Storage Performance Engineering

June 2003

invent

technical white paper

EVA Best Practices

Cost, Performance, and Availability

Table of contents

Introduction

One of the many design objectives for the HP StorageWorks Enterprise Virtual Array (EVA) program was to provide maximum real-world performance and to reduce the cost of storage management. This resulted in a design that minimized the number of user controllable options. This is in sharp contrast to traditional disk arrays, which typically have a plethora of tunable settings for both individual LUNs and the controller itself. Although tunable settings appear to be quite desirable at first glance, there are a few of problems associated with them:

- As the I/O workload changes, many of the parameters that were previously set may no longer be appropriate. Overcoming this problem requires continual monitoring, which is impractical.
- It is difficult for users to set the parameters appropriately. Many of the settings require in-depth knowledge of controller internals. Storage administrators don't have the time and resources to attain this level of expertise given their personal workload.

Because of concerns such as the above, the algorithms within the EVA were designed to reduce the number of parameters that could be set by the user, opting instead to embed heuristics and intelligence within the controller itself. The controller has a better view of the workload than most administrators and can be far more dynamic in responding to changes in the workload. The result is an array system that is both easy to configure and high performing.

Although the array is designed to work in a wide range of configurations, some of the configuration options may influence the performance, usable capacity or availability. Armed with a little information about the internal operation of the array, the storage administrator can control the configuration to maximize a prioritized attribute for a specific application.

A summary of these options are:

- Number of disk groups
- Type and number of disks in a disk group
- VRaid levels (0, 1, and 5)
- Disk failure protection level (none, single, and double)
- Cache settings

Each of these will be covered in more detail in the following sections. It should be noted that it is not always possible to simultaneously optimize a configuration for cost, performance, and availability. In the sections that follow, there may be contradictory advice, since one of the three choices must yield to demands made by another. As an example, VRaid 0 is clearly the best from a cost standpoint, since 100% of the storage is available for user data. It is also clear that from an availability standpoint, VRaid 1 is a much better choice, although storage utilization is only 50%. Other tradeoffs are minor by comparison, but must sometimes still be made. There is no "best" choice in these situations, since "best" depends on the needs of a particular environment.

Cost of ownership

The term "cost" as used in this paper refers to the cost of ownership of the storage. Simplistically, it is the cost per MB (or GB, TB, or more) of the entire storage subsystem. It is obtained by dividing the total cost of the storage by the useable data as seen by the customer. Items affecting cost of ownership will be examined one at a time, in no particular order.

Protection level

The protection level is reserved space used to rebuild the data on a failed disk A protection level of none, single, or double is assigned for each disk group at the time the disk group is created. Conceptually, it reserves space to handle 0 (none), 1 (single), or 2 (double) disk failures. The space reserved is specific to a particular group, and cannot span group boundaries.

The algorithm for reserving protected space is to find the largest disk in the disk group, double that size, multiply the result by 0, 1, or 2, and then remove that capacity from free space and distribute it across all disks in the disk group. Unlike traditional arrays, it does not reserve physical disks; all disks remain in use. The reason for finding the largest disk is that even though there may only be a few large disks in a group, they must be protected as well as the smaller ones, so space must be preserved. That size must be doubled, since the recovery algorithm for VRaid 1 failures involve transferring the data to two new disks; hence twice the capacity must be reserved.

As can be deduced from the preceding, having one or two large disks in a group means that "extra" space will be reserved. As an example, if a disk group of 168 disks consisted of all 36 GB disks with the exception of two 72 GB disks, a protection level of 2 would reserve 288 GB of space (72 GB * 2 * 2), even though the vast majority of the disks require only 144 of protection (36 GB * 2 *2). As such, cost considerations dictate that a disk group should consist of disks that are all the same size.

In addition to the preceding, VRaid 1 mirror partners can only use the capacity of the smallest disk in a mirror pair. If, for example, a 36 GB drive was mirrored with a 72 GB drive, only 36 GB of data could be used for mirroring on the 72 GB drive, potentially wasting large amounts of space.

Cost of ownership best practice: Do not mix disk sizes in a single disk group.

Since protected space cannot be shared across disk groups, having multiple disk groups can result in excessive protected space. As an example, consider a 2C12D EVA with 168 disks. With a single disk group and a protection level of 2, there will be a total of 4 disks worth of space set aside. With two disk groups, each group will have 4 disks worth of reserved space, or a total of 8 disks worth of space reserved. As more and more groups are added, the amount of reserved space increased, reducing the user data capacity, and increasing the effective cost of ownership.

Cost of ownership best practice: Use a single disk group.

Number of disk groups

Spare capacity is the space in a disk group that remains after space has been reserved for the protection level. It is used for creation of virtual disks, snapshots, and snap clones, as well as temporary space used by the EVA for certain internal operations. Spare capacity decreases when virtual disks, snapshots or snap clones are created, as well as when physical disks are removed. Spare capacity increases when virtual disks or snapshots are deleted, as well as when additional physical disks are added to the disk group.

Spare capacity is also used by the system in the event of a physical disk failure to reconstruct the data from redundancy information (VRaid 1 and VRaid 5 only). If there is spare capacity available, the system will use this before accessing the space reserved by the protection level.

Spare capacity, like protected space, exists within a disk group, and cannot be shared across groups. If there are two or more disk groups, this can lead to what is known as "stranded capacity". This happens when it is desired to create a LUN in a disk group, and although the total spare capacity of all disk groups in the EVA is sufficient for the creation of this LUN, there is insufficient capacity in any one of the individual groups to create it. The only solution for this is a massive reconfiguration of the storage, or the addition of disks to increase the spare capacity.

Cost of ownership best practice: Use a single disk group.

Disk quantity

Since a fixed price is paid for the EVA controller and supporting infrastructure, it makes sense to amortize this cost over as much storage as possible. To this end, it is reasonable to use as many disks as possible in a single EVA configuration.

Cost of ownership best practice: Fill the EVA with as many disk drives as possible.

Disk types

When looking at disks, larger disks usually offer better price per capacity. Although prices continuously change, it is reasonable to assume that at any point in time, you can purchase more capacity for the same price when using 72 GB drives than with 36 GB drives. Similarly, higher performance drives, such as 15K RPM drives, are generally more expensive than their lower performance 10K RPM counterparts.

Cost of ownership best practice: Use lower performance, larger capacity disks wherever possible.

Availability

Although the EVA is rated at five nines availability (99.999%), there are a few configuration guidelines that must be met in order to achieve this high level. In order to understand why certain guidelines are present, a discussion of the availability features of the EVA is in order.

Within a disk group, the EVA will create sub groupings of disks for redundancy purposes. Each of these individual redundancy sets contains sufficient redundancy information to allow continued operation in the event of a disk failure within that redundancy set. In this manner, the EVA can sustain multiple disk failures while not losing user data. These redundancy sets are built when a disk group is created, and additional sets are created as necessary by the controller firmware when disks are added to the disk group.

The target size of each redundancy is eight disks, with a minimum size of six, and a maximum size of eleven. As disks are added to a disk group, a redundancy set will be automatically be increased in size until such time as it reaches twelve disks. At that point, it will be split into two sets of six disks each. As more disks are added, one set will increase from six to eight (the target size), and the remaining set will then increase. After all disks have been added to a disk group, each redundancy set will contain eight disks, with the possible exception of the last set, which will contain between six and eleven disks. All of this is controlled by the EVA firmware, with no user intervention required.

In order to protect against loss of data access in the event of a shelf failure, there must be no more than one disk in a specific redundancy set per shelf. This requires a minimal amount of effort by the user at the time the disk group is created. The steps taken to accomplish this vary, depending on whether a LUN is configured as VRaid 1 or VRaid 5. If a mixture of VRaid 1 and VRaid 5 are contemplated for a single disk group, then the guidelines for VRaid 5 should be followed. If only VRaid 1 will be used, then the VRaid 1 guidelines are appropriate.

VRaid 5 availability best practice:

1. **There should be a minimum of 8 shelves in a configuration for VRaid 5.**
2. **All disks should be arranged in a vertical fashion, i.e., distribute the disks among the shelves such that the same bay in each shelf has a disk.**
3. **The total number of disks in a disk group should be an integer multiple of eight.**
4. **When creating a disk group, let the EVA choose which disks to place in the group.**

With VRaid 1, the EVA firmware will attempt to place the individual members of a mirror pair on different shelves. Because of this, the guidelines are much simpler, and there does not have to be a minimum of 8 shelves.

VRaid 1 availability best practice:

1. **All disks should be arranged in a vertical fashion, i.e., distribute the disks among the shelves such that the same bay in each shelf has a disk.**
2. **When creating a disk group, let the EVA choose which disks to place in the group.**

Replacing a failed disk

Although following the above rules will protect against loss of data access in the event of a shelf failure, there are specific steps that must be taken to help continue this protection after a disk fails. When a disk fails, the EVA will rebuild the failed disk data though a process known as "sparing". This sparing action will rebuild the original level of redundancy, but may place two members of the redundancy set on the same shelf. To restore the disk group to the original configuration, specific steps must be followed:

Availability disk replacement best practice:

1. **Wait for the sparing to be completed. This will be signaled by an entry in the event log (VCS versions 2.002 and above).**
2. **Remove the failed disk from the shelf and replace with a new one.**
3. **Add the new disk into the original disk group.**

Following this action, the EVA will initiate a leveling operation to evenly distribute data across all disks in the disk group, which implicitly results in restoring the original configuration.

When a disk is inserted into a shelf, there will be some transient activity on the back-end fire bus. In order to keep this from causing a false indication of excessive errors, insertion of multiple disks should be done carefully and slowly, with a pause between inserting disks.

Availability best practice: After inserting a disk drive into a shelf, wait 60 seconds before inserting another disk.

Number of disk groups

Although the EVA offers numerous levels of data protection and redundancy, a catastrophic failure can result in loss of a disk group. This is extremely unlikely, and requires multiple simultaneous disk failures of disks in the same redundancy set. In spite of this very low probability, installations that demand the ultimate in data availability might consider creating two separate disk groups. Although two groups will result in a slightly higher cost of ownership and potentially lower performance, the increase in availability may be the right decision for a very high availability application.

In order for two disk groups to prevent data loss, each disk group must contain sufficient independent information to reconstruct the entire data set. A practical example of this is a database that contains both data and log files. In this instance, placing the data files in one group and duplexing the log files (a typical feature of the database) to both the data file group and another group ensures that loss of an entire disk group will not prevent recovering the data. If there are multiple databases, then alternating this placement will ensure an even distribution of data for both capacity and performance balancing. As an example:

Disk Group 1: Database 1, DB 1 log files, DB2 log files

Disk Group 2: Database 2, DB 2 log files, DB1 log files

In the above configuration, each database is protected, since failure of either disk group will leave sufficient information to recover both databases[1]. Additionally, the capacity is probably split fairly equally between the two groups, so performance should be relatively well balanced also.

Availability best practice: For critical database applications, consider placing data and log files in separate disk groups.

Performance

As might be expected, performance and price are generally incompatible. The underlying virtualization technology of the EVA goes a long way towards masking this, but there are still cases where higher performance comes at the cost of both price and availability. The following sections explore how to obtain the best performance, albeit sometimes at the expense of both cost and availability.

Disk count

Since most random access applications are limited by the physical disk speeds, increasing the numbers of disks under a LUN will translate directly to an increased performance potential. With the high transfer rates of modern disk drives, maximum sequential performance can be attained with only a few disks. Having additional disks does allow multiple sequential streams, or even intermixed random and sequential streams to co-exist with minimal interaction, so for most applications, there will be a direct relationship between the number of disk drives and I/O performance.

Performance best practice: Fill the EVA with as many disk drives as possible.

Number of disk groups

For typical workloads, an increased number of disk drives under a LUN imply increased performance potential, and since a LUN can only exist within a single disk group, it follows that having a single disk group maximizes the performance capability. Similarly, the larger numbers of disk drives associated with a single disk group imply less interaction among multiple I/O streams, as well as the ability to "share disk resources" as the I/O load shifts from one LUN to another.

Performance best practice: Use a single disk group.

[1] A loss of the disk group with the data files will require a reload from a backup, but in both cases, the database will be consistent and current.

Disk RPM

For applications that perform large block sequential I/O, such as data warehousing and decision support, disk RPM has little or no effect on performance. As such, large capacity 10K RPM disks make the most sense.

For applications that issue small block random I/O, such as interactive databases, file and print servers, and mail servers, higher RPM disk drives offer a substantial performance advantage. Workloads such as these can see gains of 30% to 40% in the request rate when changing from 10K to 15K RPM disks. Although it seems contradictory to use 10K RPM disks for better performance in these circumstances, there are instances where it may make sense.

Although not guaranteed, it is a likely assumption that 15K RPM disks cost more than the equivalent capacity 10K drives. Since the gain from a 15K drive is in the range of 30% to 40%, if the 15K drives are more than 30-40% more expensive than the 10K drives, then it makes sense to purchase a larger number of 10K drives. As an example (and remember, this is only an example), consider a system with a total budget of $100K. The goal is to get the best performance possible for this amount of money. The results are shown in Table 1, with an assumed cost for disk storage, and where the "IOPS" field represents the request rate at a response time of 30 ms for a typical I/O workload.

Table 1 – Disk RPM Tradeoffs

Disk type	IOPS	Cost	$/IOP	Disk count at $100K budget	Total IOPS at $100K budget
72 GB 10K RPM	110	$2,300	$20.91	43	4,730
72 GB 15K RPM	150	$3,400	$22.67	29	4,350
Improvement	15K = 36%				10K = 9%
Best choice	**15K**	**10K**	**10K**		**10K**

As can be seen, although a 15K RPM disk offers higher performance than the equivalent capacity 10K RPM disk, the lower cost of the 10K disk means that more can be purchased for the same total amount, and the increased number of disks translates directly into higher overall performance. Although the above is clearly a contrived case to illustrate a point, it should also be clear that there are times when a "lower performance" disk such as a 10K RPM one can, for the same dollar amount, yield not only higher performance, but also more capacity (with an attendant lower cost of ownership).

Performance best practice: Using 15K RPM disks is generally best, but carefully consider cost and quantity tradeoffs between 10K and 15K RPM disks.

Cache mirroring

The purpose of cache mirroring is to prevent data loss in the unlikely event of a cache board failure. In operation, writes from the host are sent to the controller that has the LUN online. That controller will store the write data in its non-volatile cache, and then send that data across a 2 Gb mirror port to the other controller. The second controller will store a copy of the data in its cache, and then return completion to the original controller. The original controller will then signal completion of the request to the host. Since there are two independent copies of the data maintained in separate caches, failure of a cache memory board does not result in loss of the data.

Because there is a data copy operation involved in this mirroring operation however, there is an impact on the performance of writes. The amount of the impact depends on many factors, including the percentage of write operations, the size of the write requests, and the overall write request rate as presented by the various hosts.

From a performance perspective, it is possible to obtain significant gains in performance when a LUN is created with cache mirroring disabled. The clear disadvantage of disabling cache mirroring is, of course, the possibility of data loss if a cache board that contains data yet to be written to disk fails. There are applications that are not concerned with this possibility, such as databases that are reloaded every night, or other applications where loss of write data is a secondary concern in relation to performance, and these applications may be candidates for disabling write cache mirroring.

Performance best practice: Under certain, carefully considered circumstances, disabling write cache mirroring will result in significantly increased write performance.

Mixed disk speeds

Although it is esthetically preferable to have all disks in a group be of the same type, there are instances where mixed drive speeds may be unavoidable. From a performance standpoint, 15K RPM disks will perform 30% to 40% faster than 10K RPM disks with a small block, random access workload. Mixing drive speeds within a single group would then result in host level performance that will vary depending on which drive processes the I/O request. It is important to note that even when mixing drives of different speeds in the same disk group, the disk group does not slow down to the speed of the slowest drive in the group.

Although variations in performance are generally not desirable, the overall performance of a LUN is dependent on the number of physical disks underneath that LUN. If drives of differing speeds are separated into different disk groups, then it follows that LUNs in each disk group will have fewer drives. As such, the performance of each LUN will be lower that it would have been with all drives in the same group. Since there is almost always some amount of I/O load imbalance between different LUNs, the total performance of two LUNs in separate disk groups will be less than the total performance of those same two LUNs in a single large disk group.

Performance best practice: Drives with different performance characteristics may be placed in the same disk group, which will result in higher system performance than if they were in separate disk groups.

Mixed disk capacities

In a manner similar to disk speeds, disks with different capacities would seem to be better placed in separate disk groups. As with the preceding advice however, more disks will usually produce better performance, so placing disks of differing capacities in a single disk group will result in overall higher performance of the LUNs in that group.

There is a minor issue dealing with leveling on the EVA however. The EVA will attempt to ensure that the amount on each physical disk drive is proportional to that drive's contribution to the overall capacity. This means that larger drives will have more data on them than smaller drives. As an example, a 72 GB disk will have twice as much user data on it as a 36 GB drive. In a random access type of application, this implies that the larger drives will have twice as much I/O as the smaller drives, resulting in an I/O load imbalance at the disk drive level.

Although more I/O to a drive implies a higher response time, this must be weighed against the fact that for a given I/O load, more drives under a LUN equate to a lower per-drive I/O rate. This is important when there is a load imbalance between LUNs, since one LUN in a disk group may be at the maximum sustainable I/O rate, while a LUN in a different disk group might be operating at a fraction of its potential.

As an example, assume that both 36 GB and 72 GB disks are capable of 150 requests per second at a reasonable response time. If an EVA were configured with 84 of each of these drives in a separate disk group, then a single LUN in a single disk group would be capable of 84 * 150, or slightly over 14,000 IOPS. Although the total of both groups would be twice that value, this would only happen if there were a perfect balance of I/O requests between the two groups. Since this rarely happens in practice, combining all disks in the same group would allow the I/O rate of a single LUN to increase over that of a LUN in a single, smaller disk group. The total I/O rate may not be doubled, but the overall realizable rate would be much higher than that of two smaller disk groups.

Performance best practice: Drives with different capacities may be placed in the same disk group and will usually result I higher performance than if they were in separate disk groups.

Read cache

One of the parameters that may be set at either LUN creation or dynamically at a later date is read caching. The parameter affects both random access read caching and sequential (prefetch) caching, although the algorithms and cache usage are completely different.

Both algorithms are designed to come into play only when they will have a positive effect on performance. Random access caching will be enabled and disabled automatically as the I/O workload changes, while prefetch caching will only come into play if a sequential read stream is detected. Because of this dynamic response to changing I/O workload conditions, cache efficiency is maintained at a high level, and there will be no negative impact on either cache usage or performance in the presence of an I/O workload that is "cache unfriendly".

Since there is no negative impact of leaving cache enabled, and there is always a chance of a performance gain through caching, read cache should always be left enabled.

Performance best practice: Always leave read caching enabled on a LUN.

LUN balancing

Although both controllers in an EVA can access all physical disks, a LUN is online to only one controller at a time. Because of this, a single LUN can only use the cache of a single controller, may be constrained by the processing capability of a single controller, and can only use two out of the four host ports on the EVA. As such, it makes sense to ensure that each controller in an EVA pair has an equal share of the I/O load.

Although the default controller preference is established via the element manager, this is only a preference, and is usually controlled instead by the host operating system. Because of this, attention should be paid at the operating system level to ensure that the I/O load on both controllers is reasonably balanced.

Performance best practice: Always attempt to balance LUNs between the two controllers on an EVA based on the I/O load.

Summary

All of the preceding recommendations can be summarized in tabular format. This not only makes it relatively easy to choose between the various possibilities, it also highlights the fact that many of the "best practice" recommendations contradict each other. In many cases, there is no correct choice, since the best one depends on what the goal is; cost, availability, or performance. Note also that in some cases, a choice has no impact.

Table 2 – Best Practices Summary

	Cost	Availability	Performance
Mixed disk capacities in a disk group	No	No	Acceptable
Number of disk groups	1	1 – 2	1
Number of disk in a group	Maximum	Multiple of 8	Maximum
Total number of disks	Maximum	Multiple of 8	Maximum
Higher performance disks	No	–	Probably
Write cache mirroring	–	Yes	Maybe
Mixed disk speeds in a disk group	–	–	Acceptable
Read cache	–	–	Enabled
LUN Balancing	–	–	Yes