Dell EMC VMAX All Flash

Note: The Dell EMC VMAX All Flash Product Analysis is no longer being updated as of September, 2023 due to a lack of significant product updates.

The Dell EMC VMAX All Flash systems is a continuation of the VMAX line and is offered along with the VMAX³, which supports both flash and spinning disk devices. The VMAX All Flash system, as the name implies, is a high-end enterprise storage system that is optimized for flash in the form of SSDs. Recent generations of the Symmetrix system were:

- 2003 – DMX
- 2009 – VMAX
- 2014 – VMAX³ and 2016 – VMAX All Flash

VMAX All Flash continues the evolution of the Virtual Matrix architecture with HyperMax as the embedded operating environment. As a high-end enterprise class storage array, the VMAX All Flash has extensive reliability and availability features that complement broad set of advanced software features. The VMAX supports IBM mainframe environments, open systems, and IBM System i. File access support is available using the implementation of a hypervisor as part of the embedded operating system and VNX Data Mover and Control Station software as Data Services executing as containers in the hypervisor virtual machines and called eNAS.

The VMAX all flash series includes three models that differ in hardware configurations including processor engines, DRAM cache memory, connectivity, and capacity: the 250F and 950F were added to the existing 450F and 850F models. There are also models with the X suffix that include additional software features only. Competitors for the different VMAX models include IBM's DS8880 and A9000 systems, Hitachi's VSP G1500/F1500 and F800 and Hewlett Packard's 3PAR StoreServ 20000 systems.

CONTENTS

Highlights

- Virtual Matrix Architecture
- HyperMax embedded operating environment
- Scaling to 1,920 drives
- RAID 5 and 6 support across SSDs
- FICON (except for VMAX 250F), Fibre Channel, iSCSI host attach
- Scalability from 1 to 8 VMAX Engines in the largest models (850F and 950F)
- Integrated Controller - Host and Disk interfaces, CPU, and Global Cache\Matrix Interconnect
- Virtual Provisioning (Thin Provisioning)
  - Reclamation using administrator initiated scan for zero data blocks
  - Reclamation with API for software to UNMAP deleted space such as with Symantec Veritas Foundation Suite
- PowerPath multi-pathing support
  - Supports VMware, Hyper-V and Xen
  - Proactive failover management
  - Provides symmetric access, load balances path access through VMs
- Advanced Features
  - SRDF Synchronous, Asynchronous and four site configurations
  - Multiple replication modes
- Detection of unmapped and zero blocks on SRDF transfers
- Support for FC and iSCSI (10GigE or 1GigE) remote replication
  - TimeFinder family for local replication
  - RecoverPoint Write Splitter
  - Federated Live Migration – non-disruptive migration of volumes
  - Encryption of data at rest with FIPS 140-2 certification
- Direct backup of volumes to Data Domain systems with Dell EMC ProtectPoint optional software
- VMware integration
  - VAAI support
  - VASA 1.0 support
  - VMware vCenter storage plug-in
  - VASA 2.0 and VVOL support
- Embedded Unisphere for VMAX as a uniform management solution
- Unisphere 360 software available to manage up to 200 VMAX systems
- File access support – eNAS Data Service
    - Unisphere for VMAX extension with file management dashboard and storage provisioning
    - Software Data Movers and Control Stations configured as Data Service virtual machines

Overview of System Design

The Dell EMC VMAX arrays are multi-platform, enterprise class, modular storage arrays supporting virtualized and physical servers, including open systems, mainframe and IBM System i hosts.

Figure 1: Dell EMC VMAX All Flash System (Source: Dell EMC)

The Dell EMC VMAX architecture continues to advance from earlier generations. The concept of a "VMAX Engine" is the basis for the "Virtual Matrix Architecture." A single engine provides the foundation for the entry-level model, scaling the system by adding engines and racks. The system scales from the single enclosure VMAX All Flash 250F to the 950F with up to 1,920 devices.

The VMAX architecture includes storage resource pools with the advanced abilities to thinly provision volumes, non-disruptively add devices, and redistribute data for capacity and I/O balancing. The storage pool architecture provides the foundation for tiering of data based on different tiers of devices with different characteristics.

Dell EMC VMAX All Flash Models

There are currently unique models of VMAX All Flash system. The models expand the coverage of the market by Dell EMC with VMAX at different capacities and price points.

VMAX All Flash 250F – is an entry system for high-end enterprises for open systems environments with smaller capacity demands but still needing enterprise capabilities. VMAX engines can scale from one to two with 1 or 2 TB of DRAM cache per engine. The VMAX 250F is a 4U rack mountable system. The 250F does not support mainframe environments.

VMAX All Flash 450F – is a high-end enterprise system for open systems and mainframe environments. VMAX engines can scale from one to four with 1 or 2 TB of DRAM cache per engine. All VMAX models come in 19" standard racks and allow mounting of other, third party hardware in the rack. The 450F scales to 2.3PB of raw capacity.

VMAX All Flash 850F – is an entry system for high-end enterprises for open systems environments with smaller capacity demands but still needing enterprise capabilities. VMAX engines can scale from one to eight with 1 or 2 TB of DRAM cache per engine. The 850F scales to 4.4PB of raw capacity.

VMAX All Flash 950F – is a high-end enterprise system for open systems and mainframe environments. VMAX engines can scale from one to eight with 1 or 2 TB of DRAM cache per engine. All VMAX models come in 19" standard racks and allow mounting of other, third party hardware in the rack. The 950F scales to 4.4PB of capacity.

The FX models are the same as the corresponding F models with additional software features included.

Dell EMC VMAX All Flash Model Comparison

Feature / Function	250F	450F	850F / 950F
Number of Engines	1 – 2	1-4	1 – 8
RAID Levels		5, 6	
Device (min – max)	24 - 480	24-960	48 – 1,920
Max usable capacity	.55 PB	1.536 PB	3.584 PB
Device Support – 2.5”		960TB, 1.92TB, 3.84TB, 7.68TB, 15.36TB SSDs	
Connectivity per system			
FICON	N/A	4 - 128	4 – 256
FCoE			
Fibre Channel – 16 Gb/s	4 – 64	4 - 128	4 – 256
iSCSI – 10 Gb/s or Ethernet for NAS	4 – 64	4 - 128	4 – 256
Remote Replication – iSCSI - 10 Gb/s	Shared with iSCSI	2 – 64 FC 4 – 64 iSCSI	2 – 128 FC 4 – 128 iSCSI
Cache (min – max)	512 GB – 4 TB	1 – 8 TB	1 – 16 TB

Table 1: Model Comparison

Product Architecture

The foundation of the VMAX All Flash system is the Virtual Matrix Architecture utilizing the VMAX Engine as the core element. The engines are comprised of multi-core processors, cache, front-end and back-end connectivity allowing the Virtual Matrix to scale from the entry-level configuration with one engine to an aggregate up to 8 engines and 576 processor cores. Each VMAX engine contains two directors, cross-director communication path linking them and redundant interfaces to the Virtual Matrix Interconnect. Each director consolidates front-end, global memory, and back-end functions, enabling direct memory access to data. The linkage between VMAX engines is called Dynamic Virtual Matrix Interconnect, which uses InfiniBand switches or in the case of VMAX 250, direct interconnect, operating at Quad Data Rate - 56 Gb/s. The VMAX All Flash 950F supports 16 ports.

The VMAX has a system bay containing one to eight engines and separate storage bays. The system scales by the addition of engines and drive bays from the entry-level having a single-engine and storage bay configuration to the maximum two systems and eight storage bay configuration. All are 19" rack configurations. System bays can be separated up to 25 meters apart.

Two major design points for the VMAX architecture are the Virtual Matrix Interconnect and Global Memory Architecture. The engines and global memory are connected by the Virtual Matrix interconnect providing dual-active connections to all directors within the system using InfiniBand.

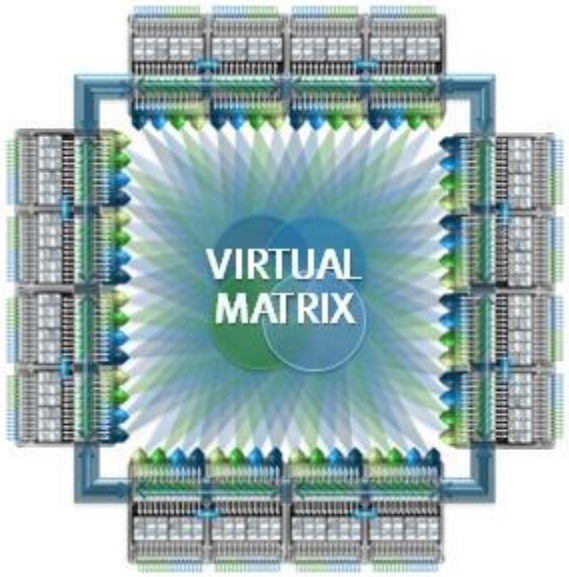


Figure 2: Dell EMC VMAX Virtual Matrix Architecture (Source: Dell EMC)

VMAX All Flash Specifications

Feature	VMAX All Flash 250F	VMAX All Flash 450F	VMAX All Flash 850F	VMAX All Flash 950F
Processor Cores (max)	96	192	384	576
Processor Type	Intel Xeon E5-2650-v4 2.6 GHz 12 core	Intel Xeon E5-2650-v2 3.0 GHz 8 core	Intel Xeon E5-2697-v2 2.3GHz 3.0 12 core	Intel Xeon E5-2697-v4 2.3GHz 18 core
Engines (min – max)	1 – 2	1 – 4	1 – 8	1 – 8
Memory	512 GB – 4 TB	1 – 8 TB	1 – 16 TB	1 TB – 16 TB
Max Capacity usable	1.1 PB	1.536 PB	3.584 PB	3.584 PB
Interface Ports	64	128	256	256
Dell EMC Measured Bandwidth	N/A			150 GB/s
Dell EMC Measured IOPs	1M @ 200 µsec			4M @ 500 µsec

Table 2: VMAX All Flash specifications

Data is available to all processors within each engine director-pair when a write is committed to memory. Processors in all director-pairs can access the data in memory autonomously without de-staging to disk and re-staging to a separate region in memory.

The multi-core processors can be dedicated to front-end ports, the HyperMax Operating System, or backend ports allowing the performance to be adjusted based on the I/O profiles, overriding the default setting for core mappings.

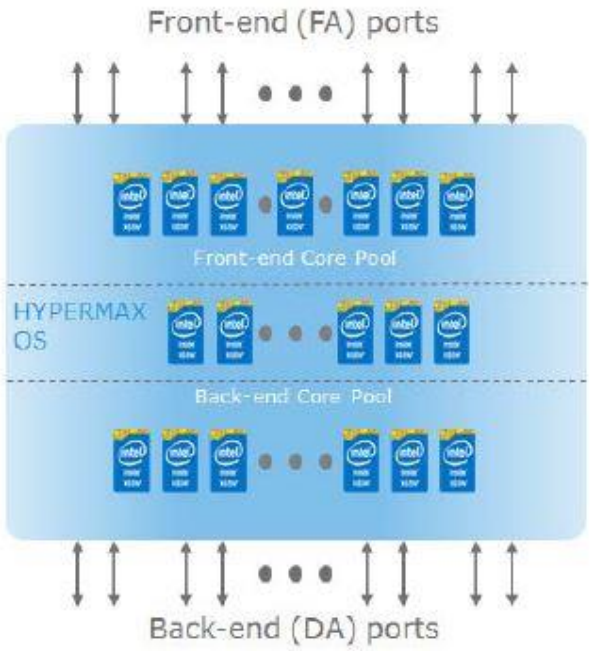


Figure 3: Multi-core Mapping source: Dell EMC

Devices

The VMAX All Flash system supports Flash SSD only. Each storage bay can have up to a max of 960 devices per storage bay. The maximum system configuration is 1,920 devices. Device Enclosure components contain devices (120 2.5" SSDs), link control cards and power and cooling elements supporting dual-ported SAS attached devices. Device Enclosures are connected to backend adapters with 12 Gb/s SAS with the 950F. The VMAX 250F uses 12 Gb/s connections. The VMAX 250F has a 25-slot Device Enclosure with a base capacity of 11 PB and a maximum of 100 devices.

Management Module Control Systems (MMCS)

The Service Processors in earlier Symmetrix systems have been replaced with two MMCSs (two processor boards) in system bay 1 for management access including remote, remote recovery, and hardware diagnostics.

Reliability, Availability, Serviceability Features

Dell EMC VMAX components are fully redundant, including directors, data paths, power supplies, standby power supplies and all back-end components. Internal redundancy includes multiple power supplies connected to independent power sources, battery backup of all power systems ensuring data in memory is written to disk prior to orderly shutdown in the event of a complete power loss, permanent spares, and environmental monitoring.

HyperMax upgrades and updates take advantage of the multiprocessing and redundant architecture. Release levels can be loaded non-disruptively online.

The VMAX utilizes global memory mirroring for protection from memory component failures. A single logical image of memory is actually two physical images. Cache integrity checks include Error Checking and Correction (ECC) and all cache memory locations are periodically verified by memory-correction routines. Pre-configured vault drives, designated disk storage, are used to de-stage data from global memory in the event of a power down or failure. Standby power supplies protect global memory and data from global memory is de-staged redundantly to vault devices. On power up global memory is restored from the vault devices. Normal operation resumes when the standby power supplies are sufficiently recharged to support another vault operation. Vault drives may be EFDs or HDDs and there is not minimum number of drives.

Replacement of a failing drive with a spare standby is by automated permanent sparing. Recommendations include two spare drives per 100 drives, per drive type with a minimum of eight spare drives in a system.

Embedded Unisphere for VMAX is available as the default management console. Unisphere for VMAX provides a common user experience across Dell EMC storage systems. Unisphere for VMAX supports REST APIs, which enables consolidated management as would be used in cloud environments or orchestration software. Much of the self-service capabilities for cloud environments are implemented through the REST interface of VMAX.

Unisphere 360 is separate software that runs on a server or virtual machine and can aggregate the management for up to 200 VMAX systems.

HyperMax Operating System Overview

Beginning with VMAX³ and continuing with VMAX All Flash, the embedded operating environment HyperMax changes the fundamental operation of enterprise family formerly known as Symmetrix. HyperMax provides a storage hypervisor with internal virtual machines. A virtual machine will run the storage system control function, advanced features, potentially specialized applications. Storage function such as block storage, management, data protection, and other future advanced capabilities are the primary focus for storage. HyperMax is also responsible for the dynamic allocation of resources in meeting defined performance levels for storage pools.

Resources virtualized by HyperMax include the storage volumes presented to hosts. HyperMax supports non-disruptive upgrades for both firmware and hardware.

HyperMax continues the features of Dell EMC TimeFinder and Symmetrix Remote Data Facility (SRDF). It also supports storage protocols and applications and enables simultaneous connection to mainframe, UNIX, System-i, Windows and Linux platforms.

Data Services

A hypervisor is included with the HyperMax OS to allow data services functions to execute as containers internal to VMAX. File access support is a data service that is adapted from the VNX file access support where data movers and control stations can execute as containers within the HyperMax OS on front-end adapters. The file services are called eNAS.

Evaluator Group Comment: Additional data services are expected to be added over time. It is unclear whether any will execute in backend adapters (BA).

Dynamic Cache Partitioning

This partitioning allows allocation of as many as 16 partitions of cache for specific application groups. These partitions can be static or dynamic, expanding or contracting according to performance criteria for the specific workloads among applications. More resources allocated for cache friendly workloads increase utilization and performance. Non-cache effective workloads can be fenced in smaller partitions reducing the utilization of the aggregate workload.

VMAX Priority Controls

Up to 16 user-defined priority levels can be set to define different priorities to device groups allowing higher-priority applications better response times during peak utilization periods. This enhances tiered storage management for read I/O and SRDF/S transfers.

Host I/O Limits

With VMAX Host I/O limits, maximum performance quotas can be set on a per storage I/O group basis. The settings can be the maximum IOPs or bandwidth or a combination to ensure that a single application does not dominate the storage system resources and impact other operations. Setting limits is a management capability in multi-tenancy operations for service guarantees. The volumes in a storage group share the same performance quota. The quota is evenly split between all directors in the storage group's masking view. When reaching a limit on a storage group, the VMAX will throttle I/O to that group to allow I/O to other groups to proceed.

Enhanced Virtual LUN

Enhanced Virtual LUN allows the relocation of data to different storage tiers, devices, and RAID types non-disruptively with no impact to replication. Upon completion of the transfer, the previously occupied physical storage space will be returned to the free pool. Up to 16 virtual migrations are supported. The new enhancement also removes impacts on remote replication, which previously was suspended during data migrations.

Data At Rest Encryption

The VMAX All Flash system uses hardware encryption on the backend SAS I/O modules and RSA management infrastructure for data at rest encryption on all attached devices. With the RSA key manager being available both internal and external to the VMAX, all key management is handled without additional administrator overhead. This allows key management to be selected to be internal to the VMAX system or to an external RSA key management system. Both selections are KMIP standard compatible.

It should be noted that this type of encryption does not use encrypting devices. Encryption of all data on each device addresses security concerns for data in the event when a device is removed from the system. The encryption must be enabled at installation time and cannot be turned on later. NIST has certified the encryption to meet FIPS 140-2.

Service Level Provisioning

An application is defined by a set of logical volumes (thin devices in the prior diagram) as a storage group, which is associated with a predefined Service Level Objective. Application data for those logical devices are allocated resources across the storage devices in different device groups. The default Service Level Objective (SLO) is Optimized, which has not explicit response time target. Other SLO settings (Diamond, Platinum, Gold, Silver, Bronze) have defined average response time targets for each. Administrators will choose the setting when provisioning storage (volumes) for an application. The I/O size, replication – remote or local, and read/write ratio will vary the actual response time.

Space Reclamation

VMAX HyperMax includes APIs (Application Program Interfaces) to support the UNMAP SCSI commands allow system elements such as databases, file systems, and volume managers to release unneeded blocks for a volume. An example of usage of the API is the Veritas Foundation Suite. Additionally, the storage administrator can invoke a storage system task called symconfigure or use Solutions Enabler to scan thinly provisioned volumes for chunks of all zero data written as part of pre-allocation by databases or file systems to return that space to the pool.

Dell EMC has also written utility software called "storreclaim" that can be executed on Windows and Linux systems to scan filesystems for deleted files and issue SCSI UNMAP commands to VMAX to reclaim the deleted space.

Inline Compression

Inline compression for VMAX all flash is selectable at the storage group level. Data stored that is not compressed may be compressed as a background task. The system automatically selects data that is not highly active to be compressed within the storage group. Highly active data will not be compressed. Compression can work independently of all other VMAX software features. A pretest of data is performed and if data is not compressible, it is left uncompressed. A compression estimator tool is available from Dell EMC.

Cloud Tiering

Integrated in the VMAX embedded software is the ability to tier data to cloud locations/systems over Ethernet using an S3 protocol. Supported cloud targets include the VirtuStream cloud and Elastic Cloud Storage object systems.

VMware VAAI Support

The VMAX supports the VMware vSphere APIs for Array Integration (VAAI), which began with vSphere 4.1. VAAI is a set of mechanisms that allow processing for certain data-related services—copying data when creating a new VM, for example—to be offloaded from the ESXi host to a storage array. The intent of these APIs is to streamline the functioning of the ESXi server and speed-up delivery of storage-supported services.

- Full copy — Enables the storage system to make full copies of data within the storage system without having the ESXi host read and write the data.
- Block zeroing — Enables storage systems to zero out a large number of blocks to speed provisioning of virtual machines.
- Hardware-assisted locking — provides an alternative means to protect the metadata for VMFS cluster file systems, thereby improving the scalability of large ESXi host farms sharing a datastore.

Beginning with vSphere 5.0, enhancements added for thin provisioning include:

- Dead Space Reclamation — Reclaims blocks from VMFS deleted files. Physical space is free with the SCSI Unmap command.
- Out of Space Condition Support — if disk space is exhausted, a virtual machine is paused as reported by the storage system. This allows administrators to mitigate the situation rather than causing the virtual machine to fail.

VMware VASA Support

The VMAX supports the VMware vSphere APIs for Storage Awareness (VASA), which is in vSphere 5.0. VASA can display the features of the physical storage devices, allowing vSphere administrators insight into storage capabilities. The VMAX has a plugin called a vendor providers, which is the integration 'glue' that sits between vCenter and the storage array. Vendor providers retrieve the storage capabilities from the array and pass these onto vCenter, which in turn can display these capabilities in the user interface. VASA will also provide information about storage health status, configuration information and capacity.

VVOLs and VASA 2.0 support are available and include up to 64,000 VVOLs.

Advanced Features

SRDF/EDP

SRDF (Symmetrix Remote Data Facility) with the VMAX includes a feature called "Extended Distance Protection" or SRDF/EDP which is a variation of a four site or bunker type of replication similar to Cascaded SRDF (disk or diskless). With SRDF/EDP, synchronous replication occurs to a nearby pass-thru system, which requires no hosts. Due to synchronous replication, the RPO is essentially zero or "no data lost", which operates by caching data enroute to a third SRDF replication system.

With the newest SRDF, a limited set of WAN compression occurs with SRDF and the total number of SRDF replication groups is 256. The highlights of SRDF include:

- Family of products that offers various levels of remote replication solutions to maintain copies of data on another VMAX at same or physically separate location
  - Three base product options
    - SRDF/Synchronous (SRDF/S)
    - SRDF/Metro – active-active stretched clusters between two sites.
    - SRDF/Asynchronous (SRDF/A)
    - SRDF/Data Mobility (SRDF/DM)
  - Four optional add-on products
    - SRDF/Automated Replication (SRDF/AR)
    - SRDF/Star
    - SRDF/Consistency Groups (SRDF/CG)
    - SRDF/Cluster Enabler for MSCS and VCS
- Supports connectivity over Fibre Channel, Gigabit Ethernet or 10/1 Gigabit Ethernet
- Can be integrated with TimeFinder for local replication
- Support of IBM Geographically Dispersed Parallel Sysplex (GDPS)
- GDDR AutoSwap provides continuous availability for storage systems and GDDR provides DR automation in two and three site mainframe configurations.

SRDF Overview

Dell EMC was first to market with a remote replication alternative for monolithic storage systems when it introduced the initial version of SRDF in April 1995. Since then it has evolved into a family of software products for Dell EMC's VMAX and earlier Symmetrix disk arrays that provides various levels of remote replication solutions for disaster recovery and business continuance.

SRDF consists of a combination of HyperMax software (microcode) and optional hardware components in the VMAX. The current SRDF family consists of three base products and five add-on options. Each base product as well as add-on option will require a license key for each system that it will be running on.

SRDF Base Products

VMAX customers can select one or more of the following base SRDF products. For a detail explanation of the SRDF capabilities and architectural design, Dell EMC reference documents should be consulted. A quick overview of SRDF follows here.

- SRDF/Synchronous (SRDF/S): SRDF/S provides synchronous remote replication from one VMAX to one or more VMAX systems at one or two different locations.
  - Synchronous mode transfers data to a remote VMAX and waits for completion response before acknowledging write completion to host.
  - Adaptive Copy mode will accumulate writes in cache or on disk and a background process will destage the writes to disk and send the writes to the secondary (remote) system. A maximum is set to limit the amount of writes accumulated.
  - Non-disruptive migration is supported with SRDF/S to or from other VMAX systems.
- SRDF/Asynchronous (SRDF/A): provides periodic asynchronous replication using Dell EMC's Delta Set architecture
  - Asynchronous mode will indicate completion after data is written to the local (primary) volume and before the completion response is received from the remote (secondary) system. With Delta Sets, sets of data are transferred to the remote site.
  - SRDF/AR – Multi-Hop mirroring allows SRDF users to mirror data that has changed since the last update onto a VMAX system in a third location.
- SRDF/DM (Data Mobility): provides asynchronous remote replication for the purpose of distributing or migrating content
- SRDF/Metro: a variation on the SRDF/S implementation, SRDF/Metro provides an active-active stretched cluster implementation for business continuity and dynamic workload migration. SRDF/Metro can have either another VMAX system as the third party witness or a Virtual Witness. A Virtual Witness is an application running in a Virtual Machine to be the arbiter between systems in a failover and takeback scenario. A third site copy is supported using asynchronous replication. The primary site to secondary site uses synchronous replication.

Optional for SRDF

One or more of the following can be added to the above base products.

- SRDF/Star: provides multi-site remote replications solutions by supporting synchronous as well as asynchronous replication from the same source volume.
- SRDF/Automated Replication (SRDF/AR): provides support for multi-hop configurations using combinations of TimeFinder/Mirror and SRDF/S or SRDF/DM.
- SRDF/Cluster Enable for MSCS or VCS: supports failover through storage-based replication and server clustering via SRDF/S and MSCS or VCS.
- SRDF/Consistency Groups: maintains data consistency for write dependent data across multiple volumes and/or multiple VMAX systems.

- SRDF/S – SRDF/A Mode Change: enables the customer to dynamically switch between SRDF/S and SRDF/A modes of operation.

SRDF Prerequisites

Mainframe Enablers is a prerequisite for the SRDF family of products in mainframe environments and is downloadable through Powerlink at no charge.

SRDF Licensing

Licensing for each of the SRDF products is based upon the registered or raw capacity of the VMAX on which they are running.

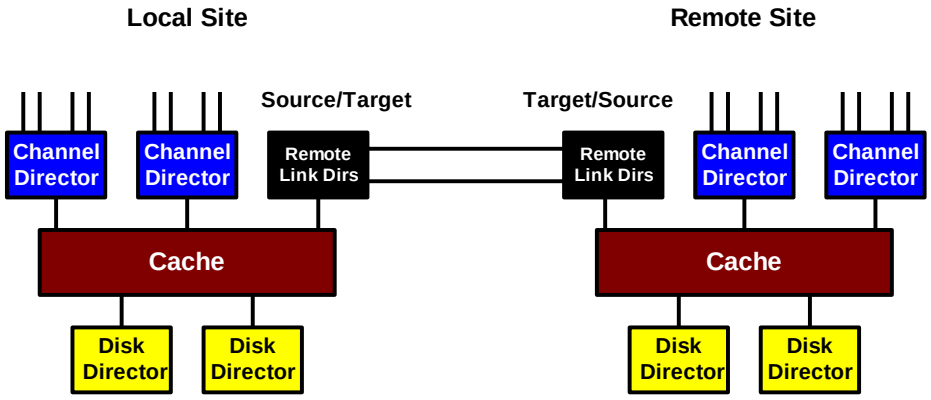


Figure 4: Basic SRDF Architecture

SRDF Director Hardware

SRDF director hardware consists of a director that provides the link connections, fiber-optic protocol support, and communications control between the VMAX systems. SRDF uses a storage protocol based on the Fibre Channel or Gigabit Ethernet (1Gb/s and 10Gb/s) specifications. The host attachment, I/O protocol, and disk data structures that each host requires are independent of the SRDF operation between systems.

SRDF Volume Types

SRDF refers to VMAX volumes (or devices) as:

- Primary or Source volumes (R1)
- Secondary or Target volumes (R2)
- Local volumes
- Dynamic SRDF devices
- PPRC mode devices
- XRC mode devices

SRDF Primary or Source Volumes (R1)

A Primary or Source volume contains production data that is remotely mirrored in one or more different VMAX systems. They are also sometimes referred to as R1 volumes. Primary volumes are locally protected by RAID and direct sparing. Updates are automatically propagated to a Secondary (Target) volume in the other system(s). Primary volumes can also be paired with a TimeFinder Clone or Snap to provide an additional working copy of the data at the same location.

SRDF Secondary or Target Volumes (R2)

Secondary or Target volumes contain the remote copy of data from a Primary volume. They are also sometimes referred to as R2 volumes. Secondary volumes can also be locally protected by RAID and direct sparing, and they can also be paired with a clone or a snap to provide an additional working copy at the same location.

Local Volumes

A Local volume is a volume on a VMAX that is not participating in SRDF activity. They are normally protected by TimeFinder or RecoverPoint CDP as well as RAID and direct sparing. SRDF can be used simultaneously with RecoverPoint CDP volumes.

Dynamic SRDF Devices

Dynamic SRDF is supported over Fiber Channel point-to-point, switched Fibre channel, and GigE/10GigE connections. Dynamic SRDF provides the following capabilities:

- Personality swap between Primary and Secondary volumes
- Terminate and re-establish a relationship with a new Secondary volume
- Create a new Primary/Secondary pair relationship from non-SRDF devices

Dell EMC Compatible Peer

A Dynamic SRDF Device will become a PPRC or XRC mode device upon receipt of a PPRC CESTPAIR command. Once the device enters PRC mode, no host-based SRDF control software can operate on the device, and it can only be controlled by PPRC or XRC commands received from the host.

SRDF Groups

Logical volumes can be assigned to SRDF Groups, which are used to define relationships between systems. An SRDF Group is a set of SRDF director port connections configured to communicate with another set of SRDF director ports in another system. An SRDF Group that is configured through VMAX configuration is called a Static Group. Each system can support up to 256 Static Groups.

Users can also dynamically create empty SRDF Groups, and dynamically associate the Groups with Fibre Channel or GigE SRDF directors. It is also possible to remove dynamic SRDF Groups. However, dynamic groups are not supported for PPRC. When combined with dynamic SRDF devices, the user has complete control over the SRDF configuration.

Special SRDF Volume Considerations

There are some special considerations for volumes when used in an SRDF environment.

Dynamic Sparing with Remotely Mirrored Pairs

When the Direct Sparing option is invoked for a remotely mirrored SRDF pair, the VMAX automatically activates an available spare in the VMAX containing the failing device and copies data to the spare. The system will continue processing I/O requests with the spare functioning as one volume of a mirrored pair while the failing device and its remote mirror operate without interruption. If all of the data cannot be copied from the failing device to the spare, it will retrieve the unavailable data from the good member of the remote pair.

VMAX RAID-10 (mirrored striped mainframe volumes with DMSP)

RAID-10 is supported only in mainframe environments, and is intended to improve performance by striping data of a logical device across four logical devices.

PPRC (Metro Mirror) Command Support

VMAX systems to support native IBM Peer-to-Peer Remote Copy (PPRC) commands through the Compatible Peer feature, and support for PPRC Version 1, architectural levels 3 and 4 (Hyper-swap) which allowed the support of IBM's Geographically Dispersed Parallel Sysplex (GDPS). Compatible Peer is available on VMAX systems with connections to FICON hosts.

SRDF Link Configurations

The links between a pair of VMAX systems in an SRDF configuration can use one of the following methods for transmitting data.

Unidirectional

When all Primary volumes reside in one VMAX system and all Secondary volumes reside in another system, write operations move in one direction, from Primary to Secondary. Data moves in the same direction over every link in the SRDF group.

Bidirectional

When an SRDF group contains both Primary and Secondary volumes, write operations move data in both directions over the links for that group.

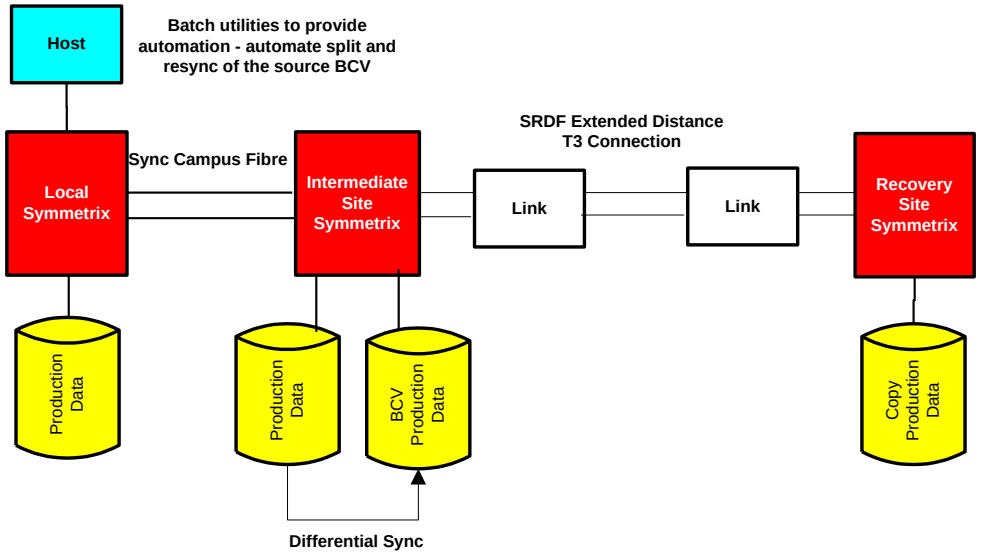


Figure 5: SRDF Multi-hop Operation

GDPS Support

Dell EMC announced SRDF support of IBM's Geographically Dispersed Sysplex (GDPS) mainframe clustering solution in November 1999. Automatic site failover can be invoked via compiled REXX execs. An Dell EMC supplied STARTIO exit traps PPRC commands and issues the SRDF equivalents. RCMF, PDAS and DEVSERV are fully supported.

PowerPath

PowerPath provides improved load balancing for virtual environments than previous versions. It supports VMware but also HyperV and Xen hypervisor environments. Load balancing across multiple active paths is now possible. PowerPath highlights include:

- Supports up to 32 paths (16 for CLARiiON/VNX) to a device
- Supports heterogeneous storage platforms
- Supports AIX, HP-UX and Tru-64, Linux, Solaris, and Windows server platforms
- Supports SCSI, iSCSI and Fibre channel for both switched fabric and arbitrated loop environments
- Provides automatic detection and recovery of failed paths
- Provides dynamic load balancing across multiple paths
- Provides volume management capabilities
- Provides data mobility across heterogeneous storage platforms
- Options include PowerPath, PowerPath Base, PowerPath iSCSI and PowerPath/SE

PowerPath Overview

Originally introduced in February 1998, Dell EMC PowerPath is server resident software that provides path management between storage systems and operating systems. It also provides an integrated logical volume manager, PowerPath Volume Manager (PPVM), which became generally available in June 2003 with the release of PowerPath V4.0. PowerPath supports Fibre channel, iSCSI and SCSI interfaces.

PowerPath resides on the server and uses the SCSI protocol with Ethernet, SCSI or Fibre Channel attachments. The application issues I/Os to a logical device and PowerPath manages the path selection to the physical device.

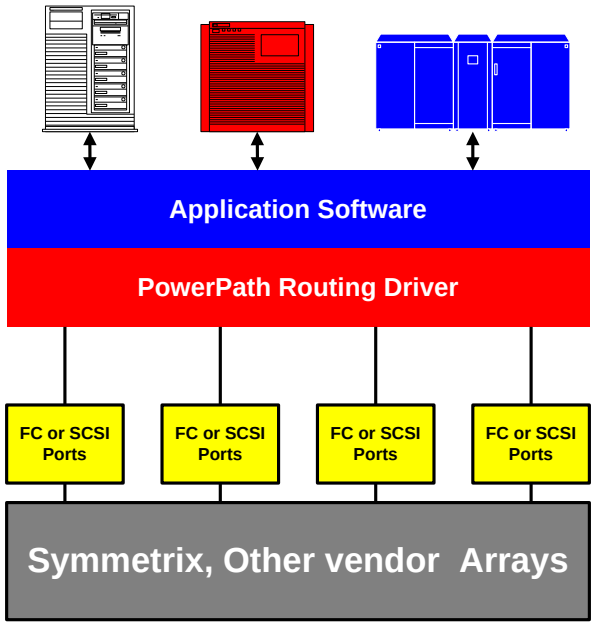


Figure 6: PowerPath Software Diagram

Multiple Ports

The user can configure a logical device as a shared device using two or more interfaces. Examples of an interface include a Fibre Adapter on a VMAX system, or a Storage Processor (SP) on a CLARiiON or VNX system. This allows all logical devices to be visible on all ports.

PowerPath works with two types of storage systems: active-active and active-passive. A host that is part of a cluster cannot have both active-active and active-passive storage devices in the same disk group.

Active-active

Active-active means all interfaces to a logical device are active simultaneously. Examples of active-active storage systems are the VMAX, Hitachi VSP G1000, and the HP 3PAR. If there are multiple interfaces to a logical device, they all provide equal access to the logical device.

Active-passive

With an active-passive system, if there are multiple interfaces to a logical device, one is designated as the primary route to the device, and that device is assigned to that interface card. Normal access to a device through any interface other than its assigned one is either impossible (for example, on VNX systems), or possibly much slower than through the assigned interface card.

In the event of a failure of an interface card or all paths to an interface card, the logical devices must be moved to another interface. If an interface card fails, logical devices are reassigned to another interface. This reassignment is initiated by the other, functioning interface, and is referred to as "trespassing". Trespassing can take several seconds to complete. However, the I/Os do not fail during this process. After the devices are trespassed, PowerPath detects the changes and seamlessly routes the data via the new route.

If PowerPath's periodic autorestore feature is enabled, logical devices can be reassigned to their original assigned interface. Otherwise, the reassignment can be activated manually.

Path Sets

All paths to the same logical device are grouped into a "path set". PowerPath creates a path set for each logical device, and then populates the path set with all usable paths to that logical device. In an active-active system, any path in the set can be used to service an I/O request. In the event of a path failure, PowerPath can redirect an I/O request to any other viable path in the set. This redirection is transparent to the application.

In an active-passive system, path sets are divided into two "load balancing groups". The active group contains all paths to the interface to which the logical device is assigned. The other group contains all paths to the other, non-assigned interface. Only one of these load balancing groups can process I/O requests at a time, and PowerPath load balances the I/O activity across all paths in the active group. In the event of a failure of a path in the active group, PowerPath redirects the I/O request to another path in the active group. If all paths in the active group fail, PowerPath reassigns the logical device to the other interface, and then redirects the I/O request to a path in the newly activated group.

Automatic I/O Path Failure Detection

PowerPath provides automatic detection of failed I/O paths. In the event the server loses a Host Bus Adapter (HBA), storage device, or a cable, the I/O request is completed through another available path. The PowerPath "multipath module" is responsible for selecting the best path to a logical device. It detects failed paths and retries failed application I/O requests to other paths.

A "path test" is used to determine if a path is operational. After a path failure, PowerPath continues to test it periodically. If the path passes a test, PowerPath restores it to service and resumes sending I/O to it. PowerPath also periodically tests "live paths". This process is invoked if the path has not been tested for at least one hour, or has not been used within the last minute.

Dynamic Load Balancing

The dynamic load-balancing feature ensures that no single path can become overloaded while other paths have underutilized activity, causing a bottleneck. When one or more paths become busier than others, PowerPath shifts the I/O traffic from the busy paths to the less busy paths. Users can choose from five types of load balancing and two types of maintenance policies.

For example, the VMAX/VNX optimization and HDS Adaptive load-balancing policies enables each path to be chosen intelligently for every I/O request, resulting in an approximately equal load on all paths. These policies also allow the setting of device priorities.

Application Tuning

PowerPath can be used to tune application performance by assigning priorities to logical devices, or manual load balancing with channel groups.

Logical Device I/O Priority

Logical Device I/O Priority allows the user to tune the bandwidth for specific logical devices. For example, the user could assign a high priority to logical devices running critical applications, and a lower priority to devices used by less critical applications.

Policy	Description	Characteristics
VMAX/CLARiiON/VNX Optimization	Basis of path load and device priority (default)	▪ Distributes I/O based on estimated completion time ▪ Calculates depth of each queue in units of time (queue with shortest time gets the request) ▪ Considers sequential activity, channel speed, cache read and write speeds, volume loading, channel groups
Adaptive	Basis of path load and device priority (default)	▪ Distributes I/O based on estimated completion time ▪ Calculates depth of each queue in units of time (queue with shortest time gets the request) ▪ Considers sequential activity, channel speed, cache read and write speeds, volume loading, channel groups
Round-robin	Each available path in rotation	▪ Does not select optimal path ▪ Synchronous behavior can impact performance ▪ Lockstep behavior in Fibre channel can impact performance ▪ Single physical device can impact all paths ▪ Does not optimize sequential activity for a path
Least I/Os	Path with fewest pending I/O requests	▪ May not select optimal path when used with mixed block sizes ▪ Large number of small block requests may adversely impact performance ▪ Does not optimize sequential activity for a path
Least Blocks	Path with fewest blocks	▪ May not select optimal path when used with mixed block sizes ▪ Large number of small block requests may adversely impact performance ▪ Does not optimize sequential activity for a path
Basic Failover (CLARiiON / VNX Only)	Fails over the storage processor only	▪ Required for non-disruptive upgrade ▪ Protects against storage processor and back-end failures

Table 3: PowerPath Load Balancing Policies

Channel Groups

The user can form a "channel group" of dedicated paths to a logical device to increase application performance. However, the user must keep in mind that reserving paths for one application makes those paths unavailable for other applications, which can impact their performance. Channel groups maintain a second set of paths in reserve in the event that the first set fails.

Maintenance Policy	Characteristic
Request	Used for diagnostic purposes in a failover environment only
No Redirect (VMAX only)	Used to disable path management (during maintenance mode)

Table 4: PowerPath Maintenance Policies

Dynamic Reconfiguration

PowerPath can automatically resize or move a volume. If a logical volume is expanded, it recognizes the growth and automatically expands the corresponding volume group.

PowerPath Volume Manager (PPVM)

PowerPath Volume Manger (PPVM) offers the user an integrated logical volume manager and volume mobility. It is a host-resident, logical volume manager that provides the user with a flexible view of data, independent of the underlying physical storage. Using PPVM the user can group physical storage devices that share a common purpose or configuration into logical pools of storage called "volume groups". The user is then able to use these volume groups to manage the physical storage independently of how users and applications access the data. An example is the ability to dynamically add additional storage resources to a volume group in response to an application's need for more storage capacity.

Using PPVM in conjunction with PowerPath provides the user a higher level of data security that is not available when using PPVM in combination with other path managers, such as Symantec Veritas DMP. However, the user does not have to use PPVM with PowerPath. It will work with other path managers, such as VERITAS' path management product, DMP.

Host Platform Support

PowerPath Volume Manager is supported only on HP-UX. AIX, Windows, and Sun Solaris platforms.

Coexistence with Third-Party Volume Managers

PPVM can coexist on the same host with third-party volume managers. However, each must manage a distinct set of physical storage devices.

Volume Mobility

The Volume Mobility extension enables the user to move, or to change the structure of a PPVM volume without application disruptions. This feature is not available with a PowerPath Base license.

Evaluator Group's perception of the strengths and potential concerns of the Dell EMC PowerPath product are as follows.

Perceived Strengths - The support of up to 32 paths is greater than many competitive offerings.

Potential Concerns - Dell EMC has stated that failover can only occur on like interfaces (i.e. SCSI to SCSI or Fibre to Fibre). This could be an issue to a small set of users.

TimeFinder and SnapVX

TimeFinder was first introduced in April 1997. It provides a means of creating duplicate, or point-in-time, copies of an entire logical volume or a dataset on either the same subsystem, or another system that is part of a Symmetrix Remote Data Facility (SRDF) configuration.

Since its introduction in 1997, Dell EMC has made numerous enhancements and now offers the TimeFinder Family of products, which consists of the following components.

- **TimeFinder/Mirror:** create copies of standard volumes
- **TimeFinder/Snap:** create pointer-based copies that utilize significantly less space than TimeFinder/Mirror copies
- **TimeFinder/Consistency Groups:** enables other TimeFinder family products to create multi-volume sets of copies that are consistent
- **TimeFinder/Exchange Integration Module:** automates the process of creating and managing TimeFinder operations in a Microsoft Windows Exchange Server environment
- **TimeFinder/SQL Integration Module:** automates the process of creating and managing TimeFinder operations in a Microsoft windows SQL Server environment
- **TimeFinder /SnapVX:** space efficient snapshots of thinly provisioned volumes (from thin storage resource pools) can share extent allocations within the thin pool. SnapVX supports up to 256 snapshots per source drive. SnapVX represents a new storage pooling implementation that allows for thin provisioning and redirect-on-write snapshots. Additionally, SnapVX provides Secure Snapshots where a snapshot will be retained until a retention period has expired. This prevents inadvertent or intentional deletion and provides business governance over data copies.

Competitive alternatives for cache centric systems include the Hitachi Data Systems ShadowImage, HP Business Copy, IBM FlashCopy, and other snapshot products.

TimeFinder Highlights

- Means of creating a point-in-time copy of a logical volume and/or a dataset for backup and/or application testing
- Pointer-based replication to minimize space requirements
- Data consistency across volumes and/or systems

TimeFinder Overview

Business Continuance Volumes (BCVs)

The point-in-time copies are maintained on a logical volume(s) that are referred to as Business Continuance Volumes (BCVs). The BCVs are classified as either mirrors, snaps, or clones. There must be one set of BCVs for volume level duplication and another set for dataset level duplication. They can be switched back and forth. The major difference is that the BCVs used for dataset level duplication are on-line to the host and controlled by standard host allocation and management routines such as SMS and/or HMS. The number of BCVs will depend upon the requirements of each environment, and requires careful consideration since they cannot be changed dynamically (changed by the customer).

Volume-level Duplication

The process for creating a duplicate point-in-time copy of a logical volume is as follows:

1. Establish a relationship between a standard volume and a logical volume in the BCV pool through the SRDF host component or Unisphere
2. Once the relationship is established, VMAX begins to create a mirrored copy of the standard volume with all updates written to both the standard volume and the BCV.
3. When they are fully synchronized, notification will be given
4. The BCV can be accessed (or split) once the first block of data has been written to the target

A record of all updates to the volume is maintained after the BCV has been split. This allows the BCV to be resynchronized at a later time.

Dataset-level Duplication

Dataset-level duplication is controlled by a Dell EMC supplied Snap utility that is very similar to the IBM SIBBATCH utility for Snapshot. Dell EMC states that the same type of syntax is utilized, and IBM Snapshot customers can implement the Dell EMC Snap utility with little or no JCL changes unless the customer uses the Dell EMC wildcard feature, which is not supported with Snapshot.

The allocation and management of a duplicate dataset on a BCV is under the control of the MVS host. This includes re-catalog and rename operations. With no support for resynchronization for dataset level duplication, this eliminates many of the management issues associated with dataset level duplication, such as insuring that sufficient space exists.

Configuration Restrictions

Dell EMC has defined the following configuration restrictions for TimeFinder.

1. Any BCV can be assigned or reassigned to any standard volume provided no other active BCV mirror exists for that volume
2. RAID-1, RAID-S or SRDF must protect a standard volume
3. The standard volume and BCV must have equal track geometry and an equal number of cylinders
4. A maximum of three concurrently active BCVs per production volume

RecoverPoint CDP and CRR

Evaluator Series Research Data Protection section describes RecoverPoint as a replication and data protection solution. It is an integral part of protection for Dell EMC storage. Supported storage systems include XtremIO, Unity, VNX, VMAX, and VPLEX.

RecoverPoint Splitter

RecoverPoint data protection and replication requires a write splitter implemented in the array, as a host driver, or in the fabric switch. VMAX includes array-based write splitter capability on the VMAX. The splitter supports replication from smaller devices to larger devices. There are some special characteristics or limitations regarding VMAX write splitter. Dell EMC documentation will provide complete details but a few notable ones are:

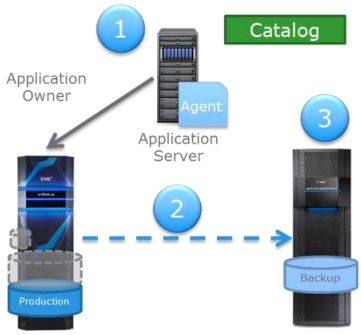
- LUN maximum size is 32TB
- iSCSI and FCoE connections are not supported
- Direct (switchless) configurations are not supported
- 4,096 Open Replicator sessions are supported
- Each volume can be attached to only one write splitter
- Support for VAAI hardware-assisted locking commands

Dell EMC ProtectPoint

VMAX enables direct backup of data with the optional software feature ProtectPoint. With ProtectPoint, agents on application servers allow application administrators to direct the VMAX system to backup volumes directly to a Data Domain system. The VMAX uses changed block tracking to send only changed blocks (from the last full backup) to the Data Domain system. The Data Domain system will create a full backup from the changed data. Application administrators can direct the VMAX to restore a volume or a specific object in a volume and VMAX will communicate directly with the Data Domain system to perform the restore.

EMC ProtectPoint (optional feature)

- Integrated data protection
- Direct backup of changed data to Data Domain
  - Initiated by application owner
  - VMAX sends data over FC to Data Domain – changed data only with CBT
  - Data Domain creates full backup in native format (LUN backup)
- Application owner directs restores
  - VMAX reads backup image data from Data Domain
  - Can recover specific object – Instant Access
- Integrated with EMC Data Protection Advisor



1. App owner triggers backup at an application consistent checkpoint
2. Primary storage sends changed blocks directly to Data Domain
3. Data Domain uses the changed blocks to create full backups in native format

Figure 7: Dell EMC ProtectPoint VMAX Direct Data Backup

zDP Mainframe Data Protection

zDP is a data protection feature for mainframe data where a virtual snapshots (SnapVX) can be taken up to every 10 minutes to create a near CDP (Continuous Data Protection) for a volume of data. Up to 256 versions may be kept for a volume. The snapshot works by capturing pointers to the data and maintaining pointers for changed data.

zBoost Mainframe I/O Acceleration

zBoost is a mainframe feature where chained I/O operations are broken into parallel I/Os to improve performance.

File Access - Embedded NAS (eNAS)

eNAS Data Services in VMAX is implemented by running the NAS software from VNX in hypervisor virtual machines of HyperMax. Data movers from the VNX implementation are called Software Data Movers as instances in the hypervisor. Control stations are also run as instances in the hypervisor.

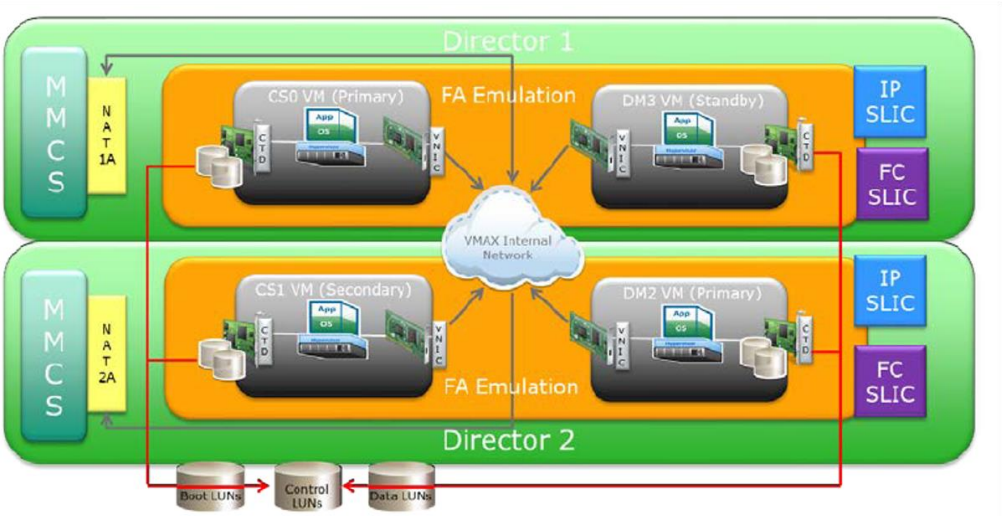
There can be from 2 to 4 Software Data Movers configured and 2 Control Stations for management.

NAS Specifications

	Value
CIFS Connections	64 K
Number CIFS Shares	10K
Number of NFS exports	2,048
Max concurrent NFS/CIFS TCP sessions	64K
Number of File/Directories opened	200,000
Max VDMs	29
Max filesystems	4,096 / system (including VDM and checkpoints)
Max filesystems / DM	2,048
Max Disk Volume Size	2 TB (minus 2 MB)
Max filesystem size	256 TB
Max file size	16 TB
Max files / directory	500,000
Max number of SnapSure Checkpoints	96 / PFS (production filesystem)
Snapshots / LUN	2,000
Max Configured # Replication Sessions	1,024
Max # Active Replication Sessions	64
Disk Volumes (LUN's) used with MPFSi / filesystem	128
Max # of VLANS	4,094
Max number of tree Quotas	2,047 / file system
Max size of tree Quota	4 TB
Max number of groups for Quota	64 K / file system

Table 5: NAS Usage Specifications

VMAX eNAS Data Service



MMCS – Management Module Control System
FA – Front-End Adapter
CTD – Cut Through Device
DM – Data Mover
CS – Control Station

Figure 8: eNAS Architecture Components source: Dell EMC

Element	250F	950F
Max Data Movers (1 is reserved as standby)	4	8
Max NAS capacity per array (TB usable)	1,150	3,584

Table 6: eNAS Models

eNAS Protocols

eNAS supports a number of protocols including the following:

NFS v2, v3, & v4/V4.1 with pNFS, CIFS/SMB-3, FTP, SNMP, NDMP, NTP, SNTP, TFTP, iSCSI, dNFS

NFS

eNAS supports multi-level NFS file system export with all versions of NFS (v2, v3 and v4/V4.1 with pNFS). This allows specific portions of a path to be exported separately, with different permissions.

Multi-level NFS exports are not supported across file-systems, or with nested mount filesystems (nmfs).

Clients accessing mounts get different permissions only if they mount a different access point with v2 and v3 protocols, but do not have to utilize different mount points to receive the correct permissions with NFS v4 clients.

For NFS v4, permissions are calculated for each client access based upon the exported path.

NFS v4 Enhancements:

- The addition of Access Control Lists (ACLs) provides fine grained user access control
- Maintenance of client context (stateful protocol) allows recovery of locks etc.
- Delegation of locking, caching and metadata information to the clients
- Secure NFS
- File Locking
- Internationalization support (I18N, uses UTF-8 characters)
- pNFS support with V4.1

CIFS/SMB

eNAS implements native SMB-3 with the added performance and functionality included with SMB-3. eNAS supports the following Microsoft Windows domain environments:

- Windows must utilize Active Directory (AD)
- Kerberos or NTLM support
- DNS server must be enabled, along with DDNS
- NTP time protocol must be enabled
- SMB-3

Automated Volume Management

Automatic Volume Management (AVM) is an eNAS feature that automates volume creation and management. By using AVM, system administrators can create and expand file systems without creating and managing the underlying volumes. Virtual provisioning, which works with Automatic File System Extension (AFSE), allows the file system to grow on demand. With virtual provisioning, the space presented to the user or application is the maximum size setting, while only a portion of that space is actually allocated to the file system.

With AVM, the size of an eNAS filesystem may be extended in the following ways:

- A filesystem may be extended if the filesystem has a user defined storage pool by allocating space from a storage pool
- A file system may be extended using a different pool than the one originally used to allocate space for the filesystem
- Enable AFSE to automatically extend the filesystem according to the AFSE policy rules

Automatic File System Extension

Dell EMC has added the ability to extend volumes and filesystems on eNAS without active administrative actions. An administrator may use both system-defined and storage pools and those defined by system administrators to create or extend the file system. Automatic File System Extension (AFSE) will increase the size of the filesystem as long as space is available in the storage pool. Included within AFSE is a feature Dell EMC refers to as virtual provisioning, also known as thin provisioning. This allows an administrator to configure a volume with a specific amount of size, and allocate less storage than the volume reports.

If AFSE is enabled, administrators may set the maximum size to which the file system may grow. Thin provisioning, is also a feature of AFSE, which permits the file system to grow as required. A high water mark is set by the administrator at the time AFSE is enabled. When the high water mark is reached, additional space is allocated to the filesystem. The high water mark (HWM) is established between 50 and 99% of the filesystems capacity. When the HWM is reached, a predefined policy action occurs with AFSE.

When a file system has automatic extension enabled and it reaches the high water mark, an automatic extension event notification is sent to the ***syslog*** and the file system is automatically extended. File systems that have Automatic File System Extension disabled continue to send event notifications when the high water mark is reached allowing the administrator the opportunity to take manual actions.

With AFSE, a file system that is smaller than 10 GB extends by its size when it reaches the high water mark. For example, a 3 GB file system, after reaching its high water mark, automatically extends to 6 GB.

When a file system larger than 10 GB reaches its high water mark, AFSE grows the file system by 5 percent of the file system size or 10 GB, whichever is larger. For example, a 100 GB file system extends to 110 GB, and a 500 GB file system extends to 525 GB.

There are a number of limitations in using AFSE including the type of file-system used, the RAID group configurations and a variety of other restrictions. Dell EMC documentation should be checked for details on restrictions. The most significant restrictions are listed below:

- Automatic filesystem extension is only supported with a file system created using from the Dell EMC AVM (automatic volume manager) pool
- AFSE only works with UxFS filesystems, including mgfs and dhsm filesystems
- Filesystem types such as rawfs, ckpt, nmfs and others will not work with AFSE
- Automatic file system extensions may not be used with iSCSI LUN's that are provisioned and allocated in a traditional manner (e.g. those that are not thinly provisioned)

Naming Services

The following naming services are supported by eNAS:

- Local files (passwd, group, hosts and netgroup)
- NIS (Network Information Services) for Unix systems
- DNS (Domain Name Service)
- LDAP (Lightweight Directory Access Protocol) with support for signed or encrypted messages
- WINS (Windows Internet Name Services)
- NT Domain and Active Directory

DNS is required for NT domains along with dynamic DNS support DDNS.

Active Directory is a directory service used in Windows that provides management of user and group accounts, security, and distributed resources. eNAS uses the LDAP protocol to query the Active Directory for domain information.

If the Active Directory schema has been extended to include UNIX attributes for Windows users and groups, administrators may configure a Software Data Mover to query the Active Directory to determine if the user and group have UNIX attributes assigned. If so, information stored in these attributes is used for file access authorization. The permissions are created in files manually on eNAS and then transferred to a control station.

Access Control

Microsoft has a function known as Access-based Enumeration (ABE) within Windows Server. This enhances file security by hiding files and folders from listing by users who do not have access rights to those files or directories. This prevents users from seeing the name of files or directories for which they do not have access rights. An executable developed by Dell EMC allows an administrator to enable or disable the state of ABE.

Multiprotocol Support

The support of multiple file management protocols (NFS and CIFS) together is critical item for NAS products as it is impossible to directly map Windows ACLs to NFS permissions. Therefore, eNAS maintains sets of both Windows ACLs and NFS permissions for every file and directory. In order to control the interaction between Windows ACLs and NFS permissions there are four access-checking policies that control how files are accessed in a multi-protocol environment.

eNAS maintains CIFS attributes and ACL's along with NFS ACL information in a multi-protocol directory (MPD) within the file system. MPD is a three-name directory that uses NFS names along with the short and long version of CIFS file names.

eNAS also utilizes UNIX-style user IDs (UIDs) and group IDs (GIDs) to record the ownership of files and directories; therefore, all Windows users and groups must be assigned UNIX-style UID and GIDs. This is accomplished by mapping Windows usernames and group names to UNIX-style UIDs and GIDs through several different methods.

A "Usermapper" service is one method to map user id's that will automatically map Windows users and groups to their equivalent UNIX ids. Additionally, eNAS provides a Unix User Management snap-in to the MMC (Microsoft Management Console). This allows an administrator to assign and modify UNIX ids for Windows users or groups within a Windows domain.

Three primary user authentication methods for eNAS systems:

- NT user authentication (default)
- UNIX user authentication (not recommended)
- SHARE user authentication (not recommended)

Only the NT user authentication method utilizes secure transmission of usernames and passwords. Using NT authentication, user id and password are sent encrypted from the client to the domain controller for authentication. Both UNIX and Share authentication methods utilize plain text or no passwords, with no checking of file or directory ACLs. Under certain conditions, NFS symbolic links are resolved for CIFS clients in order to provide access to the same files. Symbolic links are followed except in the following two cases:

- When the target file or directory is not visible to the CIFS client
- If the link does not begin with reference to the parent directory

File Locking

File locking in multi-protocol environments is an issue of concern. In a shared NFS and CIFS environment, a file may have locks set by both types of clients or users. NFS locks and CIFS deny modes and oplocks are not directly equivalent, therefore eNAS translates CIFS deny modes and oplocks to NFS locks and translate NFS locks into CIFS deny modes and oplocks. Two examples are:

- A CIFS deny read/write mode request is translated into an NFS exclusive read/ write lock.
- An NFS shared read lock is translated into a CIFS deny write mode.

eNAS implements the following locking policies, which may be used by system administrators.

	NOlock	Wlock	RWlock
Definition	No locking, permits all access	Denies write access when locked	Denies read, write or execute access when locked
Security Level	Lowest	Moderate security	Highest security
CIFS Client	Ignores locks set by NFS	CIFS clients cannot open files locked by NFS for exclusive access	CIFS clients denying concurrent access for read or write cannot open files locked by NFS for exclusive or shared access
NFS Client	NFS client can read and write files locked by CIFS	NFS clients may read but not write or delete files locked by NFS or CIFS clients	NFS clients cannot read, write, or delete files locked by CIFS

Table 7: eNAS Locking Interoperability

CIFS Oplocks

Opportunistic file locks (oplocks) are configured per file system and are on by default. This type of locking works with CIFS clients and servers and can improve network performance by allowing CIFS clients to locally buffer file data before sending it to the server. Some applications such as database application access via CIFS recommend oplocks are turned off, or for other critical data access. eNAS supports level II, exclusive, and batch oplocks.

Access Policies

For Windows: Access control lists (ACLs) control file access by explicitly allowing or denying actions against an object. An ACL contains a list of user and group accounts and the specified actions these accounts can perform. ACLs are supported by the VNX at the share, directory, and file level.

For NFS: Access classes (user/owner, group, other) are used in conjunction with access modes (read, write, and execute) to control file access. Access rights apply to directories and to files.

Access policies only apply when the user authentication method on a Data Mover is set to NT. When mounting a file system one of the following access-checking policies may be specified:

- NATIVE (default)
- NT
- UNIX
- SECURE

Permission Modes

UNIX and Windows handle access control in very different ways, making it difficult to set the same security on a file system object in a multi-protocol environment. eNAS MIXED and MIXED_COMPAT policies synchronize UNIX and Windows permissions as closely as possible by using an algorithm that translates UNIX rights into ACL entries and ACL entries into UNIX rights.

The MIXED and MIXED_COMPAT polices differ in the way they translate a UNIX Group into an ACE and how they perform access checking. The MIXED policy always performs access checking against an ACL independent of the protocol accessing a file system object, as explained in the following example. The MIXED_COMPAT policy uses the permissions from the protocol that last set or changed the permissions on a file system object.

When the MIXED and MIXED_COMPAT policies are enabled for a file system object, the ACL and UNIX mode bits are automatically synchronized. Changes to an ACL result in modifications to the mode bits and changes to the mode bits reconstruct the ACL.

	File Permissions			Directory Permissions		
	R	W	X	R	W	X
Traverse Folder/Execute File			X			X
Read Data	X					
Read Attributes	X			X		X
Read Extended Attributes	X			X		
Write Data		X				
Append Data		X				
Write Attributes		X			X	
Write Extended Attributes		X			X	
Delete		X			X	
Read Permissions	X					
List Folders				X		
Create Files					X	
Create Folders/Directories					X	
Delete Subfolders and Files					X	

Table 8: eNAS CIFS Permissions

File Permissions	R	W	X
Traverse Folder/Execute File			X
Read Data	X		
Read Attributes	X		
Read Extended Attributes	X		
Write Data		X	
Append Data		X	
Write Attributes		X	
Write Extended Attributes		X	
Delete		X	
Read Permissions	X		
List Folders			
Create Files			
Create Folders/Directories			
Delete Subfolders and Files			

Table 9: NFS to CIFS mappings in eNAS

Permission Inheritance

When files or directories are created, eNAS generates both CIFS ACL and UNIX permissions for the new object. For CIFS clients creating objects, the ACL's are inherited from the enclosing directory or share. The UNIX permissions are determined by the umask set for the entire share or mount point. The mount point umask is set through an option of the ***server_export*** command.

File Change Notification

Microsoft provides a Win32 API that allows applications to register with the CIFS server to be notified when specific actions occur. This capability is known as Change Notify and allows applications to receive a notification when changes are made, rather than to continually check (or poll) for updates. Update notification is made via asynchronous updates, which for UNIX experts, is similar in concept to using the Berkeley based ***select ()*** system call rather than the SystemV based ***poll ()*** system call. Using the Windows API, applications may register to receive change notifications asynchronously (without waiting) for any directory. Dell EMC has implemented the change notify facility, so that CIFS clients can register for notification on changes to directories located on ENAS based systems. The file change notification mechanism will asynchronously notify CIFS clients when registered directories are updated.

The limitations with this feature are that updates from other file clients, such as NFS, FTP, HTTP or MPFS clients will not trigger the change notification action to occur. Additionally, the user authentication method must be set to NT, which is the recommended setting.

Backup

eNAS offers several methods to perform a backup including the following:

- Local Backup
- Client Backup (via NFS and/or CIFS)
- NDMP Backup

A local backup may be performed which uses either a Dump or PAX format. Dump is an open UNIX format for data backup. PAX is similar and is an IEEE and POSIX standard archive format

Performing a client backup via either an NFS or CIFS from a mounted filesystem from a host (client) is also possible. The advantage with this method is that it is similar in operation to other backup operations on the host, and typically can leverage the backup infrastructure in place. The disadvantage is that a CIFS mount share will only backup the CIFS permissions, losing the NFS permissions. In the case of an NFS mount, the CIFS permissions are not backed up, and thus not available when restored.

> ***Evaluator Group Comments: For environments that use CIFS or NFS exclusively, utilizing client backups via a mounted filesystem can work well. However, most mixed mode environments should choose either local or NDMP backups in order to preserve the file permissions.***

NDMP

The Network Data Management Protocol (NDMP) was developed primarily for backing up NAS systems from remote management applications. NDMP not only can retain permissions for both NFS and CIFS clients, but it also provides the ability to share backup infrastructure, including the backup applications along with SAN tape drives and libraries.

ENAS requires a read-only filesystem for an NDMP backup, which may be accomplished most easily by utilizing SnapSure to create a point in time consistent copy that is used to backup. A SnapSure checkpoint is automatically created when an NDMP backup is initiated.

ENAS backups, including NDMP backups, operate as a traditional type, known as PAX, or a new volume based backup method known as Volume Backup (VBB). The PAX based method creates backups at the file level, and recursively descends through each directory, backing up each file in that directory.

NDMP Volume Backup reads a set of disk data blocks in an efficient manner compared to the method used for traditional, file-based backups. NDMP Volume Backup works only with Dell EMC-qualified vendor backup software, as listed in the Dell EMC NAS Interoperability Matrix. eNAS NDMP Volume Backup can be used when performing a full backup or an incremental backup.

Types of NDMP backup

- Three way backup – the backup is created on tape drives that are SAN attached to the VNX
- Local Backup – the backup is created on tape drives attached to the VNX system
- Remote Backup – the backup is created on remote VNX systems tape drives that are attached to the remote system

The following type of restore operations are supported with NDMP volume based backups (VBB):

- Full destructive restore
- File Level restore

Types of PAX Restore:

- Normal
- DAR - direct access restore
- DDAR – directory DAR, and improved form of DAR that enhances performance

The Directory Direct Access Restore (DDAR) optimizes recovery by allowing the NDMP client to directly access backed up data anywhere on the tape set when used with a backup vendor that supports DDAR. DDAR is not compatible with volume backups.

Each Data Mover supports four concurrent NDMP sessions at one time. For example, a system can run three backup sessions and one restore session, simultaneously.

NDMP restrictions:

- NDMP based backups and restores of iSCSI LUNs is not supported
- NDMP does not follow symbolic links within filesystems, therefore NDMP backups do not include the target of the link

VAAI Support for NAS

The VNX system support VMware vSphere APIs for Array Integration (VAAI) for NAS introduced starting with vSphere 5.0. The features are exploited by VMware to accelerate performance and provisioning time.

- Full File Clone – enables VNX to make copies of virtual disks (VMDK files) in the storage system
- Extended Statistics – enables vSphere to query space utilization on VNX. Information returned includes the size of the file and the space consumed by the file.
- Space Reservation – enables thick virtual disk files to be provisioned with the lazy-zeroed option.

Advanced Features for eNAS

SnapSure

SnapSure is a feature within eNAS that provides point in time copy or a snapshot of a file system. Dell EMC refers to their point in time copies as a checkpoint, which is a read-only logical point in time image of a filesystem. A SnapSure checkpoint is designed to provide file recovery in production and testing environments. It is not designed for business continuance or disaster recovery scenarios, since the original data must be accessible in order for the point in time copy data to be available.

The way SnapSure is implemented is to track changes at the block level. When a block within a production filesystem (PFS) is modified, a copy containing the block's original contents is saved to a separate volume called the "SavVol". Subsequent changes made to the same block in the PFS are not copied into the SavVol. The original blocks in the SavVol and the unchanged blocks in the PFS are read by SnapSure according to a bitmap and blockmap tracking structure. These blocks combine to provide a complete point-in-time file system image called a checkpoint, which contains two views: the current view and the previous point in time image.

eNAS supports multiple types of filesystems including a SnapSure checkpoint (ckpt) filesystem type. eNAS supports up to 96 SnapSure checkpoints of a filesystem. Checkpoints may be refreshed, which will reuse the SavVol and block map indices to recreate a current PIT copy of the filesystem.

Snapshot virtual filesystems (also known as Checkpoint virtual filesystems) permit read only access to a directories previous SnapSure checkpoint through a virtual entry hidden directory point named ".ckpt". Only the latest checkpoint has a bitmap, but all checkpoints (including the newest) have their own blockmap.

Additionally, SnapSure can be registered as a hardware VSS provider for Microsoft Windows, thereby supporting Shadow Copy for Shared Folders for filesystems accessed from Microsoft server systems.

Quotas

Both soft (advisory) and hard (restrictive) quotas may be set for the amount of storage space consumed, or the number of files utilized. These quota restrictions may be applied on a per user, a per group basis or in combinations.

eNAS can track quotas using either a block or a file quota policy. The default policy is set to track the number of blocks utilized. Any file smaller than a filesystem block size of 8KB is counted as 8KB towards the quota since the file consumes one 8KB block on disk. If used with the file mover, as the content of files migrate to secondary storage, quota usage decreases (as files are migrated off the eNAS) or increases (as they are brought back), because the number of blocks occupied in the primary file system on the eNAS changes.

If quota policy is set to file-size, disk usage is calculated in logical file size, which means a 1 KB file counts as 1 KB in the quota. When using the file-size quota policy, eNAS quota usage is unaffected by file migration and recall operations because these operations do not impact the logical size of files.

File Mover

File Mover enables administrators to place file based data according to polices. Administrators may create rules, known as policies that determine whether files should be migrated from primary storage to slower, less-expensive storage devices and back again. Commonly used migration policies are those that migrate older, less frequently used files or files of a particular type.

File mover reads file attributes and will move files by matching the rules or polices that have been established. File mover is able to utilize NFS, CIFS or HTTP to scan files. File Mover will not operate on a read-only filesystem, nor does it support the use of the NFS v4 protocol.

File Mover has added support to migrate files to clouds. Cloud targets include VirtuStream cloud Amazon S3 and Microsoft Azure.

When the policy and migration software finds a match, it migrates or moves the file from the eNAS. It does so by reading the file from the eNAS and then writing it to the secondary file server using NFS or CIFS. Next, the policy and migration software issues a command via the File Mover API to convert the file to a stub file. This releases the storage associated with the file and completes the migration operation. When a file is migrated, its data is moved to a file on secondary storage, and the original file is replaced with a stub file. The stub file contains all the metadata associated with the original file including permissions, timestamps, size and other attributes. The stub file also contains additional data associated with the File Mover functionality, including the location of the file content on secondary storage, the identity of the policy and migration software that migrated the file's content, and an information field in which the policy and migration software can store any additional information. A stub file consumes one inode (one 8KB data block), plus the size of any alternate data streams associated with the file, in the primary file system on the eNAS system.

When a write operation occurs, eNAS uses the information in the stub file to locate the file on secondary storage. It then compares the modification time (mtime) and size it has stored to the current mtime and

size of the file on secondary storage. This comparison ensures the stub file and migrated file remain synchronous. When these attributes match, file data is returned to the client to satisfy the read request. If the attributes differ, an error message appears.

There are two ways in which client applications can write to migrated files on a file server.

1. The client can either modify a portion of a file or replace the entire file
2. With exception of CIFS backup, all NFS and CIFS applications cause file data to be recalled during read operations

A read migration may occur in three ways:

1. Pass-through – eNAS reads the requested data from secondary storage and returns it to the client immediately. None of the file's data is recalled to primary storage; the file remains on secondary storage.
2. Full – eNAS recalls all the file data to primary storage when the file is first accessed for read, which maximizes performance of subsequent read or write operations.
3. Partial – eNAS recalls only as much of the file's contents as it needs to satisfy the client read request. In this case, read requests of large files complete faster than a full migration as only the required portion of the data is recalled

MPFS (formerly HighRoad)

Dell EMC's Multi Path File System (MPFS) allows command and control of data movement to be separated. With MPFS, control for data movement including locking occurs via NFS protocols, while data movement flows over Fibre Channel. In the case of MPFSi, the data flows over iSCSI connections. By using MPFS/MPFSi, data movement may occur in parallel and with less latency and overhead than movement via typical NAS protocols only. In some instances, data throughput may be increased up to 4X that of using NAS only for data movement.

Parallel data access may occur in the following ways with MPFS/MPFSi:

- MPFS may read and write data concurrently
- Clients may access multiple back-end storage systems in parallel
- Clients may access multiple ENAS concurrently, although one file may not be shared across multiple systems or back-end storage devices
- Files in an mpfs filesystem are striped across multiple LUNs of back-end storage
- The LUN's utilized for an mpfs filesystem are comprised of stripes across a back-end RAID3 or RAID5 group

A piece of code must also reside on the client (the application server that requests the data) in order to interoperate with the MPFS controller running on the ENAS. Currently, Dell EMC MPFS supports Windows servers, Linux, HP-UX, Solaris, AIX and IRIX operating systems. MPFSi is supported on Linux only.

Evaluator Group Comment: MPFSi has additional network requirements, along with the potential for network congestion within the iSCSI to FC gateway if sufficient ports are not available. These limitations, plus the support for Linux as the only host system significantly limits the deployment scenarios for MPFSi. When Dell EMC reduces or removes these limitations, MPFSi may offer advantages over MPFS for many environments.

File Level Retention

With File Level Retention (FLR) permissions to modify or delete files may be overridden at file creation time. This provides a WORM-like feature for underlying storage residing upon any storage medium, and accessed via any supported eNAS file access method, including NFS and CIFS. FLR is a characteristic given to a filesystem that safeguards data integrity and immutability by providing an entire filesystem, which has file level retention enabled. FLR effectively changes the permissions of the file to read only for a specified amount of time. The length of time, which a file is to be retained, is the period during which file changes, including changes to permissions are not permissible.

FLR is enabled at the time the filesystem is created. Files created within an FLR filesystem have one of three states, Clean, FLR or Expired. The files and the paths to the files are locked from modification for the amount of time specified by the retention period. A clean file has not yet been designated for FLR status. It may be modified in the same way as a regular file. A FLR file is in a WORM status, and may not be altered, deleted, renamed or overwritten (including its full path) until the retention period has expired. The Expired status indicates that the retention period has ended, and the file may again be modified.

The retention date is set by setting the last access time of a file to a future date. By the setting of the last access time, for a file within an FLR enabled filesystem, the file is set to the FLR state with an expiration date as set by the last access meta data field. Once a files retention time is set, it may not be shortened, only increased. The second part of establishing a file as a WORM file, is to change its permissions to read-only. These operations may be accomplished through NFS and CIFS file operations.

eNAS does permit the administrator to delete an entire filesystem, regardless of its retention status.

File Level retention does not support the use of the NFS v4 protocol.

Additionally, it is not possible to enable File Mover features on a file systems with FLR enabled for the source, as it is not possible to modify these files. However, FLR filesystems may be used as the target or destination for File Mover.

Anti-Virus Scanning

eNAS supports the use of external virus (AV) scanning engines through a mechanism known as AVA (or an Anti-Virus Agent). AVA does not actually perform anti-virus scanning; instead, it acts as an interface and method for allowing an AV agent to scan a file. AVA is a Dell EMC proprietary interface and method of virus scanning. Another developing industry standard is the ICAP interface (Internet Content Adaptation Protocol).

The AV software on an AVA server in conjunction with virus scanning software such as Symantec Norton, McAfee or others is configured to scan network drives on access.

There are two primary methods for utilizing AV scanning, either real-time (as files are accessed) or scheduled scanning. These methods may be used individually or together.

The order of operations is:

1. File open via CIFS to eNAS
2. eNAS locks the file and ships it to the AVA server(s) via the agent service (AVA) on a remote server
3. Real-time AV software scans the file
4. Once determined clean, AVA agent tells eNAS to release the lock

AVA is not a feature that is installed on the operating environment, it is an agent installed on a separate server, together with anti-virus software on a standalone server. CAVA by itself does not perform virus scanning. Using the Antivirus Agent with File Mover is possible using a method of reading and writing through (also known as data pass-through) migration method.

File Replicator

File Replicator produces a read-only, point-in-time copy of a source file system and periodically updates this copy, making it consistent with the source file system. This read-only copy can be used by an eNAS system in the same cabinet, or an eNAS at a remote site for content distribution, backup, and application testing. Failover from local to the remote site is possible in the event the source system is unavailable. Similar to MirrorView, Replicator may be used to replicate data. In the event of a failover, replication is stopped and the outstanding changed data (known as delta sets) are applied to the destination site. Once the appropriate delta sets are applied, the destination system is changed to read/write mode.

Resynchronization is also possible by re-establishing the replication, this time in reverse. At some point, after the two systems are in synch, the replication may again be reversed to establish the entire environment back to the status prior to the disaster.

Two parameters used to control replication include a high-water mark and a time-out. The high-water mark establishes the size in MB of the file system changes that may accumulate in the delta set prior to being replicated. Replicator utilizes a disk storage area known as a ***SavVol*** to store changes that occur on the source file system that have not yet been updated on the destination file system. Replicator requires a SavVol at both source and destination, with the size must be larger than 1 GB with the maximum 500 GB. The default size of a SavVol is set to be 10% of the size of the source filesystem, however for environments with either slow connections or high rates of change; this setting should be adjusted upwards. Policies may also be implemented to determine how often delta sets are created on the source, and how often they are updated on the destination.

Evaluator Group Comments

The Dell EMC VMAX All Flash system is a logical extension of the VMAX line to capitalize on flash technology while maintaining the high-end enterprise system characteristics that VMAX has delivered. There have been optimizations for flash both in physical packaging and in embedded software. Customers of Dell EMC that have invested in the high end systems can now move to the higher performance, new technology of all flash systems for primary storage with the VMAX All Flash system.

We expect current customers to move to the VMAX All Flash as their next technology update and, very importantly, continue the current operational procedures without interruption. The All Flash system should be a popular choice for those customers.

Strengths:

The Dell EMC VMAX All Flash system provides the advantages of solid-state storage to a family of systems that has been a leading high-end enterprise storage system. The use of flash will improve reliability, power consumption, physical space required, and most importantly, increase performance. The simplicity of a single technology of devices has many advantages; one being it simplifies the administrative choices. Dell EMC will be able to demonstrate increased simplicity with the VMAX All Flash system.

The VMAX All Flash system has the advantage of the maturity of the system design and investments Dell EMC has made to continue to improve the system with adoption of new technology in an evolutionary fashion. This will be a low risk acquisition for customers. Dell EMC can be expected to continue major updates every 12 to 18 months, increasing the value of the system.

As a very feature-rich, high-end system, the complexity can get overwhelming for customers at times. The use of embedded Unisphere for VMAX is a great step in help to hide or automate that complexity and allow system management by administrators with less experience. This is a trend in the industry and Unisphere is a good start. Unisphere 360 as an optional software product to manage multiple VMAX systems will be very useful for large enterprise environments.

Direct backup of volumes from VMAX to Data Domain with Dell EMC ProtectPoint is a major improvement for data protection that will produce long-term changes in how data is protected and recovered that can be exploited by customers.

The support for multiple operating environments including mainframe, open systems, and system-i is excellent and provides a unique solution for many customers.

Potential Concerns:

The VMAX All Flash system has optimizations for use of flash technology. Still, it is an adapted system meaning that the original design was for spinning disk. Additional optimizations should be expected over time to change some of those earlier characteristics.

Standard SSDs are used with the VMAX All Flash system. There may be advantages in cost and performance from using flash chips with a custom controller rather than SSDs from another vendor.

Attachment of the SSDs to the VMAX controllers is through SAS, which is a disk-based interface and protocol. Use of NVMe as a protocol and another interface such as PCIe would provide greater performance and will be a competitive issue. We would expect Dell EMC to make a transition in this area with the next major update for VMAX All Flash.

There will still be some overlap and potential customer confusion (at least initially) with the XtremIO all flash system. Salespersons can effectively deal with this but there may be some preconceptions of customers that require explanations to clear up any confusion.

More detailed information is available at http://evaluatorgroup.com