TECHNICAL NOTES

EMC® VMAX3

Service Level Provisioning with Fully Automated Storage Tiering™ (FAST)

REV A03
February, 2015

This technical notes document contains information on these topics:

EMC²

Executive Summary

Organizations around the globe need IT infrastructures that can deliver instant access to the huge volumes of data intrinsic to traditional transaction processing/data warehousing and to a new generation of applications built around the world of social, mobile, cloud, and big data. EMC® is redefining Data Center Cloud Platforms to build the bridge between these two worlds to form the next generation Hybrid Cloud.

Essential to this is the ability to achieve predictable performance at large scale for extreme-growth hybrid cloud environments. EMC Fully Automated Storage Tiering™ (FAST) for VMAX3 arrays (VMAX100K, 200K and 400K), running HYPERMAX OS 5977, dynamically allocates workloads across storage technologies by non-disruptively moving application data to meet stringent Service Level Objectives.

Audience

This technical notes document is intended for anyone who needs to understand Service Level Provisioning and FAST, best practices, and associated recommendations to achieve the best performance for VMAX3 configurations. This document specifically targets EMC customers, sales, and field technical staff who have either installed and implemented a VMAX3 data services platform or are considering a future implementation.

Introduction

VMAX3 is the first enterprise data services platform purpose-built to deliver and manage predictable service levels at scale for hybrid clouds. It is based on the world's first and only Dynamic Virtual Matrix™ that delivers agility and efficiency at scale. Hundreds of CPU cores are pooled and allocated on-demand to meet the performance requirements for dynamic mixed workloads. The VMAX3 provides up to three times the performance of previous generation arrays with double the density.

Running on the Dynamic Virtual Matrix is the new HYPERMAX OS—the industry's first converged storage hypervisor and operating system that runs mission-critical data and application services with radically simple "single click" service level provisioning. The new storage hypervisor enables VMAX3 to redefine the data center by embedding data services (file/object, replication, data mobility) and application services (e.g., database tools, analytics, ETL) that traditionally would have run external to the array, bringing new levels of efficiency to enterprise workloads.

VMAX3 arrays deliver mission-critical storage with the scale, performance, availability, and agility to meet the high demands of extreme data growth in today's and tomorrow's hybrid cloud.

The VMAX3 family delivers unmatched ease of provisioning for your specific service level objectives. These service levels are tightly integrated with EMC's FAST technology to optimize agility and array performance across all drive types in the system. FAST improves system performance while reducing cost by leveraging high performance Flash drives combined with cost effective high capacity drives.

EMC FAST dynamically allocates workloads across storage technologies, non-disruptively moving workloads to meet stringent service level objectives. FAST moves the most active parts of your workloads to high-performance flash drives and the least-frequently accessed data to lower-cost drives, leveraging the best performance and cost characteristics of each different drive type. FAST delivers higher performance using fewer drives to help reduce acquisition, power, cooling, and footprint costs.

Service Level Provisioning

VMAX3 radically simplifies management of provisioning by eliminating the need to manually assign physical storage resources to applications. Instead, the storage performance required for the application is specified during the provisioning process, with the storage array then provisioning the workload appropriately. The performance requirement is specified by associating a pre-defined service level objective to the application. Application data is then dynamically reallocated across storage resources of differing performance characteristics to achieve the performance requirement of the application.

This ability to provision to service levels is inherently available to all VMAX3 storage arrays: all arrays are virtually provisioned with FAST permanently enabled.

Virtual Provisioning

Virtual Provisioning enables the ability to increase capacity utilization by enabling more storage to be presented to a host than is physically consumed, and by allocating storage only as needed from a shared virtual pool. Virtual Provisioning also simplifies storage management by making data layout easier through automated wide striping, and by reducing the steps required to accommodate growth.

Virtual Provisioning uses a type of host-accessible device called a virtually provisioned device, also known as a thin device, which does not need to have physical storage completely allocated at the time the devices are created and presented to a host. The physical storage that is used to supply drive space for a virtually provisioned device comes from a shared storage pool, also known as a storage resource pool. The storage resource pool is comprised of one or more data pools containing internal devices called data devices. These data devices are dedicated to the purpose of providing the actual physical storage used by virtually provisioned devices.

When a write is performed to a portion of the virtually provisioned device, the VMAX3 array allocates a minimum allotment of physical storage from the pool and maps that storage to a region of the virtually provisioned device, including the area targeted by the write. The storage allocation operations are performed in small units of storage called virtually provisioned device extents. Extents may also be called chunks.

The virtually provisioned device extent size is 1 track (128 KB).

When a read is performed on a virtually provisioned device, the data being read is retrieved from the appropriate data device in the storage resource pool to which the virtually provisioned device is bound. Reads directed to an area of a virtually provisioned device that has not been mapped do not trigger allocation operations. The result of reading an unmapped block is that a block in which each byte is equal to zero will be returned. When more storage is required to service existing or future virtually provisioned devices, data devices can be added to existing virtually provisioned data pools within the storage resource pool.

Fully Automated Storage Tiering

Fully Automated Storage Tiering (FAST) automates the identification of active or inactive application data for the purposes of reallocating that data across different performance/capacity pools within a VMAX3 storage array. FAST proactively monitors workloads to identify busy data that would benefit from being moved to higher-performing drives, while also identifying less-busy data that could be moved to higher-capacity drives, without affecting existing performance.

This promotion/demotion activity is based on achieving service level objectives that set performance targets for associated applications, with FAST determining the most appropriate drive technologies, or RAID protection types, to allocate data on.

FAST operates on virtual provisioning thin devices, meaning data movements can be performed at the sub-LUN level. In this way, a single virtually provisioned device may have extents allocated across multiple data pools within a storage array

Data movement executed during this activity is performed non-disruptively, without affecting business continuity and data availability.

Elements of FAST

There are five main elements related to the use of FAST on VMAX3 storage arrays. These are:

- Disk groups

- Data pools
- Storage resource pools
- Service level objectives
- Storage groups

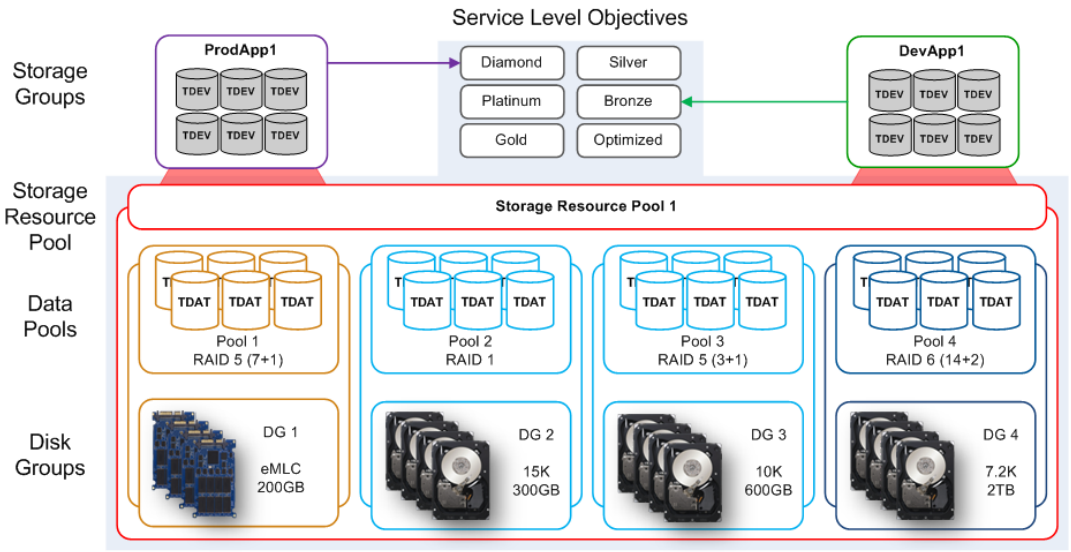

Figure 1. FAST elements

Note Disk groups, data pools and data devices, storage resource pools, and service level objectives all come preconfigured on the VMAX3 storage array when shipped to a customer site.

Disk groups

A disk group is a collection of physical drives sharing the same physical and performance characteristics. Drives are grouped based on technologies, rotational speed, capacity, form factor, and desired RAID protection type.

Each disk group is automatically configured with data devices (TDATs) upon creation. A data device is an internal logical device dedicated to providing physical storage to be used by thin devices.

All data devices in the disk group are of a single RAID protection type, and all are the same size. Because of this, each drive in the group has the same number of hypers created on them, with each hyper also being the same size. There are 16 hypers configured on each drive.

The VMAX3 storage array supports up to 512 internal disk groups.

Disk groups are preconfigured within the storage array and their configuration cannot be modified using management software.

Data pools

A data pool is a collection of data devices of the same emulation and RAID protection type. All data devices configured in a single physical disk group are contained in a single data pool. As such, all the data devices are configured on drives of the same technology type and capacity, and, if applicable, rotational speed.

The VMAX3 storage array supports up to 512 data pools.

Data pools are preconfigured within the storage array and their configuration cannot be modified using management software.

Storage Resource Pools

A Storage Resource Pool is a collection of data pools constituting a FAST domain. This means that data movement performed by FAST is done within the boundaries of the storage resource pool. Application data belonging to thin devices can be distributed across all data pools within the storage resource pool to which it is associated. TimeFinder snapshot data and SRDF/A DSE (delta set extension) data is also written to pools within a storage resource pool.

A storage resource pool can contain up to 512 data pools. By default, a VMAX3 storage array has a single storage resource pool containing all the configured data pools.

There is no restriction on the combination of drive technology types and RAID protection. When moving data between data pools, FAST will differentiate the performance capabilities of the pools based on both rotational speed (if applicable) and RAID protection.

While a storage resource may contain multiple data pools, individual data pools can only be a part of one storage resource pool.

The VMAX3 storage array supports a maximum of 2 storage resource pools. When multiple storage resource pools are configured, one of the storage resource pools must be marked as being the default storage resource pool.

Storage resource pools are preconfigured within the storage array and their configuration cannot be modified using management software.

Service Level Objectives

A service level objective defines an expected average response time target for an application. By associating a service level objective to an application, FAST automatically monitors the performance of the application and adjusts the distribution of extent allocations within a storage resource pool in order to maintain or to meet the response time target.

There are five available service level objectives, varying in expected average response time targets. There is an additional Optimized SLO which has no explicit response time target associated with it. The base Service level names are customizable. The Customized SLO name cannot exceed 32 characters in length. Only alphanumeric characters, hyphens (-) and underscores (_) are allowed. The name cannot start with a hyphen or underscore and cannot include spaces.

The base name of any SLO cannot be used; all changed SLO names must be unique.

A complete list of the available service level objectives is shown in Table 1.

Table 1. Service Level Objectives

Service Level Objective	Behavior	Expected Average Response Time
Diamond	Emulates EFD performance	0.8 ms

Service Level Objective	Behavior	Expected Average Response Time
Platinum	Emulates performance between EFD and 15K RPM drive	3.0 ms
Gold	Emulates 15K RPM performance	5.0 ms
Silver	Emulates 10K RPM performance	8.0 ms
Bronze	Emulates 7.2K RPM performance	14.0 ms
Optimized (default)	Achieves optimal performance by placing most active data on higher performing storage and least active data on most cost-effective storage	N/A

By default, all data in the VMAX3 storage array is managed by the Optimized SLO. However, an explicit response time target may be set for an application by associating it with one of the other service level objectives.

The actual response time of an application associated with each service level objective will vary based on the actual workload seen on the application and will depend on average IO size, read/write ratio, and the use of local or remote replication.

Note The expected average response times shown in Table 1assume a small average IO size for the workload (less than 64 KB), with no local or remote replication.

There are four workload types that may be added to the chosen service level objective, with the exception of Optimized, to further refine response time expectations. In addition there is the ability to save and create a customer workload types for future use. The custom workload type can be used while provisioning

storage. Within the reference workload section you will also be able to view the workload skew for the given service level selected.

The standard available workload types are shown in Table 2.

Table 2. Workload Types

Workload	Description
OLTP	Small block IO workload
OLTP with Replication	Small block IO workload with local or remote replication
DSS	Large block IO workload
DSS with Replication	Large block IO workload with local or remote replication

Service level objectives, and workload types, are predefined within the storage array and they cannot be modified using management software.

Storage groups

A storage group is a logical collection of VMAX devices that are to be managed together, typically constituting a single application. Storage group definitions are shared between FAST and auto-provisioning Groups.

Storage groups can be associated with a storage resource pool, or a service level objective, or both. Associating a storage group with a storage resource pool defines the physical storage to which data in the storage group can be allocated on. The association of a service level objective defines the response time target for that data.

By default, devices within a storage group will be associated with the storage resource pool that is designated as the default, and managed under the default Optimized SLO.

While all data in the VMAX3 storage array is managed by FAST, a storage group is not considered 'FAST managed' if it is not explicitly associated with a storage resource pool or a service level objective.

Devices may be included in more than one storage group, but may only be included in one storage group that is 'FAST managed'. This ensures that a single device cannot be managed by more than one service level objective or have data allocated in more than one storage resource pool.

The VMAX3 storage array supports up to 16,384 storage groups. Storage groups may contain up to 4,096 devices. Each storage group name may be up to 64 alphanumeric characters, hyphens (-), and underscores (_). Storage group names are not case sensitive.

FAST implementation

FAST runs entirely within HYPERMAX OS, the storage operating environment that controls components within the array.

Management

Management of service level provisioning and FAST is performed by using either Unisphere for VMAX or Solutions Enabler. Both management interfaces provide the ability to associate both service level objectives and storage resource pools to storage groups.

In addition, Unisphere for VMAX provides the ability to perform workload planning, by allowing a suitability check on newly provisioned applications to determine if there is performance bandwidth available in the storage array to accept the additional workload.

Runtime implementation

The goal of FAST is to deliver defined storage services, namely application performance based on service level objectives, based on a hybrid storage array containing a mixed configuration of drive technologies and capacities.

Based on the configuration of the array, FAST balances the capabilities of the storage resources, primarily the physical drives, against the performance objectives of the applications consuming storage on the array. This means that FAST will aim to maintain a level of performance for an application that is within allowable response

time range of the associated service level objective while understanding the performance capabilities of each disk group within the storage resource pool.

The data movements performed by FAST are determined by forecasting the future system workload at both the disk group and application level. The forecasting of future workload is based on observed workload patterns.

The primary runtime tasks of FAST are:

- Collect and aggregate performance metrics
- Monitor workload on each disk group
- Monitor storage group performance
- Execute required data movements

All runtime tasks are performed continuously, meaning performance metrics are constantly being collected and analyzed and data is being relocated within a storage resource pool to meet application service level objectives.

Performance metrics collection

Performance metrics collected for the purposes of FAST are collected at three separate levels:

- Disk group
- Storage group
- Thin device subLUN

Thin device performance collection

At the sub-LUN level, each thin device is broken up into multiple regions, known as extent groups and extent group sets. Each thin device is made up of multiple extent group sets which, in turn, contain multiple extent groups.

Figure 2 shows a graphic representation of a thin device divided into each of these separate regions.

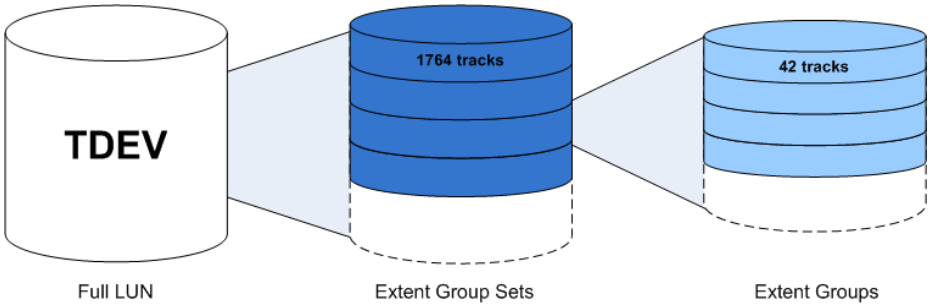

Figure 2. Thin device sub-LUN performance collection extents

Each extent group is made up of 42 contiguous thin device extents. With each thin device extent being a single track in size, an extent group represents 42 tracks of the device.

Each extent group set is made up of 42 contiguous extent groups, representing 1,764 tracks of the device.

The metrics collected at each sub-LUN level allow FAST to make separate data-movement requests for each extent group of the device, 42 tracks.

Performance metrics

The primary performance metrics collected for use by FAST relate to:

- Read misses
- Writes
- Prefetch (sequential reads)
- Cache hits
- IO size
- Workload clustering

The read miss metric accounts for each DA read operation that is performed. That is, data that is read from a thin device that was not previously in cache and so needs to be read directly from a drive within the storage resource pool.

Write operations are counted in terms of the number of distinct DA operations that are performed. The metric accounts for when a write is destaged.

Prefetch operations are accounted for in terms of the number of distinct DA operations performed to prefetch data spanning a thin device extent. This metric considers each DA read operation performed as a prefetch operation.

Cache hits, both read and write, are counted in terms of the impact such activity has the front-end response time experienced for such a workload.

The average size of each IO is tracked separately for both read and write workloads.

Workload clustering refers to the monitoring of the read-to-write ratio of workloads on specific logical block address (LBA) ranges of a thin device or data device within a pool.

Note In the case of local replication sessions where host read I/Os to a target device are redirected to the source device, the read activity is counted against the source device. This is typically the case for TimeFinder/Clone, TimeFinder VP Snap when the session status is *CopyOnWrite*. This is also true for linked TimeFinder SnapVX targets.

Performance metrics analysis

FAST uses four distinct algorithms, two capacity-oriented and two performance-oriented, in order to determine the appropriate allocation of data across a storage resource pool. These are:

- SRP capacity compliance
- SLO capacity compliance
- Disk resource protection
- SLO response time compliance

The SRP and SLO capacity compliance algorithms are used to ensure that data belonging to specific applications is allocated within the correct storage resource pool and across the appropriate drive types within a storage resource pool, respectively.

The disk resource protection and SLO response time compliance algorithms consider the performance metrics collected to determine the appropriate data pool to allocate data in order to prevent the overloading of a particular disk group and to maintain the response time objective of an application.

SRP capacity compliance

The SRP capacity compliance algorithm is a capacity based algorithm that ensures all data belonging to thin devices within a particular storage group is allocated within a single storage resource pool. This algorithm is only invoked when a storage group's association to a storage resource pool is modified.

All data allocated for the devices within the storage group will be moved from the original storage resource pool to the newly associated storage resource pool. During this movement, data for the thin devices will be allocated across two storage resource pools.

Note The removal of a storage resource pool association from a storage group may also result in data movement between storage resource pools if the storage group was previously associated with the non-default storage resource pool.

SLO capacity compliance

The SLO capacity compliance algorithm is a capacity based algorithm that ensures all data belonging to thin devices within a particular storage group is allocated across the allowed drive types based on the associated service level objective. This algorithm is only invoked when a storage group's association to a service level objective is modified and data currently resides on a drive type not allowed for the new service level objective.

The allowed drive types for each service level objective are shown in Table 3.

Table 3. SLO capacity compliance drive types

Service Level Objective	EFD	15K RPM	10K RPM	7.2K RPM
Diamond	Yes	No	No	No
Platinum	Yes	Yes	Yes	No

Service Level Objective	EFD	15K RPM	10K RPM	7.2K RPM
Gold	Yes	Yes	Yes	Yes
Silver	Yes	Yes	Yes	Yes
Bronze	No	Yes	Yes	Yes
Optimized	Yes	Yes	Yes	Yes

As an example, if a storage group's service level objective association is changed from Gold to Diamond, any data allocated for that storage group on any spinning drives would be promoted to data pools configured on Flash drives, as this is the only drive type allowed for the Diamond SLO.

Disk resource protection

The disk resource protection algorithm is a performance-based algorithm that aims to protect disk groups and data pools from being overloaded, with particular focus on the higher capacity, lower performing drives.

Each disk group can be viewed as having two primary resources – performance capability and physical capacity.

The performance capability is measured in terms of IOs per second (IOPS) and reflects the workload the disk group is capable of handling. Affecting this measurement is the number of drives configured in the disk group, along with the drive type, rotational speed (if applicable), capacity, and configured RAID protection.

The physical capacity is measured in terms to the total amount of data that can be allocated within the data pool configured on the disk group.

This algorithm aims to maintain an operating buffer in regards to both resources for each diskgroup. This is done in such a way as to have overhead available in each disk group to both accept additional data and additional workload should data be moved to the disk group.

The diagram shown in Figure 3 illustrates this concept. The vertical axis displays a disk group's ability to accept additional workload or its need to have workload

removed from it (measured in IOPS). The horizontal axis displays a disk group's ability to accept additional data from a capacity perspective.

The ideal operating quadrant for each disk group is the upper right-hand quadrant, where the disk group has the ability both to accept additional allocations and additional workload.

The remaining quadrants show situations where FAST will attempt to move data out of a disk group. Greater priority is placed on moving data from disk groups that need to remove IOPs.

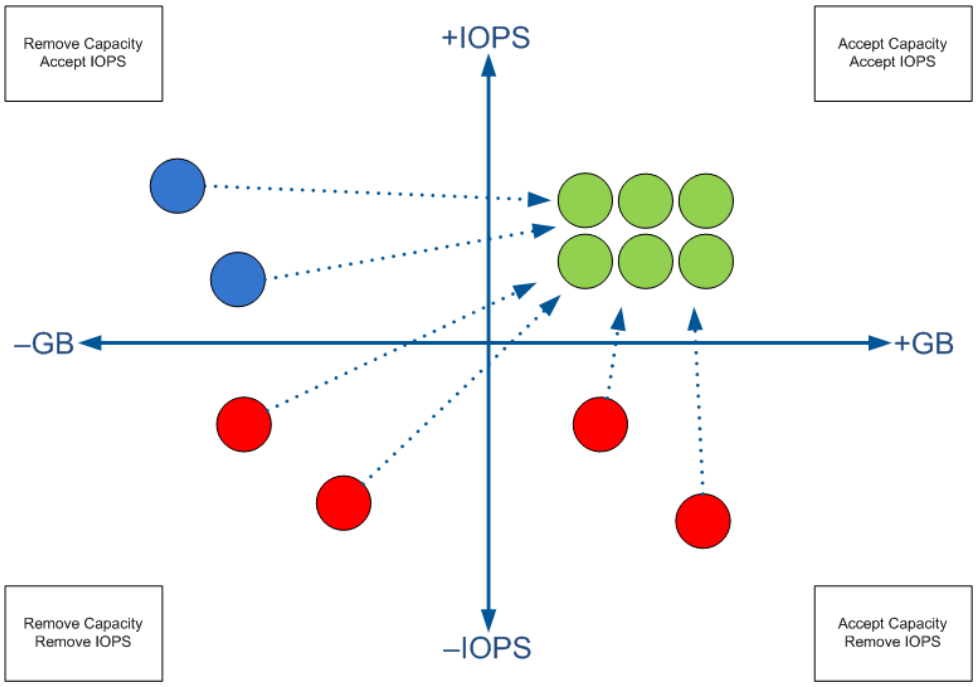

Figure 3. Disk resource protection algorithm

When moving data between disk groups to protect these resources, FAST attempts to place data on the most appropriate media, based on drive technology and RAID protection.

Heavy read workloads are targeted for movement to higher performing drives, for example EFD. Write heavy workloads are targeted for movement to more write-friendly data pools, for example a RAID 1 protected data pool configured on 15K RPM drives.

Allocated extents with little to no workload will be targeted for movement to higher capacity, but lower performing, drives.

The disk resource protection algorithm provides the basis for the default Optimized service level objective.

SLO response time compliance

The SLO response time compliance algorithm is a performance-based algorithm that provides differentiated performance levels based on service level objective associations.

This algorithm tracks the overall response time of each storage group that is associated with a service level objective and then adjusts data placement to the achieve or maintain the expected average response time target.

FAST uses a response time compliance range when determining if data needs to be relocated. When the average response time for the storage group is above the desired range, FAST will promote active data to highest performing data pool, based on available resources in that pool. This promotion activity continues until the average response time is back within the desired operation range.

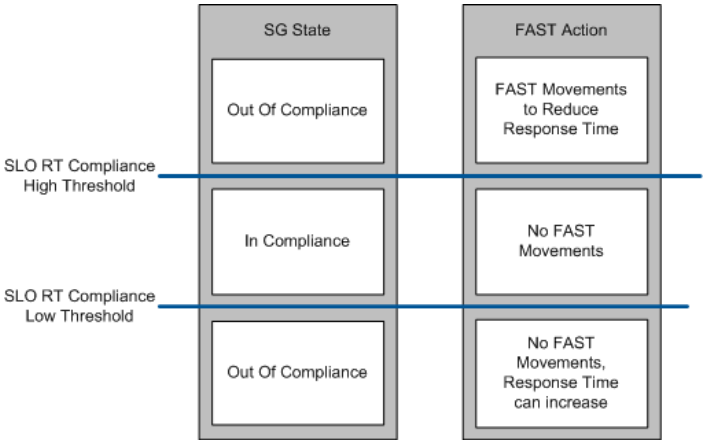

Figure 4. SLO response time compliance algorithm

Data may also be relocated between spinning drives to achieve the service level objective response time target, but this movement will be determined by the disk resource protection algorithm.

Note The use of the SLO response time algorithm only applies to storage groups that are associated with the 'metal' service level objectives – Platinum, Gold, Silver, and Bronze.

Allocation management

New extent allocations, as a result of a host write to a thin device, can come from any of the data pools within the storage resource pool to which the device is associated. FAST directs the new allocation to come from the most appropriate pool within the storage resource pool. This is based on each data pool's ability to both accept and handle the new write as well as the service level objective to which the device the allocation is being made for is associated with.

Each data pool within the storage resource pool has a default ranking based on drive technology and RAID protection type in terms of their ability to better handle write activity within the array. This default ranking is used when making allocations for devices managed by the Optimized SLO.

Due to the drive types that are available for each service level objective, the default ranking is modified for devices managed by service level objectives s other than Optimized. For more information, see Table 3 SLO capacity compliance drive types on page 17.

As an example, consider a storage resource pool configured with the following data pools:

- RAID 5 (3+1) on EFD
- RAID 1 on 15K RPM drives
- RAID 5 (3+1) on 10K RPM drives
- RAID 6 (6+2) on 7.2K RPM drives

Table 4 shows examples of the data pool ranking for new allocations for each service level objective based on the configuration of the storage resource pool.

Table 4. Example data pool ranking for new allocations

Default	Diamond	Platinum or Gold	Silver or Bronze
1. RAID 1 on 15K	1. RAID 5 on EFD	1. RAID 1 on 15K	1. RAID 1 on 15K
2. RAID 5 on 10K	2. RAID 1 on 15K*	2. RAID 5 on 10K	2. RAID 5 on 10K
3. RAID 5 on EFD	3. RAID 5 on 10K*	3. RAID 5 on EFD	3. RAID 5 on EFD*
4. RAID 6 on 7.2K	4. RAID 6 on 7.2K*	4. RAID 6 on 7.2K	4. RAID 6 on 7.2K

- *used only as necessary

As the Diamond SLO only allows extents to be allocated on EFD, the remaining pools in the ranking will only be used in the event that the EFD data pool is full. After the allocation is made to a non-EFD pool, the SLO capacity compliance algorithm will attempt to move the extent into EFD after space has been made available in that pool.

Somewhat similarly, as the Bronze SLO does not allow extents to be allocated on EFD, new extent allocations will only come from the EFD pool when the 15K and 10K pools within the storage resource pool are full. However, the EFD pool will be allocated from before the 7.2K pool as this is more beneficial to the overall performance health of

the storage array. In the case where an allocation for a Bronze SLO managed device is made in an EFD pool, the SLO capacity compliance algorithm will later move that data to the 7.2K pool.

New allocations will always be successful as long as there is space available in at least one of the data pools within the storage resource pool to which the device is associated.

SRDF and EMC FAST coordination

While FAST has the ability to manage devices replicating between VMAX family storage arrays, it must be considered that FAST data movements are restricted to the storage array upon which FAST is operating.

While an SRDF R1 device typically undergoes a read-and-write workload mix, the corresponding R2 device only sees a write workload (reads against the R1 device are not propagated across the SRDF link). A potential consequence of this is that the R2 device data may not be located on the same drive type as the related data on the R1 device.

SRDF coordination for FAST allows R1 performance metrics to be used when making movement decisions for data belonging to an R2 device.

For VMAX3 storage arrays, SRDF coordination is enabled at the system level and is always enabled.

Note SRDF coordination for FAST requires that both the local and remote VMAX storage arrays are running a minimum of 5977 code.

SRDF coordination is supported for single and concurrent pairings in synchronous, asynchronous, or adaptive copy mode.

Effective performance score

On a periodic basis, the collected FAST performance metrics for R1 devices are transmitted across the SRDF link to the corresponding R2 devices.

On the R2 devices, the R1 performance metrics are merged with the actual R2 metrics. This creates an effective performance score for the data on the R2 devices.

When data movement decisions are made on the remote array, the effective performance score is used for the R2 data, thereby allowing the R1 workload to influence the movement of the R2 data. Data that is heavily read on the R1 device is likely to be moved to higher performing drives, depending on the service level objective to which the R2 device is associated.

SRDF coordination considerations

When the SRDF link between the R1 and R2 devices is not ready, the R1 performance metrics are not transmitted to the R2 device. When the link is restored, the metrics are transmitted again.

The SRDF link is considered to be not ready when the SRDF pair state is in one of the following states:

- Split
- Suspended
- Failedover
- Partitioned

During the period that the metrics are not being sent, the workload being seen locally on the R2 device will influence data movements performed by FAST on that device.

If an SRDF personality swap is performed, performance metrics will automatically be transmitted in the opposite direction, provided the SRDF link is in a ready state.

FAST Configuration Parameters

FAST has multiple configuration parameters that control the interaction of FAST with both local and remote replication. These parameters manage the use of capacity within the storage resource pool for use with TimeFinder features as well as SRDF features..

Reserved Capacity

Both TimeFinder snapshot data and SRDF/A DSE related data are written to data pools within a storage resource pool. The reserved capacity parameter allows for the reservation of a percentage of the storage resource pool for use for thin device host write IO activities. No further allocations will be made for TimeFinder or DSE activities when the free capacity of the storage resource pool falls below the reserved capacity value.

As an example, if the reserved capacity on a storage resource pool is set to 10%, new allocations related to TimeFinder or DSE activity will stop when the used capacity of the storage resource pool reaches 90%.

The reserved capacity is set as a percentage on each storage resource pool. Valid values for this percentage range from 1 to 80, or can be set to NONE to disable the reserved capacity. By default, the reserved capacity is set to 10 percent.

Usable by SRDF/A DSE

As previously stated, all SRDF/A DSE related data is written to data pools within a storage resource pool. As such, there needs to always be a storage resource pool that is assigned to be usable by SRDF/A DSE for related allocations.

Usable by SRDF/A DSE is Enabled or Disabled at the storage resource pool. It may only be enabled on one storage resource pool at a time.

By default, the storage resource pool that is designated as the default in the system will be enabled for use by DSE.

Enabling Usable by SRDF/A DSE on a different storage resource pool will automatically set the flag on the storage resource pool currently being used to disabled.

DSE Maximum Capacity

While the use of capacity within a storage resource pool for allocations related to TimeFinder and DSE can be restricted using the reserved capacity parameter, the capacity available for DSE can be further restricted using DSE maximum capacity.

Set at the array level, this parameter sets the maximum capacity that can be consumed by use of SRDF/A DSE in the storage resource pool designated for use in a spillover scenario.

The DSE maximum capacity is set as an absolute capacity in GigaBytes (GB). Valid values for this capacity are from 1 to 100,000, or can be set to NOLIMIT to disable it.

FAST Interoperability

FAST operates seamlessly alongside all VMAX3 and HYPERMAX OS features including Virtual Provisioning and Auto-provisioning groups.

FAST is also fully interoperable with all VMAX3 replication technologies: EMC SRDF, EMC TimeFinder/Clone, TimeFinder VP Snap, TimeFinder SnapVX, and Open Replicator. Any active replication on a VMAX3 device remains intact while data from that device is being moved. Similarly, all incremental relationships are maintained for devices for which data has been moved.

Virtual Provisioning

Each thin device may only be associated with a single storage resource pool. All host-write-generated allocations, or user-requested pre-allocations, are performed on this storage resource pool. FAST data movements only occur within the associated storage resource pool.

It is possible to change the storage resource pool association of a thin device. Doing so results in all extent allocations belonging to that device being migrated to the newly associated storage resource pool.

Thin device creation

When a thin device is created, it is implicitly associated with the default storage resource pool and will be managed by the default Optimized SLO. As a result of being associated with the default storage resource pool, thin devices are automatically in a ready state when they are created.

Optionally, during the creation of a thin device, the device may be added to an existing storage group. The thin device will then be associated to the storage resource pool managed under the service level objective associated with the storage group, if any.

No extents are allocated during the process of creating a thin device or associating the device to a storage resource pool. Extents are only allocated as a result of a host write to the thin device or a preallocation request.

Virtual Provisioning space reclamation

Space reclamation may be run against a thin device that is associated with a service level objective or a storage resource pool, or both.

Virtual Provisioning T10 unmap

Unmap commands can be issued to thin devices associated with a service level objective or a storage resource pool, or both.

The T10 SCSI unmap command for thin devices advises a target thin device that a range of blocks are no longer in use. If this range covers a full thin device extent, then that extent can be deallocated, and the free space is returned to the relevant data pool.

If the unmap command range covers only some tracks in an extent, those tracks are marked Never Written by Host (NWBH). The extent is not deallocated. However, those tracks do not have to be retrieved from the drive should a read request be performed. Instead, the VMAX3 array immediately returns all zeroes.

Auto-provisioning Groups

Storage groups created for the purposes of Auto-provisioning FBA devices may also be used for FAST. However, while a device may be contained in multiple storage groups for the purposes of Auto-provisioning, the device may only be contained in one storage group that is associated with a service level objective or a storage resource pool.

If separate storage groups are created for the purposes of applying separate FAST service level objectives, then these groups can be added to a parent storage group, using the cascaded SG feature. A masking view can then be applied to the parent SG, provisioning both sets of devices.

Note Service level objectives and storage resource pools may only be associated to storage groups containing devices. A parent SG, containing other storage groups, cannot be associated to a service level objective or a storage resource pool.

SRDF

SRDF devices, R1 or R2, can be associated with a service level objective and a storage resource pool. Extents of SRDF devices can be moved between drive types within a storage resource pool while the devices are being actively replicated, in synchronous, asynchronous, or adaptive copy mode.

Note For more information on using FAST with SRDF devices, see *SRDF Coordination* on page 23.

TimeFinder/Clone

Both the source and target devices of a TimeFinder/Clone session can be associated with a service level objective and a storage resource pool. However, both the source and target are managed independently, and, as a result, may end up with a different distribution across drive types.

TimeFinder VP Snap

Both the source and target devices of a TimeFinder VP Snap session can be associated with a service level objective and a storage resource pool.

Target devices sharing allocations may be associated with different service level objectives.

TimeFinder SnapVX

The source device in a TimeFinder SnapVX session can be associated with a service level objective and a storage resource pool. Similarly, linked target devices can be associated with a service level objective and a storage resource pool.

Extent allocations related to snapshot deltas are managed by the Optimized service level objective.

Note For more information on TimeFinder SnapVX, see the *VMAX3 Local Replication Suite TimeFinder SnapVX and TimeFinder Emulation* technical note available at http://support.emc.com.

Open Replicator

The control devices in an Open Replicator session, push or pull, can be associated with a service level objective and a storage resource pool.

Best practice recommendations

The following sections detail best practice recommendations for planning the implementation of a FAST environment.

The best practices documented are based on features available in Enginuity 5977.497.471, Solutions Enabler V8.0.1, and Unisphere for VMAX V8.0.1.

Storage resource pool configuration

Storage resource pools are preconfigured within the storage array and their configuration cannot be modified using management software. As such, it is important that the design created for the storage resource pool during the ordering process uses as much information as is available.

EMC technical representatives have access to a utility call Sizer that can estimate the performance capability and cost of mixing drives of different technology types, speeds, and capacities, within a VMAX3 storage array.

Sizer can examine performance data collected from older-generation VMAX and Symmetrix storage arrays and can model optimal array configurations (both for performance and cost). It will also include recommendations for service level objectives for individual applications, dependent on the performance data provided.

The configurations created by Sizer include the disk group/data pool configurations, including drive type, size, speed, and RAID protection, required to provide the performance capability to support the desired service level objectives.

EMC recommends the use of a single storage resource pool, containing all disk groups/data pools configured within the VMAX3. In this way, all physical resources are available to service the workload on the storage array.

Creating multiple storage resource pools will separate, and isolate, storage resources within the array. Based on use case, however, this may be appropriate for certain environments. EMC representatives should be consulted in determining the appropriateness of configuring multiple storage resource pools.

Service level objective selection

The more information that is available for the applications being provisioned on the VMAX3 storage array, the easier it will be to select a service level objective to associate with that application.

Applications that are being migrated from older storage should have performance information available, including average response time and average IO size. This information can simply be translated to a service level objective and workload type combination, thereby setting the performance expectation for the application and a target for FAST to accomplish.

If little is known about the application, having the default Optimized SLO allows for FAST to take the most advantage of the resources within the storage array and provide the best performance for the application based on the availability and workload already running on those resources.

Associating a non-default service level objective to an application, thereby setting a response time target for that application, can limit the amount of capacity allocated on higher performing drives. Once an application is in compliance with its associated service level objective, promotions to higher performing drives will stop. Subsequent movements for the application will look to maintain the response time of the application below the upper threshold of the compliance range.

Storage group configuration

In order to provide the most granular management of applications, it is recommended that each application be placed in its own storage group to be associated to a service level objective. This provides for more equitable management of data pool utilization and ensures that FAST can manage to the response time target for the individual application.

Cascaded storage groups

In some cases, there may be a need to separately manage different device types within a single application. For example, it may be desired to apply different service level objectives to both the redo log devices and the data file devices within a database application. To accomplish this, it is recommended that cascaded storage be used.

Cascaded storage groups allow devices to be placed in separate child storage groups which can then be associated with the same parent storage group. Separate service level objectives can be associated with each child storage group, while the parent storage group is used in a masking view for the purposes of provisioning the devices to host.

As additional capacity is added to the application by way of adding devices, the devices may be added to the appropriate child storage group (based on the desired service level objective). Upon adding the devices to a child storage group, they will be automatically mapped and masked based on the parent storage group's masking view.

Storage group device movements

Depending on requirements, it may be necessary to change the service level objective of an individual device which may require moving the device to another storage group.

Device movement between storage groups with differing service level objectives is allowed and may be performed non-disruptively to the host if the movement does not result in a change to the masking information for the device being moved between groups. That means, following the move, the device is still visible to the exact same host initiators on the same front-end ports as before the move.

Devices may also be moved between child storage groups who share the same parent, where the masking view is applied to the parent group.

SRDF

While FAST has the ability to manage devices replicating between VMAX family storage arrays, it must be considered that FAST data movements are restricted to the storage array upon which FAST is operating.

SRDF coordination for FAST allows R1 performance metrics to be used when making movement decisions for data belonging to an R2 device.

For VMAX3 storage arrays, SRDF coordination is enabled at the system level and is always enabled. However, SRDF coordination for FAST requires that both the local and remote VMAX storage arrays are running a minimum of 5977 code.

Each SRDF configuration presents its own unique behaviors and workloads. As SRDF coordination is only supported between VMAX3 arrays, the information in the following sections should be considered prior to implementing FAST in an SRDF environment where the remote array is running a version of Enginuity less than 5977.

FAST behavior with SRDF

While an SRDF R1 device typically undergoes a read-and-write workload mix, the corresponding R2 device only sees a write workload (reads against the R1 device are not propagated across the SRDF link). A potential consequence of this is that the R2 device data may not be located on the same drive type as the related data on the R1 device.

In this scenario, if there are R1 device extents that only experience read activity, then the corresponding extents on the R2 devices will see no I/O activity. This will likely lead to these R2 device extents being demoted to the lowest tier included in the FAST VP policy associated with the R2 device.

SRDF operating mode

EMC best practices, for both synchronous and asynchronous modes of SRDF operation, recommend implementing a balanced configuration on both the local (R1) and remote (R2) Symmetrix arrays. Ideally, data on each array would be located on devices configured with the same RAID protection type, on the same drive type.

As FAST operates independently on each array, and also promotes and demotes data at the sub-LUN level, there is no guarantee that such a balance may be maintained.

In SRDF synchronous (SRDF/S) mode, host writes are transferred synchronously from R1 to R2. These writes are only acknowledged by the host when the data has been received into cache on the remote R2 array. These writes to cache are then destaged asynchronously to disk on the R2 array.

In an unbalanced configuration, where the R2 data resides on a lower-performing tier than on the R1, performance impact may be seen at the host if the number of write pendings builds up and writes to cache are delayed on the R2 array.

In a FAST environment, this typically does not cause a problem, as the promotions that occur on the R2 side are the result of write activity. Areas of the thin devices under heavy write workload are likely to be promoted and maintained on the higher-performing drives on the R2 array.

In SRDF asynchronous (SRDF/A) mode, host writes are transferred asynchronously in predefined time periods or delta sets. At any given time, there are three delta sets in effect: The capture set, the transmit/receive set, and the apply set.

A balanced SRDF configuration is more important for SRDF/A, as data cannot transition from one delta set to the next until the apply set has completed destaging to disk. If the data resides on lower-performing drives on the R2 array, compared to the R1, then the SRDF/A cycle time may elongate and eventually cause the SRDF/A session to drop.

This is similar to SRDF/S mode in most environments, so this may not be a large issue, as the data under write workload is promoted and maintained on the higher-performing drives.

SRDF/A DSE (delta set extension) should be considered to prevent SRDF/A sessions from dropping in situations where writes propagated to the R2 array are being destaged to a lower tier, potentially causing an elongation of the SRDF/A cycle time.

SRDF failover

As FAST works independently on both the R1 and R2 arrays, it should be expected that the data layout will be different on each side if SRDF coordination is not enabled. When an SRDF failover operation is performed, and host applications are brought up on the R2 devices, it should then also be expected that the performance characteristics on the R2 will be different from those on the R1.

In this situation, it can take FAST some period of time to adjust to the change in workload and start promotion-and-demotion activities based on the mixed read-and-write workload.

SRDF bi-directional

Considerations for SRDF change slightly in a bi-directional SRDF environment. In this case, each Symmetrix array has both R1 and R2 devices configured. This means that each array has a local workload sharing resources with a remote workload.

In the situation where both the local and remote array are VMAX3, with SRDF coordination enabled, it is possible that data belonging to an R2 device could displace data belonging to an R1 device on higher-performing drives. This may happen if the R2 device's corresponding R1 device has a higher workload than the local R1 device or a better performing service level objective.

In this scenario, the R2 devices may be associated with a lower performing service level objective in order to reserve higher-performing resources for the local R1 workload. Should a failover need to be performed, the service level objective associated with the R2 devices can be changed to one providing better performance.

TimeFinder

FAST is fully interoperable with all VMAX3 local replication technologies, including EMC TimeFinder/Clone, TimeFinder VP Snap, and TimeFinder SnapVX. Any active replication on a VMAX3 device remains intact while data from that device is being moved. Similarly, all incremental relationships are maintained for devices for which data has been moved.

Note For more information on local replication on VMAX3, see the *VMAX3 Local Replication Suite TimeFinder SnapVX and TimeFinder Emulation* technical note available at http://support.emc.com.

TimeFinder/Clone

Both the source and target devices of a TimeFinder/Clone session can be associated with a service level objective and a storage resource pool. FAST manages both the source and target devices independently, based on the workload being seen on each device.

Extent allocations owned by the source device are managed by the service level objective associated with the source device, while those owned by the target are managed by the target device's service level objective.

For clone sessions in the *CopyOnWrite* state, read I/Os to the target device may be redirected to the source device. This read activity is counted against the source device.

Because of this, heavy read activity against a target device may influence movement decisions made by FAST on the source device. The result of this may be to move data belonging to the source to a set of higher performing drives in its storage resource pool.

Similarly, read I/Os to the target device that are redirected to the source will be subject to the source's service level objective. This may impact the performance of the workload running against the target and cause the clone target to miss its service level objective.

To avoid this, it is recommended that the target devices in a clone session created in *nocopy* mode should be managed by a service level objective of equal or lesser performance than the source devices.

In the case of clone sessions created in *copy* mode, where all the allocations for the device will be owned by the target device, the appropriate service level objective can be set based on the performance requirements for the target.

TimeFinder VP Snap

Both the source and target devices of a TimeFinder VP Snap session can be associated with a service level objective and a storage resource pool.

Similar to TimeFinder/Clone, extent allocations owned by the source device are managed by the service level objective associated with the source device, while those owned by the target are managed by the target device's service level objective.

Read I/Os to the target device may be redirected to the source device. This read activity is counted against the source device. As such, heavy read activity against a target device may cause extents owned by the source device to be moved to higher performing drives.

Redirected read I/Os to the target device are subject to the service level objective associated with the source device which may have an impact on the performance of the target device's workload.

To avoid the target device missing its associated service level objective, it is recommended that the target be managed by a service level objective of equal or lesser performance than that of the corresponding source devices.

TimeFinder SnapVX

The source device in a TimeFinder SnapVX session can be associated with a service level objective and a storage resource pool. Similarly, linked target devices can be associated with a service level objective and a storage resource pool.

As with TimeFinder/Clone and TimeFinder VP Snap, extent allocations owned by the source device are managed by the service level objective associated with the source device, while those owned by the linked target are managed by the target device's service level objective.

Read I/Os against the target device that are redirected to the source device will be counted against the source device with regard to the FAST performance metrics. This read activity against the linked target can influence the data placement of extents owned by the source device.

These redirected read I/Os to the target device are also subject to the service level objective associated with the source device.

To avoid performance impact to the target device, causing it to miss its associated service level objective, it is recommended that the target be managed by a service level objective of equal or lesser performance than that of the corresponding source devices.

Extent allocations related to snapshot deltas (unlinked snapshots) are managed by the Optimized service level objective. These extents will see no host I/O until a target is linked. As such, over time, these allocations will be moved to the more cost effective drives within the storage resource pool the source device is associated with.

Conclusion

The VMAX3 family delivers unmatched ease of provisioning for your specific service level objectives. These service levels are tightly integrated with EMC's Fully Automated Storage Tiering (FAST) technology to optimize agility and array performance across all drive types in the system. FAST improves system performance while reducing cost by leveraging high performance Flash drives combined with cost effective high capacity drives.

EMC FAST dynamically allocates workloads across storage technologies, non-disruptively moving workloads to meet stringent service level objectives. FAST moves the most active parts of your workloads to high-performance flash drives and the least-frequently accessed data to lower-cost drives, leveraging the best performance and cost characteristics of each different drive type. FAST delivers higher performance using fewer drives to help reduce acquisition, power, cooling, and footprint costs.