HP StorageWorks 8000 Enterprise Virtual Array and Microsoft Exchange Server 2003 – storage performance and configuration

Executive summary

Microsoft® Exchange Server 2003 is a very random and aggressive I/O intensive application that often necessitates dedicated storage to provide high performance. Successful Exchange Server 2003 deployments have a critical dependence on proper sizing and configuring of the storage subsystem to meet requirements for performance, fault tolerance, and reliability.

This white paper addresses best practice considerations when deploying Exchange Server 2003 on the HP StorageWorks 8000 Enterprise Virtual Array (EVA8000) storage array. Testing was conducted on the EVA8000 2C12D model with 168 disks (72GB, 15K-rpm drives). This document includes the following information:

- Best practices for the configuration of the EVA8000 storage array for Exchange Server 2003 to achieve high performance while balancing the demands for high availability
- Analysis of test results conducted on the EVA8000 using the Microsoft Jetstress utility

Key Findings: The following best practice recommendations serve as guidelines for deployment of Exchange Server 2003 on the EVA.

1. For the best combination of performance and availability, Exchange transaction logs and databases should be isolated into separate disk groups on the EVA.
2. As long as the number of spindles are properly sized and configured to support Exchange database activity (providing a recommended 15-20% overhead to handle peak I/O), either a single large database disk group or multiple smaller database disk groups can be configured to support the Exchange database I/O.
3. If the EVA is hosting additional workloads such as non-Exchange application or backup workloads, an additional disk group should be configured to isolate this I/O.
4. For Exchange transaction logs, VRAID1 should be used to provide the best performance and availability.
5. For Exchange databases, VRAID1 should be used to provide the best performance and availability.
6. Leave the default value, **Shortest Queue Service Time**, as the load balancing algorithm when using the HP Full Featured Multi-Pathing Input/Output (MPIO) driver.

Target audience

This document is intended for experienced Exchange storage architects and administrators and requires previous knowledge regarding Exchange and SAN technologies, specifically the HP StorageWorks Enterprise Virtual Array. Additional background Exchange storage information is available at www.microsoft.com/exchange and www.hp.com/solutions/exchange.

Overview

When deploying Microsoft Exchange Server 2003, there are a number of factors that must be addressed to ensure a successful deployment and implementation of the application. To provide a high performance Exchange server environment, the initial design and planning phase is critical, and key decisions must be made around the type of server hardware, the amount of required memory, the network infrastructure, and Active Directory design. One area that requires extensive knowledge and discussion is the proper sizing and configuration of the storage subsystem.

Microsoft Exchange Server is a very I/O intensive application and is extremely sensitive to I/O latencies. Understanding proper storage sizing and best practices is vital, as one of the leading causes of poor Exchange Server performance is under- or mis-configuration of the storage subsystem.

This white paper focuses on providing guidelines for achieving a high performance Exchange configuration while balancing the requirements for high reliability and availability.

The tests documented in this white paper were driven by the Microsoft Jetstress utility, which simulates Exchange I/O load on the Jet database level and is a useful tool for analyzing storage performance.[1] Tests were performed to determine the performance and scalability of the EVA8000 disk subsystem in various configurations, which included testing and quantifying the impact of:

- Changing disk group configurations
- Comparison of VRAID1 vs. VRAID5
- Changing load balancing algorithms with the full featured MPIO DSM

This white paper is meant to serve as a reference guide for the design and configuration of the EVA8000 storage array for Exchange Server 2003 deployments. The best practices outlined in this document are recommended based on the analysis and observations from this testing. Keep in mind that other factors may impact the design criteria based on specific customer or environmental considerations. Other EVA configurations may yield higher performance under different circumstances than those tested.

HP StorageWorks 8000 Enterprise Virtual Array

The HP StorageWorks 8000 Enterprise Virtual Array is an enterprise class, high performance, high capacity, and high availability "virtual" array storage solution. The EVA8000 is designed to meet the needs of the data center to support critical application demands for consistently high transaction I/O.

This document assumes that the reader has previous knowledge of the EVA family of storage arrays. However, there are several terms that are exclusive to the EVA and referenced throughout the document. These terms are defined below to provide a knowledge baseline. For further information on the HP StorageWorks EVA8000 visit, http://www.hp.com/go/eva8000.

Disk Group (DG) – A grouping of physical disks from which virtual disks are created. The minimum disk group size is 8 disks.

Virtual Disk (vdisk) – A logical disk created within a disk group. When a virtual disk is presented to a host, then the term LUN is used. The document will utilize LUN, virtual disk, and vdisk interchangeably.

Exchange Server 2003 workload overview and storage considerations

In order to fully comprehend the design decisions made in regards to configuring the EVA subsystem, an understanding of the Exchange Server 2003 workload characteristics is important. Exchange Server 2003 is a two-phase commit, transactional workload; Exchange data is first written to a log file and then committed to a database. This leaves two main workloads to characterize: database and transactional logging. Exchange Server 2003 contains two database file types: edb and stm. Data is stored in edb files when accessed by MAPI clients and is the focus of this white paper. An Exchange stm database is utilized by Internet protocol clients such as OWA, POP3 and IMAP.

In normal operations, transactional logging is a 100% sequential write activity, with an average data block transfer size of 4-8 KB. For Exchange database activity, the workload is a mixture of reads and writes. Generally, the breakdown is between 60/40 and 75/25 read/write ratio, and is heavily randomized. The data block sizes for reads and writes are small, in the 4-16 KB range. Exchange database write activity will also exhibit a bursty behavior, resulting in small periods of low activity followed by a burst of hundreds to thousands of I/O operations per second against the storage subsystem.

---

[1] Microsoft recommends the use of Jetstress testing prior to placing any Exchange storage configuration into production.

One of the primary indicators of Exchange storage performance is to measure latencies on the database and log LUNs. The threshold level for acceptable latencies will vary depending upon the requirements of the customer environment. For a properly configured environment, a typical value for read latency on the database drives is sub 20 milliseconds (ms). For write latency, this value should not exceed 10 ms. As the latencies on the drives increase, Exchange performance will start to degrade and client response times increase.

Given the nature of the workload, the storage array cache provides limited benefit for Exchange. Thus, Exchange is often classified as a "spindle dependent" application. In simple terms, this means that Exchange performance is bounded by the amount of I/O that can be achieved from a given number of physical disk spindles in the array. As an administrator, smart choices need to be made on how best to configure the SAN to optimize Exchange performance. This includes dedicating the right number of spindles to support Exchange I/O and capacity requirements, configuring an optimal disk arrangement, isolating Exchange I/O activity from other applications, and other design decisions. With direct attached storage (DAS), configuring storage for Exchange is a simpler process as there is no concern with competing applications and dividing up SAN resources. However, a properly configured SAN will not only provide high performance for the Exchange environment, there will also be incremental benefits in storage management and resource utilization, resulting in a lower cost of ownership.

Performance testing

The goals for the performance testing were to identify the best practices for deploying Microsoft Exchange Server 2003 on the EVA8000 storage array to achieve optimal performance while maintaining high availability and recoverability. The analysis in this white paper focused on three key design decisions for the EVA8000 device: the disk group configuration, VRAID redundancy level, and the Multi-Pathing I/O (MPIO) load balancing algorithm setting.

Testing was conducted using the Microsoft Jetstress utility to drive Exchange database and transaction log I/O load against the storage subsystem and to determine the maximum performance of the various EVA8000 configurations. The performance was measured in terms of I/O per second (IOPS) throughput within a pre-defined threshold for disk latency. As previously stated, Exchange is very sensitive to I/O latency; under- or mis-configured SAN infrastructures can result in an I/O bottleneck within the Exchange infrastructure, causing high disk latencies that will reduce the overall throughput of the server and negatively impact the client experience. As noted above, a generally accepted threshold for Exchange latency is an I/O response time of less then 20 ms for database reads and 10 ms for database and log write operations. This will be the pass/fail criteria used to examine results in the performance section below.

For more information on specific Exchange performance counters discussed in this document see, Appendix B – Performance counters.
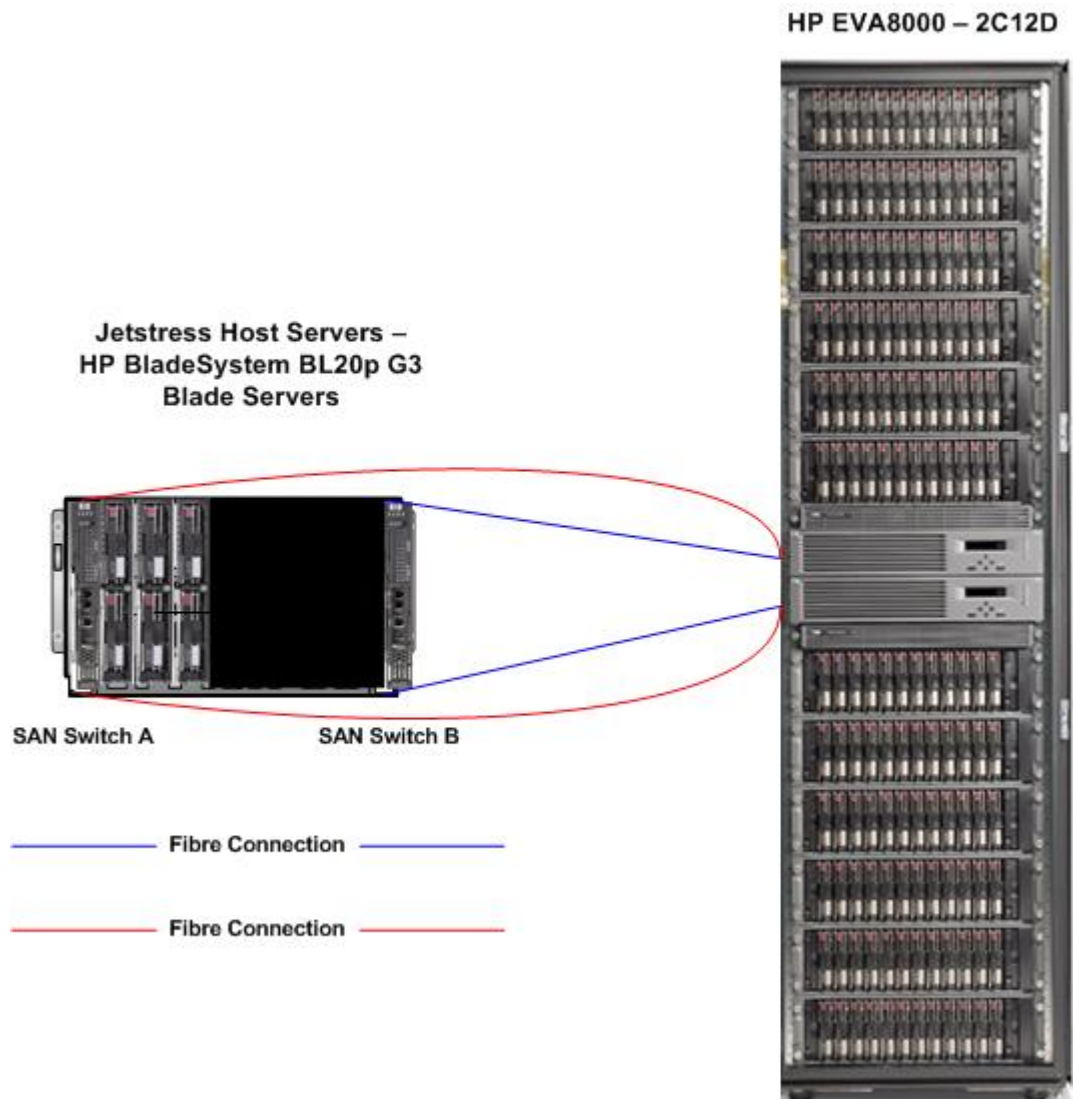
Note: The focus of this testing is to understand the best practices and optimal performance configuration for the EVA8000. Analysis of server performance is not addressed in this white paper. For additional information regarding server, network, and AD performance for Exchange please read the following document, *HP Performance and Configuration Guide for Microsoft Exchange Server 2003* available at
http://www.hp.com/solutions/activeanswers/exchange

Test architecture and configuration

The test environment utilized three HP ProLiant BL20p G3 blade servers to drive the I/O load against the HP StorageWorks EVA8000 (2C12D model) SAN as depicted below in Figure 1. The EVA8000 2C12D model contained a total of 168 72GB, 15K-rpm spindles. To connect the servers to the fabric, the Brocade 4Gb SAN Switch for HP p-Class BladeSystem, an integrated 0U switch which is seated in the back of the BladeSystem enclosure, was utilized. Each host blade server is equipped with dual QLogic host bus adapters (HBAs), one HBA connected to each SAN switch for optimal performance and redundancy.

Figure 1. Architectural diagram

Software and drivers

Each host server in the environment was configured identically with the same set of software and driver versions:

- Microsoft Windows® Server 2003 Enterprise Edition
- Microsoft Exchange Server 2003 Enterprise Edition[2] SP1
- HP MPIO Full Featured DSM for Windows for EVA Disk Arrays v1.00.01
- QLogic HBA Storport Driver – 9.0.2.16
  – Default HBA driver settings utilized in the testing
- Version 5.020 for the EVA Virtual Controller Software (VCS) firmware code.

LUN ownership

When a virtual disk is created, one of the properties that can be configured is LUN ownership. The default value is "No Preference," which allows the EVA controllers to mange LUN ownership. This will provide a balance between the two controllers based on the number of LUNs each controller owns. However, since the EVA is not aware of the type of workload generated on each LUN, no workload balancing will occur.

For Exchange, the best practice is to configure LUN ownership so that an even number of database and log LUNs are balanced between the two controllers. This will provide the highest degree of controller performance scalability and is discussed in more detail in the *Load Balancing* section below. There are 8 LUNs presented to each host: 1 LUN per storage group (4 storage groups) and the accompanying 4 LUNs for each storage group's transaction logs. For each host, the following LUN ownership configuration is utilized:

Controller 1

- SG1
- SG1 Logs
- SG3
- SG3 Logs

Controller 2

- SG2
- SG2 Logs
- SG4
- SG4 Logs

Load balancing

HP MPIO Full Featured DSM for Windows for EVA Disk Arrays v1.00.01 is utilized to provide host load balancing. Unless otherwise indicated, each host is configured to use the default load balancing algorithm, **shortest queue service time** (SQST).

EVA sparing level

One available parameter when creating a disk group is the sparing level. This determines the amount of spare space that is reserved to rebuild a failed drive. There are three sparing levels:

---

[2] A full install of Microsoft Exchange Server is not required for Jetstress. The utility only requires several DLLs from the version of Exchange that is being tested. However, in this environment, a full install of Exchange is present.

- Zero – No space is reserved for the drive rebuild process
- Single – Capacity is reserved to rebuild a single failed drive
- Double – Capacity is reserved to rebuild two failed drives

When configuring EVA sparing with Exchange, a protection level of single is the recommended and is the configuration used in this testing.

Storport vs SCSIport

Testing was conducted using the QLogic HBA Storport Driver, version 9.0.2.16. Both Storport and SCSIport drivers are supported on this configuration. However, the Storport driver is optimized for SAN configurations and can provide improved performance over the SCSIport driver. The default HBA configuration parameters were utilized in the testing. If the SCSIport driver is utilized in an Exchange environment, several registry parameters controlling the HBA queue depth may need to be increased to support the I/O performance demands of a high end Exchange server.

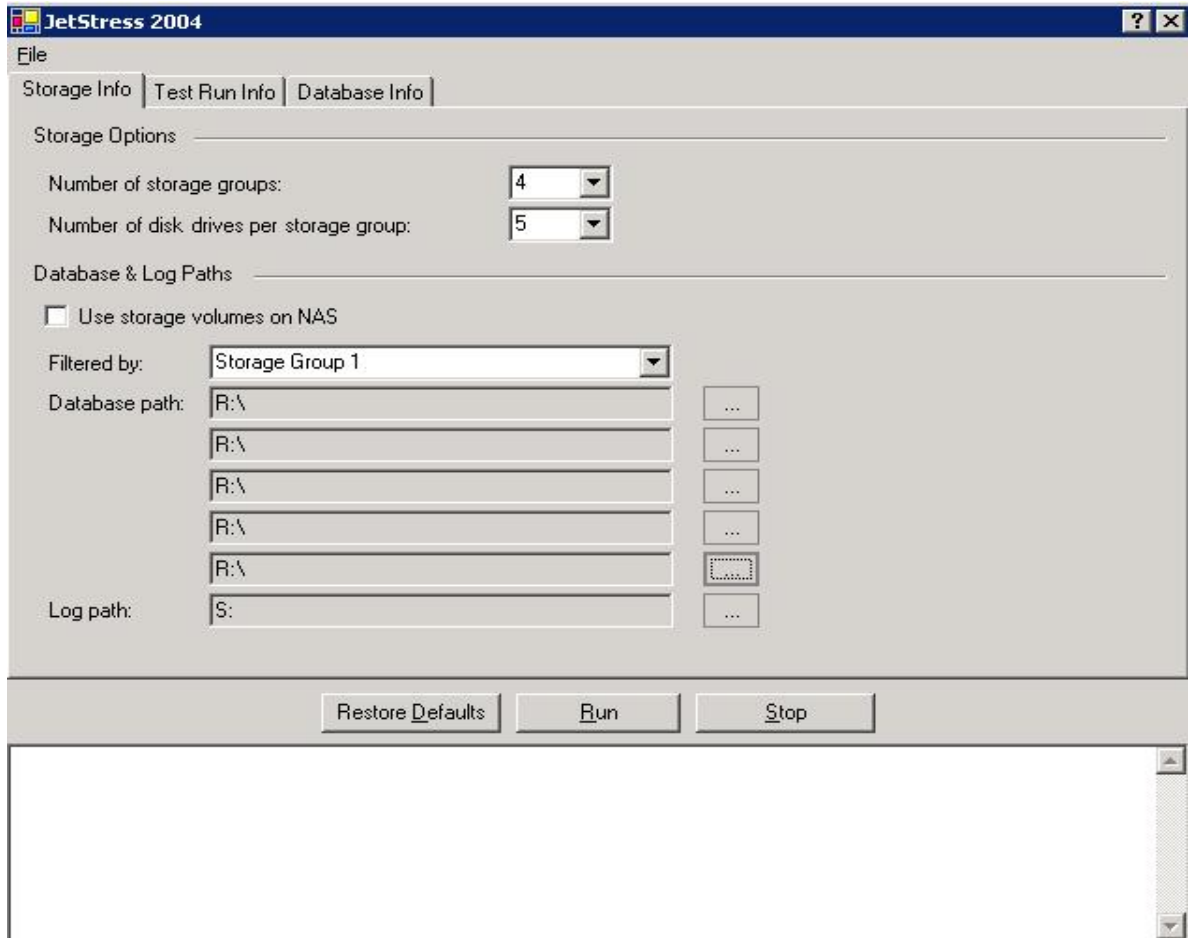### Sector alignment configuration

The EVA8000 no longer requires diskpar or diskpart for sector alignment to increase server performance.

Microsoft Jetstress 2004

The Microsoft Jetstress 2004 utility is used to simulate Exchange disk I/O load with low-level Exchange Server I/O operations including page inserts, deletions, modifications, and seeks in order to verify both performance and stability of the SAN. There are various parameters that can be modified within Jetstress to perform unique data tests against the storage. However, in order to accurately compare the values obtained between multiple tests, the workload for each test must be equivalent. Thus, the testing relied on a standardized workload that was utilized to run against each storage configuration. There were a number of assumptions that were made in the testing that are described with the accompanying Jetstress utility screenshots below.
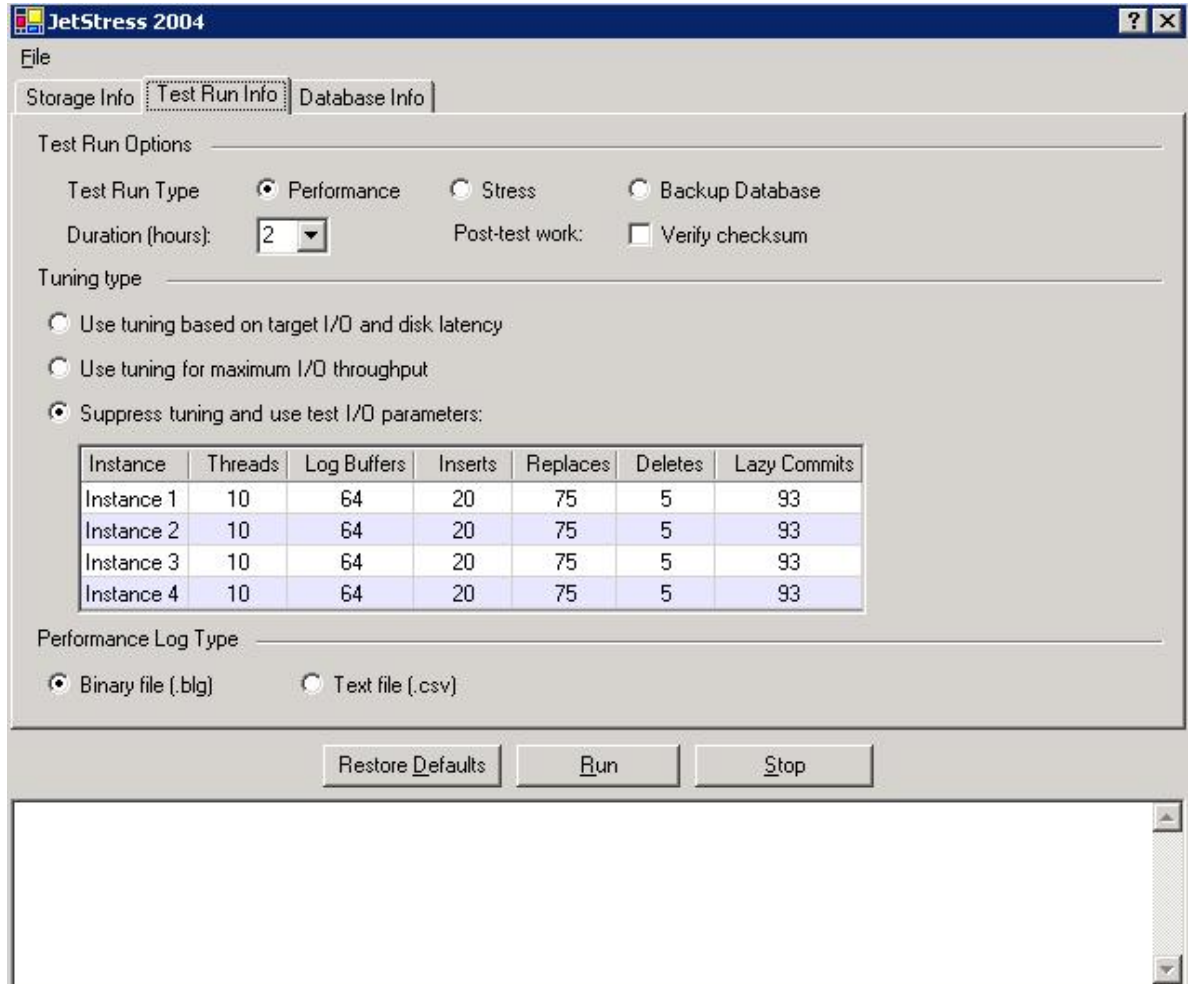
The first screen below depicts the "Storage Info" tab. Each instance of Jetstress simulated four storage groups and five databases per storage group for a total of twenty databases per host, the maximum allowed for Exchange Server 2003 Enterprise Edition. All five databases per Storage Group resided on the same LUN presented by the EVA8000.

Figure 2. Storage Info tab

On the "Test Run Info" screen (Figure 3), a number of parameters are set that determine the duration of the test and the specific workload. Each performance test was executed for a total of two hours. The default operation of the Jetstress utility is to self-tune the workload to meet the required number of IOPS as calculated from the values on the "Database Info" tab. However, this can result in variances between tests that invalidate test-to-test comparisons. Instead, a specific workload is selected to ensure consistency for each trial and the **Suppress tuning and use test I/O parameters** option is selected.

Figure 3. Test Run Info tab

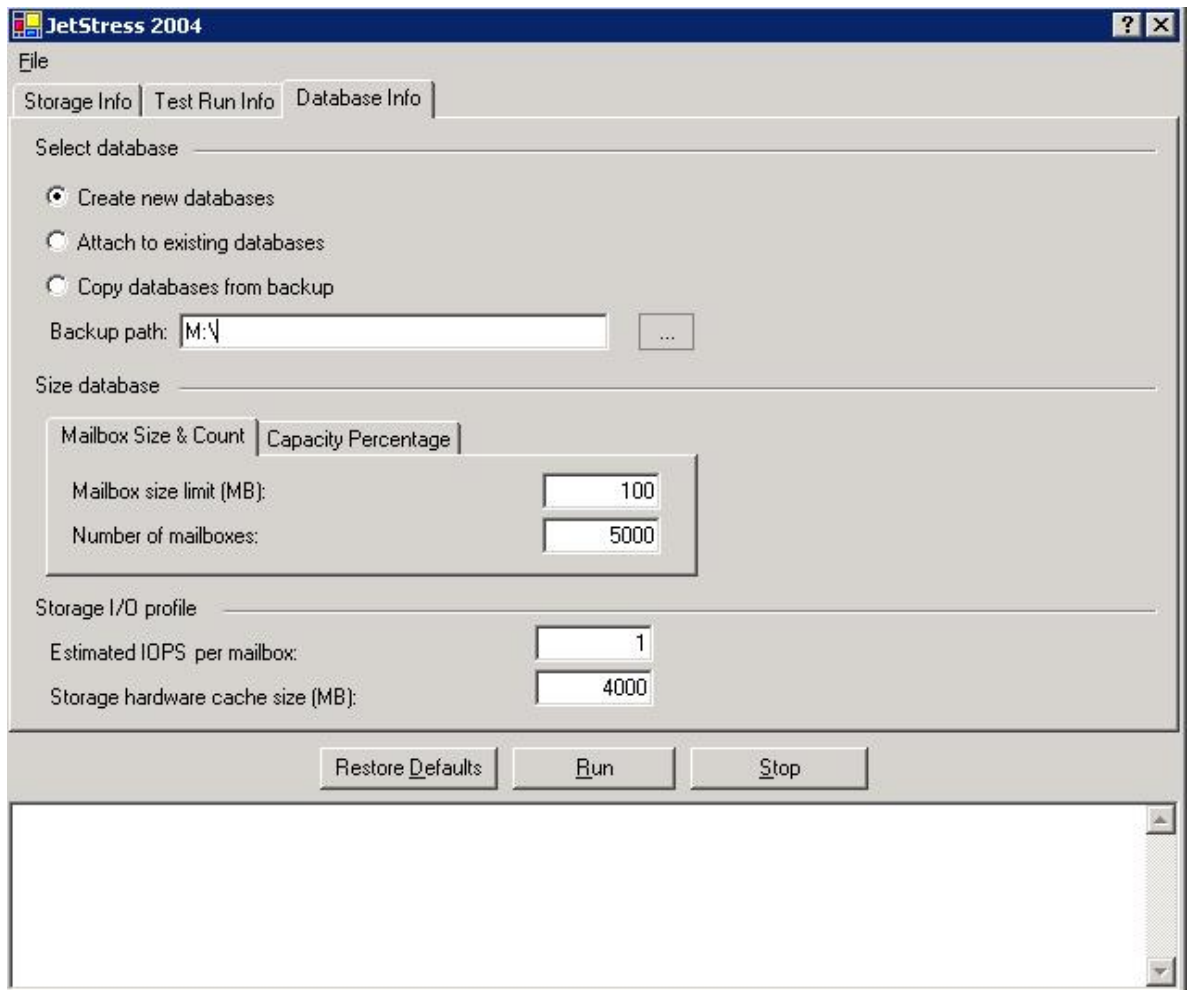As shown in Figure 3, the following tuning parameters were utilized:

- Threads – 10 (Variable)
  - This value will vary and is used to increase and decrease the I/O load from a host
- Log Buffers – 64
- Inserts – 20
- Replaces – 75
- Deletes – 5
  - The values for inserts, replaces, and deletes results in an approximate 65/35 read/write workload ratio that represents a typical corporate workload
- Lazy Commits – 93

This maintains a consistent workload between each trial run, with the only variant being the number of threads that are generating I/O.

On the "Database Info" tab (Figure 4), mailbox size and estimated IOPS per mailbox are collected to determine the type of environment that is being emulated. The mailbox size is necessary to calculate

the required amount of storage capacity for the databases. The estimated IOPS per mailbox value is only utilized if the auto-tune feature of Jetstress is turned on. In this testing, this feature is turned off.

Figure 4. Database Info tab

As with the workload parameters, these values were held constant throughout the testing in order to provide a direct test-to-test comparison. Each host server simulated 5000 users, each user with a 100MB mailbox quota. When the databases are originally created, the **Capacity Percentage** option is used to select the ratio of the size of the Jetstress databases to the full size of the databases in production. For this testing, a 75% ratio was chosen. This means that the Jetstress databases are 75% of their full capacity size. The goal is to ensure that the size of the databases is sufficiently larger than the storage hardware cache size.

In the case of the cache size, the total size of the databases must be large enough to ensure that the EVA cache isn't providing an unfair performance advantage in the testing. Exchange database I/O is a very random workload, and when a read request is made, there is a very small percentage of the I/O that is satisfied from cache. If the capacity percentage value is set too low, then performance will be artificially inflated due to increased read requests being satisfied from cache instead of disk. The EVA8000 has 4GB of read/write cache available, 2GB per controller. With 5000 users at 100MB per user, the total size of all the databases would be 500GB at 100%. At a 75% capacity percentage, this number drops to 375GB; however, this is still much larger than the size of the cache.

A general rule-of-thumb is the total size of the databases should be at least 20 times the size of the storage subsystem cache when utilizing Jetstress for performance analysis.

Note: In a production environment, the size of the user's mailbox can impact the number of IOPS/u. Previous testing has shown that an increase in the size of the mailbox can cause an increase in the IOPS/u value and impact performance. This is due to a potential increase in the number of items per mailbox with the increase of the size of the mailbox. However, since Jetstress is a low-level Jet database stress tool, there is no concept of mailboxes, folder, or items. Increasing the size of the databases, by increasing the size of the user's mailbox, will not impact performance. For this performance testing, the size of a user's mailbox is held constant and is not a variable of the test.

Performance results

The analysis of the performance results is broken down into different sections that focus on key design decisions for deploying Exchange Server 2003 on the EVA8000. The first section focuses on a comparison between disk group configurations (see the "*Comparing disk group configurations*" section). The second section highlights the performance differences between VRAID1 and VRAID5 virtual disks (see the "*Comparing VRAID1 and VRAID5*" section). The third section discusses load balancing configurations using the HP full featured MPIO load balancing software (see "*Load balancing*" section). Each of these configuration choices will play a significant role in providing an optimized and high performance Exchange environment and must be properly evaluated in order to avoid potential post-deployment performance issues.

The testing documented in this white paper varies between single- and multi-server tests and is specifically defined in each section. Each server is simulating four Exchange storage groups with five databases per storage group. The five databases per storage group will reside on the same virtual disk.

Note: To simulate four storage groups, Jetstress executes four instances, one per SG LUN. The number of threads determines the amount of I/O pushed against the array and varies in each test iteration. In the data results below, when referencing the number of threads, the value indicated is per SG instance; to calculate the total number of threads per host, multiply the number by 4.

Comparing disk group configurations

When an EVA is first initialized, the storage administrator is faced with the task of creating a disk group ranging from a minimum of 8 spindles to a maximum of the number of spindles in the EVA. A disk group serves as an isolation boundary for both performance and fault tolerance. The disk activity in one disk group or a failure of a disk group will be isolated from other disk groups in the EVA.

When a virtual disk is created, the vdisk will be allocated slices of storage from each spindle in the disk group. This can provide a performance benefit for applications (like Exchange) that are very I/O intensive by creating larger disk groups and leveraging more spindles to service the I/O requests to the vdisks. Exchange Server 2003 is a transaction workload with two main workloads for normal operation: writing to transaction logs, and database read/write activity. Thus, it is important to consider how to carve out the EVA to maximize the performance benefits of larger disk groups while balancing the demands for workload isolation of log and database activity.

The disk group design performance testing utilized a raid level of VRAID1 for the virtual disks and is predicated on the following recommendation:

Best Practice: For the best combination of performance and availability, Exchange transaction logs and databases should be isolated into separate disk groups on the EVA.

This recommendation is based upon the best practice of transaction log isolation from the accompanying data source, as well as the performance impact on a production Exchange environment from a mixed workload environment. An Exchange Server 2003 write operation is first written to the transaction log and then written to the database as part of the two-stage commit process. When a full backup is taken, the transaction logs are deleted, and a full copy of the Exchange database is created. During a restore operation, the backup sets can be utilized to restore Exchange to the point in time when the backup was taken; however, to restore to the moment of failure, the Exchange logs are required to replay the transactions to the restored database. If the EVA is configured such that a disk group contains both database and transaction logs from the same server, then, in the extremely unlikely event that a disk group is lost, there is no way to restore the data from a point in time later than the last backup.

Performance to the transaction log LUNs is also improved in an isolated configuration. A single transaction log I/O stream is a 100% sequential write workload. However, because there are multiple storage groups, there are several log I/O streams which randomize the workload on the spindles and negate some of the benefits that would be observed with a purely sequential data stream. But, by eliminating the aggressive random read data activity associated with the database workload, performance to the log drives improves dramatically. This will also improve disk availability as well. Due to the nature of Exchange database I/O, there is substantial disk head movement incurred to support the aggressive random read workload. Anecdotal evidence has demonstrated that for the same disk type, spindles isolated solely to Exchange transactional logging may have a longer lifespan than spindles handling both log and database activity. Isolating the two workloads can provide a higher level of protection for the log data.

Since Exchange is a very sensitive application to I/O latency, performance can be significantly impacted when other workloads are running on the same disk group as production Exchange databases. This includes any additional applications supported by the EVA, data management activities such as large mailbox moves or copy operations, and backup considerations.

Best Practice: If the EVA is dedicated to additional applications, such as SQL for example, then a separate disk group should be configured to host the non-Exchange application workload. If the requirement for additional application I/O and capacity is not significant enough to warrant a separate disk group, then proper planning must be made to ensure that the disk group supporting Exchange production databases has the resources to support the non-Exchange application requirements. In the case of other transaction workloads, log LUNs can be placed in the same disk group as Exchange transaction logs.

Disk group comparison design

To achieve higher Exchange performance from the EVA subsystem while maintaining optimal data reliability, the best approach is to increase the number of spindles isolated for the database workload. The following testing analyzes both the performance impact when splitting the database disk group into multiple smaller disk groups and the scalability of the isolated configuration when increasing the size of the database disk group. The disk group configuration decisions were based upon best practice recommendations for designing Exchange on the EVA.

The layout of each configuration is the following:

Primary Isolated configuration:

- 88 disks – DBs
- 16 disks – Logs
- 64 disks – "Backup"

Secondary Isolated configuration:

- 44 disks – Host 1 DBs
- 44 disks – Host 2 DBs
- 16 disks – Logs
- 64 disks – "Backup"

Other variations of the primary configuration will contain a different number of spindles as indicated in the results section. For each configuration tested in this white paper, a 16-disk log disk group is utilized. Both the primary and secondary isolated configurations also contained an additional disk group, defined as "backup," containing the remaining 64 spindles of the EVA. There was no concurrent testing performed on this disk group as part of this performance and characterization guide. However, this disk group was isolated to allow for data backup considerations such as Exchange Server 2003 disk-to-disk backup operations, snapshot and snapclones, or other data requirements. The existence of this disk group is not a requirement for the EVA in production environments, and the size of the backup disk group was selected based upon deployment observations and can/will vary. However, if backup-to-disk operations are being implemented, a separate disk group to isolate the workload from production I/O is considered a best practice.

Best practice: When deploying Exchange Server 2003 on an EVA, analysis of the backup and restore operations necessary to meet SLA requirements should be performed during the initial design phase. If spindles are required on the EVA for disk-disk backups, Fast Recovery Solution implementation (FRS), Continuous Access replication targets, or other non-Exchange I/O, a separate disk group(s) should be created to isolate this workload.
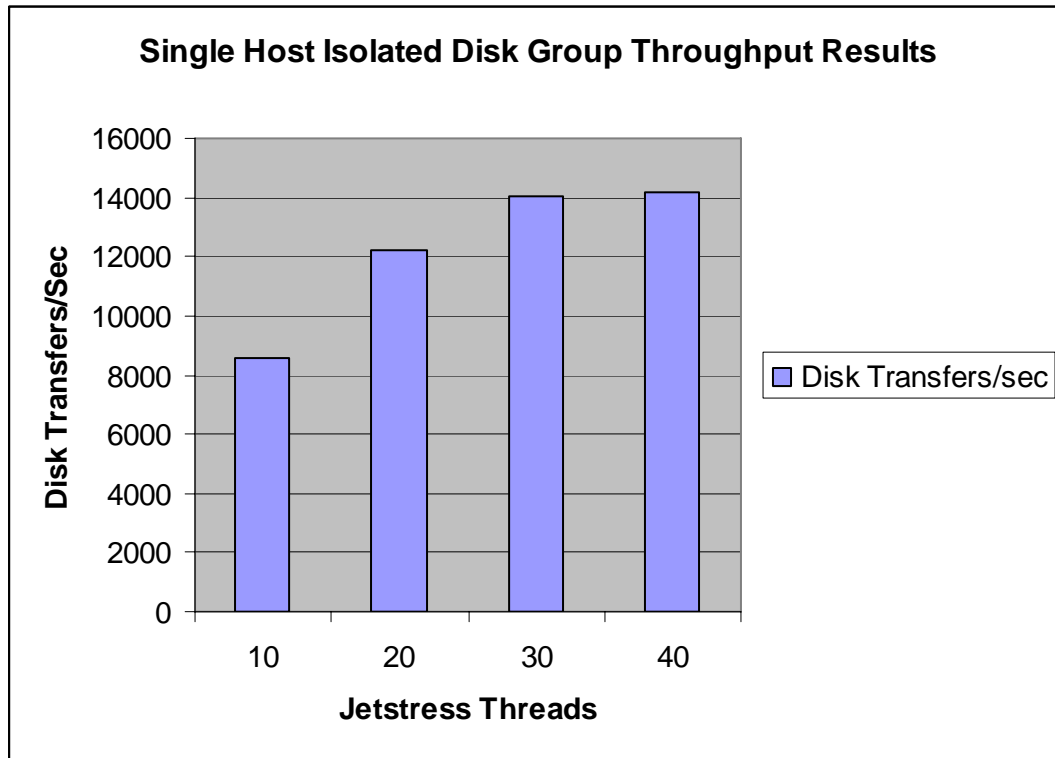
For more information regarding the configuration and design of FRS please read the "HP StorageWorks Fast Recovery Solution for Microsoft Windows 2003, Exchange Server 2003 and HP StorageWorks EVA8000" available by clicking on "More technical documents" at www.hp.com/solutions/exchange.

Disk group comparison results

The data outlined in Figures 5 and 6 below show the performance results of testing on the primary isolated configuration. This testing was conducted using a single Jetstress test server to drive load against the storage array.
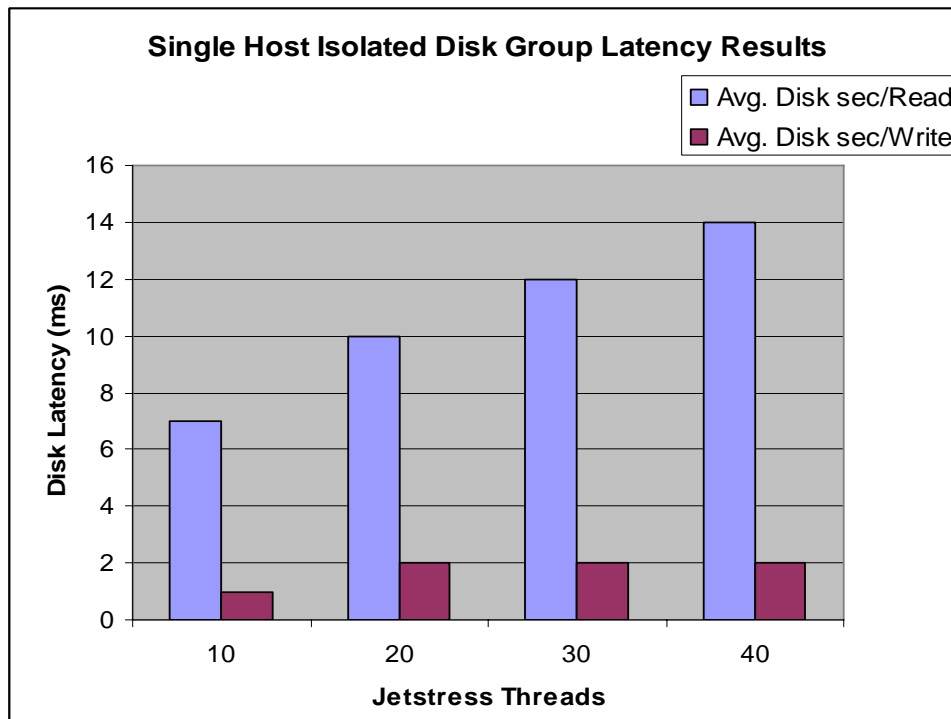
Figure 5 below highlights the cumulative database disk transfers/sec throughput, as the number of Jetstress threads is incremented.

Figure 5. Cumulative database disk transfers/sec throughput

As the I/O load against the array increases, the throughput increases to slightly over 14,000 IOPS. Figure 6 below depicts the average database LUN read and write latencies for the 4 test points. As is depicted in the graph, there is a steady increase in latency as the load on the storage subsystem is increased. However, both the read and write latency levels remain below the acceptable criteria for disk response time.

Figure 6. Average database LUN read and write latencies

Single Host Isolated Disk Group Latency Results

- ☐ Avg. Disk sec/Read
- ■ Avg. Disk sec/Write

Disk Latency (ms)

Jetstress Threads

The average read latency tops out at 14 ms, below the defined threshold level of 20 ms for Exchange server performance. The database write latencies for each of the test points were less than or equal to 2 ms, indicating that there was negligible cache saturation impacting the performance. If the demand on the EVA increased to a level at which the cache could not flush data to disk at the same rate that write requests were being received from the host, the cache would fill up until becoming saturated, and I/O write requests would queue up at the host level and incur additional latency.

These test results provide a good indication of the baseline performance of the array for the primary isolated disk group configuration. However, generally, production Exchange servers will not push 14,000 IOPS per server. As a broad guideline, the IOPS per server will range between 4000-7000 IOPS for a high performance, enterprise Exchange server, heavily depending on the specific user profile. Additional testing was conducted utilizing multiple host Jetstress servers to analyze the impact of increased I/O activity on the array.

As one might expect, there is overhead associated with adding an additional host. As the number of hosts, or I/O streams, increases, the throughput of each spindle will decrease slightly. Table 1 below highlights the performance of a single-host test vs a dual-host test for the primary isolated configuration. The single host executed 30 Jetstress threads against the EVA, while each server in the dual-host configuration pushed 15 threads. Both servers were configured identically, so the actual number of Exchange databases supported in the dual-host configuration doubled, while the load being generated by each Jetstress host was cut in half. This is analogous to comparing one 5000-user Exchange server with two 2500-user Exchange servers.
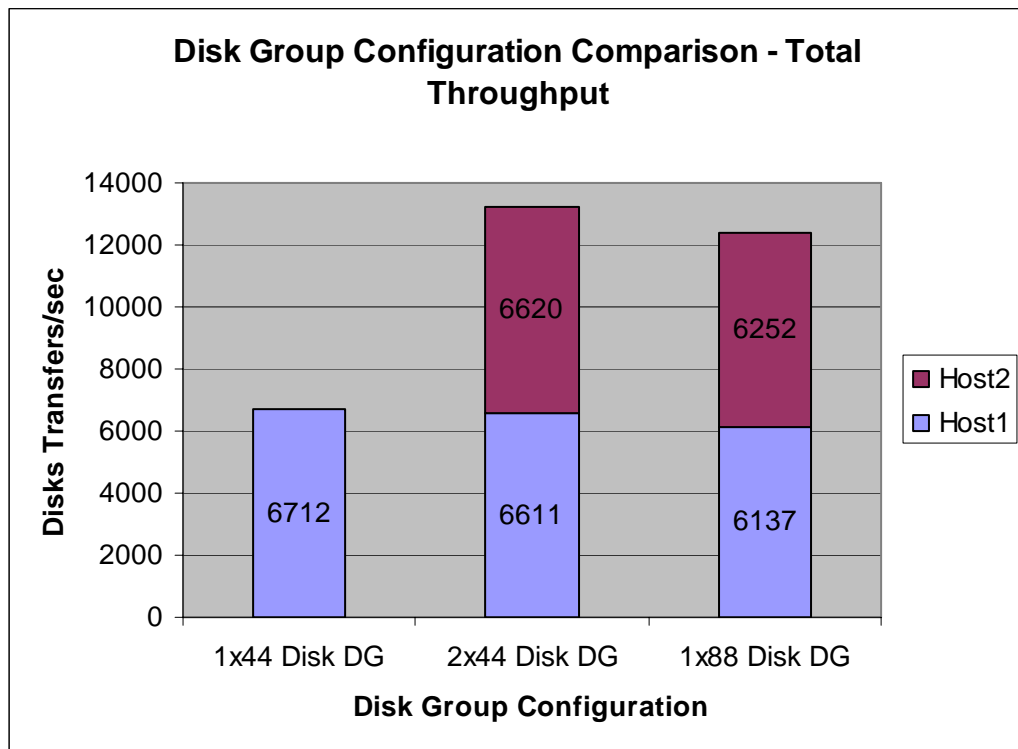
Table 1. Single host vs. Dual host data comparison

Single host		Dual host	
Disk Transfers/sec	14021	Disk Transfers/sec	12458
Avg. Disk Sec/Read	12.2	Avg. Disk Sec/Read	15.1
Avg. Disk Sec/Transfer	8.5	Avg. Disk Sec/Transfer	10
Avg. Disk Sec/Write	2	Avg. Disk Sec/Write	2

The table indicates that the throughput performance is 12% higher for the single-host configuration versus the dual-host, with lower latencies.

So, how will this impact the design decisions for deploying Exchange on the EVA? The graph below (Figure 7) shows a comparison of several configurations for the database disk group(s), with an isolated 16-disk log disk group.

Figure 7. Comparison of several configurations for the database disk group

The first configuration in the graph is a single host running 15 Jetstress threads against a 44-spindle variant of the primary isolated disk group. The second configuration is the secondary isolated configuration with two 44-spindle disk groups, each dedicated to a separate host, and each host running 15 Jetstress threads. The third configuration is the standard primary isolated configuration from above, with both hosts again running 15 threads.

Each test point had comparable latencies:

- 14-15 ms avg. disk sec/reads
- 10 ms avg. disk sec/transfers
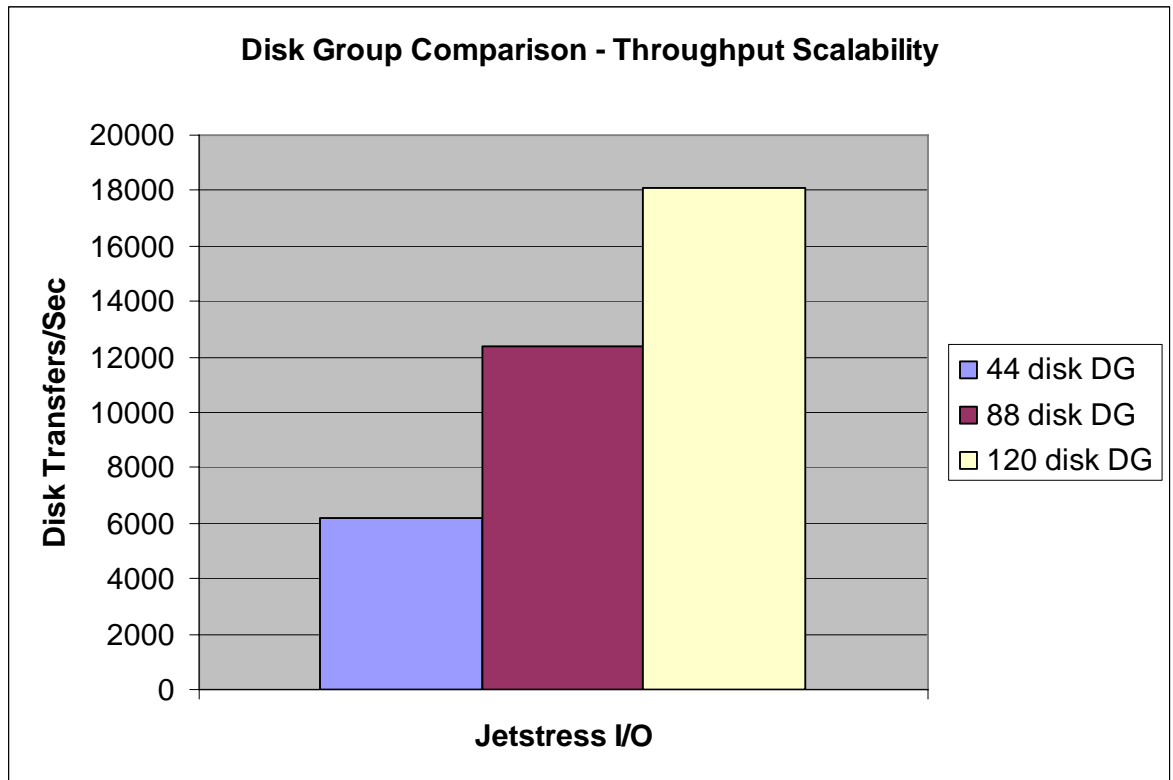- 2 ms avg. disk sec/writes

The throughput from the single host (1x44 disk) configuration dropped only slightly (1.5%) when adding a second isolated 44-spindle disk group to host Exchange databases from a second server. This indicates that the EVA8000 controller shows negligible performance degradation when adding a second host to the subsystem. However, there is a more significant throughput drop of 7 percent, when comparing the secondary isolated vs. the primary isolated configuration. Generally, the larger the disk group, the greater the aggregate I/O; i.e., the more spindles, the more I/O available to each virtual disk. However, keep in mind that the primary isolated configuration, while having double the number of spindles for I/O, also has double the number of I/O streams than either disk group in the secondary isolated configuration. Thus, there is an apparent trade-off between the benefits of virtualization in a larger disk group and the performance degradation with multiple hosts and an increased number of virtual disks. This document delves further into this observation later in this section.

The EVA family of controllers has improved from the first revision of the EVA5000 to the EVA8000. There are faster processors, faster PCI buses, more cache memory, and higher throughput on the mirror port to name a few hardware architecture upgrades. However, one of the major changes that positively impacts Exchange is the further optimization of code to handle the type of workload Exchange generates. The EVA8000 has better performance when utilizing large disk groups in high transaction environments than the previous EVA models. The sweet spot for the maximum disk group for Exchange on the EVA5000 tended to be around 88 spindles for a corporate enterprise workload with a 15-20 ms read latency threshold. The maximum number of spindles supported in a single disk group is primarily a function of the required latency thresholds and the amount of work being performed. With the EVA8000 code optimization, disk group sizes can be increased, and performance of the storage subsystem is controlled by the addition and subtraction of spindles. Figure 8 below, shows the throughput results to the database disk group for several variations (44, 88, and 120 disks) of the primary isolated configuration. For each configuration, a 16-disk transaction log disk group was configured, and the remaining disks were left idle. This testing was conducted using several host configurations in order to emulate a production Exchange server environment. Each host generated approximately 6000 IOPS with a breakdown as follows:

- 44 disk primary configuration – one host
- 88 disk primary configuration – two hosts
- 120 disk primary configuration – three hosts

Each data point represents the maximum throughput achieved while maintaining LUN response times between 14-16 ms for reads and 2-3 ms for writes. Higher throughput can be achieved if higher latency thresholds are tolerated.

Figure 8. Throughput scalability

The graph indicates the linear scalability of the EVA as the spindle count in the database disk group increases. Moving from the 44- to 88-spindle DG, there is an increase from 6175 to 12389 disk transfers/sec, respectively, or a jump of nearly exactly two times the throughput. Interpolating the disk group scalability with an additional 44 spindles, to a 132-spindle disk group yields a value of 18525 disk transfers/sec. Since the 120-spindle disk group is 10% smaller than a 132 spindle disk group, the theoretical number of disk transfers/sec with linear scalability would be roughly 16700. However, the observed data shows a much higher throughput rate at 18089 disk transfers/sec. This reinforces the notion that the controller is not a bottlenecking factor, and that the larger disk group size is getting additional scalability benefits from the increased number of spindles dedicated to the I/O.

Looking deeper into the IOPS per spindle values, the 44- and 88-spindle disk group have values of 140 host IOPS per spindle. The 120-spindle disk group increases to 150 host IOPS per spindle. By spreading the I/O across a larger number of disks, each spindle in the array is able to handle the additional workload while maintaining acceptable latency values.

Keep in mind that there is a distinction made between host IOPS per spindle as measured at the OS/Application layer and the array IOPS per spindle, which is measured at the storage array level. This is an important distinction to understand when looking at Exchange sizing examples and is explained in greater detail in, Appendix A – IOPS and sizing considerations.

Disk Group data results analysis

On one hand, the data in Figure 8 above indicates favorable scalability with a single database disk group when increasing the size of the database disk group to 120 spindles. With the additional spindles, the overhead of multiple host I/O streams in that testing is negated. On the other hand, the

data in Figure 7 indicates that higher performance can be achieved with the secondary isolated configuration by separating database disk groups and isolating multiple hosts with two 44-spindle disk groups versus a single 88-spindle DG for databases. So how does one interpret the data?

Unlike the recommendations for isolating workloads (databases, logs, backups, etc) into separate disk groups, there is no clear-cut best practice recommendation for the size or number of database disk groups. However, there are a number of performance and reliability factors that may play a role in determining what the best fit is for a particular customer environment. A single large disk group provides the benefits of a large number of spindles for all I/O activity and can absorb the impact of multiple servers more efficiently than smaller disk groups. Also, with the bursty behavior of Exchange, a larger disk group has more spindles to handle and absorb random bursts of hundreds to thousands of I/Os and can potentially support higher peak I/O rates per server. The theory is that all Exchange servers won't be operating at a 100% peak I/O rate all the time; this is especially true in a consolidated infrastructure with Exchange servers hosting mailboxes from multiple time zones. There is also a general non-application-specific EVA best practice to minimize the number of disk groups, again so as to maximize the performance benefits of virtualization.

However, the data has shown that splitting a larger database disk group into smaller disk groups provides some of the performance benefit of fewer I/O streams against each of the physical disks. There is also the benefit of domain isolation between multiple Exchange servers in the environment. In the unlikely event of a disk group failure or during a backup/restore operation, there is a minimized impact on other production servers outside that disk group. In situations where there exists a number of small Exchange servers in terms of required disk I/O (not necessarily in terms of the number of users), the administrator should avoid creating small 8-16 spindle disk groups for each server's databases. This not only maximizes the impact of spare space (which is configured on a disk group level) but also minimizes any potential virtualization benefits from the EVA. In this case, the best practice is to combine the servers (all or a subset of) into larger disk group(s).

Best Practice: As long as the number of spindles are properly sized and configured to support Exchange database activity (providing a recommended 15-20% overhead to handle peak I/O), either a single large database disk group or multiple smaller database disk groups can be configured to support the Exchange environment.

Comparing VRAID1 and VRAID5

The previous section of this document took an in-depth look at several configuration design decisions focused on disk group layout for the EVA. Once a disk group is configured, virtual disks are created within that disk group. When creating virtual disks, the administrator is tasked with the decision of whether to utilize VRAID1 or VRAID5 to provide data redundancy. With any implementation of a VRAID level, 1 or 5, there is a subsequent performance penalty to implement the redundancy, specifically in write requests to the disk. For VRAID1, two writes are required for each write request from the Exchange server, to write out the data block and to write the corresponding mirrored block for redundancy. With VRAID5, there is a greater performance penalty of up to four additional disk transactions to complete a single write request. For each write operation issued from the application, the following sequence occurs:

- Read the original data and parity block (two requests)
- Calculate the new parity block
- Write the new data and parity block (two requests)

The EVA8000 controllers utilize advanced cache optimization and write-gathering techniques to minimize the performance penalty of VRAID5 write activity. With write-gathering, multiple write operations are grouped together to minimize the performance penalty of the parity update that would occur for each individual write request to disk. In the best case scenario, with multiple writes grouped together from a data stripe, the parity block need only be updated once.

> **Note:** Virtual disks on the EVA can also be configured as VRAID0 LUNs, which would provide the highest Exchange performance. However, because there is no redundancy in the event of a disk failure, this configuration is not recommended for any production deployments and is not tested as part of this white paper.

Disk group design

For the VRAID1 versus VRAID5 comparison tests, the EVA was configured in the primary isolated configuration with 88 spindles for the database disk group and 16 spindles for the log disk group. The analysis looks at the performance differences between changing the VRAID setting on the database virtual disks. The transaction log LUNs are configured as VRAID1 only. Since transaction logging is a 100% write activity, the VRAID5 penalty would yield a significant performance overhead on the log disks.
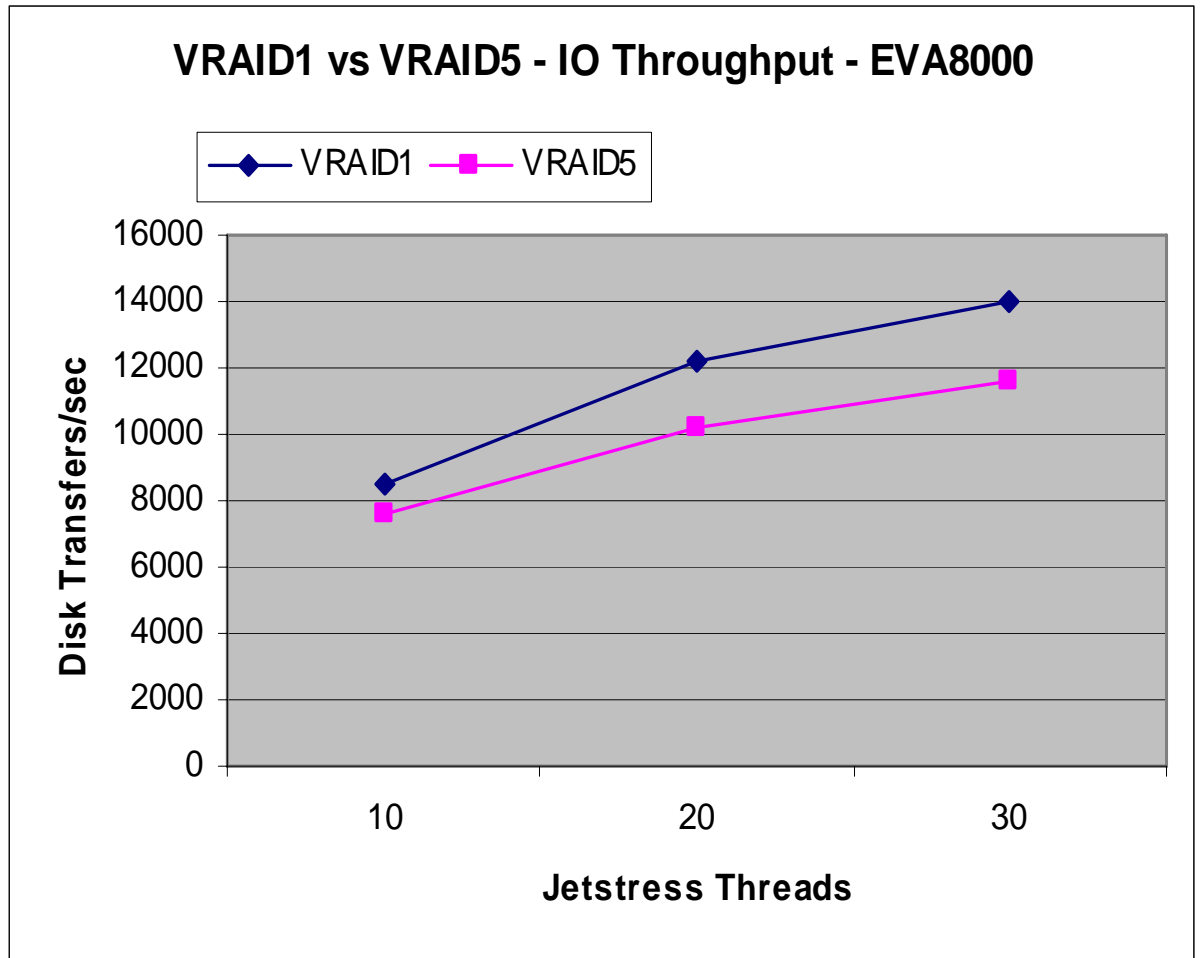
Best Practice: For Exchange transaction logs, VRAID1 should be used to provide the best performance and availability.

The tests were conducted using a single host to drive the Jetstress load against the array.

VRAID comparison results

Figure 9 illustrates the throughput in terms of disk transfers/sec to the EVA8000 for both the VRAID1 and VRAID5 configuration.
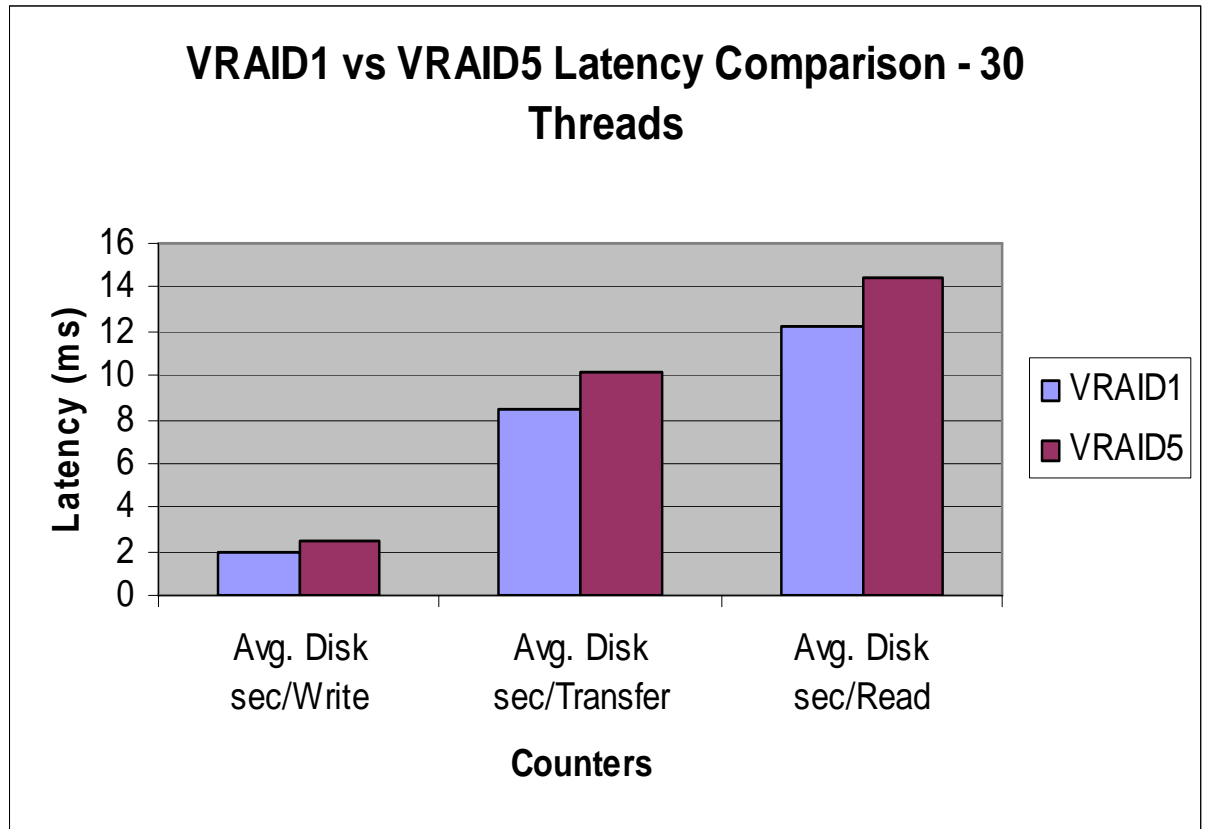
Figure 9. I/O Throughput

The graph shows that at 30 threads, the VRAID1 configuration outperforms the VRAID5 configuration by a significant margin. The VRAID1 databases can support 14021 IOPS while the VRAID5 databases only scale to 11633 IOPS, an 18% difference in performance.

Correspondingly, there are higher read latencies associated with the lower throughput of the VRAID5 configuration. Figure 10 shows the latency results for the 30 threads Jetstress test. The VRAID1 configuration has an average disk sec/read latency of 12 ms, while the VRAID5 configuration has a latency of 14 ms.
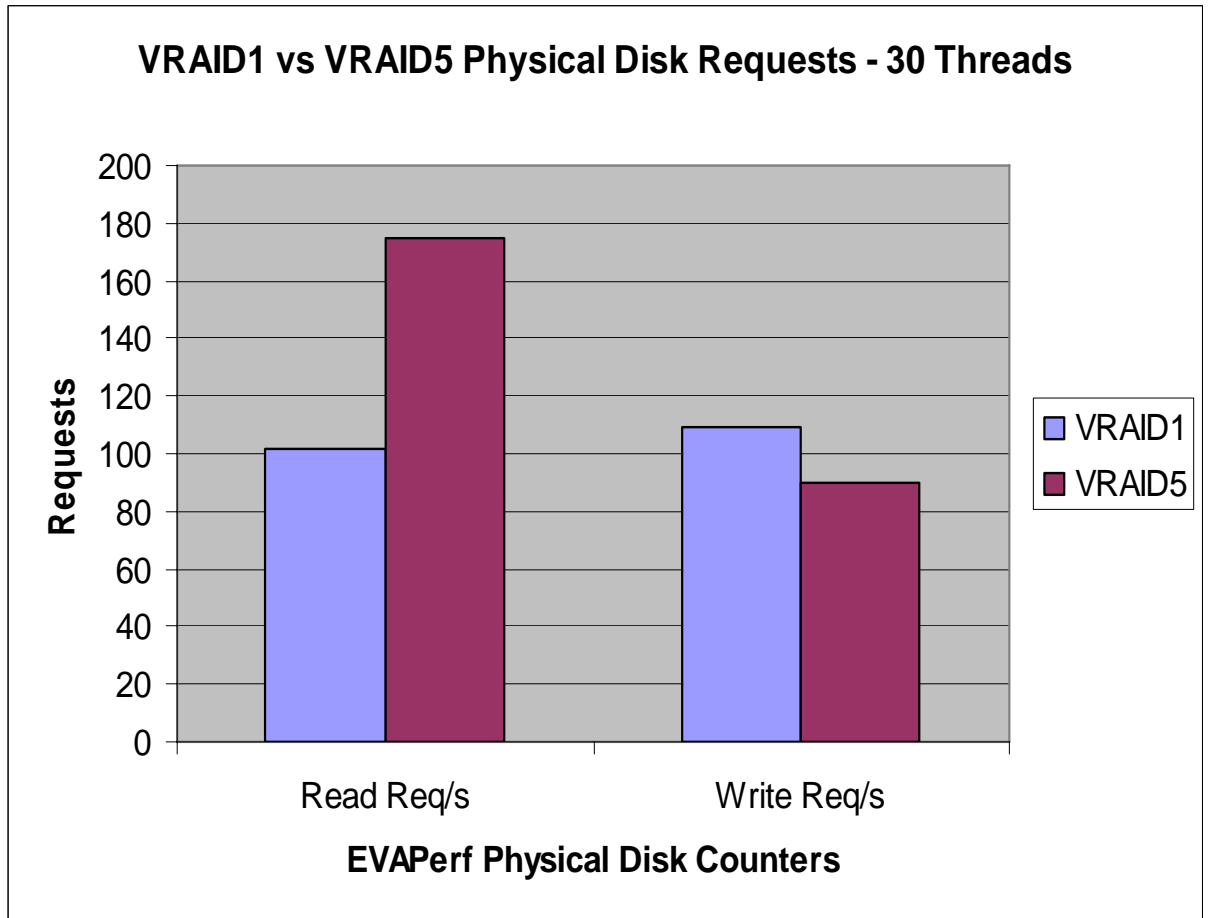
Figure 10. Latency results for the 30 threads Jetstress test

VRAID1 vs VRAID5 Latency Comparison - 30 Threads

While the VRAID5 latency is within the required threshold for acceptable Exchange performance, this again highlights the impact on spindle performance when using VRAID5 versus VRAID1.

The next figure highlights additional performance statistics captured by the HP EVAPerf utility, a tool for monitoring performance of the EVA subsystem via the Windows Perfmon utility or the command line. For additional information on the EVAPerf utility, download the *Performance analysis of the HP StorageWorks Enterprise Virtual Array storage systems using HP StorageWorks Command View* guide from www.hp.com. Looking at Figure 11, the EVAPerf Physical Disk Requests counter shows the average number of requests across the physical spindles in the database disk group for both the VRAID1 and VRAID5 vdisks.

Figure 11. Additional performance statistics

Notice that the average number of read requests is significantly higher on the physical spindles for VRAID5. Remember that to implement a single write request for VRAID5 requires two read and two write operations. By looking below the host level and at the array level, the EVAPerf counter is capturing this additional overhead. While VRAID1 spindles are processing on average about 100 read requests, the VRAID5 spindles are processing closer to 180. This increased overhead on the disks causes the latencies and disk queues to increase and impedes the ability to process more host application requests.

Turning the attention back to Figure 9, there is an increase in the performance delta between VRAID1 and VRAID5 as the load increases. When the load on the configuration is low, i.e., 10 threads, less activity is occurring on the spindles, and there is more I/O available to absorb the increased overhead imposed with VRAID5; the performance difference is 12%. As the load increases, this grows to 18% as the disks are busy responding to additional host requests. This reinforces the notion that to support VRAID5 disks, more spindles are required to support the increased I/O requirements.

The data shows that, from a performance standpoint, there is a clear advantage to using VRAID1 over VRAID5. Unless the environment consists of large mailboxes that are infrequently accessed, a general rule-of-thumb is that Exchange will require more spindles to support the I/O requirements of VRAID5 than VRAID1.

Outside of performance, there are other considerations that favor selecting VRAID1 over VRAID5. VRAID5 provides higher capacity utilization of the drives because there is less disk space consumed for redundancy support as compared to a mirrored VRAID1 configuration. However, due to the nature of Exchange, in most cases, sizing for I/O performance will require more spindles than sizing for capacity. When sizing for I/O, there will typically be a significant amount of free disk capacity that remains unallocated due to the I/O requirements, effectively negating the benefits of VRAID5 capacity gains. VRAID1 also provides higher availability than VRAID5. The EVA controllers implement a technology called Redundancy Storage Sets (RSS) to increase the number of simultaneous spindle failures that can be handled within a disk group prior to losing data. RSS sets are internal groupings of spindles from 6 to 11 disks that form failure domain barriers, which isolate the impact of a disk failure to the RSS set it belongs to. The RSS configuration is handled internally by the controller and is transparent to the administrator. If a disk failure occurs, the EVA controllers will begin to rebuild the data for a failed disk on spare disk space immediately. Once the data has been rebuilt, a second failure can occur in an RSS without the risk of a loss of the group. However, problems can arise in the very unlikely event of simultaneous failures or failures occurring before the first drive can be rebuilt, especially with a VRAID5 configuration. The other disadvantage of using VRAID5 is that to recover a failed disk requires reading the chunk of parity data from each disk in the RSS set. This will impact more disks that will be servicing production data than recovery of a VRAID1 disk failure which will be rebuilt from the mirrored drive.

Best Practice: For Exchange databases, VRAID1 should be used to provide the best performance and availability.

One area where VRAID5 virtual disks may be considered for an Exchange environment is for hosting disk-to-disk backup operations. The sequential nature of the workload when performing the streaming backup allows the advanced controller algorithms of the EVA to offset the performance impact of VRAID5. Testing has shown that comparable performance can be achieved writing to either a VRAID5 or VRAID1 LUN. In this case, additional capacity is also beneficial, and VRAID5 will provide more efficient storage utilization and increase the amount of data that can be stored.

Best Practice: When performing disk-to-disk backups on the EVA, VRAID5 virtual disks should be configured for the backup target.

Load balancing

To achieve high availability for an Exchange server connected to a SAN, multiple HBAs must be configured to maintain system uptime in the event of an HBA failure. This requires host software to support load balancing and failover on multiple data paths for drives presented from the EVA. With the EVA3000/5000, multi-pathing on the host was achieved using the HP SecurePath software. With the EVA4000/6000/8000 family, SecurePath is no longer supported. To implement load balancing, on the EVA4000/6000/8000 family, the HP MPIO Full Featured DSM for Windows for EVA Disk Arrays is utilized. The HP EVA-specific multi-pathing I/O driver (commonly referred to as MPIO) is based on Microsoft's MPIO architecture.

The MPIO driver contains a number of different load balancing algorithms that can be implemented by the administrator. These include the following:

- SQST - Shortest Queue Service Time (default)
- SQB – Shortest Queue Bytes
- RR - Round Robin
- SQR -Shortest Queue Requests
- No Load Balancing (requires manual setting)

Another significant change from the older EVA3000/5000 series is that the newer EVA controllers operate in true Active/Active mode. This means that a host can issue a request on a virtual disk via

either controller. In the EVA3000/5000 series, both controllers are active concurrently; however, a specific LUN is owned and accessed from a single controller. Only in the event of a failure would the second controller receive LUN ownership.

Testing was conducted to understand the performance differences between each of the MPIO algorithms on the Active/Active EVA8000 controllers. The data was generated using a single host to drive Jetstress load (30 threads) against the array configured in the primary isolated configuration.

Figures 12 and 13 depict the throughput and latency results of the tests on the four load balancing algorithms (the no load balancing setting is discussed later in the section). The first graph indicates that there are no discernible differences between the numbers of disk read and write operations being processed for any of the four tests. In fact, there is less than 1% difference in the number of total disk transfers. The second graph highlights the latency and again, there is no difference in the latency results between the tests

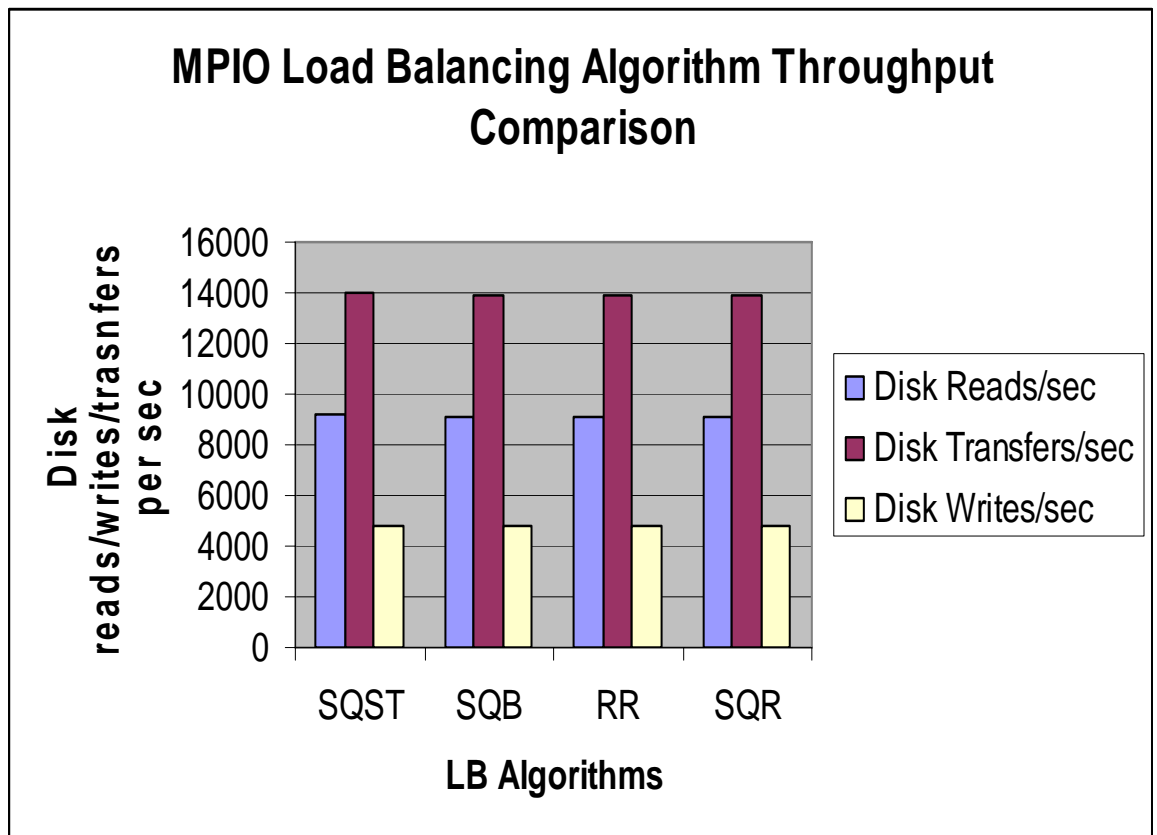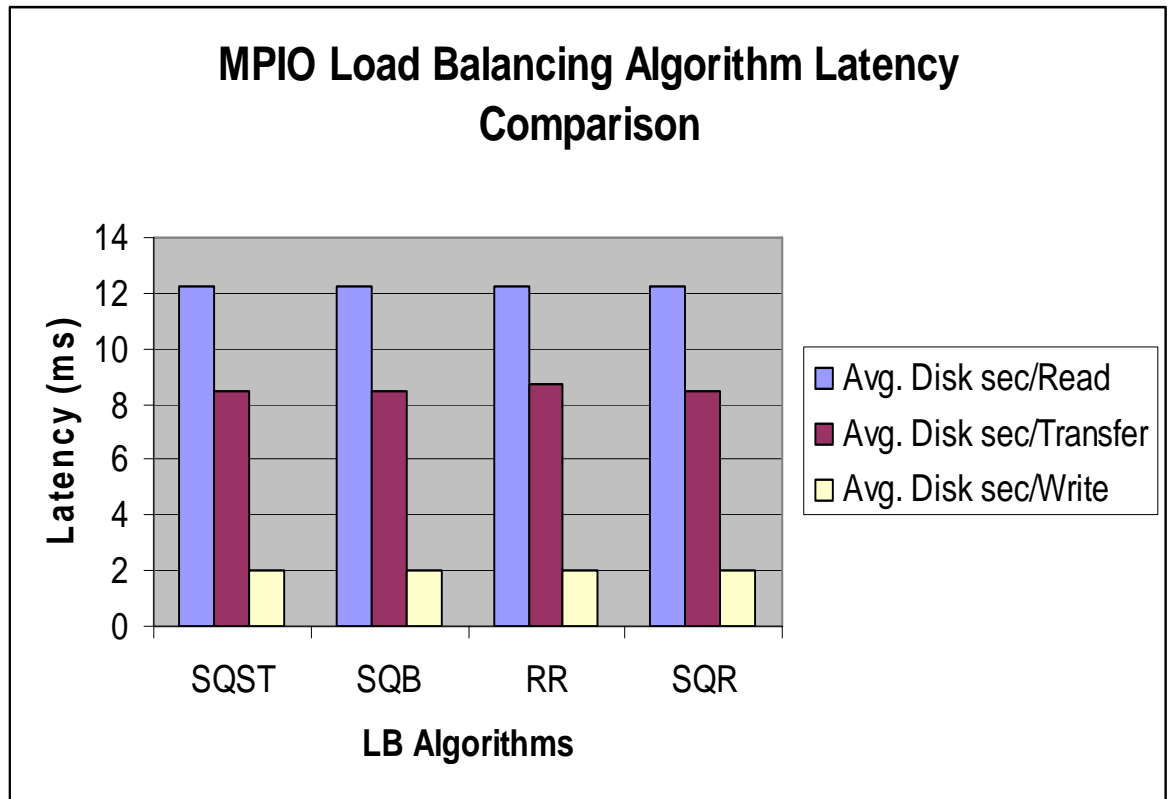**Figure 12.** Throughput results of the tests on the four load balancing algorithms

MPIO Load Balancing Algorithm Latency Comparison

These results indicate that for Exchange, there is no benefit of one load balancing algorithm over another. (Testing was also conducted with multiple hosts and yielded the same comparative results as the single host testing.) This conclusion is very dependent on the workload and will vary with different applications that have different workload footprints.

For Exchange production operations, both read and write requests need to be considered when evaluating the lack of a performance difference between different controller cache algorithms.

With the EVA controller, write-back caching is enabled (by default) when the Exchange virtual disks are created. When a write request is issued from the host, the request is written to either controller cache and the controller responds to the host with a write complete response. Since battery-backed write-back caching is enabled, Exchange doesn't have to wait for the data to be written to the physical disks, which can be a significant cause of data latency. Thus, as long as neither controller's cache is saturated, a write request preference for either controller won't impact performance or response time.

When a read request is issued from the host, a target controller is selected based upon the algorithm specified to control load balancing. The following sequence of events can then occur:

1. If the controller selected is the owning controller and the data resides in the cache, then the request is satisfied immediately and sent to the host. If not, then the data is requested from the disk and sent back to the host. (The owning controller is a property set on each virtual disk and can be manually configured with the EVA management interface).

2. If the controller selected is the proxy controller, then the proxy controller will send the request across the mirror port to the owning controller.
    a. If the data resides in the owning controller cache, then the data is retrieved from cache and sent back across the mirror port to the proxy controller and back to the host.
    b. If not, the data is retrieved from disk and sent back across the mirror port to the proxy controller and back to the host.

This document has addressed the random nature of the Exchange database workload. Because of the type of database activity, the likelihood of Exchange requesting data that is satisfied from cache is minimal. And since the Exchange transfer sizes are small, in the 4K – 16K range, transferring data of this size through the PCI buses and mirror ports is extremely fast. As long as neither controller is bottlenecked, the latency will be determined almost entirely by the time required to read from the spindles. This is because the latency associated with accessing data from a disk is substantially greater than the latency associated with accessing data from the cache or the latency incurred when transferring small Exchange data across the mirror port. This is why Exchange is termed a "spindle dependent" application. If the workload exhibited more benefit from cache read requests or exhibited larger size data blocks, then the algorithms that selected the owning controller more frequently would generally provide quicker response times, as the overhead of transferring the request and data across the mirror port would be negated. This would favor the SQST algorithm for Exchange.

Best Practice: The recommendation is to leave the default value, **Shortest Queue Service Time**, as the load balancing algorithm with the HP Full Featured MPIO for EVA DSM.
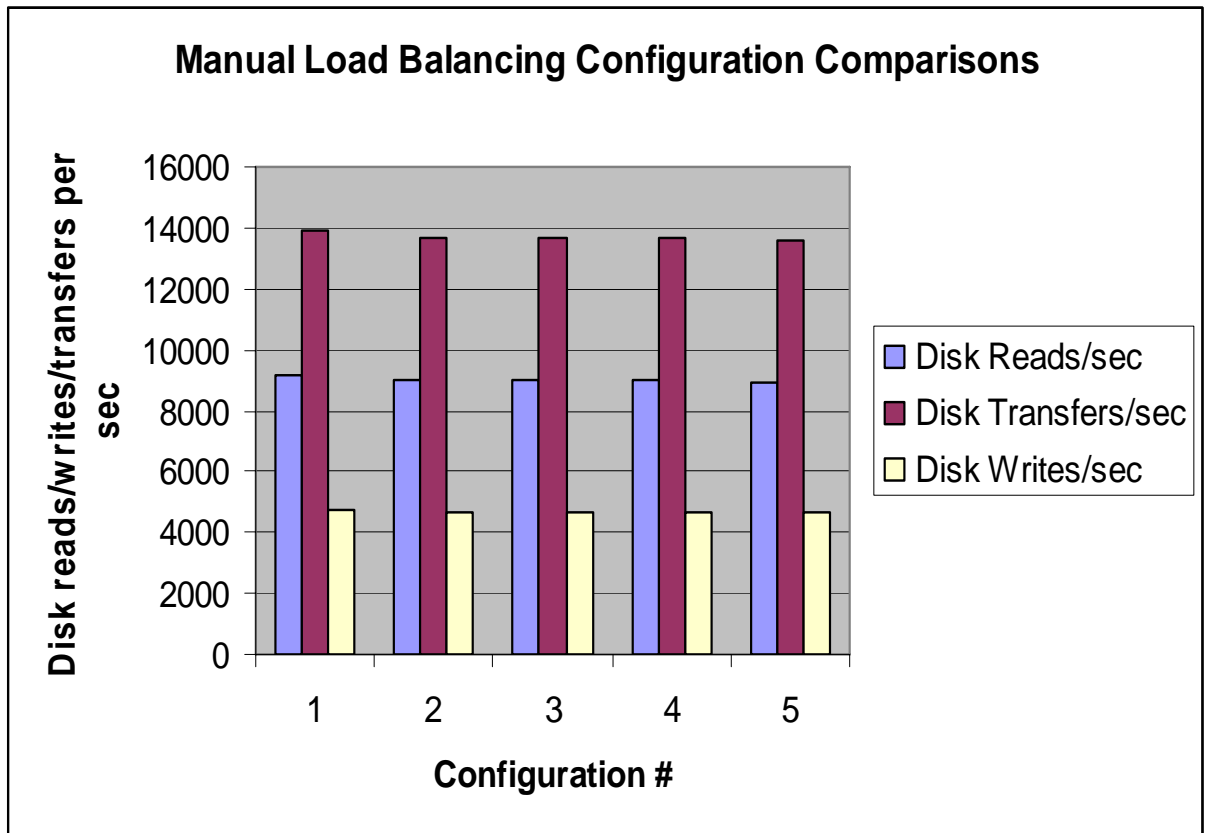
No load balancing

Further testing was also performed to examine the impact of setting the MPIO load balancing option to **no load balancing** and manually configuring load balancing. The rationale behind this set of tests was to measure the performance impact of oversubscribing a single controller or host port on the controller and measure any observed performance differences between manual load balancing configuration and the algorithms discussed above.

This testing also analyzed performance changes when manipulating LUN ownership and HBA path load balancing together. As described above, LUN ownership determines which controller owns the LUN and issues the data requests to the physical disks. The MPIO HBA load balancing algorithms determine which path to the controller the request is processed on.

The figure below, Figure 14, highlights the following five data points:

1. Virtual disks are balanced manually across both controllers A and B through the MPIO DSM manager. LUN ownership preference is configured to match the controllers specified through the MPIO DSM manager.

2. Virtual disks are balanced manually across both controllers A and B. LUN ownership preference is set for only a single controller B.

3. Virtual disks are manually configured to access a single controller A. LUN ownership preference is set for each virtual disk to the same controller A.

4. Virtual disks are manually configured to access a single controller A, through a single controller port A1. LUN ownership preference is set for each virtual disk to the same controller A.

5. Virtual disks are manually configured to access a single controller A. LUN ownership preference is set for each virtual disk to the opposite controller B.
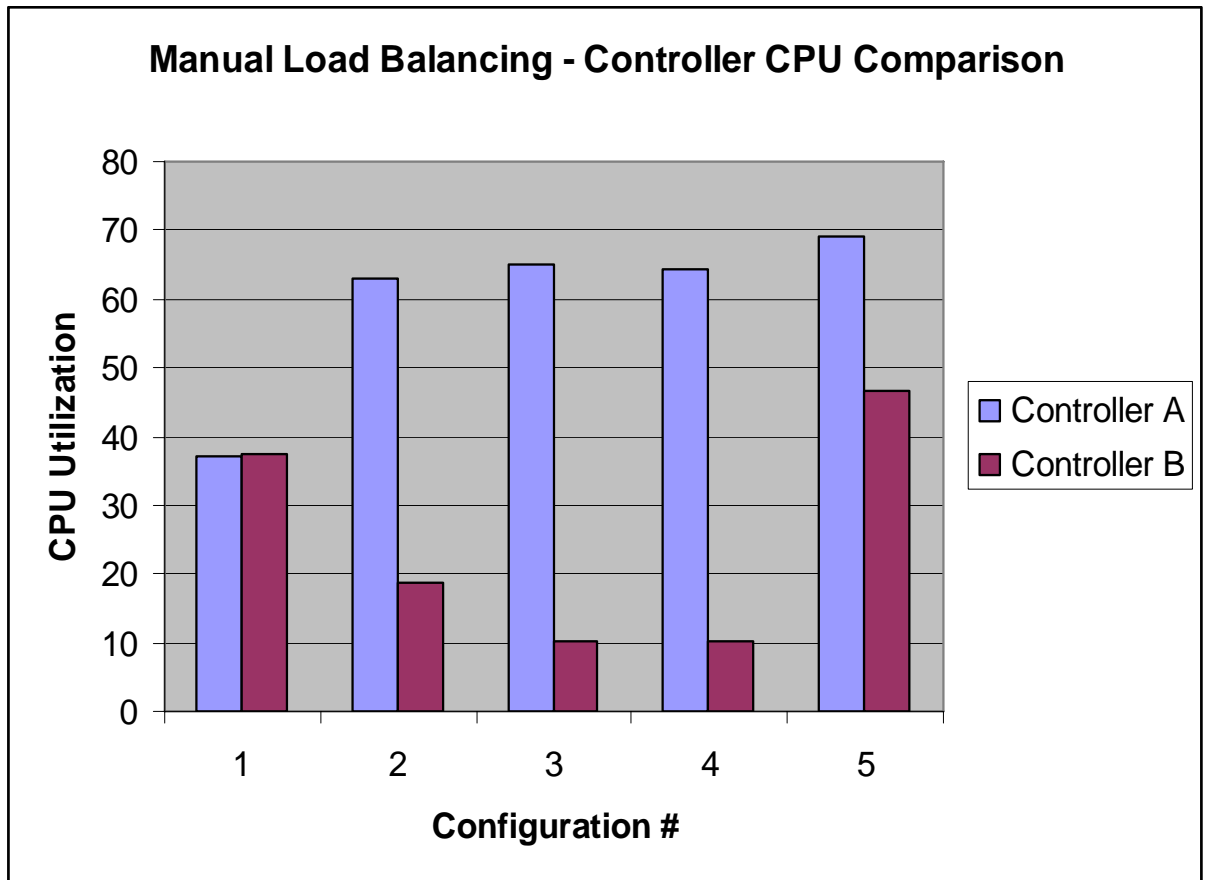
Figure 14. Various configuration throughput datapoints

From the graph, the throughput data indicates that nearly the same level of performance is achieved with all five configurations. Taking a look at the hypothetical optimal configuration, number 1, in which load is balanced evenly across each path and every request goes to the owning controller, as compared to the least optimal configuration, number 5, in which all traffic goes to the proxy controller and must traverse the mirror port, there is only a slight decrease from 13890 IOPS to 13587 IOPS, a difference of less then 5%.

Figure 15 depicts the EVAPerf Controller CPU Utilization counter stats for both controllers A and B for each configuration. This graph indicates the performance overhead from implementing non-optimal configuration MPIO and LUN ownership configurations.

Figure 15. EVAPerf Controller CPU Utilization counter stats

Again, the data reinforces that configuration number 1 is the most optimal in terms of EVA storage array utilization with both controllers averaging CPU utilizations less than 40%. This is as compared to the least optimal configuration number 5, which is 100% proxy activity, and both controller CPU averages are significantly higher. For reference, each MPIO load balancing algorithm tested above in Figure 12, averaged controller CPU utilizations evenly balanced at approximately 50%.

So why is this important for Exchange? First, the data reinforces that, given the footprint of the Exchange workload (small transfer sizes and little read cache success rate); there is very little performance difference in terms of host throughput between any of the configurations for the tested workload. However, the EVA controllers for sub-optimal configurations operate at a significantly higher CPU utilization percentage. As the load increases on the controller, or other I/O streams such as streaming backup/restore operations are initiated, the scalability of the controller will be impacted in less than optimal configurations.
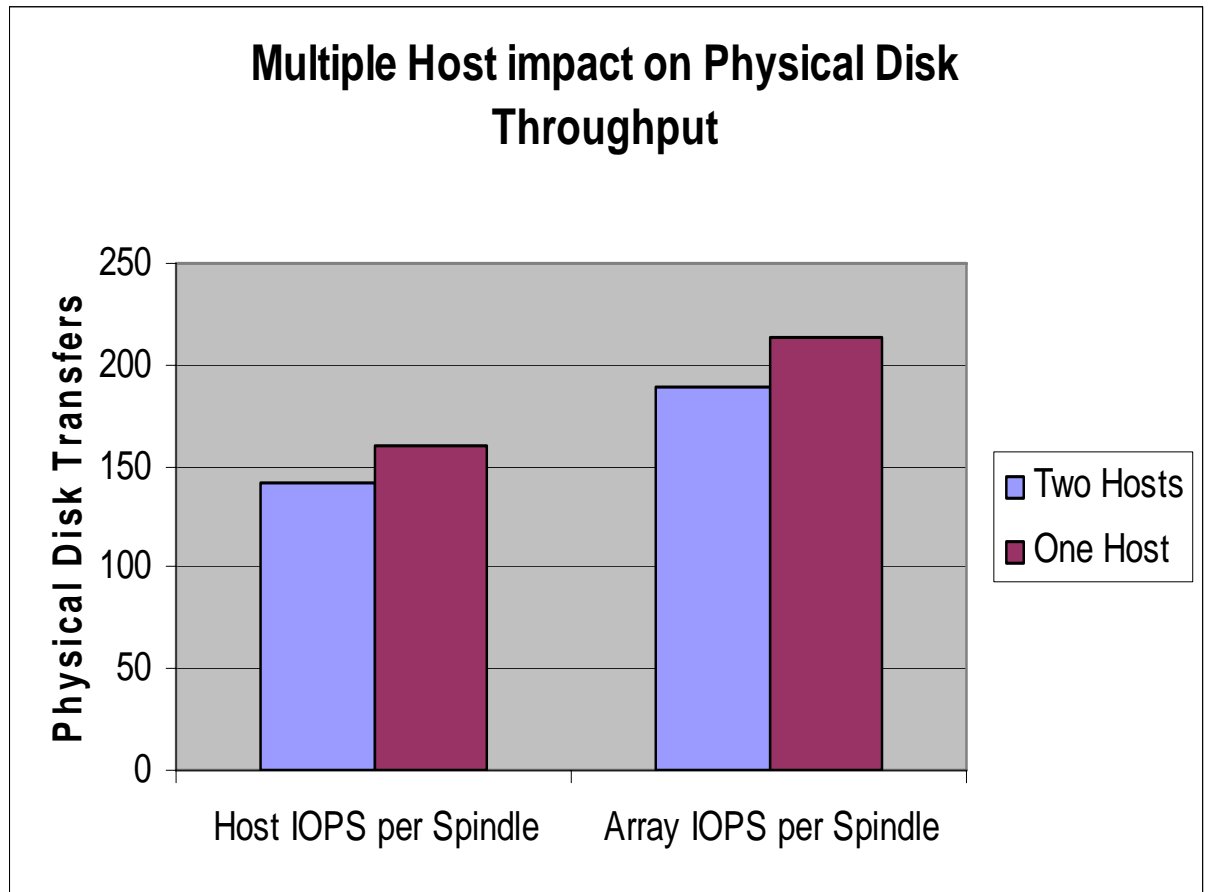
Best Practice: In configurations in which manual load balancing is utilized (such as with Microsoft clustering), ensure the following:

1. Paths to the virtual disks are load balanced evenly between both the HBAs on the host and the EVA controllers.

2. Host paths are configured to send data requests to the owning controller as configured in the virtual disk LUN ownership preference.

Appendix A – IOPS and sizing considerations

The data results outlined in this white paper focus mostly on the results from the application level as captured by the Windows perfmon utility. However, statistics captured via the EVAPerf utility reveal additional activity on the storage array from implementing RAID level redundancy. Figure 16 highlights this difference below.

Figure 16. Multiple host impact graph to examine host vs. array IOPS

This graph is an excerpt from the single- vs. multi-host configuration testing. In both cases, the array IOPS per spindle value exceeds the host IOPS per spindle value as is expected. This variable is one of several key inputs used in sizing and calculating spindle requirements for Exchange storage. Normally, the sizing algorithm will utilize the array IOPS per spindle value. However, this process can be skewed if an administrator utilizes the number of host IOPS per spindle, easily calculated from the graphs above (Figures 5 and 9 for example) by dividing the number of spindles into the total throughput. The following back-of-the-napkin sizing examples, A and B, show how differentiating and understanding the difference between these two values can help prevent improper storage sizing for Exchange. Example A uses the array IOPS per spindle value and example B uses the host IOPS per spindle value.

The following variables are constant:

- Disk Type = 72 GB-15K
- Average IOPS/u = 1
- Number of users = 5000
- Read/Write ratio = 66/33
- Concurrency = 100%
- Redundancy level = VRAID1

Example A – 180 array IOPS per spindle

1. The first step in the process is to calculate the total required number of IOPS.
   a. 5000 users * 1 IOPS/u = 5000 IOPS
2. Then, calculate the additional requirements to implement redundancy.
   a. (5000*.66) + (2*(5000*.33)) = 6600 IOPS
3. Finally, divide the new total number of IOPS by the number of IOPS per spindle.
   a. 6600 IOPS / 180 IOPS per spindle = **37 spindles**

Example B – 150 host IOPS per spindle

1. The first step in the process is to calculate the total required number of IOPS.
   a. 5000 users * 1 IOPS/u = 5000 IOPS
2. Then, calculate the additional requirements to implement redundancy.
   a. (5000*.66) + (2*(5000*.33)) = 6600 IOPS
3. Finally, divide the new total number of IOPS by the number of IOPS per spindle.
   a. 6600 IOPS / 150 IOPS per spindle = **44 spindles**

The host IOPS per spindle value is measured at the application layer and has no knowledge of underlying disk operations. The value simply measures the number of disk requests Exchange processes per second and has absorbed the redundancy overhead into the result. The general sizing algorithm above takes as input the read/write ratio and adds the overhead of RAID redundancy into the total number of IOPS that the storage array must support. Thus, using the host IOPS per spindle value in example B over-factors the RAID impact and computes more spindles than are required (44 vs. 37) to support the necessary throughput.

There are a number of white papers and tools available from the ActiveAnswers website from HP, www.hp.com/solutions/activeanswers/exchange to assist in this sizing and planning process.

Appendix B – Performance counters

To capture the necessary statistics for analysis, Windows Performance Monitor was utilized along with the EVAPerf add-in that enables monitoring of specific EVA subsystem counters.

Windows Performance Monitor counters

Windows Performance Monitor (Perfmon) is an MMC snap-in that enables monitoring of the utilization of operating system resources such as CPU, memory, and disk. The counters that are discussed in this white paper are described below.

Physical disk counters

The physical disk counters keep track of information pertaining to each instance of a disk presented to the server. There is an instance of these counters for each physical disk presented to the Windows operating system on the server.

- Disk Transfers/sec—The rate of read and write operations on the disk
- Disk Bytes/sec—The rate bytes are transferred to or from the disk during write or read operations
- Disk Writes/sec—The rate of write operations on the disk
- Disk Reads/sec—The rate of read operations on the disk
- Avg. Disk sec/write—The average time, in seconds, of a write of data to the disk
- Avg. Disk sec/read—The average time, in seconds, of a read of data from the disk
- Avg. Disk sec/transfer—The average time, in seconds, of the average disk transfer

For more information

For further information on HP Microsoft Exchange Server solutions, please visit: http://www.hp.com/solutions/exchange.

For additional information on HP StorageWorks products including the EVA8000, please see HP StorageWorks - Enterprise Class Storage Products, http://h18006.www1.hp.com/storage/enterprisestorage.html.

For general information on Exchange sizing and best practices, there are several white papers and planning tools that are free to download from, http://www.hp.com/solutions/activeanswers.

To learn more about HP products, services and solutions visit http://www.hp.com.

For more information about the HP ProLiant servers, visit http://h18000.www1.hp.com/products/servers/.

11/2005 -1