RAID 5 rebuild performance in ProLiant

technology brief

Abstract

This technology brief provides an overview of factors that affect RAID 5 array rebuild performance. Factors discussed include array size, RAID stripe size, rebuilding priority settings, and concurrent host I/O requests. Performance testing results are presented to demonstrate the relative performance impact of different configurations.

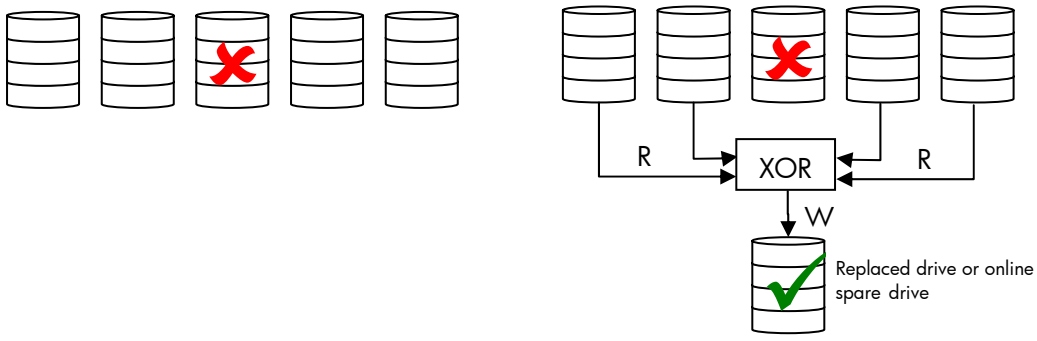
Overview of the RAID 5 rebuild process

In a redundant array of independent disks (RAID) configuration, data are stored in arrays of drives to provide fault tolerance and improved data access performance. In a RAID 5 array configuration, the user data and parity data (encoded redundant information) are distributed across all the drives in the array (data striping). By striping the user data and distributing the parity data across all drives in the array, optimum performance is achieved by preventing the slowdown (bottleneck) caused by constant hits on a single drive.

If a drive fails in a RAID 5 array configuration, the data can be reconstructed (or rebuilt) from the parity data on the remaining drives. If the array is configured with an online spare drive, the automatic data recovery process (or rebuild process) begins immediately when a failed drive is detected; otherwise, the rebuild process begins when the failed drive is replaced.

To rebuild lost data on a failed drive, each lost segment is read from the remaining drives in the array (where "N" is the total number of drives in the array, "N-1" is the remaining drives). The segment data is restored through exclusive-OR (XOR) operations that occur in the array controller XOR engine. After the XOR engine restores the lost segment, the restored segment data is written to the replacement or online spare drive. The rebuild process involves (N-1) reads (R) from the operational drives in the array and a single write (W) to the replacement or online spare drive (See Figure 1). When a segment is fully restored, the rebuild process proceeds to restore the next lost segment.

Figure 1. RAID 5 rebuild process



During the rebuild process, the array remains accessible to users; however, performance of data access is degraded. An array with a failed drive operates in "degraded mode." During the rebuild process, the array operates in "rebuild mode." If more than one drive fails at any given time, or any other drive fails during the rebuild process, the array becomes inaccessible.

Upon completion of the rebuild process, the rebuilt drive contains the data it would have contained had the original drive never failed. In configurations using an online spare drive, the status of the

online spare configuration is restored when the failed drive is replaced. After the failed drive is replaced, the content of the online spare drive will be copied to the replaced drive. After the completion of disk copy, the online spare configuration is restored.

Estimating the mean-time-to-failure (MTTF)

Mean-time-to-failure (MTTF) indicates the average time a drive will operate from start of use to failure. A higher MTTF value indicates that a device is less likely to fail.

Mean-time-to-repair (MTTR) indicates the total time (in hours) required to repair a failed drive in the array. To achieve high reliability, it is generally desirable to minimize MTTR.

The MTTF for RAID 5 array configurations can be estimated using the following equation derived by Patterson et al. (1988):

$$\frac{MTTF_{disk}^{2}}{N \times (G-1) \times MTTR_{disk}}$$

Calculation variables are defined as follows:

- $MTTF_{disk}$ —MTTF of a single drive
- N—total number of drives in the array
- G—parity group size
- $MTTR_{disk}$ —MTTR of a single drive

For further information on the probability of data loss, refer to the "RAID 6 with HP Advanced Data Guarding technology: a cost-effective, fault-tolerant solution" technology brief at: http://h200001.www2.hp.com/bc/docs/support/SupportManual/c00386950/c00386950.pdf.

Factors affecting RAID 5 array rebuild performance

The time required to rebuild a RAID 5 array is affected by the following factors:

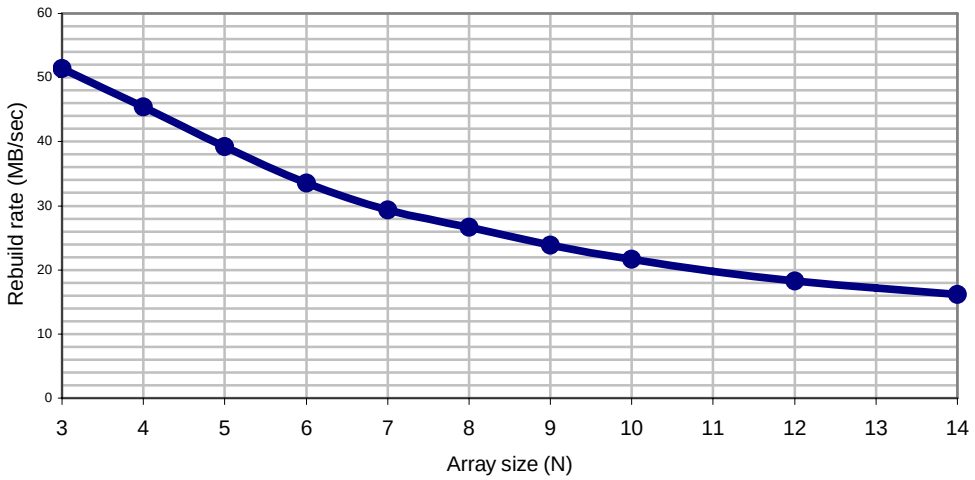
- Array size (total number of drives in the array [N])
- RAID stripe size
- Rebuild priority setting
- Drive capacity
- Concurrent host I/O activities during the rebuild process

The following sections examine the dependency of array size, RAID stripe size, rebuild priority settings, and concurrent host I/O requests on the RAID 5 rebuild performance. The affect of drive capacity and controller characteristics are not discussed; however, the rebuild rates reported can be used for estimating the rebuild times required for different drive capacities.

Array size

To restore each lost stripe, the RAID 5 rebuild process requires one read request from each of the operational drives (N-1 read requests), and one write request to the replacement drive. Therefore, there is approximately an N:1 inefficiency in the rebuild process[1]. Consequently, RAID 5 arrays with many drives will require longer rebuild times (see Figure 2).

Figure 2. Affect of array size on rebuild rate

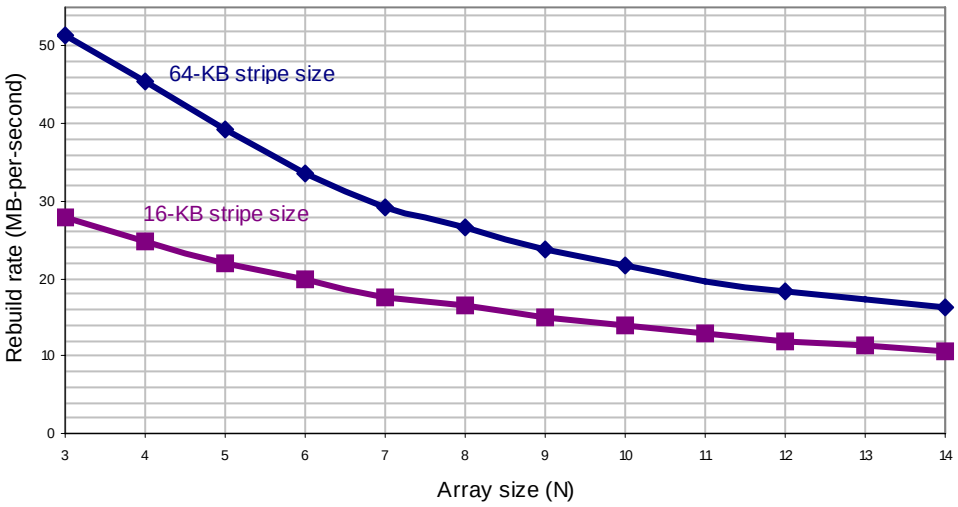


RAID stripe size

A significant factor affecting rebuild performance is the drive data block size (the stripe size). Because the RAID 5 rebuild process restores the failed drive one (or more) stripes at a time, the overall data rate of the read and write operations over the entire array depends on the efficiency of transporting stripe(s) of data over the SCSI bus. With approximately the same SCSI overhead, larger stripe sizes yield higher SCSI bus efficiency and faster data transfer rates. Therefore, performance of the RAID 5 rebuild process improves because the array controller can retrieve data from the operational drives and then restore the lost data to the replacement or online spare drive more efficiently (see Figure 3).

[1] The N:1 inefficiency is an approximation and is most apparent when the array size (N) is sufficiently large so that the data rate achieved by the N (a single write and N-1 reads) requests approaches the SCSI protocol limitation.

Figure 3. Affect of RAID stripe size on rebuild rate



RAID 5 rebuild performance

The rebuild process is significantly affected by host I/O activities. To balance rebuild and host I/O activity performance on an HP Smart Array RAID Controller, the rebuild process can be given a priority setting of high, medium, or low. The rebuild priority setting can be dynamically configured in the HP Array Configuration Utility (ACU).

When the rebuild priority setting is set to high, significant portions of system resources are devoted to the RAID 5 rebuild process. Servicing the rebuild process is the highest priority and servicing the host I/O activities becomes secondary.

When the rebuild priority setting is set to low, all system resources are devoted to serve host I/O activities. Minimal system resources are devoted to the RAID 5 rebuild process. The rebuild priority setting of low provides the best performance for serving host I/O activities during the rebuild process. Virtually no rebuild will take place as long as the host I/O activities persist. However, the rebuild process automatically proceeds at full speed when the host I/O activity is light (for example, during off-peak hours).

Rebuild rate for different rebuild priority settings

To understand the rebuild priority settings affect on the RAID 5 rebuild rate and the performance of host I/O activity, testing was completed for an online transaction processing (OLTP) environment, which provides concurrent host I/O requests during the RAID 5 rebuild process.

The following tables summarize the impact of rebuild priority settings on the rebuild rate for concurrent host I/O requests for the following modes:

- Normal mode—no failed drive
- Degraded mode—with a failed drive
- Rebuild mode—during the rebuild process

While operating in rebuild mode, there is a tradeoff between performance of concurrent host I/O activities and the rebuild rate. The performance of host I/O activities varies as the rebuild process progresses. The results are captured approximately one to two minutes after the rebuild process begins. Throughout the testing, Iometer issues simulated host I/O requests to the array with one outstanding I/O request at any given time.

Table 1 lists the affect of the rebuild priority setting on host I/O activity for a RAID 5 array containing six drives and 2-KB OLTP with one outstanding I/O request.

Table 1. Affect of rebuild priority setting on host I/O activity (I/O-per-second [IOPS]) and the rebuild rate

Rebuild priority setting	Normal mode (IOPS)	Degraded mode (IOPS)	Rebuild mode (IOPS)	Rebuild rate (MB/sec)
High	195	175	100	5.73
Medium	195	174	138	3.36
Low	195	173	170	Not applicable

Table 2 lists the affect of the rebuild priority setting on host I/O activity for a RAID 5 array containing six drives and 64-KB sequential read requests with one outstanding I/O request.

Table 2. Affect of rebuild priority setting on host I/O activity (MB/sec) and the rebuild rate

Rebuild priority setting	Normal mode (MB/sec)	Degraded mode (MB/sec)	Rebuild mode (MB/sec)	Rebuild rate (MB/sec)
High	105.13	42.18	6.26	21.83
Medium	107.06	42.39	18.47	12.58
Low	106.00	42.87	42.43	Not applicable

Rebuild rate with no host I/O activity

With no host I/O activity during the rebuild process, the rebuild rates are as follows (Testing was performed using an array of six drives. See "Testing configuration" section for further details.):

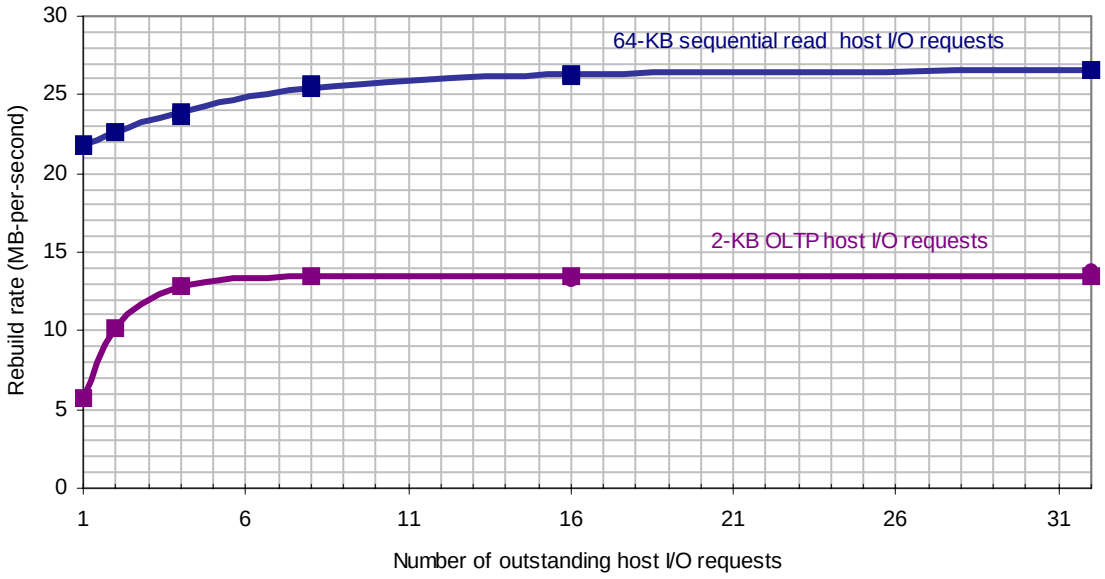
- High—33.07 megabytes-per-second [MB/sec]
- Medium—33.07 MB/sec
- Low—31.29 MB/sec

With no concurrent host I/O activities, the rebuild process progresses at full speed. Within the measurement error, the rebuild rate is virtually independent of the rebuild priority setting.

Rebuild rate for concurrent host I/O requests

Controller/array resources are shared between host I/O activities and the rebuild process. At a given rebuild priority setting, the effective rebuild rate strongly depends on the characteristics of host I/O activities. For example, Figure 4 illustrates the dependency of measured rebuild rates on the concurrent host I/O activities. The rebuild priority setting is set to high.

Figure 4. Dependency of concurrent host I/O activity on the rebuild rate



Since resources are shared between serving host I/O requests and rebuilding the array, concurrent host I/O activity performance degrades when the array transitions from degraded mode to rebuild mode. The performance impact depends on the characteristics of the host I/O requests. Figures 5 and 6 illustrate the performance of concurrent host I/O activity when the array is operating in the different rebuild modes.

Figure 5. Performance of concurrent 2-KB OLTP host I/O requests for different operating modes

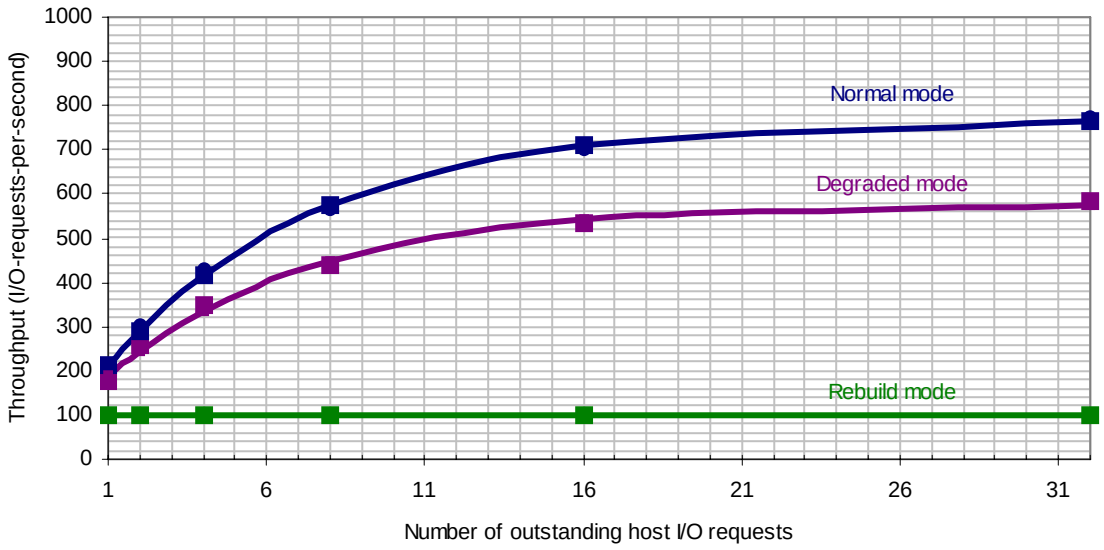
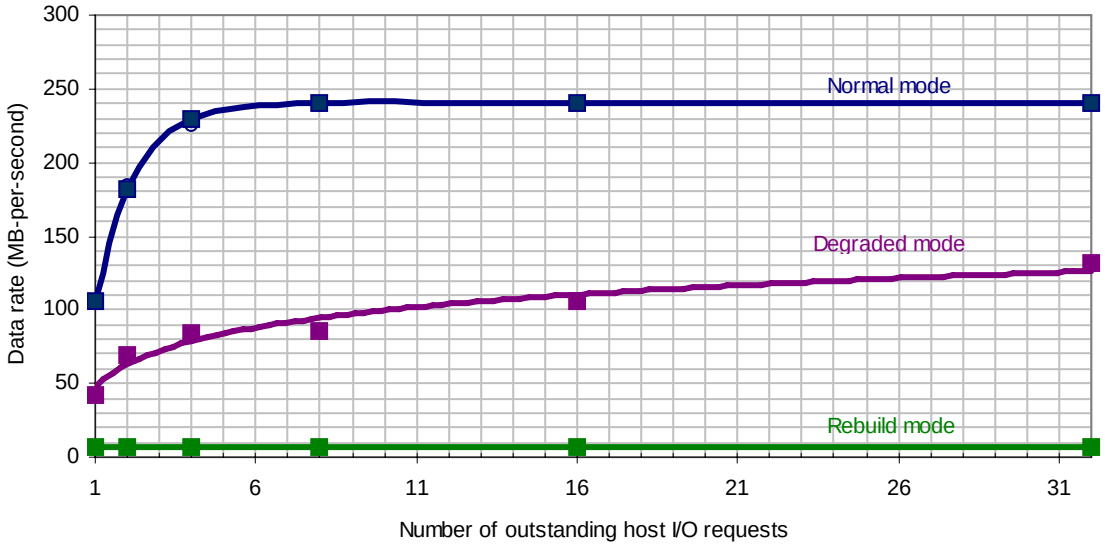


Figure 6. Performance levels of concurrent 64-KB sequential read host I/O requests for different operating modes



Testing configurations

Hardware configuration

The testing configuration used for testing the RAID 5 rebuild process is as follows:

- HP ProLiant DL380 Generation 3 (G3) server configured as follows:
  - System ROM P29 (07/25/2003)
  - 3.06-GHz Intel Xeon Processor with 512 KB of Level 2 cache memory and 1 MB of Level 3 cache memory
  - 1024 MB of system memory
- Microsoft Windows Server 2003, Enterprise Edition
- Array Configuration Utility Version 7.15.17.0
- HP Smart Array 6402 Controller configured as follows:
  - Firmware Version 1.92
  - 320 MB of cache memory
- HP StorageWorks Modular Smart Array 30 Storage Enclosure
- 36.4-GB Ultra320 low-voltage-differential (LVD) hard drives with firmware Version HPB6

Software configuration

To emulate the different environments, Iometer was used to generate the controlled host I/O activity.

Iometer test parameters for 2 KB OLTP host activity were configured as follows:

- I/O request size: 2 KB
- I/O request type: 67 percent Read requests and 33 percent Write requests
- Randomness: 100 percent random over the whole array capacity
- Outstanding I/O count: 1, 2, 4, 8, 16, 32

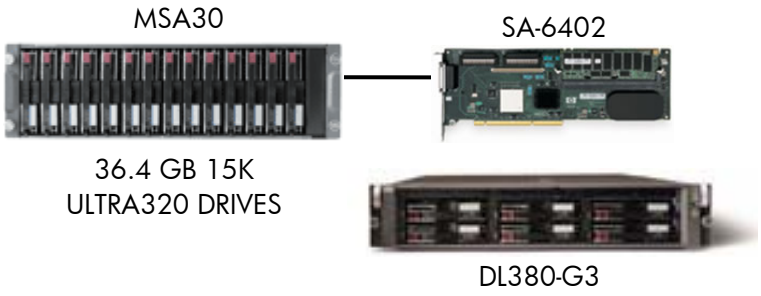
Iometer test parameters for 64 KB sequential read host activity were configured as follows:

- I/O request size: 64 KB
- I/O request type: 100 percent Read requests
- Randomness: 100 percent sequential starting from the beginning of the array
- Outstanding I/O count: 1, 2, 4, 8, 16, 32

Drive failure emulation process

To emulate a drive failure, the drive in the second bay was removed from the HP StorageWorks Modular Smart Array 30 Storage Enclosure. A replacement drive is inserted to the same drive bay to replace the failed drive. The insertion of the replacement drive triggers the rebuild process to begin. The time stamps recorded in the Microsoft Windows Server 2003 Event Log were examined to calculate the total RAID 5 rebuild time. Figure 7 illustrates the testing configuration.

Figure 7. Testing configuration



For more information

For more information on the ACU, refer to the website http://h18000.www1.hp.com/products/servers/proliantstorage/software-management/acumatrix/index.html.

For more information on Iometer, refer to the website www.sourceforge.net or www.iometer.org.

Call to action

To help us better understand and meet your needs for ISS technology information, please send comments about this paper to: TechCom@HP.com.

TC050702TB, 07/2005

Printed in the US