RAID EE Technology

White Paper

July 2022

ANNOUNCEMENT

Copyright

QSAN believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

Trademarks

- QSAN, the QSAN logo, QSAN.com, XCubeFAS, XCubeSAN, XCubeNXT, XCubeNAS, XCubeDAS, XEVO, SANOS, and QSM are trademarks or registered trademarks of QSAN Technology, Inc.
- Microsoft, Windows, Windows Server, and Hyper-V are trademarks or registered trademarks of Microsoft Corporation in the United States and / or other countries.
- Linux is a trademark of Linus Torvalds in the United States and / or other countries.
- UNIX is a registered trademark of The Open Group in the United States and other countries.
- Mac and OS X are trademarks of Apple Inc., registered in the U.S. and other countries.
- Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and / or its affiliates.
- VMware, ESXi, and vSphere are registered trademarks or trademarks of VMware, Inc. in the United States and / or other countries.
- Citrix and Xen are registered trademarks or trademarks of Citrix Systems, Inc. in the United States and / or other countries.
- Other trademarks and trade names used in this document to refer to either the entities claiming the marks and names or their products are the property of their respective owners.

TABLE OF CONTENTS

FIGURES

TABLES

Notices

Information contained in this document has been reviewed for accuracy. But it could include typographical errors or technical inaccuracies. Changes are made to the document periodically. These changes will be incorporated in new editions of the publication. QSAN may make improvements or changes in the products. All features, functionality, and product specifications are subject to change without prior notice or obligation. All statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

PREFACE

Executive Summary

The RAID architecture which has been in existence for over 30+ years now is undergoing a wave of transformation. The original RAID technology has failed to address the problem of excessive rebuilding time due to TB class large capacity hard disk drives. The new generation RAID technology based on the traditional block technology, which we call RAID EE, is seen as a path to solving the traditional RAID flaw.

Audience

This document is applicable for QSAN customers and partners who are interested in learning about RAID EE for solving the problem of excessive rebuilding time. It assumes the reader is familiar with QSAN products and has general IT experience, including knowledge as a system or network administrator. If there is any question, please refer to the user manuals of products, or contact QSAN support for further assistance.

Technical Support

Do you have any questions or need help trouble-shooting a problem? Please contact QSAN Support, we will reply to you as soon as possible.

- Via the Web: https://www.qsan.com/technical_support
- Via Telephone: +886-2-77206355
- (Service hours: 09:30 - 18:00, Monday - Friday, UTC+8)
- Via Skype Chat, Skype ID: qsan.support
- (Service hours: 09:30 - 02:00, Monday - Friday, UTC+8, Summer time: 09:30 - 01:00)
- Via Email: support@qsan.com

Information, Tip, and Caution

This document uses the following symbols to draw attention to important safety and operational information.

INFORMATION

INFORMATION provides useful knowledge, definition, or terminology for reference.

TIP

TIP provides helpful suggestions for performing tasks more effectively.

CAUTION

CAUTION indicates that failure to take a specified action could result in damage to the system.

1. OVERVIEW

RAID (Redundant Array of Independent Disks) is to combine multiple independent physical disks based on a certain algorithm to form a virtual logical disk that provides a larger capacity, higher performance, or better data error tolerance. RAID has been the basic technology of storage system as a mature and reliable data protection standard. However, with the rapid growth in demand of disk drive for data storage and the advent of high-performance applications in recent years, traditional RAID has gradually revealed its defects.

As hard disk capacity increases, the amount of time required to rebuild RAID data has also dramatically increased. This makes one of the most troubles for enterprise storage management today. In the past days when the hard disk capacity was only 10 GB to 100 GB, RAID built was a job that could be completed in 10 minutes or even more than 10 minutes, which was not yet a problem without special concern. However, as disk capacity grows to hundreds of GB and even TB, RAID rebuild times have increased to hours or even days, it becomes a major problem in storage management.

For example, a traditional RAID 5 with 8 and 1 parity on 6 TB NL-SAS disk drives takes 2.5 days to rebuild data. The rebuild process consumes system resources, reducing the overall performance of the application system. If users restrict the rebuild priority, the rebuild time will be even longer. The important of all, during time-consuming rebuilding, a large number of access operations could cause the failure of other disk drives in the pool, greatly increasing the probability of disk failure and the risk of data loss.

1.1. Limitations of Traditional RAID Architecture

The traditional RAID architecture is composed of a certain number of disk drives which are selected to form a disk group (also known as RAID group). You may also assign some disk drives as idle hot spare disk drives. A storage pool is grouped to provide capacity for volumes, and then finally map the LUN to the host to become the storage space on the host.

There are several limitations in such RAID architecture:

- First of all, when a disk drive of the disk group is damaged and the rebuild is required, only the member disks of the disk group participates in the rebuild job, and the data writing loading at the time is concentrated on the spare disk to form a bottleneck.

- Second, volume data access is limited to the member disks belonging to a disk group; this restricts the performance available to the host because the storage is executing both accessing and rebuilding I/O.

1.2. Why RAID Rebuild Time-Consuming

As drive capacity grows, RAID rebuild time grows linearly, raising the rebuild time required by traditional RAID architectures to tens of hours when using RAID disks with more than 4TB HDD capacity.

There are several factors that affect the RAID rebuild time:

- **HDD Capacity:** The HDD capacity makes up the disk group, the larger the HDD capacity, the longer the rebuild time is required.
- **Quantity of Disk Drives:** The quantity of disk drives included in a disk group affects the amount of time it takes for the system to read data from the remaining healthy disk drives and write them to the hot spare disk drives. The more disks, the longer the rebuild time.
- **Rebuild Job Priority:** During RAID rebuild, the system still has to assume I/O access to the front-end host. The higher the priority assigned to the RAID rebuild job, the faster the rebuild, but the less the front-end host gains I/O performance.
- **Fast Rebuild:** Enabling fast rebuild function only need to rebuild the actual capacity of the volume, unused disk group space has not to rebuild. If only part of the space in a disk group is used by the volume, the rebuild time will be shortened.
- **RAID level:** RAID 1 and RAID 10 with direct block-to-block replication will rebuild faster than RAID 5 and RAID 6 with parity calculations.

Given the potential for failure on each disk drive, the more disk drives contain in a disk group, the more possibility of cumulative failure increase, so there is an upper limit on the quantity of disk drives in a disk group. Compared with the previous factors, the increasing impact of the disk drive capacity on the rebuild speed has become the primary factor. Such a long rebuild time is apparently not acceptable to any user. To solve the problems of traditional RAID, we implement RAID EE technology.

2. THEORY OF OPERATION

RAID EE adds more spare disks in a disk group, we call them **RAID EE spares** to separate the original global, local, and dedicated spares. Spare areas are preserved in each stripe of the disk group and are distributed in the disk group by means of disk rotation. When disks have failed in the disk group, missing data is rebuilt into the preserved spare areas. Since all disks in the set are destination of rebuilt data, the bottleneck of traditional RAID rebuild is gone, rebuild performance dramatically improved. If new disks are added in, data in spare areas are copied back to new joined disks.

Four new RAID levels are provided for RAID EE, there are:

- **RAID 5EE** (E stands for Enhanced), requires a minimum of 4 disk drives with one RAID EE spare disk which can tolerate 2 disk drives failure. Adding more RAID EE spares will tolerate more disk drives failure.
- **RAID 6EE** requires a minimum of 5 disk drives.
- **RAID 50EE** requires a minimum of 7 disk drives.
- **RAID 60EE** requires a minimum of 9 disk drives.

INFORMATION

The RAID EE spare quantity in a disk group is 1 to 8 disk drives.

2.1. Example of RAID 5EE with 1 RAID EE spare

Now we take an example to describe how it works. The following example is a RAID 5EE with 5 disks. 4 disks are for RAID disks, and additional one disk is for RAID EE spare. After initialization, data block distribution is as follows. P is stands for parity, S is stands for RAID EE spare, and it is empty now.

D1	D2	D3	D4	D5
1	2	3	P	S
S	4	5	6	P
P	S	7	8	9
10	P	S	11	12
13	14	P	S	15

Figure 2-1 Data Block Distributed of RAID 5EE

Assume that disk 2 has failed. RAID 5EE is under degraded mode.

D1	D2	D3	D4	D5
1	2	3	P	S
S	4	5	6	P
P	S	7	8	9
10	P	S	11	12
13	14	P	S	15

Figure 2-2 Disk 2 has Failed

The spare areas are rebuilt with data from the failed disk drive. This action is called **EE Rebuild**. After rebuild, data distributed is like RAID 5 and it can tolerate another failed disk drive. As we can imagine, the more RAID EE spare disks, the faster it rebuilds.

D1		D3	D4	D5
1		3	P	2
4		5	6	P
P		7	8	9
10		P	11	12
13		P	14	15

Figure 2-3 Empty Blocks are Rebuilt from the Failed Disk Drive

When a new disk drive is joined into the RAID EE disk group, the data rebuilt in the spare area will be copied back to the new disk. This action is called **Copyback**.

D1	D2	D3	D4	D5
1	2	3	P	S
S	4	5	6	P
P	S	7	8	9
10	P	S	11	12
13	14	P	S	15

Figure 2-4 Data is copied back

After copied back, it is back to RAID 5EE normal state.

2.2. Example of RAID 60EE with 2 RAID EE spares

Take another example of a RAID 60EE with 10 disks. 8 disks are for RAID disks, and 2 disks are for RAID EE spares. After initialization, data block distribution is as follows.

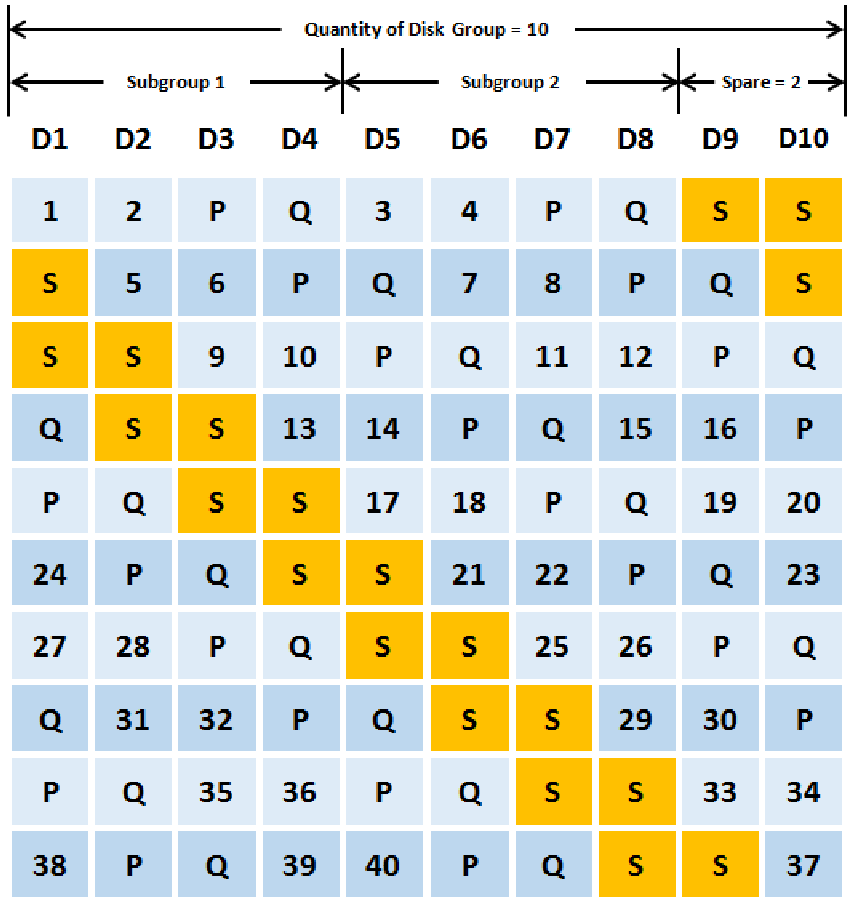

Figure 2-5 Data Block Distributed of RAID 60EE

Rebuild and copy back of RAID 60EE is similar as the above; it will not be repeated here.

2.3. RAID EE Level Summary

The following is the summary of the RAID EE levels.

Table 2-1 RAID EE Level Summary

	RAID 5EE	RAID 6EE	RAID 50EE	RAID 60EE
Min. # Drives	4	5	7	9
Fault Tolerance **(G = subgroups, S = RAID EE spares = 1 ~ 8)**	2 ~ 9 drive failures (e.g., 1 (RAID 5) + S spares)	3 ~ 10 drive failures (e.g., 2 (RAID 6) + S spares)	G+1 ~ G+8 drive failure (e.g., 2 subgroups (RAID 50) + S spares)	2xG+1 ~ 2xG+8 drive failures (e.g., 2x2 subgroups (RAID 60) + S spares)
Read Performance	Very Good	Very Good	Very Good	Very Good
Write Performance	Good	Fair to Good	Good	Fair to Good
Capacity **(N = drive quantity, M = drive capacity, G = subgroups, S = RAID EE spares)**	(N-1-S) x M (e.g., (10 drives – 1 – 2 spares) x 1TB = 7TB)	(N-2-S) x M (e.g., (10 drives - 2 - 2 spares) x 1TB = 6TB)	(N-G-S) x M (e.g., (10 drives - 2 subgroups - 2 spares) x 1TB = 6TB)	(N-2xG-S) x M (e.g., (10 drives - 2x2 subgroups - 2 spares) x 1TB = 4TB)
Capacity Utilization **(Min. ~ 26 drives)**	18% ~ 92% (e.g., 7/10 = 70%)	17% ~ 88% (e.g., 6/10 = 60%)	29% ~ 88% (e.g., 6/10 = 60%)	25% ~ 80% (e.g., 4/10 = 40%)
Typical Applications	Data warehouse, Web service, Archive	Data archive, High Availability solution, Server with large capacity requirement	Large database, File server, Application server	Data archive, High Availability solution, Server with large capacity requirement

3. TEST RESULTS

3.1. Case 1: RAID 5 vs. RAID 5EE

This test provides the comparison of rebuild time and copyback time between RAID 5 and RAID 5EE. We assume that the more RAID EE spare disks will have less rebuild time.

Test Equipments and Configurations

- Server
  - Model: ASUS RS700 X7/PS4 (CPU: Intel Xeon E5-2600 v2 / RAM: 8 GB)
    iSCSI HBA: Intel 82574L Gigabit Network Connection
    OS: Windows Server 2012 R2
- Storage
  - Model: XS5224D
    Memory: 16 GB (2 x 8 GB in bank 1 & 3) per controller
    Firmware 1.4.1
    HDD: 24 x Seagate Constellation ES, ST500NM0001, 500 GB, SAS 6 Gb/s
  - HDD Pool:
    RAID 5 Pool with 16 x NL-SAS HDDs in Controller 1
    RAID 5EE Pool with 17 (16+1 x RAID EE spare) x NL-SAS HDDs in Controller 1
    RAID 5EE Pool with 18 (16+2 x RAID EE spares) x NL-SAS HDDs in Controller 1
    RAID 5EE Pool with 20 (16+4 x RAID EE spares) x NL-SAS HDDs in Controller 1
    RAID 5EE Pool with 24 (16+8 x RAID EE spares) x NL-SAS HDDs in Controller 1
  - HDD Volume: 1 TB in Pool
- I/O Pattern
  - Tool: IOmeter V1.1.0
  - Workers: 1
  - Outstanding (Queue Depth): 128
  - Access Specifications:
    Backup Pattern (Sequential Read / Write, 256 KB)
    Database Access Pattern (as defined by Intel / StorageReview.com, 8 KB, 67% Read, 100% Random)

File Server Access Pattern (as defined by Intel, refer to the Figure 3-1, http://www.storagereview.com/articles/200003/20000313OSandBM_5.html)
Idle

Access Patterns			
% of Access Specification	*Transfer Size Request*	*% Reads*	*% Random*
File Server Access Pattern (as defined by Intel)			
10%	*0.5 KB*	*80%*	*100%*
5%	*1 KB*	*80%*	*100%*
5%	*2 KB*	*80%*	*100%*
60%	*4 KB*	*80%*	*100%*
2%	*8 KB*	*80%*	*100%*
4%	*16 KB*	*80%*	*100%*
4%	*32 KB*	*80%*	*100%*
10%	*64 KB*	*80%*	*100%*
Workstation Access Pattern (as defined by StorageReview.com)			
100%	*8 KB*	*80%*	*80%*
Database Access Pattern (as defined by Intel/StorageReview.com)			
100%	*8 KB*	*67%*	*100%*

Figure 3-1 Access Patterns by StorageReview.com

- Test Scenario
  - First we create a RAID 5 pool. After initialization, plug out and then plug in one disk drive. Count the rebuild time with different I/O access patterns.
  - Continue to create RAID 5EE with 1 / 2 / 4 /8 x RAID EE spare disks in sequence. After initialization, plug out one disk drive. The RAID EE starts rebuilding. Count the rebuild time with different I/O access patterns. Then plug in one disk drive and set it as dedicated spare, it starts copying back. Last, count the copyback time.

Test Result

Table 3-1 The Test Result of RAID 5 and RAID 5EE

	RAID 5 (x16)	RAID 5EE (x16+1)	RAID 5EE (x16+2)	RAID 5EE (x16+4)	RAID 5EE (x16+8)
Sequential Read, 256 KB	40'43"	11'39"	10'09"	6'46"	4'58"
Improved		71%	75%	83%	88%
Copyback		34'08"	30'57"	28'30"	30'56"
Sequential Write, 256 KB	24'54"	10'37"	9'08"	6'12"	4'02"
Improved		57%	63%	75%	84%
Copyback		22'29"	30'23"	28'50"	31'54"
Database Access Pattern	507'33"	62'23"	53'25"	34'37"	19'50"
Improved		88%	89%	93%	96%
Copyback		1320'37"	1082'21"	829'00"	754'35"
File Server Access Pattern	431'18"	58'24"	45'54"	27"19"	25'42"
Improved		86%	89%	94%	94%
Copyback		1363'17"	1093'20"	736'87"	705'50"
Idle	11'20"	8'24"	7'20"	5'02"	3'22"
Improved		26%	35%	56%	70%
Copyback		21'36"	22'26"	24'31"	26'05"

Take an example, the rebuild time of RAID 5 with sequential read 256 KB is 40 minutes and 43 seconds. Compare to the RAID 5EE with 1 x RAID EE spare disks, the rebuild time is 11 minute and 39 seconds. It improves ((40 x 60 +43) - (11 x 60 + 39)) / (40 x 60 +43) = (2443 – 699) / 2443 = 0.7138 == 71%.

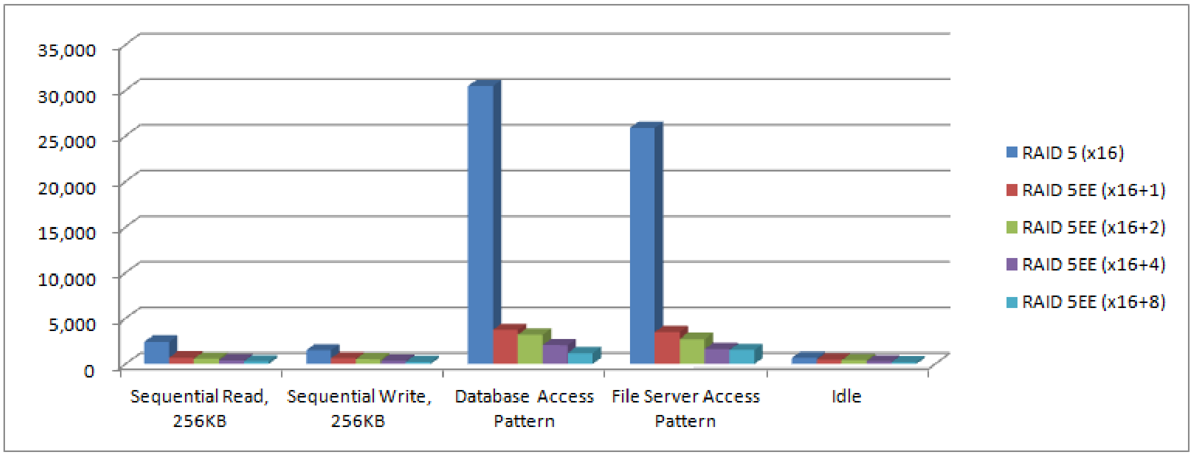

Figure 3-2 The Chart of RAID 5 and RAID 5EE

Summary

- RAID EE can improve rebuild time by up to 96%.
- The more RAID EE spare disks are used, the less rebuild time is.
- Rebuild time is more effective when there are reading accesses.
- If the access pattern is random, the copyback time is longer.
- When the data is copying back, the system resources will be reserved preferentially to the front end I/O.

3.2. Case 2: RAID 60 vs. RAID 60EE

This test provides the comparison of rebuild time and copyback time between RAID 60 and RAID 60EE. The same, we assume that the more RAID EE spare disks will have less rebuild time and RAID 60EE will have better efficiency.

Test Equipments and Configurations

- Server
  - Model: ASUS RS700 X7/PS4 (CPU: Intel Xeon E5-2600 v2 / RAM: 8 GB)
    iSCSI HBA: Intel 82574L Gigabit Network Connection
    OS: Windows Server 2012 R2
- Storage
  - Model: XS5224D
    Memory: 16 GB (2 x 8 GB in bank 1 & 3) per controller
    Firmware 1.4.1
    HDD: 24 x Seagate Constellation ES, ST500NM0001, 500 GB, SAS 6 Gb/s
  - HDD Pool:
    RAID 60 Pool with 16 x NL-SAS HDDs in Controller 1
    RAID 60EE Pool with 17 (16+1 x RAID EE spare) x NL-SAS HDDs in Controller 1
    RAID 60EE Pool with 18 (16+2 x RAID EE spares) x NL-SAS HDDs in Controller 1
    RAID 60EE Pool with 20 (16+4 x RAID EE spares) x NL-SAS HDDs in Controller 1
    RAID 60EE Pool with 24 (16+8 x RAID EE spares) x NL-SAS HDDs in Controller 1
  - HDD Volume: 1 TB in Pool
- I/O Pattern
  - Tool: IOmeter V1.1.0
  - Workers: 1
  - Outstanding (Queue Depth): 128
  - Access Specifications:
    Backup Pattern (Sequential Read / Write, 256 KB)
    Database Access Pattern (as defined by Intel / StorageReview.com, 8 KB, 67% Read, 100% Random)
    File Server Access Pattern (as defined by Intel, refer to the Figure 3-1, http://www.storagereview.com/articles/200003/20000313OSandBM_5.html)
    Idle
- Test Scenario
  - First we create a RAID 60 pool. After initialization, plug out and then plug in one disk drive. Count the rebuild time with different I/O access patterns.
  - Continue to create RAID 60EE with 1 / 2 / 4 /8 x RAID EE spare disks in sequence. After initialization, plug out one disk drive. The RAID EE starts rebuilding. Count the rebuild time with different I/O access patterns. Then plug in one disk drive and set it as dedicated spare, it starts copying back. Last, count the copyback time.

Test Result

Table 3-2 The Test Result of RAID 60 and RAID 60EE

	RAID 60 (x16)	RAID 60EE (x16+1)	RAID 60EE (x16+2)	RAID 60EE (x16+4)	RAID 60EE (x16+8)
Sequential Read, 256 KB	24'58"	13'04"	11'37"	9'23"	4'02"
Improved		48%	53%	62%	84%
Copyback		46'41"	41'38"	34'01"	35'43"
Sequential Write, 256 KB	20'16"	13'59"	12'28"	9'40"	4'26"
Improved		31%	38%	52%	78%
Copyback		33'32"	44'33"	31'54"	35'04"
Database Access Pattern	623'22"	93'24"	58'17"	35'34"	16'52"
Improved		85%	91%	94%	97%
Copyback		843'11"	876'59"	492'25"	383'16"
File Server Access Pattern	458'03"	63'36"	55'23"	34'24"	15'51"
Improved		86%	88%	92%	97%
Copyback		1215'34"	1087'47"	673'21"	478'57"
Idle	13'47"	10'49"	10'50"	7'24"	3'30"
Improved		22%	21%	46%	75%
Copyback		26'46"	27'32"	29'30"	33'22"

Take an example, the rebuild time of RAID 60 with sequential read 256 KB is 24 minutes and 58 seconds. Compare to the RAID 60EE with 1 x RAID EE spare disks, the rebuild time is 13 minute and 4 seconds. It improves ((24 x 60 +58) - (13 x 60 + 4)) / (24 x 60 +58) = (1498 – 784) / 1498 = 0.4766 == 48%.

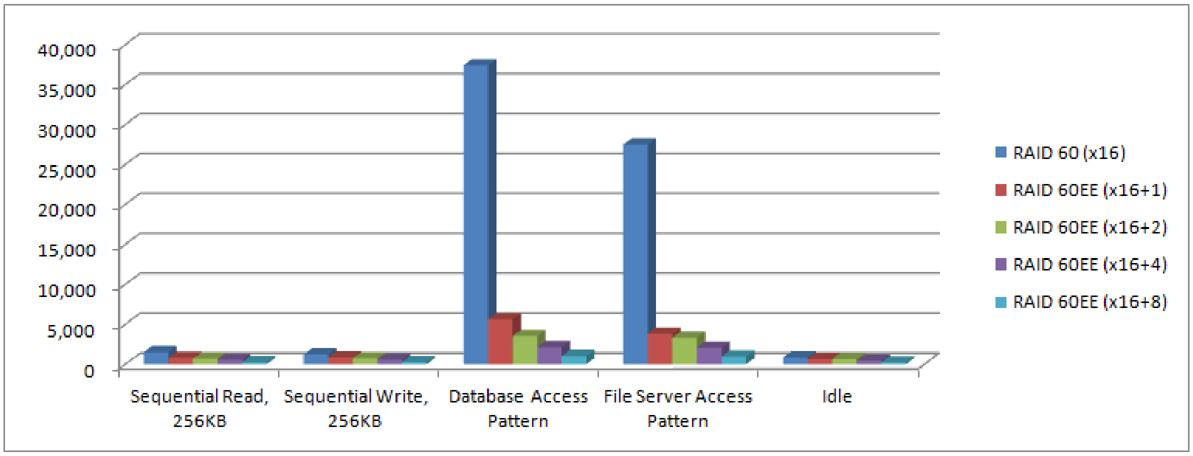

Figure 3-3 The Chart of RAID 60 and RAID 60EE

Summary

- RAID EE can improve rebuild time by up to 97%.
- The more RAID EE spare disks are used, the less rebuild time is.
- Rebuild time is more effective when there are reading accesses.
- If the access pattern is random, the copyback time is longer.
- When the data is copying back, the system resources will be reserved preferentially to the front end I/O.

4. CONCLUSION

As drive capacity grows, RAID rebuild time grows linearly. The more disk drives contain in a disk group, the more possibility of cumulative failure increase, so does the increasing impact of the disk drive capacity on the rebuild speed. Using RAID EE technology will greatly reduce these risks.

5. APPENDIX

5.1. Apply To

- XEVO firmware 1.0.0 and later
- SANOS firmware 1.4.1 and later

5.2. Reference

- XEVO Software Manual
- RAID EE Tutorial in XEVO
- SANOS Software Manual
- RAID EE Tutorial in SANOS