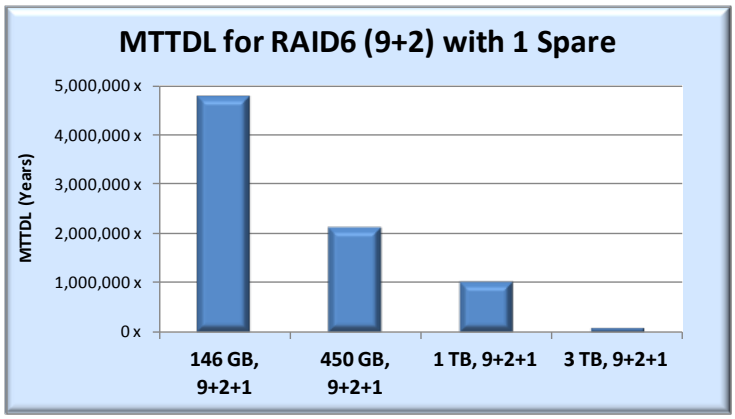

FIGURE 1. MTTDL FOR DIFFERENT DISK CAPACITIES

While the high capacity drives enjoy a reduced cost, they also have **dramatically** lower reliability. In fact, one RAID6 group with 3 TB SATA drives will have 60 times lower data integrity than the RAID6 group with 146 GB FC drives. That's 6000 percent! And that's just one RAID group of drives. In reality, Mainframe enterprise storage systems usually employ hundreds of HDDs in multiple RAID6 groups so the probability of data loss is much higher. In addition, drive capacities will continue to increase with 4 TB drives already on the near horizon. We believe that a RAID6 strategy is not adequate in today's environment. Clearly, a new data protection paradigm is necessary.

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

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 3 TB drive has 20 times more capacity than its 146 GB counterpart, so the risk of data loss is also 20 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 RAID6 group (11 total drives with 2 parities) when a rebuild is necessary, all eight surviving HDDs and one parity drive would need to be read successfully for the rebuild of the failed drive to be successful[2]. If one bit error returned during the read process, the rebuild would be deemed a failure by the RAID controller. In a RAID6 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 for a 3 TB drive is:

$$1 - (1 - \frac{1}{10^{15}})^{(9\times8\times3E12)} = 19.41\%$$

In contrast, the rebuild failure probability for a RAID6 group consisting of the same number of 146 GB FC drives is only 1.04 %. This is a huge degradation from the FC drives to the SAS and SATA 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.

Time to Rethink RAID6

There are several potential alternatives to mitigate the MTTDL issues that result from the traditional RAID6 strategy. Possible mitigations include:

- Assigning fewer data drives in RAID6
- Limiting drive capacities
- Implementing one or more global hot spare drives
- Increasing the number of parities from two to three

[2] For illustration purposes only. The actual data in RAID6 are striped, and two parity blocks are distributed across all 11 member drives.

Assigning Fewer Data Drives in RAID6

Allocating fewer drives per RAID6 group can be an easy technical alternative. Reducing the number of drives in a group of 3 TB drives from 11 to 9 increases the MTTDL by a factor of 2. However, a factor of 2 doesn't come close to making up for the MTTDL impact due to higher capacity drives. Further reductions in the number of drives per RAID6 group will increase costs while lowering capacity, with only limited success in mitigating the data loss risk.

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) and VSM6 products. These products were designed to present 85% of the HDD's physical capacity to the customer, after analysis of Customer Usage Profiles. This does not impact the customer's normal backup operation, but improves the RAID data reliability by at least 10%. 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 85% gives us a chance to bypass this potential hot spot.

Implementing a Global Hot Spare Drive

Once a failed drive is detected, a hot spare drive will allow the RAID controller to rebuild the failed drive automatically without requiring 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 benefit to reliability is limited. Globally sharing spares versus dedicating spares to specific RAID groups increases the time to data loss by 18%

Increasing the Number of Parities from Two to Three

The Oracle Zettabyte File System (ZFS) increases the number of parities from two to three in a RAID group. This new RAID strategy is called RAIDZ3 and it optimizes the drives' capacity while minimizing the MTTDL risk. Figure 3 compares RAID6 and RAIDZ3, each with one shared spare. The RAID6 group has the equivalent of nine data drives and two parity drives plus a spare (9+2+1). The RAIDZ3 group has the equivalent of nine data drives and three parity drives plus a spare (9+3+1).

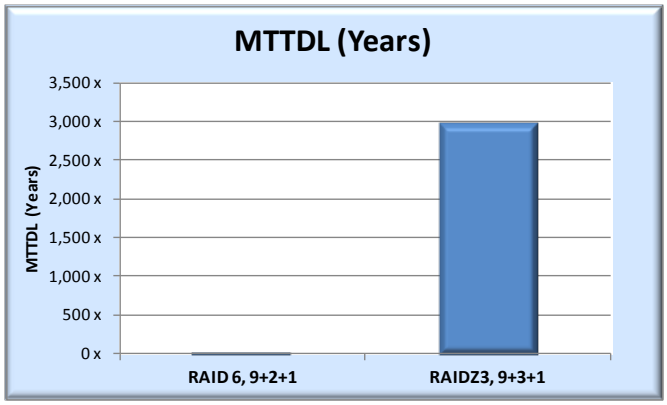

FIGURE 3. MTTDL FOR RAID6 & RAID Z3

The MTTDL increases by a factor of 300! It is plain to see the tremendous data protection benefits of RAIDZ3. RAIDZ3 can now meet the Mainframe enterprise storage customer's data integrity requirements.

Oracle's Approach

Oracle's VLE and VSM6 systems combine the huge data protection improvements of RAIDZ3 with globally shared spares. Additionally, VLE and VSM6 present only 85% of the drive capacity to the end user to further minimize the risk of data loss. All of these features enable VSM6 and VLE to easily meet the Mainframe enterprise storage customer's data integrity requirements.

The increases in MTTDL are shown in Figure 3. "RAID6, 9+2+1, 85%" represents nine data plus two parity drives plus one shared spare, and the drive capacity is confined to 85%. "RAIDZ3, 9+3+1" represents nine data plus three parity drives plus one shared spare. VLE and VSM6 have slightly different RAID Z3 configurations, with eleven (not twelve) drives per RAID group with three rotating parities, plus a global hot spare. Therefore, the final configuration, "VLE & VSM6, 8+3+1, 85%" represents eight data plus three parity drives, plus one shared spare, and confines the drive capacity to 85%.

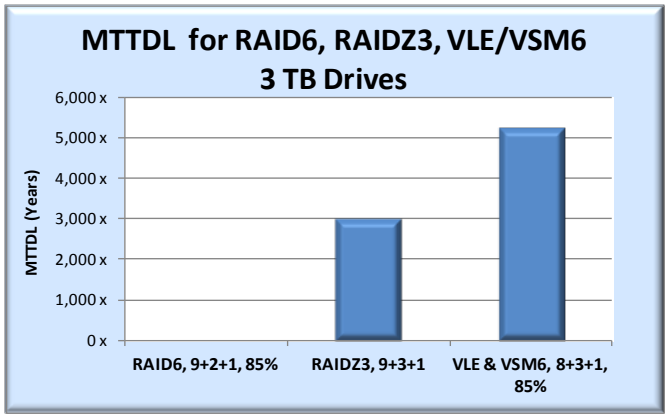

FIGURE 3. MTTDL FOR RAID6, RAIDZ3 AND VLE/VSM6 CONFIGURATIONS

Conclusion

Our study demonstrates that adopting large capacity SATA and SAS drives into Mainframe enterprise storage systems requires a more reliable system design to tolerate HDD and rebuild failures. RAID6, the former standard for data reliability, is incapable of meeting Mainframe enterprise customer's data integrity requirements with large capacity SAS or SATA drives. RAIDZ3 increases the number of parity drives per RAID group by one, and increases the mean time to data loss (MTTDL) by a factor of 300 over a similar RAID6 configuration. The combination of RAIDZ3 and other options is an excellent alternative to RAID6. Without a fault tolerant system design, simply packaging large capacity HDDs into the old RAID6 configurations will leave Mainframe enterprise customers' with a high probability of data loss. For a smoother experience with today's drives, a modern RAIDZ3 implementation is required.

A Better RAID Strategy for High Capacity Drives in Mainframe Storage
January 2013

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

Copyright © 2013, Oracle and/or its affiliates. All rights reserved.
This document is provided for information purposes only and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. UNIX is a registered trademark licensed through X/Open Company, Ltd. 0910

SOFTWARE. HARDWARE. COMPLETE.