

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Upon completion of this module, you should be able to:

- Relate backend architecture to performance capabilities
- Use the Symmetrix architecture and performance features to create tiers of performance
- Explain the performance issues related to Symmetrix data protection schemes
- Measure backend performance and utilization
- Take advantage of Command Tag Queuing optimization



This module focuses on analyzing the performance of Symmetrix VMAX backend directors.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 1: Symmetrix Backend Architecture

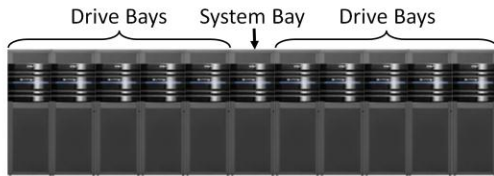
During this lesson the following topics are covered:

- Backend architecture to performance capabilities
- Symmetrix architecture and performance features to create tiers of performance
- Disk size and speed when choosing disks to meet capacity and performance needs



This lesson covers concepts of the architecture of the Symmetrix VMAX backend adapters.

Symmetrix VMAX Backend Configuration



Largest VMAX has 10 Drive Bays = 2,400 Drives
(3200 2.5" drives)



Drive Bay = 16 DAEs
= 240 Drives



DAE = 15 Drives
(25 if 2.5" drive)

EMC²

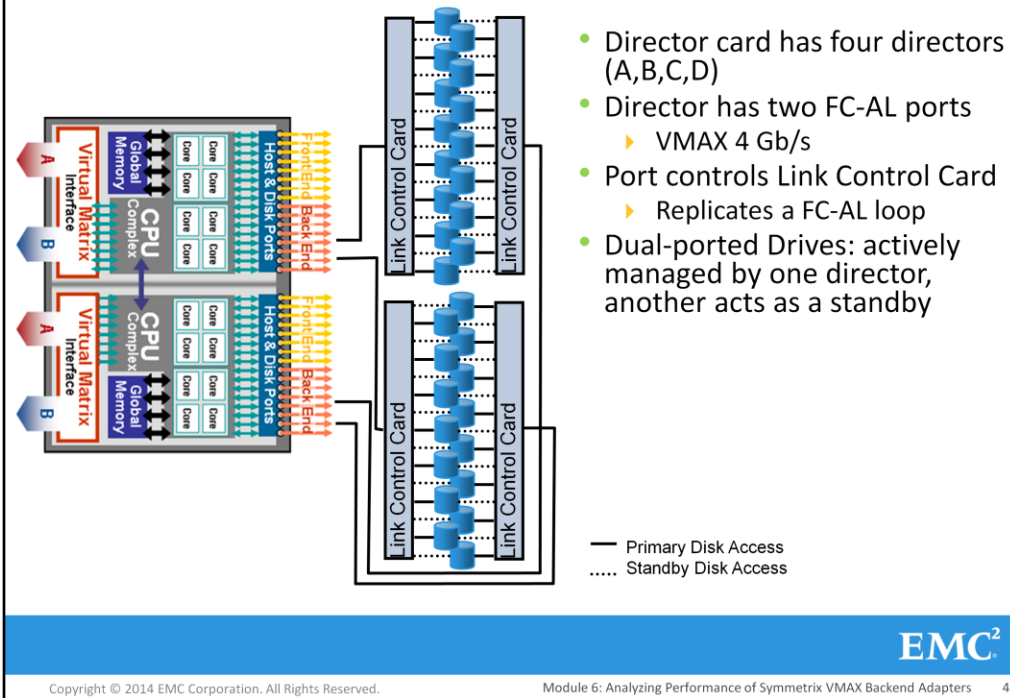
Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 3

The drives in a VMAX array are stored on the array in Disk Array Enclosures (DAE). Each enclosure holds 15 3.5" drives or 25 2.5" drives.

A single drive bay holds 16 enclosures for a total capacity of 240 drives. A full-sized array can have up to ten drive bays containing a total of 2,400 drives (3200 if the 2.5" drives are used).

Symmetrix VMAX Backend Architecture



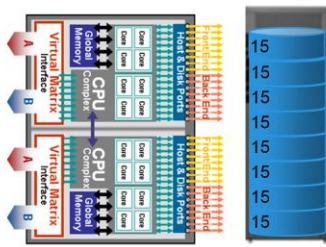
The backend architecture for the Symmetrix VMAX family is illustrated here. Each disk director card has four backend directors, labeled A through D. Each director controls two Fibre Channel Arbitrated Loop ports. Each port is connected to a device that contains circuitry, which replicates the traffic to each drive just as if the drives were physically connected in an arbitrated loop.

Up to the DMX-2, the device was called a Port Bypass Card; with the DMX-3 it is a Link Control Card. Unlike a physical loop, one drive can be replaced without interrupting the activity of the others.

Only half of the drives connected to the Link Control Card are actively managed by the primary director. A similar Link Control Card is connected to the dual-ported drives so that a standby director can also manage those drives. If the primary director fails, the standby director will take responsibility for its drives.

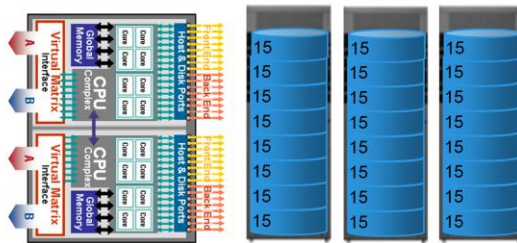
The Link Control Cards replicated the functionality of an arbitrated loop by sending every message from the director to every disk in the loop. All disks would have to view the message even though only one would respond. The DMX-4 array replaced the Link Control Card with Point-to-Point circuitry. The new circuitry performs the same function, but acts like a switched connection, routing a message from the director only to the receiving disk.

Configuration for Performance or Capacity



1 Storage Bay

- Few Drives per Director
 - ▶ 1 Bay = 120 drives (at most) per director pair
- Drive limited
 - ▶ Processors can send enough I/O to fully utilize drives
- Best for performance



3 Storage Bays

- Many Drives per Director
- Add bays to add more drives to backend ports
- Processor limited
 - ▶ Processors may reach high utilization without fully utilizing drives
- Best for Capacity

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 5

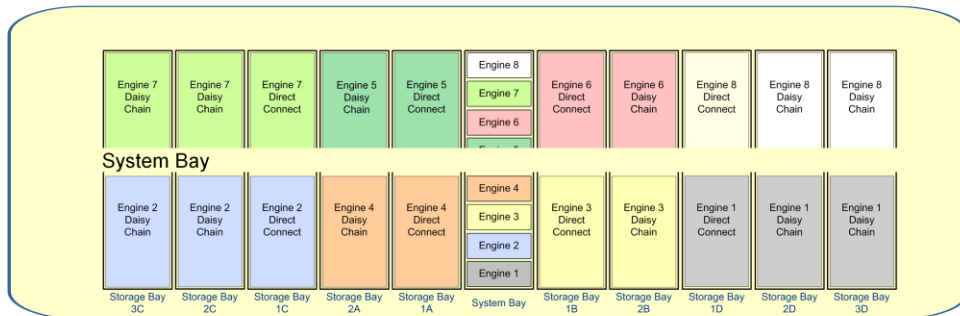
You can configure the VMAX drive subsystem for high performance, large capacity, or strike a balance between the performance and capacity. The number of drives in a director's loop determines the performance/capacity tradeoff.

If the number of drives managed by a director is small, the director processing power will be large compared to the number of drives. The processors are able to issue enough I/Os to keep all of the drives fully utilized, while not becoming over-utilized themselves. An I/O will be serviced by the processor quickly, and only has to wait for the drive action.

If the number of drives per director is large, the processing power may not be enough to fully utilize each drive. Performance suffers, as some of the drives sit idle while waiting for the director to assign them work. I/Os have to wait in a processor queue before being serviced, and then have to wait again for the drive action. Of course, much more capacity can be managed this way.

The point at which the backend goes from being drive-limited to processor-limited is subject to the workload. A medium number of drives per director might be processor limited or drive limited, depending on the I/O characteristics of the workload.

Director-to-Bay Relationships



Symmetrix VMAX: Eight Engines, 2,400 Drives

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 6

VMAX arrays have specific director-to-bay relationships, as shown here. When you insert a DAE of drives into a bay, you are assigning them to be managed by a specific director pair.

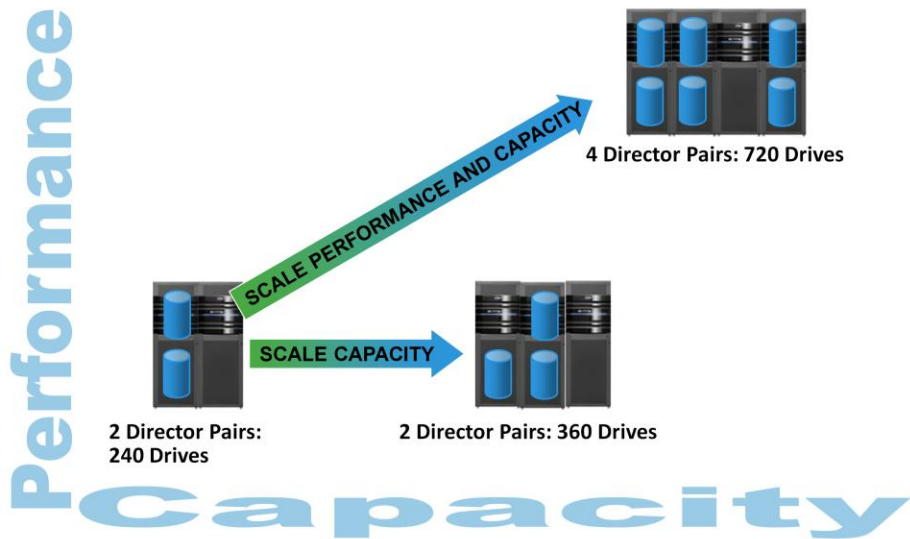
In a VMAX array, each engine of two director cards is directly connected to one-half of one drive bay. For instance, engine four is directly connected to the bottom half of the bay, to the immediate left of the system bay. The other seven engines have their own direct connect half-bays. To add more drives to a director pair, add a “daisy chain” bay and fill it with drives.

A similar configuration was used in a DMX array. Each director pair was directly connected to the drives in one half-bay. Adding “daisy chain” bays extended the number of drives managed by a port.

Since a VMAX has twice as many back-end ports, the number of drives per daisy chain was reduced. This provides greater backend performance without reducing the number of drives.

The segmentation of the directors gives us a way to isolate workloads on the backend. I/O targeted to drives managed by one engine will not interfere with backend resources of any other engine. This is an important factor in creating multiple tiers of storage on a single array.

Symmetrix Configuration Flexibility



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 7

Since the director-to-bay configuration of the Symmetrix DMX-3+ and VMAX isolates the activities of the director pairs, you can configure for performance or capacity in the same array. You might start with a single bay of 240 drives. You will need two VMAX engines to manage them. This would result in a high-performance configuration. Over time, you might add another bay and only fill one half with drives. That would double your capacity on that director pair, but leave the other pair untouched. The double-capacity pair might start to become processor limited.

By adding more bays of storage to the existing directors, you create a more capacity-oriented environment. By adding more directors with the storage bays, you maintain a high-performance environment.

Segregating Workload Tiers

“Capacity” Tier Applications



“Performance” Tier Applications

Benefits:

- Director resources needed by one tier won't impact another tier's performance
- Ensures some level of performance for each tier
- Can isolate lower speed drives on a capacity tier

Disadvantages:

- Director resources assigned to lower tiers are not available for tier-1 bursts
- Reduces maximum possible performance across the entire array

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

8

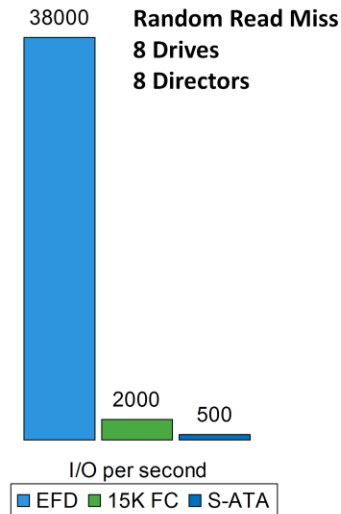
There are several considerations regarding tiered storage. The illustration on this slide shows an array with three storage bays configured to create tiers. The director pairs managing the bays on the left are configured toward capacity because there are thirty drives in their loops. The pairs managing the bays on the right have only fifteen drives per port, making them performance-oriented.

If you segregate your data set so it is entirely managed by one director pair, the workload impact on the drive and director resources does not affect any other pair. You can keep the data that requires the highest backend performance on a separate, performance-configured pair. Of course, other resources, such as the matrix channels and cache, are universally shared by all workloads. This ensures a consistent level of performance for the tier, regardless of the activity in other tiers.

However, the isolated data set cannot “borrow” processing time from other quadrants at times of peak utilization. By isolating the data set on just one tier, you lose the ability to use all of the power the array can provide. For example, an application experiencing a peak load on the A quadrant cannot use the processor resources of the other quadrants, even if they happen to be idle.

Factors That Affect Performance: Drive Type

- Enterprise Flash Drives (EFD)
- Fibre Channel Drives (FC)/SAS
 - ▶ 15 K RPM
 - ▶ 10 K RPM
 - ▶ 7.2 K RPM
- Serial ATA Drives (S-ATA)



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

9

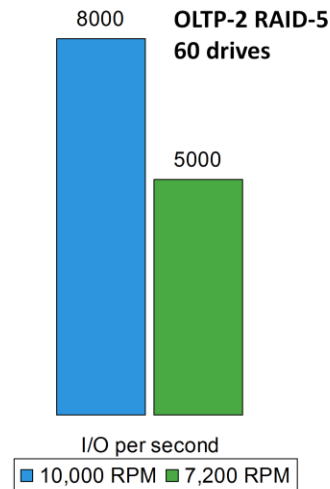
Symmetrix arrays currently use four types of drives: Enterprise Flash Drives, Fibre Channel Drives, SAS, and Serial ATA Drives. Enterprise Flash Drives, or solid state drives, are significantly faster because they do not use a rotating magnetic platter. The drives used in Symmetrix arrays employ write leveling technology to mitigate the effects of electrical burn. These drives are expected to have a usable lifetime of about five years.

Fibre Channel, SAS, and Serial ATA drives are traditional rotating platter drives. Fibre Channel drives use faster connection technology than Serial ATA. Fibre Channel drives are available in three different speeds. Currently SAS and FC drives have the same performance for the same disk RPM. In Symmetrix arrays, Serial ATA are only used at the largest sizes, typically 1 TB drives or larger.

There is a clear performance difference across these drive types. The graph on the right shows the result of a test performed on eight Enterprise Flash Drives, 15 K RPM Fibre Channel Drives, and Serial ATA drives. Each drive was managed by a single backend director, eliminating the chance that a processor bottleneck would affect the results. An 8 KB random read miss workload was applied to the eight drives as fast as possible. The results show that Enterprise Flash Drives can process significantly more I/Os than the traditional hard drives. The limiting factor for EFD is the director. Fibre Channel drives are also significantly faster than Serial ATA drives.

Factors That Affect Performance: Drive Speed

- Drive performance scales nearly linearly with drive speed
- Performance comparisons show:
 - ▶ 10,000 RPM is ~40% better than 7,200 RPM
 - ▶ 15,000 RPM is ~50% better than 10,000 RPM



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

10

Within a drive type, performance scales nearly linearly with drive speed. In Symmetrix arrays, the only type of drives that are offered at different speeds are Fibre Channel drives, which are available at 7,200 RPM, 10,000 RPM, and 15,000 RPM. Benchmarking shows that the performance of a faster drive will be better than the performance of a slower drive. The amount of improvement is approximately equal to the ratio of the difference in speeds. Naturally, this holds only within the same drive type (do not compare Fibre Channel to Serial ATA drives this way).

The benchmark results displayed on the right are one example of this factor. An OLTP-2 test was performed on RAID-5 devices spread across 60 drives. The 10,000 RPM drives outperformed the 7,200 RPM drives. While the calculation predicts a 40% improvement, this test actually showed a 60% improvement with the faster drives.

Physical Drive Characteristics



- **Seek Time:** Time to Move heads to appropriate track
- **Latency:** Time until data rotates under heads

	15,000 rpm	10,000 rpm	7,200 rpm
Internal Transfer Rate (MB/s r, MB/s w)	68, 114	47, 94	47, 94
Time to Transfer 8 KB (r, w)	0.11 ms, 0.07 ms	0.16 ms, 0.08 ms	0.16 ms, 0.08 ms
Avg. Seek Time (r, w)	3.5 ms, 4 ms	4.7 ms, 5.4 ms	8.5 ms, 9.5 ms
Avg. Single Track Seek Time (r, w)	0.27 ms, 0.45 ms	0.20 ms, 0.50 ms	0.80 ms, 1.00 ms
Avg. Latency	2.0 ms	3.0 ms	4.7 ms

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

11

Some useful performance observations can be drawn directly from the characteristics of a physical drive. The main components of the response time in drive operations are the seek and latency times. The seek time is the time taken to move the read/write heads to the appropriate track. Once there, the latency time measures how long it takes for the data to rotate to the head position.

To help get an idea of the lengthy times involved, typical values are given for the three typical drive speeds used in the Symmetrix array: 15,000 rpm, 10,000 rpm, and 7,200 rpm. For a given speed, the times are very similar for all drive manufacturers and drive sizes. Note how much longer the seek and latency times are, when compared to the time to (internally) transfer an 8 KB I/O. Positioning the disk heads is a very time consuming process for a disk.

The numbers shown here are for the version used in the Symmetrix, which will be slightly different from the values Seagate publishes. In particular, the average transfer rate is lower than the drive manufacturer's published values because the write cache of the drive is disabled in a Symmetrix array to prevent data loss in the case of a power failure.

Factors That Affect Performance: I/O Location

Action	Time Calculation for 15,000 rpm	Time
One 8KB read I/O	1 * 3.5 ms (seek) + 1 * 2.0 ms (lat) + 1 * 0.11 ms (xfer)	5.61 ms
One 32KB read I/O	1 * 3.5 ms (seek) + 1 * 2.0 ms (lat) + 4 * 0.11 ms (xfer)	5.94 ms
Four adjacent track read 8KB I/Os	1 * 3.5 ms (seek) + 3 * 0.27 ms (1 trk seek) + 4 * 2.0 ms (lat) + 4 * 0.11 ms (xfer)	12.75 ms
Four random read 8KB I/Os	4 * 3.5 ms (seek) + 4 * 2.0 ms (lat) + 4 * 0.11 ms (xfer)	22.44 ms

Note: Number and location of I/O have large impact on response time; much more than I/O size has.

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

12

Using the numbers on the previous page, we can compare the effects of different I/O activities on a 15,000 rpm physical drive. A single 8 KB I/O has a 5.61 ms response time using the average seek and latency values. A single 32 KB I/O requires only 0.33 ms additional time even though it is four times the size. The actual data transfer time is small compared to the time taken to move the heads into position.

Accessing four 8 KB I/Os from adjacent tracks adds an additional 0.27 ms seek time for each additional I/O after the first, and 2.0 ms of latency for each I/O. The total transfer time is 0.11 ms for each I/O, bringing the total to 12.75 ms.

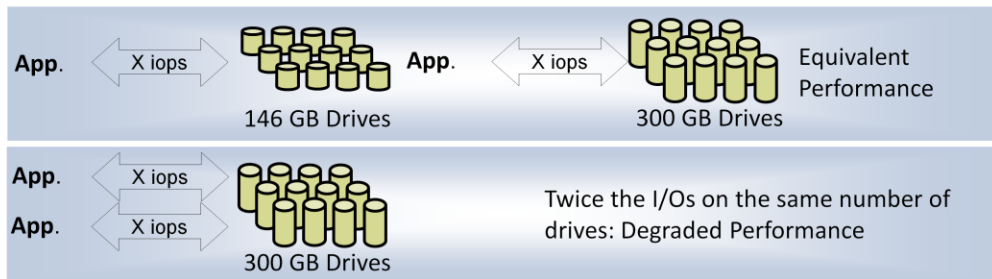
Accessing four 8 KB I/Os from completely random locations requires the average time for a single random I/O each time. On average, the four transfers will take 22.44 ms.

From these observations, we see that it is the number of I/Os and their location that have the greatest impact on the disk throughput. The last three examples each transferred 32 KB of data, but the total time ranged from 6.60 ms to 24.60 ms, depending entirely on the number of I/Os needed and their relative location.

Bear this example in mind when considering Symmetrix backend optimization. Minimizing the numbers of disk I/O and the distance between them is an important step in optimizing the backend. Caching, prefetching, and the Command Tag Queuing topics that will be discussed in this module all assist in this way.

Factors That Affect Performance: Drive Size

- All other factors being equal, different sized drives have approximately the same performance
- Performance comparisons show:
 - ▶ 146 GB is ~ equal to 300 GB
 - ▶ 300 GB is ~ equal to 500 GB, and so on...
- However, administrators tend to crowd more I/O on larger drives, affecting performance



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

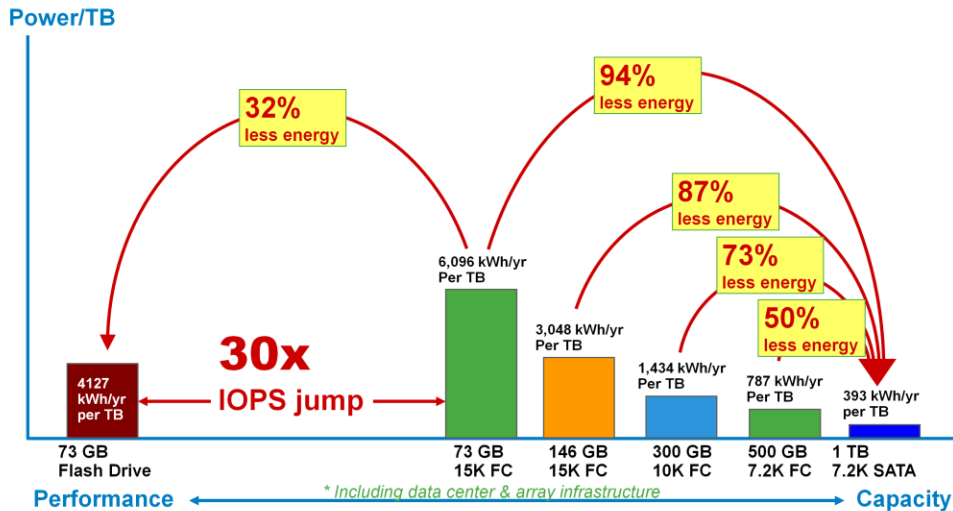
13

All other factors being equal, drive size does not significantly affect performance. A 146 GB drive will have approximately the same performance as a 300 GB or 500 GB drive of the same type and speed.

However, intuition tells us that larger drives should be used for low performance applications. This is because our tendency is to crowd more applications on larger drives, causing a higher I/O rate. Any component will tend to have poorer performance when you increase the number of I/Os. Unless you carefully control the number of active applications on the drives, you can assume that larger drives will end up providing poorer performance, simply because they have more traffic.

Power to Capacity Comparison

Power / GB: Energy to Store 1TB of Information*



EMC²

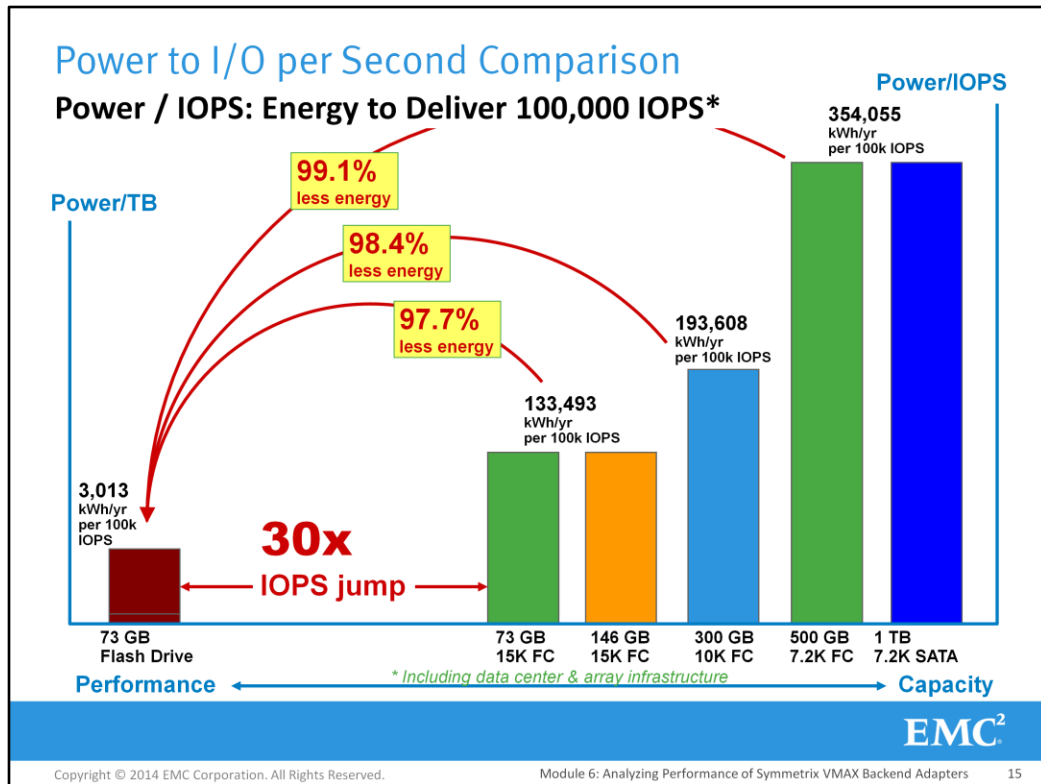
Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

14

When Enterprise Flash drives were introduced, EMC carefully studied their power consumption, and the power characteristics of the other drives used in Symmetrix arrays. Their study showed that for a given capacity, Enterprise Flash drives demand an amount of power that is similar to the other drives. The Enterprise Flash drives consumed slightly more power than the 146 GB 15,000 RPM Fibre Channel drives.

The lowest speed drives had the lowest power requirements, of course.



When considering Enterprise Flash drives, the number of I/O per second, not the capacity, is most likely the main concern. If you compare the amount of power consumed per I/O per second delivered, Enterprise Flash drives are considerably more power efficient than any other drive type. The fastest Fibre Channel drives consume 44 times the power, and Serial ATA drives consume over 100 times the power when delivering the same I/O performance. The long term cost of powering the drives needs to be considered when purchasing resources for critical applications.

Enterprise Flash Drive Best Practices

Enterprise Flash Drives Enhance *Drive* performance.

The Symmetrix array is designed to mitigate the effects of Drive Performance!

- **Best choice for EFD:** Any read miss heavy applications
- **Good Choice for EFD:** Database indexes and temp
- **Bad Choice for EFD:** Database logs
- **Best Practice:** All engines have same # of EFDs (and same size)



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

16

It is important to remember that an Enterprise Flash drive is much faster than other drives, but does not improve any other Symmetrix component. The Symmetrix architecture was designed to mitigate the effects of drive performance. Many I/O operations, preferably as many as possible, are serviced by the directors and cache, not the drives. For those operations, a faster drive does not improve performance.

Since read miss operations depend on the drive speed, put heavy read miss applications on Enterprise Flash drives. In many database applications, access to the indexes and temporary space tends to be read miss heavy, making them good candidates.

Databases tend to write to log devices almost exclusively. If the Symmetrix has enough cache to buffer these writes, you will get no performance gains at all from Enterprise Flash drives. If you have insufficient cache for the log devices, a less expensive solution can probably be achieved by spreading the workload among more devices.

A best practice is to have the same number of Enterprise Flash Drives on each director loop and engine. This spreads a number of drives among more processors, helping increase the number of I/Os driven to the drives. The alternative, putting many Enterprise Flash drives on a single disk director loop, is likely to produce a processor bottleneck. The director will not be able to issue enough I/Os to fully utilize so many fast drives.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 1: Summary

During this lesson the following topics were covered:

- Backend architecture to performance capabilities
- Symmetrix architecture and performance features to create tiers of performance
- Disk size and speed when choosing disks to meet capacity and performance needs



This lesson covered concepts of the architecture of the Symmetrix backend.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 2: Analyzing Symmetrix Backend Utilization

During this lesson the following topics are covered:

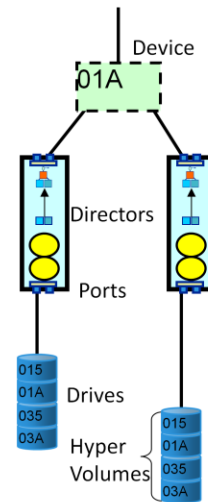
- Using the correct metric to monitor backend hardware components
- Detecting backend workload imbalance



This lesson covers analyzing Symmetrix backend utilization.

Backend Traffic Measurements

- Logical Device Level:
 - ▶ Traffic routed to that device, no matter how it is physically stored
- Disk Director Level:
 - ▶ Traffic handled by the processor for all disks, including protection writes and business continuity synchronization
- Drive Level:
 - ▶ Traffic to/from all hyper volumes on the drive, including protection, replication, and internal scrubbing



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

19

There are several levels at which to consider I/O traffic at the backend:

The Logical Device Level:

For the most part, I/Os from the host to a particular logical device are observed and measured at the device level. So, a standard measure like I/O per second indicates the number of I/Os generated by the hosts accessing the device. However, some device measures show the backend traffic associated with the device. These measures are typically prefixed with “DA.”

The Disk Director Level:

Traffic measured at the disk director level shows all traffic to/from all of the physical disks controlled by the director. This includes host I/O, protection writes to mirrored and RAID devices, and internal synchronizations like those performed by TimeFinder and Optimizer.

The Drive Level:

All traffic for the hyper volumes stored on a physical drive are measured here. As with the disk director level, traffic due to protection and internal synchronizations will also be visible in these metrics. Additionally, traffic performed by internal scrubbing (error detection and correction) will also be apparent here.

Measuring Backend Director Performance

- BE Directors
 - ▶ Monitor > All BE Directors (DA)
IOs/sec, Throughput, Prefetched Tracks/sec, % Busy
 - ▶ Analyze>Host MB/sec - All Backend Directors
- Single Director
 - ▶ Monitor>BE Director (DA)> DA
Port IOs per Sec, Port MBs per sec, Prefetched Tracks, % Busy
 - ▶ Monitor>BE Directors (DA) Summary>
%Busy, %Non-IO Busy, %Busy Port 0&1, IOs/sec, Reqs/sec, MBs
Read & Written/sec
 - ▶ Analyze>Array>BE Director – Tabular view Sort by % Busy
 - ▶ Analyze>Reqs/sec, Read & Write Reqs/sec (area chart)
 - ▶ Analyze>Array>BE Director (DA)>DA> IOs/sec



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

20

There are many backend director metrics. With dashboards, time series, and snapshot views you can show I/Os per second and throughput across all directors and for each director. You can also break the I/Os down into read and writes for a single director.

You can also measure utilization for each director. These measures will come from the busy counters on the directors.

Measuring Disk Performance

- Analyze>Array>Disk Group | Disks
 - ▶ % Busy
 - ▶ Avg Response Time (ms)
 - ▶ IOs /sec
 - ▶ MBs/sec
- Analyze>Array>Disk Group>group>Disk
 - ▶ Seeks/sec
 - ▶ Seek Distance/sec



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

21

Performance Viewer also contains many measures at the drive level. You can show snapshot and time series graphs for all drives by clicking on the Disks Groups Tab, or drill down to select the measures for a single disk. The bar graph (Snapshot Chart) comparisons are especially useful in showing the traffic on each drive.

Utilization measures will come from the busy counters on the drives.

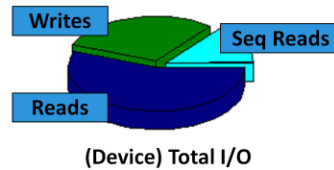
Average response times are also available for reads and writes individually.

Measuring I/O Per Second – SCSI Commands

- Analyze>Array>Disk Group>group>Disk
 - ▶ Total SCSI Commands/ sec
 - ▶ IOs/sec

SCSI Commands Per Sec =

$\sum$ Read Commands Per Sec
Write Commands Per Sec
Skip Mask Commands Per Sec
Verify Commands Per Sec
XOR Writes Per Sec
XOR Write-Reads Per Sec



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

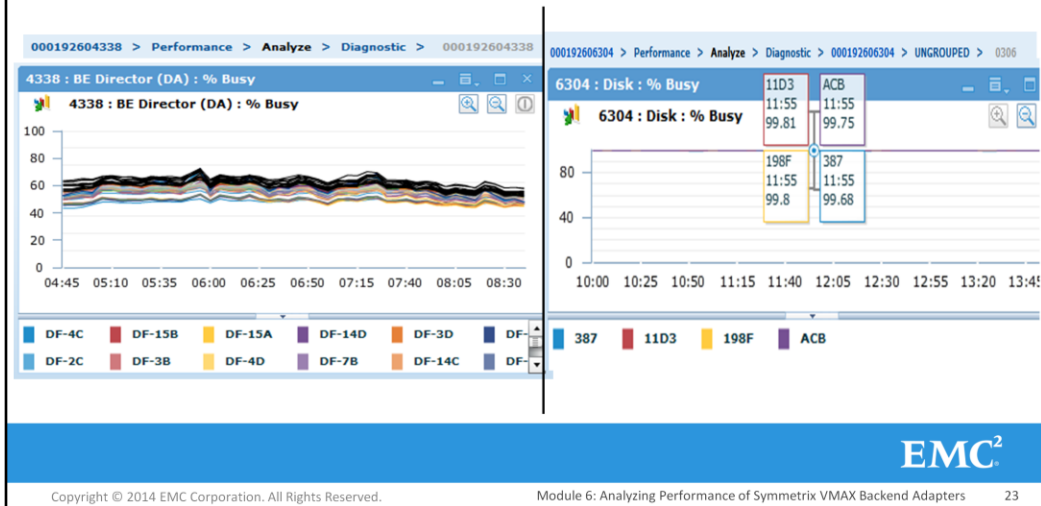
Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

22

At the disk level the primary measure of traffic rate is “SCSI commands per sec.” This sums a number of other measures (listed) that represent all of the traffic a disk experiences. Read and write commands are standard data operations. Skip mask commands are used to manipulate the disk controller’s “skip mask,” which is used to perform several operations from non-sequential disk locations by “skipping” the drive heads to a new location before returning the result. Verify commands are used when scrubbing the disks for bit errors. The XOR commands are used during RAID parity calculations.

What to Look For

- High utilization on disk directors or disks
- Imbalance among disk directors or disks
- Use device measures to trace source of high utilization or imbalance back to a host application



Your first step in examining backend performance is to look for overly high utilization on disk directors or disks. This means that additional response time is probably being experienced for this component.

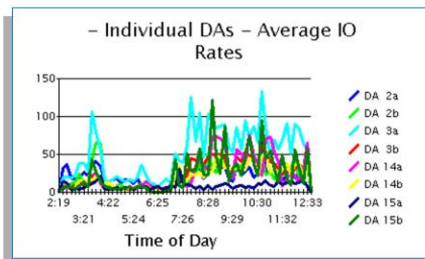
Having found examples of high utilization, you should then check the balance among the disk directors and disks. If some are being highly utilized at the same time that others are not, then that extra response time is being experienced for nothing. Moving the activity to the underutilized components will help reduce the issue.

The graph on the left of this slide shows the percent busy across all disk directors on a particular array. The disk directors show medium utilization and are fairly well balanced. Strong balance among disk directors is very common in Symmetrix arrays. Any significant imbalance at this level should be investigated.

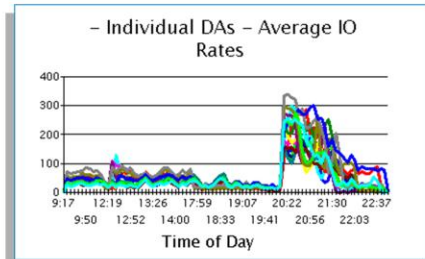
The graph on the right of this slide shows the utilization for some disks on a different array. All of these disks are operating at a sustained peak capacity. If it was only for a short period, such spikes might ordinarily be ignored. However, it is clear that these disks are at 100% utilization and have no more to give. This represents an imbalance, possibly some heavy workload is concentrated on just a few disks. Better balance of that data across as many disks as possible would be preferable.

I/O per second and % Busy measures can be ignored for devices since they are virtual components. There is no maximum number of I/O per second a device can experience in isolation. Any limits for a device are actually limits on the front-end hardware, cache, or backend hardware; examine those for over utilization. However, examining device measures is useful in determining what might be causing problems on the hardware components. In the example graphs on this slide, the over utilized disks are probably peaking due to increased traffic on some devices (it could also be due to business continuity activity). Your next steps would be to determine how those disks are being utilized and see what could be done to get things more balance, and in this case, probably on more or faster disks.

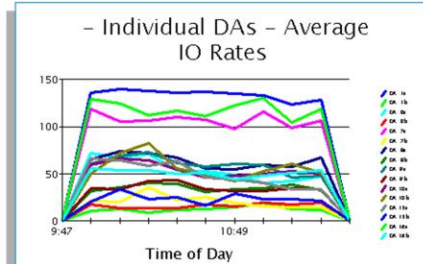
Disk Director I/O Examples



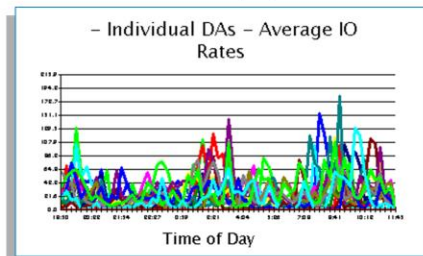
Oracle/SAP (UNIX) OLTP



SQL Server/SAP (NT) OLTP



Lotus Notes (NT) Mail



Oracle (UNIX) Banking-OLTP

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

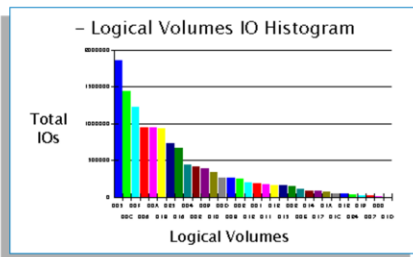
Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

24

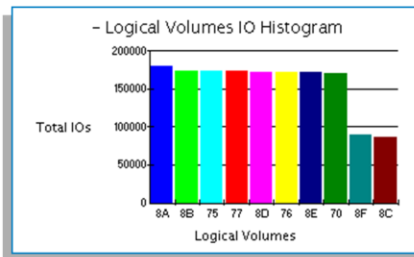
It is more difficult to balance disk director activity than it is to balance front-end activity. Disk directors service traffic from all the hosts connected to the array. Sudden I/O spikes from a single host might only affect one or two directors, or write pending status might cause a few directors to rapidly destage data.

In general, the best performance is achieved when the directors are generally balanced. Ignore momentary spikes and concentrate on steady state performance. Directors that are continuously used more than the others should be investigated.

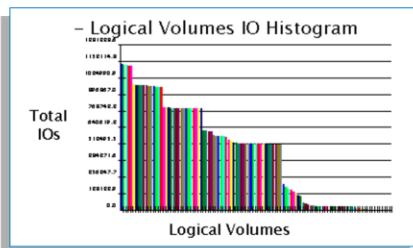
Observations About Device I/O



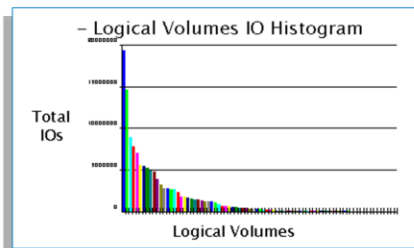
Oracle/SAP (UNIX) OLTP



MS Exchange (NT) Mail



SQL Server/SAP (NT) OLTP



Oracle (UNIX) Banking-OLTP

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

25

Although we are not as concerned about utilization or balance among devices, some interesting observations can often be made by examining the basic measures. Rounded slopes, like the ones in the Oracle/SAP OLTP and Oracle Banking-OLTP graphs, are common when no striping or balancing is in effect. In mixed environments it would be very difficult to evenly balance the traffic across all devices. Mixing different filesystems, applications, and even operating systems tends to produce very unbalanced traffic. Nothing can typically be done to improve the balance under these circumstances.

Plateaus like the ones displayed in the SQL Server/SAP OLTP and MS Exchange Mail graphs indicate balancing efforts. Host or Symmetrix based striping might be employed, or the application itself might write data evenly across all devices.

The very active devices shown in the Oracle (UNIX) Banking-OLTP graph are common in database environments. Database log files are often placed on separate devices for backup purposes, and the logs are typically accessed more heavily than any other data element in the application. Those highly used devices are probably holding the database log files.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 2: Summary

During this lesson the following topics were covered:

- Using the correct metrics to monitor backend hardware components
- Detecting backend workload imbalance



This lesson covered analyzing Symmetrix backend utilization.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 3: Analyzing I/O Imbalance

During this lesson the following topics are covered:

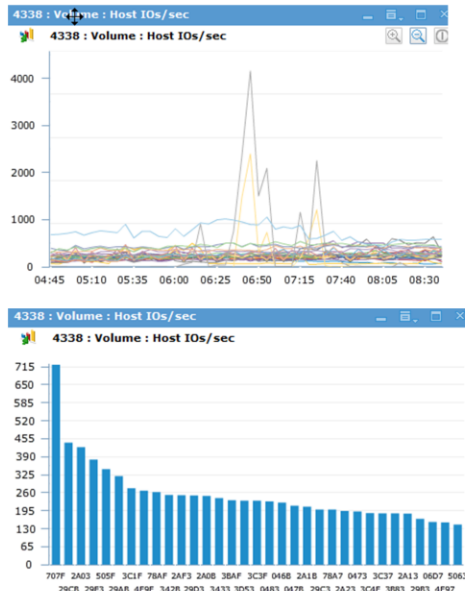
- Symmetrix devices to achieve workload balance
- Striping to distribute workloads over many backend components
- Virtual Provisioning
- FAST VP



This lesson covers the choices available for balancing workload on the backend of the Symmetrix.

Imbalance Scenario – Devices

- Device I/O per sec
- Devices unbalanced
- Heavily used devices might be balanced across physical disks
- Unused devices may be BCV, gatekeeper, etc.
- Perfectly acceptable in mixed environments



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

28

As part of continuing performance monitoring, the activity of this Symmetrix is currently under study. No performance problems are being reported, but a good analyst doesn't wait for delays to happen before checking the environment.

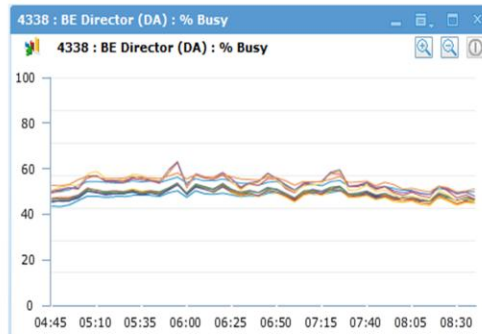
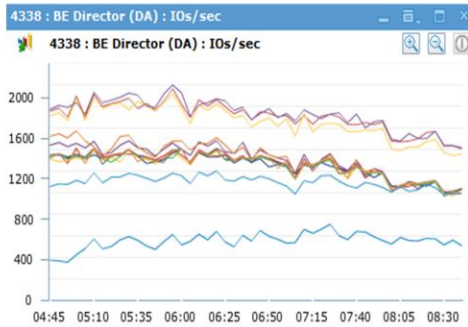
This scenario concentrates on the disk activity of the Symmetrix. Assume that the other issues discussed throughout this course have already been checked, or will be considered later.

As a first check, the I/O per second for each device has been plotted in a snap chart histogram and a regular time series graph.

The devices look unbalanced, but that alone is not a concern. This is not unusual; all workloads will have a wide range of I/O rates on the devices. Symmetrix arrays contain a large number of devices that are not regularly used, such as BCVs and gatekeepers. Performance problems are a result of physical component limitations, not logical limitations. If the heavily used devices are spread across physical disks properly, there should be no problem.

Imbalance Scenario – Disk Directors

- Disk Director % Busy, IOs per second
- Near perfect balance?
- No need to fix this?



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

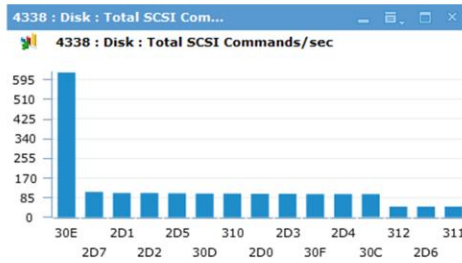
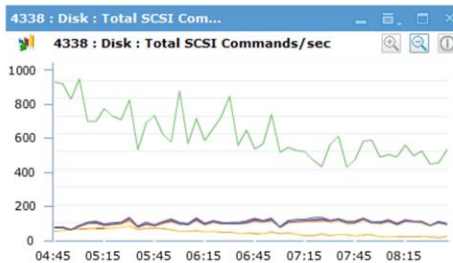
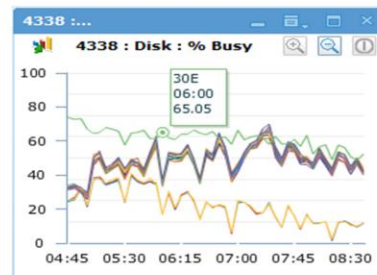
Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

29

Another health check involves the disk director utilization and IOs per second. A single metric doesn't always tell the whole story. The utilization (% Busy) looks fairly balanced, however, when looking at the IOPS the unbalanced activity is apparent.

Imbalance Scenario – Disks

- Disk %Busy, SCSI commands per sec
- Significant difference in utilization among disks, high utilization
- Disk access speed is the slowest link in the performance chain
- Balancing activity here could have a dramatic impact on performance



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

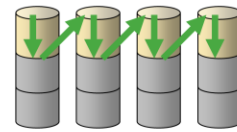
30

A third check plots the disk SCSI commands per second (effectively, the I/O per second measured at the disk level).

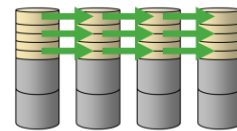
The heavy utilization and imbalance on the disks could be a serious concern. One of the disks show extended utilization of over 60%, with many peaking over 60%. However, the “SCSI commands per sec” histogram shows that many of the disks are largely unused. By Little’s Law, we know that the overall response time would be lower if we could balance this traffic. Putting some of the workload from the heavily utilized disk on to the lightly utilized disks will help bring the overall response time down.

Striping

- Evenly distributes data across all drives
- Eliminates “hot spots” that concentrate on just one drive
- Not necessary if workload is already very random and fully distributed across all drives
- Can hinder prefetching for very sequential workloads (**Exception:** Symmetrix-striped devices)



Regular Devices



Striped Devices

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

31

Having chosen the devices for your data set so that they are distributed across the array, you might further consider striping to help even the workload. Instead of addressing the data linearly on the drives (first 10 GB on first 10 GB device, etc.), striping addresses small blocks of data across each drive (first 64 KB on first device, next 64 KB on next device, etc.). This helps eliminate “hot spots” that can focus all of the I/O on a single drive, canceling out the hard work you have done to choose distributed drive.

Striping is not necessary in situations when the workload is already very random and fully distributed across all drives. This situation is already ideal from a backend point of view. Unfortunately, real world workloads are rarely this random; it is even rarer for a workload to be completely distributed across the several drives the data set resides on.

Striping can actually hinder prefetching if the workload is very sequential. This is because the disk directors consider I/O to a single device as separate from another device’s I/Os. A back-end director will never consider separate device I/Os when deciding what to prefetch. However, if the striping is implemented through Symmetrix RAID-5, RAID-6, or striped meta devices, the directors will recognize that the I/Os are part of the same device and create prefetch tasks to pre-populate cache with the data.

Striping Tips and Best Practices

- Multiple levels of striping (host + array)
 - ▶ Acceptable for random workloads
 - ▶ Eliminates sequential gains
 - ▶ Use, at most, two levels of striping
- When host striping, 128-256 KB stripe width is good
- Number of drives to stripe a single data set over
 - ▶ Four or Eight is best
 - ▶ Too many physicals in a stripe set just overlaps directors



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 32

Here are some general striping tips and best practices.

You can stripe at the Symmetrix level by using a variety of methods. You can also stripe at the host level using volume management software. You can also do both to more completely disperse hot spots of data. If your workload is random, but still has hot spots, you might stripe at both the host and array level. But, realize that this will eliminate any gains from sequential I/O. None of the Symmetrix prefetching algorithms work well if the host is striping the data.

At most, use just two levels of striping. This usually means host striping with a volume manager, and one type of Symmetrix striping. Two levels of array striping can be achieved in a variety of ways. If you do this, do not also stripe at the host level as it is unlikely to do any more good.

Benchmarking shows that when host striping, a stripe width (stripe size) of 128 to 256 KB works best with the Symmetrix array. If you would like to use a larger size, at least use a multiple of 256 KB.

Benchmarking shows that striping a single data set over four or eight physical drives gives the best performance. Using more drives does not increase performance gains, and can even hinder performance. Once the number of physicals exceeds the number of available disk directors, I/O for the device will be competing for director utilization with other I/O from the same device.

Choosing Stripe or Meta Members

- Be sure hypers are distributed
- If striped, be sure to use devices of the same size



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 33

When choosing stripe or meta members, make sure all hyper volumes are distributed across the backend. No two hypers should be on the same physical. If possible, no two hypers should be on the same disk director port.

It is acceptable to have the hypers for the meta or stripe set appear on devices or ports with hypers from other metas or applications. On a fully utilized Symmetrix, you will not have the luxury of allocating completely unused disks for each task. Your goal in choosing devices is not to separate the hypers from all other traffic, it is to ensure that two hypers carrying the same traffic do not compete for disk resources. This prevents temporary spikes in a single application's work flow from hitting common resources. Different applications will not usually spike at the same time, so data from different applications can typically share resources without significant impact.

Striped Symmetrix Devices

Technology	Stripe Width	Number of Drives
RAID-5 3+1 or 7+1: 1 parity block for every 3 or 7 data blocks	256 KB*	4 or 8
RAID-6 6+2 or 14+2: 2 parity blocks for every 6 or 14 data blocks	256 KB	8 or 16
Striped Meta: Striped collection of devices of any protection	960 KB	8-16 devices are best



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

34

Here is a comparison of some of the standard device striping technologies available in the Symmetrix array.

RAID-5 and RAID-6 devices stripe their data over a number of drives. The stripe size is 256 KB for these devices.

Meta devices can also be striped. The stripe size of a meta is 960 KB. The number of drives the data is spread over depends on the number of meta members and the device protection. In most cases, it is best to use no more than eight to sixteen members in a meta.

*Stripe width in (5876) VMAX 10K (987) and 40K other than R5 3+1 and R6 is 128K.

Meta Devices Considerations

- Meta devices are used when the host needs larger devices:
 - ▶ Maximum Symmetrix device size limit of 262668 cylinders (240 GB)
- Striped metadevices (recommended in most cases):
 - ▶ Better performance with Synchronous SRDF
 - ▶ Better read performance
 - ▶ More I/Os can be queued to striped metas because of higher write pending limits caused by each member getting a percentage of cache
- Advantages of concatenated metadevices:
 - ▶ Non meta devices can be converted to meta devices without destroying existing data
- Concatenated or striped thin meta devices can be expanded without destroying data



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

35

In most cases, EMC recommends using striped rather than concatenated metadevices with Virtual Provisioning because they offer the following advantages:

- With Synchronous SRDF, Enginuity allows one outstanding write per thin device per path. With concatenated metadevices, this could cause a performance problem by limiting the concurrency of writes. This limit will not affect striped metadevices in the same way, however, due to the small size of the metavolume stripe (1 cylinder or 1920 blocks).
- Potential performance benefits to using striped thin metas instead of concatenated thin metas.
- The Symmetrix Enginuity operating environment has a logical volume write pending limit to prevent one volume from monopolizing writeable cache. Because each meta member gets a small percentage of cache, a striped meta is likely to receive more writeable cache for the metavolume.

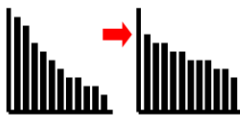
However, concatenated thin metadevices do offer two important operational advantages over striped thin metadevices:

1. Non-metadevices can be converted to concatenated metadevices without destroying the existing data. It is possible to convert a thin device to a concatenated meta without first unmapping the device.
2. Concatenated or striped thin metadevices can be expanded without destroying existing data by adding metamembers to an existing metadvice.

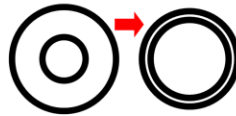
Symmetrix Optimizer – Automatic LUN Migration for Thick Devices

- Automatically study performance and swap devices within a disk group to balance drive performance for thick devices
- Can control:
 - ▶ Devices eligible for swap, swap device priority
 - ▶ Analysis time window, swap time window
- Can manually swap devices
- One DRV device per swap

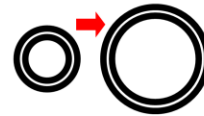
Optimizer Goals:



1. Balance Drive Load



2. Minimize seek between active devices



3. Move active devices to drive edge (faster)

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 36

Optimizer is an obvious solution to just about any backend imbalance situation. It automatically analyzes drive performance to find heavily loaded drives. It analyzes device performance to find heavily and lightly utilized devices, then it swaps heavily utilized devices with lightly utilized ones to optimize performance. Devices swapped must be in the same disk group and have the same RAID protection.

Optimizer attempts to achieve a number of goals by swapping. Most importantly, it is trying to balance the drive load. It will also consider a swap to minimize seek time between active devices on a disk. Finally, it will consider moving an active device to the outer edge of a drive where there is a slight performance increase.

You can control many Optimizer factors. You can eliminate any devices that have been carefully configured from swapping, and set a device priority to make some devices more prone to optimization. You can also set an analysis and swap time window to control Optimizer's activities, as well as use Optimizer to manually swap devices in anticipation of a change in I/O traffic.

Optimizer requires a dedicated DRV (Dynamic Reallocation Volume) device as a temporary place for swapped data. The device must be at least as big as the swapped devices. Prior to code version 5774, two DRVs were required, and they were required to be exactly the same size as the swapped devices.

VP Performance Considerations

- Extent allocation is striped across all enabled data devices in the pool:
 - ▶ With regular Symmetrix volumes, wide striping can be achieved using striped meta devices
 - ▶ Thin devices are widely striped automatically
 - » Makes it easier to avoid hot spots
 - » Potential performance improvement for some workloads
- Pool-based view of storage layout:
 - ▶ Thin pools should contain storage of a particular class (or tier) with a defined business purpose
 - ▶ Planning consists largely of deciding what tier of storage is required and which business group's resources should be used
 - ▶ Storage is viewed at a higher level when planning new devices
 - ▶ By managing Thin pools with FAST VP storage can be quickly and dynamically relocated to a different tier when needed



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 37

Thin device performance can be better than regular devices. However, real world application performance is highly dependent on the applications that are sharing the Thin pool.

In addition to potential contention for access to shared data device, response time and throughput can be impacted by the overheads incurred the first time a write is performed on an unallocated region of a thin device. This will disappear once the “working set” of a thin device has been written.

When managed with FAST VP, thin pool extents can be dynamically migrated to higher or lower performing storage tiers depending on the performance needs of an application.

Thin Pool Performance – All Dashboards

- All Thin Pools
 - ▶ Monitor>EMC Dashboards>All Thin Pools
 - ▶▶ Total Pool Capacity
 - ▶▶ Used Pool Capacity
 - ▶▶ Read Miss Response Time
 - ▶▶ Write Response Time
 - ▶▶ BE Read Req/sec
 - ▶▶ BE Write Req/sec

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 38

You will find excellent metrics for Thin pool performance. You can get a quick overview of all of the pools or a specific pool by using the dashboards.

You can show a dashboard of the I/O activity and the capacity allocation of all pools. The Bar charts are the default, but you can change this to Pie or Tabular by going to the dropdown menu then, “Change chart properties and settings” in the top right of each chart.

Of particular interest is the capacity allocation graph that appears in both the Snapshot and Trend views. This compares the allocated capacity (sum of the tracks allocated to thin devices) to the total capacity of the pool. To avoid write errors, add more devices to the pool when the allocated capacity approaches the total capacity.

Thin Pool Performance – Individual Pools

- Monitor>EMC Dashboards>Thin Pool> Select Thin Pool
 - ▶ Capacities (shows Total, Enabled, Used, and Allocated)
 - ▶ BE Read & Write Response Times
 - ▶ MBs Read & Written
 - ▶ IOs/Second (BE Read & Write Reqs/second)
- Analyze
 - ▶ Analyze> Thin Pool> Click one or more Thin Pools
 - ▶▶ % Used
 - ▶▶ Ingress Tracks
 - ▶▶ Egress Tracks



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 39

For specific pools, you can use the Thin Pool dashboard and select individual Thin Pools.

In addition to the graphs from the dashboards there are many metrics available that you can graph for one or multiple pools at a time.

From the Analyze view, click on the Thin Pool tab. It will show you all of the pools in a tabular view along with BE Requests per second, BE Megabytes Transferred per second, BE Average Response Time in milliseconds, and the Allocated Pool Capacity in gigabytes.

Of particular interest might be the % Used graph, which will show the current capacity. It will be a good indicator of how close you are to your next threshold. To avoid write errors, be sure to add more devices to the pool when the allocated capacity approaches the total capacity, or sooner.

Virtual LUN Migration

- Manually migrate thick or thin devices to change RAID level, disk type
 - ▶ Balance drive performance
 - ▶ Move data to another storage tier
- Source can be anything but Snap device, any Pool device, Vault device
- Target can be
 - ▶ Other, unused devices of the same size
 - ▶ Unconfigured space in a disk group

Performance Tip:

- Migrations move very slowly at 50% system write pending
- Migrations almost stop entirely at 70% system write pending



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters 40

The Enhanced Virtual LUN Migration feature allows you to manually move devices to different storage. You might move some devices to unused drives to better balance drive performance, or you might need to change the tier that the devices are stored on. You can change both the protection type and the drives the devices are stored on. By adding the new storage as a mirror of the original storage, the migration is performed non-disruptively.

You can migrate devices by copying them to unused devices of the same size. In this case, you would specify the source device numbers and target device numbers to the migrate utility (a Solutions Enabler command). The target devices would most likely have a different protection level, different drive properties, or both. You can also migrate devices by creating new targets out of unused storage in a disk group. You can specify a new protection type.

Since a migration requires cache locations to transfer data, the array will de-prioritize the migration when system write pendings reach 50% of the available limit, causing the migration to slow. At 70% of the system write pending limit, migrations slow to almost stopping to allow the directors to handle the heavy I/O load. You should avoid performing migrations when the number of write pendings are expected to increase.

Symmetrix Disk Groups

- Default disk groups identify drives by
 - ▶ Type (SATA, FC, EFD), Speed, and Size
- Create additional disk groups to identify drives by
 - ▶ Director Pair
 - ▶ Application
 - ▶ Etc.
- When creating devices, choose disk group to control type of drive
- Internal disk groups can only be created via SymmWin
 - ▶ They cannot be created using SYMCLI or Uni4VMAX



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

41

Use disk groups to control the characteristics and backend location of drives. By default, drives are arranged into separate disk groups by their characteristics. Groups are used to identify drives of different type, speed, and size. For example, 300 GB Fibre Channel 10,000 RPM drives would go in a separate group from the 300 GB Fibre Channel 15,000 RPM drives. You might create additional disk groups that organize disks by disk director pair.

When creating devices for an application, you can specify a disk group from which to allocate the storage. Use the default groups to allocate storage based on type, speed, and drive size. Use the custom disk director groups to allocate storage that is isolated on the backend. Since even custom disk groups cannot contain drives of different type, speed, and size, you will also be choosing drives of a specific level of service.

Fully Automated Storage Tiering – FAST

- FAST (also referred to as FAST DP) supports disk group provisioning for Symmetrix VMAX:
 - ▶ Full LUN movement of disk group provisioned (“thick”) devices
 - ▶ Supports FBA and CKD devices
- FAST VP supports virtual provisioning for Symmetrix VMAX:
 - ▶ SubLUN movement of thin devices
 - ▶ Supports FBA devices only
- Can promote/demote devices/SubLUNs across Enterprise Flash Drives, Fibre Channel Drives, Serial ATA (SATA) Drives



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

42

FAST DP moves application data at the volume (LUN) level. Entire devices are promoted or demoted between tiers based on overall device performance. Fully Automated Storage Tiering (FAST DP) is conceptually similar to Optimizer, but its main goal is to move devices across storage tiers to improve overall performance. Optimizer is restricted to moving devices to drives having the same characteristics, or in other words, in the same tier. FAST monitors device performance to find active devices that would benefit from being moved to higher performing drives. FAST will also identify inactive devices that could be relocated to lower performing drives without performance being affected. Data movement executed during this activity is performed non-disruptively, without affecting business continuity and data availability.

FAST will perform one of two actions when relocating devices. It will move a device to a new storage tier if there is unconfigured space in the tier. If there is no space in the tier, it will find a candidate in the new tier that can be swapped out, and will swap the devices much as Optimizer does.

FAST VP adds finer granularities of performance measurement and data movement. The data from a single thin device under FAST control can be spread across multiple tiers. The FAST controller is free to relocate individual sub-extents of a thin device, based on performance data gathered at the extent level.

When to use FAST and FAST VP

- Workloads with a higher skew will benefit more from FAST or FAST VP:
 - ▶ Workloads with skew above 80/20 are considered good candidates
 - ▶ Unbalanced workloads direct a higher percentage of I/O to a small percentage of the storage allocated
 - ▶ Heavily utilized devices are moved to faster technologies, to reduce response time
 - ▶ Under utilized devices are moved to less expensive technologies, to reduce cost
- Workloads with a lower skew may not benefit:
 - ▶ Workloads with a skew closer to 50/50 (uniform workload) are less likely to contain candidates for promotion/demotion



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

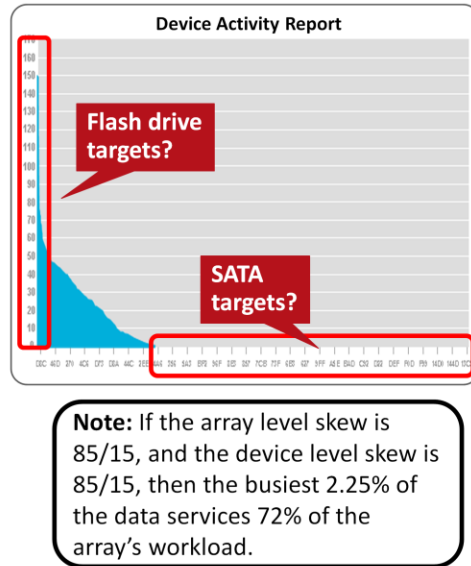
43

FAST is most effective in an environment where the workload tends to be skewed and the skew patterns are dynamic. For example, if devices that are heavily accessed this month may be underutilized next month, a suitable FAST policy will continually perform the necessary relocations to locate the heavily used devices on faster performing drives.

FAST is not beneficial if the workload is not skewed, or if the I/O workload is static, since workload skews can then be addressed using manual methods such as Symmetrix VLUN Migration.

Workload Skew

- Skew defines an asymmetry in device usage over time:
 - ▶ 85% of the workload serviced by 15% of the devices
- FAST accounts for array level skew:
 - ▶ Relocates busiest devices to higher performing tiers
 - ▶ “Idle” devices moved to more cost-effective tiers
- FAST VP accounts for device level skew:
 - ▶ Relocates busiest thin device extents to higher performing tiers
 - ▶ “Idle” data moved to more cost-effective tiers



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

44

The chart on the left is a Device Activity Report that shows a small number of active devices that are candidates for higher performance drives and a large number of inactive devices that are candidates for lower performance drives.

FAST moves complete logical devices to higher or lower performing tiers. In contrast, FAST VP moves thin device extents as necessary.

FAST DP Usage Considerations

- Use Performance Viewer to determine applicable workloads for FAST
- Use many SATA drives spread across the backend to reduce I/O traffic on individual drives
- FAST device moves and swaps can impact to disk response times



SATA drive technology is suited for data that is not accessed often, or data that is accessed mostly sequentially. It is best for I/O bandwidth to be shared by the larger pool of SATA disks (in comparison to FC disks). In this way, on average, each SATA disk will be accessed less often.

FAST device move and swap operations may result in a short term impact to disk response time for a very limited period. In the case of swaps, the impact will be greater due the greater number of devices involved. This impact is not specifically caused by FAST; however, it is a consideration for FAST data movement.

FAST DP Usage Considerations (Continued)

- Configure a single standard size of device on all tiers:
 - ▶ Increases probability of like-sized devices being available to perform swap
 - ▶ Smaller LUN sizes will benefit more from FAST
 - ▶ Larger LUN sizes may also benefit if workload is sufficiently skewed
- Meta devices are moved/swapped as complete entity:
 - ▶ Meta device members cannot exist across physical disk groups
 - ▶ Members of striped meta devices may not exist on same physical disks for FC and SATA
- For optimal system performance use Optimizer in concert with FAST:
 - ▶ Optimizer will balance load within a tier



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

46

Using a single size of device within the array allows greater flexibility in executing device swaps. Smaller LUNs tend to experience greater workload skews and can benefit more from the implementation of FAST.

Meta device members are not permitted to exist on different disk groups. FAST will not cause a striped metavolume to end up with two members of the stripe on the same physical disk.

Symmetrix Optimizer complements the functionality of FAST. The former does load balancing within disk groups, whereas the latter performs load and capacity balancing across disk groups.

What Makes FAST VP Different?

- FAST VP reacts quickly:
 - ▶ Initial review period as short as 2 hours
 - ▶▶ Ongoing reviews every 10 minutes
 - ▶ Ongoing movement when thresholds are crossed
- FAST VP tiers easily:
 - ▶ Up to 4 tiers of storage (EFD, FC, ATA, External(FTS))
 - ▶ Tiering across discrete pools
 - ▶ Tiering controlled at the storage group (or LUN) level
- FAST VP moves small chunks:
 - ▶ 7.5 MB moves are very efficient
 - ▶ New performance profiles are more readily noticed



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

47

FAST VP is dramatically different from other tiering tools.

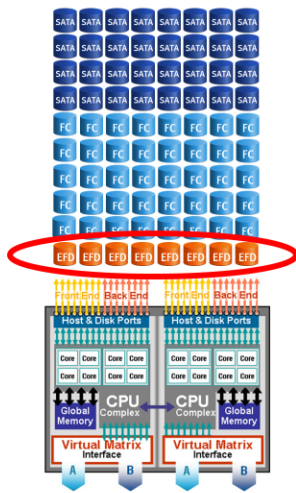
First, FAST VP is more responsive. After the initial review period, there are ongoing reviews to set the thresholds every 10 minutes.

More importantly, the system looks at performance continuously. So, when the performance of a sub-LUN data chunk crosses one of the set thresholds, it can be moved immediately, there is no waiting for a new analysis period.

FAST VP moves data easily across 4 tiers of storage. By having EFD, FC, ATA, and External tiers available, FAST VP can place the right data in the most effective storage class. And, by keeping the pools physically separate, different policies can easily be applied at the storage group level, or even down to a single LUN.

FAST VP also moves data in very small chunks. By moving 8 MB (really 7.5, but close enough) chunks of capacity, the system is more agile. High I/O levels are able to be much more isolated, and the system can take action more quickly.

Balancing EFDs Across the Backend



- A balanced backend EFD configuration is a recommended part of a successful FAST VP deployment with even distribution of EFDs across DAs and backend FC loops
- Multiples of 8 EFDs per VMAX 20k engine
- Multiples of 16 EFDs per VMAX 40k engine
- All engines to be configured with same number and size of EFDs:
 - ▶ Total of 16, 32, 48 etc. EFDs with 2 engine VMAX
 - ▶ Total of 32, 64, 96 etc. EFDs with 4 engine VMAX
 - ▶ Total of 64, 128, 192 etc. EFDs with 8 engine VMAX
 - ▶ This does NOT include EFD spares !
- Same size of EFDs to be used, assuming EFDs to be configured in a single pool and used by one or more FAST policies

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

48

A balanced backend EFD configuration is an essential part of a successful FAST VP deployment with even distribution of EFDs across DAs and backend FC loops.

The recommendation is to implement 8 EFDs per VMAX engine, or multiples of 8, across all VMAX engines, as is shown in the diagram highlighted within the red oval.

Below are the recommended number of Flash drives for various size VMAXs:

<u>1 engine</u>	<u>2 engine</u>	<u>4 engine</u>	<u>8 engine</u>
8, 16, 24, 32 ...	16, 32, 48, 64 ...	32, 64, 96, 128 ...	64, 128, 192, 256 ...

The above does NOT include spare EFDs. The use of the same size EFDs within a single pool is critical to ensure even loading of EFDs and DAs. Also, as shown, a balanced configuration of FC and SATA disks is recommended, but is less critical, as each FC and SATA can only place a relatively small load on a DA, whereas a single EFD can place a significant load on the DA. Balancing EFDs across DAs is critical, as a single EFD drive is capable of consuming a DA. As you add more EFDs, performance will not scale linearly if the DAs become saturated.

General Recommended RAID Levels for FAST VP

- EFD Tier RAID 5 (3+1) recommended for EFD Tier:
 - ▶ RAID 1 is better for performance, but costs more
 - ▶ Other RAID 5 and RAID 6 levels cause the DAs to get very busy because of the number of backend I/Os and parity calculations
- RAID 1 for FC/SAS Tiers offers the optimal balance between cost, performance, and reliability:
 - ▶ No parity calculations needed by the DAs
- RAID 6 (6+2) should be considered as default SATA Tier:
 - ▶ RAID 6(6+2) should be used as it provides better granularity for capacity expansion (in groups of 8 drives at a time)
 - ▶ A disadvantage of RAID 6(14+2) is the need to add 16 drives at a time for balanced configurations



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

49

For the EFD Tier: RAID 5 offers a good balance between cost and performance. Although RAID 1 is better and may occasionally be right for 400 GB EFDs, it requires twice the amount of needed capacity, which represents a large expense for this class of storage. RAID 6 for the EFD tier should be avoided unless the customer insists on it and understands its impact. Each host write and FAST VP promotion will create 3 reads and 3 writes, significantly lower the host I/O load the EFD tier can support. Also the load on the DAs will be significantly higher, meaning more VMAX engines will be required to support a given host I/O load.

For the FC/SAS Tier: RAID 1 offers the right balance between performance, cost, and reliability. Using RAID 1 relieves the DAs of parity calculations and extra I/Os during writes. RAID 5(3+1) can also be considered. RAID 6 for the High Performance tier should be avoided unless the customer insists on it and understands its performance and cost impact.

For the SATA Tier: RAID 6 is recommended because it offers the best reliability. Since the SATA tier is expected to host storage, which is rarely accessed, the high demands on the DAs caused by RAID 6 is not expected to create an unduly heavy load. If 14+2 protection is chosen, 16 drives will need to be added to maintain DAs symmetry whenever storage needs to be added to the SATA tier. With a single engine, VMAX RAID 6 (6+2) should be used. RAID 6 (14+2) will result in 2 disk members configured on the same DA slice.

Choosing FAST Policies

- A FAST policy associates a storage group with up to 4 tiers (DP is still 3 tiers)
- Individually, each FAST tier is allowed to have a maximum participation of 100%, thus allowing for an aggregate of 300%:
 - ▶ The effect of a 100%+ allocation permits the Symmetrix to optimize data placement dynamically across tiers
 - ▶ While this may be good for application performance, it is difficult for planning and charge back
- The percentage of storage specified for each FAST tier in the policy, when aggregated, must total at least 100%:
 - ▶ This means that the allocations within tiers is static and a FAST related promotion must be matched with a corresponding demotion within the storage group



A FAST policy associates a storage group with up to four tiers (three tiers for DP). Each tier may have a maximum participation of 100%. At a minimum, the percentage of storage specified for each tier in the policy when aggregated must total at least 100 percent.

If the sum of all tier Allocations in a policy is 300%, then it means that at any given time each associated storage group can be placed entirely on any tier and FAST will locate the devices based on performance criteria. In this case, all storage of the tiers is shared dynamically by all storage groups associated with that FAST policy.

While this is good from a performance perspective, it is unpractical for planning purposes, because you don't want to have to provision 300% of the needed storage capacity in your array. In addition, it is difficult for chargeback, because an application can be using different amounts of storage from different tiers at different times.

Consider a different case where a policy is configured with three tiers and the three tiers have a maximum allocation of $i\%$, $j\%$, and $k\%$ respectively. When the sum of the tier allocations ($i\% + j\% + k\%$) is 100%, there can be no flexibility of storage allocations between the different storage groups.

This policy definition restricts FAST movements such that any promotion of a device or a thin device extent must be matched by a corresponding demotion of a device or a thin device extent from the same storage group. This method facilitates chargeback because we can be certain that a storage group always uses the pre-assigned percentages of storage from each tier.

Of course, the sum of $i\%$, $j\%$, and $k\%$ can be higher than 100, but not as much as 300. This would give the Symmetrix some flexibility while not requiring that 300% of needed storage capacity be provisioned.

FAST VP – FAST by Storage Group

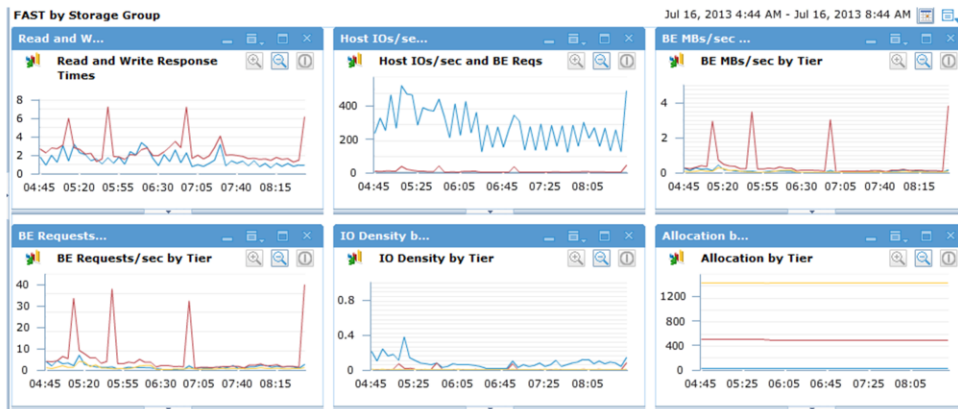
- Monitor>EMC Dashboards>
 - ▶ FAST by Storage Group> select a Storage Group
 - ▶▶ Read and Write Response Times
 - ▶▶ Host IOs/second and BE Requests
 - ▶▶ BE MBs /sec by Tier
 - ▶▶ BE Requests/sec by Tier
 - ▶▶ IO Density by Tier
 - ▶▶ Allocation by Tier



There are three FAST Dashboards in the EMC Dashboards section. All of the FAST Dashboards are FAST VP based. These dashboards are a good way of monitoring your Thin pools and FAST policies.

FAST VP – FAST by Storage Group – Example

- Monitor>EMC Dashboards>Fast by Storage Group



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

52

This is an example of the EMC Dashboard - Fast by Storage Group. This is where you want to monitor if investigating the performance of your storage groups.

FAST VP – FAST by Tier

- Monitor>EMC Dashboards>
 - ▶ FAST by Tier> select Virtual Pool Tier
 - ▶▶ Read and Write Response Times
 - ▶▶ IO Density by SG
 - ▶▶ BE Requests by SG
 - ▶▶ BE MBs Transferred by SG
 - ▶▶ Allocation by SG
 - ▶▶ Tier Ingress/Egress

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

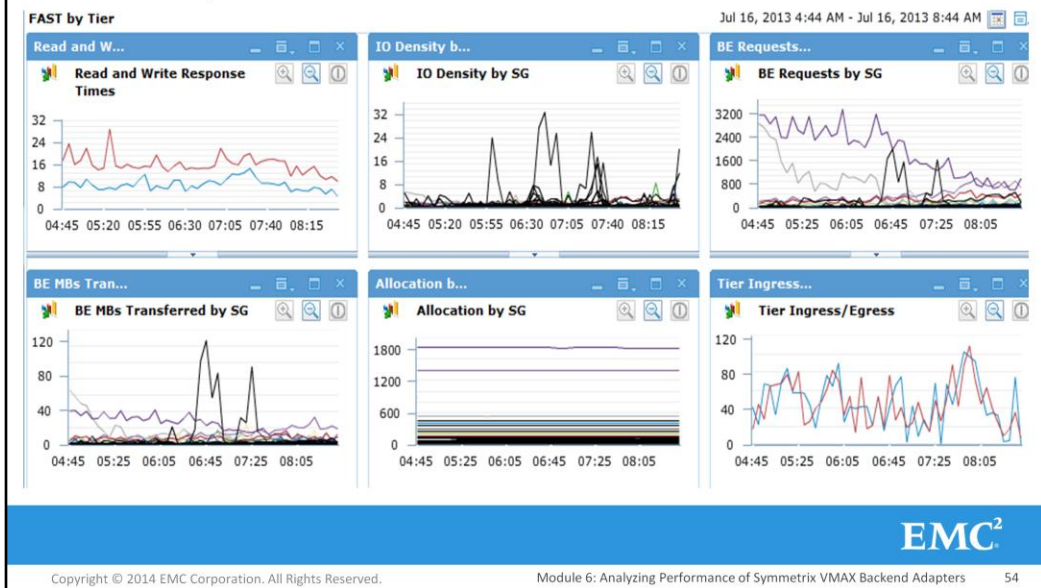
53

These are the graphs shown for the Fast by Tier Dashboard. It allows an alternate view of the system by tier. I/O Density is a measure of I/Os in a gigabyte of storage, your EFD drives should have the highest density.

Each of the FAST Dashboards shows six graphs.

FAST VP – FAST by Tier – Example

- Monitor>EMC Dashboards> FAST by Tier
 - ▶ FAST by Tier> select Virtual Pool Tier



This is an example of the EMC Dashboard - Fast by Tier. Changing the time range may be of value when evaluating peak periods.

FAST VP – FAST Policy

- Monitor>EMC Dashboards>
 - ▶ FAST Policy>select FAST VP Policy
 - ▶▶ IO Density
 - ▶▶ FAST Internal Activity
 - ▶▶ BE Throughput
 - ▶▶ BE Requests
 - ▶▶ FAST Policy Tier Capacity
 - ▶▶ FAST Policy SF Capacity

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

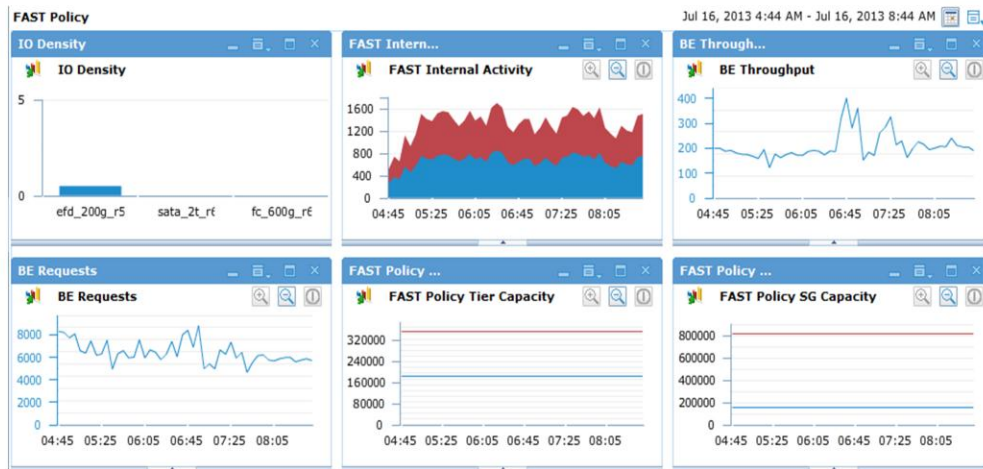
55

These are the graphs shown for the Fast Policy Dashboard. This is probably the dashboard you will want to start your monitoring with, then you could move on to the FAST by Storage Group Dashboard.

Many of the graphs plot multiple metrics.

FAST VP – FAST Policy – Example

- Monitor>EMC Dashboards> FAST Policy
 - ▶ FAST Policy>select FAST VP Policy



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

56

This is an example of the EMC Dashboard - Fast Policy. This is useful for capacity planning and trends. As with the other dashboards, you have a choice between Diagnostic and Historical.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 3: Summary

During this lesson the following topics were covered:

- Symmetrix devices to achieve workload balance
- Striping to distribute workloads over many backend components
- Virtual Provisioning
- FAST VP



This lesson covered the choices available for balancing workloads on the backend of the Symmetrix.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 4: Symmetrix Backend Optimization Algorithms

During this lesson the following topics are covered:

- Monitoring the effectiveness of prefetching
- Symmetrix Priority Control to prioritize application performance



This lesson covers Symmetrix backend optimization algorithms.

Symmetrix Prefetch

- Trigger: Host Director detects reads from two consecutive *drive tracks*, begins prefetch task for device
- Can have many simultaneous prefetch tasks
- RAID-5, RAID-6, Meta, and Thin Device prefetch tasks follow the stripes across devices and drives
- Two *tracks* are read into cache ahead of read stream
- As sequential stream continues, more read ahead
- Prefetch task remains for an additional 120 seconds after sequential I/O stops to avoid thrashing



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

59

Enginuity's prefetch algorithm collects statistics about the sequential patterns of each logical volume. Based on the addressing of past I/Os and system utilization factors, Enginuity decides whether to prefetch more data, how many tracks to prefetch, when to prefetch, and for how long to keep the prefetched data in cache.

The host directors detect prefetching opportunities and notify disk directors to begin prefetch tasks. When reads from two consecutive disk tracks are detected, a prefetch task is started for the device. This means that a prefetch task will only begin when the read stream has crossed over a track boundary; multiple consecutive reads within the same track will not trigger prefetch.

Eight or more prefetch tasks can be active on a single device at a time. Considering the number of devices in a modern Symmetrix array, this means that the disk directors can be monitoring a considerable number of prefetch tasks at once. Prefetch tasks for RAID-5 and RAID-6 devices follow the stripe set across disks to prevent the task from halting each time a stripe boundary is crossed.

Once the task begins, the disk directors try to prefetch data into cache at least two tracks in advance of the current read address. As the sequential stream continues, more read ahead will be performed.

Once the sequential read stream stops, the prefetch task will remain monitoring the activity for another 120 seconds before terminating. If the sequence begins again before the 120 seconds has elapsed, the same task will continue the prefetch work. This prevents tasks from becoming a performance issue by starting and stopping frequently.

Measuring Prefetching

- Volume
 - BE Prefetched Tracks/sec
 - BE Prefetched Tracks Used/sec
 - % Sequential I/O
 - Seq Reads/sec (prefetch trigger)
- BE Director
 - Prefetched Tracks/sec
- System
 - Prefetched tracks/sec

Tracks used:

Incremented *each time* a read from a track that was prefetched happens

% Sequential I/O:

$100 * (\text{total seq IOs per sec} / \text{total IOs per sec})$

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

60

Several measures are available to track prefetch activity in the Symmetrix:

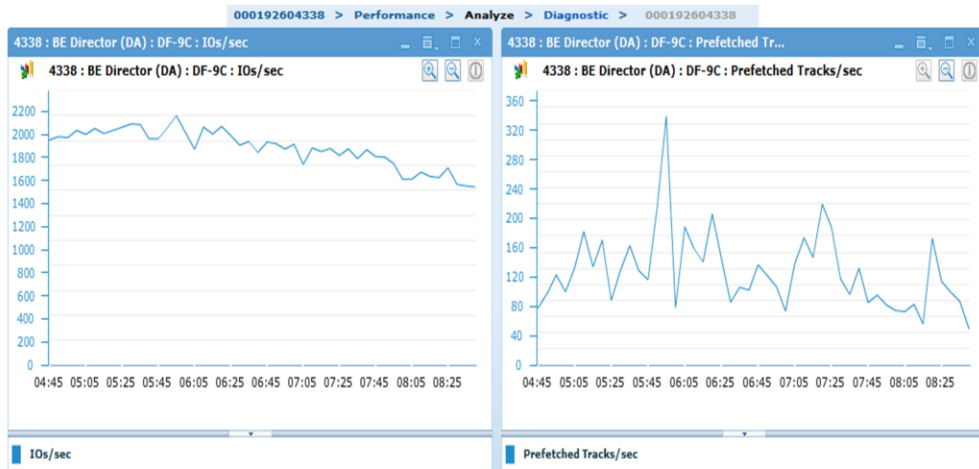
1. **Sequential Reads Per Second:** As mentioned earlier, this measure is really used internally to signal to a disk director when a prefetch task should start. It is not a good measure of sequential workload behavior. You should see increases in this measure when many prefetching tasks are started, but otherwise the Sequential reads per second measure will not help your analysis much.
2. **Prefetched Tracks Per Second:** Number of tracks prefetched per second, measured at the volume, BE, and system level.
3. **Tracks Used Per Second:** This measure is incremented whenever data is read from a track that was loaded into cache by a prefetch task. This metric will not strictly be counting in units of tracks. If several small I/Os are read from a single prefetched track, the counter will increment several times. If one large I/O that spans several tracks is read, the counter will increment several times. So, it should not be relied on to correspond directly to either I/Os or prefetch activity.

Since “BE Prefetch tracks per sec” count actual tracks, while “BE Prefetched Tracks used per sec” does not, comparing them directly can be confusing. It is not unusual to find the “Tracks used per sec” measure significantly higher than the “Prefetched tracks per sec” measure, for instance, and in rare circumstances, it can be even higher than the system-wide I/O per second measure.

% Sequential IO: A calculated value: $100 * (\text{total seq IOs per sec} / \text{total IOs per sec})$

What to Look For – Backend

- General Sequential Activity
 - ▶ I/Os per sec vs. Prefetched Tracks per sec (Tracks vs. I/Os)



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

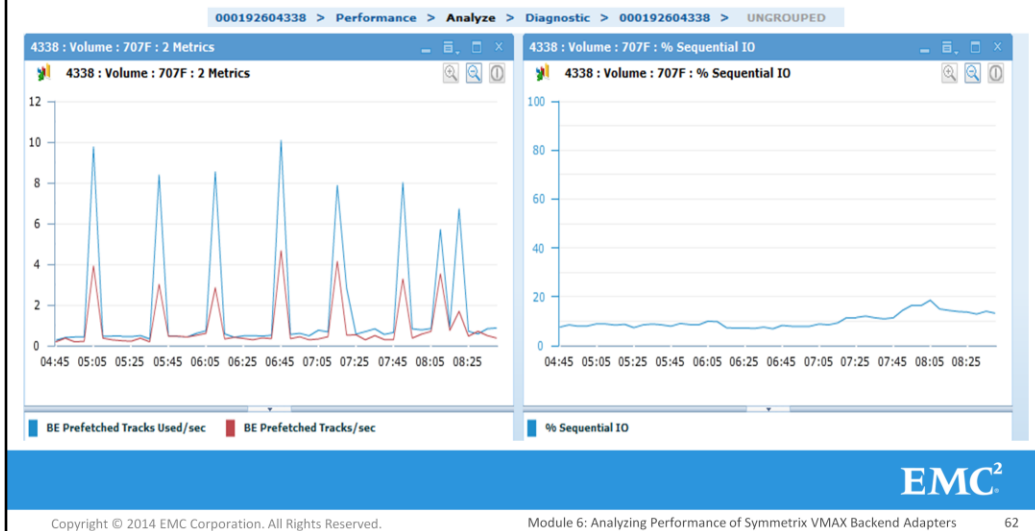
Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

61

You can evaluate general sequential activity on the array by comparing the I/O per second and the Prefetched tracks per second. Note that the chart has two scales. Since one is measuring in units of I/Os (perhaps 8 KB) while the other is measuring in units of tracks (always 64 KB), do not expect an exact mathematical correspondence. Rather, use the two measures to detect increases or decreases in sequence; at least sequence that is detected by the prefetch algorithm. If the I/O per second increases but the prefetched tracks per second has remained the same, then the amount of sequence has actually gone down as a percentage of the total I/O.

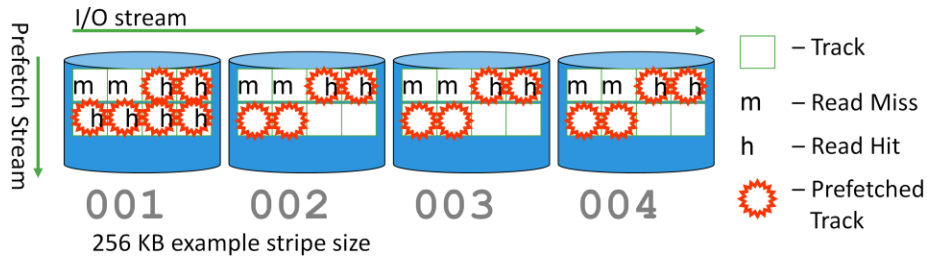
What to Look For – Volume

- General Sequential Activity
 - ▶ Prefetched Tracks Used vs. Prefetched Tracks per second
 - ▶ % Sequential I/O



You can also evaluate general sequential activity at the volume level by comparing the Prefetched Tracks per second and the Prefetched Tracks Used per second. Charting the % Sequential IO may be useful as well.

How Host Striping Can Hinder Prefetch



- Striping applies sequential I/O stream *across* devices
- Front-end directors detect and prefetch *within* a device
- Each time the sequential stream jumps to a new device, a new prefetch task must start over again
- Use host striping when benefits of distributing load outweigh loss of prefetch optimization

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

63

This class has discussed striping as an effective way to spread workload over many physical drives as a performance enhancement. But, while it is very good at reducing the utilization of individual drives, striping will typically reduce the performance of prefetching tasks.

The reason is that the directors view an I/O stream that continues over more than one device as a series of separate streams. Host-based volume manager striping is built on a collection of individual Symmetrix devices. A sequence can be detected on a device and a prefetch task started, but the read ahead will be of data from the device only. Each time the sequential stream jumps to a new device, a new sequence has to be detected before prefetching can start. This guarantees several misses.

If the sequential stream is long enough, it can “wrap” around to the first physical disk again. If the prefetch task is still active (remember, it will terminate after 120 seconds of activity), it will detect further sequential activity and continue the prefetching. This will cause more hits on that device as desired.

In the example above, 64 KB sequential reads are being striped across a device using a 256 KB stripe size (four reads per stripe). Since individual prefetch tasks must be started on each new device, the first two reads of each new device are a miss, even though the sequence is unbroken from the point of view of the host. The read stream will return to the first disk again after 4x256 KB of I/O, or 1024 KB. Assuming the prefetch task on the first device is still active, these reads will be hits.

Symmetrix Priority Controls for Thick Devices

- Affects the ordering of queues within the Symmetrix
 - ▶ High priority I/Os moved to the front if they spend too long in the queue
 - ▶ Consistent quality of service for high priority applications
 - ▶ Prevents low-priority applications from affecting high-priority applications when queues build up
- Same setting affects Two Queues:

Read Requests in Disk I/O Queue:

- Only affects devices that share a disk

Write Requests in SRDF/S Link Queue:

- Only affects devices that use the same SRDF links



In an active environment, the directors will build up queues of I/Os that need to be satisfied. Through unlucky timing, an I/O from a critical, tier one application might be stuck at the end of a queue behind a series of lower-tier applications. This is another example of a lower priority application “stealing” resources from a tier one application.

Symmetrix Priority Controls is a way of giving priority to critical thick devices in queues. A single setting controls the device priority in two queues. One queue is the Disk I/O Queue, used by a disk director to order all I/Os for a disk. Each disk has its own queue. The priority control only affects read requests, the I/O type most dependent on the backend hardware speed, promoting them to the head of the queue under certain circumstances.

The other queue is the SRDF/S Link Queue, used by an RDF director to order all I/Os sent across an SRDF link. Each link has its own queue. The same priority control affects write requests in this case. In fact, only writes are sent across the SRDF links when synchronizing normally. Higher priority I/Os may be promoted to the head of the queue under certain circumstances. Since the application must wait for the write to be transmitted and acknowledged in SRDF/S, moving an I/O to the front of the queue can represent a significant performance improvement.

Symmetrix Priority Control Operation

- Assign each Device a Priority
 - ▶ 1-16 or Time to Become Urgent (ms)
- I/O from device promoted to front of queue when time expires

I/O Priority	Time in Queue
4 (10 ms)	9.0 ms
4 (10 ms)	8.0 ms
8 (40 ms)	6.0 ms
4 (10 ms)	3.0 ms
2 (2 ms)	2.0 ms
8 (10 ms)	1.0 ms

← Promote to front

Pri	TBU (ms)
1	0
2	2
3	5
4	10
5	15
6	20
7	30
8	40
9	60
10	80
11	120
12	160
13	240
14	320
15	480
16	640

Note: A disk or link queue must have at least four I/Os for prioritizing to take effect.

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

65

Under Symmetrix Priority Controls, every device can have a single priority setting. Although it is displayed as a number from 1 to 16, the priority more accurately represents a Time to Become Urgent, in milliseconds. The priority and time relationship is shown on the right of this slide. When an I/O has been in a queue for longer than its Time to Become Urgent, it is promoted to the head of the queue. You can see that the lower priority numbers have shorter time periods. Devices with lower priority numbers are more likely to be promoted, and therefore have higher performance priority.

A disk or SRDF link queue must contain at least four I/Os for the Symmetrix Priority Controls to start rearranging the ordering.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 4: Summary

During this lesson the following topics were covered:

- Monitoring the effectiveness of prefetching
- Symmetrix Priority Control to prioritize application performance



This lesson covered Symmetrix backend optimization algorithms.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 5: Impact of RAID Protection on Backend Performance

During this lesson the following topics are covered:

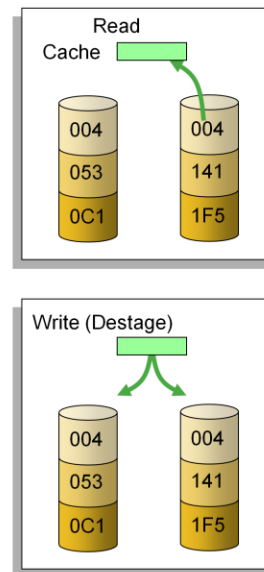
- Relate each device protection type to its backend performance impact
- Compare the relative performance of each device protection type
- Estimate the amount of traffic from the workload type and device protection



This lesson covers impact of RAID protection on the Symmetrix backend performance.

Mirror (RAID-1) Protection

- Two HVEs for every device
 - ▶ Read from either HVE
 - ▶ Write to both HVEs during destage
 - ▶ Drive failure: reads and writes access alternate HVE
- Good:
 - ▶ Potential double read performance with DMSP
 - ▶ Unprotected performance when drive fails
- Bad:
 - ▶ Twice as many drives needed for storage
 - ▶ Two physical writes for every destage



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

68

Mirroring, or RAID-1 protection, allocates two hyper volume extents to store the data for a single device. The duplicate copies protect the data: if one drive fails, the data is still available on the other drive.

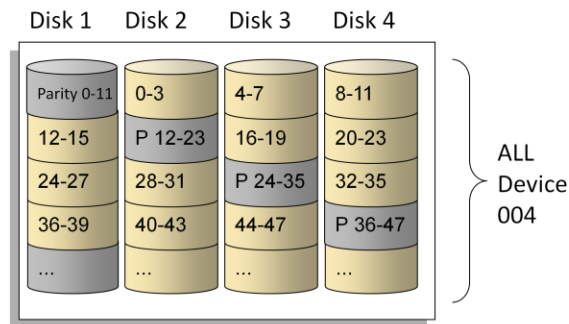
One of the performance benefits of mirroring comes from the ability to read data from two locations. The Symmetrix Dynamic Mirror Service Policy (DMSP) attempts to improve I/O performance by carefully selecting the HVE to read from. This algorithm can double the performance when compared to an unprotected device under the right circumstances.

Another performance improvement comes from the fact that performance falls only to that of an unprotected device if one of the disks fails. While this means DMSP is no longer able to benefit performance, a mirrored pair performs better with a drive failure than a parity protected device.

Just one operation is needed to read a value into cache, even if it's mirrored, due to the Selective Cache Mirror Optimization.

RAID-5 3+1, RAID-5 7+1 Protection

- Parity striped across all HVEs
 - ▶ Stripe size: 4 tracks (256 KB)
- Three or seven data elements for each parity element
 - ▶ Four or eight HVEs total
- Data from **single device** spread across multiple drives
- **Drive failure:** Reads calculate missing data from remaining elements + parity.
 - ▶ Writes update parity only



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

69

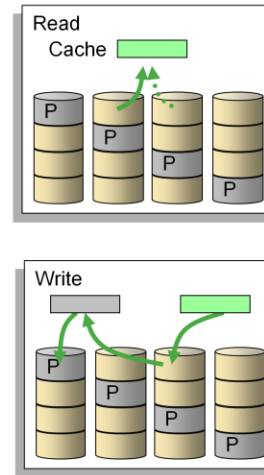
Symmetrix RAID-5 devices follow the standard RAID configuration rules: the data for a single device is written in stripes across multiple physical drives, with each stripe protected by a parity block. The stripe size is fixed at four tracks (256 KB). RAID-5 devices can be configured with three data elements for each parity element (3+1), or seven data elements per parity (7+1). You need either four or eight hyper volume extents to configure these devices, and they must all be on separate physical drives.

During a drive failure, missing data can be calculated from the remaining data elements and the parity. Writes can still be committed to the failed drive data space by just updating the parity. It requires more I/Os to calculate the missing data, so performance is reduced during the time that the drive has failed.

Typically, Dynamic Sparing is configured in a Symmetrix array to help this issue. If a spare drive is available, it will logically move into the place of the failed drive, and the data will be rebuilt. Once the rebuild is finished, performance returns to normal. At some future time, the failed drive will be replaced, and the data again rebuilt, releasing the dynamic spare.

RAID-5 Read Write Operations

- **Read:**
 - ▶ Retrieve directly from HVE[s]
(more than one backend I/O if read crosses stripe boundary)
- **Write:**
 - ▶ Write command to data HVE
 - ▶ Calculate XOR differences of old and new data
 - ▶ XOR-write differences to parity HVE
- **XOR calculations performed by:**
 - ▶ Drive or director depending on load



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

70

A read from a RAID-5 device simply retrieves the necessary data. Since each data element is only 256 KB in length, the data from a single read might be spread across more than one hyper volume. When this happens, you can expect a minor performance degradation when compared to single drive reads, since two drives will experience seek and latency delays. Properly aligned data (discussed in the Cache module) will help avoid this problem.

In a write operation, data needs to be written to the appropriate data element and new parity written to the parity element that protects it. Traditionally, the parity calculation was always performed by the drive itself, using an “XOR read-write” operation. In this operation, the drive reads and locally buffers the existing data, compares it to the data to be written, and then writes the data. The drive then returns the differences, or “XOR” value of the old and new data. To update the parity, the director sends the differences data to the parity element with an “XOR-write” command to change the necessary parity bits.

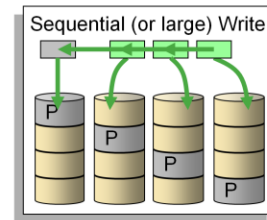
In 2007, a parity calculation optimization was introduced. By default, parity is calculated by the drive, as we just discussed. But, if the drives are heavily utilized, the parity calculation will be done by the directors in cache. This helps reduce the work done by the drive, improving the performance of the slower, heavily-utilized components.

Enterprise Flash drives never calculate parity. The directors always perform the calculation for them.

RAID-5 Optimizations

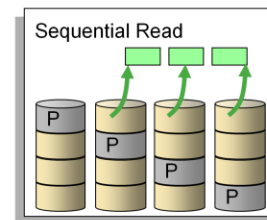
- Write Full Parity Set

- ▶ If write spans *whole stripe set*, no need to read old data to calculate parity
- ▶ Parity calculated entirely in cache
- ▶ Optimizes *large or long sequential* (min. 4 track X # data elements) writes



- Parallel Prefetch

- ▶ Prefetch task initiated on each director that monitors device
- ▶ Excellent sequential read performance for single device; not so effective with many devices on same drives



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

71

Several optimizations have been implemented for RAID-5 devices. The disk directors can optimize long sequences or large writes to these devices. If the write stream overwrites an entire stripe set (tracks 0-11 in a 3+1 device, for instance), then the entire parity block will need to be re-calculated from the new data. Since the new parity can be entirely determined from the writes in cache, there is no reason to read the existing data from the disks. The parity is calculated in cache and the extra reads from the disk controller is not needed.

As we have seen, striping schemes can hinder prefetch effectiveness. This is because a sequential I/O read from the host will not all go to the same drive when striping is used. Separate disk directors need to start separate prefetch tasks on each of the drives. Without optimization, each director would wait for a few I/Os before recognizing the sequential nature of the workload before starting the task.

With the optimization, the directors treat all the data elements as one device with regard to prefetching. A sequence detected anywhere in the address space of the device triggers a prefetch task. The task will continue reading data ahead of the read stream, even when it jumps to a stripe on a different physical drive. The directors keep the task running by passing the execution block to the next director in the data space.

The parallel prefetch optimization actually produces better sequential read performance for a single RAID-5 device than for a single RAID-1 device. However, wider testing shows that when several devices on the same drives are active at the same time, the disk directors have to swap their attention across a large number of prefetch tasks. In this more realistic case, the performance is comparable to RAID-1.

RAID-5 Protection

- Good
 - ▶ Lower cost of raw storage
 - ▶ Automatically stripes data across four or eight drives
 - ▶ Outperforms Striped Meta Devices for sequential read streams
- Bad
 - ▶ Additional operation needed for write (except for full-stripe write)
 - ▶ Read from all devices required during drive failure or rebuild
 - ▶ Remaining RAID set is unprotected during rebuild
 - ▶ DMSP not available



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

72

The main advantage of RAID-5 protection is the lowered storage cost. With RAID-1 protection, twice as much raw drive space is needed to protect the data. With RAID-5 3+1, and with RAID-5 7+1 much more raw space is needed.

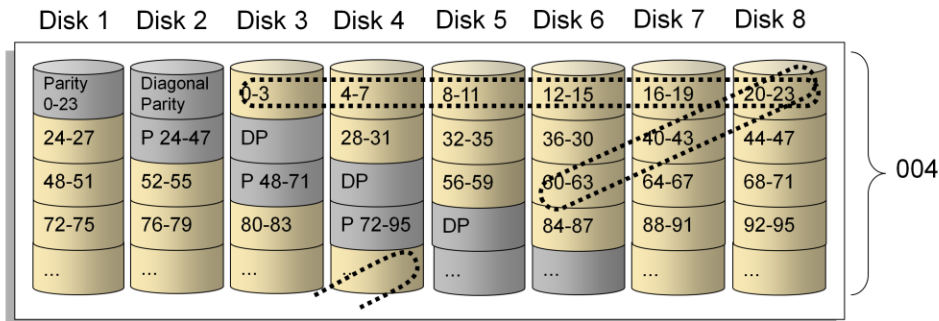
Data is automatically striped in RAID-5 devices, helping to spread the workload on to more physical disks. The parallel prefetch feature enhances sequential read performance, making them a better choice than striped RAID-1 meta devices or host-striped devices.

RAID-5 devices do have some detriments. A normal write operation requires an additional transfer on the backend. While this will not immediately affect write hit performance, it takes longer to destage data in a busy array, and that could cause cache or backend issues.

When a drive fails, it must be rebuilt from the remaining drives. This long process puts extra traffic on the backend, and leaves the other drives in the RAID set unprotected.

RAID-6 6+2, RAID-6 14+2 Protection

- **Horizontal Parity:** Across rows
- **Diagonal Parity:** Diagonally
- Six or fourteen data elements for each Horizontal and Diagonal parity
 - ▶ Eight or sixteen HVEs total
- Data from single device spread across multiple drives
- **Availability:** No data loss with one or two failed drives



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

73

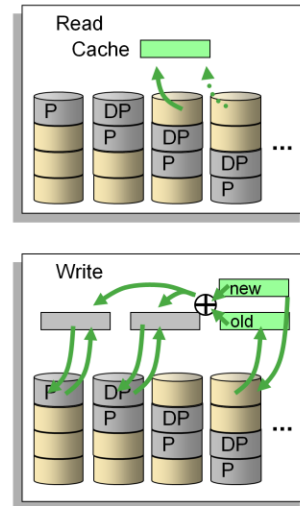
RAID-6 devices are similar to RAID-5 devices, but data is protected by two independent parity elements. The standard, or “horizontal” parity element protects data “across” the drives, just as RAID-5 does. The diagonal parity element protects data diagonally across the drives. This dual coverage ensures that data can still be recovered even if two drives in the set have failed.

As with RAID-5, the data from a single logical device is spread across several physical drives. RAID-6 devices are available in two sizes: six data elements to two parity (6+2), and fourteen data elements to two parity (14+2). These require eight and sixteen drives, as each element must be on a separate drive.

The real advantage of RAID-6 is availability. The dual independent parity schemes protect against data loss if one or even two of the drives fail. When a single drive fails, a spare drive automatically replaces it and the data rebuilt (or rarely, a technician replaces the drive manually). In RAID-1 and RAID-5, the devices are unprotected during the rebuild time. As drive sizes get larger, this rebuild time can stretch to hours. With RAID-6, the devices are fully protected even if two drives fail. Even though it actually takes longer to rebuild a RAID-6 drive, as two parities have to be calculated, it is orders of magnitude less likely to experience data loss since three drives from the same set have to fail before that will happen.

RAID-6 Read Write Operations

- Read
 - ▶ Retrieve directly from HVE[s]
(more than one backend I/O if read crosses stripe boundary)
- Write:
 - ▶ Read old data and parity from drives into cache
 - ▶ Director calculates new parity in cache
 - ▶ Write new data and parity to drives
- Write full parity set optimization
- Parallel prefetch optimization



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

74

As with RAID-5, a read from a RAID-6 device simply retrieves the necessary data. One operation is used unless the array mirrors its cache. The I/O may cross from one drive to the next, causing a minor performance degradation.

In a write operation, the old data, horizontal parity, and diagonal parity values are all read into cache. The directors compare the old data with the new data to be written to compute the differences, or XOR value. The XOR value is then used to calculate a new horizontal and diagonal parity, and the new parity values written to the drives. All parity calculations are done in cache for RAID-6 devices.

The full parity set optimization discussed for RAID-5 devices is also available for RAID-6. Whenever a complete parity stripe is available in cache, there is no reason to read the old values from the drive because the new parity can be calculated entirely from the data already in cache. As with RAID-5, this enhances sequential write performance for one device, but, the effects are limited when writing to many devices on the same disks.

The parallel prefetch optimization discussed for RAID-5 is also available for RAID-6. Benchmarking shows that the sequential read performance of RAID-6 is nearly identical to that of RAID-5.

As you can see from the description of the RAID-6 options, its read performance will be similar to that of RAID-5, but its write performance will be much slower.

RAID-6 Protection

- Good
 - ▶ Lower cost of raw storage than RAID-1
 - ▶ Data available even if two drives from the set are lost
 - ▶ Automatically stripes data across eight or sixteen drives
 - ▶ Outperforms Striped Meta Devices for sequential read streams
- Bad
 - ▶ Slightly more cost of raw storage than RAID-5
 - ▶ Six operations needed for write (except for full-stripe write)
 - ▶ Read from all devices required during drive failure or rebuild
 - ▶ DMSP not available



Copyright © 2014 EMC Corporation. All Rights Reserved.

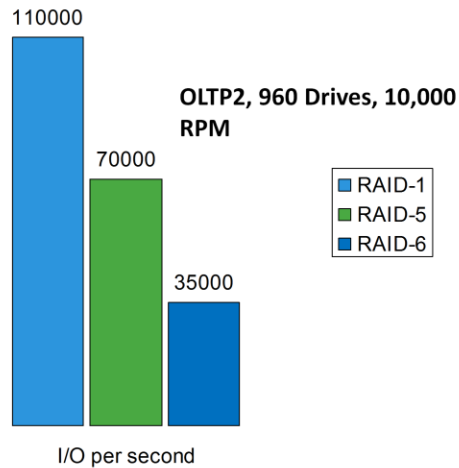
Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

75

The main advantages of RAID-6 protection are the lower storage cost when compared to RAID-1 and high availability. Its other benefits are similar to those of RAID-5.

RAID-6 costs slightly more than RAID-5 in terms of raw storage. The large number of operations needed to perform a write is a serious downside of this technology. It is not generally considered when high-performance is an issue. Its other disadvantages are similar to those of RAID-5.

Protection Type Comparison



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

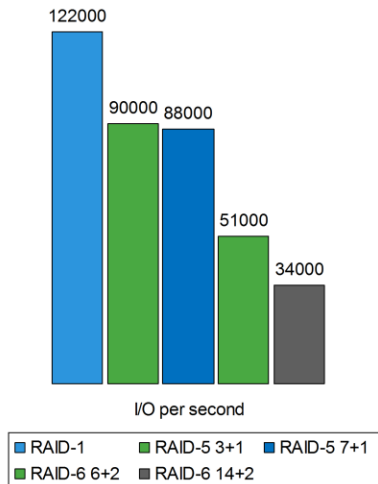
76

One test performed on a array highlights the performance differences across the protection types for a more realistic workload. The test used an OLTP2 workload (Oracle like) serviced by approximately 960, 146 GB, 10,000 rpm drives. The only difference in the test runs was the protection of the devices.

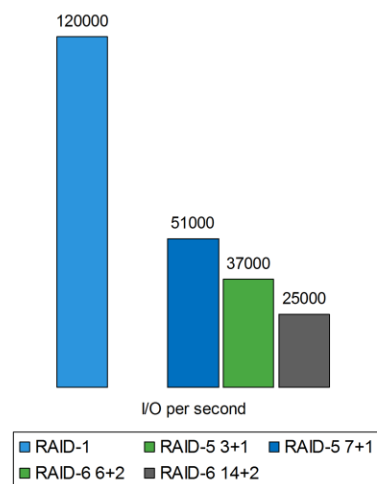
You can see from the chart that the RAID-1 test achieved approximately 50% more I/Os than the RAID-5 test, and more than triple the I/Os of the RAID-6 test.

Protection Type and Capacity Comparison

**Fixed drive count: 480
OLTP1 32 Director, 64 Port**



**Fixed capacity: 60 TB
OLTP1 32 Director, 64 Port**



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

77

A different test performed on a larger array compared protection types over drive count and capacity. In the first test (left), the devices tested were spread across the same number of drives. Since RAID-1 devices occupy the most space (counting protection), they consumed more capacity of the drives than the other protection types. You can see that RAID-1 still outperformed RAID-5, which outperformed RAID-6. The two different RAID-5 protections were nearly identical, while RAID-6 6+2 devices outperformed the RAID-6 14+2 devices.

In the second test, the same capacity of devices were packed onto the minimum number of drives. This helps simulate the effects of a fully-utilized array. The RAID-1 test used the most number of drives, while fewer drives were needed to store the same capacity with RAID-5 and RAID-6 devices. You can see a more pronounced difference between RAID-1 and the other protection types. This is mostly due to the fact that more drives were needed to meet the capacity. The RAID-1 results are nearly the same in the second test because the number of drives was nearly the same. A repeat RAID-5 3+1 test was not performed, probably because the results were similar to the RAID-5 7+1 test.

Protection Performance Summary

	# BE I/Os Per Op.	Random Read Miss	Sequential Read Miss	Write (Destage)
RAID – 1	Read: 1 Write: 2	Best	Best when many active devices on same drive	Best
RAID – 5	Read: 2 Write: 2	Best	Best when one active device per drive	Better
RAID – 6	Read: 3 Write: 3	Best	Best when one active device per drive	Good

Note: Comparisons valid for *same* number of I/Os on *same* number of drives only!

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

78

Let us summarize the work done on the backend and the performance characteristics of the protection types. Reads take just one backend I/O for each protection type, although with RAID-5 and RAID-6 more than one I/O might be required if the I/O crosses from one stripe element to another.

This is reflected in the similar random read miss performance of the protection types. The sequential performance is affected by the prefetching issues we discussed. RAID-1 has the best performance when many devices are active on the same drive, but RAID-5 and RAID-6 will get the best performance when only one device is active on a drive at a time.

There is a wide difference in the number of backend I/Os required for writes, and that is reflected in the write performance. RAID-6, which requires the highest number of I/Os to write data, shows only “good” write performance in benchmarking. RAID-5 has “better” performance, and RAID-1, which requires only two write operations, gets the “best” performance.

Remember that these performance comparisons, and nearly all others, are valid only when comparing the same number of I/Os on the same number of drives. So benchmarking the performance of one RAID-1 device against one RAID-5 7+1 device is not a fair comparison. The RAID-5 device has eight physical drives working in parallel to service the I/Os, while the RAID-1 device has only two. It would be fair to compare eight drives filled with RAID-1 devices against eight drives filled with RAID-5 7+1 devices.

Quick Calculation: How Many Physical Drives Do I Need?

- New database application to be added to array
- Will use mirrored devices
- Expected to produce 1,200 I/O per second
- Similar applications showing 45% read hit, 25% write
- Assume drives at 100% utilization with 100 SCSI commands (read/writes) per second
- Existing applications already sending 20 SCSI commands per second to all drives
 - ▶ Say we don't want them to go beyond 70 SCSI commands with new application

How many Physical Drives do I need to distribute this data over?

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

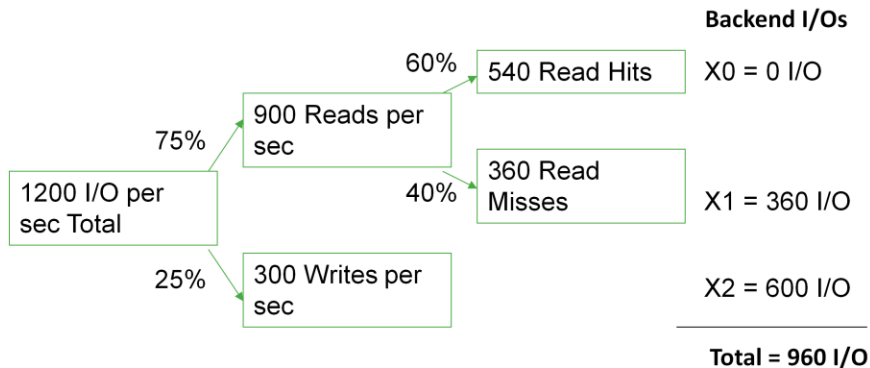
79

Here is a quick planning exercise to determine the number of disks needed to support a new application. The database application will be using mirrored devices, and is expected to be producing a very active 1,200 I/Os per second. Similar applications on similar arrays are showing a 45% read hit ratio and a 25% write ratio.

Assume the drives in the array will reach 100% utilization when they are experiencing 100 SCSI commands per second. Furthermore, we note that all of the drives in the array are currently experiencing about 20 SCSI commands per second under the current workload. To leave a little room for growth, let us ensure the drives will not exceed 70 SCSI commands per second with the new workload.

How many physical drives will be needed to distribute this data over in order to keep the SCSI commands from exceeding 70 per second?

A Rough Estimation on the Number of Drives



$$960 / (70 - 20) = 19.2 \text{ (Round to 20)}$$

Caveat: High locality of reference will reduce actual disk writes.

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

80

Since different I/O types cause different numbers of reads and writes on the drives, we will have to break down the 1,200 I/O per second into different types. First, we will note that 75% of the I/O will be reads (900), while 25% will be writes (300). 45% of the total I/O is expected to be read hits (60% of the reads alone) (540), while another 30% (40% of the total reads) will be misses (360).

- Read hits cause no I/Os on the backend; they can be omitted from this calculation.
- Read misses cause one I/O on the backend if cache is not mirrored. Multiply the 360 read misses by 1.
- Mirrored writes (either hits or misses) cause two I/Os on the backend. Multiply the 300 writes by 2.

The sum of the number of backend I/Os is 960 per second. We want our drives to experience no more than 70 SCSI commands per second, and they are already experiencing 20 per second. No more than 50 additional SCSI commands per second should be directed at the drives. Dividing 960 by 50 gives 19.2. Round this to 20 for convenience.

This shows that if the data set is distributed evenly over 20 drives, each drive should experience no more than 50 additional SCSI commands per second. Since the data is intended to be on mirrored devices, a total of 10 devices are needed. No two hyper volumes from this data set should be on the same physical (a good recommendation for any data set or filesystem).

As with all the rough estimations presented in class, this will not exactly predict the needs of the database. If the database has a high locality of reference, the number of writes actually written to disk will be lower than the number issued from hosts (re-writes, combining several adjacent writes into a single destage). This will cause our estimate to be a bit high.

Command Tag Queuing

- Standard SCSI protocol allows multiple concurrent I/Os to be sent to disk array without waiting for response
- Total outstanding I/Os limited by Throttle Depth or Queue Depth setting (on host)
- Symmetrix Disk Directors can rearrange reads from a single drive to minimize seeking (read hits and writes gain no benefit from this optimization)

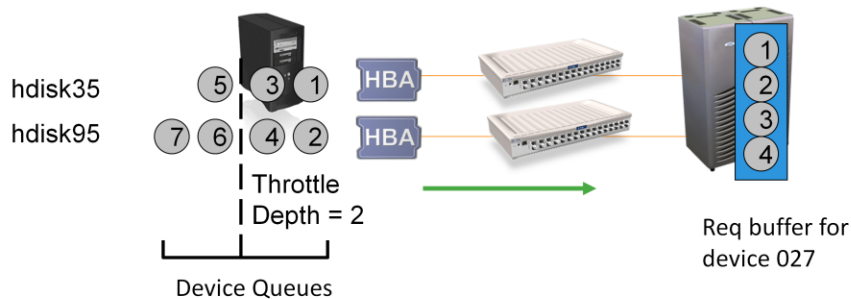


The Symmetrix array makes use of the standard SCSI protocol of Command Tag Queuing. This feature allows a host to send multiple I/Os to an array without waiting for a response. Ten I/Os might be sent to the array before the first one returns under this protocol. The host setting that limits the number of outstanding I/Os is often called the Throttle Depth or Queue Depth, but different manufacturers might use other names.

A Symmetrix array can optimize the I/O retrieval when several outstanding I/Os have been requested for the same drive. In particular, read misses can be rearranged to minimize the seek distance using elevator algorithms. Seek time is a significant part of the delay incurred with random disk access; this optimization technique can raise the maximum number of I/Os processed per second by 30%.

This optimization technique has no effect on read hits or writes. Read hits do not require disk access, so of course, any drive-based optimization techniques will not enhance these operations. The speed of writes is largely governed by the cache access. The response time in a write hit depends on the time required to store the data in cache, not on any drive time. Once the write is in cache, disk director algorithms already optimize the destaging process to move the data onto drives with a minimum delay. These algorithms will destage large blocks of data in any order that optimizes the access time. For this reason, the Command Tag Queuing topic will not benefit writes.

Command Tag Queuing – Example



- Two host devices (hdisk35, hdisk95) represent parallel paths to Symmetrix device 027
- Host submits I/O up to the Throttle Depth setting (2)
- I/O requests stored in the request buffer for the device
- Maximum request buffer: 4 per device

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

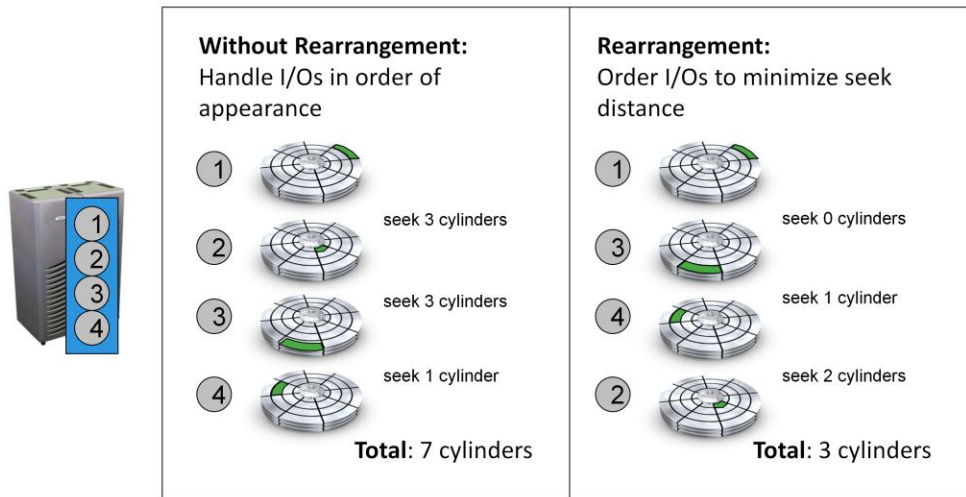
82

This illustration shows an example of the Command Tag Queuing process. Devices “hdisk35” and “hdisk95” on the host represent the two parallel paths to Symmetrix device 027. Some active applications have queued up a total of seven reads distributed across the two host devices. “hdisk35” has three I/Os in its device queue and “hdisk95” has four I/Os in its queue.

Since the throttle depth on this system is set to 2 per device, the first two I/Os in each queue will be sent directly to the Symmetrix without waiting. These four read requests go into the request buffer for the device. Symmetrix systems allow a maximum of four outstanding requests per device, so no more would be accepted until one was serviced. Note that the example Throttle Depth of 2 is for illustration only, typical values are higher.

Had any of the requests been read hits or writes, they would have been serviced directly from cache, and would not have been a part of this example.

Processing Reads from the Request Buffer



EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

83

Now that the disk directors can see all four of the I/O requests simultaneously, they can be handled in any order. Rearrangement can happen to minimize the seek time.

The example on the left above illustrates what might happen if the I/Os were just processed in the order received with no rearrangement. The drive heads must seek over three cylinders between servicing I/O 1 and I/O 2 since they are three cylinders apart. Three more cylinders must be crossed to get I/O 3, and then one more for I/O 4. A total of seven cylinders were crossed to satisfy the requests.

Rearranging the accesses in the order shown on the illustration on the right cuts the seeks by more than half. Since I/O 1 and I/O 3 are both on the outside cylinder, they can both be accessed without seeking. I/O 4 is just one cylinder away, requiring one more seek. I/O 2 is on the inside cylinder, requiring another two seeks to reach. In this case, only 3 cylinders needed to be crossed.

Once the I/O is returned to the host, the tag identifiers will allow the operating system to arrange them back in the original order if they were requested by a single application thread. This is important since a single application expects to receive data in the order requested. This would not be necessary if several threads or processes requested the I/Os, since they are considered independent I/O tasks.

Of course, these illustrations are only for the purpose of expressing this point. Real drives have more than just four cylinders and more than 8 tracks per cylinder.

Calculating Throttle Depth

$$\text{throttle depth} = \frac{4 * \text{meta members}}{\text{number of parallel paths}}$$

- Setting Throttle Depth by Device
 - ▶ Solaris (one setting for all devices), AIX, HP-UX
 - ▶ “meta members” is number of meta members in device; count non-meta as 1
 - ▶ Use recommended default if result is very low
- Setting Throttle Depth by HBA Channel
 - ▶ Windows (via Emulex configuration)
 - ▶ “meta members” is number of meta members on entire channel; count non-meta as 1



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

84

The Throttle Depth calculation is simply a mathematical way of saying this: make sure that the host will send a total of four I/Os to every Symmetrix device. Since each meta member is treated as a separate device, count meta members separately. Remembering that I/O can be queued simultaneously on all paths by which the device can be accessed, be sure to divide the number of devices by the number of paths.

The same formula is used whether the throttle depth is being set by device or by HBA channel. If it is being used to calculate the depth by device, multiply the number of meta members in the device by four and divide by the number of parallel paths. If the device is not a meta, just divide four by the number of paths. If there are many paths, the result will be very low, or even fractional. You might just use the default setting (8 for AIX, HP-UX, 20 for Solaris) rather than lowering it.

Solaris software is somewhat problematic in that it does not give a way to set the Throttle Depth individually for each device, one value is used for all devices. You would typically want regular devices to have a fairly low number, while meta devices with many members should have a large number. The default value of 20 is a compromise that should work in most situations.

Emulex software used to configure Windows hosts sets' Throttle Depth by channel. In this case, the “meta members” part of the formula should be the total number of devices or meta members accessed by each channel.

Quick Throttle Depth Calculation

- A Sun Solaris host is assigned 10 LUNs from the Symmetrix array
- Each host addressable LUN is a meta volume that consists of 8 devices each
- The devices are visible over two paths, and PowerPath is being used to balance the I/O traffic

What Throttle Depth setting is suggested by the formula?

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

85

Here is a quick scenario to illustrate the Throttle Depth calculation. On a Sun Solaris host, ten devices have been assigned from a Symmetrix array. Each device is an eight member meta volume. Each device is visible over two paths.

What Throttle Depth setting is suggested by the formula?

Note: Remember to count the meta members separately: each has its own request buffer, capable of handling up to four read miss requests at a time.

Caveats on Throttle Depth Settings

- Too few means no read miss optimization
- Too many means array returns “queue full” response
 - ▶ Older HBAs could experience long delays or freeze when receiving this response
- EMC Host Configuration Guides give approved, conservative numbers for Throttle Depth on all supported systems
 - ▶ When in doubt, stick with those
- Calculation assumes that queued I/O will apply evenly across all devices
 - ▶ Not realistic



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

86

The correct Throttle Depth setting can help optimize read misses, but in certain circumstances the wrong setting can produce significant problems. If the Throttle Depth setting is very high, the host may send a large number of I/O to the array; too many for it to handle. When this happens, it will return a “queue full” error back to the host. The host should simply retry the request again, possibly after waiting for some outstanding I/O to return. However, some older HBA software would not handle this situation properly, and would experience long delays or would even freeze entirely. Therefore, do not just set the throttle depth to some arbitrarily large number.

The Host Configuration Guides published by EMC give Throttle Depth recommendations for each host that is supported. These numbers are typically much more conservative (lower) than the values you would calculate using the formula. Ordinarily, EMC technicians will just use these recommended values when connecting hosts to Symmetrix arrays.

You will note from the analysis performed in this section that the formula expects that I/Os will be applied to each of the devices so that exactly four end up in the request buffer for each device. In real world environments, data never balances out this evenly. Expect that some of the devices will have a large number of outstanding I/Os while others will have none at all. There is no more exact way to handle the situation.

Do I Really Need to Adjust The Throttle Depth?

- Is the host queue greater or less than Throttle Depth?
 - ▶ Measure host queuing by device
Analyze>Array>Disk Group>Disk>Average Queue Depth
- Are there a significant number of read misses?

- 50 Symmetrix devices, two paths (100 total host physical devices)
- Windows throttle depth currently = 64/path
- Does the formula suggest a higher throttle depth?
- Measured Average queue for each path:
 - ▶ ~1.5 on 20 host devices
 - ▶ ~0.5 on 30 host devices

Will raising the Throttle Depth have any effect?



Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

87

Given that changing the throttle depth can have serious implications to the host, and in any event cannot be done while the device is in use (most systems at least require the device to be closed and reopened to make the setting take effect), you should consider whether a change is really needed.

Check the current size of the queue on the host: if it is smaller than the current throttle depth setting, then you can assume that all of the I/Os currently in the queue are being sent to the array and being optimized. If the size of the queue is larger than the throttle depth setting, then perhaps raising it would do some good.

Also consider checking the number of read misses. Since these are the only types of I/Os this optimization will enhance, there is no need to consider the setting if there are very few. Most workloads do have a significant percentage of read misses, so in most cases some attention should be paid to this issue.

As an example, consider the scenario presented on this slide. The host accesses 50 devices (no meta volumes) on two paths. 100 host physical devices are therefore viewed in a host archive. The current Throttle Depth is set to 64 per path (Windows host). The formula suggests that it should be raised to 100 per path ($4 \times 50 / 2$).

However, looking at the current queuing on the host suggests otherwise. 40 of the devices are measured at an average queue of 1.5, while the other 60 are measured at only 0.5. Assuming the heavily used and lightly used devices are evenly distributed over the two paths (PowerPath will certainly make sure of this), then each path has a total queue of 45 ($40 \times 1.5 + 60 \times 0.5$). The current Throttle Depth allows up to 64 I/Os from each path to be sent directly to the array without waiting, but only 45 I/Os are in the queue. Raising the Throttle Depth will have no effect here. In fact, it could conceivably be lowered to 45 without changing the way I/O is handled.

Roadmap – Directors and Disks

Disk Director and Disk Traffic:

Metric	Issue
• Monitor>All BE Directors(DA) • IOs/sec, Throughput, % Busy • Monitor>BE Directors (DA) Summary	• Are directors balanced?
• Analyze>Array>BE Director Tab>DA • Analyze>Array>Storage Group Tab>SG • Analyze> BE Directors(DA)> DA	• Are disks balanced? • Is utilization low? • Is any disk group heavily utilized?

Note: The number ONE Symmetrix performance problem is disk overutilization caused by imbalance!



Metrics useful for analyzing the backend.

Roadmap –Prefetch

Prefetch:

Metric	Issue
- Analyze>Volume>Select - BE Prefetched Tracks/sec - BE Prefetched Tracks Used/sec	Prefetched tracks used: as high as
- Analyze>Volume>Select % Sequential IO	Correlate sequential activity on devices to the source application

EMC²

Copyright © 2014 EMC Corporation. All Rights Reserved.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

89

Metrics useful for analyzing the backend.

Module 6: Analyzing Performance of Symmetrix VMAX Backend Adapters

Lesson 5: Summary

During this lesson the following topics were covered:

- Related each device protection type to its backend performance impact
- Compared the relative performance of each device protection type
- Estimated the amount of traffic from the workload type and device protection



This lesson covered impact of RAID protection on the Symmetrix Back-end performance.

Module 6: Summary

Key points covered in this module:

- Symmetrix backend architecture and data protection schemes
- Backend performance measures
- Performance benefits of data protection types
- Prefetch operations
- Command Tag Queuing optimization



This module covered on analyzing performance of Symmetrix back-end directors.

This slide is intentionally left blank.



This slide is intentionally left blank.