Technical Paper

Best Practices for SAS® on EMC® SYMMETRIX® VMAX™ Storage

Table of Contents

Introduction

The EMC® SYMMETRIX® VMAX™ Storage System is a powerful, flexible, and easy-to-manage storage subsystem answering the needs of performance, consolidation, and automation for today's SAS® workloads. Virtualization, tiered storage, and automation provide flexible arrangements for the continuum of workloads the typical SAS® shop employs. These workloads have specific storage needs and requirements for maximum application performance. This technical paper will outline best practices for architectural setup and tuning to maximize SAS Application performance with EMC® SYMMETRIX® VMAX™ storage.

An overview of the storage system will be discussed first, including physical and virtual architecture, pooling and virtualization, storage tiers and management. This will be followed by a list of practical recommendations for implementation with SAS®.

BRIEF OVERVIEW OF VMAX™ ARCHITECTURE

The EMC® Symmetrix VMAX™ provides a Virtual Matrix™ architecture to scale performance and capacity via common building blocks called Symmetrix VMAX™ engines. Each engine has dual integrated Virtual Matrix Directors providing its own CPU, Memory, and Cache resources, along with front end (to host) and back end (to physical storage) ports. Each Directory front end FA supports 4 front end ports, serviced by 2 CPUs (2 ports per CPU). FA ports have finite Write I/O limitations (FA Port Performance has noticeable degradation beyond 30,000 IOPs, and is maximized at 50,000 IOPs); FA resourcing is extremely important for large workloads and will be discussed in more detail later.

Up to 8 engines can be employed in system for scale-out, completely interconnected between the Virtual Matrix™, providing local engine CPU, Memory, Cache utilization, as well as globally sharing those engine resources across the system. Most systems generally start with 1 to 2 engines, and scale as capacity and performance requires. Balancing systems is crucial as engines are added to avoid performance bottlenecks. See Figure 1 Below.

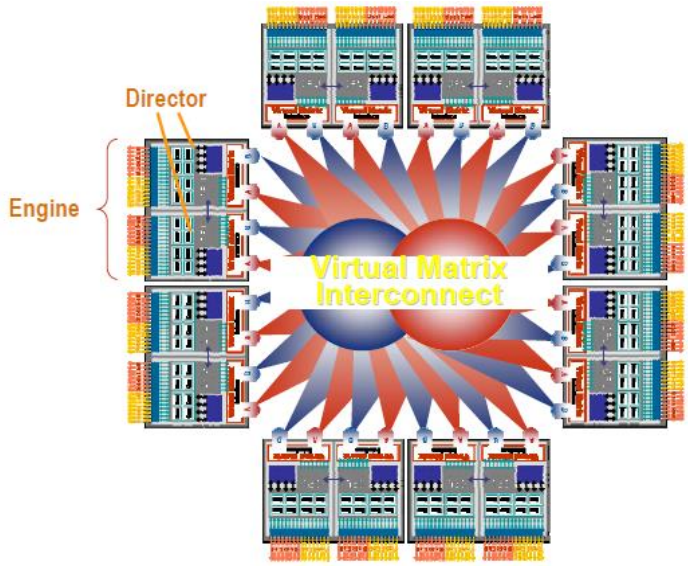

Figure 1. EMC® Virtual Matrix™ Architecture

PHYSICAL STORAGE – DISK TYPES, FA PORTS, STORAGE ADAPTERS

The underlying physical storage in the VMAX™ system consists of 3 storage basic tiers, listed below from highest performing to lowest performing drive technologies in terms of performance:

- Tier 1 – EMC Flash Drives, (EFD)
- Tier 2 – Fibre Channel Drives
- Tier 3 – SATA II 7200 RPM

Note: ** Drive capacities may vary by installation. Check with EMC representative for your configuration

These storage tiers provide the least (Tier 3 - SATA) to the highest (Tier 1 - FLASH) performance and cost per Gigabyte. SATA devices are large, slower drives, with much higher response times than FC or FLASH devices. Read Miss response times are in the 12ms range compared to Flash Drives at 1 ms or 15K FC devices at 6 ms.

File Systems can be placed on tiers appropriate for the performance requirements they service. Automated tiering is available through EMC® Fully Automated Storage Tiering (FAST™), which will be discussed in detail later. The storage tiers can be implemented with various RAID levels (RAID 1, RAID5 (3+1) or (7+1) and RAID 6 (6+2) or (14+2)). Disks can also be placed into a RAID virtual architecture, in which RAID levels can be virtually switched.

The following configuration “best practices” should be considered:

- Tier 2 – Never inter-mix FC rotational speeds. Use all 15K FC or all 10K FC
- Use the same RAID protection type within a tier
- Use the same storage capacity drives within a tier

BASIC LUN PRESENTATION

There are two basic LUN types, Back End and Front End. Back End LUNs are created from physical drives, and can be presented to the server as a physical LUN.

Front End LUNs are grouped from back-end LUNs to create larger front-end entities (similar to a logical volume created by a volume manager). They can be concatenated or striped.

LUNs have finite sizes. In order to create a high capacity LUN (e.g. > 240 GB), a Metavolume can be created. A Metavolume is a type of front-end LUN. It is composed of two or more Hypervolumes (logical volumes configured from slices of physical drives). Metavolumes can be either striped or concatenated. We recommend striped Metavolumes, when Metavolumes are employed for SAS usage.

A Striped Metavolume is created by combining back end Raid 1 LUNs into a single front end volume. One the front end volume the data is then striped using a 960KB stripe size. The net result is a R1/0 LUN, (mirrored and then striped with no parity). This is commonly called a R10 Striped Metavolume

Single LUNs are usually presented from single FA Ports. Striped Metavolumes can be spread across multiple FA ports for throughput aggregation. This can have significant performance ramifications which will be discussed below.

VMAX™ VIRTUAL PROVISIONING

VMAX™ Virtual Provisioning is based on thin pools. The EMC® SYMMETRIX® VMAX™ system introduces 2 new device types to support virtualization:

- TDAT, or thin data device, is an internal LUN which is assigned to a thin pool. Each TDAT LUN is comprised of multiple physical drives configured to provide a specific data protection type. An example of a TDAT device might be a Raid-5(3+P) LUN. Another common example would be a Raid – 1 mirrored LUN. Multiple TDAT devices are assigned to a pool. When creating a thin pool LUNs for presentation to a host the data is striped across all of the TDAT devices in the pool. The pool can be enlarged by adding devices and rebalancing data placement (background operations with no impact to the host). Care must be taken to monitor pools as filling up pools will freeze them.

- TDEV – (thin pool LUN), is a host accessible (redundantly presented to an FA port) back-end LUN device that is “bound” to a thin device pool (TDATs) for its capacity needs. As stated above, a TDEV is a host presentable device that is striped across the back end TDAT devices. The stripe size is 768K. Each TDEV is presented to an FA port for host server allocation. When utilizing thin provisioning, Thin Pool LUNs are employed. The utilization of TDEVs is required to use EMC® Fully Automated Storage Tiering Virtual Provisioning (FAST/VP™) features.

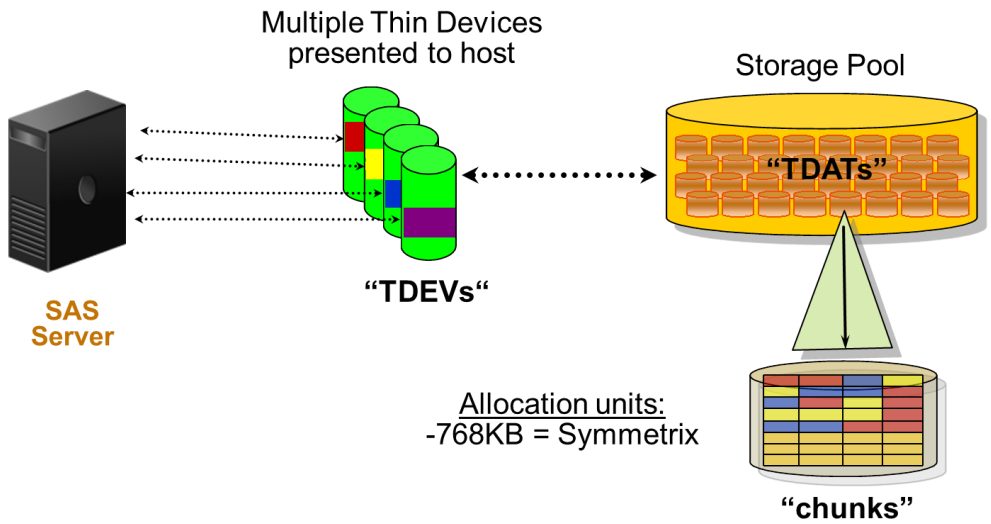

Figure 2. EMC® Thin Provisioning

The wide striping across the virtual pools automatically distributes the data stripe across the back-end devices. Storage Admins no longer have to do this manually. Pool rebalancing evenly redistributes the chunks when necessary, without changes to the virtual LUNs presentation. See Figure 2 above.

MANAGEMENT POOLS

FAST/VP™ Management Pools are setup and managed via the Symmetric Management Console™ (SMC) or the SYMCLI™ (command line interface). They can represent any combination of thin provisioned LUNs you wish to define them with, and are generally capacity based. These management pools form the entities that FAST/VP™ uses to monitor and manage automated tiering within.

EMC FULLY AUTOMATED STORAGE TIERING/VIRUTAL PROVISIONING – FAST/VP™

HOW IT WORKS

FAST/VP™ allows automated storage tiering that sets up quickly, and allows tier promotion and demotion based on live experience via sub-LUN level migration (see diagram below). The three main elements within a FAST/VP™ Management Pool are storage tiers, storage groups, and policies. It is important to note this migration can involve differing RAID protection types transparently. Please use the RAID protection type most suited to your data safety and recovery needs, as well as your performance requirements. Most SAS® shops employ their LUNs in RAID 5 or RAID 1/0 as their safety level.

FAST/VP™ monitors VP LUN utilization and moves the busiest thin extents to appropriate VP pools located on various drive technologies. It also moves underutilized thin extents to pools located on high-capacity drives. Because the unit of analysis and movement is measured in thin extents, this sub-LUN optimization is precise and efficient.

AUTOMATED TIERING

The **storage tiers** are the combinations of drive technology (type – e.g. SATA II, FC, SSD) and RAID protection level (e.g. RAID 5, RAID 1, RAID 6). A tier 1 may consist of SSDs striped in a RAID 5 configuration, a Tier 2 with 15K rpm Fiber Channel Disks striped in a RAID 1 or RAID 5, and a Tier 3 with 7K rpm SATA II Disks striped in a RAID 6 Configuration. The Disk technologies and the RAID levels can be mixed and matched as allowed by the system. The general idea is that you can get higher performing tiers of storage with the various disk types and the RAID protection levels chosen for them. You can choose which types of tiers to employ and construct to fit your data operations. Fast tiers can be used for production applications strict service levels, and slower performing tiers can be allocated to less crucial operations that aren't as time dependent.

Storage Groups are a collection of Symmetrix™ host-addressable devices (e.g. the TDEVs described above). An example of a storage group may be all the devices (LUNs) provisioned to a SAS System.

Lastly the **Policies** are the rules and regulations put into place to manage movement of data across the tiers. FAST policies tie Storage Groups to Storage Tiers; and define the configured capacities as a percentage, that a Storage Group is allowed to consume on that tier (e.g. SSD, FC, and SATA). For example a capacity policy for a storage group might read as:

Tier 1: 20 -30%
Tier 2: 100%
Tier 3: 30%

The above policy would be interpreted as maximum of 20 – 30% of the storage groups extents could be migrated to Tier 1 (SSD) storage if the policy interpreted it could benefit from it; 100% could reside on Tier 2 (FC) and up to 30% could reside on Tier 3 (SSD).

The combination of the percentages applied to the three tiers must add up to at least 100% of there will be a shortage of storage allocation. In nearly all cases the combination of the percent storage allocations will exceed 100%.

Scoring Process

FAST/VP™ monitors the storage group usage and performance on the tiers and up-tiers parts of the data when they meet the pre-set policy for promotion (moving to a higher performing tier), and vice-versa for demotion. If part of the policy states that certain data cannot be moved, it will not be migrated. FAST/VP™ policies can be changed by the Storage Administrator at any time to react to changing demand or performance preferences. Movement is managed by capacity via Management Pools.

The FAST/VP™ Engine monitors performance at the Extent Level, and migrate promoted or demoted data at the sub-extent level. See Figure 3 below.

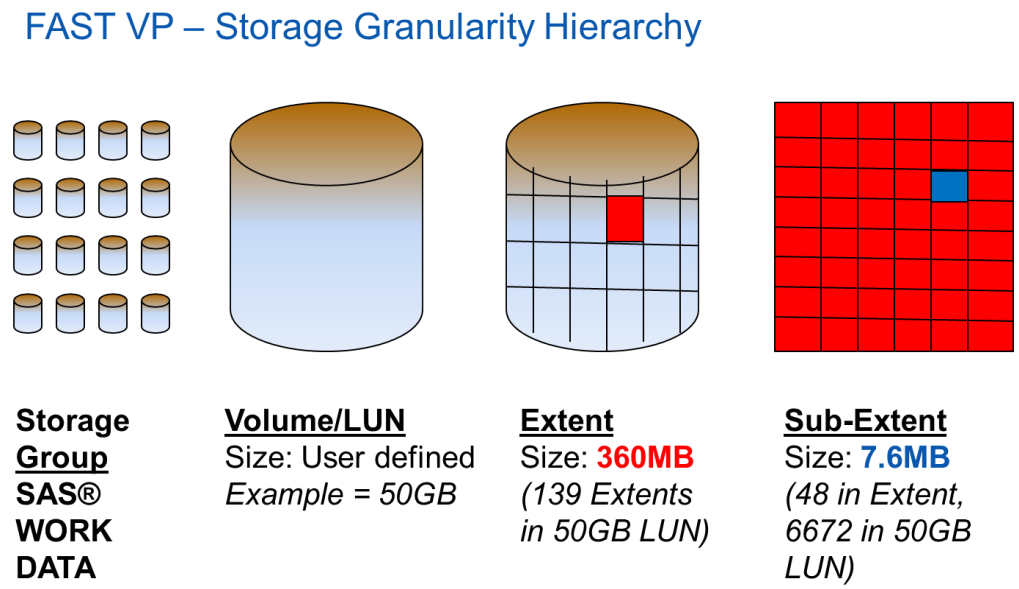

Figure 3. Migrations are made in small chunks at the Sub-Extent level

Host Usage counters are accumulated in 10 minute intervals for each extent for three types of I/O operations:

- Read Misses (RM)
- Pre-Fetches (PF)
- Disk Writes (W)

Those three variables are used to create an interval score. I/O operations for creating clones, snaps, etc. are not included in the score. Interval scores are then added to sets of short-term (for promotion) and long-term (for demotion) scores, which then guide the engine on tier promotion or tier demotion of extents. The operation is biased towards making it easier to up-tier than down-tier, to not have an undue effect on performance. Short-term scores for promotion provide quick responsiveness by FAST™ to immediately changing needs, while the long-term score for demotion remembers experience of the past weeks days and hours, and down-tiers low-demand extents.

It is important to note that the FAST moves in promotion and demotion are performed via sub-LUN level extents migration by the Storage Array, and not the FAST engine itself. In the event that the Array does not complete a promotion or demotion within the next 10 minute scoring period, new scores are collected and given to the array. The array will then finish all moves that were previously in progress, but adjust to the new scores for activity.

Beyond FAST/VP™

There are additional technologies in the VMAX™ array that provide resource management and allocation to isolate resources to highly needed areas. Two of these include Symmetrix™ Priority Controls™ and Dynamic Cache Provisioning™.

Symmetrix Priority Controls™ enhance tiered storage management by allowing prioritization of host application read I/O and SRDF/S transfers by assigning a priority level to specific device groups. The task's priority determines its position in the queue. During low demand all requests are satisfied quickly, but in situations of heavy disk contention, the controls provide service differentiation.

Dynamic Cache Partitioning™ allows portions of the cache to be allocated to specific allocation groups. The default in the array is equal distribution. DCP allows application groups that experience heavy cache re-use to benefit by allocating a higher amount of memory to them. Work with your EMC Storage Engineering Team to determine if and when you should employ the above features, based on a careful examination of your workload.

ARCHITECTING THE VMAX™ FOR SAS WORKLOADS

Given the extreme flexibility, automation, virtualization, and power of the EMC®/VMAX™ system, what are the best practices for storage support on SAS Applications? This next section will give generally recommended practices by SAS® and EMC® resulting from multiple field experiences.

Many long-term SAS® Customers have migrated through years of storage technology, across increasingly complex storage systems architectures. Over the long haul, there are standard paradigms that have not changed in terms of successful storage architecture for SAS. These paradigms still apply today, regardless of what storage technology you attempt to implement. The following paper details those considerations: **http://support.sas.com/rnd/papers/sgf07/sgf2007-iosubsystem.pdf** .

There are two pertinent issues when using thin provisioning to virtualize storage for SAS® applications. The first is under-provisioning on spindle alone; the second is over-subscription of the physical capacity. Thin provisioning is often implemented on much larger, slower drives to gain capacity cheaply. Doing this reduces the actual number of disks needed. Since SAS® is I/O throughput oriented (see the paper noted in the previous paragraph), this can be detrimental. Throughput aggregation is attained by striping across many disks, and aggregating the throughput of each disk in a single "stripe" when reading or writing. When fewer disks are involved, lesser throughput can be attained. This is detrimental to large SAS® workloads. If utilizing thin pools, use wide striping to attain higher throughput.

Another goal of thin-provisioning virtualization typically includes providing thin pools of storage that multiple applications (and hence varying workloads and types) will share. The physical provisioning underneath the thin pools is not sufficient to cover the defined "logical space" on defined LUNs. Hence the term "thin". This under-provisioning of the actual total physical space available, to the sum of the Logical LUN spaces provided is called over-subscription. The paradigm banks on the fact that not all virtual customers will hit the thin pool at once, with a workload that would simultaneously demand all their defined logical LUN extents. Unfortunately that is exactly what SAS® ad hoc workload environments are easily capable of engendering. Granted this can happen in any type of array setup, thin or not, the thin pool construction typically makes it more vulnerable for SAS if oversubscription occurs. Overhead, safety, and expansion space on the old thick LUN definition has been replaced by a pool that may be quickly become "oversubscribed". Care and attention must be paid to peak load demand, and adequate coverage provided if thin pool construction is chosen. Otherwise, you may wish to follow the route of not using thin-pool provisioning.

Some large SAS® shops, even when provisioning new virtualized storage arrays have chosen to create thin pools that were not over-subscribed, across most or all the spindles in their array. This works well as long as your demand load and performance expected fit within the architecture you create. Depending on your shop's storage expertise, it will require working closely with your EMC® Storage Engineer for desired performance.

WHEN THIN PROVISIONING AND FAST/VP™ ARE EMPLOYED:

- FAST VP best performance is achieved with the following tier configuration

    - Tier 1 EFD Raid-5(3+P)
    - Tier 2 FC Raid-1 mirrored
    - Tier 3 SATA Raid-6(6+2P)

- A single large thin pool should be used for all devices to be presented as FAST VP Thin Pool LUNs to the host server(s). This pool should be created from the Fibre Channel tier 2. This becomes the "bind pool" in the FAST VP process

- Separate storage groups are created out of the same pool of thin LUNs described above. The storage groups are used to apply FAST VP policies to differing workloads within the SAS implementation

- Utilize as many FA Ports as possible and balance via EMC Powerpath™. Be aware of FA Port performance and provide additional capacity as needed.

- SASWORK is a special use file system that often has a very high, equally mixed, read/write IO content. SASWORK should still use thin pool LUNs and still be part of a FAST VP storage group as described below. SASWORK is usually allocated on a single file system that is often a single LUN. Due the high write content this single LUN can overrun an FA CPU ability to process the write IO. For the SASWORK only, the IO needs to be distributed across multiple FA's. There are two common ways of accomplishing this. First is to simply create a Stripe Meta out of the thin pool LUNs that encompasses multiple FA's, (at least 4). Another method would be to place the thin pool LUNs in a volume, or disk group, using the OS volume manager and stripe the LUNs across the FA's

- Build a minimum of 2 – 3 thin FAST VP Storage Management Pools :

    - A SASWORK Pool for SAS® working space with the following initially recommended FAST Policy tier spreads:
        - 20 – 30 % Tier 1 – Weblogic Flash Drives,
        - 100% - Tier 2 – FC Drives
        - 0% - Tier 3 – SATA II, 7K rpm
    - A SASDATA Pool for SAS® Permanent Data space with the following initially recommended FAST Policy tier spreads:20 – 30 % Tier 1 – Weblogic Flash Drives
        - 100% - Tier 2 – FC Drives
        - 30 - 60%** - Tier 3 – SATA II, 7K rpm i.
    - A Potential 3rd Pool - UTTILOC Pool for SAS Working Utility space. This would segregate the Utility space from the SASWORK file space. It would have the same initially recommended FAST Policy tier spreads as SASWORK:20 – 30 % Tier 1 – Weblogic Flash Drives
        - 100% - Tier 2 – FC Drives
        - 0% - Tier 3 – SATA II, 7K rpm

- Note that in the above allocations, SATA is not used for SAS Working Utility Space. It can be used for SAS®

DATA in smaller SAS® workloads that do not have a high throughput requirement (e.g. >75 MB/sec). SATA is more appropriate for SAS workloads exhibiting a random access performance profile (e.g. heavy INDEX usage, SAS/OLAP™, and random traversal of data).

THROUGHPUT TESTING

It is wise to perform system throughput testing with new configurations. This ensures you have a finite idea where the systems throughput boundaries are. A throughput program provided by SAS® Technical Support is here: **http://support.sas.com/kb/51/660.html**

CONCLUSION

Modern array technology has incorporated virtualization, enabling thin provisioning and offering automated storage tiering. These offerings are intended to combat high array costs by driving high utilization of resources, and automating much of the array Administration. Unfortunately the goal of driving high utilization is often diametric to having sufficient throughput resources quickly available for large-block I/O applications such as SAS®.

When employing such arrays, the advice in this paper, and the specific vendor papers listed in the Further Reading Section below should be carefully considered. In many instances the new technologies can be utilized effectively, and in others it must be mitigated with appropriate architecture changes, and usage.

Work with your Storage Vendor to ensure you are employing their storage technology to its best affect, and providing high performance to your SAS® applications. There is no substitute for careful planning, and testing, to ensure adequate performance will be provided.

REFERENCES

EMC® Technical Notes – EMC® Symmetrix® V-MAX™ Best Practices. Technical Note P/N 300-009-149 REV A02. November 18, 2009. Copyright © 2009 EMC Corporation All Rights Reserved.

EMC® Symmetrix VMAX™ with ENGINUITY, EMC® Product Description Guide H6544.5. September, 2011. Copyright © 2009 EMC Corporation All Rights Reserved.

RECOMMENDED READING

Usage Note 51660: Testing Throughput for your SAS 9 File Systems: UNIX and Linux platforms
http://support.sas.com/kb/51/660.html

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Steven Bonuchi
Enterprise: EMC Inc., US Central Division
United States
Work Phone: +1(913) 708-3265
E-mail: **steven.bonuchi@emc.com**

Name: Tony Brown
Enterprise: SAS Institute Inc.
Address: 15455 N. Dallas Parkway
City, State ZIP: Dallas, TX 75001
United States
Work Phone: +1(214) 977-3916
Fax: +1 (214) 977-3921
E-mail: **tony.brown@sas.com**

Name: Margaret Crevar
Enterprise: SAS Institute Inc.
Address: 100 SAS Campus Dr
Cary NC 27513-8617
United States
Work Phone: +1 (919) 531-7095
Fax: +1 919 677-4444
E-mail: **margaret.crevar@sas.com**