High Performance Tier Implementation Guidelines

A Dell Technical White Paper

PowerVault™ MD3200, MD3200i, and MD3600i Series Storage Arrays
THIS WHITE PAPER IS FOR INFORMATIONAL PURPOSES ONLY, AND MAY CONTAIN TYPOGRAPHICAL ERRORS AND TECHNICAL INACCURACIES. THE CONTENT IS PROVIDED AS IS, WITHOUT EXPRESS OR IMPLIED WARRANTIES OF ANY KIND.

© 2010-2011 Dell Inc. All rights reserved. Reproduction of this material in any manner whatsoever without the express written permission of Dell Inc. is strictly forbidden. For more information, contact Dell.

Dell, the DELL logo, and the DELL badge, PowerConnect, and PowerVault are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell Inc. disclaims any proprietary interest in trademarks and trade names other than its own.

March 2011
Contents

What is High Performance Tier? ...................................................................................... 2
What are the Performance Advantages of High Performance Tier? ........................................... 2
When Should High Performance Tier be Used? .................................................................... 2
  Number of Disk Drives........................................................................................................ 3
  Workload Dependencies ..................................................................................................... 3
  RAID Types ........................................................................................................................ 3
How is High Performance Tier Enabled and Disabled? ........................................................... 4
What are the Best Tuning Practices for High Performance Tier? .............................................. 5
Performance Analysis ........................................................................................................ 6
  System Configuration for iSCSI for MD32xxi ................................................................. 6
  System Topology for iSCSI for MD32xxi .......................................................................... 6
  System Configuration for iSCSI for MD36xxi ................................................................. 7
  System Topology for iSCSI for MD36xxi .......................................................................... 7
  System Configuration for SAS ......................................................................................... 8
  System Topology for SAS ................................................................................................. 8
  Back-End SAS Cabling Diagram ..................................................................................... 9
MD3200 iSCSI Performance Data Base vs. High Performance Tier ........................................... 9
MD3200 SAS Performance Data Base vs. High Performance Tier ........................................... 11
MD3600 iSCSI Performance Data Base vs. High Performance Tier ........................................... 13
What is High Performance Tier?

High Performance Tier (also known as High Performance Tier Mode) is an optional upgrade to increase performance of the MD3200i, MD3220i, MD3600i, and MD3620i arrays that have a high drive count, solid state drives (SSDs), or high data transfer workloads. This implementation is based on an enhanced firmware algorithm and does not require any new hardware dependencies. Several factors determine the potential performance increase, including the array configuration, host, operating system, host bus adaptor (HBA), number and types of drives, and application workload.

What are the Performance Advantages of High Performance Tier?

High Performance Tier uses the enhanced firmware with the existing hardware to extend the performance of the storage array. Enabling High Performance Tier improves the performance of the array in both data transfer rate (bandwidth) and I/O transactions per second (IOPS). The main advantages are increased bandwidth for sequential and random workloads and increased IOPS for random workloads. For data transfer intensive workloads, both read and write bandwidth can be doubled, while for random access workload patterns, an increase of one-third has been measured.

When Should High Performance Tier be Used?

High Performance Tier is most beneficial when the workload intensity exceeds the controller’s performance capability in base mode and the system configuration supports additional performance. Two primary aspects of workload are IOPS and data rate (bandwidth). These factors combine in varying ratios depending on the workload mix.

To present the MD3200i, MD3220i, MD3600i, and MD3620i arrays with a workload that exceeds the base product limits, several factors must be realized. For high bandwidth applications, the following are essential for optimizing data transfer capability:

- Enough host links to provide the data rates needed to reach the controller’s limit. A single host link to a controller may not provide enough bandwidth to exceed the base performance capability.
- A workload from the host servers adequate to exceed the controller’s base performance specifications. This is usually accomplished with large transfer sizes and enough outstanding IOs to keep the link saturated.
- Sufficient disk groups properly configured to support the required data rates.
- Enough host interface capability to sustain at least 1000 MB/s of data transfer rate per controller (this is a rough guideline).

IO transactions per second are maximized for smaller transfer sizes. If the controller CPU is not fully utilized, maximum IOPS will not be achieved. If the workload is not sufficient to exceed base product performance specifications, or if target configuration is not optimized for performance, then High Performance Tier may not provide any benefit. Maximum IOPS might not be obtained under the following conditions:

- The number of hard disk drives is insufficient to provide enough responses to keep the controller busy.
- The workload does not have broad enough access characteristics to keep enough disks active. If the workload is highly localized (i.e., accesses a relatively small number of disks at any given time) then it will be disk-limited due to the insufficient number of active spindles servicing the controller’s IO requests, regardless of how many disks are behind the controller.
Number of Disk Drives
Maximum performance can be limited if the number of drives is not sufficient. The number of disk drives depends on the workload being supported. A workload of primarily large sequential transfers requires fewer disk drives than does a workload dominated by small transfer, random access requests. As the transfer size increases for random workloads, the dominant array workload characteristic will shift from transactions per second to bandwidth, and disk requirements will morph accordingly.

For strictly large sequential workloads (file server), approximately 24 drives provide sufficient disk bandwidth to saturate a dual controller configuration in High Performance Tier. In a dual configuration, more than 12 drives are required to exceed base mode performance thresholds; a system with a drive count smaller than this is drive-limited and cannot take advantage of High Performance Tier. A single controller can see saturation while streaming to only 12 drives in High Performance Tier.

Disk requirements for small transfer size transaction workloads (email, database server) are harder to assess. Because the performance is often limited by the storage device and not by the controller, the performance is significantly determined by applied workload. For example, if a workload is concentrated on a single controller, that controller may benefit from the High Performance Tier feature, while a dual controller with the same configuration will not. For a workload balanced across both controllers, High Performance Tier can be expected to maximize IOPS with 48 drives in base mode, but make full use of 96 drives for a small transfer workload. As the transfer size increases, the bottleneck shifts back to the path bandwidth, even for random workloads.

Solid state disks (SSDs) are known for their ability to deliver high rates of small transfers and thus require a smaller number to saturate a controller. High Performance Tier is essential for delivering the full capability of these drives. To take advantage of High Performance Tier, a minimum of seven SSDs is required for sequential workloads, and a minimum of three is required for random workloads in a RAID 5 disk group configuration.

Workload Dependencies
The performance of the MD3200i, MD3220i, MD3600i, and MD3620i arrays are dependent not only on configuration but on the entire system, which includes the host, host operating system, host HBA, and application workload. If the workload does not exceed base I/O performance limits, enabling High Performance Tier will provide no benefit. If you are unsure if the current configuration exceeds the base limits, please contact Dell for assistance.

RAID Types
High Performance Tier can provide an improvement in performance for all RAID types supported by the MD3200i, MD3220i, MD3600i, and MD3620i arrays. Characteristic performance for individual RAID types will not change—for example, RAID 0 will generally have better WRITE performance than RAID 5; these types of inter-RAID relationships will remain the same.
How is High Performance Tier Enabled and Disabled?

High Performance Tier is enabled using the MDSM management application by applying a Premium Feature key purchased from Dell. To obtain this key, you must provide the Feature Enable Identifier referenced in MDSM.

1. Select **Premium Features** from the **Storage Array** main menu. The following window displays.

![Premium Features Window]

2. Obtain the number in the **Feature Enable Identifier** field and provide it to a Dell representative, who will generate a key based on this information.

3. Select the High Performance Tier option from the **Premium Features** drop-down list and click the **Enable** button.

4. Select the High Performance Tier key provided by Dell. Once the key has been selected, all MD32xx/MD36xx controllers will perform a reboot automatically. *Please plan accordingly.*

5. Verify that the High Performance Tier option in the **Premium Features** drop-down list now shows “Enabled”. 

---

*Figure 1: Premium Features Window*
What are the Best Tuning Practices for High Performance Tier?

Proper array and server tuning is essential for realizing the advantages of High Performance Tier. If a storage system is not optimally tuned, the benefits from High Performance Tier may not be fully realized. For sequential workloads, use a controller cache block size of 32KB (the default setting is 4KB). Change cache settings by selecting Change ➔ Cache Settings from the Storage Array main menu.

For random workloads, use a controller cache block size of 16KB, and disable cache pre-fetch for each virtual disk. Change virtual disk cache settings by selecting Change ➔ Cache Settings from the Virtual Disk main menu.

If the workloads are a mix of sequential and random IO, we initially recommend using the 16KB cache block with cache pre-fetch disabled, but encourage the system administrator to adjust cache block size as necessary to achieve the optimal results. Cache block size may be changed “on the fly” without requiring a controller reset, and live system tuning may be required in some cases.
Performance Analysis

The purpose of this data and these tests is to show how MD3200i, MD3220i, MD3600i, and MD3620i array configurations using default settings (cache block size=4KB, read cache enabled, write cache enabled, cache mirroring enabled, cache pre-fetch enabled, and segment size=128KB) may benefit from High Performance Tier. It is not the intent of this report to show maximum performance characteristics of the MD3200i and MD3600i series.

System Configuration for iSCSI for MD32xxi

The iSCSI system configuration consisted of a single R710 PowerEdge Server running Windows 2008 x64 SP2 with 32GB of system memory connected to a duplex MD32xxi via a 10GBase-T Extreme x650 switch. The server used two single port Intel 10G AT Server Adapters installed in the R710’s PCIe GEN2 x8 electrical slots. In the server BIOS setup, the Intel I/O Advanced Technology (IOAT) port was enabled (disabled by default). In the operating system, RSS support was enabled, Windows firewall was disabled on all 10G Intel interfaces, and MD32xxi host failover software was installed. Adapter defaults were used except for the following: number of RSS queues set to match number of processor cores (16 in this case) and jumbo frames enabled (9014 bytes). The native Microsoft Initiator that comes with SP2 was used to connect to the MD32xxi with multi-path enabled. On the switch, default settings were used except jumbo frames were enabled, and flow-control was verified enabled. On the target, default settings were used in base mode except jumbo frames were enabled using the maximum 9000 byte MTU size. In High Performance Tier, the cache block size was changed to 32KB for sequential workload testing, cache block size was changed to 16KB and cache pre-fetch was disabled for random workload testing, and jumbo frames were enabled in all cases. The MD32xxi consisted of 96 drives; eight RAID 5 disk groups (12 drives per group) with one virtual disk per disk group were used for the sequential workloads, and 16 RAID 5 disk groups (6 drives per group) with one virtual disk per disk group were used for the random workloads. Disk groups were evenly distributed across both controllers. IoMeter was used to collect the performance data using a single worker per raw disk with a variable queue depth setting.

System Topology for iSCSI for MD32xxi

Eight IPv4 iSCSI sessions were created between the server and target, one connection per session. Each MD32xxi consisted of two controllers (controller 0 and controller 1) having four 1G iSCSI host ports per controller (ports 0 thru 3). In the following, C00 corresponds to controller 0, iSCSI host port 0, and so on. The server network interfaces are referred to as N0 and N1. Four iSCSI sessions were made between N0 and Controller 0, and four iSCSI sessions were established between N1 and Controller 1—only one connection was made to each iSCSI target host port. The following describes these connections: N0<->C00, N0<->C01, N0<->C02, N0<->C03, N1<->C10, N1<->C11, N1<->C12, N1<->C13.
The arrows represent iSCSI sessions, not physical connections. The server has two physical 10G interfaces, and the MD32xxi has eight physical 1G interfaces (four per controller).

**System Configuration for iSCSI for MD36xxi**

The iSCSI system configuration consisted of a single R910 PowerEdge Server running Windows 2008 x64 SP2 with 32GB of system memory connected to a duplex MD36xxi via a 10GBase-T Power Connect 8024 switch. The server used five single port Broadcom 57710 adapters installed in the R910’s five PCIe GEN2 x8 electrical slots using the iSCSI offload feature. The basic operating system configuration had Chimney and RSS enabled, Windows firewall disabled on all 10G Broadcom interfaces, and MD36xxi host failover software installed. Adapter defaults were used except jumbo frames were enabled with an MTU size of 4500 bytes. The native Microsoft Initiator that comes with SP2 was used to connect to the MD36xxi with multi-path enabled. On the switch, default settings were used except jumbo frames were enabled, and flow-control was verified enabled. On the target, default settings were used in base mode except jumbo frames were enabled using the maximum 9000 byte MTU size. In High Performance Tier, the cache block size was changed to 32KB for sequential workload testing, cache block size was changed to 16KB and cache pre-fetch was disabled for random workload testing, and jumbo frames were enabled in all cases. The MD36xxi consisted of 96 drives; eight RAID 6 disk groups (12 drives per group) with one virtual disk per disk group were used for the sequential workloads and random workloads. IoMeter was used to collect the performance data using a single worker per raw disk with a variable queue depth setting.

**System Topology for iSCSI for MD36xxi**

Seven IPv4 iSCSI sessions were created between the server and target, one connection per session. Each MD36xxi consisted of two controllers (controller 0 and controller 1) having two 10G iSCSI host ports per controller (ports 0 and 1). In the following, C00 corresponds to controller 0, iSCSI host port 0, and so on. The server network interfaces are referred to as N0 thru N4. The following describes these connections: N0<->C00, N1<->C01, N2<->C10, N3<->C11, N4<->C00, N4<->C01, N4<->C10.
**System Configuration for SAS**
The 6G SAS system configuration consisted of a single R710 PowerEdge Server running Windows 2008 x64 SP2 with 32GB of system memory connected to a duplex MD32xx using two Dell 6G SAS HBAs with MD32xx host failover software installed. The HBAs were installed in the R710’s PCIe GEN2 x8 electrical slots and default adapter settings used. On the target, default settings were used in base mode. In High Performance Tier, the cache block size was changed to 32KB for sequential workload testing, the cache block size was changed to 16KB, and cache pre-fetch was disabled for random workload testing. The MD32xx consisted of 96 drives; four RAID 5 disk groups (24 drives per group) with one virtual disk per disk group were used for the sequential workloads, and 16 RAID 5 disk groups (6 drives per group) with one virtual disk per disk group were used for the random workloads. Disk groups were evenly distributed across both controllers. IoMeter was used to collect the performance data using a single worker per raw disk with a variable queue depth setting.

**System Topology for SAS**
Four connections were made between the server and MD32xx. Each MD32xx consisted of two controllers (controller 0 and controller 1) having four 6G SAS host ports per controller (ports 0 thru 3). In the following, C00 corresponds to controller 0 SAS host port 0, and so on. Each server HBA interface has two SAS ports—S0 and S1 reside on one HBA, and S2 and S3 reside on the other HBA. Two SAS connections were made between S0, S1 and Controller 0, and two SAS connections were established between S2, S3 and Controller 1. The following describes these connections: S0<->C00, S1<->C02, S2<->C10, S3<->C12.
Back-End SAS Cabling Diagram
An asymmetric cabling method was used on both the SAS and iSCSI systems to interconnect controller and drive shelves. There is no evidence to show asymmetric vs. symmetric cabling improves performance, but it is added here for consistency and best cabling practices. For asymmetric cabling, controller 0 (TOP (0)) connects to the top drive shelf controller (TOP (1)), and controller 1 (BOTTOM (0)) connects to the bottom drive shelf controller (BOTTOM (N)). The drive shelves are then subsequently daisy chained as normal.

<table>
<thead>
<tr>
<th>TOP(0)</th>
<th>BOTTOM(0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP(1)</td>
<td>BOTTOM(1)</td>
</tr>
<tr>
<td>TOP(2)</td>
<td>BOTTOM(2)</td>
</tr>
<tr>
<td>TOP(N)</td>
<td>BOTTOM(N)</td>
</tr>
</tbody>
</table>

The “N” refers to the last shelf in the configuration.

MD3200 iSCSI Performance Data Base vs. High Performance Tier
The following charts show Base vs. High Performance Tier performance for sequential READ (100/0 R/W), sequential WRITE (0/100 R/W) in MiB/s (1MiB=1024*1024B), random READ (100/0 R/W) IOPS, and random WRITE (0/100 R/W) IOPS for transfer sizes ranging from 512B to 4MB.
**MD3200 SAS Performance Data Base vs. High Performance Tier**

The following charts show Base vs. High Performance Tier performance for sequential READ (100/0 R/W), sequential WRITE (0/100 R/W) in MiB/s (1MiB = 1024*1024B), random READ (100/0 R/W) IOPS, and random WRITE (0/100 R/W) IOPS for transfer sizes ranging from 512B to 4MB.
**MD3600 iSCSI Performance Data Base vs. High Performance Tier**

The following charts show Base vs. High Performance Tier performance for sequential READ (100/0 R/W), sequential WRITE (0/100 R/W) in MiB/s (1MiB = 1024*1024B), random READ (100/0 R/W) IOPS, and random WRITE (0/100 R/W) IOPS for transfer sizes ranging from 512B to 4MB.
RANDOM WRITE IOPS

![Graph showing RANDOM WRITE IOPS for Base and Turbo tiers. The graph compares IOPS across different data block sizes, with Base and Turbo tiers depicted by different markers. The Base tier shows consistently higher IOPS across all block sizes compared to the Turbo tier.](image-url)