

Implementing EMC Symmetrix VMAX in a Cloud Service Provider Environment

Applied Technology

Abstract

This white paper documents the usage of the EMC® Symmetrix® VMAX™ in a cloud service provider environment. It describes a number of VMAX capabilities and features as well as best practices that help facilitate the cloud model. Perspective is provided in describing EMC's work in configuring and deploying Symmetrix VMAX systems with the cloud service provider Terremark.

August 2010

Copyright © 2010 EMC Corporation. All rights reserved.

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED “AS IS.” EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

For the most up-to-date listing of EMC product names, see EMC Corporation Trademarks on EMC.com

All other trademarks used herein are the property of their respective owners.

Part Number h7335.1

Table of Contents

Executive summary	4
Introduction	4
Audience	4
Cloud computing and cloud service providers	4
Rapid, elastic, and on-demand resource allocation	5
Multi-tenancy	5
Lack of visibility into the application layer	6
Monitoring and chargeback	6
Terremark.....	7
Symmetrix VMAX deployment at Terremark	8
Using VMware vSphere with EMC Symmetrix VMAX.....	8
End-to-end visibility integrated into VMware with EMC Storage Viewer	10
Symmetrix VMAX connectivity	11
Symmetrix VMAX HBA port flags.....	12
Automating load balancing with PowerPath/VE.....	13
PowerPath/VE features.....	14
PowerPath/VE in vCenter Server.....	15
On-demand storage provisioning with Auto-provisioning.....	16
Storage groups	16
Port groups.....	17
Initiator groups	17
Cascaded initiator groups	18
Masking views.....	18
Dynamic LUN addressing	19
Rapid elasticity and resource pooling with Virtual Provisioning	19
Considerations for vSphere environments on thin devices	21
Automated Pool Rebalancing	22
Space reclamation	23
Considerations for space reclamation in VMware environments.....	23
Thin metavolumes.....	24
Delivering measured service with FAST	24
Enterprise Flash Drives.....	25
How FAST is configured	26
FAST algorithms	26
Device movement	27
Conclusion	28
References	28

Executive summary

The cloud computing model is changing the way IT resources are delivered into organizations. As an operational model, it has a range of benefits, including reductions in cost, flexibility in scaling, rapid deployment of new services, and improved efficiency. By aligning with these needs, cloud computing has generated great interest. A growing number of organizations are deploying or experimenting with this model either internally or externally via services such as Terremark's vCloud™ Express and Enterprise Cloud™.

By leveraging EMC technologies such as Virtual Provisioning™, Auto-provisioning Groups, Fully Automated Storage Tiering (FAST), and PowerPath®/VE in a VMware® vSphere™ environment, the EMC® Symmetrix® VMAX™ is well suited to enable the cloud service provider model, increasing flexibility and simplifying management while maintaining service level agreements (SLAs), performance, and availability.

This white paper expands on this list of VMAX capabilities as well as best practices for deployment in a cloud environment. In particular, the deployment of the VMAX in Terremark's cloud environment is cited as a reference.

Introduction

This white paper begins by introducing cloud service providers, identifying their common elements and challenges, and offering a number of Symmetrix VMAX features that help meet these needs. Next, Terremark's Enterprise Cloud is cited as a real example of VMAX deployment in a cloud service provider model. This section will include Terremark's requirements and the VMAX configuration deployed to meet these needs. The white paper closes by detailing a number of VMAX features and capabilities that enable environments operating in a cloud service provider model. These are delivered with best practices and recommendations.

Audience

This white paper is intended for storage architects and administrators who are responsible for designing, managing, and operating cloud service provider environments leveraging the Symmetrix VMAX. In addition, field personnel who are tasked with implementing the VMAX in a cloud environment may leverage this paper.

Cloud computing and cloud service providers

The cloud is an operational model for the enablement of convenient, on-demand network access to a shared pool of computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. There are a number of essential characteristics that help define the cloud.

- **On-demand self-service** – As resources are needed, end users are able to automatically provision these resources with little to no interaction with IT or the service provider.
- **Standardized network access** – Capabilities of the cloud are readily accessible via standard network mechanisms that are widely used, such as web browsers.
- **Location independent resource pooling** – Resources are abstracted from the underlying infrastructure and pooled logically. This is often implemented using virtualization technology. As such, end users are not tied to infrastructure and thus need no visibility into the physical locations of their resources as long as their service level agreements (SLAs) are upheld.
- **Rapid elasticity** – Clouds have the capability to rapidly and dynamically provision and release resources at any time. Customers are billed only for resources as they are needed. Afterwards, these resources are returned to the pool and made available once more for allocation.

-
- **Measured service** – Clouds can automatically control, fine-tune, and optimize the usage of resources. Key elements such as computing power, bandwidth, and storage capacity are monitored, controlled, and reported, providing transparency for both provider and consumer.

There are primarily two classes of clouds: private clouds and public clouds.

In a private cloud, infrastructure is deployed and operated exclusively for an organization or enterprise. This may be managed internal to the organization or by an independent service provider. Additionally, this infrastructure may exist on- or off-premise at a hosting facility, as is the case with managed service providers.

In contrast, a public cloud consists of infrastructure shared across many customers. In this model, the infrastructure is owned by the organization that is offering the service. This is the cloud model deployed by cloud service providers.

Whether public or private, the cloud model has a number of benefits that align with the needs of application end users as well as IT managers. These include:

- Cost reductions due to the consolidation of physical resources, which also leads to a reduction in physical and carbon footprint.
- Avoidance of capital expenditure by reducing the number of underutilized resources or shifting infrastructure out of the organization as exemplified in the cloud service provider model.
- Flexibility in the dynamic scaling of IT and the ability to roll out new services rapidly.
- Transparency of costs and chargeback.

While not part of the definition, the cloud model has largely been enabled by the proliferation of virtualization technology into almost all aspects of IT. With emphasis on rapidity, flexibility, and dynamic capabilities, the cloud is a logical progression in the virtualization paradigm. Purpose-built for the virtual data center, Symmetrix VMAX helps address many of the challenges facing cloud service providers.

Rapid, elastic, and on-demand resource allocation

Cloud service providers must be able to provision and release resources in a pool at any given time. Symmetrix VMAX facilitates rapid elasticity through a number of mechanisms and features.

Auto-provisioning Groups allow storage to be rapidly and simply provisioned to and deallocated from clusters of servers. Changes to groups are automatically propagated to existing masking views. Symmetrix Management Console (SMC) further simplifies the storage provisioning process via the use of the masking view wizard and storage group expansion template.

Virtual Provisioning allows administrators to create pools of shared storage from which thinly provisioned devices are presented. Physical capacity is provisioned and consumed in extents striped across all the underlying data devices in the pool as they are written to by the host. Virtual Provisioning allows for the oversubscription of actual capacity compared to host-presented storage. Data devices can be added or removed from the pool dynamically; Symmetrix VMAX can automatically rebalance the used data extents across the resulting data devices. In addition, Symmetrix VMAX can reclaim any extents containing all zeros.

Additionally, nondisruptive expansion of presented LUNs is supported by Symmetrix metavolumes and complemented by VMware Virtual Machine File System (VMFS) Volume Grow.

Multi-tenancy

Cloud service providers offer the use of virtual infrastructures as a service. Often, these reside on the same physical infrastructure. Multi-tenancy is thus a principal challenge for cloud service providers.

At the VMware level, virtual disk files of virtual machines for various customers may be stored on the same VMFS datastore. In the storage array, data from multiple file systems may be striped across the same physical spindles in a large pool of data devices. Data can be kept on dedicated spindles by creating a number of separate thin pools. Keeping separate thin pools also allows for differentiated service levels. For

example, performance-sensitive applications may be run on premium storage residing on a Flash-based thin pool, while test and development virtual disks reside on VMFS datastores on SATA drives.

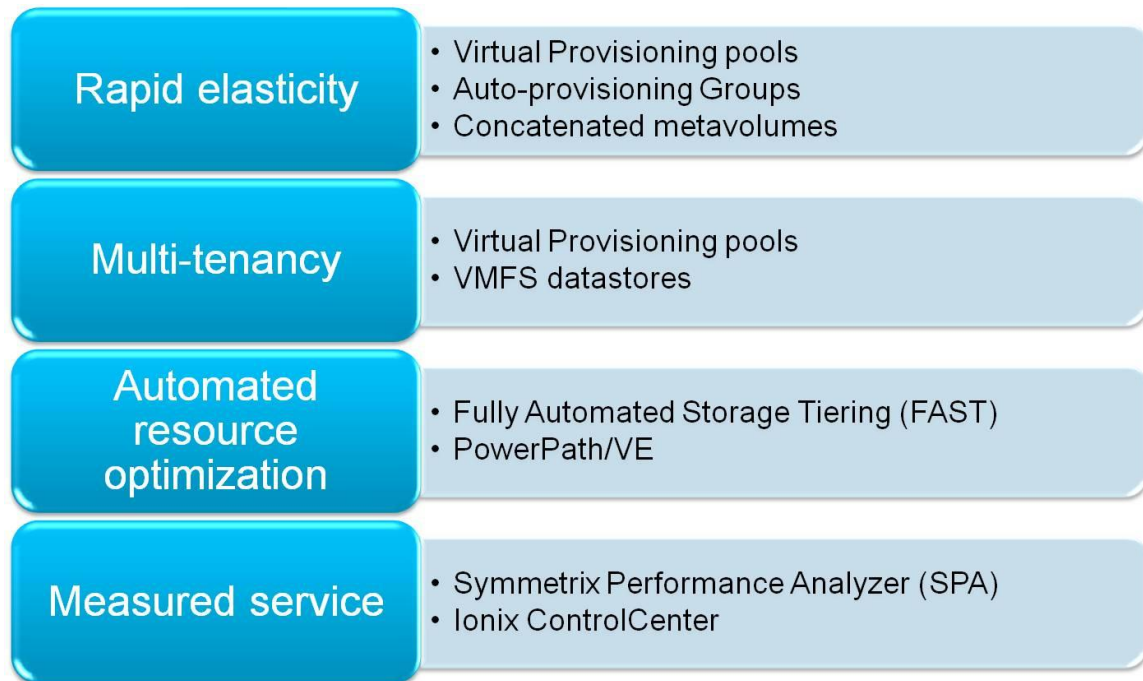


Figure 1. Meeting cloud service provider challenges

Lack of visibility into the application layer

Once virtual resources have been allocated through self-service mechanisms, customers may configure virtual machines, install operating systems, and run applications. It is the responsibility of the service provider to uphold the required SLAs. This involves fine-tuning and configuring the storage array to optimize performance given on a particular workload. However, cloud service providers in general do not have visibility into the application layer. In addition, the virtual resources may be rapidly and dynamically reconfigured. These issues make it difficult to manually tier storage.

Wide striping mechanisms via metavolumes and Virtual Provisioning help alleviate some of these problems by maximizing the utilization of back-end resources. In addition, the introduction of FAST with the Enginuity™ 5874 Q4 '09 service release allows Symmetrix VMAX to monitor and dynamically move thick LUNs across user-defined tiers according to user-defined policies. Symmetrix VMAX automatically accounts for changing workloads by enabling the placement of heavily utilized LUNs onto better performing tiers while moving less accessed LUNs onto lower tiers. This process is nondisruptive and completely invisible to the host. Among others capabilities, fine controls allow administrators to toggle between Automatic and User Approved mode, set time windows for data analysis and LUN movement, and dictate the maximum number of concurrent moves and moves per day. With the current release of Enginuity, FAST only supports the movement of thick devices.

Monitoring and chargeback

The multi-tenancy nature of cloud service provider environments dictates the use of sophisticated reporting tools with the capability to monitor the consumption of virtual resources. In complement, EMC Ionix™ ControlCenter® and Symmetrix Performance Analyzer (SPA) provide capacity and performance planning as well as reporting and analysis. Enhanced capabilities give Ionix ControlCenter StorageScope™ and SPA visibility into the use of storage types in FAST environments and the resulting impact on performance.

With SPA, storage administrators can quickly view key performance indicators such as IOPS and response time of a storage group before and after the execution of a FAST change plan to assess the impact of performance.

With EMC Ionix ControlCenter, storage teams can schedule, execute, and distribute detailed reports to support capacity planning, reporting, and analysis, including:

- Host capacity consumption by storage type to enable application chargeback or show-back processes
- Capacity by storage type for a storage group to support performance troubleshooting and analysis
- Allocated and unallocated capacity by storage type within an array to enhance capacity planning

For more information on managing FAST environments with EMC Ionix ControlCenter and Symmetrix Performance Analyzer, see the white paper *Managing Your FAST Environment with EMC Ionix ControlCenter and Symmetrix Performance Analyzer*.

Terremark

Terremark is a leading global provider of IT infrastructure services. Leveraging purpose-built data centers in the United States, Europe, and Latin America, Terremark delivers to government, enterprise, and Web 2.0 customers a comprehensive suite of managed solutions, including managed hosting, collocation, network, and security services. Of these services, Terremark currently has two cloud computing offerings: vCloud Express and Enterprise Cloud.

Partnered with VMware, Terremark's vCloud Express offering is designed for development teams and department needs. Boasting a quick setup for flexible, high-performance computing, vCloud Express allows customers to dynamically configure resources. Payments are made via credit cards and customers only pay for the resources they use. As such, it lends itself to use cases as a dynamically scalable and flexible development platform. The management console can be accessed via the web interface using a browser. In addition, the vCloud Express API allows programmatic access to many of the functions available in the user interface.

The Enterprise Cloud offering provides precise, dynamic allocation of computing resources with the scale and performance required for enterprise-wide applications. Multiple users with role-based security can be given access to dedicated resource pools. The Enterprise Cloud can be seamlessly integrated into a customer's existing private network. This hybrid environment is facilitated by connections made via secure network channels. Leveraging this mechanism, physical resources such as existing dedicated servers may be combined with the Enterprise Cloud to allow for management from the same web-based interface. This feature is especially powerful for organizations that must adhere to compliance policies and need to maintain infrastructure on dedicated servers and strictly internal networks. Additional features such as "burst mode" allow a pool of additional resources to be dynamically used on a metered basis for dealing with spikes in activity.

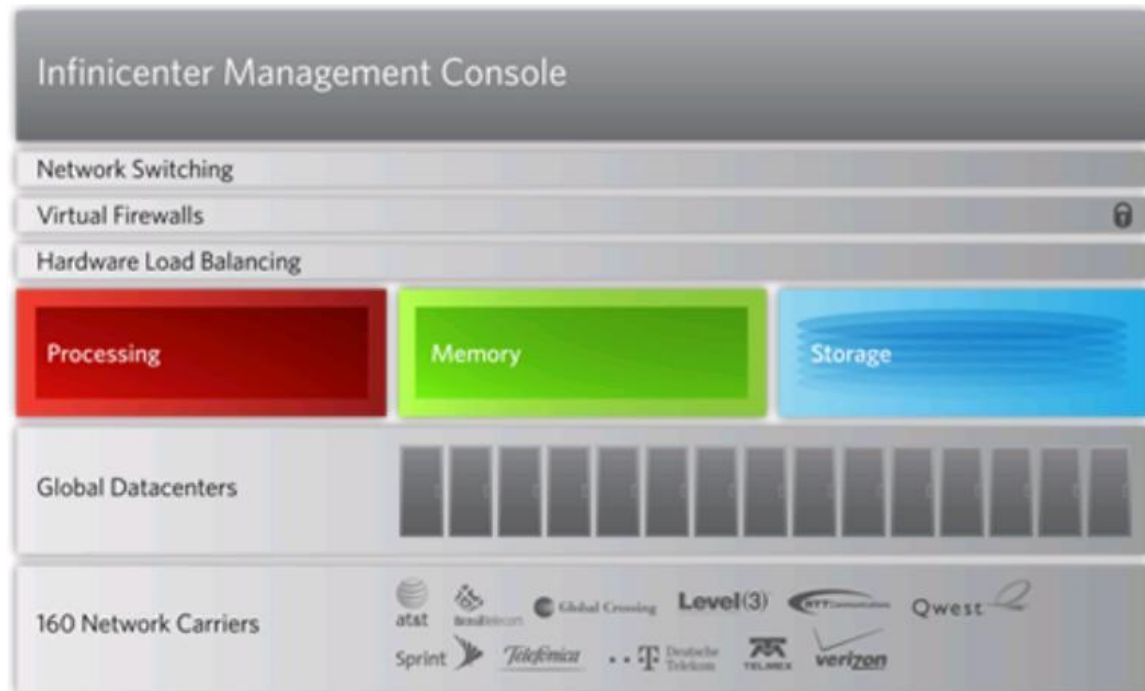


Figure 2. Terremark Enterprise Cloud architecture

For more information on Terremark’s cloud offerings, go to this website:
<http://www.terremark.com/services/cloudcomputing.aspx>

Symmetrix VMAX deployment at Terremark

In deploying the Symmetrix in its cloud infrastructure, Terremark leveraged the cloud-enabling capabilities of the VMAX. The key tenet in their deployment was to spread the data as widely as possible. To do this, Terremark’s VMAX configuration largely features virtually provisioned storage. Virtual Provisioning greatly simplifies data layout and reduces planning and labor overheads. By taking advantage of its wide striping capabilities, Terremark is able to achieve equal or potentially greater performance than standard “thick” provisioning.

Currently, Terremark utilizes a single pool of storage to support its VMware datastores. Data devices for the storage pool were created using large hypervolumes. The drives are controlled by a single VMAX Engine. As the cloud infrastructure continues to grow, Terremark has the flexibility to scale storage by capacity or performance via cache and bandwidth by nondisruptively adding additional drives and engines. Virtual Provisioning facilitates dynamic operations on the underlying storage by enabling administrators to add, remove, and automatically rebalance the data laid out across the pool.

Using VMware vSphere with EMC Symmetrix VMAX

The VMware vSphere virtualization suite consists of various components including ESX/ESXi hosts, vCenter Server, vSphere Client, vSphere web access, and vSphere SDK. In addition to this, VMware vSphere offers a set of distributed services including distributed resource scheduling, high availability, and consolidated backup. VMware vSphere virtualizes the entire IT infrastructure, including servers, storage, and networks. VMware vSphere aggregates these resources and presents a uniform set of elements in the virtual environment (Figure 3). With VMware vSphere, you can manage IT resources like a shared utility and dynamically provision resources to different business units and projects.



Figure 3. VMware vSphere 4.0 architecture

Using vCenter Server, the key elements like hosts, clusters, resource pools, datastores, networks, and virtual machines can be viewed, configured, and managed. vCenter Server aggregates physical resources from multiple ESX/ESXi hosts and presents a central collection of simple and flexible resources for the system administrator to provision to virtual machines in the virtual environment.

vSphere uses the VMware Virtual Machine File System (VMFS) to store virtual machines. VMware VMFS is a high-performance cluster file system that provides storage virtualization that is optimized for virtual machines. Each virtual machine is encapsulated in a small set of files; and VMFS is the default storage management interface for these files on physical SCSI disks and partitions.

VMFS allows for simplified virtual machine provisioning by efficiently storing the entire machine state in a central location, while also allowing multiple ESX host servers to access shared virtual machine storage concurrently. Additionally, VMFS enables virtualization-based distributed infrastructure services that provide a groundwork that allows the scaling of virtualization beyond the confines of a single system. VMFS further simplifies environments by providing a model to allocate storage space to the individual virtual machines without exposing them to the complexity of the physical storage technologies such as Fibre Channel SAN, iSCSI SAN, direct-attached storage, and NAS.

Key features in VMware vSphere 4 include the following:

- **VMware DRS** dynamically load balances server resources to deliver the right resource to the right application based on business priority; this allows applications to shrink and grow as needed.
- **Pluggable Storage Architecture** allows for third-party multipathing software to be loaded into the ESX kernel to better leverage parallel storage connectivity.
- **Hot add** enables CPU and memory to be added to virtual machines when needed without disruption or downtime.
- **Hot plug** enables virtual storage and network devices to be added to or removed from virtual machines without disruption or downtime.

-
- **Hot extend** of virtual disks allows virtual storage to be added to running virtual machines without disruption or downtime.
 - **VMware vNetwork Distributed Switch** simplifies and enhances the provisioning, administration, and control of virtual machine networking in VMware vSphere environments. It also enables third-party distributed virtual switches such as the Cisco Nexus 1000v to be used in vSphere environments, providing network administrators with familiar interfaces to control quality of service at the virtual machine level.
 - **VMware vStorage Thin Provisioning** provides dynamic allocation of storage capacity, allowing storage to be leveraged at a much higher efficiency.
 - **VMware VMotion** eliminates the need to schedule application downtime due to scheduled server maintenance through live migration of virtual machines across servers with no disruption to users or loss of service.
 - **VMware Storage VMotion** eliminates the need to schedule application downtime due to planned storage maintenance or during storage migrations by enabling live migration of virtual machine disks with no disruption to users or loss of service.
 - **VMware High Availability (HA)** provides cost-effective, automated restart within minutes for all applications in the event of hardware or operating system failures.
 - **VMware Fault Tolerance (FT)** provides continuous availability, without any data loss or downtime, to any application, in the event of hardware failures.
 - **VMFS Volume Grow** allows for a nondisruptive VMFS volume to grow to fill a storage device that has been increased in size since the datastore was originally created on it.

For a more detailed listing of important benefits and features of VMware vSphere 4, please refer to the *VMware vSphere 4 Key Features and Benefits* document available at <http://www.vmware.com>.

End-to-end visibility integrated into VMware with EMC Storage Viewer

The VMware virtualization suite is the most widely deployed software suite for optimizing and managing IT environments through virtualization. With the maturing of this technology as a platform of choice for deploying mission-critical x86 applications, business activities are frequently completed from within the VMware vSphere Client. To aid in this goal, EMC now provides EMC Storage Viewer for vSphere Client. This free tool (available to download on Powerlink[®]) enables additional capabilities to the vSphere Client, so that users may now view detailed storage-specific information, previously impossible to obtain without using one or two separate applications outside of vSphere Client. EMC Storage Viewer provides simple, read-only storage mapping functionality for various EMC storage-related entities that exist within vSphere Client, including datastores, LUNs, and SCSI targets. The storage information displayed through EMC Storage Viewer allows distinction among the types of storage used, the specific arrays and devices presented, the paths that are used for the storage, and the individual characteristics of the existing storage.

EMC Storage Viewer provides two main views: the VMware ESX context view and the Virtual Machine context view.

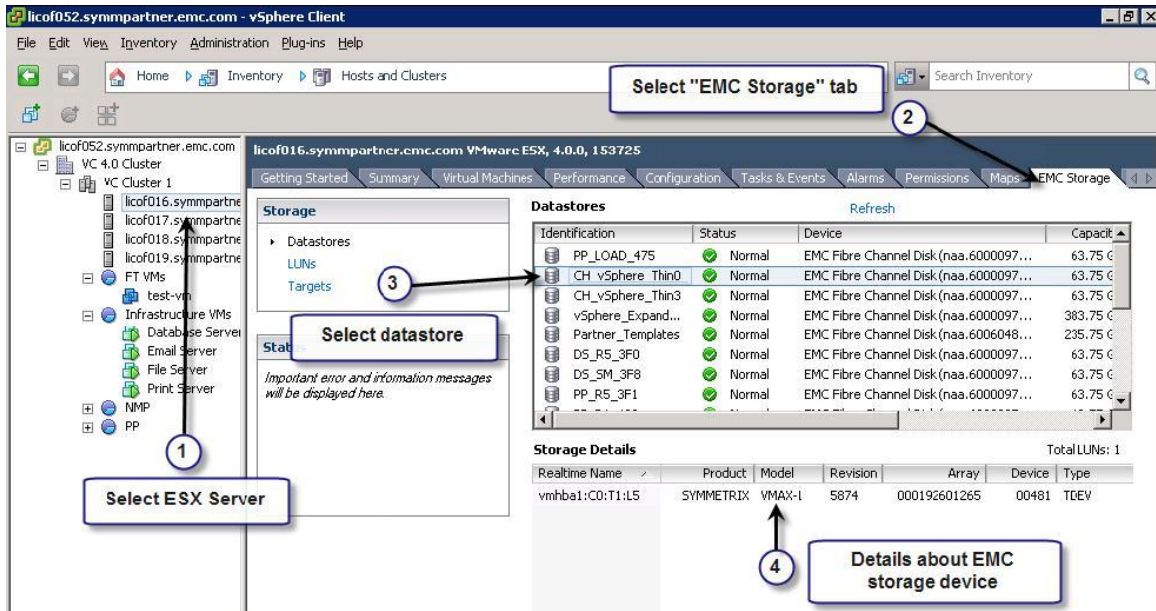


Figure 4. Datastore subview of the VMware ESX context view of EMC Storage Viewer

The VMware **ESX context view** contains three subviews:

- The **datastore subview** provides detailed information about EMC storage associated with VMFS datastores visible on the selected ESX host.
- The **LUN subview** displays detailed information about EMC LUNs provisioned to the ESX host server. For Symmetrix VMAX, this information includes LUN protection type, serial number of the array, Engenuity level, device type, metavolume information, Virtual Provisioning information, and more.
- The **Targets subview** drills down to the SCSI targets level and displays detailed information on each, allowing for the most efficient way to map specific Symmetrix arrays to their respective SCSI target numbers on the ESX host system.

The **Virtual Machine context view** displays detailed information about EMC devices associated with the selected virtual machine. This is sorted by virtual disks (VMFS) and raw device mappings (RDM) in use by the virtual machine.

Without EMC Storage Viewer, the vSphere Client by itself provides an inadequate amount of detail for mapping data sources presented to the ESX host system or virtual machines to their underlying storage devices. EMC Storage Viewer allows users to view and decipher all storage entities in ways that the vSphere Client has never allowed before. Due to the combination of the EMC Storage Viewer and VMware vSphere Client, managing a virtual environment attached to EMC storage is easier and far more efficient than ever before. For these reasons, it is highly recommended that EMC Storage Viewer be installed to manage any VMware vSphere environment making use of Symmetrix VMAX arrays. An in-depth description of the functionality provided by EMC Storage Viewer for VMware vSphere Client is available in the white paper *Using EMC Storage Viewer Version 2.0 for vSphere Client* on Powerlink.

Symmetrix VMAX connectivity

To ensure high availability, each ESX host attached to a Symmetrix VMAX system should have visibility to its LUNs through at least two separate paths. This means that there should be at least two physical HBAs, with each HBA connected to a different Symmetrix front-end port. This configuration ensures continued operations in the event of a single HBA failure, link failure, or Symmetrix VMAX front-end port failure.

The goal is to avoid a single point of failure. As much as possible, multiple connections from a host should be kept on distinct subsets of components. On the Symmetrix front end this means spreading connections from the same host to separate directors, separate slices within these directors, and separate ports per slice. For example, in a Symmetrix VMAX with a single engine, each HBA for a host may be connected to the same port on the two directors (for example, 7E:0 and 8E:0) through two SANs. Thus, the host is connected to the Symmetrix VMAX via four connections – two per HBA – to a total of four processors on two directors, ensuring resilience and availability.

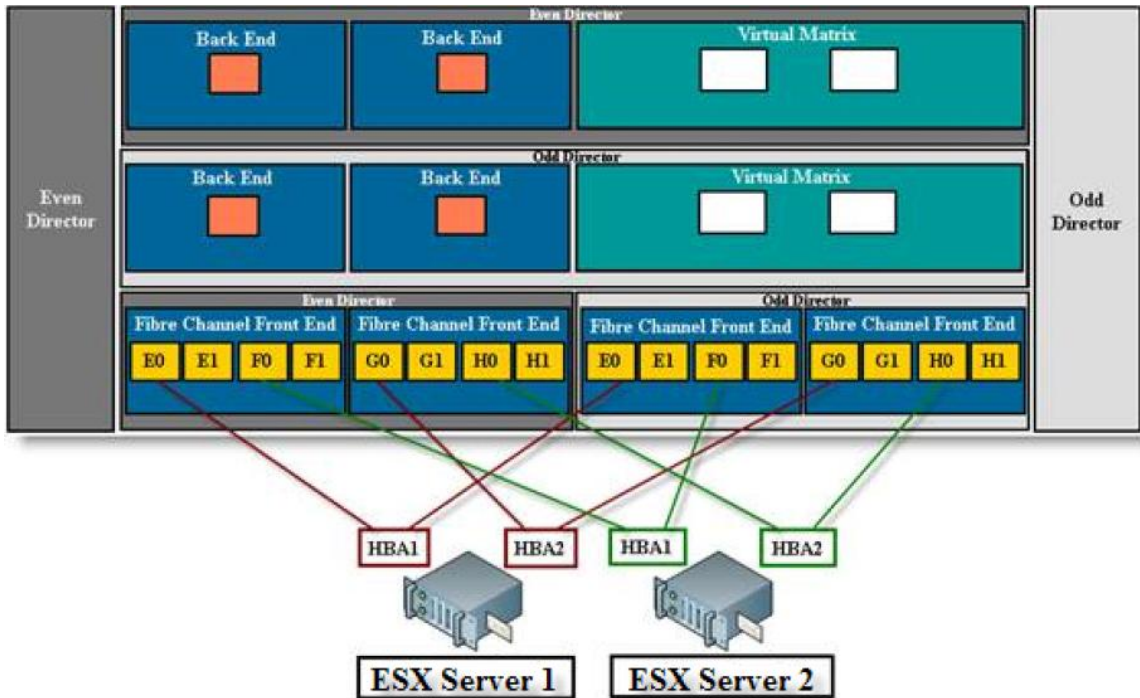


Figure 5. Connecting ESX servers to a single-engine Symmetrix VMAX

As the Symmetrix VMAX system expands and more engines are added, connectivity can be scaled as needed. Since both ports on a slice share processor resources, hosts should be connected to the 0 ports before connections are made for additional hosts to the 1 ports of the same director. These methodologies for connectivity ensure all front-end directors and processors are utilized, providing maximum potential performance and load balancing for VMware vSphere environments connected to the Symmetrix VMAX storage arrays.

Symmetrix VMAX HBA port flags

For proper functioning of a Symmetrix VMAX/ESX environment, it is important to have multiple bits set on the front-end directors. Table 1 lists the required and optional bit settings. For detailed information, refer to the *EMC Support Matrix* on EMC.com.

Table 1. Symmetrix VMAX HBA port bit settings for ESX 4.0

Bit	Optional/Required	Description
Common_Serial_Number	Required	This flag should be enabled for multipath configurations or hosts that need a unique serial number to determine which paths lead to the same device.
Unique_WWN	Required	When enabled (default) for all environment configuration changes and new environments to ensure unique World Wide Names (WWN) within the fibre environment (uses Symmetrix serial numbers and port numbers). When disabled, you don't have to change WWNs.
SCSI_3	Required	When enabled, the Inquiry data is altered when returned by any device on the port to report that the Symmetrix supports the SCSI 3 protocol. When this flag is disabled, the SCSI 2 protocol is supported.
SPC-2	Required	Provides compliance to newer SCSI protocol specifications.
ACLX	Required	When enabled, allows storage provisioning using initiator groups. This flag is applicable for Engenuity 5874 and later.
OS2007	Optional	When enabled, this flag provides a stricter compliance with SCSI standards for managing device identifiers, multiport targets, unit attention reports, and the absence of a device at LUN 0.

Port flags can be set in a number of ways. HBA port flags can be set on a per-initiator basis or by hostname. This feature allows specific host flags to be enabled and disabled on a specified director and port or on all directors and all ports. Alternatively, port flags can be set at the initiator group level. This provides flexibility for connecting the Symmetrix VMAX to a number of different operating systems. Note that a flag cannot be set for the group if it conflicts with any initiator in the group. After a flag is set for a group, it cannot be changed on an initiator basis. Initiator groups are further explained in the “On-demand storage provisioning with Auto-provisioning” section on page 16. Port flags can be set using Solutions Enabler as well as Symmetrix Management Console.

Automating load balancing with PowerPath/VE

EMC PowerPath/VE delivers PowerPath Multipathing features to optimize VMware vSphere virtual environments. With PowerPath/VE, you can standardize path management across heterogeneous physical and virtual environments. PowerPath/VE enables you to automate optimal server, storage, and path utilization in a dynamic virtual environment. With hyper-consolidation, a virtual environment may have hundreds or even thousands of independent virtual machines running, including virtual machines with varying levels of I/O intensity. I/O-intensive applications can disrupt I/O from other applications and before

the availability of PowerPath/VE, load balancing on an ESX host system had to be manually configured to correct for this. Manual load-balancing operations to ensure that all virtual machines receive their individual required response times are time-consuming and logistically difficult to effectively achieve.

PowerPath/VE works with VMware ESX and ESXi as a multipathing plug-in (MPP) that provides enhanced path management capabilities to ESX and ESXi hosts. PowerPath/VE is supported with vSphere (ESX4) only. Previous versions of ESX do not have the PSA, which is required by PowerPath/VE.

PowerPath/VE installs as a kernel module on the vSphere host. PowerPath/VE will plug in to the vSphere I/O stack framework to bring the advanced multipathing capabilities of PowerPath – dynamic load balancing and automatic failover – to the VMware vSphere platform (Figure 6).

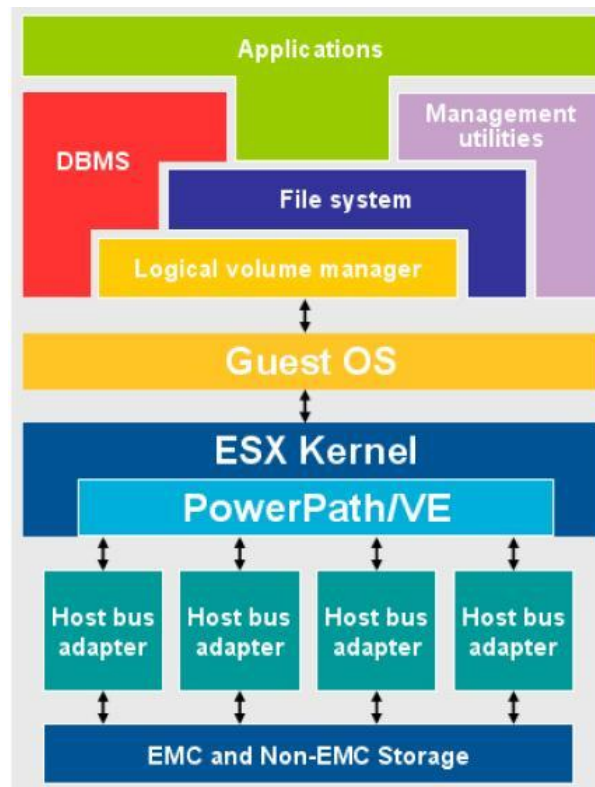


Figure 6. PowerPath/VE vStorage API for multipathing plug-in

At the heart of PowerPath/VE path management is server-resident software inserted between the SCSI device-driver layer and the rest of the operating system. This driver software creates a single “pseudo device” for a given array volume (LUN) regardless of how many physical paths on which it appears. The pseudo device, or logical volume, represents all physical paths to a given device. It is then used for creating virtual disks, and for RDM, which is then used for application and database access.

PowerPath/VE’s value fundamentally comes from its architecture and position in the I/O stack. PowerPath/VE sits above the HBA, allowing heterogeneous support of operating systems and storage arrays. By integrating with the I/O drivers, all I/Os run through PowerPath and allow for it to be a single I/O control and management point. Since PowerPath/VE resides in the ESX kernel, it sits below the guest OS level, application level, database level, and file system level. PowerPath/VE’s unique position in the I/O stack makes it an infrastructure manageability and control point – bringing more value going up the stack.

PowerPath/VE features

PowerPath/VE provides the following features:

- **Dynamic load balancing** – PowerPath is designed to use all paths at all times. PowerPath distributes I/O requests to a logical device across all available paths, rather than requiring a single path to bear the entire I/O burden.
- **Auto-restore of paths** – Periodic auto-restore reassigns logical devices when restoring paths from a failed state. Once restored, the paths automatically rebalance the I/O across all active channels.
- **Device prioritization** – Setting a high priority for a single device or several devices improves their I/O performance at the expense of the remaining devices, while otherwise maintaining the best possible load balancing across all paths. This is especially useful when there are multiple virtual machines on a host with varying application performance and availability requirements.
- **Automated performance optimization** – PowerPath/VE automatically identifies the type of storage array and sets the highest performing optimization mode by default. For Symmetrix, the mode is SymmOpt (Symmetrix Optimized).
- **Dynamic path failover and path recovery** – If a path fails, PowerPath/VE redistributes I/O traffic from that path to functioning paths. PowerPath/VE stops sending I/O to the failed path and checks for an active alternate path. If an active path is available, PowerPath/VE redirects I/O along that path. PowerPath/VE can compensate for multiple faults in the I/O channel (for example, HBAs, fiber-optic cables, Fibre Channel switch, storage array port).
- **Monitor/report I/O statistics** – While PowerPath/VE load balances I/O, it maintains statistics for all I/O for all paths. The administrator can view these statistics using rpowermt.
- **Automatic path testing** – PowerPath/VE periodically tests both live and dead paths. By testing live paths that may be idle, a failed path may be identified before an application attempts to pass I/O down it. By marking the path as failed before the application becomes aware of it, timeout and retry delays are reduced. By testing paths identified as failed, PowerPath/VE will automatically restore them to service when they pass the test. The I/O load will be automatically balanced across all active available paths.

PowerPath/VE in vCenter Server

PowerPath/VE for vSphere is managed, monitored, and configured using rpowermt as discussed in the previous section. This CLI-based management is common across all PowerPath platforms and presently, there is very little integration at this time with VMware management tools. However, LUN ownership is presented in the GUI.

As seen in Figure 7, under the ESX Configuration tab and within the Storage Devices list, the owner of the device is shown.

View:

Devices Refresh

Runtime Name	LUN	Type	Transport	Capacity	Owner
vmhba1:C0:T2:L94	94	disk	Fibre Channel	64.00 GB	PowerPath
vmhba1:C0:T1:L82	82	disk	Fibre Channel	64.00 GB	PowerPath
vmhba1:C0:T2:L92	92	disk	Fibre Channel	64.00 GB	PowerPath
vmhba1:C0:T1:L0	0	disk	Fibre Channel	384.01 G	PowerPath
vmhba1:C0:T1:L4	4	disk	Fibre Channel	384.01 G	PowerPath
vmhba0:C1:T0:L0	0	disk	Block Adapter	67.05 GB	NMP

Figure 7. Device ownership in vCenter Server

Figure 7 shows a number of different devices owned by PowerPath. A set of *claim rules* are added to the vSphere PSA, which enables PowerPath/VE to manage supported storage arrays. As part of the initial installation process and claiming of devices by PowerPath/VE, the system must be rebooted. Nondisruptive

installation is possible by using VMware VMotion to temporarily migrate running VMs to another ESX host in the cluster.

On-demand storage provisioning with Auto-provisioning

Auto-provisioning is the improved way to present storage from the Symmetrix with the VMAX. With Auto-provisioning, mapping and masking are accomplished using initiator groups, port groups, and storage groups. A masking view is constructed with one of each of these groups. At this point, the devices are mapped to each FA port and masked to each initiator. These groups and their associations to views are kept in the Access Logix™ database, which resides in the internal Symmetrix file system.

This new approach significantly decreases the number of steps needed to perform all aspects of provisioning including initial provisioning, adding or removing capacity, adding or removing front-end ports, and adding or removing HBAs to and from a host, or adding or removing hosts to and from a cluster. Reducing steps reduces the overall time and risk of error. The benefits are ideally suited for today's virtual data center where servers have multiple HBAs, and are often organized into databases, virtual servers, and high-availability clusters. In addition, many application environments require the ability to isolate workloads to a specific set of HBAs and other storage resources. The flexibility of Auto-provisioning allows an administrator to create the required masking views to easily accommodate these requirements.

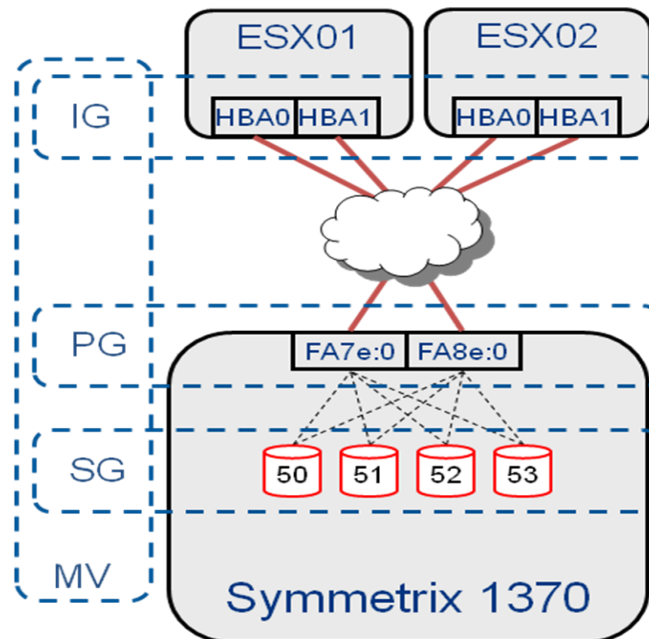


Figure 8. Auto-provisioning Groups

The following are some considerations when planning for Auto-provisioning.

Storage groups

A storage group is a logical grouping of Symmetrix devices.

- A group may contain up to 4,096 devices.
- A Symmetrix device may be a member of multiple storage groups.
- The same storage group could be part of multiple masking views.
- A single Symmetrix VMAX can have up to 8,192 storage groups.

-
- When a masking view is created, or when additional devices are added to an existing storage group that is part of a masking view, each device in the storage group is assigned a LUN address using the dynamic LUN addressing feature. Dynamic LUN addresses are the device addresses used by the host and are independent of the channel address assigned when a device is mapped to a front-end port. The LUN address assigned to each device will be consistent across all paths.
 - Storage groups are not only used for masking operations, but they are also used for other operations that require grouping such as device migration and FAST.

Port groups

A port group is a logical grouping of Fibre Channel and/or iSCSI front-end director ports.

- A single Symmetrix VMAX may be configured with up to 8,192 port groups.
- The number of ports in a port group is limited by the number of ports in the Symmetrix VMAX.
- A port may be a member of multiple port groups.
- The same port group can be part of multiple masking views.
- Front-end ports can be segregated to isolate workloads by grouping one or more ports into separate port groups and using the different port groups for different masking views.
- For availability and performance, devices should be mapped to two or more front-end director ports. These ports should be on different directors, and if possible, different engines. With the exception of a single engine system, the “Rule of 17” can still apply.
- Front-end ports are typically shared between multiple servers and HBAs. Reference E-Lab™ Navigator on Powerlink for supported fan-out ratios.
- Consider actual or estimated workloads to ensure front-end ports are not oversubscribed.
- Ports can be added or removed from a port group as workloads change and the view is updated automatically. This allows a storage administrator to easily balance the workload across front-end directors
- To be a member of a port group the ACLX flag must be enabled on the port through the Set Port Attributes dialog of SMC or the `symconfigure` command.
- When a masking view is created and the devices in the associated storage group are not currently mapped to the ports in the port group, they will automatically be mapped using the next available channel address. On the Symmetrix VMAX, the channel address assigned to the device on the port is not visible to the host. Instead, the assigned dynamic LUN address is what is visible.

Initiator groups

An initiator groups is a logical grouping of a combination of Fibre Channel initiators and iSCSI names.

- A group may contain up to 32 FC initiators, eight iSCSI names, or a combination of both.
- A single Symmetrix VMAX may be configured with up to 8,192 initiator groups.
- An initiator may belong to only one initiator group.
- An initiator group may also contain the name of another initiator group to allow the groups to be cascaded to a depth of one. The next section, “Cascaded initiator groups,” has more information.
- Port flags may be set on an initiator group basis, with one set of port flags applying to all initiators in the group.
- Generally the more paths in an initiator group, the greater the management efficiency. While it is possible to define an initiator group with a single HBA and create a view for each HBA, additional efficiency is gained by grouping related HBAs.
- An empty initiator group may be created as a placeholder and used to create a masking view. HBAs could be added to the initiator group later and the masking view would be updated and the required mapping and masking performed automatically. (This approach is useful when planning a new server

installation and the actual WWNs are unknown, yet is desirable to build the groups and preallocate the storage.)

Cascaded initiator groups

An initiator can only belong to a single initiator group. However, cascaded initiator groups can be configured to allow an initiator group to be a member of another initiator group.

A cluster is a good example of where cascaded initiator groups might be appropriate. Figure 9 illustrates an example of cascaded initiator groups. Here, initiator groups are created for each server containing all HBAs for that server. Each child initiator group is used to create a masking view that contains devices used exclusively by that server. For example, it may contain the boot device, dedicated gatekeeper devices, and devices used for nonclustered applications. Next, a cascaded initiator group is created to contain the initiator groups for each server in a cascaded configuration. Using the cascading initiator group, another masking view is created using a storage group with devices shared by all servers in the cluster.

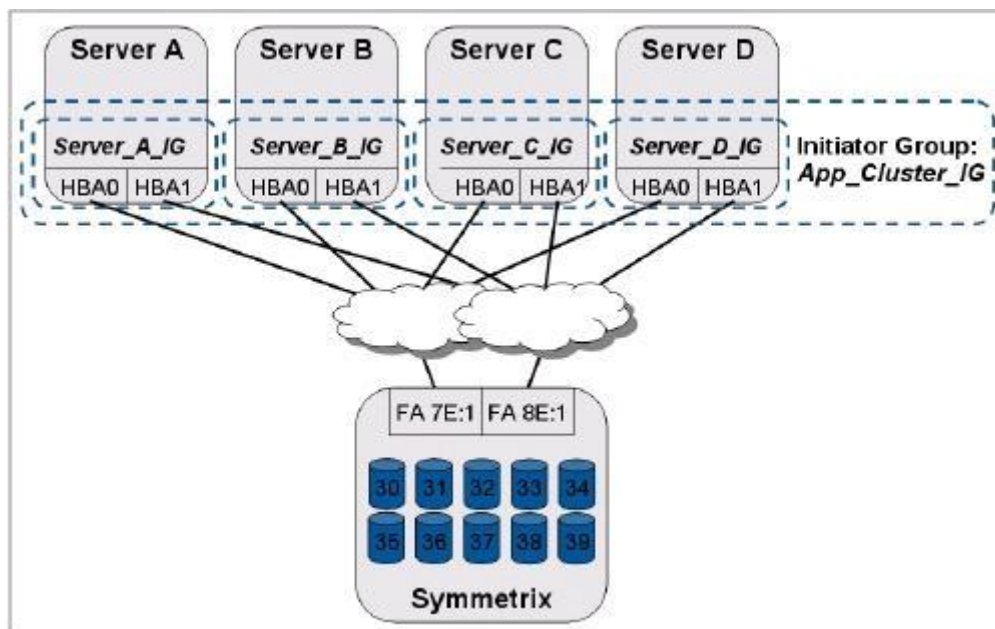


Figure 9. Cascaded initiator groups

Cascading of initiator groups is only allowed to a depth of one – that is, an initiator can be part of an initiator group, which in turn can be part of another initiator group. An initiator group can be a member of only one masking view; however, an initiator group can be a member of multiple cascaded initiator groups and each cascaded initiator group can be associated with different masking views. A cascaded initiator group can have a maximum of 1,024 initiators total.

When configuring cascaded initiator groups in a cluster environment, it is possible to have a different number of devices assigned to each host. When a view is created that spans multiple servers, a consistent dynamic LUN address will be applied but this may result in nonconsecutive addresses and holes in the address range.

Masking views

Masking views associate a set of devices to Symmetrix front-end ports and server HBAs. A masking view is constructed from an initiator group, a port group, and a storage group. Once a view is created, the necessary mapping and masking operations are performed automatically to provision storage through the specified paths. Masking views are dynamic in the sense that any changes made to a storage, port, or initiator group are automatically propagated to any existing views that the group participates in. In practice,

this enables the flexible provisioning of storage assets into and out of clusters, reducing the effort in operating and maintaining a dynamic environment. This makes Auto-provisioning ideal for virtualized environments. It also reduces complexity, execution time, labor cost, and the risk of error. Depending on server and application requirements, each server or group of servers may have one or more masking views that associate a set of Symmetrix devices to an application, server, or cluster of servers. The key to ease of management is descriptive names for all groups and views. As a best practice, the name of a masking view should readily identify the application and/or server names. To be effective, naming standards must be consistently utilized and understood by all administrators. This helps keep information on the allocation of storage assets organized and readily accessible.

Dynamic LUN addressing

Starting with Enginuity 5772, the optional dynamic LUN addressing feature provides a more flexible approach to device addressing by eliminating the direct connection between the channel address that is assigned when a device is mapped to a front-end director port and the logical unit number (LUN) that is visible to the host. Instead, when creating a masking entry, a storage administrator assigns either a specific LUN address or the next available LUN address is automatically assigned, and this LUN address is what is visible to the host system. Dynamic LUN addressing eliminates many host-addressing issues as it provides each host its own address space that is independent of the channel address on the front-end port. For example, many hosts are limited to 256 devices with addresses in the range of 00-FF (0-256). It also enables all hosts that share a FA port to have a LUN with an address that starts with 0.

Dynamic LUN addressing is an integral part of Auto-provisioning. It provides for provisioning flexibility by abstracting the LUN that is assigned to a particular device at the FA level from the LUN that is seen by the host. A dynamic LUN address is assigned either automatically or manually to a device on a per-initiator basis regardless of the assigned LUN value at the time of mapping the device to the FA. When provisioning to a cluster of servers, dynamic LUN addressing helps eliminate the potential for inconsistent LUN addresses for the same device on different paths and thus simplifies the provisioning process and management. LUN addresses can be set manually using the `-lun` flag for the `symaccess` command.

```

C:\Documents and Settings\Administrator>symaccess -sid 70 list
Symmetrix ID          : 000192601370

Group Name            Count  Uiew Count  Type
-----
LICOD232              -      -          Initiator
LICOD056              -      -          Initiator
ports_1               -      -          Port
LICOD056_GK          -      -          Storage
R5_DS                 -      -          Storage

C:\Documents and Settings\Administrator>symaccess -sid 70 create view -name D56_DS -sg R5_DS -pg ports_1 -ig LICOD056 -lun 5
C:\Documents and Settings\Administrator>_
  
```

Figure 10. Manually setting dynamic LUN addresses

Rapid elasticity and resource pooling with Virtual Provisioning

Symmetrix thin devices are logical devices that can be used in many of the same ways that Symmetrix devices have traditionally been used. Unlike traditional Symmetrix devices, thin devices do not need to have physical storage completely allocated at the time the device is created and presented to a host. A thin device is not usable until it has been bound to a shared storage pool known as a thin pool. Multiple thin

devices may be bound to any given thin pool. The thin pool is comprised of devices called data devices that provide the actual physical storage to support the thin device allocations.

When a write is performed to a part of any thin device for which physical storage has not yet been allocated, the Symmetrix allocates physical storage from the thin pool for that portion of the thin device only. The Symmetrix operating environment, Enginuity, satisfies the requirement by providing a unit of physical storage from the thin pool called a thin device extent. This approach reduces the amount of storage that is actually consumed, enabling more applications to be stored and more internal or external customers to be supported on a single array in a cloud environment.

The thin device extent is the minimum amount of physical storage that can be reserved at a time for the dedicated use of a thin device. An entire thin device extent is physically allocated to the thin device at the time the thin storage allocation is made as a result of a host write operation. A round-robin mechanism is used to balance the allocation of data device extents across all of the data devices in the pool that are enabled and that have remaining unused capacity. The thin device extent size is 12 tracks (768 KB). Note that the initial bind of a thin device to a pool causes one thin device extent, or 12 tracks, to be allocated per thin device. If the thin device is a metavolume, then one thin device extent is allocated per meta member device. So a four-member thin metavolume would cause 48 tracks (3078 KB) to be allocated when the device is bound to a thin pool.

When a read is performed on a thin device, the data being read is retrieved from the appropriate data device in the thin pool. If a read is performed against an unallocated portion of the thin device, zeros are returned to the reading process.

When more physical data storage is required to service existing or future thin devices, for example, when a thin pool is running out of physical space, data devices can be added to existing thin pools dynamically without needing a system outage. Starting with Solutions Enabler 7.1 in conjunction with the Enginuity 5874 Q4'09 service release, the Automated Pool Rebalancing feature allows users to dynamically rebalance allocated extents over the data devices in a thin pool as new data devices are added. New thin devices can also be created and bound with existing thin pools.

When data devices are added to a thin pool they can be in an enabled or disabled state. In order for the data device to be used for thin extent allocation it needs to be enabled. For it to be removed from the thin pool, it needs to be in a disabled state. Beginning with Enginuity 5874, active data devices can be disabled, which will cause any allocated extents to be drained to the other enabled devices in the pool. They can then be removed from the pool when the drain operation has completed.

The following figure depicts the relationships between thin devices and their associated thin pools. There are nine devices associated with thin Pool A and three thin devices associated with thin pool B.

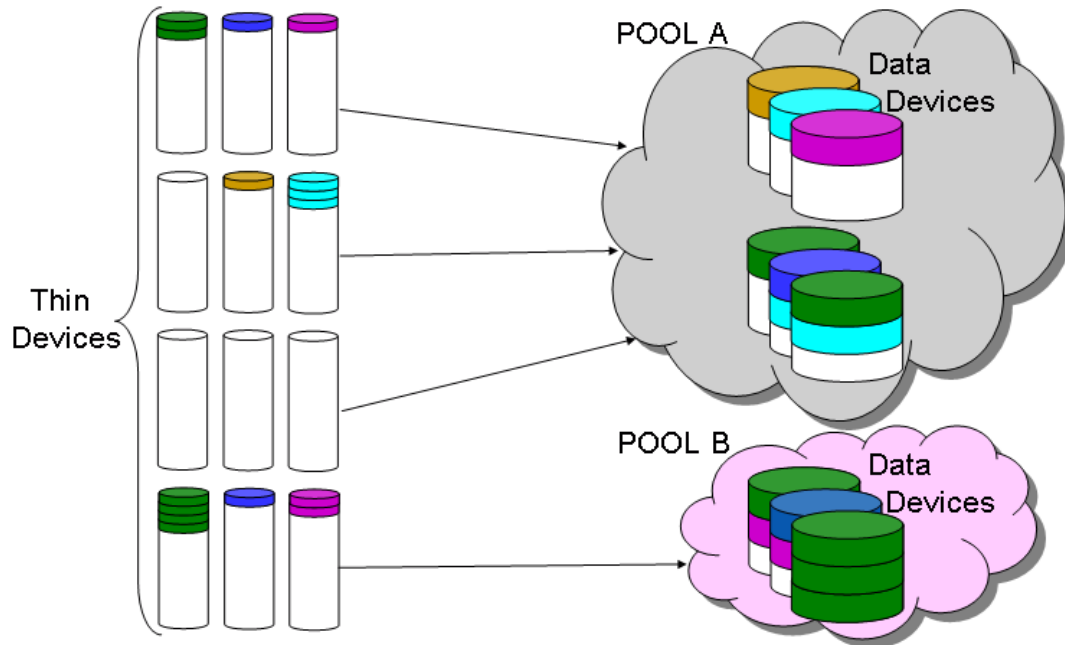


Figure 11. Thin devices and thin pools containing data devices

The way thin extents are allocated across the data devices results in a form of striping in the thin pool. The more data devices that are in the thin pool, the wider the striping is, and therefore the greater the number of devices that can participate in application I/O.

Considerations for vSphere environments on thin devices

Bound thin devices appear like any other SCSI-attached device to VMware vSphere. A thin device can be used to create a VMware file system, or assigned exclusively to a virtual machine as an RDM.

One thing to note is that when a thin device is presented in RDM (either physical or virtual compatibility) to a virtual machine, the VMware kernel does not play a direct role in the I/O to the thin device generated by the guest operating system running in the virtual machine. In this configuration, the considerations for using thin devices are no different from the ones for physical servers running the same operating system and applications.

VMware vSphere 4 offers multiple ways of formatting virtual disks and for the first time, has integrated these options into vCenter itself. When creating, cloning, or converting virtual disks, the default option is “thick.” The “thick” selection, as it is referred to within the vSphere Client, is actually the “zeroedthick” format. This is the recommended disk format for use with Symmetrix Virtual Provisioning. In this allocation scheme, the storage required for the virtual disks is reserved in the datastore but the VMware kernel does not initialize all the blocks. The blocks are initialized by the guest operating system as write activities to previously uninitialized blocks are performed. The VMware file system will return zeros to the guest operating system if it attempts to read blocks of data that it has not previously written to. This is true even in cases where information from previous allocation is available—the VMware file system will not present stale data to the guest operating system when the virtual disk is created using the “zeroedthick” format. Since the VMFS volume will report the virtual disk as fully allocated, the risk of oversubscribing and running out of space in the thin pool is reduced.

As shown in Figure 12, when creating a new virtual machine, the default allocation mechanism is “zeroedthick”. Other options include “eagerzeroedthick” and “thin”: With “eagerzeroedthick”, space required for the virtual disk is completely allocated and written to at creation time; with “thin”, capacity is neither reserved nor initialized at the VMFS level. The use of “zerodthick” is recommended for deployment with Symmetrix Virtual Provisioning. For more information on these different mechanisms, refer to the white paper *Implementing EMC Symmetrix Virtual Provisioning with VMware vSphere*.

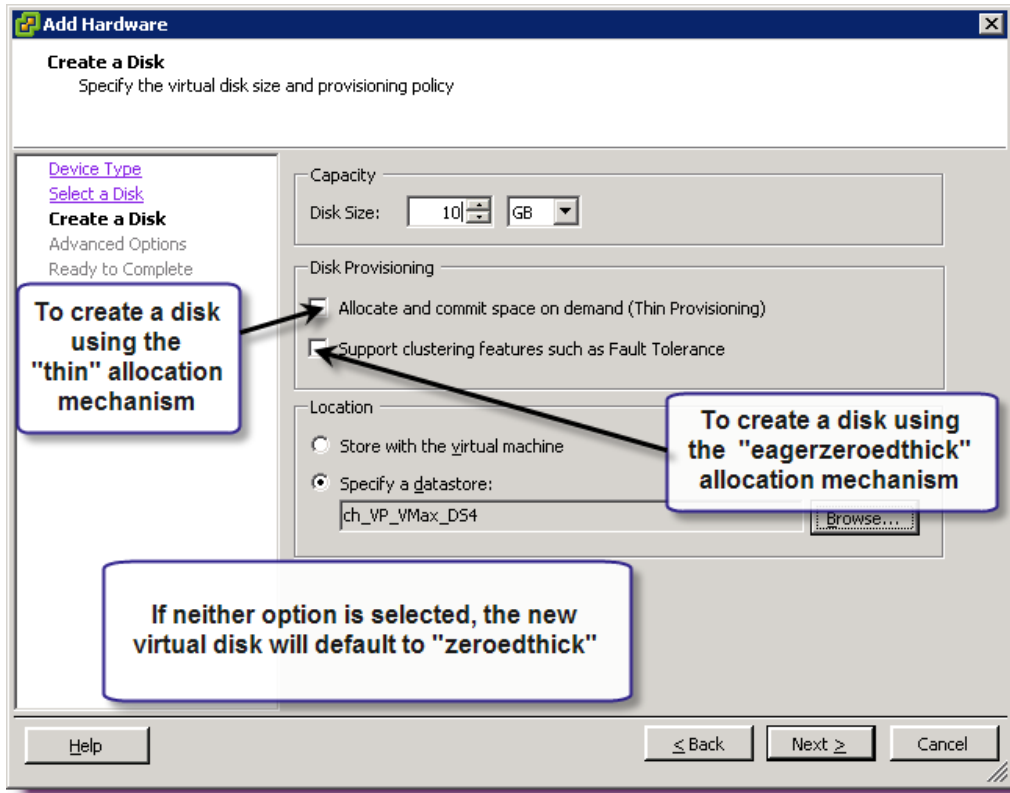


Figure 12. Virtual disk allocation mechanism options in vSphere 4.0

Automated Pool Rebalancing

Users of Symmetrix Virtual Provisioning are able to automatically and nondisruptively rebalance used capacity across all devices (and drives) in order to maximize the performance benefits of wide striping. This helps reduce total cost of ownership (TCO), because users are able to add capacity to a pool in small increments. That is, a small number of data devices can be used to expand a pool while still ensuring data is striped widely across the thin pool.

Automated Pool Rebalancing allows the user to run a command against a thin pool that will rebalance the allocated “chunks” or 768 KB extents (12 tracks) across all enabled data devices in the pool. The balancing algorithm will calculate the minimum, maximum, and mean used capacity values of the data devices in the thin pool. The Symmetrix will then move thin device extents from the data devices with the highest used capacity to those with the lowest until the pool is balanced. As many as eight thin pools may be rebalanced at any one time.

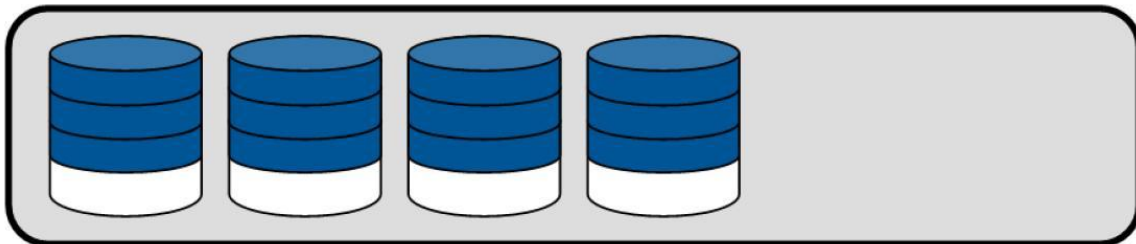


Figure 13. Thin pool before adding new data devices

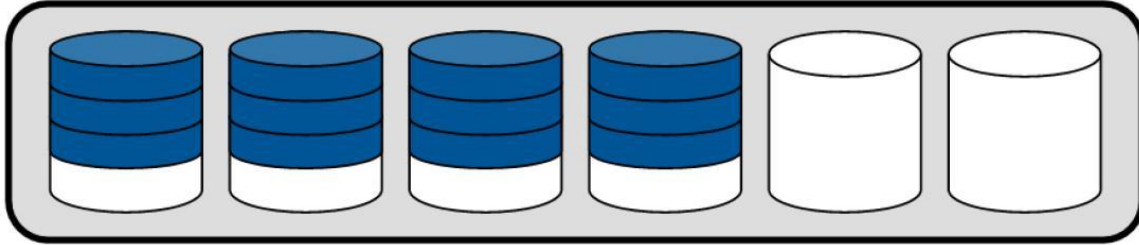


Figure 14. Thin pool after new data devices are added

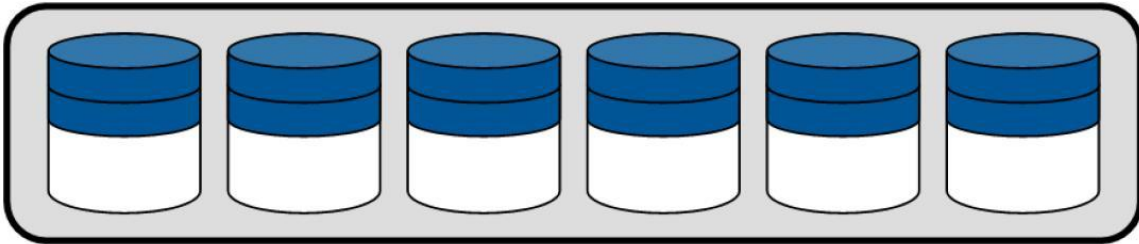


Figure 15. Thin pool after Automated Pool Rebalancing

Space reclamation

Users are able to automatically reclaim Virtual Provisioning chunks (extents) that contain all zeros in order to reduce capacity requirements and TCO. This feature will provide the greatest benefit after migrating standard volumes to thin volumes in a VMAX array, for example with host-based migration tools (for example, Open Migrator for non-Windows environments), Open Replicator, TimeFinder®/Clone, or SRDF®. Space reclamation will thin out the new thin volumes by identifying and returning to the thin pool any “chunks” that contain all zeros. By reclaiming that allocated, “trapped” storage, users can reclaim unused capacity for use by other applications.

Space reclamation is an extension of the existing Virtual Provisioning space de-allocation mechanism. Running the space reclamation command will spawn a DA background task that will examine the allocated extents on specified thin devices. For each allocated extent, all 12 tracks will be brought into cache and examined to see if they contain all zero data. If the entire extent contains all zero data, the extent will be de-allocated and added back into the pool, making it available for a new extent allocation operation.

Space reclamation is not supported on actively replicating SRDF volumes. The link must be suspended prior to running the reclamation operation. Reclamation will also not be performed on tracks that participate in a local replication session.

Considerations for space reclamation in VMware environments

For both Virtual Infrastructure 3.x and vSphere 4.0, there are important considerations that should be evaluated before performing reclaims.

If a thin device stores multiple virtual machines, a zero reclaim function cannot be performed on just one of the virtual disks. It is currently not possible to determine the specific blocks each virtual disk consumes on the thin device and therefore it is not possible to limit a reclaim to a single virtual disk out of many on a given VMware file system.¹ Consequently, a reclaim should only be performed on a thin device if all virtual machines hosted by the VMware file system volume on that thin device can and should be reclaimed. In particular, in vSphere environments, virtual machines that are set up in a Fault Tolerant configuration should not share Symmetrix thin devices (or the VMware file systems that are hosted by

¹ EMC is actively working with VMware on the vStorage initiative that may allow in the future for zero reclaim to be performed on a single virtual disk.

them) with other virtual machines. This will reduce the risk of reclaiming the zeros from a Fault Tolerant virtual machine.

Additionally, zeroing out by guest operating systems must be taken into account before performing a zero reclaim. If the virtual machines' operating system or its hosted applications zero out files for a particular reason, care must be taken to consider any implications of removing those zeros before reclaiming them.

The use of EMC Storage Viewer (available for Virtual Infrastructure 3.5 and vSphere 4.0) is highly recommended in order to easily map virtual machines and their corresponding VMFS volumes to the correct underlying Symmetrix thin device. Double-checking storage mapping information with the EMC Storage Viewer will eliminate the possibility of performing a zero reclamation on the incorrect thin device.

For more information please refer to the white paper *EMC Symmetrix VMAX Virtual Provisioning Space Reclamation and Application Considerations*.

Thin metavolumes

Metavolumes or metadevices are utilized for three primary reasons: support for larger devices, easy expansion by adding meta members, and wide striping. The maximum size of a Symmetrix thin device is approximately 240 GB. In order to present LUNs of greater sizes, metavolumes may be utilized. A metavolume consists of up to 255 Symmetrix devices acting as meta members. Depending on the metavolume's configuration, data is distributed different across the meta members.

A concatenated metavolume is organized with the first byte at the beginning of the first device, or meta head. Addressing continues to the end of a device before moving on to the next. When writing to a concatenated device, the first meta member receives all the data until it is full before data is directed to the next member and so on. A concatenated metavolume has the advantage of supporting nondisruptive meta expansions. This is accomplished by adding additional devices at the tail end of the metavolume. This capability is complemented at the vSphere level with the use of the VMFS Volume Grow feature.

A striped metavolume divides each meta member into a series of stripes with a depth of one cylinder. Capacity is distributed by addressing a stripe from each member before starting over from the first device in a round-robin manner. When writing to a striped volume, equal size stripes of data are written alternately to each member of the set. While there are potential performance benefits, striped metavolumes are not as flexible as concatenated metavolumes. Currently, it is not possible to expand a thin striped metavolume.

In thick environments using standard devices, striped metavolumes help maximize resource utilization by spreading I/O workloads across a large base of disk drives and directors on the back end. This can potentially have a very good performance impact. Since wide striping is implemented when thin devices carve 12 track extents out of the data devices in a thin pool, this benefit is largely inherent to Virtual Provisioning already.

In most cases, EMC recommends using concatenated rather than striped metavolumes with Virtual Provisioning.

For more information please refer to the technical note *Best Practices for Fast, Simple Capacity Allocation with EMC Symmetrix Virtual Provisioning* available on Powerlink.

Delivering measured service with FAST

EMC Symmetrix VMAX FAST for standard provisioned environments automates the identification of data volumes for the purposes of allocating or reallocating application data across different performance tiers within an array. FAST proactively monitors workloads at the volume (LUN) level and in order to identify "busy" volumes that would benefit from being moved to higher performing drives. FAST also identifies less "busy" volumes that could be relocated to higher capacity drives, without existing performance being affected. This promotion/demotion activity is based on policies that associate a storage group to multiple drive technologies, or RAID protection schemes, based upon the performance requirements of the application contained within the storage group. Data movement executed during this activity is performed nondisruptively, without affecting business continuity and data availability.

The primary benefits of FAST include:

- Automating the process of identifying volumes that can benefit from Enterprise Flash Drives and/or that can be kept on higher capacity, less expensive drives without impacting performance.
- Improving application performance at the same cost, or providing the same application performance at lower cost. Cost is defined as space, energy, acquisition, management, and operational expense.
- Optimizing and prioritizing business applications, allowing customers to dynamically allocate resources within a single array.
- Delivering greater flexibility in meeting different price/performance ratios throughout the lifecycle of the information stored.

Management and operation of FAST are provided by SMC, as well as the Solutions Enabler Command Line Interface (SYMCLI). Also, detailed performance trending, forecasting, alerts, and resource utilization are provided through Symmetrix Performance Analyzer (SPA). EMC Ionix ControlCenter provides the capability for advanced reporting and analysis that can also be used for chargeback and capacity planning.

Figure 16 shows the relationship between user-created Symmetrix tiers, FAST policies, and storage groups. For more information, refer to the *Implementing Fully Automated Storage Tiering (FAST) for EMC Symmetrix VMAX Series Arrays Technical Note*.

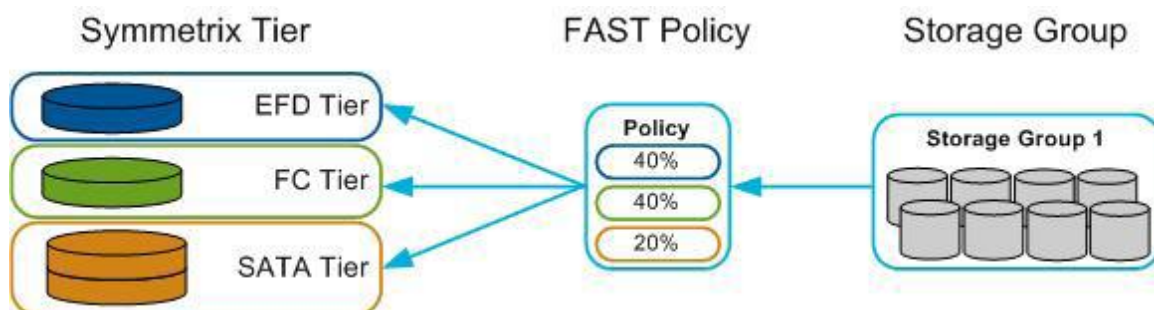


Figure 16. FAST managed objects

Enterprise Flash Drives

With the Symmetrix VMAX support of Enterprise Flash Drives, EMC has created a new ultra-high performance capability that removes previous performance limitations imposed by magnetic disk drives.

For years, the most demanding enterprise applications have been limited by the performance of magnetic disk media. Performance in storage arrays has been constrained by the physical limitations of hard disk drives. Enterprise Flash Drives deliver unprecedented performance and response times, which are benefits well suited for applications.

Enterprise Flash Drives dramatically increase performance for latency-sensitive databases like Oracle. Enterprise Flash Drives, also known as solid state drives (SSDs), contain no moving parts, which remove much of the storage latency delay associated with traditional magnetic disk drives. A Symmetrix VMAX with Enterprise Flash Drives can deliver single-millisecond application response times and up to 30 times more I/O operations per second (IOPS) than traditional Fibre Channel hard disk drives (HDD). Additionally, because there are no mechanical components, Enterprise Flash Drives consume significantly less energy than hard disk drives. When replacing a larger number of HDDs with a lesser number of Enterprise Flash Drives, energy consumption can be reduced by up to 98 percent for a given IOPS workload.

The high-performance characteristics of Enterprise Flash Drives eliminate the need for organizations to purchase large numbers of traditional hard disk drives, while only utilizing a small portion of their capacity to satisfy the IOPS requirements of database. The practice of underutilizing a hard disk drive for increased performance is commonly referred to as short-stroking. Enterprise Flash Drives can increase database

performance and eliminate the need to short-stroke drives, thus keeping storage footprint and power consumption to a minimum and reducing TCO.

How FAST is configured

FAST is configured by defining three distinct objects:

- A **storage group** is a logical grouping of up to 4,096 Symmetrix devices. Storage groups are shared between FAST and Auto-provisioning Groups; however, a Symmetrix device may only belong to one storage group that is under FAST control.
- **Storage types** are a combination of a drive technology (for example, EFD, FC 15k rpm, or SATA) and a RAID protection type (for example, RAID 1, RAID 5 3+1, or RAID 6 6+2). There are two types – static and dynamic. A static type contains explicitly specified Symmetrix disk groups, while a dynamic type will automatically contain all Symmetrix disk groups of the same drive technology. A storage type will contain at least one physical disk group from the Symmetrix but can contain more than one. If more than one disk group is contained in a storage type, the disk groups must be of a single drive technology type.
- **FAST policies** associate a set of storage groups with up to three storage tiers, and include the maximum percentage that storage groups' volumes can occupy in each of the storage tiers. The percentage of storage specified for each tier in the policy when aggregated must total at least 100 percent. It may, however, total more than 100 percent. For example, if the storage groups associated with the policy are allowed 100 percent in any of the tiers, FAST can recommend for all the storage devices to be together on any one tier (capacity limit on the tiers is not forced). In another example, to force the storage group to one of the tiers simply set the policy to 100 percent on that tier and 0 percent on all other tiers. At the time of association, a storage group may also be given a priority (between 1 and 3) with a policy. If a conflict arises between multiple active FAST policies, the Fast Policy priority will help determine which policy gets precedence.

FAST can be configured to operate in a “set and forget” mode (Automatic) where the system will continually gather statistics, analyze, and recommend and execute moves and swaps to maintain optimal configuration based on policy, or in a “user approval” mode (User Approved) where all configuration change plans made by FAST must be approved for a FAST-suggested plan to be executed.

FAST algorithms

FAST uses three distinct algorithms when determining the appropriate tier a device should belong to. The algorithms, in order of priority, are:

- EFD promotion/demotion algorithm
- Capacity-based algorithm
- FC/SATA cross-tier algorithm

The goal of the EFD promotion/demotion algorithm is to maximize Flash drive utilization within the array. When complete, the algorithm will have listed all the devices in the array in order of which devices would be best served to be configured on EFD. FAST will then attempt to place those devices onto Flash drives.

The goal of the capacity-based algorithm is to enforce the FAST policy storage usage rules. A storage group is considered to be in violation when a higher percentage of devices exist on a tier than is configured in the policy for that tier. A storage group is also considered to be out of compliance if devices in the storage group are configured on a tier that is not included in the associated policy—such devices will be reported as being out of policy.

The goal of the FC/SATA cross-tier algorithm is to balance utilization across Fibre Channel and SATA technologies. Devices are sorted by disk service time, and the most utilized devices will be moved to the least utilized disks.

If Optimizer is also enabled on the Symmetrix, then the traditional Optimizer algorithm will be used to balance load within a disk group.

Device movement

There are two methods by which a device will be relocated to another tier: move or swap.

A *move* occurs when unconfigured (free) space exists in the target tier. Only one device is involved in a move, and a DRV (special Symmetrix device used for device swapping) is not required. Moves are performed by creating new devices in unconfigured space on the appropriate tier, moving the data to the new devices, and deleting the old device.

A *swap* occurs when there is no unconfigured space in the target tier, and results in a corresponding device being moved out of the target tier. In order to preserve data on both devices involved in the swap a single DRV is used.

Moves and swaps are completely transparent to the host and applications and can be performed nondisruptively, without affecting business continuity and data availability. Symmetrix metadevices are moved as a complete entity; therefore, metadevice members cannot exist in different physical disk groups.

FAST optimizes application performance in Symmetrix VMAX arrays that contain drives of different technologies. It is expected that customers will have their arrays configured with Flash, Fibre Channel, and/or SATA drives, resulting in storage tiers with different performance levels. Rather than leave applications and data statically configured to reside on the same tier, FAST allows customers to establish the definitions and parameters necessary for automating data movement from one tier to another according to current data usage. The current release of FAST with Enginuity 5874 will move data at the full device granularity.²

² In future releases of Enginuity, FAST will support operations at the sub-LUN level for virtually provisioned devices.

Conclusion

The emergence of cloud computing is enabling organizations to capture the value of virtualization and other enabling technologies while reducing costs as well as physical and carbon footprint. The deployment of new applications is accelerated as resources in pools are rapidly and elastically provisioned and reclaimed. Complementing VMware vSphere, the EMC Symmetrix VMAX is well suited to enable the cloud service provider model, increasing flexibility and simplifying management while maintaining SLAs, performance, and availability.

References

The following documents are available on EMC's Powerlink website:

- *Using EMC Storage Viewer Version 2.0 for vSphere Client* (white paper)
- *EMC Symmetrix VMAX and VMware Virtual Infrastructure* (white paper)
- *Implementing EMC Symmetrix Virtual Provisioning with VMware vSphere* (white paper)
- *EMC Symmetrix VMAX Series Product Guide*
- *Best Practices for Fast, Simple Capacity Allocation with EMC Symmetrix Virtual Provisioning Technical Note*
- *Implementing Fully Automated Storage Tiering (FAST) for EMC Symmetrix VMAX Series Arrays Technical Note*
- *EMC Support Matrix*

VMware vSphere documentation can be found on VMware.com:

- [VMware vSphere 4 documentation](#)
- [VMware vSphere 4 Key Features and Benefits](#)

Information on Terremark's cloud computing services can be found here:

- [Terremark cloud computing services portal](#)