whitepaper

An Architectural Overview of the ContinuStor Director

Introduction

The ContinuStor Director is a storage management system that provides a number of second-generation SAN features. These include:

- Storage virtualization
- Storage pools
- Mirroring
- Dynamic storage allocation
- Performance optimization
- Data synchronization
- Availability

This paper focuses on the architectural elements of the ContinuStor Director design. Implementation features vary by model.

The ContinuStor Director

The ContinuStor Director design consists of a storage management application supported by a compact real-time operating system that is hosted on a high-performance multiprocessor hardware platform. Both host server connectivity and storage connectivity may use either SCSI or Fibre Channel. Fibre Channel support includes point-to-point, Arbitrated Loop, and Switched Fabric connections. The ContinuStor Director does not require software drivers or modifications for either host servers or storage systems.

Heterogeneous Storage Environments

The ContinuStor Director supports heterogeneous storage systems from a variety of vendors. The storage systems may vary by:

- Size
- Channel connections (single, dual, multiple)
- Configuration (JBODs, arrays)
- Technology/protocol (SCSI, Fibre Channel)
- Access (single channel, dual channel, multiple channel)
- Operational access (active/passive, active/active).

The ContinuStor Director creates a single storage domain from multiple storage systems and provides a single interface for administration, growth, and access.

Storage Virtualization

Storage virtualization is a second-generation SAN feature provided by the ContinuStor Director. Storage virtualization removes the necessity to understand the physical characteristics of the storage in order to access it. Storage virtualization also creates logical connections to storage. The host server simply requests a LUN of a desired size.

Note: The terms volume, device, and LUN are used interchangeably.

To reach this goal, the ContinuStor Director removes the need to know the following physical characteristics of the storage it manages:

- Geometry (Cylinder, Head, Sector)
- Protocol (SCSI, Fibre Channel)
- Topology (RAID and striping levels)
- Operating parameters (size, speed)
- Location (bus, loop, fabric)

Storage virtualization provides a uniform interface to storage regardless of its physical access characteristics and removes the server's need to understand and manage the underlying protocols. The ContinuStor Director provides the ability to create storage objects, LUNs, that are universal across the storage domain.

Storage Pool

Storage pooling is the ability to manage dissimilar storage to provide a common means of storage access. Through storage virtualization, independent storage systems are collectively grouped into a single storage image from which LUNs of varying sizes are provided to the requesting host servers. Storage from this pool is classified as primary devices, mirrors, spares, and unassigned.

Primary devices are LUNs used by host servers for data access using normal I/O commands. *Mirrors* are LUNs that are identical copies of the primary devices they are assigned to. *Spares* are LUNs that may become mirrors to provide an additional level of availability. Unassigned storage may be placed into service in roles as primary devices, mirrors, or as spares.

Server access to the storage pool is controlled through standard SAN masking, zoning, and partitioning features. In addition, enhanced features such as sharing, write-protection, and mirror access are also provided by the ContinuStor Director.

Mirroring

The key design feature of the ContinuStor Director is the ability to provide mirroring on an individual LUN basis. Mirroring starts with defining a *mirror set* which consists of a primary LUN and one or more mirrors. Write operations to the primary device are automatically written by the ContinuStor Director to each mirror. The ContinuStor Director keeps the mirror set in exact synchronization - each mirror is logically identical to the primary device.

The mirror set provides uninterrupted access to the primary's data. In the event that the primary device becomes unavailable, a mirror is *promoted* to primary status and I/O operations continue as it would normally. Host write operations continue, but data is now written to the promoted mirror and read operations are serviced from the promoted mirror. The host is not aware that the original primary device has become unavailable.

Each device in the mirror set, whether primary or mirror, has an associated bitmap that tracks changes (missing writes) should it become unavailable and therefore, go out-of-synchronization. Thus, a primary device that is unavailable would have its bitmap track updates made to the promoted mirror that is replacing it. Similarly, an unavailable mirror has its bitmap track changes made to the still available primary.

When the unavailable device becomes available again, the bitmap (indicating missing updates) is used to read missing blocks from the primary device to re-synchronize it to the primary. This process is used to bring each device back into synchronization.

The ContinuStor Director provides further protection by using an available spare to act as a mirror to substitute for a promoted mirror.

For example, if a primary device goes offline, a mirror is promoted in its place and a spare replaces the promoted mirror. The bitmap for the now out-of-service primary is updated for each new write operation to the new primary (the promoted mirror). When the original primary returns to service, its bitmap is used to update it and then resumes the role of primary device, replacing the promoted mirror. The promoted mirror reverts back to mirror status, and finally, the spare returns to the spares pool.

Mirror Usage

In addition to providing an added level of protection against outages, the mirror set has several other features. Mirrors may be larger than the primary device and used to replace the primary device. This allows dynamic LUN expansion subject to the availability of the corresponding feature in the host operating system.

Mirrors may also be used to provide data replication for application testing or conversion by simply terminating the mirror and using it as an identical copy of the primary data.

Mirrors may be used as point-in-time copies of primary data. On either an ad-hoc or scheduled basis, mirrors may be suspended or terminated at various time intervals to provide alternative access to a copy of the primary data as of the time of suspension.

Lastly, suspended mirrors may be used to perform offline backup of primary data. This relieves the host system and the primary device of the burden of an online backup. It is also much faster due to data movement at device speeds rather than at file systems speed.

Remote Mirrors

Multiple ContinuStor Directors may be used to create a networked storage domain that provides remote mirrors at remote locations. The data links between ContinuStor Directors can be any combination of TCP/IP, Fibre Channel, or SCSI. Primary devices and associated mirrors may be defined across this networked storage domain without restriction.

Cross mirroring capability is available to systems where primary LUNs at each ContinuStor Director location have mirrors at other locations.

In local environments, a single point-to-point Fibre Channel connection provides bi-directional data flow. A Switched Fabric may also be used to connect two ContinuStor Directors.

Beyond local distances, Ethernet connections are used to connect multiple ContinuStor Directors.

Application requirements and distances will determine the bandwidth, latencies, and data synchronization modes required.

Multiple links between ContinuStor Directors provide additional bandwidth as well as additional protection against outages.

Cascaded Mirrors

The ContinuStor Director provides the ability to *cascade* mirrors. A cascaded mirror is a mirror of a mirror. Normally, a primary device with two mirrors will have two data streams, one for each mirror. For remote mirrors, this doubles the network bandwidth required. Using cascaded mirrors, the first remote mirror uses half the normal bandwidth resources for mirroring while the second remote mirror uses local channel resources to mirror from the first mirror.

This is particularly useful in remote mirroring applications where the bandwidth limitations prevent fast re-synchronization. In these applications, the first remote mirror is in continuous synchronization with the primary device while the second mirror, at the same remote location as the first mirror, can be in and out of synchronization with the first mirror. Should the second mirror require it, a full re-synchronization occurs at channel speeds rather than at network speeds.

Dynamic Storage Allocation

The ContinuStor Director provides dynamic storage allocation at the LUN level and at the array level.

Mirrors have only a single requirement: that they be the same size or larger than the primary device. The mirror set can have a mirror that is substantially larger than the primary, thus allowing a larger mirror to replace a smaller primary. This permits online volume expansion, subject to the host operating system's ability to do the same.

Dynamic storage allocation is also possible when storage systems are dynamically added to the ContinuStor Director storage domain. Storage assignments (LUN definitions) may be pre-defined in situations where the growth requirements are known and planned, or may be dynamically defined at the time of storage addition.

Both features allow host servers to expand storage requirements without the usual system interruptions.

Performance Optimization

The ContinuStor Director is designed to provide performance optimization at multiple levels.

Data caching in the ContinuStor Director allows reads (of cached data) to be processed without a corresponding access of the physical media. This also reduces the I/O command traffic. Asynchronous writes of sequential blocks may be combined for efficiency and higher throughput. Anticipatory prefetch algorithms enhance sequential read operations.

Through storage virtualization, devices are identified and managed by their respective Fibre Channel (WWD) and SCSI (serial number) ids. Devices may migrate across storage system boundaries, permitting *data migration* to higher bandwidth channels or systems and eliminating the need to re-configure the system for a location change. This feature also permits higher host server throughput as technology and storage configurations change without requiring a storage re-configuration on the host.

The ability to mix and match SCSI and Fibre Channel storage through the ContinuStor Director's storage virtualization capabilities extends the SAN into legacy storage. This allows SCSI-only host servers to access Fibre Channel storage and vice versa.

Multiple host channel support allows multiple I/O command streams from one or more host servers to one or more LUNs. Storage channel optimization includes load-balancing I/O traffic thru multiple channels to common LUNs.

Multiple ContinuStor Directors providing local and remote mirrors may have multiple links between the ContinuStor Director nodes for higher throughput as well as an added level of protection against line outages.

Scaling capabilities within the ContinuStor Director node uses multi-processor technology to provide higher bandwidth. Multiple ContinuStor Director nodes extends this scaling capability.

Data Synchronization and Protection

The ContinuStor Director provides several levels of data synchronization to optimize the balance between protection and performance.

Synchronous mode offers the highest level of protection for write operations where completion status is not returned until the data is written to the disk. *Asynchronous* mode write operations to return completion status when the data is in the cache. *Adaptive synchronous* mode allows a pre-defined number of asynchronous write operations, after which, become synchronous writes.

The ContinuStor Director provides the ability to convert asynchronous write operations to synchronous write operations for an added level of data protection.

Read errors from a device are automatically retried from another device in the mirror set. In addition, a *read recovery process* is initiated to repair the bad read block using the correct data from an alternate device.

Availability

Using the ContinuStor Director generally raises the availability level of the entire system. Storage virtualization, mirroring and the dynamic multipathing features of the ContinuStor Director are used to provide the redundancy for protection against contingencies and outages. Data synchronization and protection options provide data integrity enhancements.

Storage virtualization allows optimal placement of critical data within the ContinuStor Director storage domain. This process allows LUN movement from less available storage (such as a JBOD) to more reliable storage (such as RAID arrays), both initially and over time as the storage environment is refreshed.

Placement of primary, mirror, and spare devices across multiple independent storage systems raises availability and prevents loss of access should a single storage system fail.

Mirroring provides an additional level of availability and protection for access of critical data on the primary device. This allows uninterrupted and continued data access when the primary device is unavailable. Availability is also extended through the use of spares.

Point-in-time copies of critical data reduces the time to recover by eliminating the time to restore from slower, offline tape copies.

Dynamic multi-pathing provides alternate path access to data during channel contingencies. This capability starts at the host server level where operating system drivers and volume managers interleave I/O commands to all available channels for load balancing and reliability requirements. The ContinuStor Director continues to process these data access commands to all available storage channels. In the event of a path failure, the ContinuStor Director uses alternate paths to access data.

Data synchronization options allow dynamic adjustments to the levels of data protection desired. This ranges from using adaptive synchronous write option to synchronous write option and apply within the local storage environment and at the remote ContinuStor Director locations.

Read availability is enhanced through mirroring using a variety of options ranging from simple, round-robin read rotations, to channel-balancing read algorithms.

Summary

The ContinuStor Director's storage virtualization and mirroring capabilities provide the basis for designing systems to meet availability, access and data protection requirements of the enterprise data center. Data replication through mirroring provides access to data for protection, conversions, and migrations. Storage virtualization provides support for heterogeneous storage, extending the SAN to legacy equipment, and provides a single consistent administrative storage domain. In short, the ContinuStor Director provides the ability to plan the future for storage without performance penalties.

Solution Centre Limited
Vickers House
Priestley Road
Basingstoke
RG24 9NP

www.solutioncentre.co.uk
Tel: 01256 818600
Fax: 01256 819600
E-Mail: sales@solutioncentre.co.uk