Cluster File System

The CFS enables you to concurrently mount the same file system on multiple nodes and is an extension of the industry-standard VxFS. Unlike other file systems that send data through another node to the underlying storage, the CFS is a true SAN file system. All data traffic happens over the SAN, and only the metadata traverses the cluster interconnect.

The CFS uses a distributed locking mechanism called Global Lock Manager (GLM) to ensure all nodes have a consistent view of the file system. The GLM provides metadata and cache coherency across multiple nodes by coordinating access to file system metadata such as inodes and free lists. The role of the GLM is set on a per-file system basis to enable load balancing.

Cluster Volume Manager

14

In general terms, a volume is a unit of storage carved out of a physical disk device. The CVM presents a consistent volume state across an InfoScale cluster as nodes import and access volumes concurrently. It also enables all nodes in a cluster to access their underlying storage devices concurrently. The CVM transforms the read and write requests that CFS addresses to volume blocks into I/O commands that it issues to the underlying disks.

The primary difference between CVM and VxVM is that CVM allows disk groups to be imported on all the systems in the cluster concurrently, whereas VxVM only allows a disk group to be imported on a single node at a time.

All CVM instances in a cluster must always present the same view of disk group and volume configuration, even in the event of:

- Storage device failure—For example, if a disk is added to or removed from a mirrored volume, all CVM instances must effect the change and adjust their I/O algorithms at the same logical instant.
- Cluster node failure—If a cluster node fails while it is updating one or more mirrored volumes, CVM instances on the surviving nodes must become aware of the failure promptly, so they can cooperate to restore volume integrity.

The CVM always guarantees that all instances in a cluster have the same view of shared volumes, including their names, capacities, access paths, and "geometries." Most important, the CVM also manages volume states, including whether the volume is online, the number of operational mirrors, whether mirror resynchronization is in progress, and so forth. A volume's state may change if a device fails or a node fails or an administrative command is issued.

Kubernetes Container Storage Interface Plug-In

InfoScale can also provide software-defined storage for Kubernetes (K8s) environments. The InfoScale Container Storage Interface (CSI) plug-in allows you to use InfoScale volumes created and managed via CVM and CFS as persistent storage for your stateful containerized applications running in Kubernetes.

I/O Fencing

A condition known as "split brain" occurs when there is communication disruption between cluster nodes. This disruption can result in data corruption due to the fact that InfoScale (and other cluster software) cannot always distinguish between a system failure and an interconnect failure. The split-brain condition can also occur if a node within the cluster is so busy that it appears to be hung and pauses communication with the other cluster nodes. The split-brain condition can occur in all clustered storage implementations, including K8s. To mitigate and resolve the split-brain condition, InfoScale implements an I/O fencing system that guarantees data integrity by determining which nodes in the cluster should remain in the event of a communication disruption. When a disruption occurs, the node (or nodes) that has failed is ejected from the cluster and prevented from accessing the data disks.

Chapter 1

Recovery from hardware failure

Symantec's Veritas Volume Manager (VxVM) protects systems from disk and other hardware failures and helps you to recover from such events. This chapter describes recovery procedures and information to help you prevent loss of data or system access due to disk and other hardware failures.

If a volume has a disk I/O failure (for example, because the disk has an uncorrectable error), VxVM can detach the plex involved in the failure. I/O stops on that plex but continues on the remaining plexes of the volume.

If a disk fails completely, VxVM can detach the disk from its disk group. All plexes on the disk are disabled. If there are any unmirrored volumes on a disk when it is detached, those volumes are also disabled. 14

Note: Apparent disk failure may not be due to a fault in the physical disk media or the disk controller, but may instead be caused by a fault in an intermediate or ancillary component such as a cable, host bus adapter, or power supply.

The hot-relocation feature in VxVM automatically detects disk failures, and notifies the system administrator and other nominated users of the failures by electronic mail. Hot-relocation also attempts to use spare disks and free disk space to restore redundancy and to preserve access to mirrored and RAID-5 volumes. For more information, see the "Administering Hot-Relocation" chapter in the *Veritas Volume Manager Administrator's Guide*.

Recovery from failures of the boot (`root`) disk requires the use of the special procedures described in "Recovery from boot disk failure" on page 35.