Dell tiering vs cache-based competitor

Review of Evaluator Group test report shows that, despite competitor efficiency claims, Dell's large raw hardware advantage translated directly into better results

Dell Marketing
April 2015

A Dell Technical White Paper

Revisions

Date	Description
April 30, 2015	Initial release
May 6, 2015	A note, previously in Table 1, referencing the Evaluator Group paper Appendix A, was removed because the revised Evaluator Group paper now lists the PS6210XS 1.2TB drive capacity in Appendix D.
May 8, 2015	Fixed typo in product name in Figure 9.

Table of contents

Executive summary

Dell tiering provides greater raw flash resources at lower prices than cache-based hybrids.   To check competitor claims of having efficiency advantages that make up for the lower flash capacities, Dell engaged Evaluator Group to run tests modeled on real-world environments.

Evaluator Group summarized:

> *"Our testing found that for sustained workload performance, both Dell systems provided better performance as measured by higher I/O rates, lower latencies, and higher throughput than the competing system."*

The competitor efficiency claims came with undocumented inefficiencies such that the end result was that Dell superior economic efficiency for raw flash translated directly into large net advantages for capacity and performance.

Doing some math on the test report reveals that:

- Dell Storage SC4020 provided 8x more raw flash and 2.3x more total raw capacity at a similar price.
- Despite the competitor's efficiency claims of compression and of not consuming any flash space for parity or sparing, the Dell raw advantage led to an equal or larger net advantage in how much data could be accelerated to flash speeds.
    - 8x more raw capacity → 8x more data held in flash in SC4020
    - 2.3x more raw capacity → 2.5x more data held in flash in EqualLogic PS6210XS
- The competitor was subject to dramatic reduction in performance in the event that a flash drive was to fail and thereby reduce the cache hit rate.
    - 4x or 2.5x reduction depending on starting point.
- Despite competitor claims of extreme write optimization, Dell was faster in practice on the simulated realistic data pattern even with 40% writes:
    - 2.9x more raw HDD performance → 4x more HDD-only performance in PS6210XS
    - SC4020 Read Intensive (RI) SSD → SC4020 2.6x faster even with 40% writes
- Despite the competitor's suggestion that caching reacts quickly to changing workloads, tiering was a better choice for alternating workloads: no degradation for Dell.
    - As Evaluator Group commented: "*the caching approach of the competitor did not warm cache as quickly as literature would lead us to believe*."
    - 2 hours into Monday AM processing, competitor was 3x slower than Friday, 12x slower than PS6210XS, and 30x slower than the flash tier in SC4020.
    - For a long-term idle workload suddenly made hot, PS6210XS was faster for the first 14 hours and then again 33 hours onward.

1 At similar price points, Dell provides more flash hardware

Dell's tiering architecture allows large flash capacities to be provided with low cost overhead.   A comparison of market prices showed that the competing caching architecture provided much less raw hardware capability for the price.   The testing described in later sections showed that, despite competitor efficiency claims, the Dell advantage on raw flash capacity translated directly into advantage in usable flash capacity and system performance.

Evaluator Group determined the partially pre-discounted, government purchasing price, known in the US as GSA. Source for Table 1: Appendix D of the Evaluator Group paper.

Table 1   Configurations and Pricing

Unit	Raw flash	Total raw data capacity	GSA price
Dell Storage SC4020	19.2TB (12x1.6TB read-intensive)	19.2TB as tested	$85,908
	19.2TB (same)	55TB with add-on NL-SAS	$106,950
Dell EqualLogic PS6210XS	5.6TB (7x800GB)	26 TB (includes 17 1.2TB 10K RPM HDD)	$77,133
Competitor largest controller as of June 2014, configured as shown	2.4TB (4x600GB)	24TB (12x2TB NL-SAS). Competitor SSD is for an extra copy that does not contribute to storage capacity.	$94,710

2 Summary of efficiency claims, downsides, and results

In order to align testing to the best solutions offered to a given customer, one must test comparably priced configurations, rather than configurations of equal flash capacity. For example, the SC4020 hybrid offers so much more flash capacity that entire key applications can run in all-flash for the price of the competitor's hybrid that only fits a small portion of overall data in flash. Adding 12 3TB NL-SAS to enable automatic tiering makes the configuration 13% higher in price with 2.3x more raw capacity in addition to 8x larger raw flash capacity. Testing addressed competitor efficiency claims and downsides as shown in **Error! Reference source not found.**.

Table 2   Net effect of competitor efficiency claims and related downsides

Dell raw advantage	Competitor efficiency claims	Competitor downsides	Test results
SC4020: 8x more raw flash; PS6210XS 2.3x more raw flash	Compression and no space lost to parity or hot spares	Extra copies of data awaiting garbage collection	SC4020 fit 8x larger data sets into flash; PS6210XS fit 2.5x more data in flash
Consistency: keep both weekday and weekend hot loads in flash simultaneously	Immediate data movement; alleged agility benefit: don't need to keep both hot loads cached at once	Cache re-warms at low, HDD-small-block-random IO rates	No "Monday morning" drop for Dell. Even 2 hours into the morning, Competitor was still 3x slower than Friday, which is 12x slower than PS6210XS and 30x slower than SC4020
Data in SSD protected	Read-only caching does not require the extra space for parity or hot spares to protect data; data can be re-read from HDD	25% of formerly cached data no longer cached if 1 of 4 SSD's fail	Competitor performance drops 2.5x to 4x at the resulting, lower cache hit rate
PS6210XS 2.9x more raw disk IOPS (more HDD, 10K RPM) SC4020 Read Intensive (RI) SSD's provide higher capacity with good performance	Write optimization	Metadata burden and longer response times Cost overhead means fewer devices for the money	PS6210XS did 4x more HDD IOPS; lower latency SC4020 did 2.6x more IOPS (tests used 40% writes)

Figure 1 depicts the Dell Storage SC4020 advantages on the two most commonly requested metrics: how much data can be accelerated to flash speeds and how fast can the system go:

Dell SC4020 vs. cache-based competitor: More data accelerated to higher speeds

Net flash capacity

8x

Competitor   Dell SC

Performance

2.6x

Competitor   Dell SC

At similar price, and lower $/GB, **Dell SC4020 :** **held 8x more data** in flash and provided **2.6x more IOPS**
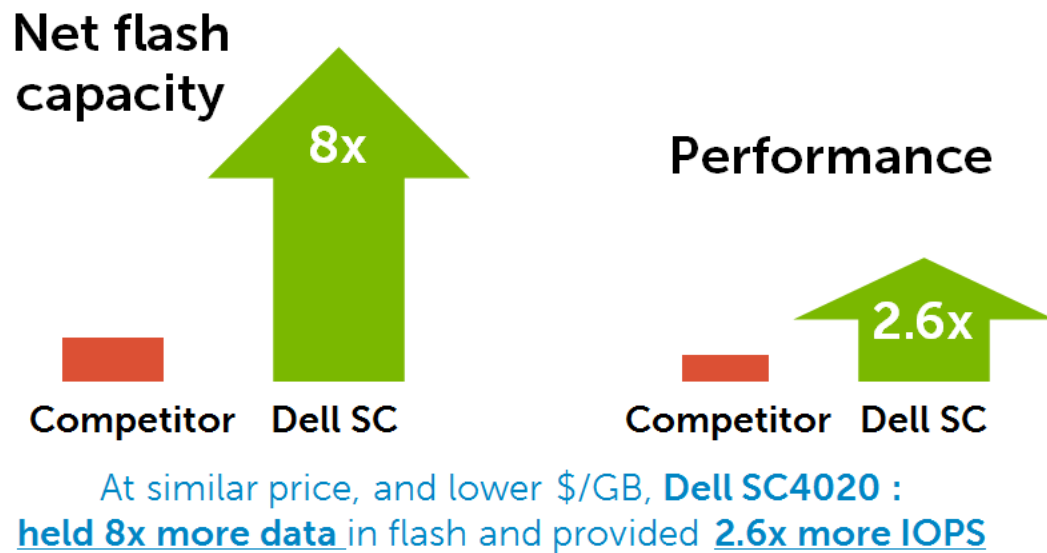
Figure 1   Dell Storage SC4020 usable flash capacity and performance vs. competitor

# 3   Flash raw and usable capacity comparisons

Despite the competitor efficiency claims, the Dell net, usable advantage was equal or greater to the raw advantage.

- SC4020 had 8x more raw flash capacity and held 8x more data in flash
- PS6210XS had 2.3x more raw flash capacity and held 2.5x more data in flash

For raw advantages, compare the capacities in the configuration.

As shown below, the Dell tested advantage was larger than the raw advantage.   How could that be, given that the competitor has compression and does not require SSD space for parity or hot spares?

When putting read misses or writes into flash, the competitor architecture writes new data to a new location rather than over-writing the location where the data had previously resided. Some process must later remove older copies, of which there may be several if the block has been accessed multiple times. Whatever the exact cause of the space overhead is, the test result was that it completely negated the claimed advantages when testing at flash performance levels.
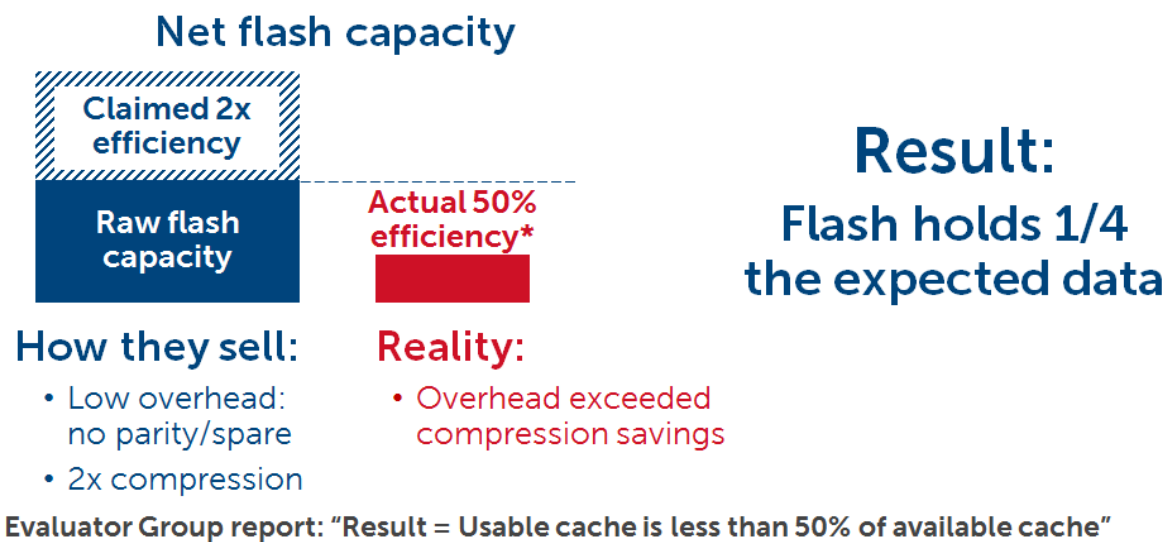
Figure 2   Flash capacity expectations vs. results

The above Evaluator Group text quote is on page 11 of the Evaluator Group paper. The graphic is by Dell. Efficiency is even lower than shown in the graphic when high cache hit rates with high performance are needed.

If performance load on the competitor flash is lower, as happens when HDD is greatly limiting performance, the space overhead was lower. As show in Figure 1 of page 11 of the Evaluator Group paper,

when data is evenly distributed across the 6TB, the cache hit rate was 26%. With 26% of 6TB being cache hits, the data cached is 60% of the raw size, at which point compression and lack of parity drives has apparently overcome the extra space overhead, but has not moved the competitor ahead of standard products that don't compress and do need space for parity. However, if performance needs are that low, the competitor hybrid is not the best solution.

The low-performance example is not particularly useful for a hybrid array given the performance achieved with 26% cache hit rate was well below the HDD-only performance for PS6210XS as shown in the "100% of IO to 6TB" data point Figure 5 on page 16 of the Evaluator Group paper.

The following subsections make the most relevant compares: flash capacity comparisons for tests in which the products could deliver flash performance levels.

3.1   SC4020 held 8x more data in flash

The testing was done based on a 6TB data set with a compressibility setting of 2:1.

As shown in Figure 1 of the paper on page 11 of the Evaluator Group paper, the size of the data had to be reduced dramatically to make the cache hit rate approach 100%.   As further documented at the end of Appendix E of that paper, the volume size that produced 99% cache hit rate was 0.75TB. At 1TB, the cache hit rate 97% as reported in Appendix E of that paper. The competing array needed the higher 99% cache hit rate to achieve its best all-flash performance.

Note: the upper limit of the SC4020 flash capacity was not tested.

The 8x claim is calculated as 6TB for SC4020 divided by 0.75TB for the competitor.

3.2   PS6210XS held 2.5x more data in flash

The data that fits in PS6210XS flash was determined to be at least 2.5TB, as documented in the final bullet on page 11 of the Evaluator Group paper, describing the net result of the alternating workload test 5b on page 19 of the Evaluator Group paper. The performance of two 1.25TB hot regions, both of which had to be in flash simultaneously to avoid performance degradation following the switch between which was heavily accessed, was as high as the all-flash testing found in Appendix E of that paper.

The competitor had similar all-flash performance when the data set was 1TB, as shown in Figure 5 on page 16 of the Evaluator Group paper. Therefore, for an apples-to-apples comparison, we compared the 2.5TB stored by EqualLogic in flash that provided similar performance to a 1TB data set for the competitor, giving Dell at 2.5x advantage.

4 Performance comparisons

4.1 Dell Storage SC4020 performance comparisons

At best, when operating as a tiny all-flash array holding only 0.75TB of data, the competitor was 2.6x slower than the SC4020.   For larger data sets that were still smaller than the raw flash size of the competitor, the SC4020 advantage was 3.2x (for 1TB of data) or 6.1x (for 1.5TB of data).
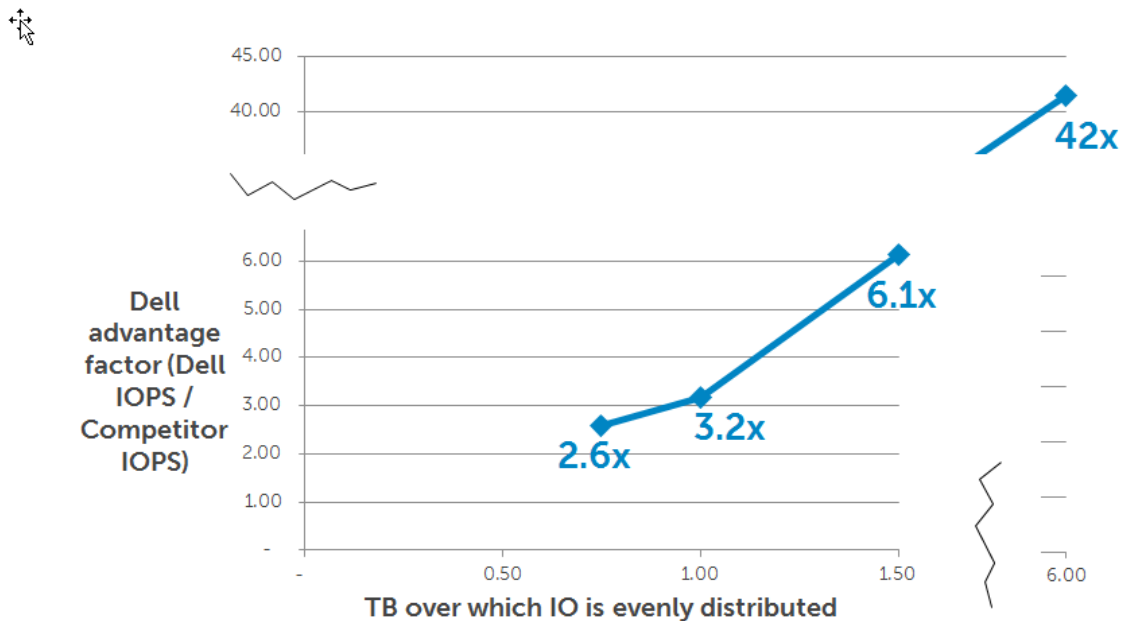
Figure 3   Dell Storage SC4020 performance advantage as a function of data set size

The comparisons shown in Figure 3 are based on the performance reported in Appendix E of the Evaluator Group paper.

4.2 Dell PS6210XS performance comparisons

For HDD-only performance, the competitor claims a great advantage due to turning logically random small block writes into fewer, large-block sequential writes to the HDD. However, doing so creates other forms of overhead. Even for 40% write data, the overhead exceeded the benefit. PS6210XS did better in the actual testing than one would have expected just based on its greater numbers of faster-spinning HDD.

Dell PS6210XS vs. cache-based competitor:
Faster raw HDD and even greater HDD-only IOPS

2.9x

Dell direct block addressing had lower overhead than competitor's, sequentialized data layout, outweighing the competitor benefit of more efficient writes

4x

More raw HDD speed
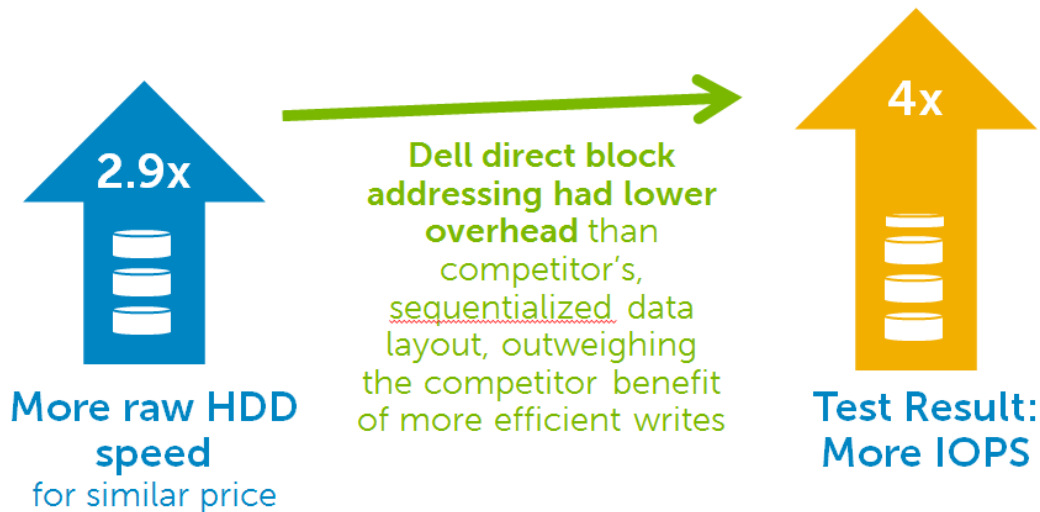for similar price

Test Result: More IOPS

Figure 4   Dell PS6210XS HDD raw performance advantage and test results

PS6210XS: 16 active 10K RPM disks, estimated 2x higher raw speed per disk than competitor's 11 active 7.2K RPM disks. 16 x 2 / 11 = 2.9.

The test result was based on max IOPS sustainable at under 20ms average read or write response time. The competitor data used for the compare was taken 45 minutes after the workload first started accessing an area that had been idle a long time. By that time there was a small cache hit rate. To calculate a truly HDD-only performance level, the data for the comparison would have to be taken immediately after the workload started, so that the cache hit rate would be zero. However, in that scenario, for the first half hour of running at maximum load, the competitor could not perform any IOPS at all with average response time below 20ms. No matter how low the queue depth was set, no matter how few IOs were done, the average response time remained above 20ms.

Figure 5 is a blow-up of the data contained in the Evaluator Group paper page 19 Figure 6, for the scenario that the data set in question has been idle for a long time. Thus for both products it starts out on HDD only.

Competitor exposure to low-performance HDD

Perception about competitor ↗

Little to worry about?
Starts faster, rapidly moves toward SSD performance?
- Write optimization causes turbo-charged HDD speeds?
- Caching reacts fast to changing workloads?
- ~~SSD is super-fast, so a little goes a long way?~~

Y axis: Relative performance that can be sustained without violating 20ms avg. read response time SLA

At hour 15 Competitor catches up; at hour 30 Dell regains lead, with 1.8x higher steady state performance

Reality about competitor

— Dell PS6210XS
— Competitor Reality

X axis: Hours of applying maximum IO rates to long-term idle 1.5TB
(40% writes, 95% random, 2:1 compressible, 22k avg. block size)

Reality about competitor:
- Slower than pure-HDD performance of PS6210XS for the first 8 hours
- 33% cache hit rate (at 8 hours) needed to match PS6210XS pure HDD performance
- Couldn't meet 20ms avg. response time SLA for first 30 minutes regardless of IO rate
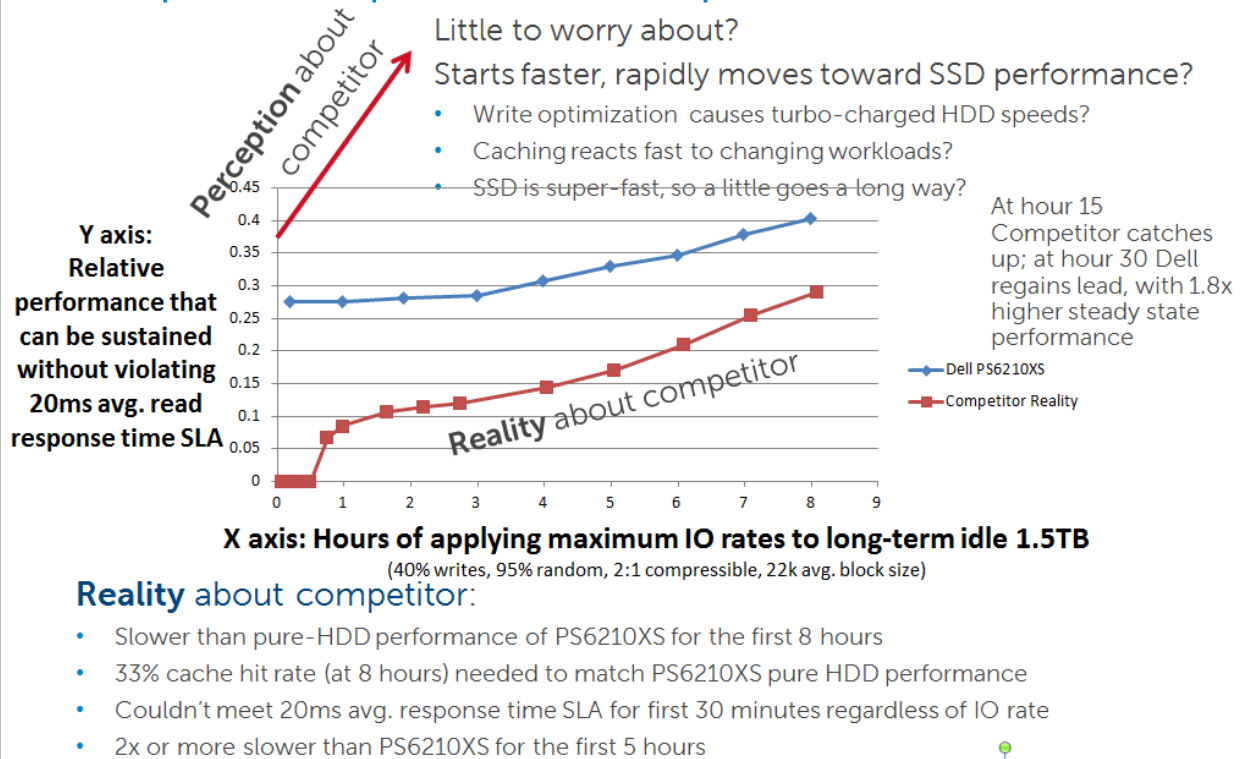- 2x or more slower than PS6210XS for the first 5 hours

Figure 5   Performance over time for a previously idle data set

Even when operated at steady state, at which point the metadata presumably is cached, the latency advantage of the faster HDD in PS6210XS resulted in lower latencies; the lower the latency SLA, the greater the advantage for PS6210XS. The below in Figure 6 is taken directly from the Evaluator Group paper.
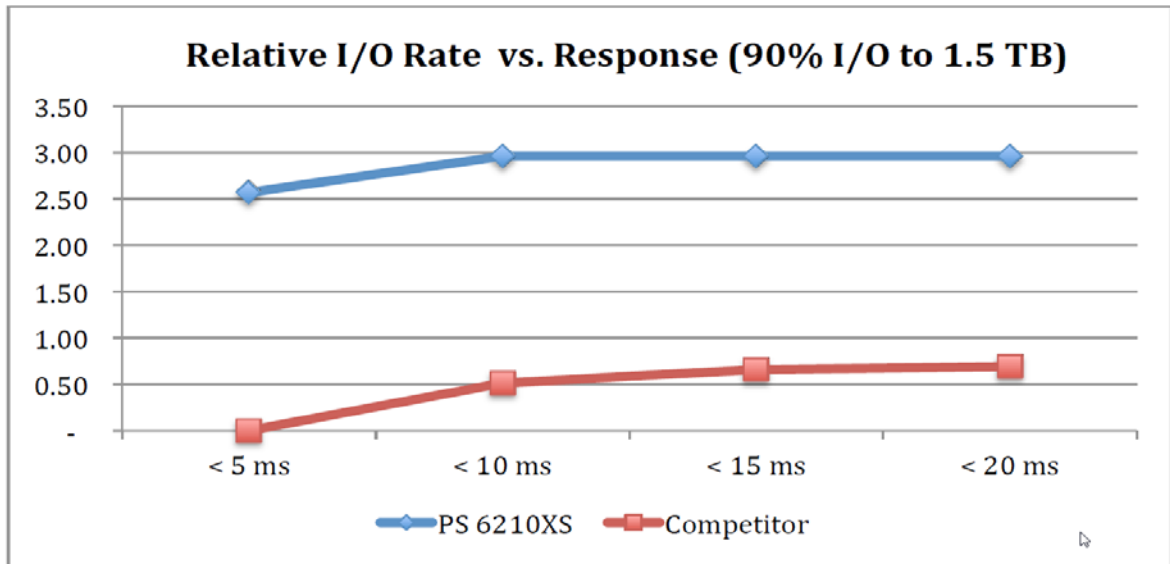
Figure 2: Dell PS vs. Competitor (Relative I/O at Response Time Thresholds)

Figure 6   IOPS rate as a function of average read or write response time requirement

Figure 6 is from page 14 of the Evaluator Group paper.

How do these effects play out in various use cases? Figure 7 shows the net result of various use cases tested in the paper.
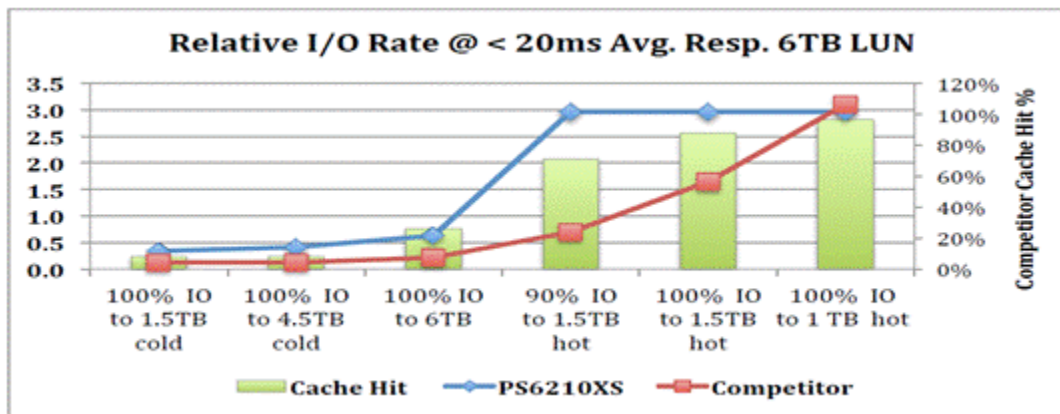

Figure 7   Relative performance in various use cases with different hot areas and IO concentrations

The graphic in Figure 7 appears as Figure 5 on page 16 of the Evaluator Group paper.

At the far left are HDD-dominated results. For the left-most two results, the competitor was tested at 8% cache hit rate that resulted as an effect of testing run immediately prior. The competitor, was dramatically slower than the PS610XS tested with HDD-only (far left), or mostly HDD (next result from the left).

The third result is for an even-IO pattern. The fourth result is for a more concentrated IO pattern. Finally, the two results at the right illustrate all IO to a small hot region. Only when all the IO was concentrated into 1TB of space did the competitor with 2.4TB of raw flash reach PS6210XS performance.   The competitor was only competitive at extreme cache hit rates that required very small hot data regions.

5   Effect of sudden loss of 25% of cache capacity

Evaluator Group did not test physical removal of a drive. However, the test data provides enough information about the relationship between performance and cache hit rate to illustrate what would happen if 25% of whatever data had previously been cached is no longer cached.

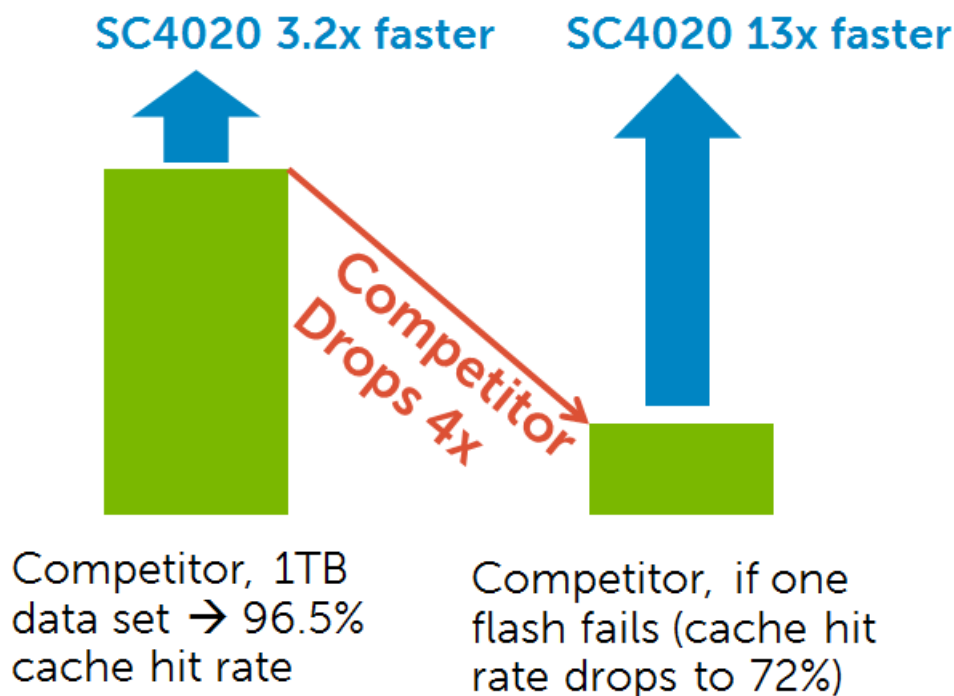Figure 8 compares a starting scenario of a 1TB data set.

Figure 8   Dell Storage SC4020 performance vs. competitor with or without all cache SSD's working

The 1TB example showing relative performance of 3.1 for the competitor and 9.8 for SC4020 is found in Appendix E of the Evaluator Group paper. Had one of the four flash drives failed, the new cache hit rate would be 75% x 96.5% = 72.4%. The Evaluator Group paper does not list the performance for that exact cache hit rate, but its Appendix E shows the performance at 71% cache hit rate is 0.7, and the performance at 88% cache hit rate is 1.6.   Linear interpolation between the observed 71% point and the 88% point leads to a calculated performance value of 0.77 at 72.4% cache hit rate, which is 4x lower than the pre-failure performance of 3.1.

If the required performance is lower, then the reduction is less. With a starting scenario using the larger 1.5TB data set in which the cache hit rate is 88% and the relative performance of the competitor is 1.6 (instead of 3.1), the resulting drop is 2.5x.

PS6210XS 1.9x faster   PS6210XS 4.5x faster

Competitor Drops 2.5x

Competitor, 1.5TB data set → 88% cache hit rate

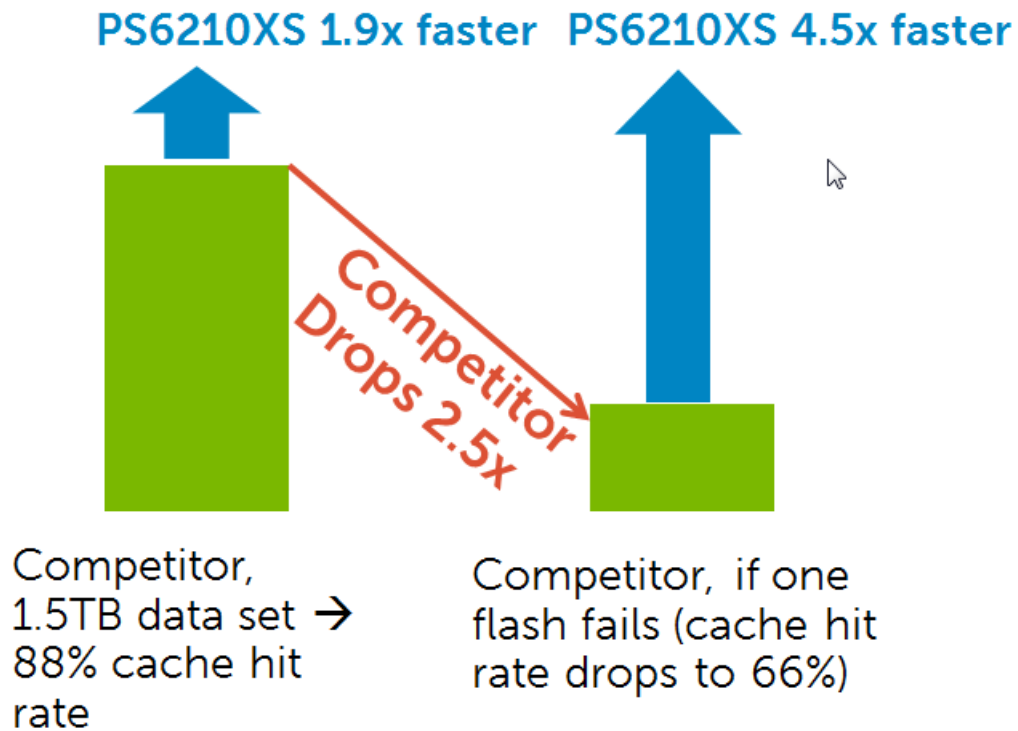Competitor, if one flash fails (cache hit rate drops to 66%)

Figure 9   Dell PS6210XS performance vs. competitor with or without all-cache SSDs working

The 66% cache hit rate relative performance is not included, but the 71% cache hit rate relative performance was discussed above; the 26% cache hit rate relative performance score is 0.2 in Appendix E of the Evaluator Group paper. Linear interpolation provides a calculated relative performance score for 66% cache hit rate to be 0.64, which is 2.5x lower than the starting performance of 1.6.

6 Alternating workload scenario

Figure 7 on page 21 of the Evaluator Group paper shows the performance of two different workloads based on how much time has passed since the system switched from running one of them to running the other of them.

To put the data into a use-case context, consider "Workload 2" to be the "weekday" scenario. There is one hot area and one cold area. Consider "Workload 1" to be the weekend scenario. The data that was accessed during the weekday is still accessed over the weekend, but is dominated 20:1 by IO done to a different hot and cold area during the weekend.

In Figure 10, the time scale is changed to depict graphically that workload 2 was active before switching to workload 1, and then "Monday morning" it shifts back from workload 1 to workload 2.

Tiering avoided the caching Monday-Morning problem

Dell PS6210XS (tiering): no change. Hot data for both periods stayed in flash
Remembered all workweek which data was hot for weekend;
remembered all weekend which data was hot for weekday

Competitor "Monday Morning" Problem

Y axis: Max relative IOPS (without regard to meeting response time SLAs)

X axis: Hours from start of 48-hours of weekend workload

Weekday "app": 1.25TB gets 90% of IOPS, 1.75TB gets 10%. Weekend: 95% to different 1.25TB hot and 1.75TB cold, 5% to the weekday "app". % random: Hot: 95%, cold: 90% Both: 40% write, 22k avg. blo
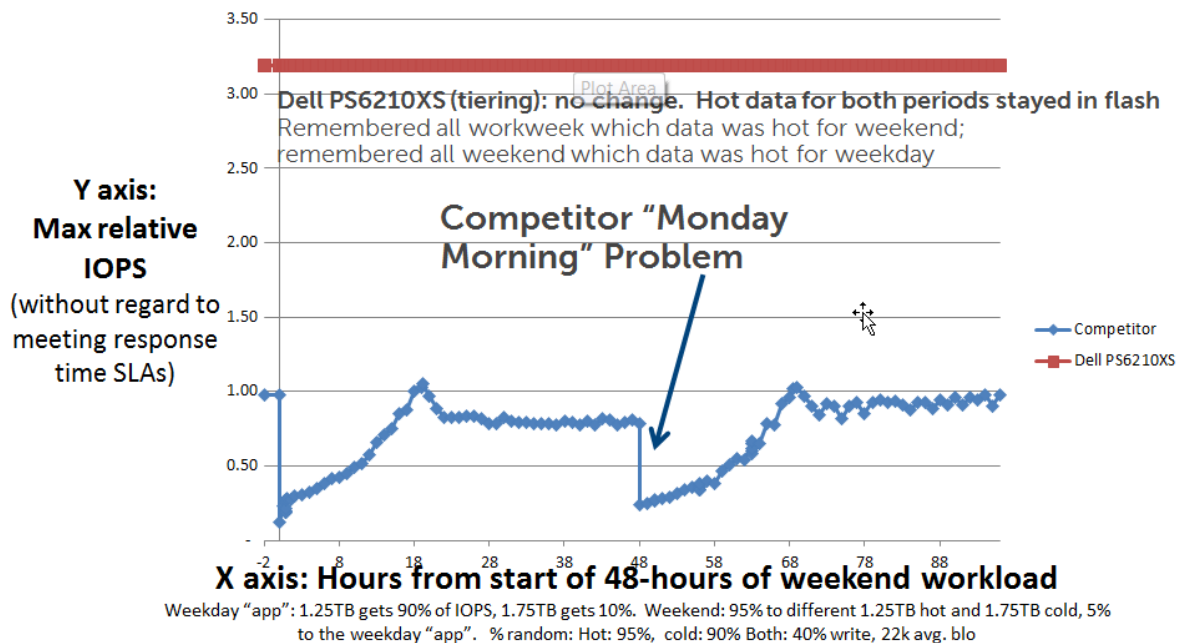
Figure 10   Monday AM scenario of alternating workloads: Dell PS6210XS vs. competitor

At hour 48, the workload switches back to the weekday load. The hour-50 relative performance is 0.27, which is 3x lower than the hour-47 performance of 0.81. The PS610XS performance was 3.18, which is 12x higher than the competitor Monday morning (hour-50) scenario. SC4020 was not tested in this scenario, but since the data was all in flash either way, appendix E of the Evaluator Group paper indicates that the performance would be between 8.3 and 10. Taking the most conservative case, SC4020 is 30.7x faster than the competitor 2 hours in to Monday morning.

Because data moves into cache in response to activity, rather than based on the hour of the day, whether there is a time gap between the weekend load ending and the users becoming active Monday morning is irrelevant. For the same reason, if the IO starts out light, data moves only slowly into flash. The effects of slower ramp-ups of activity were not tested but one can see that performance remains low until the total amount of IO that has occurred represents hours of the array running as fast as it can, or many more hours of the array being lightly loaded.

7 Bulk data movement examples

Many analytical processes involve scanning large amounts of bulk data as fast as possible. Such data typically resides in HDD because it receives very little random IO. A hybrid system for storage consolidation should handle both high IOPS and good MB/sec.

For such bulk data movement, 256KB block sequential IO is typical. To simulate more than one job running at the same time, the test used four parallel streams for writing and for subsequent read-back. The details are in Figure 8 on page 22 of the Evaluator Group paper. In Figure 11 we've restricted it to the most relevant comparison points, ignoring 1000:1 compressible data from IOmeter.

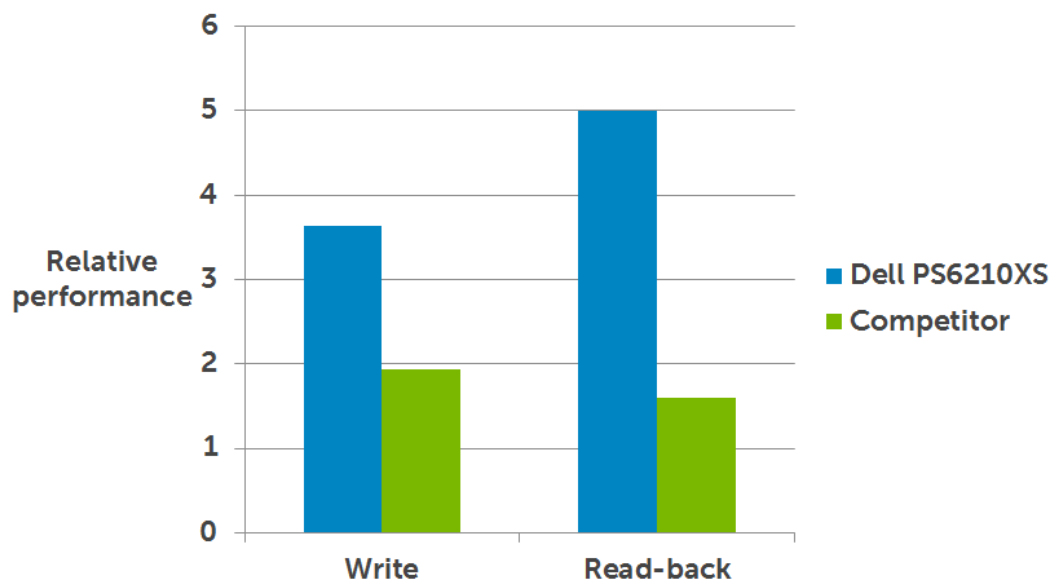The test results include a comparison of the competitor to PS6210XS:

Figure 11   Bulk data transfer rates, Dell PS6210XS vs. competitor

As noted in the paper, the SC4020 was not tested. It has Fibre Channel and iSCSI, whereas the competing system had only iSCSI, and the HDDs were not tested. However, given that the expandability of SC4020, in Figure 12 we compare its system maximums to what the paper reported for the competitor:

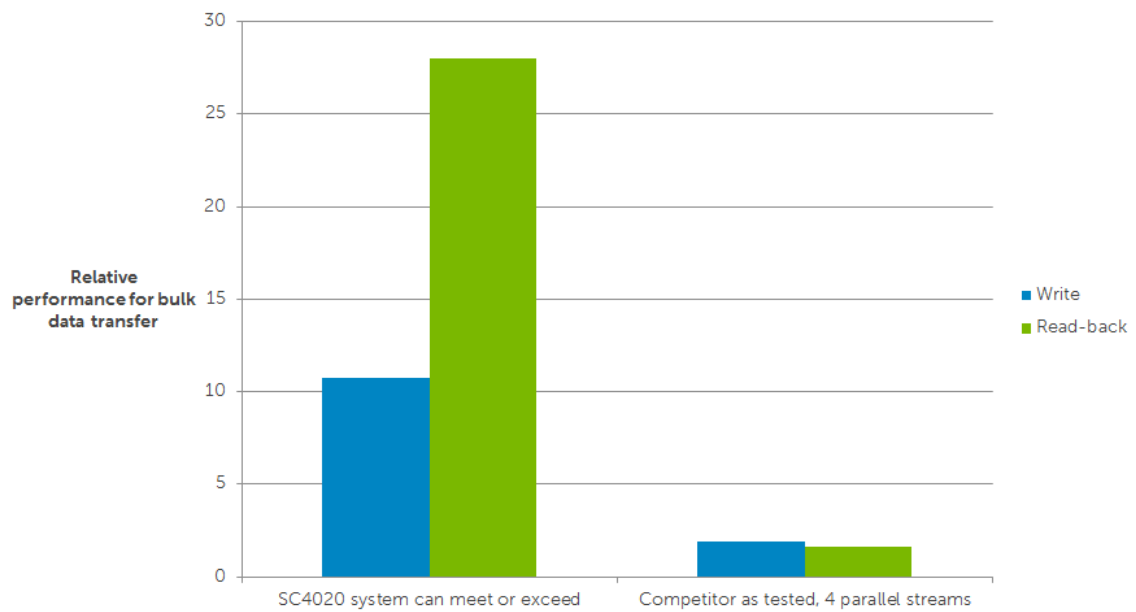# Dell SC4020 moves bulk data faster

Figure 12   Bulk data movement rates, Dell Storage SC4020 vs. competitor

A   Additional resources

The data used in this paper comes from an Evaluator Group paper, dated March 23, 2015, title:
Comparison of Hybrid Storage System Performance