Technology Paper

Reducing RAID Recovery Downtime

Introduction

Depending on the specific RAID level chosen, an array of multiple disk drives can be configured to achieve maximum data reliability, maximum I/O performance or a mix of those two objectives.

Each RAID level employs a unique combination of technologies (mirroring, bit-, byte- or block-level striping, dedicated or distributed parity) to achieve these goals, but a common thread unites these RAID solutions: the need for lengthy, error-prone RAID recovery when drive failures occur.

Recent years have only increased the need for a solution, as fantastic growth in disk drive capacities has far outpaced improvements in storage system throughput, resulting in even more time-consuming RAID recovery. It can take hours—and often days—to recover an enterprise RAID set; more worrisome, if a secondary drive failure occurs during the long recovery process, the RAID's data will be deemed unreadable and the IT manager must retreat to a backup version of this data.

Conventional RAID recoveries typically use the parity data on the remaining active drives to recover the data on the failed drive and write it to a hot spare, an inherently slow and complex process. Seagate now delivers a better solution with the Seagate RAID Rebuild™ technology's partial failure copy feature, which helps the host rapidly extract as much data from a failed drive as possible before resorting to a RAID recovery; with far less data to recreate, this type of RAID recovery is much faster and less error-prone.

Benefits of Seagate RAID Rebuild™ Technology

Based on extensive research, Seagate has determined that the most important metrics for a RAID recovery are:

- Risk/exposure to secondary failure during RAID recovery
- System-level performance degradation during RAID recovery
- Time from start to finish of RAID recovery

Reducing RAID Recovery Downtime

The RAID recovery process places a great deal of additional stress on disk drives due to the prolonged read-write activity that recoveries entail. Considering the longer recovery times required due to the quantities of data stored on today's enterprise drives, and it is no surprise that secondary disk failures during RAID recovery are an increasingly common concern.

Furthermore, these lengthy recovery activities degrade system-level performance and thus delay user access to data on the RAID's remaining active drives. Giving the recovery process a higher priority than the host I/O in the RAID array can shorten recovery times (and reduce the chance for secondary drive failure), but will further diminish system-level performance.

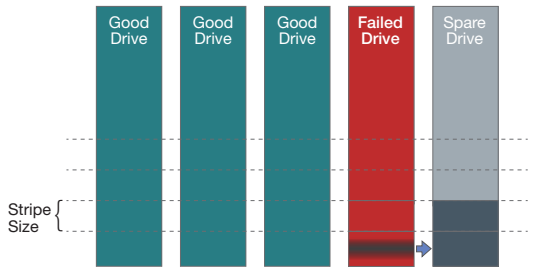
The partial failure copy functionality of disk drives equipped with Seagate RAID Rebuild technology directly addresses each of the above challenges. By enabling a failed drive to actively assist the host in retrieving the maximum amount of data possible before a RAID recovery is commenced, Seagate RAID Rebuild technology ensures:

- Faster, less error-prone RAID recovery
- Reduced impact on system performance
- Rapid access to recovered data

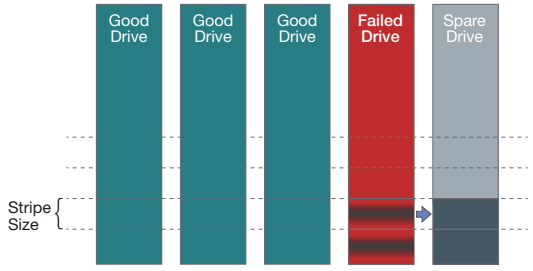
Seagate RAID Rebuild™ Technology: How It Works

Formatting PI Drives

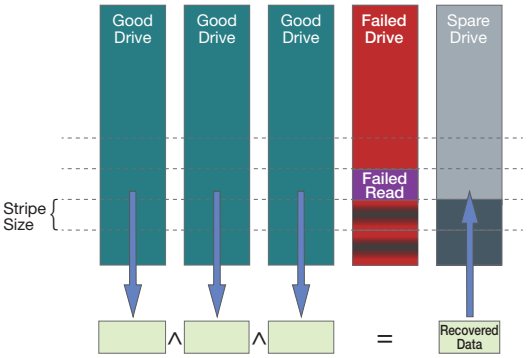
1. Attempt to read from failed drive first
2. First stripe is recovered by copying from failed drive



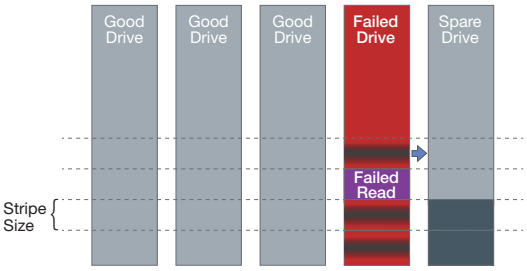
3. Second stripe is recovered by copying from failed drive.



4. Third stripe encounters error from failed drive so data is recovered using good drives.



5. Fourth stripe is recovered by copying from failed drive.



Reducing RAID Recovery Downtime

SAS Drives Equipped With Seagate RAID Rebuild™ Technology

The host will request that the failed drive characterize itself by issuing a Send Diagnostic command. If the drive is not spinning, it will attempt to spin up on whatever good heads exist. If the drive fails to spin up, it will return a HARDWARE ERROR and the drive's data cannot be extracted via the Seagate RAID Rebuild partial copy function.

If the drive successfully spins up, it will prepare itself for the partial copy function by eliminating any unnecessary background activity, determining if it contains any unusable heads, write-protecting the media and entering a special mode that includes limiting error recovery to the free retries. This mode will stay active until the drive is power-cycled and will persist through any bus resets.

The host should then issue a sequential read workload to extract the available data. During these sequential reads, the drive may choose to add to the list of unusable heads based on the rate of retries. When a command includes a head that is contained in the unusable head list, the drive will return sense data that identifies where the failing LBA is and where the next acceptable LBA is; the host should restart the sequential read work at the next acceptable LBA. (If the host is issuing queued read commands, then several of the commands may fail and point to the same next acceptable LBA.)

The host is responsible for maintaining the list of LBAs that still need to be rebuilt, that is, those LBAs that could not be read from the drive.[1]

SATA Drives Equipped With Seagate RAID Rebuild™ Technology

The host will issue a S.M.A.R.T. Offline Immediate command to the failed drive. (If the drive fails to spin up, the drive's data cannot be extracted via the Seagate RAID Rebuild partial copy function.) Upon this command, the drive will eliminate any unnecessary background activity, determine if it contains any unusable heads, write-protect the media, and enter a special mode that includes limiting error recovery to the free retries. This mode will stay active until the drive is power-cycled.

The host should then issue a sequential read workload to extract the available data using the Read FPDMA Queued command. During these sequential reads, the drive may choose to add to the list of unusable heads based on the rate of retries. When a command includes a head that is contained in the unusable head list, the drive will return the corresponding status value and error value, and the error LBA and the next available LBA must then be read by the host from log page 0×10 in order to continue. (In the SATA protocol, after an NCQ error the device will not accept any new commands until this log is read by the host.) The host should then restart the sequential read work at the next acceptable LBA.

The host is responsible for maintaining the list of LBAs that still need to be rebuilt; that is, those LBAs that could not be read from the drive.[2]

1 With SAS drives, the Force Unit Access (FUA) bit may be used on Read commands to re-enable full error recovery and to disregard the list of failed heads on a per-command basis.

2 With SATA drives, the Force Unit Access (FUA) bit may be used on Read FPDMA Queued commands to re-enable full error recovery and to disregard the list of failed heads on a per-command basis.

Reducing RAID Recovery Downtime

Conclusion

The value of disk drives equipped with the Seagate RAID Rebuild™ partial failure copy feature is clear and compelling: The more data that can be directly extracted from a failed drive, the less data that must be recovered during a time-consuming and error-prone RAID recovery.

The rapid recovery enabled by Seagate RAID Rebuild technology minimize risk/exposure to secondary drive failure, thus protecting data integrity in the RAID environment. This exclusive Seagate technology reduces system-level performance degradation and ensures that the failed drive's data can be quickly brought back online.

Company Affiliations

As a contributing member to leading industry standards bodies, Seagate has submitted an open-standards proposal of the RAID Rebuild™ functionality under the name *Rebuild Assist* to the T10[3] (SAS Proposal: 11-298) and SATA-IO[4] (SATA Proposal: SATA31_TPR_D144) committees for inclusion in their published standards specification.

3 T10 is a Technical Committee of the InterNational Committee on Information Technology Standards (INCITS) and is accredited by, and operates under rules that are approved by, the American National Standards Institute (ANSI). These rules are designed to ensure that voluntary standards are developed by the consensus of industry groups. INCITS develops Information Processing System standards, while ANSI approves the process under which they are developed and publishes them. ANSI also serves as the representative for the United States on Joint Technical Committee – 1 (JTC-1) of the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC). For more information, go to http://www.t10.org/

4 The Serial ATA International Organization (SATA-IO) is an independent, non-profit organization developed by and for leading industry companies. The SATA-IO provides the industry with guidance and support for implementing the SATA specification. The standardized SATA specification replaces a 15-year-old technology with a high-speed serial bus supporting up to 10 years of expected future. Members of the SATA-IO have the ability to influence or directly contribute to the development of the SATA specifications. For more information, go to http://www.sata-io.org/

www.seagate.com

RAID Rebuild™

AMERICAS Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, United States, 408-658-1000
ASIA/PACIFIC Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Singapore 569877, 65-6485-3888
EUROPE, MIDDLE EAST AND AFRICA Seagate Technology SAS 16–18, rue du Dôme, 92100 Boulogne-Billancourt, France, 33 1-4186 10 00