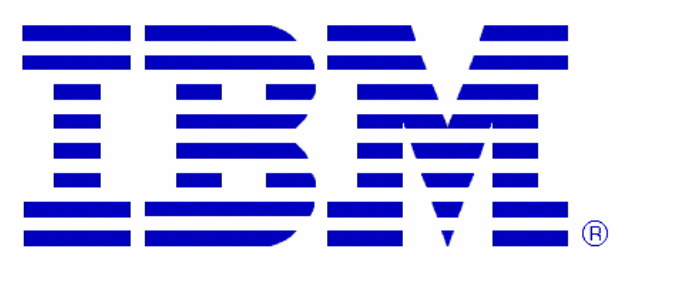

IBM® System Storage™ DS8700™ and DS8800™ Performance with Easy Tier® 2nd Generation

July 2011

Joshua Martin
Nick Clayton
Lee La Frese
Kaisar Hossain
Bruce McNutt
Yan Xu

Document WP101961

Systems and Technology Group

Notices, Disclaimer and Trademarks

Acknowledgements

The authors would like to thank the following colleagues for their comments and insight:

Andrew W. Lin – IBM Systems & Technology Group, Tucson, AZ.

Paul Muench – IBM Research, San Jose, CA.

David Sacks – IBM Systems & Technology Group, Chicago, IL.

Brian Sherman – IBM Sales & Distribution, Markham, Ontario Canada.

Alisair Symon – IBM Systems & Technology Group, Tucson, AZ.

Sonny E. Williams – IBM Systems & Technology Group, Tucson, AZ.

A Note to the Reader

This White Paper assumes a familiarity with the general concepts of Enterprise Disk Storage Systems and the DS8000 product line. Readers unfamiliar with these topics should consult the References section at the end of this paper.

Table of Contents

Acknowledgements 3
A Note to the Reader 3
Table of Contents 4
1 Introduction 5
1.1 Audience 5
2 Overview 6
2.1 New Features 6
3 Easy Tier Performance Results 6
3.1 DB2 Brokerage Transactional Workload Performance 6
3.2 Simpler and More Cost Effective Growth 11
3.3 CPW Workload Performance with IBM i 14
3.4 Intra-tier Auto Rebalance 15
4 Easy Tier Tools 17
4.1 Storage Tier Advisor Tool 17
4.2 Disk Magic 21
5 Real World Workload Analysis and Insights 21
5.1 Easy Tier Data Collection 22
5.2 Easy Tier analysis 22
5.3 Workload Skew 24
5.4 Variation over time 29
6 Best Practices 31
7 Conclusions 32
8 References 33
9 Appendix 34
8.A Appendix A: Workload Characteristics 34
8.B Appendix B: DS8000 Hardware Configurations 35

1 Introduction

Easy Tier was first introduced on the DS8700 Storage System in 2010 with the Licensed Internal Code (LIC) Release 5.1. This innovative feature provides a simpler method for managing the capacity and capability of Solid State Drives (SSD) than the manual options available prior to its release. Through automated migration of data, Easy Tier Automatic mode allows the storage system to intelligently manage its own performance capability by learning from the behavior of the host workload and identifying data that would benefit most from the performance capacity of SSDs.

This paper examines the performance of the second generation of the Easy Tier feature which is available with LIC Release 6.1 for no additional charge on both the DS8700 and the DS8800. In addition to expanding support of the Easy Tier feature to the DS8800, this update also splits Hard Disk Drives (HDDs) into two tiers allowing for more granularity of performance management. The auto-rebalance functions was also added which further simplifies storage management with Easy Tier by ensuring that data is well balanced for performance within every managed tier. These two new functions make Easy Tier a cost effective option for managing storage performance and growth.

A DB2 Brokerage environment was used to compare Easy Tier 2nd generation on the DS8700 and the DS8800 to the initial release of Easy Tier on the DS8700. Comparing the two generations of Easy Tier showed a transaction rate improvement of 40% on the DS8700. Through utilization of its enhanced device adapters, a similarly configured DS8800 running Easy Tier 2nd generation showed an additional transaction rate improvement of 60% over a DS8700 also running Easy Tier 2nd generation. Online Transaction Processing environments were used to examine Easy Tier performance with the new HDD tiers and to examine the Auto Rebalance function.

In addition to considering benchmark environments, Easy Tier data was also collected from several existing real world environments. Using that data, this paper examines the design of Easy Tier with consideration for its performance. Easy Tier was observed to perform notably well compared to how the data could have been placed with the benefit of hindsight.

This white paper focuses only on the functions in Easy Tier automatic mode.

1.1 Audience

This technical paper was developed to assist IBM Business Partners, field sales representatives, technical specialists, and IBM's clients in understanding the performance characteristics of the IBM 2107 Model 941 and the IBM 2107 951 with the updated Easy Tier feature. The IBM 2107 Model 941 is the DS8700, POWER6 model and shall be referred to as the DS8700 throughout this paper. The IBM 2107 Model 951 is the DS8800, Power6+ model and shall be referred to as the DS8800 throughout this paper. The update to the Easy Tier function may be referred to as either Easy Tier 2nd Generation or as Easy Tier 2, while the initial release of Easy Tier may be referred to as Easy Tier 1.

2 Overview

2.1 New Features

Easy Tier 2nd generation adds two new functions. First, it adds the concept of "any two tiers" to the Easy Tier paradigm. Second, it introduces an auto-rebalance feature for Easy Tier managed storage pools.

Previously, Easy Tier focused on pools containing a mixture of Solid State drives (SSDs) and traditional hard disk drives (HDDs). With the new release, hard disk drives are separated into two tiers: "Enterprise Tier" which covers 15K and 10K RPM disk drive varieties and "Nearline Tier" (also referred to as SATA Tier) which covers 7.2K RPM disks. SSDs constitute a third tier. Easy Tier can now manage any combination of two of these three tiers. Thus, Easy Tier may now manage a storage pool containing SSD + Enterprise drives, or SSD + Nearline drives, or Enterprise + Nearline drives.

The auto-rebalance function automatically balances the load on the RAID ranks over time to continuously maintain excellent performance despite variations in workload. At this time, auto-rebalance will only apply to Easy Tier managed storage pools. It focuses on balancing the load on ranks within each tier.

3 Easy Tier Performance Results

3.1 DB2 Brokerage Transactional Workload Performance

A DB2 Brokerage Transactional workload was used to evaluate the application performance improvement available with Easy Tier 2nd Generation and a POWER7 host server. This lab experiment was designed to simulate a class of applications that facilitate and manage transaction-oriented business processes and are commonly used in a broad range of industry segments including finance, retail, and manufacturing. These applications are characterized by having over 90% read hits in server memory (overall buffer pool hit ratio can be near 97%) and highly random disk I/O. Since it is highly random, this is an ideal workload for evaluating Easy Tier.

The results from these lab experiments clearly demonstrated how customers may benefit from the Easy Tier feature with similar types of transactional applications. The lab experiments also showed that the Easy Tier feature enhances the ease of storage management and offers dramatic improvements in both storage and application performance in real world customer environments without disruption to applications.

Workload Configuration:
DB2 9.7 SP1 on a POWER7 p770 server was used with backend storage DS8800/DS8700 running Release 6.1 licensed internal code. A total of 4 DB2 instances were set up, each with a 2 TB database and 54 GB Buffer Pool allocated. All four instances (total of 8 TB) ran the DB2 Brokerage workloads simultaneously with clients on a single LPAR of the POWER7 server.

To illustrate the behavior of the system with Easy Tier 1st generation, two DB2 Brokerage workload intensities were evaluated: Typical and Peak I/O intensities. A default HBA queue depth value of 20 was used to evaluate a typical I/O intensity, while a much higher queue depth value of 256 was set to evaluate the peak I/O intensity workload. In those experiments, the Overall Transaction Rate (OTR) was observed to have improved by 41% with the typical I/O intensity and 35% with the peak I/O intensity workloads with Easy Tier 1. In this paper, the experiments considered the peak I/O intensity workloads only.

Baseline measurements (tests without Easy Tier active) were taken using HDDs only and compared to Easy Tier measurements which used a pool of HDDs and SSDs. Each Baseline test was run using a 30 minute ramp up time (RUT) period for the workload plus an additional 6 hours of steady state run time. The Easy Tier experiment used two ranks of SSDs in addition to the HDDs used in the Baseline test. Similar to the baseline experiment, for each Easy Tier test 30 minutes RUT was used, but 14 hours of steady state run time was needed for Easy Tier learning and data movement to reach a steady state.

DS8800/DS8700 Configuration:
The baseline experiment was run with 13.2 TB of physical storage capacity attached to two Device Adapter (DA) pairs on a single DS8800 or DS8700. The baseline storage capacity consisted of 128 146 GB 15K RPM HDDs configured as 16 RAID-5 arrays. The Easy Tier experiment used 16 300 GB (DS8800) or 16 600 GB (DS8700) SSDs in addition to the Baseline configuration (detailed configuration information is available in Appendix B, 8.B.1). This experiment was designed to demonstrate the relative performance of Easy Tier, not to show the maximum capabilities of the systems tested.

The Easy Tier function was set to run in automatic mode so that extents in the databases were moved dynamically based on both backend disk access frequency and accumulated backend disk latency. In order to reduce experiment duration, some of the default Easy Tier settings were changed; the short-term learning window was reduced from the default 24 hours to 2 hours, the data migration rate was changed from the default of 12 extents transferred every 5 minutes to the fastest rate possible.

DB2 Brokerage Workload Performance with Peak I/O Intensity
The Overall Transaction Rates of the DS8700 and DS8800 baseline runs are shown in Figure 1. A significant improvement of 29% in transaction rates was observed in the DS8800 when compared with the DS8700 due to the enhanced POWER6+ processors and Device Adapters in the DS8800. For both of these runs, LUN queue depth was set to 256 to create peak I/O Intensity workloads.

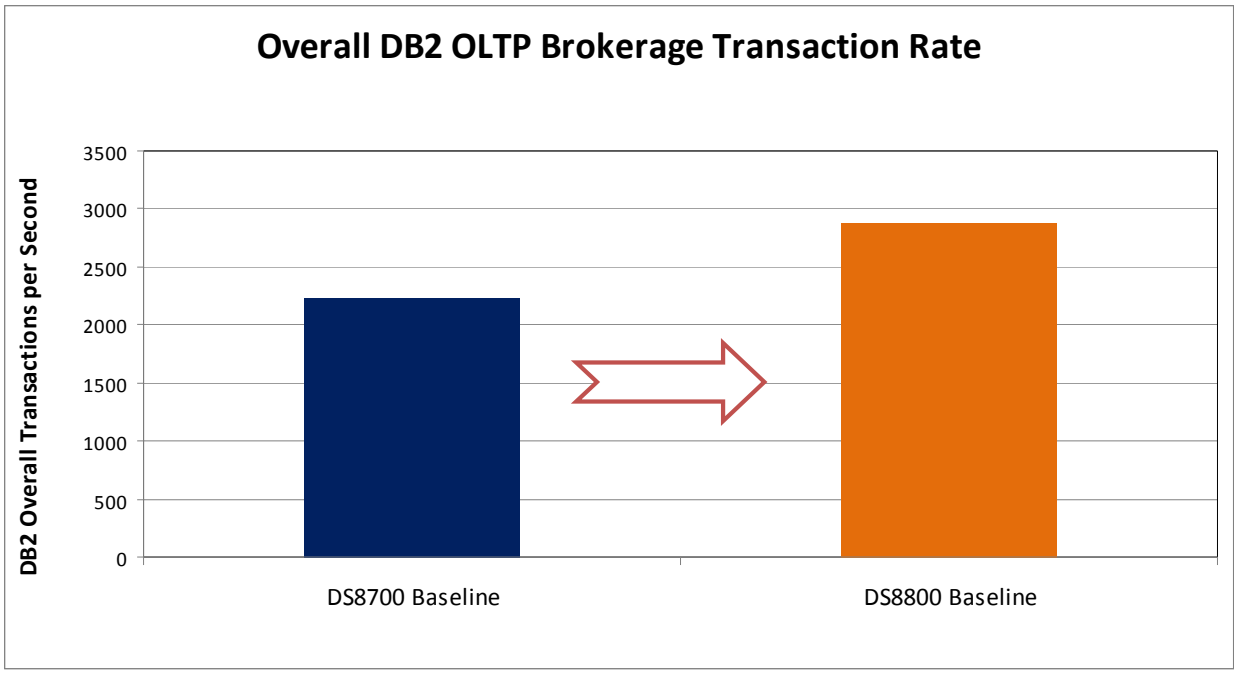

Figure 1. DS8700 and DS8800 Baseline runs

Performance improvements from Easy Tier 2 with a DS8800 are demonstrated in Figures 2-6. After Easy Tier data movement became stable between the SSD and HDD tiers, the OTR was improved by more than 2x for the peak I/O Intensity workloads when compared with the baseline run (Figure 2). Even with increased throughput, the Overall Transaction Response Time improved by 58% as seen in Figure 3.

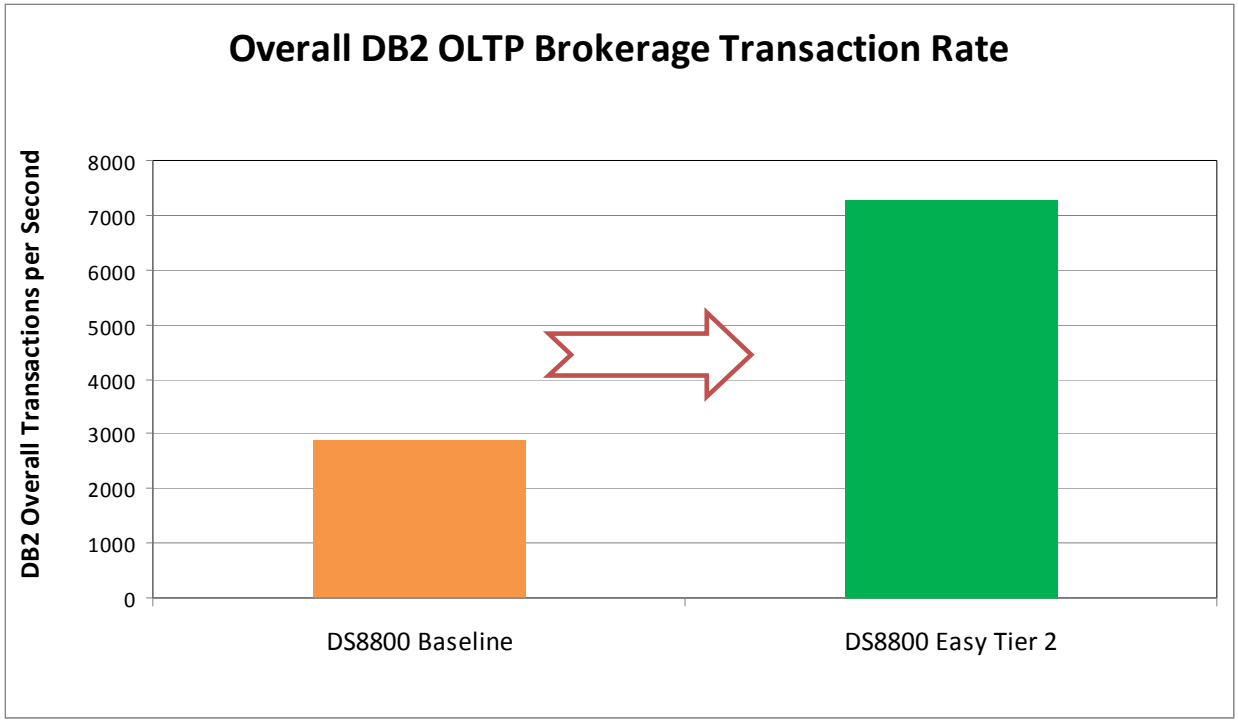

Figure 2. DS8800 Baseline and Easy Tier 2 runs

Weighted Average RT (ms)				
Trade Activity	Lookup	Order	Update	Overall Transaction
DS8800 Base	9666	290	13332	1169
Easy Tier2	3441	198	5883	493
Benefit (%)	**64.40%**	**31.72%**	**55.87%**	**57.83%**

Figure 3. Weighted Average Overall Transaction Response Time

Similar improvement was seen in the DS8800 IO activities. Figure 4 shows more than 2x increase in throughput with Easy Tier 2, while the response time was reduced by 69%.

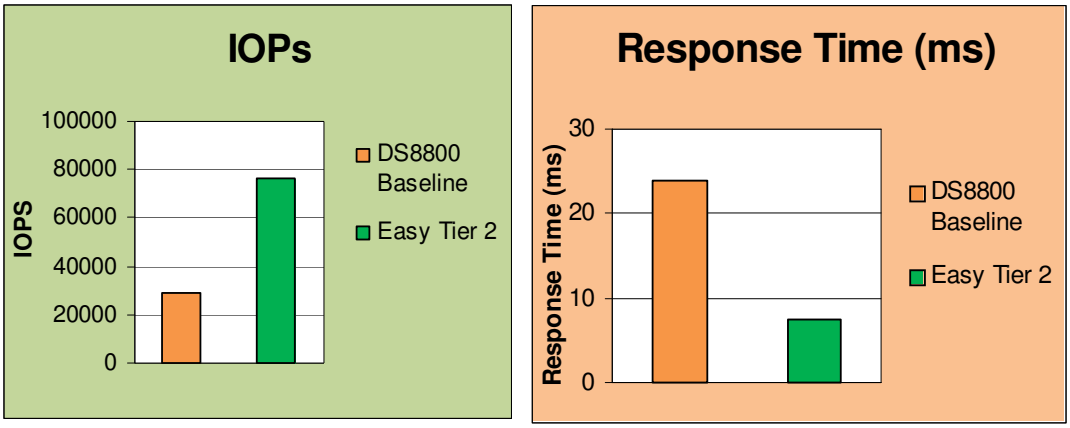

Figure 4. DS8800 Storage Throughput and Response Time

The POWER7 server played a key role in our testing by improving overall system performance allowing us to leverage performance per core, improved processor utilization, and dynamic infrastructure. Figure 5 shows CPU usage increased by 2.4X with Easy Tier 2 and POWER7 giving improved throughput and response time.

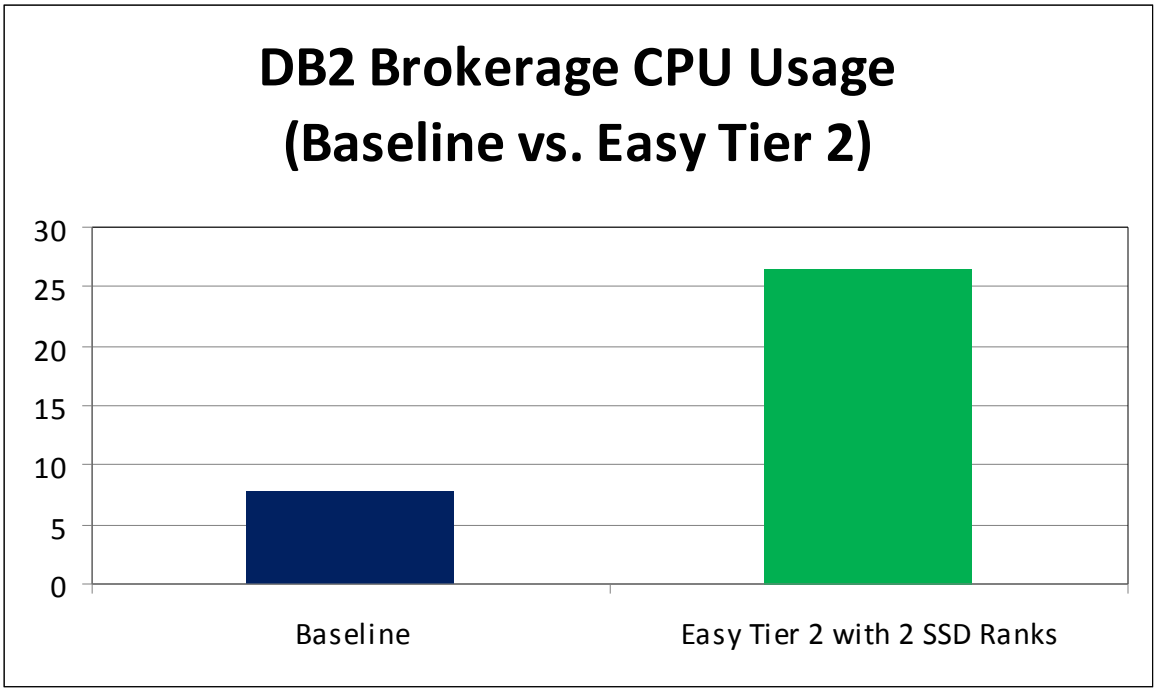

Figure 5. CPU usage at POWER7 Server: Easy Tier with DS8800

A distinct improvement was observed when we compared the lab results of the run with Easy Tier 1 on the DS8700 to the run with Easy Tier 2 on the DS8700 and to the run with Easy Tier 2 on the DS8800 (Figure 14). Two SSD ranks were used for Easy Tier with all three experiments. The DS8700 with Easy Tier 2 benefitted from better performance of the 600 GB SSDs to give the improvement over the previous run of the DS8700 with Easy Tier 1. Faster Device Adapters

combined with better performing 300 GB SSDs further improved Easy Tier performance on the DS8800 over the DS8700.

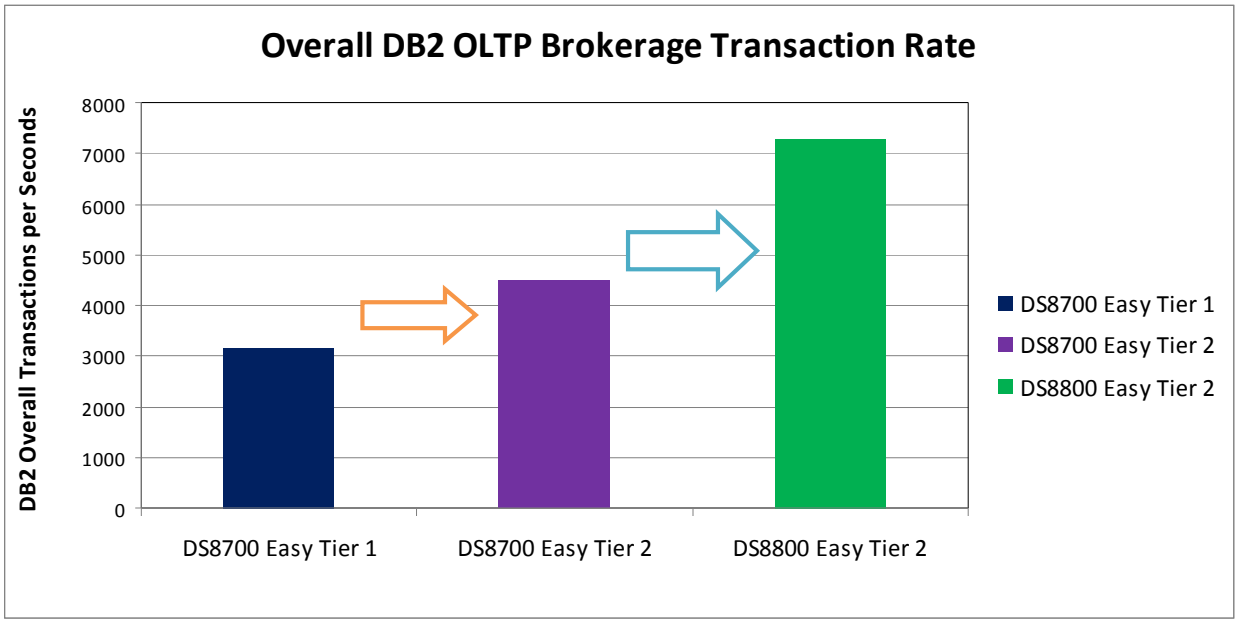

Figure 6. Measurement between Easy Tier 1 (DS8700), Easy Tier 2 (DS8700) and Easy Tier 2 (DS8800)

3.2 Simpler and More Cost Effective Growth

In the past, the introduction of a new, high-capacity disk technology typically involved a full replacement of the preceding disk technology.

In some installations, an effort has historically been made to reduce costs by maintaining a mix of disk technologies, while managing data placement so as to match applications with the appropriate technology. The effort required for such management, however, is typically very significant and tends to undercut the ability to achieve the desired cost savings.

By contrast, the net effect of the new release of Easy Tier is that growth need no longer occur via full replacement of one disk technology with another, but instead can play out in a more evolutionary manner. This results in a mix of disk technologies, one in which automation has eliminated the manual effort previously needed to maintain such a mix.

The ability to adopt an evolutionary strategy for growth is a central benefit of Easy Tier, since it offers not merely lower storage costs but also simpler planning and administration. Within this framework, it is easy to consider the idea of adding new, high density disk storage resources in response to the demands of new, storage intensive applications. This can be accomplished simply by expanding an existing storage pool with new high density disks that are suitable for the intended applications. The physical placement of the data can be allowed to occur automatically, based upon the I/O requirements seen in each storage region.

When expanding a storage system by adding new disk resources, it may also be desirable to add cache memory and/or host paths at the same time. Typical cache sizes today are large enough to accommodate large variations in cache demand. Nevertheless, the addition of new storage-intensive applications may impose new I/O requirements and the new applications may sometimes be more random (less "cache friendly"). Added cache memory may help to ensure that the average residency time of data in cache does not fall due to the new I/O requirements. Since the new applications are not typically expected to increase I/O proportional to their capacity requirements, we do not recommend that the size of cache be increased in proportion to the increase in capacity. As a very rough guideline, consider retaining the same proportional relationship between the amount of cache memory and the number of disk spindles.

In addition to cache concerns from new I/O requirements, batch processing requirements may also increase when expanding a storage system. Adding host paths may help multiple distinct applications to complete their batch processing without needing to compete for path resources. For this reason, consider retaining the same proportional relationship between the number of paths from the storage to the SAN, relative to that from the SAN to the various application servers.

The growth scenario just described sharply reduces the cost of introducing new, more storage intensive applications. The same process also reduces the average access density of the application data. At the system level, the end result is that Easy Tier manages a mixture of data with higher and lower access densities, through intelligent placement of data on the available disk resources.

To give a simple “before” and “after” comparison for this kind of growth scenario, two performance experiments were completed using an Online Transaction Processing (OLTP) benchmark. Overall, the I/O content of the OLTP benchmark consisted of 60 percent writes and 40 percent reads, with an average transfer size of approximately 8 KB. More importantly, the OLTP benchmark incorporated a mix of access densities seen in various areas of the storage, allowing it to represent the growth scenario as just described.

Figure 7 presents the results of the comparison. The “before” test (light blue curve) was done using a relatively smaller database stored on 16 RAID-5 arrays of 15K RPM, 300 GB drives. The “after” test (pink curve) was done using a relatively much larger database stored on the same 16 arrays, plus an additional 8 RAID-10 arrays of Nearline (NL), 2 TB drives (Nearline drives are available on DS8700 only with Release 6.1 LIC code).

In moving from the “before” to the “after” case, the high density of the added near line drives caused the total database size to increase by a factor of more than two times. No cache or host paths were added to the system.

Figure 7 makes clear that we have accomplished a very effective result while taking advantage of the lower cost per unit of storage possible with near line technology. In the “after” case, we have delivered not only a large increase in capacity at a low cost per unit of storage, but also, we have done so while continuing to deliver high system throughputs. A similar combination of low

cost per unit of storage and high system throughput capability can not be achieved by simply replacing the existing 15K RPM disks with similar aggregate capacity provided by larger but slower disks (dark blue curve) nor by adding NL drives used in a random manner without taking the I/O requirements of individual data extents into account (purple curve).

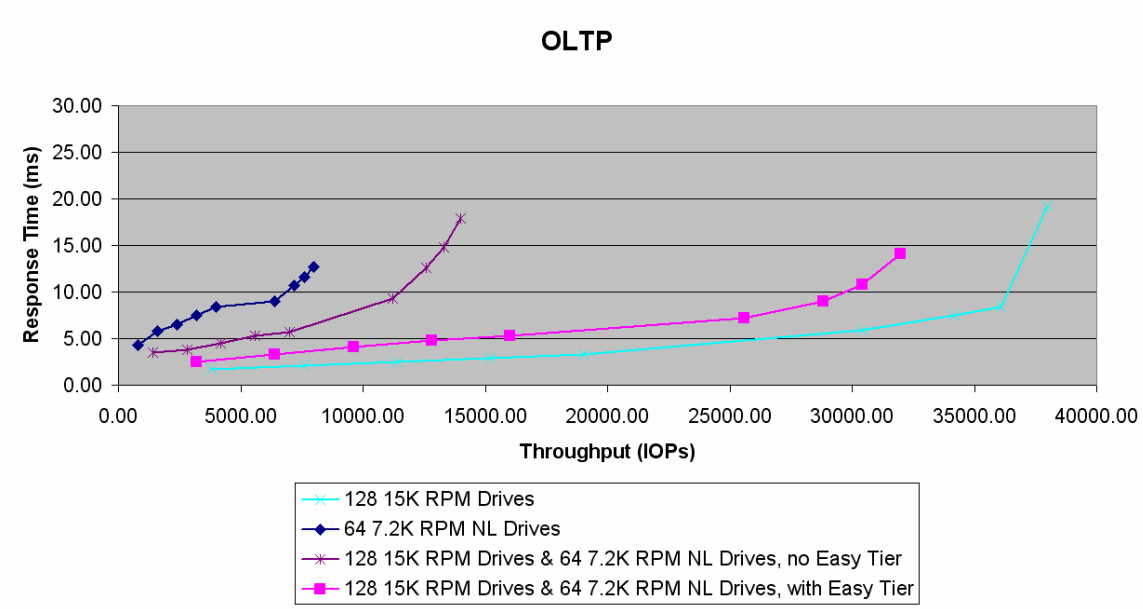

Figure 7. Capacity Growth

Figure 8 shows IO redistribution between the 15K RPM Enterprise tier and the NL Tier by Easy Tier. Over time, Easy Tier moved extents with higher access density to the 15K RPM Enterprise tier and extents with lower access density to the NL tier. These data movements resulted in more IOs on the 15K RPM tier and fewer IOs on the NL tier. For this particular experiment, some internal Easy Tier settings were changed in order to shorten the test cycle. For example, the short-term learning period was reduced from the default 24 hours to 1 hour and data migration rate was changed from the default 12 extents every 5 minutes to the maximum allowed by the system.

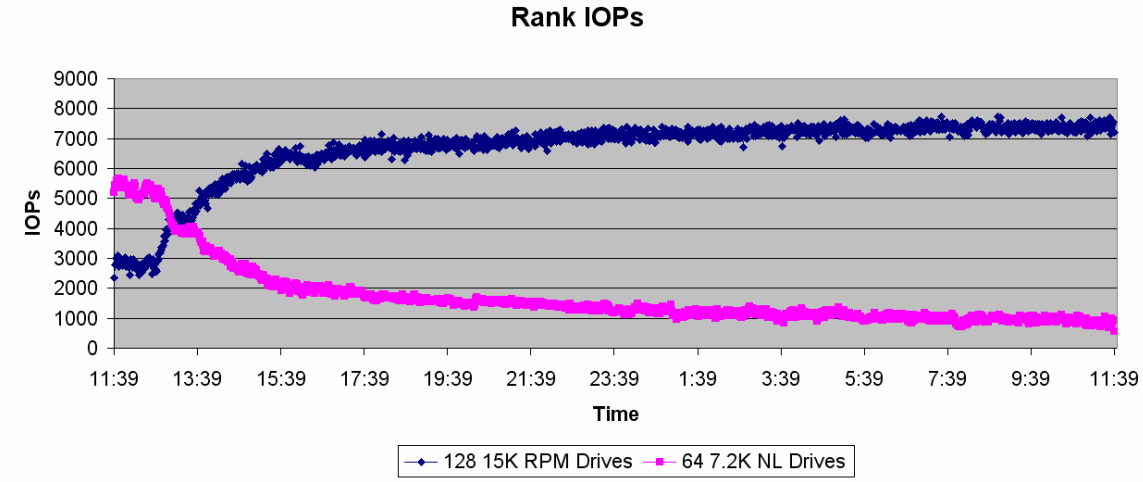

Figure 8. IO Redistribution by Easy Tier

3.3 CPW Workload Performance with IBM i

This section reviews performance results using the Commercial Processing Workload (CPW) running on the IBM i operating system attached to a DS8800 with SSDs. CPW is characterized by many jobs running brief database transactions in an environment that is dominated by IBM system code performing these database operations. CPW is used by IBM for evaluation of different IBM i processor models where the primary application is oriented toward traditional commercial business uses (order entry, payroll, billing, etc.) using commitment control. Although the CPW workload is typically used to represent the relative performance of a complex of processors, it does a significant number of read and write disk accesses making CPW useful in disk performance characterization as well.

For the purposes of this study, the storage was broken into two Auxiliary Storage Pools (ASPs). The database ASP features small, random front-end IOs (typically 8 to 12 KB in size) that are approximately 60% reads. These IOs typically are coalesced into 60 to 100KB operations on the back-end disk. The journal ASP was made up of practically all sequential writes, typically 60 to 120KB in size. About two-thirds of the available storage was allocated to the database ASP and one-third was used for the journal ASP. A larger number of smaller LUNs were used in the Database ASP to increase the overall queue depth. See Appendix B (8.B.3) for more details on the storage configuration used.

There are several ways to place the most frequently accessed data on SSDs with a combination of IBM i and DS8800. ASP Balancer is an IBM i operating system feature that moves data to SSDs based on the read I/O count statistics of each 1 MB auxiliary storage extent of the collection ASP. After the ASP Balancer job completes, “cold” data is moved off the SSDs and “hot” data is moved on. In some ways it is similar to Easy Tier but at the operating system level. DB2 Media Preference allows users to choose which media type selected database files will be stored on. In this case targeted data is placed directly on SSD’s and there is no automated data movement as opposed to Easy Tier, which is a storage-based optimization method.

Figure 9 below shows CPW performance using all three methods compared to a baseline run without SSDs. In this experiment all three methods provided significant improvements compared to the baseline. ASP Balancer provided a 33% increase in throughput while DB2 Media Preference delivered a 40% increase. When run with Easy Tier, the improvement was 34% which is very similar to the ASP Balancer result. DB2 Media Preference relies on application knowledge to decide which databases should be placed on SSDs so it is not surprising that it enabled the best performance. Both ASP Balancer and Easy Tier are automated methods. In general, specific application knowledge usually contributes to better results with SSDs than any automated method.

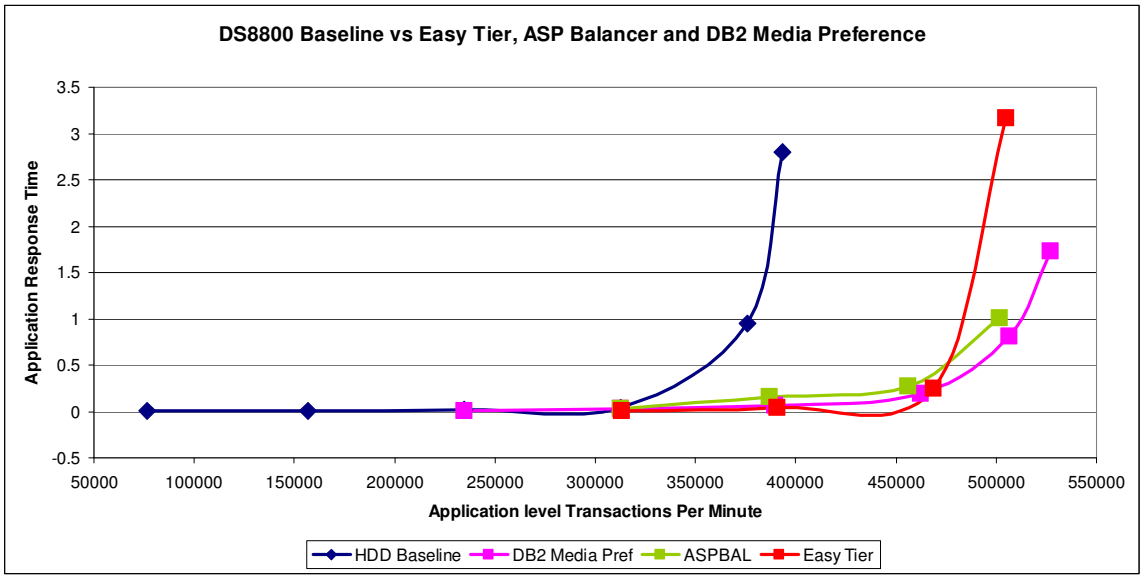

Figure 9. IBM i CPW Results with DS8800 and SSDs

3.4 Intra-tier Auto Rebalance

A new function added in Easy Tier 2 is Intra-tier Auto Rebalance. This function attempts to make the best use of the drives available in one tier of a two-tier extent pool by redistributing the data in that tier. This new function will balance I/Os across available ranks automatically for skewed workloads or when ranks are added or removed from the extent pool.

Performance test runs were made with a small two-tier pool consisting of 1 SSD rank and 3 15K RPM HDD ranks on a DS8800. In order to test the new function, 2 additional 15K RPM HDD ranks were added to the pool while a transactional workload (DB Open) was running to the existing 3 15K RPM HDD ranks. Since the experiment was executed in a lab environment, some of the default settings were changed to reduce test time: the learning period for auto rebalance was reduced from 6 hours to 30 minutes and the migration rate was set to the fastest allowed setting. Since the short term learning period for Easy Tier was left at its default of 24 hours, no data was moved to the SSD tier within test duration. This allowed the results to focus on the performance effects of the auto rebalance function in the 15K RPM HDD tier.

In Figure 10 the activity on all 5 HDD ranks is shown over the course of a 5 hour run. Shortly into the run, 2 unused ranks were added into the pool, after which the learning period begins. As expected, at the start of the run, the 2 added HDD ranks had no I/Os and no capacity allocated since all of the data and I/Os were on the 3 original HDD ranks. After the initial learning period, data migration began and the lines representing original ranks and newly added ranks began sloping toward one another. That continued steadily until balance was achieved with about 55% of the IOPS on the original ranks and the remainder on the newly added ranks.

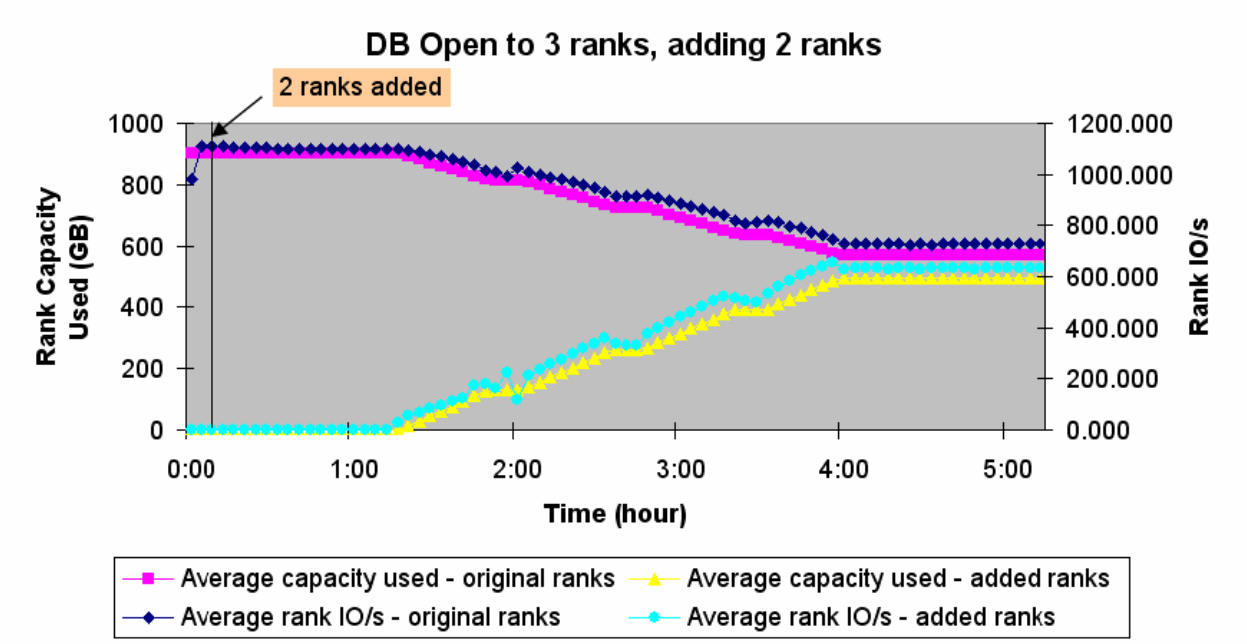

Figure 10. Auto Rebalance – I/O and Capacity Distribution

Figure 11 shows the effect on response time of the system during the data migration process. At the start of the migration process, the response time increased slightly for a short time due to the extra activity on the ranks caused by data migration, while the workload could not realize any of the benefits of the new ranks. As data was migrated to the new ranks, the response time decreased toward the improved performance expected from nearly doubling the number of ranks available to this workload. The dips in response time correspond to the flat periods in Figure 10 and represent periods where the migration paused at a threshold made to limit the impact on host performance. Reducing the migration rate (which was set to the highest possible value for this experiment) may reduce the appearance of these periods as it will be less likely that this threshold will be reached.

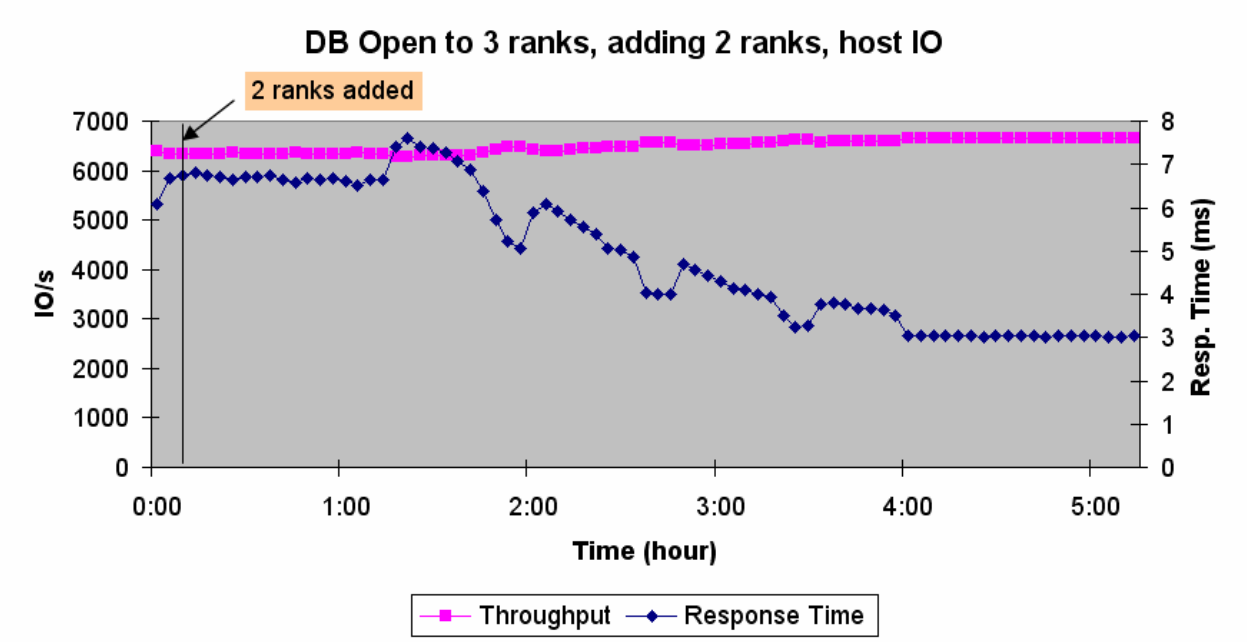

Figure 11. Auto Rebalance – Response Time Performance

4 Easy Tier Tools

4.1 Storage Tier Advisor Tool

The Storage Tier Advisor Tool (STAT) allows users to view summary data output by Easy Tier on the DS8700 and DS8800. The summary data provides estimates of the utilization of configured ranks; graphics showing volume distribution of cold, warm, and hot data; and capacity planning recommendations based on the measured distributions.

STAT is supported on Windows™[1] and can be installed similarly to DSCLI. Input data files to the tool are offloaded from either DSCLI or the DS8000 GUI, while output files from the tool can be viewed with a web browser.

The following screenshots were taken after running a sample OLTP workload. All of the screenshots were taken after the learning period had completed, but before any extents had been migrated. Figures 14-18 were taken from a system with Enterprise HDD and SSD tiers. Figures 19 and 20 come from a system with Enterprise HDD and Nearline tiers.

Figures 14 and 15 provide a high level system view. In Figure 12, the System Summary is shown which gives a high level view of the system's configuration and an estimation of the amount of Hot Data seen by Easy Tier. Figure 13 shows the system wide recommendations for SSD configurations based on the results of the learning period.

[1] Windows is a registered trademark of Microsoft Corporation.

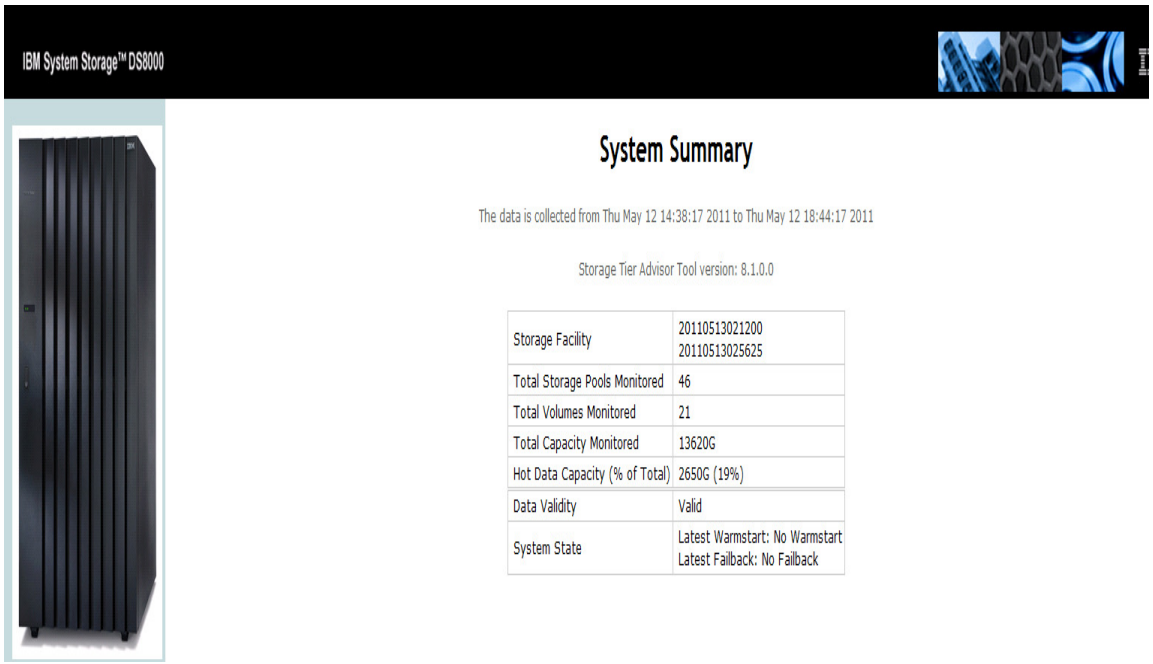

Figure 12. System Summary Output

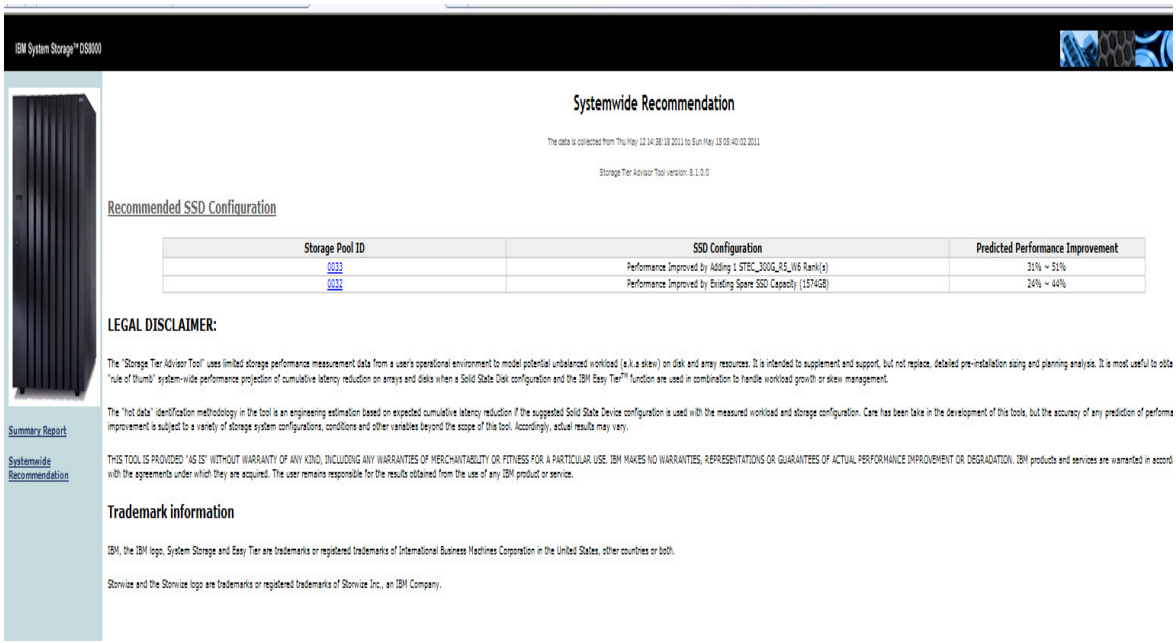

Figure 13. System wide Recommendations

Figures 16, 17, and 18 show results and recommendations at the Storage Pool level. In Figure 14, current performance of the ranks in the sample pool is shown. There, the utilization of each rank is broken down into up to three colors: blue, orange, and red. The breakdown is based on the relative performance to that tier of drives in the system with the percentage shown in blue representing the below average utilization, the percentage shown in orange represents the above average utilization which is still below the maximum threshold, and the percentage shown in red (not pictured here) represents the percentage over the maximum allowed IOPS threshold.

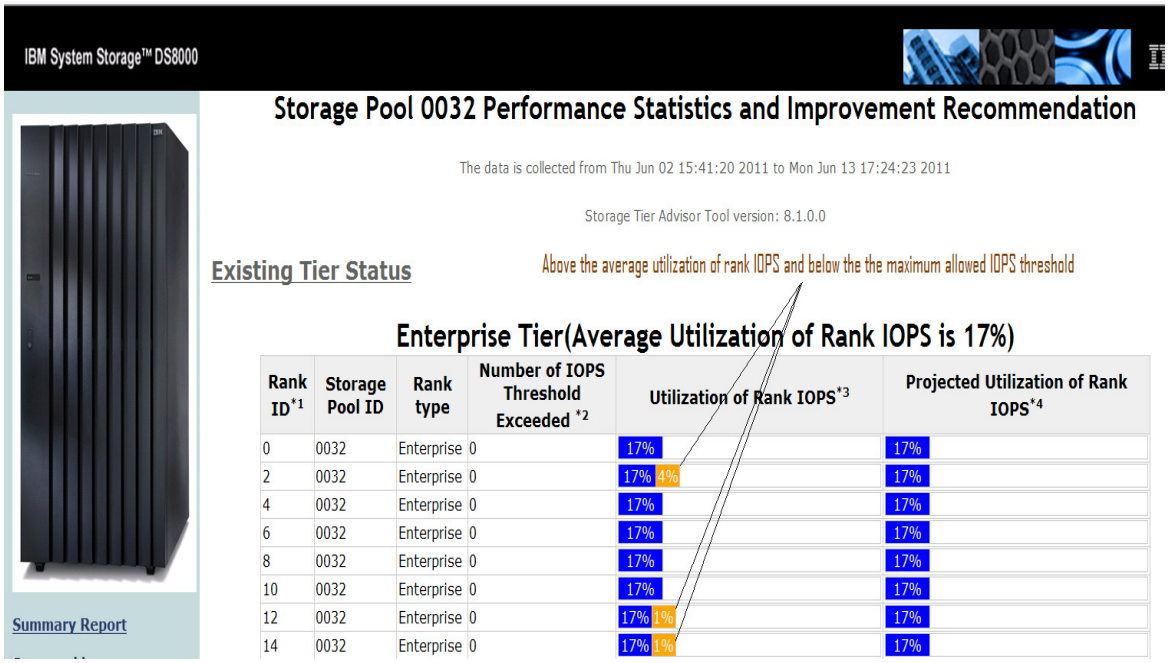

Figure 14. Storage Pool Performance Statistics

In Figure 15, recommendations are shown for the selected Storage Pool in the form of suggestions to add additional SSD ranks to the pool. The expected improvement of each recommended action is given for that pool's performance as well as the expected improvement to the system as a whole. The recommendations given are based on the heat map distribution which is calculated during the learning period. The results of that calculation can be seen in Figure 16 which gives a breakdown of the extents for each volume in the selected Storage Pool into cold, warm, and hot data. Recommendations for additional SSD ranks are made to contain data marked as hot by the Easy Tier algorithms.

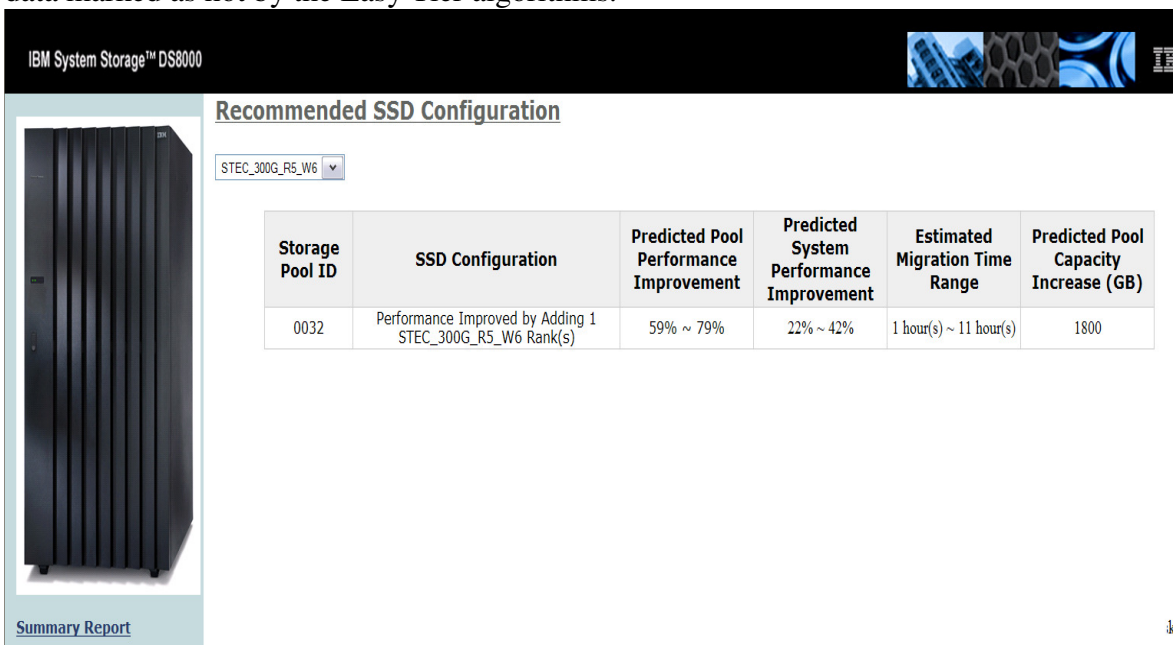

Figure 15. Storage Pool Recommendations – SSD Tier

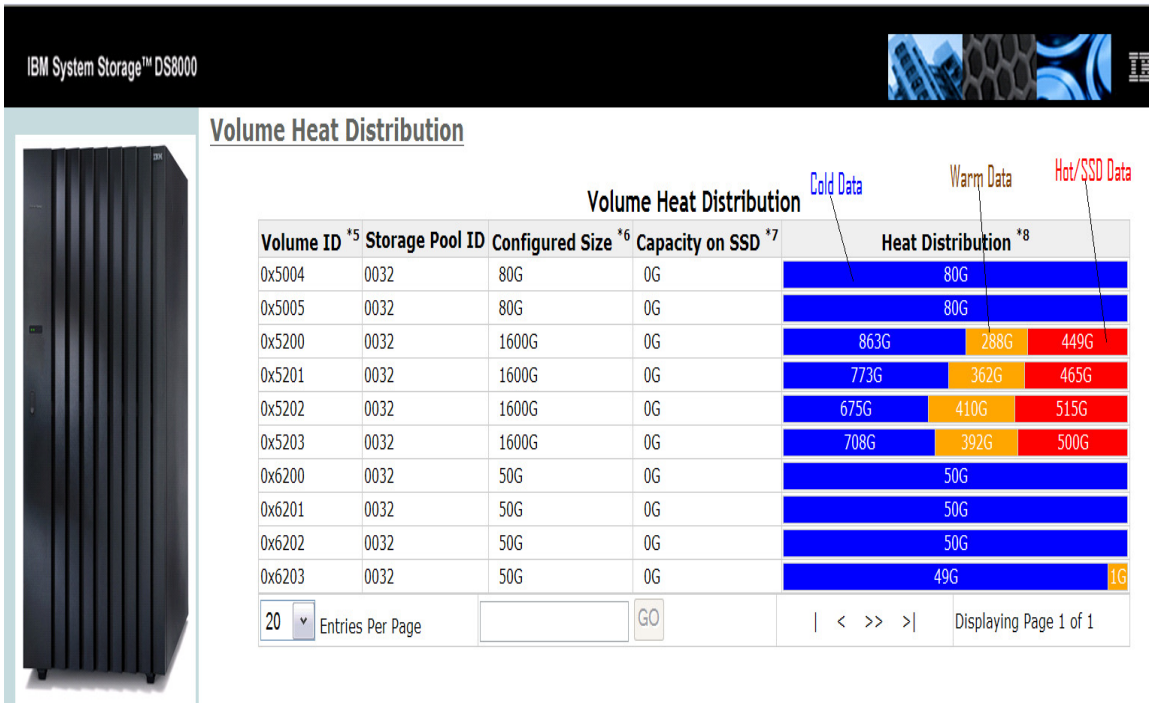

Figure 16. Storage Pool Heat Map Distribution by Volume

Figure 17 is similar to Figure 15, but is taken from a system with 15K RPM Enterprise Tier drives and Nearline Tier drives. Recommendations are given to grow the Enterprise Tier in order to improve performance. Recommendations to increase the Nearline Tier are made if additional storage could be useful for containing cold data in the pool.

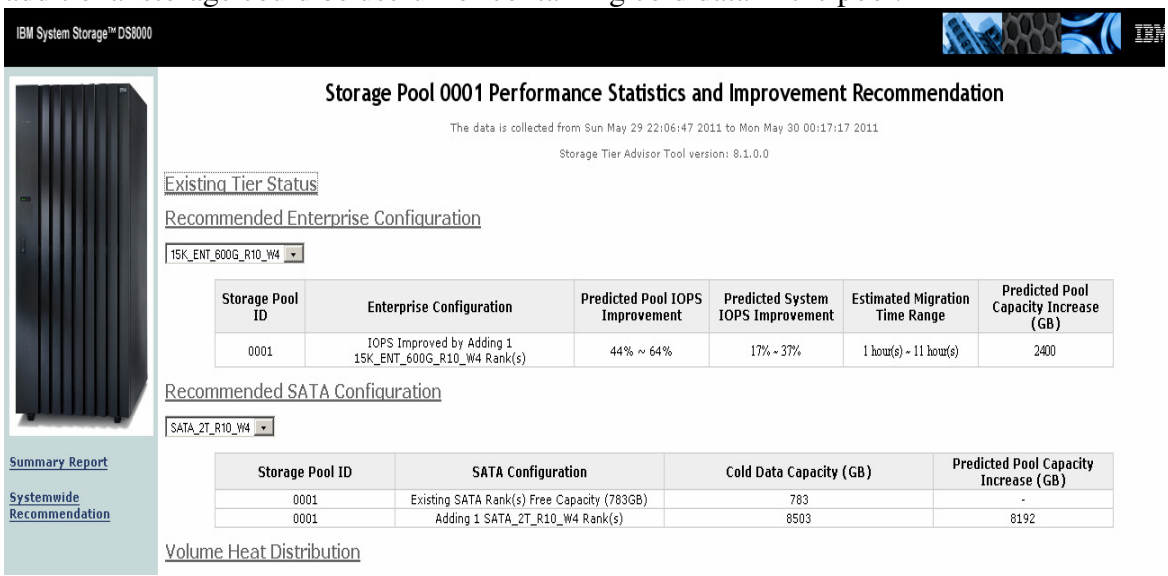

Figure 17. Storage Pool Recommendations – Enterprise and SATA Tier

Figure 18 also came from a configuration with Enterprise Tier drives and SATA disk drives and shows the System wide Recommendations. It is another form of the screen captured in Figure 13, like in Figure 6, recommendations are given to add SATA ranks if needed to contain cold

data while recommendations to increase the Enterprise tier are given for possible performance benefits.

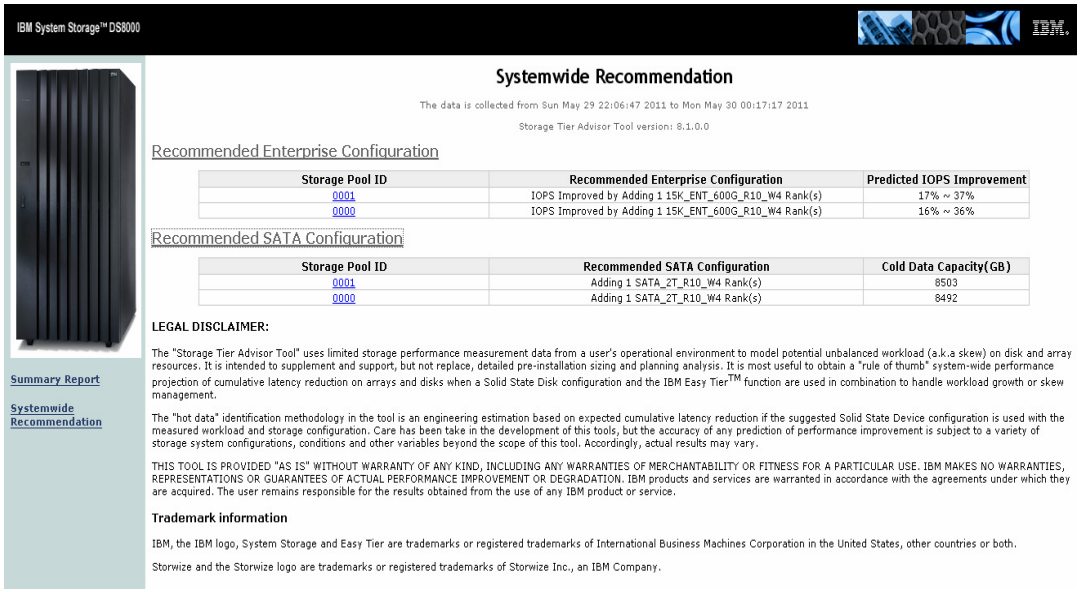

Figure 18. System wide Recommendations – Enterprise and SATA Tier

For a more detailed overview of the functionality and use of this tool, consult the IBM® System Storage™ DS8000™ Easy Tier Redpaper [1].

4.2 Disk Magic

Another tool available for Easy Tier analysis is Disk Magic™[2], a performance modeling tool used by IBM that can help predict the expected performance of storage subsystems. It includes support for models to predict the performance of DS8700 and DS8800 configurations running Easy Tier 2nd generation.

5 Real World Workload Analysis and Insights

Many synthetic benchmarks are not designed to provide realistic skew and variation of activity over a typical production sized working set of data. Even database and application benchmarks are likely to behave somewhat differently in this regard compared to a typical production installation. Hence for Easy Tier it is important to also look at real world performance data to understand the behaviour.

It is possible to run the Easy Tier software in a monitoring mode (even if the Easy Tier feature code is not installed) to collect the same data used to make decisions in a tiered environment. It will even perform the analysis to identify which extents Easy Tier would have migrated if multiple tiers of storage had existed and the migration functionality had been enabled. This extremely detailed data can be offloaded and analyzed by IBM support to provide customers with better insight into their workload and how it interacts with their storage system.

[2] Disk Magic is a registered trademark of IntelliMagic, Inc.

IBM has worked with a variety of clients to collect this data from production installations both with and without solid state drives. This includes a wide range of workloads including mainframe, VMware, Oracle, DB2, etc. The data used for this whitepaper includes three mainframe environments and five open systems environments.

5.1 Easy Tier Data Collection

Easy Tier considers only the IO activity to the backend ranks on the disk subsystem as it is aiming to optimize the workload. Hence cache read hits are ignored and the write workload is considered taking into account any write folding and destage optimisation provided by the write cache. The workload is broken down into reads and writes and then further into small and large IOs.

Figure 19 shows an example of how a production workload is seen by Easy Tier showing the different categories and the overall MB/s generated by the workload over a 24 hour period.

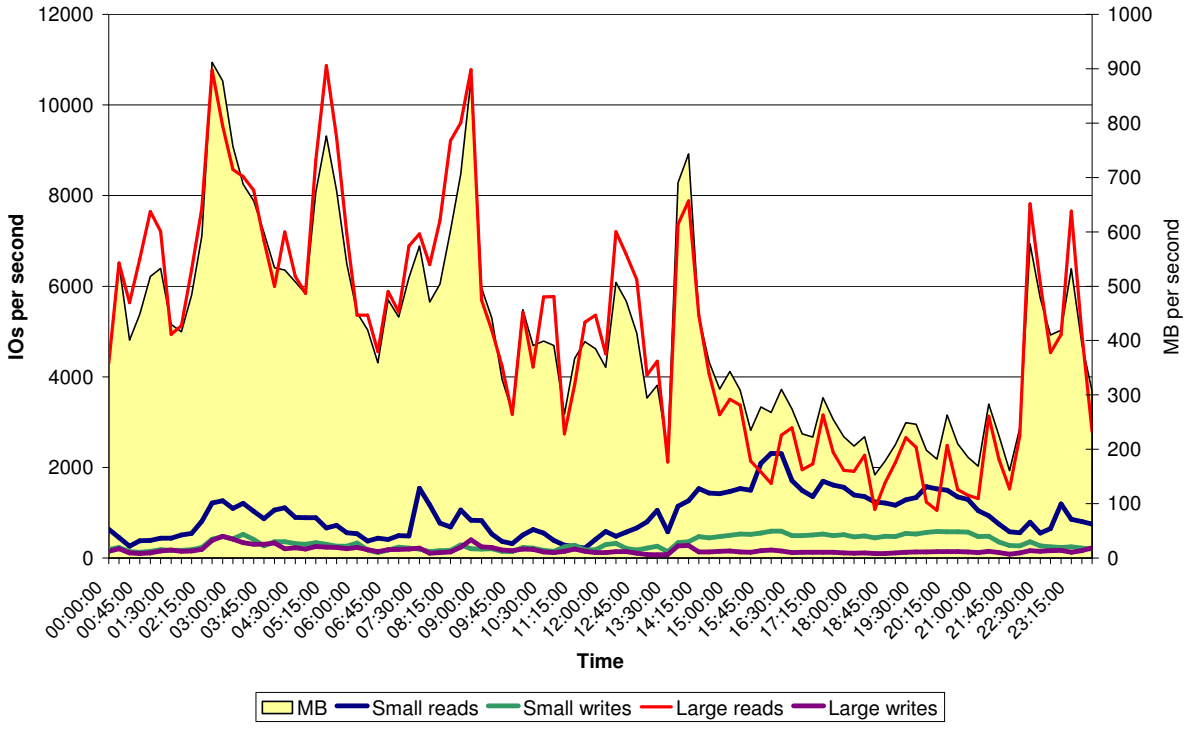

Figure 19. Easy Tier view of a Production Workload

5.2 Easy Tier analysis

Approximately every 24 hours Easy Tier will perform a workload analysis and incorporate the workload information since the last analysis time into the categorization of the extents. The precise interval where this is performed is randomized in order to prevent issues that might occur

if this was performed at exactly the same time each day. However, it will always occur at least every 24 hours.

Figure 20 shows how Easy Tier has characterized the extents of a sample production environment into cold, warm, and hot extents over a 7 day period. Extents are grouped based on the type of activity run to them.

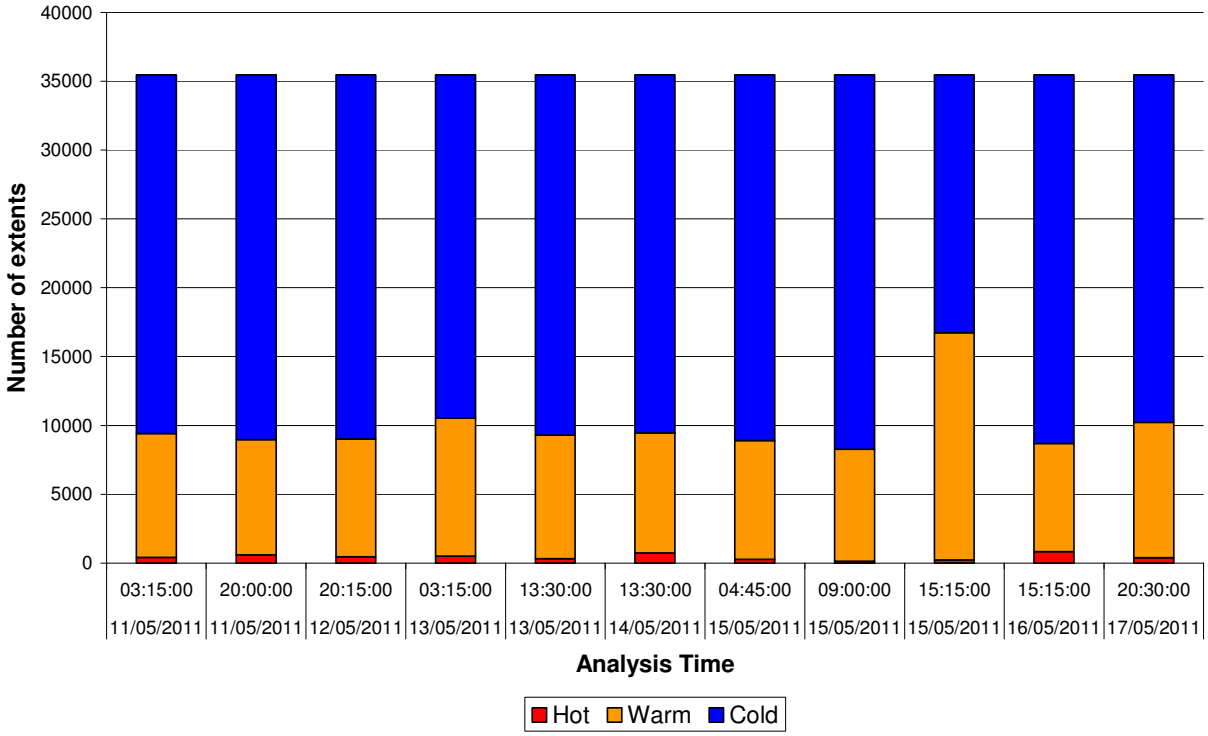

Figure 20. Cold, Warm, and Hot Characterization of a Workload

Cold data will never be promoted to SSDs as it does not have enough small IO activity to justify the promotion of the data, it may have a significant amount of large sequential IOs. In an environment with Nearline drives the cold data could be eligible for demotion to these drives. Hot data will be preferentially promoted to Solid State drives as it contains extents with the most active data in terms of small IOs.

Warm data may also be promoted to Solid State Drives assuming there is free space once the hot data has been promoted. However, it is possible that a cost-benefit analysis might conclude that it is not worth promoting some of this data depending on the level of activity.

Figure 21 shows how Easy Tier has categorized the data for a number of different environments.

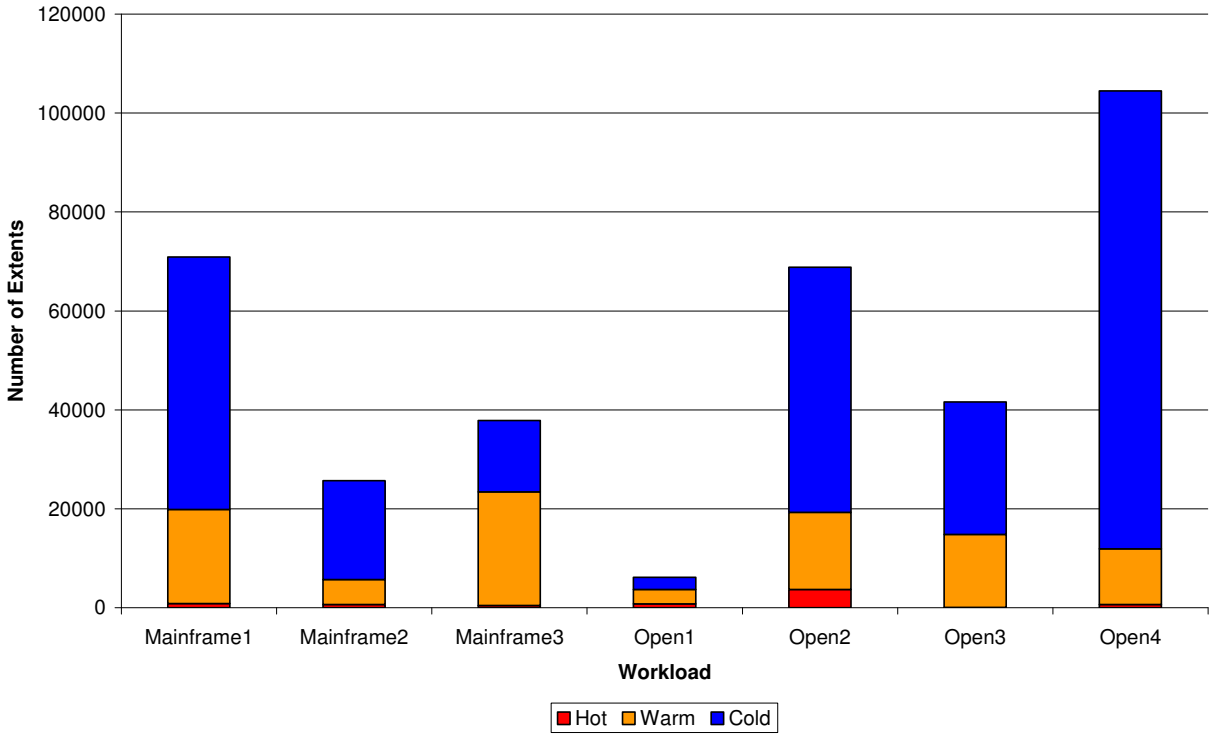

Figure 21. Cold, Warm, and Hot Characterization of several different workloads

It is possible to see here that there is some significant variation in how Easy Tier has categorized the data in different environments depending on how the workload is distributed over the extents. This shows that while we can make some generalizations about behavior not all environments will necessarily conform to these.

5.3 Workload Skew

From an Easy Tier perspective a key attribute of a production workload is the amount of skew in the environment. In an environment where the workload is evenly distributed over the capacity then there would be no hot or cold data to promote or demote to a different tier of storage.

Another aspect of skew is whether the skew of the small random IOs is similar to the skew of the large sequential IOs. In order to effectively use SSDs we want to be able to satisfy the small IOs from the SSD while not overloading them with a significant amount of large sequential IOs.

Figure 22 shows an example of the skew in a real world production environment on the DS8000. This chart is based on 24 hours worth of data so includes both online and batch periods.

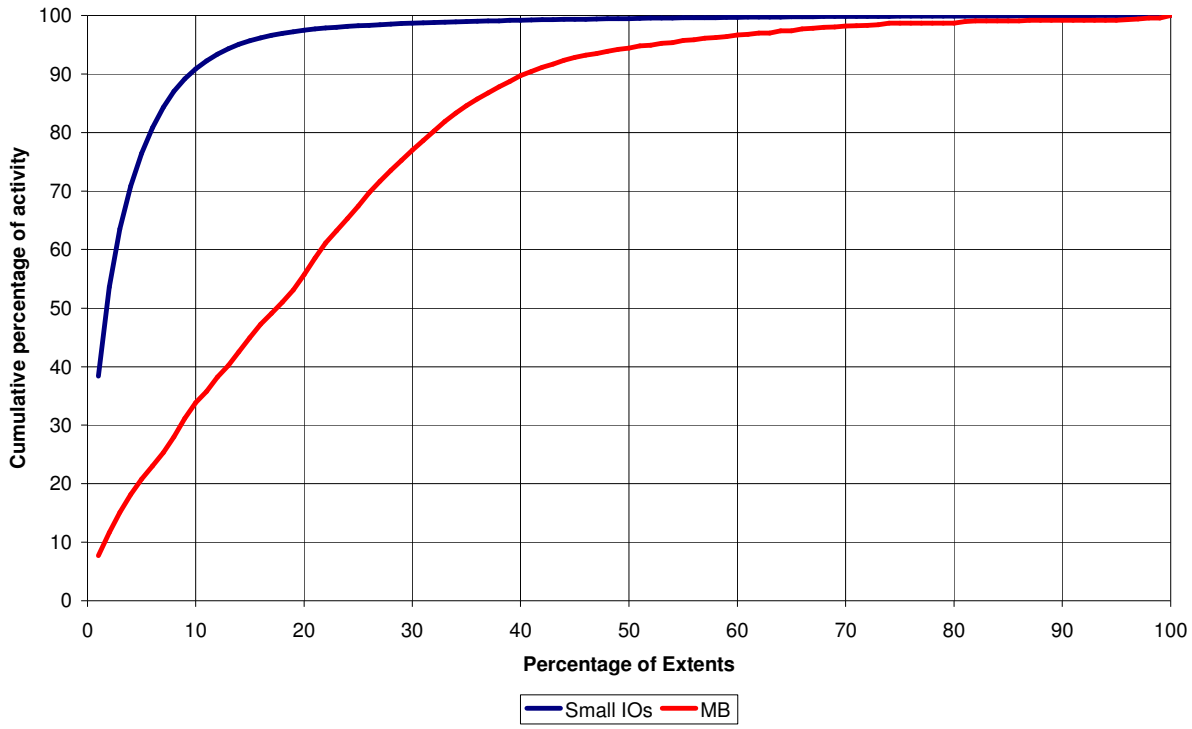

Figure 22. Production Workload Skew Example

There are a number of observations that can be made from this

1) There is significant skew in the distribution of the small IOs over the extents. 90% of the IOs are to only 10% of the extents
2) The distribution of the MB/s bandwidth over the extents is significantly less skewed with slightly more than 30% of the MB written to the top 10% of extents
3) There is a significant amount of capacity which has a low activity both in terms of small IOs and MB/s

Figure 22 shows an optimal distribution of IOs based on a 24 hour period. However, this would assume that Easy Tier was 100% accurate in predicting the hot data on the disk subsystem for this particular day. In order to see how close to the optimal distribution Easy Tier might achieve we can use the Easy Tier ranking of extents from the previous day and see how close to the optimum distribution this would achieve.

We can see from Figure 23 that the Easy Tier algorithm will be able to determine a workload distribution which is very close to the optimum which would have been achievable with hindsight. It will also continue to take into account any changes in the workload profile over this period to perform migrations and to reflect the evolving workload profile.

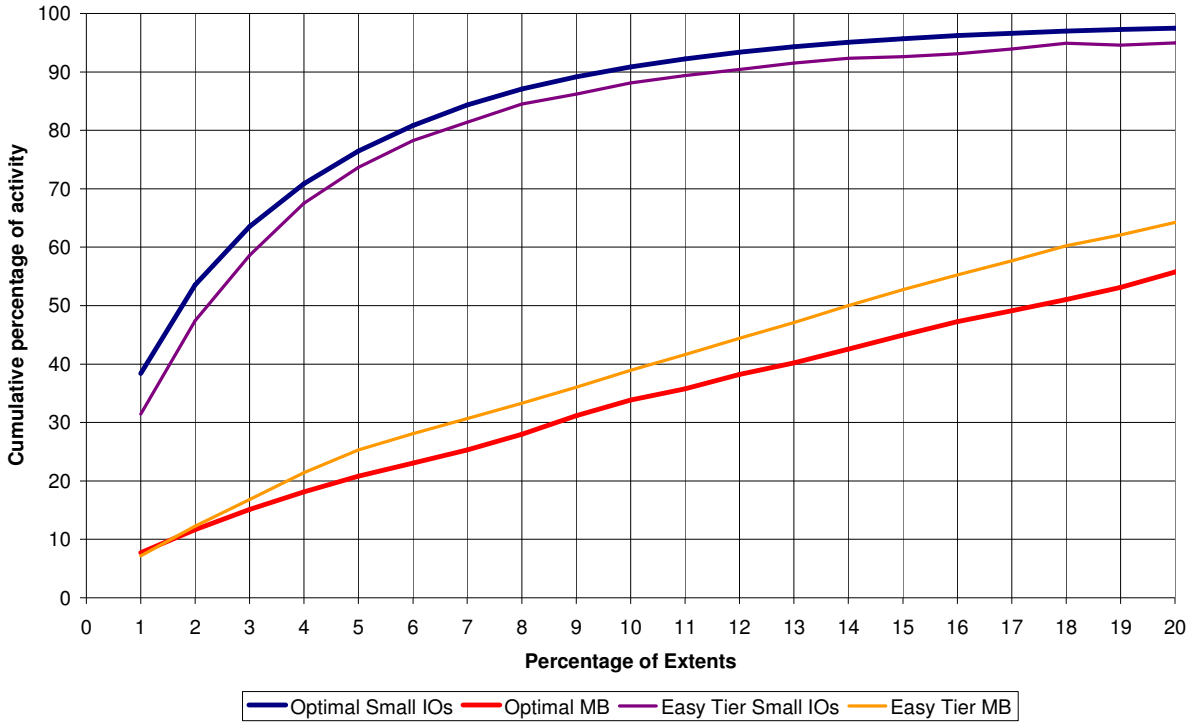

Figure 23. Easy Tier Distribution vs. Optimal Distribution

We also should consider whether the overall 24 hour profile provides an optimal distribution for the peak period during the online day where the largest amount of small IOs is performed.

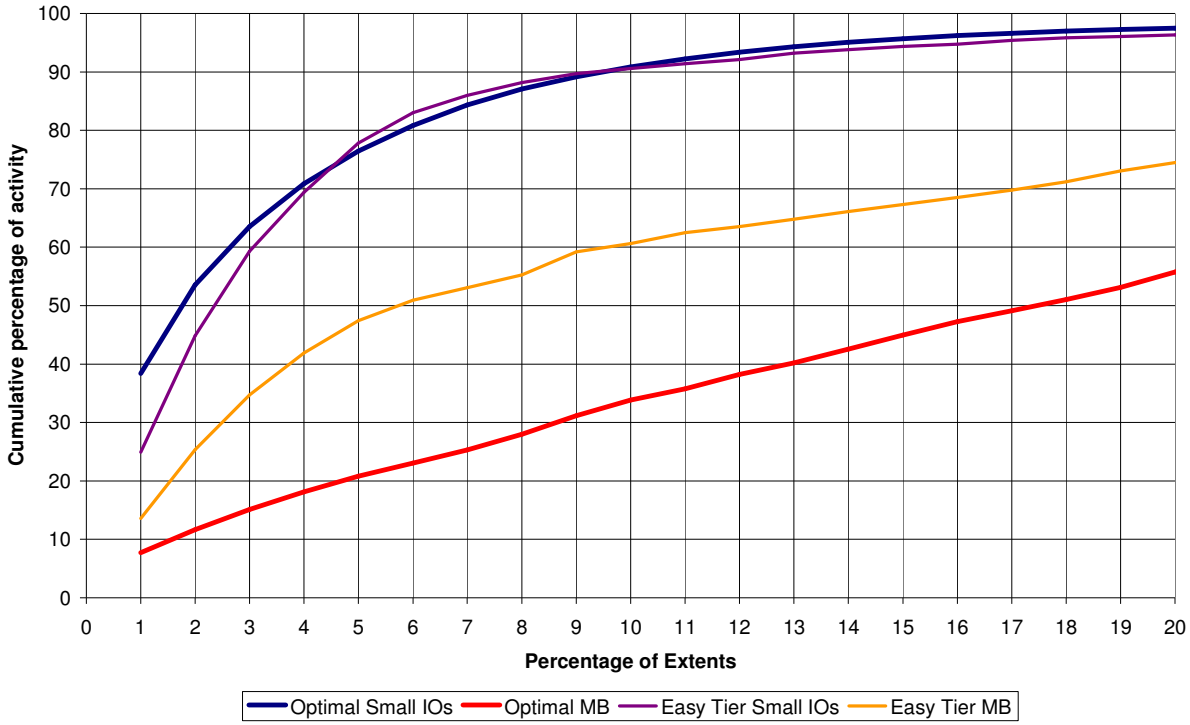

Figure 24. Easy Tier Distribution for Peak Period vs. Optimal Distribution for Peak Period

Figure 24 shows how Easy Tier would have distributed the IO for the peak online period which reflects a distribution that is very close to the 24 hour optimal distribution. Hence, using the total workload profile tends to reflect the active data in peak periods.

Figures 25 and 26 show the skew of small IOs and the skew of MB for a variety of clients. We can see that there is no significant difference in the skew of data between mainframe and open systems workloads and that many environments are similar in terms of the level of skew.

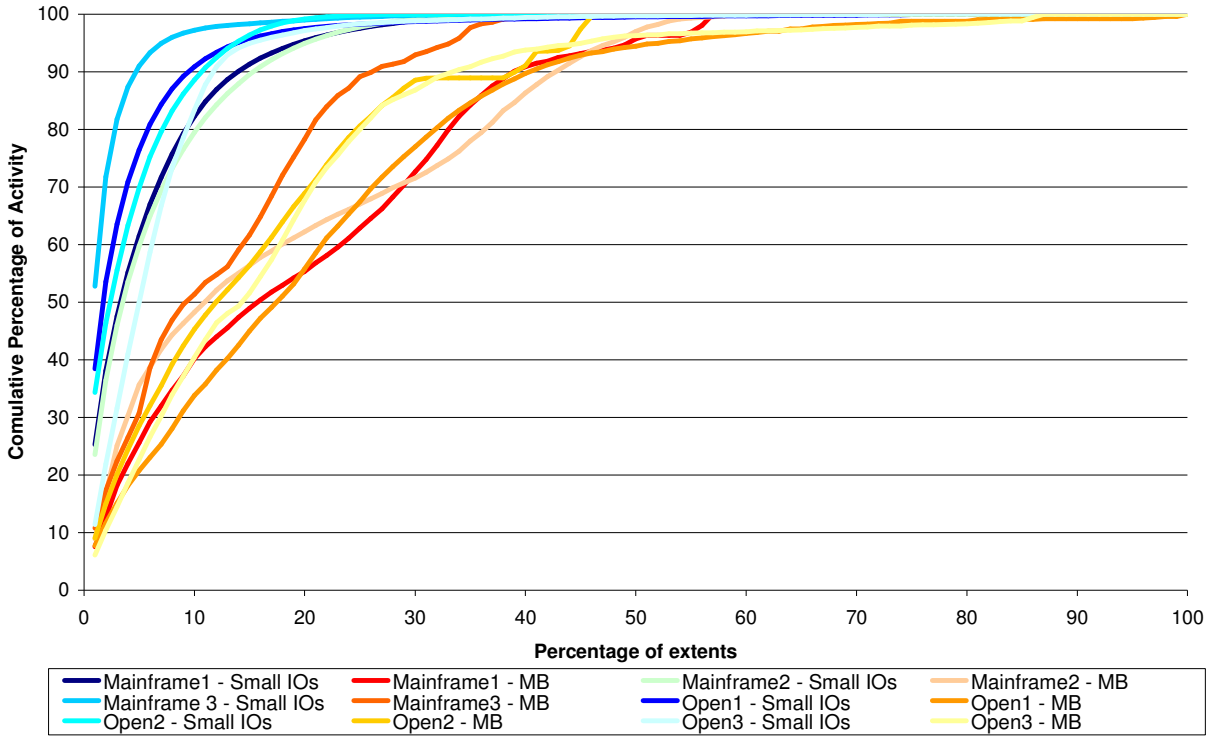

Figure 25. Comparative Workload Skew

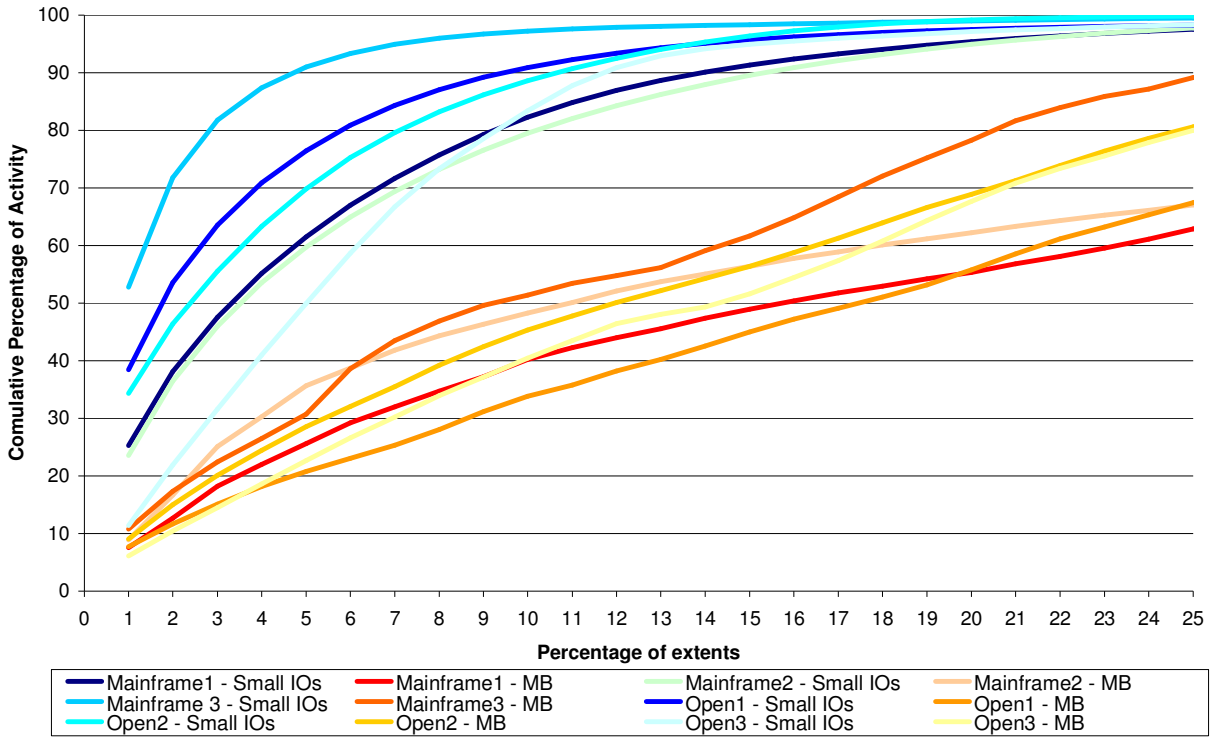

Figure 26. Comparative Workload Skew showing detail of active extents

If we look at the detail of the active extents then we can see that despite the general similarities there are some differences in the percentage of activity on the most active extents.

For example, if we look at the Open3 workload then there is a much more even distribution of the workload over the busiest 10% of extents. This client is using wide and very granular software striping for their main database workload which has the effect of reducing the skew between the extents containing the database.

5.4 Variation over time

Another aspect of the skew of the workload is how it varies over time. If the hot data in the workload varies rapidly over time the storage system would spend significant resources constantly moving data rather than handling the production workload. Additionally, as the system was reacting to workload change, the workload could change again before the system had fully optimised for the prior change.

Because of this, the design for Easy Tier assumes that we are able to find a relatively stable tiered configuration that only requires incremental changes as the workload evolves over time.

Figure 27 shows the variation in the skew of small IOs over a 24 hour period based as before on the Easy Tier categorisation from the previous day. We can see that the extents that are hot during the day are also hot during the batch period and so Easy Tier is not required to make changes between these two time periods.

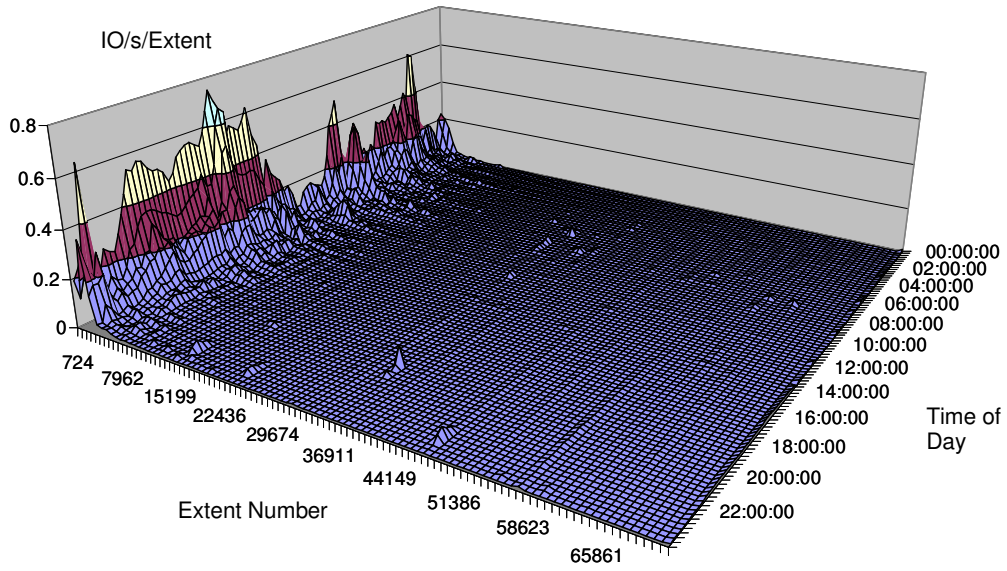

Figure 27. Small IO Skew Variation Example

This also indicates that continually collecting data over the 24 hour period and acting on this is a good approach.

In Figure 28 we look at the spread of the MB/s over the 24 hour period and can see that there is more variation of the workload both over time and over the extents in the disk subsystem. This means that it is unlikely that the movement of the hot data to Solid State will result in these drives being overloaded with large sequential IO.

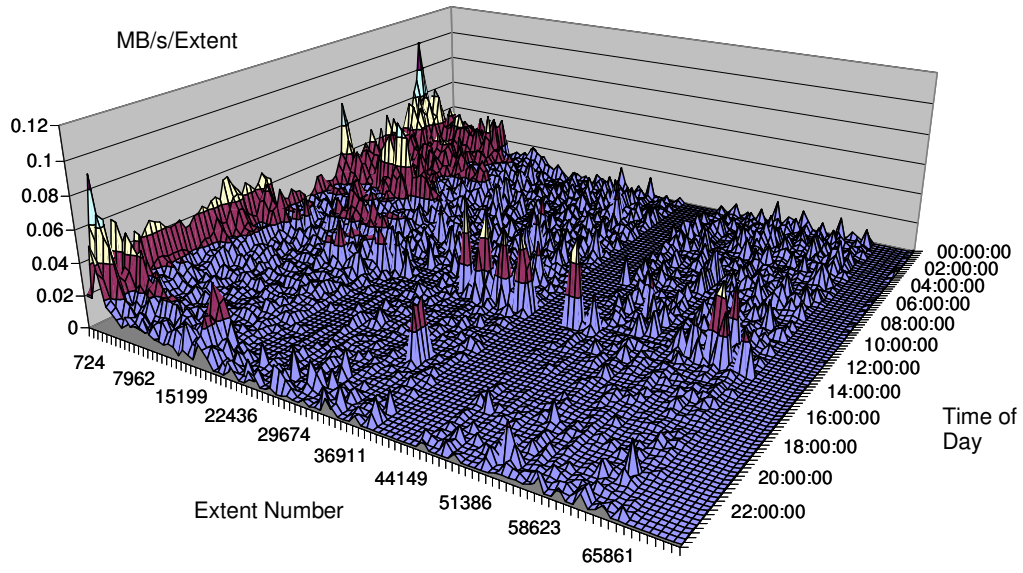

Figure 28. Sequential IO Skew Variation Example

6 Best Practices

In general, the best practices for Easy Tier have not changed much with the new functionality. The best practices included in the original DS8700 Easy Tier whitepaper still apply [2].

Some rules of thumb may be useful to help you decide which storage tiers to consider for an Easy Tier implementation depending on your objectives. If your primary motivation is to obtain high performance, a mixture of SSD and Enterprise drives would generally be your best choice. If the focus is cost reduction, a combination of Enterprise and Nearline drives may help you realize this. Finally, if you need to reduce energy consumption or physical footprint, a combination of SSD and Nearline drives may help you get there. Of course these are just rules of thumb and any Easy Tier implementation would likely provide advantages in multiple areas.

When considering an Easy Tier implementation, it is not necessary to do extensive planning to have a successful result. However, there is some up front effort that may be helpful. When possible, using STAT over a number of days will help you verify that your workload will consistently benefit from Easy Tier. You may also use the data from STAT to see if the volumes containing hotter data are critical to your business and actually require enhanced performance. Using other performance tools such as IOSTAT, TPC, etc. may provide additional information to supplement STAT. Finally, IBM or a business partner representative may use the Disk Magic tool to help estimate the performance benefits that Easy Tier would deliver in your environment.

Since the Easy Tier feature does not have a charge associated with it, we recommend including it in any new orders which may be able to make use of Easy Tier.

Adding Nearline drives and Easy Tier to an existing Enterprise drive storage system will generally enable much higher storage capacities without compromising performance. However, it is suggested that additional cache be applied when adding capacity in this manner. In general, it is recommended that cache be increased in proportion to the number of disk spindles added. For example, if 64 Nearline drives are added to an existing storage system with 64 Enterprise drives, the cache size should be doubled. When adding SSD drives, it is also important to ensure that the SSD arrays are well balanced across the DS8000 servers. For example, with two ranks of SSDs, each server should own one SSD rank.

One common question is whether Easy Tier auto mode should be turned off periodically if workload changes are expected. This is not recommended. If monitoring remains on, then the learning continues to happen and will take effect when auto mode is re-enabled. You would need to also disable monitoring to avoid this which effectively turns Easy Tier off completely.

Another common question is whether it is advisable to use multiple automated storage tier management technologies at once. For example: “When using ASP Balancer on System i, should Easy Tier be used at the same time?” Or, “When using Easy Tier on a SAN Volume Controller (SVC), should Easy Tier also be used on an underlying DS8000?” In general, this is not recommended unless there is a compelling business reason.

7 Conclusions

Easy Tier 2nd generation added additional functionality which makes the feature an even more attractive offering. With the option of selecting from multiple drive tier combinations, Easy Tier can now effectively manage storage performance and growth according to need. Whether concerned mostly for cost or performance, Easy Tier removes the many difficulties of carefully placing data and tuning to get the best performance available. With the addition of the auto-rebalance function, storage growth within a tier becomes a much simpler process as well.

In this paper we have shown the significant performance benefits which can be realized through this automated feature. And in examining real environments, we showed how the Easy Tier algorithms provide very similar performance as would be achieved through optimal data placement. Performance which was previously only available after painstaking research and manual tuning may now be realized automatically. Easy Tier brings the added benefit of being able to adjust to conform to changes in the workload without intervention by the user.

8 References

[1] Dufrasne, B., Couteau, S., Jer, M., Manthorpe, S., Murke, R., Rosichini, M. "IBM® System Storage™ DS8000™ Easy Tier V2." June 2011.

[2] La Frese, L., Hossain, K., Hyde, J., Lin, A. W., McNutt, B., Sansone, C., Xu, Y., Zhang, Y. "IBM® System Storage™ DS8700™ Performance with Easy Tier®." May 2010.

[3] La Frese, L., Sutton, L., and Whitworth, D. "IBM® System Storage™ DS8000® with SSDs: An In-Depth Look at SSD Performance in the DS8000." April 2009.

[4] Lin, A. W., Whitworth, D., Williams, S. E., Xu, Y. "IBM® System Storage™ DS8800® Performance Whitepaper." December 2010.

[5] La Frese, L., Lin, A. W., Martin, J., Williams, S., Xu, Y. "IBM® System Storage™ DS8700® Performance Whitepaper." August 2010.

[6] Ripberger, R. and Xu, Y. "IBM System Storage, DS8000 Storage Virtualization Overview, Including Storage Pool Striping, Thin Provisioning, Easy Tier", WP101550 V2.0, May 2010.

[7] Altman, J., Sutton, L., and Sutton, P. z/OS Hot Topics, article 20-48: z/OS Support for Solid State Drives in the DS8000. February 2009.

9 Appendix

8.A Appendix A: Workload Characteristics

- *DB Open*: 70% reads, 30% writes, 50% read hits. This workload is designed to be comparable to typical applications with transactional workloads. Read/Write Ratio = 2.33, Read Hit Ratio = 0.50, Destage Rate = 17.2%, Transfer size = 4 KB.
- *OLTP workload*: simulates the workload of transaction processing systems that require small, mostly random, read and write operations (for example, database systems, OLTP systems, and mail servers). It resembles the mix of I/O workload components as defined in the SPC-1 specification.

8.B Appendix B: DS8000 Hardware Configurations

8.B.1 Configuration for DB2 Brokerage Measurements

- DS8700 Configuration
  - 941 4-Way, Cache was reduced from 256 GB to 128 GB
  - 128 146 GB / 15K RPM drives configured as RAID-5
  - For experiments with SSDs: 16 600 GB SSDs configured as RAID-5. (16 146 GB SSDs were used in Easy Tier1)
  - 3 DA Pairs with SSDs on a separate DA pair
  - 4 x 4 Gb HA ports
- DS8800 Configuration
  - 951 4-way 128 GB Cache
  - 128 146 GB / 15K RPM drives configured as RAID-5
  - For experiments with SSDs: 16 300 GB SSDs configured as RAID-5
  - 3 DA Pairs with SSDs on a separate DA pair
  - 4 x 8 Gb HA ports
- DB2 Configuration
  - DB2 9.7 FP1, 4 Instances, 4 x 2 TB DBs, 4 Buffer Pools at 54 GB each
    - 8 x 1.6 TB volumes were allocated for database, Temp files and Data Generation
    - 8 x 50 GB volumes were allocated for log files
- Server Configuration
  - P770 (AIX 6.1.5.0), 8 Eight Core P7 (3GHz)
  - 512 GB Cache
  - 4 x 8 Gb FC Ports

8.B.2 Configuration for OLTP Measurements with 15K and 7.2K RPM Drives

- DS8700 Configuration
  - 941 4-Way 128 GB Cache
  - 128 300 GB / 15K RPM drives configured as RAID-5, 64 2 TB / 7.2K RPM Nearline drives with RAID-10
  - 4 DA Pairs
  - 8 x 4 Gb HA ports
- Server Configuration
  - P770 (AIX 6.1.5.0), 8 Eight Core P7 (3GHz)
  - 512 GB Cache
  - 8 x 8 Gb FC Ports

8.B.3 Configuration for CPW Measurements with IBM i

- DS8800 Configuration
  - 951 4-Way 128 GB Cache
  - 144 146 GB / 15K RPM Drives configured as RAID-10
  - 4 DA Pairs
  - For experiments with SSDs, 16 300 GB SSDs configured as RAID-5

- Server Configuration
  - 9119-FHA 64 Way Server
  - 320 GB Cache
  - 6 dual port 8 Gbps Fibre Channel Cards