HP StorageWorks Enterprise Virtual Array 5000 and Microsoft® Exchange Server 2003: storage performance and configuration — white paper

Executive summary

Successful Microsoft® Exchange Server 2003 deployments depend on properly sizing and configuring the storage subsystem. Proper configuration of the storage array is critical for supporting the aggressive random I/O requirements of a Microsoft Exchange Server 2003 database. Understanding best practices for the Exchange environment and applying them to known performance data can be useful for setting up proper storage deployments and averting issues caused by incorrect array sizing and configuration.

Testing was conducted using an HP StorageWorks Enterprise Virtual Array 5000 (EVA5000) 2C18D configuration with 240 disks (72-GB, 15K-rpm drives). Of the 240 disks, 140 disks were used for testing and 100 disks were configured in a reserve disk group for disk-to-disk backup or other usage, which did not impact the performance testing.

This document includes the following:

- Best practices for the configuration of the EVA5000 storage subsystem in an Exchange 2003 environment based on data derived from performance testing
- Storage configuration recommendations for the EVA5000, including proper disk group and virtual RAID (VRAID) configurations based on data derived from performance testing
- An analysis of testing conducted on the EVA5000 using the Microsoft Jetstress utility and varying disk subsystem configurations

Key findings: The storage architecture with the highest performance (over 14,000 I/Os per second) was achieved by separating the database and log logical unit numbers (LUNs) into two disk groups (120 disks in the database LUN disk group and 20 in the log LUN disk group) and by configuring both database and log LUNs as VRAID1.

Overview

The EVA5000 is an enterprise class, high performance, high capacity, and high availability VRAID storage solution that removes the time, space, and cost boundaries of traditional storage. The EVA5000 provides improved storage utilization and scalability and meets Exchange Server 2003 specific demands for consistently high transaction I/O and MB data rate performance.

In corporations, messaging is a critical application, and proper configuration of the EVA5000 storage array is critical for supporting a Microsoft Exchange Server 2003 deployment. This paper suggests best practices based on data produced in HP engineering labs. The tests described in this white paper were driven by the Microsoft Jetstress utility, which is designed to simulate Exchange I/O load at the database level. In addition, tests were performed to determine the performance and scalability of the EVA5000 disk subsystem in various configurations, which included testing and quantifying the impact of:

- Changing disk group configurations
- Testing with different VRAID levels
- Testing the impact of running the DiskPar utility to realign the hard disk tracks with the Microsoft Windows® Server 2003 physical disk partitions

Performance testing

The goals of the testing performed on the EVA5000 were to determine the optimal performance configurations in a Microsoft Exchange Server 2003 environment, including maximum I/Os per second (IOPS) throughput, and to quantify the differences between suboptimal configurations. The

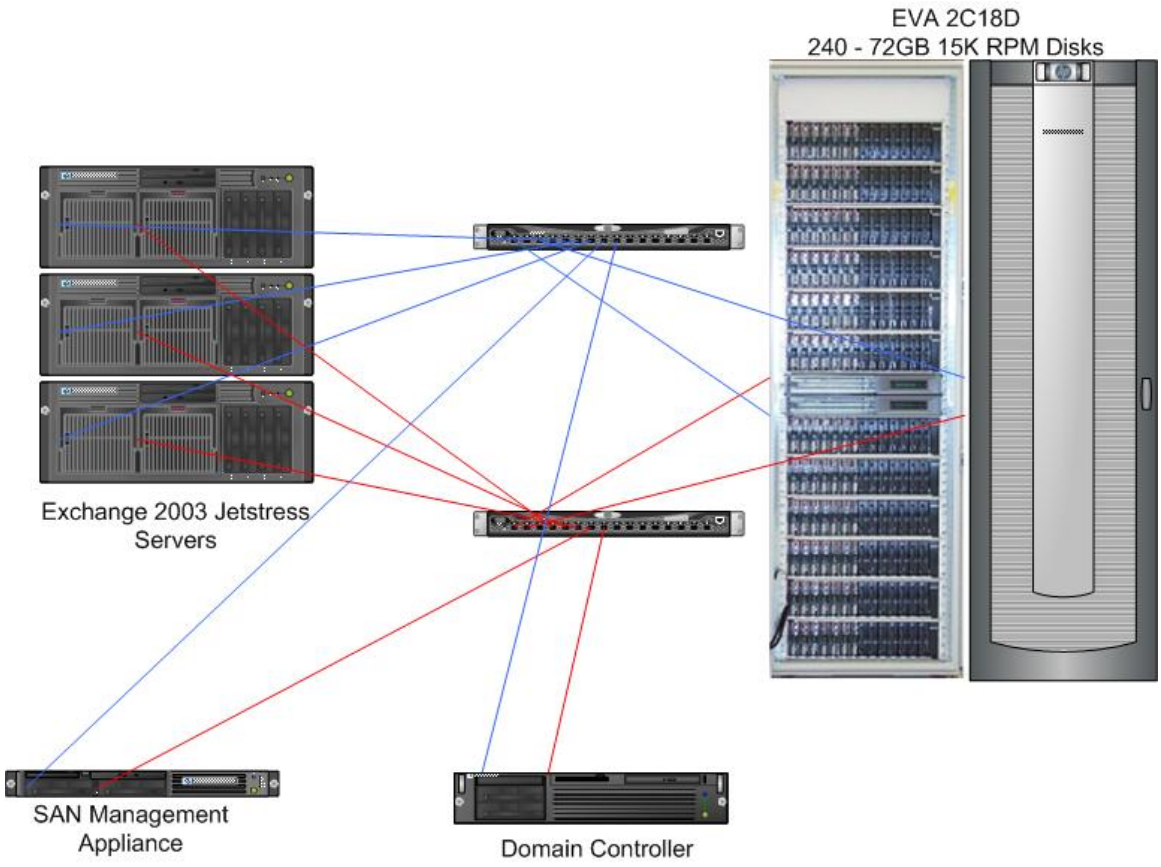
testing and analysis focused on two key design decisions when creating storage for an Exchange Server 2003 environment, which included disk group configuration and VRAID settings. A secondary analysis was performed using the DiskPar utility to realign the physical disk partition with the hard disk tracks, per Microsoft's recommendation, to determine the impact on performance.

The tests used the Microsoft Jetstress utility to drive an Exchange Server I/O load against the storage subsystem.

Test architecture and configuration

The tests used three HP ProLiant DL580 G2 servers connected to the HP StorageWorks Enterprise Virtual Array (EVA) storage area network (SAN) to drive Exchange Server I/O load. Figure 1 depicts the test configuration. See Appendix A—Reference Configuration BOM for more information.

Figure 1.



Microsoft Jetstress and test workload

Microsoft Jetstress is a command line utility that simulates Exchange Server disk I/O load with low-level Exchange Server I/O operations including page inserts, deletions, modifications, and seeks. It is used to verify the performance of a disk subsystem. To accurately compare the values obtained between multiple tests, the workload for each test must be equivalent. For each test, therefore, a consistent workload was maintained that ran against each storage configuration. The testing included:

- A 75:25 read/write Exchange Server 2003 workload ratio

- Test length of two hours
- Execution of two instances of Jetstress by each server (simulating two storage groups and two databases per storage group)
- 100-MB mailbox per user quota
- 5,000 users per server (This number was used to calculate the size of the Jetstress database for testing and is a constant in the testing. In previous testing performed with the Microsoft LoadSim 2003 utility, 5,000 users with 100-MB mailboxes had a value of .80 IOPS per user, which translates into approximately 4,000 IOPS per server to support database random I/O throughput.)
- Test capacity requirements based on 5,000 users and a 100-MB mailbox per user (500 GB per server, 250 GB per storage group, and 125 GB per database)
- 10% database size ratio (See Appendix B—Database ratio and Jetstress parameters for more information.)
- 12.5-GB Jetstress database per Jetstress instance (calculated as 10% of the size of the storage requirements)
- Standardized Jetstress parameters for test conditions and workload (See Appendix B—Database ratio and Jetstress parameters for more information on the parameters used in the testing.)

For this testing, the users per server and mailbox size per user variables were constant at 5,000 and 100 MB respectively. These values were used to calculate the size of the Jetstress databases to be used in the testing. Both parameters were selected after careful consideration and some preliminary testing. Because Jetstress has no notion of email messages or folders, there is no correlation between the Jetstress testing and a specific mailbox size or user count.

Initial thought was given to increasing the size of the Jetstress database by modifying the user mailbox size parameter from 100 to 200 and 400 MB per user. Preliminary testing indicated that these larger database sizes had negligible impact in this testing. For a given thread count, there was no difference in I/O throughput between any of the Jetstress databases; thus, no changes to this variable were made during the test.

Note
Previous testing by HP engineers with the Microsoft LoadSim 2003 utility analyzed the impact of modifying a user's mailbox size. This testing showed a correlation between an increase in the size of a user's mailbox and a decrease in the supported number of users on that particular server; however, the results of this testing are outside the scope of this document.

Reviewing the performance results

This section analyzes the differences between various EVA5000 storage configurations. The first section focuses on the comparison between disk group configurations (see the "Comparing disk group configurations" section), which used VRAID1 virtual disks. The second section highlights the performance differences between VRAID1 and VRAID5 virtual disks using the optimal disk configuration determined from the disk group testing (see the "Comparing VRAID1 and VRAID5" section). The third section discusses the results of the DiskPar testing (see the Testing disk partition realignment with DiskPar" section).

Each of the three HP ProLiant DL580 G2 servers ran two instances of the Jetstress utility for a total of six databases and six logs. For reference, the names used to describe the database and log LUNs are listed as follows:

- Server 1
  - DB1, Log1
  - DB2, Log2
- Server 2
  - DB3, Log3
  - DB4, Log4
- Server 3
  - DB5, Log5
  - DB6, Log6

Comparing disk group configurations

The EVA5000 utilized in the testing was a 2C18D configuration (two controllers, 18 disk shelves) with a total of 240 disks. The disk group (DG) comparison testing was conducted using VRAID1 LUNs for both databases and logs.

Disk group design

The disk group comparison testing analyzed the performance results of three configurations. The breakdown of the disk groups and the associated database and log LUNs are listed in the following tables.

Table 1. Single disk group—Single disk group for both databases and logs

Disk Group 1 (140 disks)*					
DB1	DB2	DB3	DB4	DB5	DB6
Log1	Log2	Log3	Log4	Log5	Log6

* See Note on page 6.

Table 2. Isolated configuration—Two disk groups in an isolated configuration, which contains a single disk group for databases and a single disk group for logs

Disk Group 1 (120 disks)*			**Disk Group 2 (20 disks)***		
DB1	DB2	DB3	Log1	Log2	Log3
DB4	DB5	DB6	Log4	Log5	Log6

* See Note on page 6.

Table 3. Crossover configuration—Two disk groups in a crossover configuration, which contains two disk groups with logs and databases in both disk groups

Disk Group 1 (65 disks)*			**Disk Group 2 (65 disks)***		
DB1	DB2	DB5	Log1	Log2	Log5
Log3	Log4	Log6	DB3	DB4	DB6

* See Note on page 6.

Each configuration also utilized a third disk group (DG3), which contained the additional 100 disks. There was no concurrent testing because this disk group served as a placeholder only. This disk group is reserved space for other considerations, such as performing Exchange Server 2003 disk-to-disk backup using snapshot and snapclone technology or other requirements. The existence of this disk group is not a requirement for the EVA5000 in a production environment.

Best practice: For both reliability and performance, the recommendation for an Exchange Server installation is to isolate the database and transaction logs. For performance considerations, HP recommends creating the largest possible disk group for the database I/O streams and isolating the transaction log I/O streams on a separate disk group.

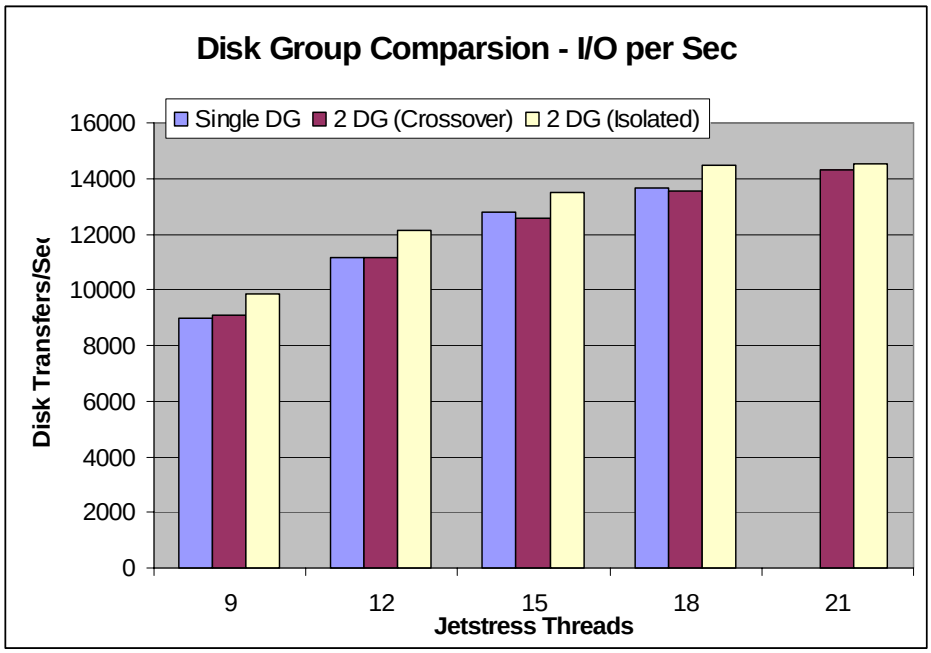
Note
In a production environment, the number of disks allocated to each disk group must be sized based on both capacity, providing sufficient space for user mailboxes or log files, and performance, providing a sufficient number of spindles to support the overall I/O load. However, HP recommends designing the EVA for performance before capacity to avoid poor performance from the disk subsystem. EVA array sizing tools are available to help with these calculations.

The data outlined in the following figure indicates an advantage to the design with two disk groups with an isolated database and log design, compared to the design with the largest single disk group and crossover reference configuration.

Disk group comparison results

Figures 2 and 3 highlight performance metrics for the sum of all the databases (two per server for a total of six) as the number of Jetstress threads are increased from nine to 21. Figure 2 depicts the performance of the three disk group configurations, the single, isolated, and crossover DG configurations. Figure 2 shows that disk transfers per second throughput is higher with the isolated two-DG configuration and that there is a performance advantage when separating database and log traffic into separate disk groups within the EVA subsystem.

Figure 2.

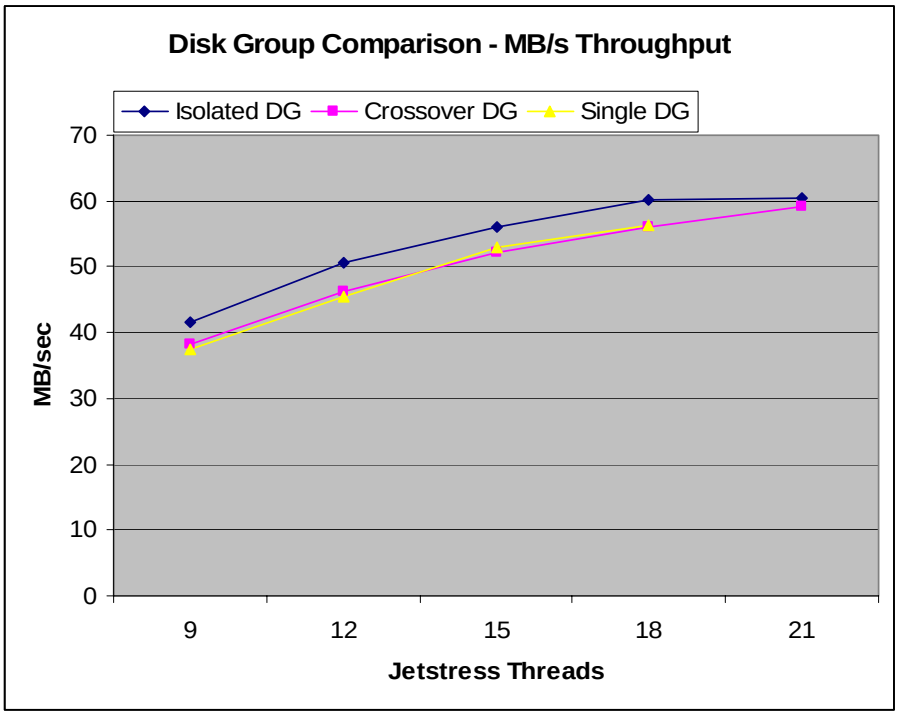


There is approximately a 5 to10% performance improvement when the EVA is configured in an isolated two-DG architecture. This configuration maintains a separation between the two-workload

streams and provides enhanced performance over the single configuration with the largest disk group, as well as the crossover configuration.

Figure 3 highlights the changes in throughput in terms of MB/s to the database LUNs in the configurations and reflects similar data as shown in Figure 2. In terms of throughput, the two-DG isolated configuration outperforms the single and crossover configurations.

Figure 3.



The EVAPerf counters capture specific information about the EVA using the Windows Perfmon utility. A full description of the counters is available in Appendix D—Performance monitors. Figures 4 and 5 analyze the cumulative impact on throughput as measured by EVAPerf for the total EVA subsystem. Figure 4 measures the impact of throughput in terms of KB/s throughput to the EVA. Figure 5 measures the performance impact on the total overall number of requests (cumulative read and write) per second handled by the EVA. There is no additional storage activity on the EVA besides the load placed on the subsystem by the Exchange Server Jetstress testing.

In looking at total host kilobytes per second (KBS), there is approximately a 7% increase in KBS in the isolated configuration over both the crossover and single disk group configurations. For total requests per second (RPS), there is an increase of about 10 to 11%. In both cases, the isolated configuration outperforms the crossover and single disk group configurations by a discernible margin.

Figure 4.

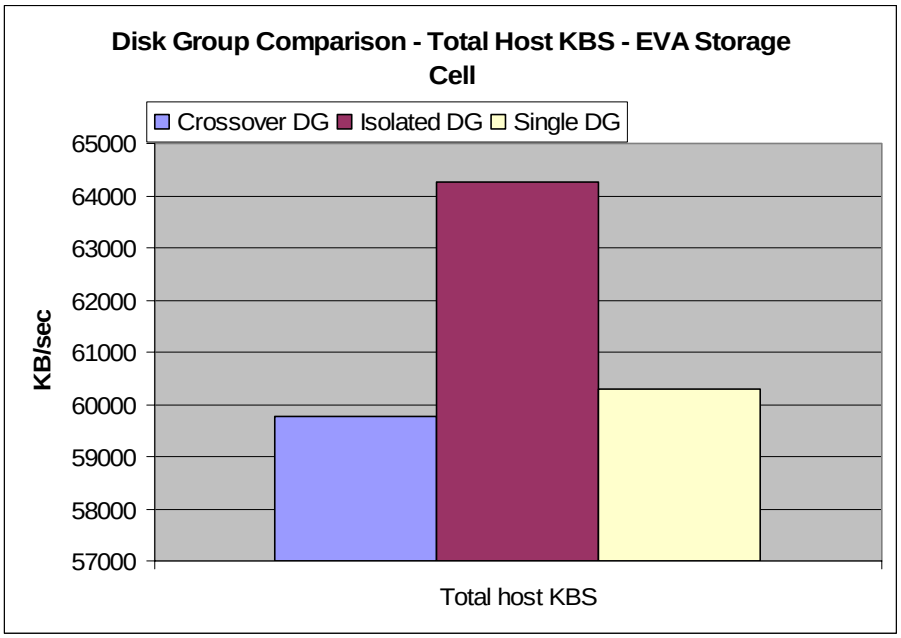
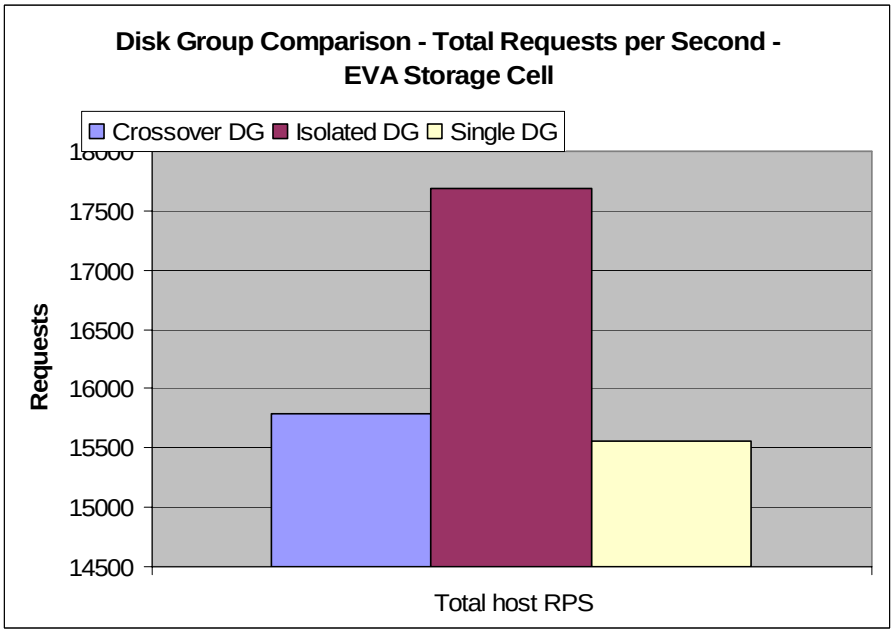


Figure 5.



As indicated in these graphs, there is little performance difference between the single DG and the crossover DG configuration. From a performance perspective, even the single disk group of 140 disks

did not have an advantage over the crossover configuration. Both the crossover and single disk group configurations incorporate mixed workload streams, random I/O from the Exchange Server databases, and the sequential I/O for the transaction logs on the same physical disks. The single disk group contains more then twice as many disks as each disk group in the crossover configuration (140 to 65). However, each disk group in the crossover configuration is only handling half of the total I/O throughput for the EVA.

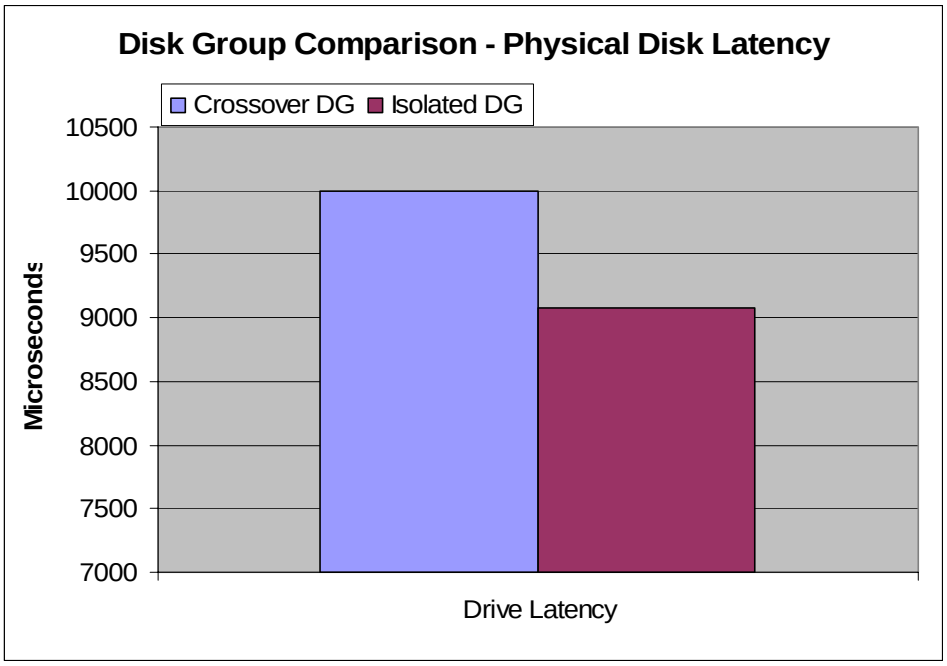
There is, however, consistently better performance from the isolated configuration. Because there are negligible performance differences between the crossover and single disk configurations and the isolated disk group configuration outperforms the single disk configuration, the size of the disk group is not the only rationale for performance improvement of an Exchange Server workload. As mentioned earlier, the best practice for Exchange Server performance is to incorporate the largest disk group possible for the database I/O but isolate the transaction log I/O streams onto a separate disk group.

Because the data outlined has shown little difference between the crossover and single disk group configurations, the remaining graphs in this section focus on a comparison between the crossover and isolated configurations, omitting the single disk group data unless otherwise mentioned and specifically focusing on the 18-thread data point for analysis.

Figures 6 and 7 highlight several of the EVAPerf physical disk counters, drive latency, read RPS, write RPS, and drive queue depth. The graphs depict an average value across a subset of the disks for each configuration. The graphs compare the physical disks from the database disk group in the isolated configuration, in which all the traffic is Exchange Server database random I/O, with the physical disks from either crossover disk group, in which there is a mixed workload on each disk, both random and sequential I/O.

Figure 6 shows the average drive latency in microseconds across the disks for the isolated and crossover two-DG configurations. The drive latency tracks the time between when a data transfer command is sent to a disk and when the command is completed.

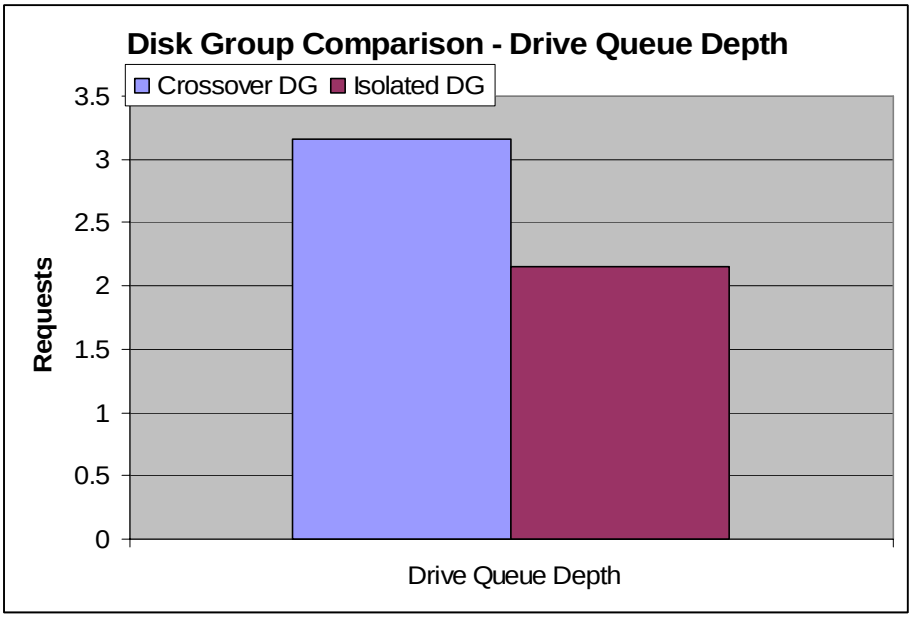
Figure 6.



The data reveals that there is more than a 10% increase in the latency time to complete a data transfer request on each physical disk with the crossover configuration, which indicates that a generic command completion is taking almost 1,000 microseconds longer on the physical disks in the crossover disk group.

As seen with the increase in physical disk latency, there is also an increase in the physical drive queue depth, which is the number of requests that have been sent to the drive but not yet completed (Figure 7). There is an increase of one request per second (50%) per physical disk within the crossover configuration disk groups compared to the disks in the isolated configuration.

Figure 7.

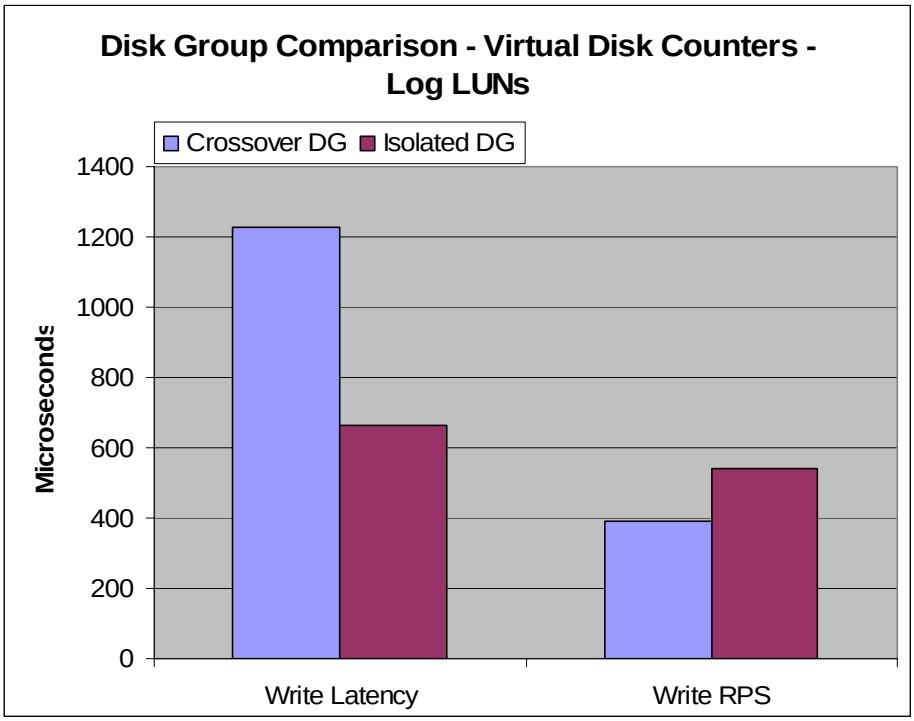


The single disk group results for physical disk latency and drive queue depth (not pictured in Figures 6 and 7) are almost equivalent to the numbers for the crossover configuration. The physical disk latency has a value of 10,121 microseconds, and the drive queue depth is 3.01 requests.

Figures 8 and 9 highlight several of the EVAPerf counters that measure performance at the virtual disk level. Figure 8 is a comparison graph of the transaction log virtual disks for the crossover and isolated disk groups. This graph illustrates the performance of the transaction log LUNs, primarily all write activity, and depicts the impact of mixing random and sequential I/O on the performance of the virtual LUN.

During the testing, the average time to complete a write request to the EVA for the transaction log LUNs in the isolated configuration is half the time required for the crossover configurations, from 662 to 1,226 microseconds. Following this observation, there is less average throughput per LUN for the crossover configuration, 850 KB/s, as compared to almost 1 MB/s throughput per LUN in the isolated configuration (data not pictured in the graphs).

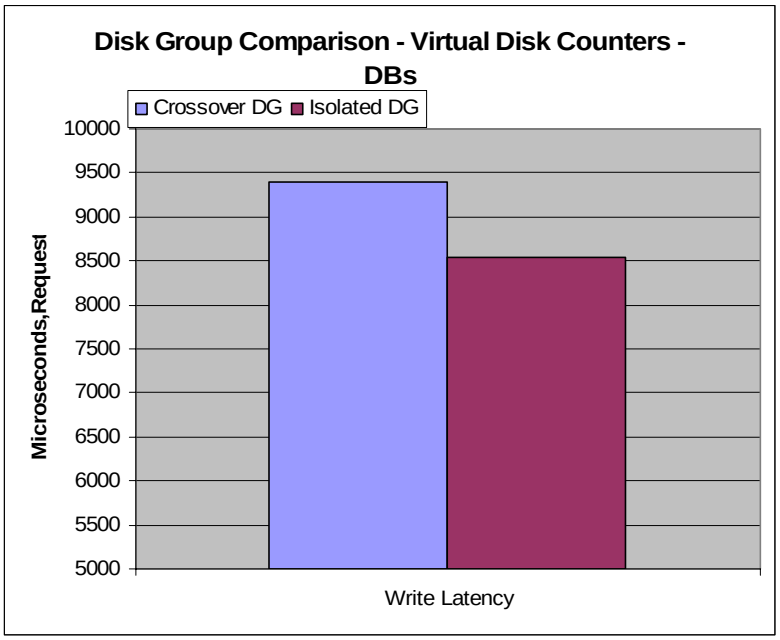
Figure 8.



Because there are additional latencies incurred in the crossover configuration, there is also a subsequent decrease in the number of write requests that are handled per second from the isolated configuration, which is a direct correlation to the throughput measurements discussed previously.

While not as dramatic, there is also an impact in the write latency on the database virtual disks. Figure 9 shows a 10% increase in the write latencies on the database LUNs.

Figure 9.



For additional graphs and analysis of the disk group comparison testing, see Appendix C—Additional performance data.

Comparing VRAID1 and VRAID5

The VRAID comparison testing analyzed the performance results of VRAID1 as compared to VRAID5 Exchange Server database virtual disks. 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 a single write request from Exchange Server to write out both the data block and 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 request, 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 HP StorageWorks Enterprise Virtual Array 5000 (EVA5000) utilizes cache optimization and write-gathering at the controller level to minimize the performance penalty of VRAID5 writes. 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. With multiple writes grouped together, the parity block need only be updated once. However, there is still a performance penalty for VRAID5 compared to VRAID1 database and log virtual disks.

Best practice: Choose VRAID1 for the best Exchange Server performance for both database and log virtual disks. The following data supports this recommendation.

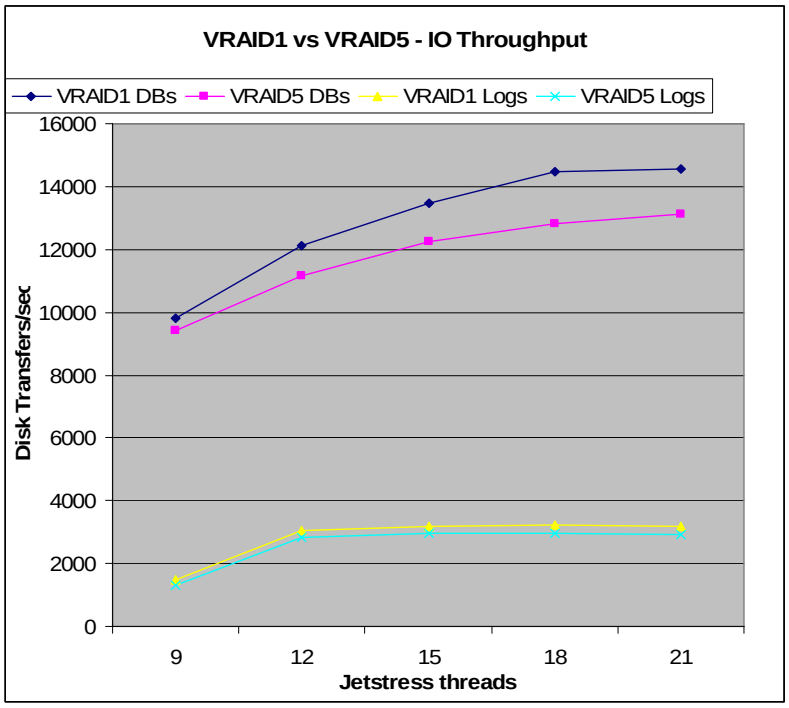
Disk group design

From the analysis performed, the isolated configuration outperforms the crossover configuration and the single disk group configuration. Therefore, the isolated configuration was used for both the VRAID1 and VRAID5 tests. See Table 2 for detailed information. For both the VRAID1 and VRAID5 test configurations, the transaction logs are VRAID1 LUNs.

VRAID comparison results

Figure 10 illustrates the total number of disk transfers per second for the sum of all the databases, as the number of Jetstress threads are increased from nine to 21, for the VRAID1 and VRAID5 configuration. When deploying VRAID5 LUNs, there is a performance penalty for write intensive applications because of the additional cost of calculating and writing out the parity bit. The graph indicates that, as expected, I/O throughput is higher when the database LUN is configured with VRAID1.

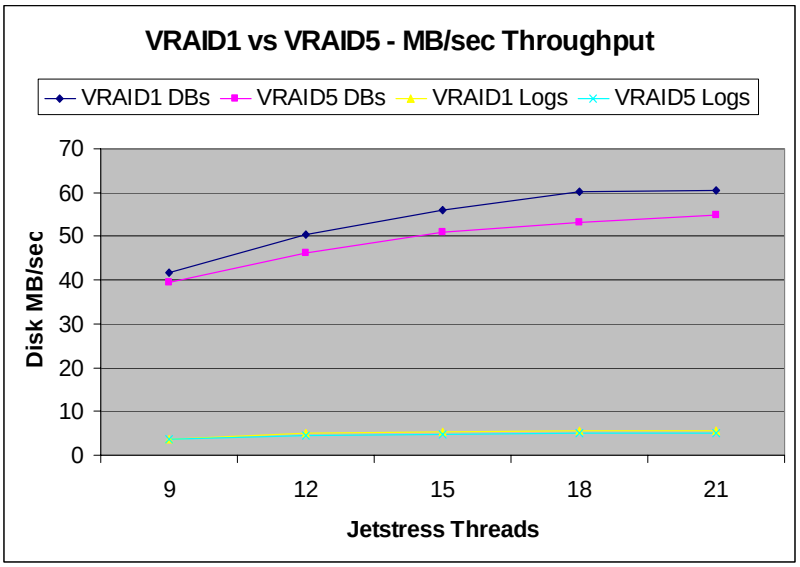
Figure 10.



As the number of Jetstress threads increase, the percentage improvement for disk transfers per second increases from roughly 5% (9,425 to 9,830) to approximately 10% (13,123 to 14,540). For both VRAID1 and VRAID5 configurations, the logs LUNs were configured as VRAID1, thus there is only a small increase in the I/O, representative of the additional throughput when configuring VRAID1 database LUNs.

Figure 11 reflects the same results as Figure 10, measured in terms of Disk MB/s throughput.

Figure 11.

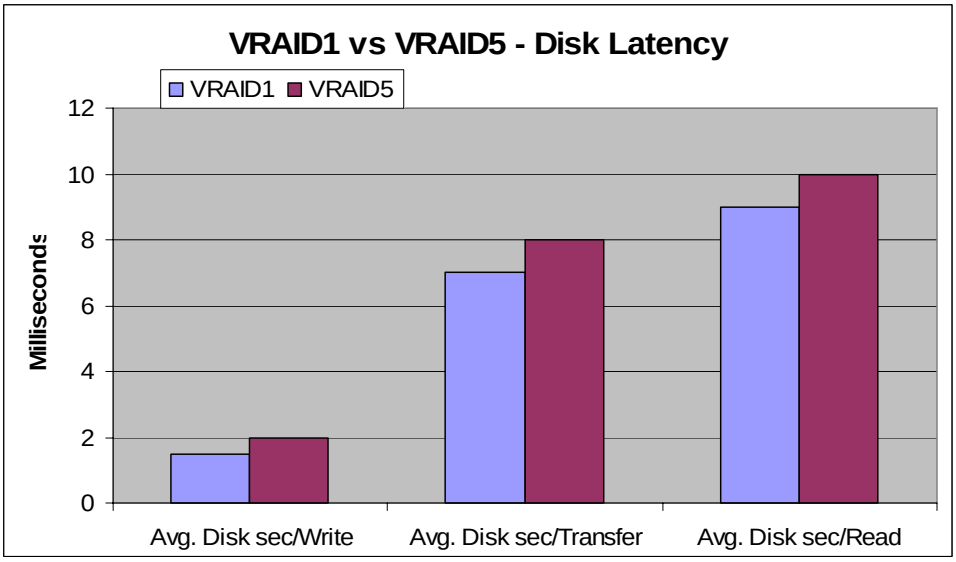


As was the case with disk transfers per second, Disk MB/s performance for the VRAID1 configuration outperformed the VRAID5 configuration. At nine threads, there is a performance gain of about 5%, scaling up to a gain of almost 10% in the 21-thread testing.

In both disk transfers per second and disk MB/s, performance for the VRAID1 configuration surpasses the VRAID5 configuration performance. However, similar to the results depicted in the disk group comparison section (see Appendix C—Additional performance data), there is minimal impact on the response time of the disks.

Figure 12 shows a small improvement when utilizing VRAID1 compared to VRAID5 LUNs. The VRAID1 testing yielded an average disk seconds per write latency of 1.5 ms, while the VRAID5 average increased to 2 ms. The average disk seconds per transfer (combining both write and read times) was also slightly higher for the VRAID5 tests at 8 ms compared to 7 ms.

Figure 12.



Note

While the data generated by this testing might not indicate a substantial performance difference between VRAID1 and VRAID5 database LUNs, HP highly recommends utilizing VRAID1 LUNs for high performance Exchange environments. Previous testing and qualitative analysis with smaller disk groups (56 disks) has shown far greater performance improvements when configuring VRAID1 compared to VRAID5 LUNs. Also, a higher percentage of write activity can occur in production environments, and as such, a greater penalty is incurred with VRAID5. Latency can be impacted when the write percentage increases above 30% with VRAID5 LUNs. For best performance and high availability, HP strongly recommends configuring the virtual disks for the EVA5000 as VRAID1.

Each of the following graphs compares only the VRAID1 and VRAID5 results during the 18-thread Jetstress tests and highlights the EVAPerf counters. Figure 13 analyzes the EVAPerf Physical Disk, Drive Queue Depth counter, measuring the total number of requests that have been sent to the disk but have not been completed. The result of the VRAID5 configuration is a jump of 50% in the number of outstanding requests increasing from approximately two to three requests.

Figure 13.

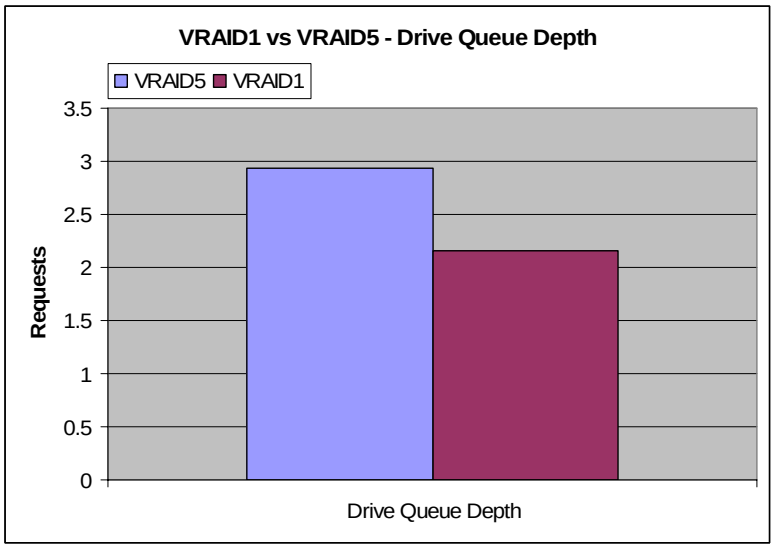
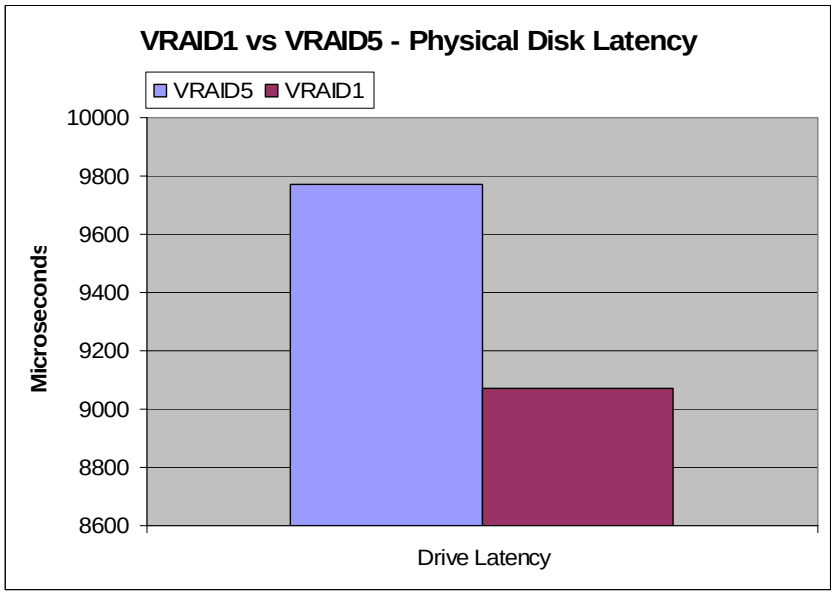


Figure 14 focuses on the latency counter for the EVAPerf Physical Disk object. The graph shows the increase in drive latency for the VRAID5 testing. The EVAPerf Physical Disk – Drive Latency counter records the time from when a command is sent to a disk until the command is completed.

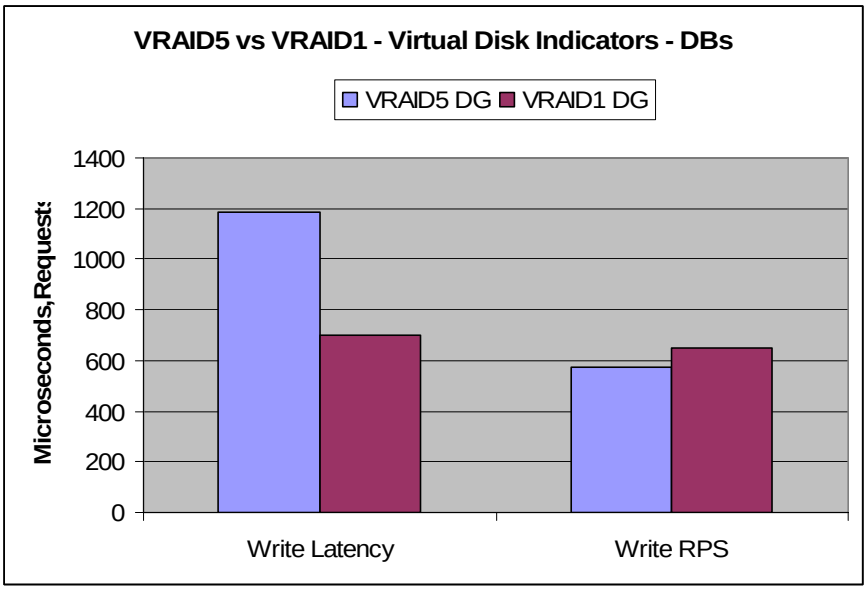
Figure 14.



In this testing, the VRAID1 configuration yields a 7% improvement from the VRAID5 configuration, a decrease from 9,771 to 9,071 microseconds per physical disk.

Figure 15 analyzes the latency on the virtual disks, highlighting the EVAPerf Virtual Disk object counters, specifically around latency and requests per second for the database virtual disks. The average write latency of the database LUNs for VRAID5 is nearly double the average latency for the VRAID1 LUNs, which is representative of the penalty incurred for VRAID5 write operations. There is also a subsequent increase in the average number of write requests per second handled by the VRAID1 virtual disks.

Figure 15.



Note
In both configurations, VRAID1 and VRAID5, the log virtual disks are configured as VRAID1. Outside of the additional write and read RPS handled by the VRAID1 log drives (a correlation to the additional work performed by the VRAID1 database drives), there is no discernible difference in log metrics to record and thus none of the data has been highlighted.

For additional data see, Appendix C—Additional performance data.

Testing disk partition realignment with DiskPar

As part of the disk group and VRAID performance testing on the EVA5000 with Exchange Server 2003, an analysis of the impact of realigning the Windows primary partitions to match the disk tracks with DiskPar was also performed.

As a quick background advisory, applications that utilize EVA VRAID disks might experience a write performance penalty with the default Windows 2003 primary disk partition alignment. Windows 2003 uses the first 63 sectors for volume information before the start of the first partition, causing the first partition to start on the last sector of the first track. Exchange Server 2003 writes out data in 4,000 chunks so every eighth I/O will cross a track boundary, resulting in additional latency on the

I/O request. Using the DiskPar utility before formatting the drive, the alignment can be set so that the first partition begins with a sector offset alignment of 64, rather than the default 63, which causes the first partition to begin on a new track without incurring any track overlapping.

For more information on configuring the partition offset with DiskPar, Microsoft's "Optimizing Storage for Exchange 2003" provides specific information at http://www.microsoft.com/downloads/details.aspx?FamilyID=c6084d20-9730-4ffc-805d-b957327604c6&DisplayLang=en.

The default for this testing was to configure a sector alignment of 64 with the DiskPar utility on all logical disks in the environment. After completing the initial series of tests, the Jetstress tests were repeated with the same parameters and workload without running DiskPar, utilizing the default alignment of 63 sectors. In analyzing the data for this testing, there was no discernible difference in throughput between the two sets of tests in either a VRAID5 or VRAID1 configuration.

Qualitative evidence has shown that sector realignment with DiskPar has the greatest impact on large block sequential writes to VRAID5 LUNs rather than random I/O data streams. A significant impact can occur when performing a disk-to-disk backup using a VRAID5 LUN for the destination volume, for which there are large block sequential writes.

Best practice: HP recommends setting the sector alignment to 64 using DiskPar per Microsoft's recommendation for new HP StorageWorks Enterprise Virtual Array (EVA) Exchange Server 2003 environments. In existing environments, the key is to properly evaluate and understand the performance of the Exchange environment. If the Exchange 2003 environment is performing well and meeting the customer's performance requirements, there is no reason to rebuild the architecture to run the DiskPar utility.

Appendix A—Reference Configuration BOM

Table 1. Exchange Jetstress servers—HP ProLiant DL580 G2 server

Exchange 2003 Jetstress servers	**Quantity**
HP ProLiant DL580 G2 server—(2) Intel® Xeon™ 2.0 GHz procs 1-MB Integrated Level 3 Cache, 2 GB ECC SDRAM NC7770 PCI-X Gigabit Server Adapter RAM upgrade, 4 GB ECC SDRAM	3 servers
36.4-GB 15,000-rpm, U320 Universal Hard Drive, 1 in.	4 per server
FCA 2101 2-Gb Fibre HBA for Windows 2003	2 per server
HP StorageWorks Secure Path 4.0C for Windows, 1 license	1 per server
Windows Server 2003, Enterprise Edition	1 per server
Exchange Server 2003, Enterprise Edition*	1 per server
Microsoft Exchange Jetstress utility	N/A

* Full Exchange 2003 install is not required to run the Jetstress utility. However, the full distribution was installed on the servers before testing.

Table 2. Storage Array—HP StorageWorks Enterprise Virtual Array 5000 (EVA5000)

Management Appliance	**Quantity**
HP OpenView Storage Management Appliance III	1
HP OpenView Storage Management Appliance software 2.1	1 per appliance
Storage – Enterprise Virtual Array	1
EVA5000—2C18D-B configuration	1
HP StorageWorks Virtual Controller Software Package 3.014 for Dual HSV110 Controllers (VCS 3.014)	1 per array
72-GB, 15,000-rpm disk drives	240
VCS 3.0 Platform Kit for EVA5000—Windows 2003	1 per array
Fibre Channel cables	As required

Appendix B—Database ratio and Jetstress parameters

Database size ratio

It is critical to create Jetstress databases that match the expected database sizes in the production environment. Exchange clients usually access only a portion of their data on a day-to-day basis. If a user has a 100-MB mailbox quota, you can assume that the user does not access all of the database in a given day. Through analysis of Exchange client usage patterns, it has been determined that an approximate ratio of accessed data to total data is 1:10. (For example, 10 MB of data is accessed on a given work day in a 100-MB mailbox). Be sure to size Jetstress databases in correlation to this 1:10 ratio. Jetstress, unlike real Exchange clients, will access 100% of the database. If the production database size (or multiple databases within a storage group) is 100 GB, be sure to build Jetstress databases of at least 10 GB to fully exercise the caching mechanisms of the disk subsystem.

After you have calculated the database size based on the production database size, it is important to consider the storage cache size. If the Jetstress database size is relatively small (equal to or less than the storage cache size), the test results of the disk latency are not a true measurement of disk performance. Because Jetstress can only run for a short time, the only way to get a good measurement of disk performance is to increase the database size to at least double the size of the storage cache size, so that the disk latency readings are a measure on the disk I/Os and not cache I/Os.

The database sizes for this testing follow the 1:10 ratio, and all are sufficiently larger than the EVA cache at 25 GB per Jetstress instance.

Jetstress parameters

The actual workload that Jetstress performs is also configurable. Four actions can be configured to execute, including record insertion, modification, deletion, and seeking. By default, Jetstress only performs seek operations, which is not indicative of many real-world workloads. Because one of the assumptions of the testing was a 75:25 read/write ratio, preliminary testing was conducted to adjust the operations to generate the consistent 75:25 read/write ratio.

Note
Where applicable, the recommended settings from Microsoft's "Verifying the Performance and Stability of the Disk Subsystem" document were utilized. For further information, this document is available as part of the Jetstress2004 utility download in the "Running Jetstress Tests with Jetstress.exe" section at http://www.microsoft.com/downloads/details.aspx?FamilyId=94B9810B-670E-433A-B5EF-B47054595E9C&displaylang=en.

The following parameters were used for the testing:

- -l = Log directory
- -n = Number of 1-KB records to insert (only used for database initialization)
- -b = 64 MB cache
- -t = Number of threads X (The thread count measures the number of Jetstress threads that are launched. The more threads that are started, the more I/O that is generated against the EVA. X indicates a variable in the testing to increment the I/O load.)
- -l = 15 (Percentage of insertion operations)
- -r = 35 (Percentage of replace operations)
- -d = 5 Percentage of deletion operations (By default, the remaining 45% of the transactions are seek operations.)

- -z = 93 (The percentage of all commits that will be committed lazily. The switch is used to better simulate the Exchange production environment and to control the log write size to the log drive.)
- -a = Attach to a previously created database (default is in the running directory), instead of creating a new one
- -q = 7200 (Length of the test in seconds. Tests will run for two hours [7,200 seconds] to provide an ample interval of time to sample the EVA performance.)
- -g = 500 (The number of transaction log buffers being used by the database engine. This value is recommended in Microsoft benchmark testing.)

Two instances of Jetstress will be executed on each server, which will simulate two storage groups on each server. To execute an instance of Jetstress the following command line parameter was used.

```
Jetstress.exe – l {log directory} –n 0 –b 64 –t X – I 15 –r 35 –d 5 –z 93
–a –q 7200 –g 500
```

Appendix C—Additional performance data

This appendix contains additional data from the performance testing.

Additional disk group comparison graphs and analysis

Figures 16 and 17 provide a quick glance inside the cumulative disk writes per second and disk reads per second breakdown for the database LUNs and the disk writes per second for the log LUNs. As expected, more writes and reads are completed per second in the isolated configuration for both the databases and logs, correlating directly to the increased throughput for the isolated configuration as compared to either the crossover or single disk group configuration.

A quick look inside the numbers shows that the read/write workload ratio is approximately 75:25 and 73:27 for both tests, reflecting the ratio desired in the parameters section.

Figure 16.

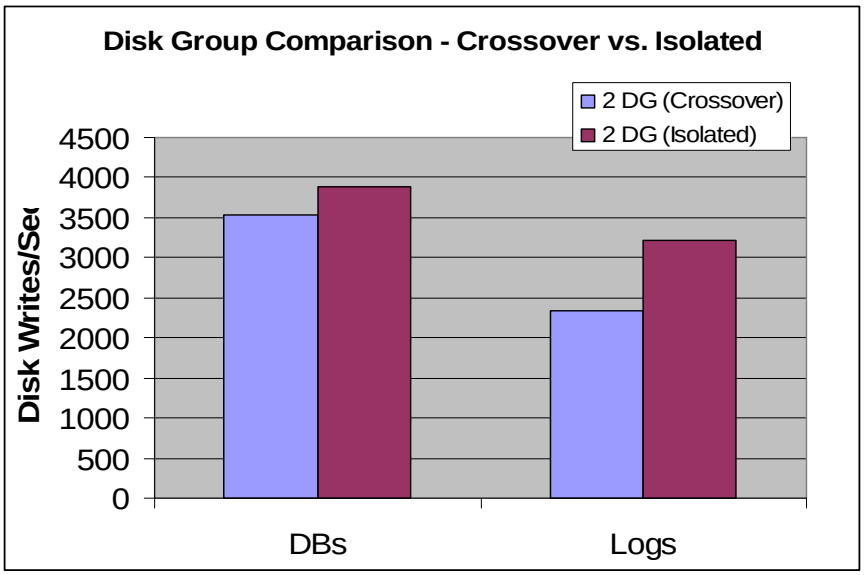


Figure 17.

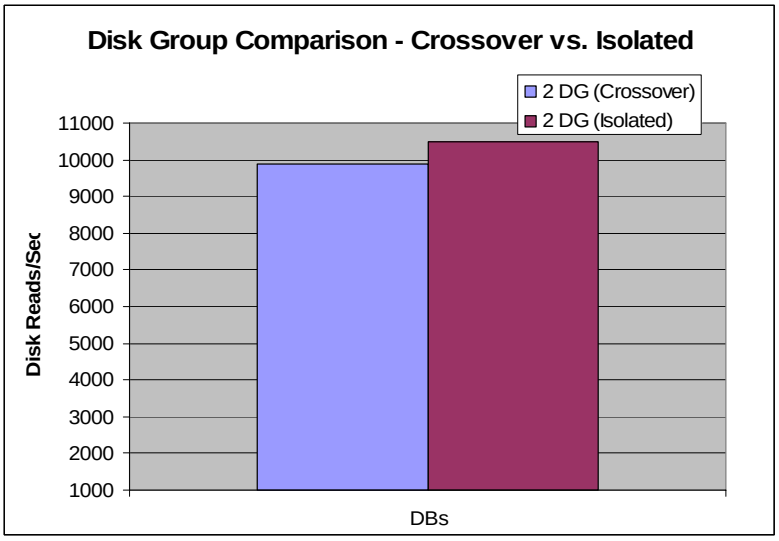


Figure 18 shows a comparison of the average response times for the database LUNs for the two disk group configurations. At 18 threads, the total disk transfers per second were approximately 13,500 and 14,400 IOPS for the EVA database LUNs. However, as the graph depicts, there are sub-10 ms response times for isolated and crossover configurations and negligible response time differences between the two configurations.

Figure 18.

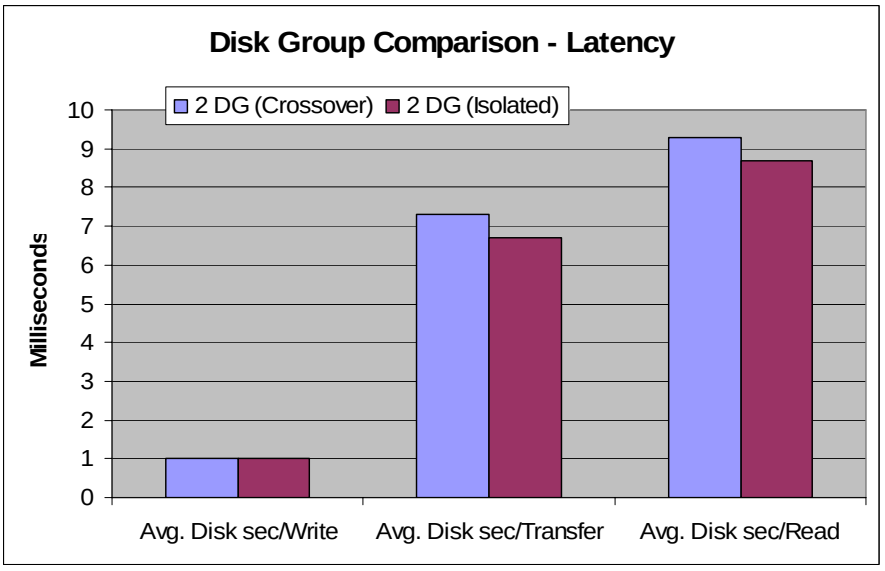
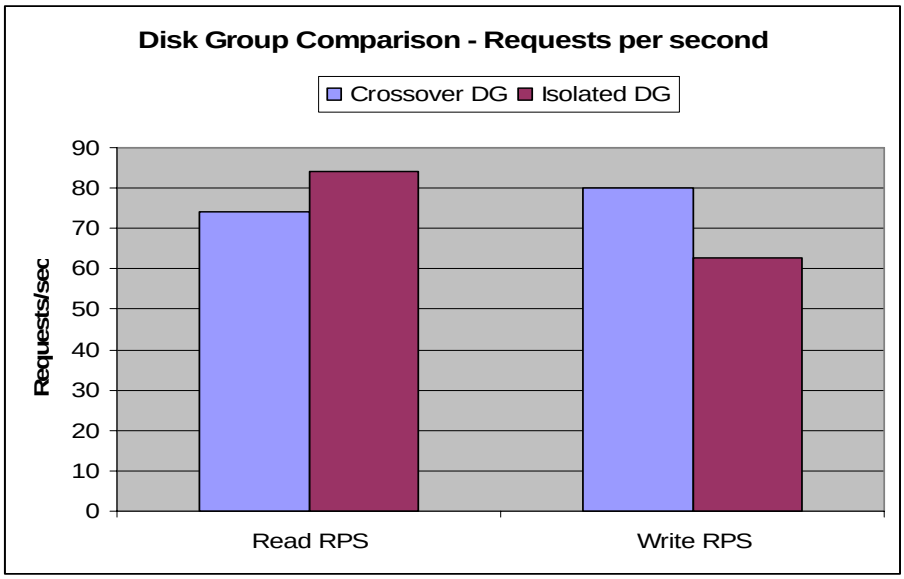


Figure 19 provides an informational view of the type of workload performed by each physical disk. The graph charts the average read and write requests processed per second by the physical disks in each disk group configuration.

Figure 19.

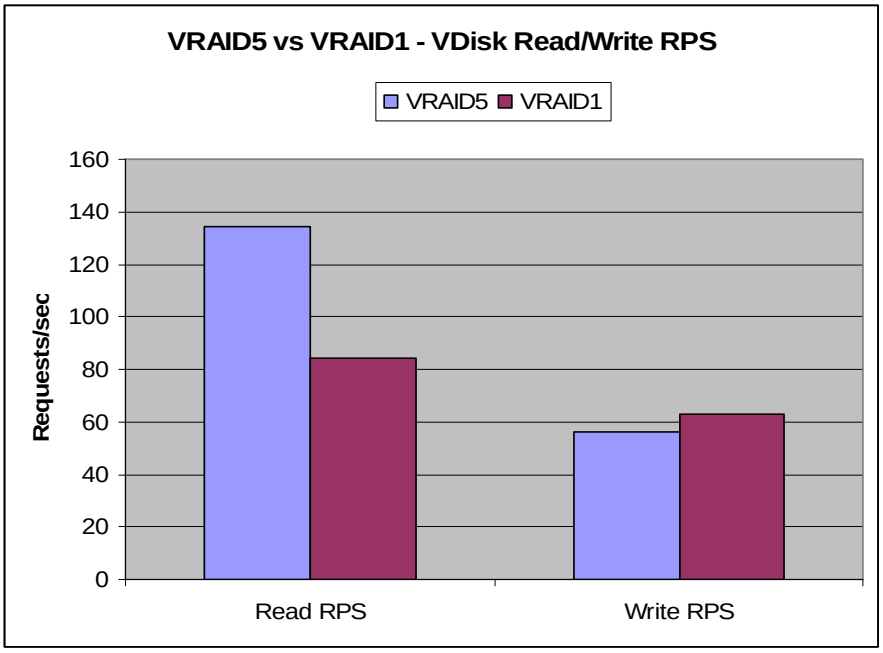


For the physical disks in the crossover disk groups, more write activity is requested per second, as a result of servicing transaction log requests. The disks can handle less read activity than the disks from the isolated configuration.

Additional VRAID comparison graphs and analysis

Figure 20 depicts the breakdown of the average read and write requests per database virtual disk. Notice that there was a substantially higher number of read requests for the VRAID5 configuration than for VRAID1, representative of the additional read requests incurred during the write penalty phase for VRAID5 LUNs. For each write request to a VRAID5 LUN, there is a maximum of two additional read requests to read the old data and to read the parity block. The EVA5000 cache optimization and write-gathering techniques reduces the number of disk I/Os as a result of the parity penalty, but there is still an overhead incurred with VRAID5 as indicated in the additional read requests per second.

Figure 20.



Figures 21 and 22 highlight the total EVA subsystem, measuring the throughput in KB/s and the number of requests (read and write) per second handled by the EVA. There is no additional storage activity on the EVA besides the load placed on the subsystem by the Exchange Jetstress testing.

For both total host KBS and total RPS, there is a 10% increase when utilizing VRAID1 database LUNs over VRAID5.

Figure 21.

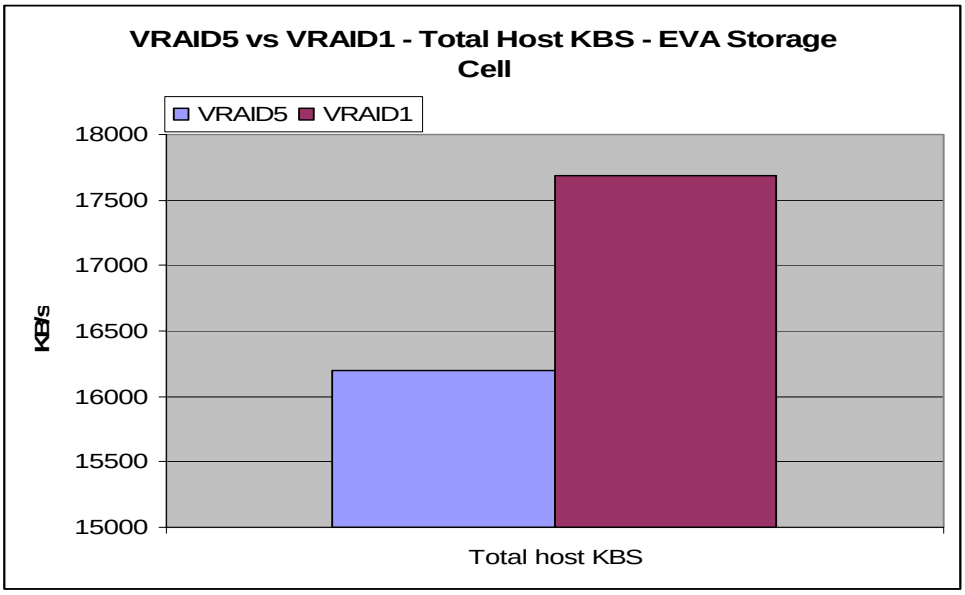
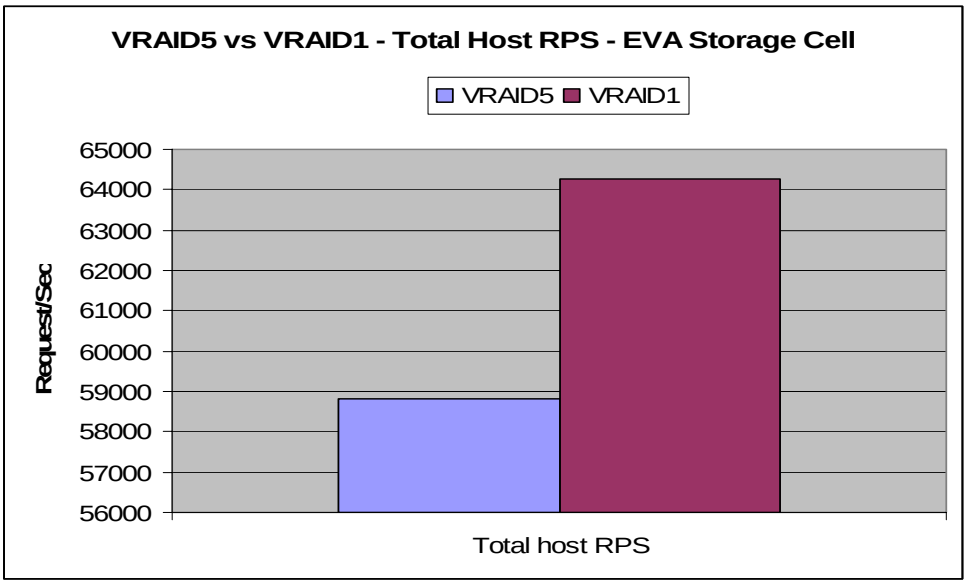


Figure 22.



Appendix D—Performance monitors

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 in the following sections.

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 to the disk
- Avg. Disk sec/transfer—The average time, in seconds, of the average disk transfer

EVAPerf counters

The EVAPerf utility is an add-in to the Windows Performance Monitor for monitoring of the EVA subsystem.

EVA physical disk counters

The physical disk counters keep track of information on each physical disk on the system. There is no information relating these disks to a specific disk group, nor is the activity broken out into the underlying cause of the I/O, such as host driven, cache flushes, read-ahead, leveling, and snapshot activity.

There is one instance of these counters for each physical disk on the EVA. Each disk is uniquely identified by a four-digit hexadecimal number. This number is an internal representation of the disk used by the EVA known as a "noid" and has no relationship to the shelf or bay where this disk resides.

- Drive latency—This counter tracks the time between when a data transfer command is sent to a disk and when command completion is returned from the disk. This time, which is measured in microseconds, is not broken into read and writes latencies but is simply a "command processing" time. Note that completion of a disk command does not necessarily imply host I/O completion because the I/O to a specific disk might be only a part of a larger I/O operation.
- Drive Queue Depth—This counter tracks the total number of requests that have been sent to the drive but not yet completed. It is incremented whenever a command is sent to the disk and decremented whenever a command completes.
- Read RPS—This counter tracks the number of read requests that have been sent to the disk drive. Because this counter is updated once per second, it translates directly into the read requests per second.

- Write RPS—This counter tracks the number of write requests that have been sent to the disk drive. Because this counter is updated once per second, it translates directly into the write requests per second.

EVA VDisk counters

The VDisk object tracks performance for each virtual disk (LUN) on the EVA. It is similar to the physical disk object, but it tracks virtual LUNs instead.

There is one instance of these counters for each virtual disk on the EVA. Each VDisk is uniquely identified by a four-digit hexadecimal number. This number is an internal representation of the LUN used by the EVA known as a "noid" and has no relationship to the LUN number.

- Read Hit Latency—This counter tracks the time taken from when a host read request is received until such time as that request has been satisfied from the EVA cache memory. The time, which is measured in microseconds, only applies to read commands that are satisfied from read cache. If the read command is a cache miss, the time is not tabulated here (see Read Miss Latency). Note that this value includes not only the latency from cache hits generated from random access activity, but also the latency associated with a cache hit as a result of a prefetch operation generated by a sequential read data stream.
- Read Miss Latency—This counter tracks the time taken from when a host read request is received until such time as that request has been satisfied from the physical disks. The time, which is measured in microseconds, only applies to read commands where the data is not in read cache and must be read from disk. If the read command results in the data being read from cache, the time is not tabulated here.
- Write Latency—This counter tracks the time, measured in microseconds, between when a write command is received from a host and when command completion is returned.
- Write RPS—This counter tracks the total number of write requests to a virtual disk that were received from all hosts. Because this data is updated once per second, it translates directly into write requests per second.

EVA storage cell counters

The storage cell object tracks information that is related to the overall storage system. It is a quick roll-up of several of the important metrics associated with overall EVA performance. There is only a single instance for these counters; this single instance represents the sum total of both controllers.

- Total host KBS—This counter tracks the total KB that has been read and written by all hosts connected to the EVA. Because this information is updated once per second, it translates directly into the total KB per second that the EVA is processing. Note that this is the sum of both read and write data.
- Total host RPS—This counter tracks the total number of I/O requests that have been issued by all hosts connected to the EVA. Because this information is updated once per second, it translates directly into the total requests per second that the EVA is processing. Note that this is the sum of both read and write requests.

For more information

- To learn more about award-winning industry hardware, visit the HP site at http://www.hp.com.
- For more information about the HP ProLiant servers, visit http://h18000.www1.hp.com/products/servers/.
- ActiveAnswers website
  http://www.hp.com/solutions/activeanswers
- HP Storage Solutions for Microsoft Exchange Server 2003
  http://www.hp.com/solutions/Microsoft/exchange/storage
- Microsoft references
  http://www.microsoft.com

5982-9586EN, 11/2004