An Oracle White Paper
May 2016

Mirroring and Failure Groups with ASM

Introduction

ASM provides the means of mirroring data as a means of protecting against its loss as a result of a disk or other component failure. Unlike storage array RAID functionality, ASM mirroring only creates copies of blocks in existing files. Conversely, a storage array implementing RAID1 functionality mirrors all blocks regardless of whether the blocks are allocated to a file or even contain data. The redundancy level controls in ASM determine how many disk failures can be tolerated without forcibly dismounting the Diskgroup or losing data. The Diskgroup redundancy type specifies the mirroring level when the database creates new files. The redundancy types are;

- **External Redundancy**

External Redundancy tells ASM not to mirror files and it is expected that underlying storage hardware will provide all necessary redundancy and resiliency to component failures. In the event of a disk failure with External Redundancy, the ASM Diskgroup becomes unavailable until the disk is made available.

- **Normal Redundancy**

Normal Redundancy tells ASM to provide two-way mirroring for files in the Diskgroup. Normal Redundancy Diskgroups require at least two Failure Groups (explained below).

- **High Redundancy.**

High Redundancy tells ASM to provide three-way mirroring for files in the specified Diskgroup. High Redundancy Diskgroups require at least three Failure Groups.

The Case for ASM Mirroring

High-end storage arrays generally provide hardware RAID protection. Use Oracle ASM mirroring when not using hardware RAID. You can use ASM mirroring in configurations when mirroring between geographically-separated sites (extended clusters). Additionally, ASM mirroring is useful in situations where you want ASM to mirror across separate physical storage arrays. In this case, data in a Diskgroup will remain available even if one storage array is lost.

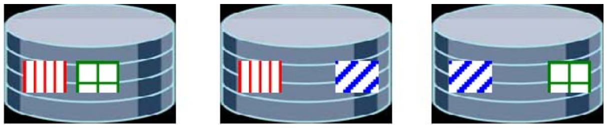
How ASM Mirrors Files

When Oracle ASM allocates an extent for a mirrored file, Oracle ASM allocates a primary copy and one or two mirror copies. Oracle ASM chooses the disk on which to store the mirror copy that is in a different Failure Group than the primary copy. Failure Groups are used to specify placement of mirrored copies so that each copy is on a disk in different Failure Groups. The simultaneous failure of all disks in a Failure Group does not result in data loss.

The customer defines the Failure Groups for a Diskgroup when they create an Oracle ASM Diskgroup. After a Diskgroup is created, you cannot alter the redundancy level of the Diskgroup. If you omit the Failure Group specification, then ASM automatically places each disk into its own Failure Group. Normal Redundancy Diskgroups require at least two Failure Groups. High Redundancy Diskgroups require at least three Failure Groups. Diskgroups with external redundancy do not use Failure Groups.

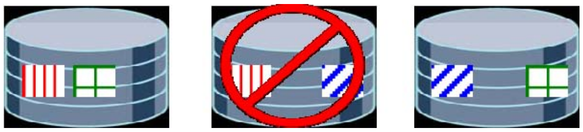
What is an ASM failure group?

ASM mirroring is done at the extent (typically 1MB) level and may be configured for two or three-way mirroring. When ASM allocates an extent for a Normal Redundancy Diskgroup it allocates a primary copy and a secondary copy. The disk for the secondary copy is chosen to be in a different Failure Group than the primary copy. Failure Groups are used to place mirror copies of data. Each copy is on a disk in a different Failure Group. Thus the simultaneous failure of all disks in a failure group will not result in data loss.



Three Failure Groups with two way mirroring

A Failure Group is a subset of the disks in a Diskgroup, which could fail at the same time because they share a common piece of hardware. The failure of that common piece of hardware must be tolerated. Four drives that are in a single removable tray of a large JBOD array should be in the same Failure Group because the tray could be removed making all four drives fail at the same time. Drives in the same cabinet could be in multiple Failure Groups if the cabinet has redundant power and cooling so that it is not necessary to protect against failure of the entire cabinet. ASM mirroring is not intended to protect against a fire in the computer room that destroys the entire cabinet.



Failure Group can fail with no data loss

What if I never create Failure Groups?

Every disk in a Diskgroup belongs to exactly one Failure Group. There are always Failure Groups even if they are not explicitly created. If the administrator does not specify a Failure Group for a disk, then a new Failure Group is automatically constructed containing just that disk. A Normal Redundancy Diskgroup must be partitioned into at least two Failure Groups. A High Redundancy Diskgroup must be partitioned into at least three Failure Groups. However it is much better to have several Failure Groups. A small number of Failure Groups, or Failure Groups of uneven capacity, can lead to allocation problems that prevent full utilization of all available storage.

Most systems do not need to explicitly define Failure Groups. The default behavior of putting every disk in its own Failure Group will work well for most customers. Failure Groups are only needed for large complex systems that need to protect against failures other than individual spindle failures.

A Normal or High Redundancy Diskgroup is composed of extent sets where all the extents in an extent set contain the same data. For Normal Redundancy there are two extents in an extent set. For High Redundancy there are three extents in an extent set. An extent set contains a primary extent and one or two secondary extent sets. The primary extents for a file are spread evenly across all the disks in the Diskgroup without considering their Failure groups. Once a primary extent is allocated on a disk a secondary extents is allocated on a disk in another Failure Group. For high redundancy another secondary extent is allocated on a disk in yet another Failure Group. Thus every copy of the data is in a different Failure Group.

When a block is written to a file, the write goes to all the extents in that block's extent set. When a block is read from a file, the primary extent is read unless it is unavailable. This ensures reads are evenly distributed across all available disks no matter how they are placed into failure groups.

How many Failure Groups should I create?

Choosing Failure Groups depends on the kinds of failures that need to be tolerated without loss of data availability. For small numbers of disks (<20) it is usually best to use the default Failure Group creation that puts every disk in its own Failure Group. This even makes sense for large numbers of disks when the main concern is spindle failure. If there is a need to protect against the simultaneous loss of multiple disk drives due to a single component failure, then Failure Group specification can be used. For example, a Diskgroup may be constructed from several small modular disk arrays. If the system needs to continue operation when an entire modular array fails, then a Failure Group should consist of all the disks in one module. If one module fails, all the data on that module will be relocated to other modules to restore redundancy. Disks should be placed in the same Failure Group if they depend on a common piece of hardware whose failure needs to be tolerated with no loss of availability.

Having a small number of large Failure Groups may actually reduce availability in some cases. For example, half the disks in a Diskgroup could be on one power supply while the other half are on a different power supply. If this is used to divide the Diskgroup into two failure groups then tripping the breaker on one power supply could drop half the disks in the Diskgroup. Reconstructing the dropped disks would require copying all the data from the surviving disks after power is restored. This can be done online but consumes a lot of I/O and leaves the disk group unprotected against a spindle failure during the copy. However if each disk were its own Failure Group, the Diskgroup would be dismounted when the breaker tripped. Resetting the breaker would allow the Diskgroup to be remounted and no data copying would be needed.

Having Failure Groups of different sizes can waste disk space. You may have room to allocate primary extents, but no space available for secondary extents. For example, suppose there is a Diskgroup with six disks and three failure groups. If two disks are each their own individual Failure Group and the other four are in one common Failure Group then there will be very unequal allocation. All the secondary extents from the big Failure Group can only be placed on two of the six disks. The disks in the individual Failure Groups will fill up with secondary extents and block additional allocation even though there is plenty of space left in the large Failure Group. This will also put an uneven write load on the two disks that are full since they contain more secondary extents that are only accessed for writes.

The unit of failure is still the individual disk even when there are multiple disks in a Failure Group. Failure of one disk in a Failure Group does not affect the other disks in that Failure Group. For example a Failure Group could consist of six disks connected to the same disk controller. If one of the six disks has a motor failure the other five can continue to operate. The bad disk will be dropped from the Diskgroup and the other five will stay in the disk group.

Once a disk has been assigned to a Failure Group it cannot be reassigned to another Failure Group. If it needs to be in another Failure Group then it can be dropped from the Diskgroup and then added back. Since the choice of a Failure Group depends on the hardware configuration, a disk would not need to be reassigned unless it is physically moved.

A Failure Group is always a subset of the disks within a single Diskgroup. Thus a Failure Group does not include disks from two different Diskgroups. However there could be disks in different Diskgroups that share the same hardware. It would be reasonable to use the same Failure Group name for these disks even though they are in different Diskgroups. This would give the impression of being in the same failure group even though that is not strictly the case.

What if I have two simultaneous failures of different Failure Groups?

This will usually force a dismount of the disk group. Any attempts to access the files in the Diskgroup will get I/O errors. If access to the Failure Groups can be restored with no data loss, then the Diskgroup can be mounted again without rebalancing. This happens if there is a failure of a piece of hardware used by multiple Failure Groups. If two disks in different Failure Groups fail, but not the entire Failure Groups, then there is a reasonable chance that the two disks do not mirror any data, and the Diskgroup will tolerate the failure.

Conclusion

ASM provides the means to protect the loss of data by mirroring the data across separate Failure Groups inside a Diskgroup. The choice of disks making up the Failure Groups is determined by the customer. That choice is determined by what components the customer wants to protect against a failure. In situations where data protection is provided by storage array hardware, then External Redundancy can be use in which case ASM will not mirror data.

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

ASM Mirroring and Failure Groups
May 2016
Author: Jim Williams

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

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.

Hardware and Software, Engineered to Work Together