the green grid®
get connected to efficient IT

Preliminary Assessment of Emerald Data

SNIA Emerald Metric Analysis Working Group
ENERGY STAR® Discussion November 18, 2015

Agenda

- Review of SNIA rationale for creating Emerald as an energy efficiency metric
- Composition of Current Dataset
- Categorization of Storage Products beyond the On-line Categories
- General Observations Regarding the Dataset and the Emerald Test Data
- Analysis of Emerald Data by On-line Categories:
  - On-line 2, 3, and 4
- Capacity Optimization Methods
- Next Steps

Note: The use of the word “score” refers to the performance/power efficiency score.

Composition of Current Dataset

- Data collected through July of 2015: Storage Products Certified to ENERGY STAR
  - 105 Configurations
  - 25 Families
  - 9 Manufacturers have equipment in the dataset.

On-line Category	Number of Families	Number of Configurations	Number of Manufacturers
2 Transactional	1	7	1
2 Sequential	2	9	1
3 Transactional	11	29	2
3 Sequential	3	9	2
4 Transactional	6	22	3
4 Sequential	7	28	4

Categorization of Storage Products underneath Online Categories

- Workload Type: Transaction, Streaming, Capacity
- Drive Type: HDD or SSD
  - HDD: Drive capacity, rpm and form factor
  - SSD Drive type - SSD, Flash, Non-volatile DIMM – and capacity
- Drive Count
- Connectivity:
  - Server to Controller
  - Controller to Storage Media
- Controller Cache Size

Note: Need these Groupings to get representative comparisons between products.

Examples of OL-3 and OL-4 with Top 25% Highlighted

Family	Config #	Device Form Factor (1.8, 2.5, 3.5)	Device Rated Speed (RPM)	Device Raw Capacity (GB)	Total Num Installed Storage Devices	Installed Solid State Devices	Installed Rotational Devices	Hot Band Workload Test (IOPS/W)	Ready Idle Workload Test (GB/W)
16	87	2.5	15000	300	24	0	24	4.2	11.3
17	91	3.5	7200	4000	12	0	12	5	122.3
12	67	3.5	7200	1000	12	0	12	5.5	19.4
11	64	3.5	7200	1000	12	0	12	6.4	23.3
14	75	3.5	7200	1000	12	0	12	8.7	28.1
13	70	3.5	7200	1000	24	0	24	8.8	29.1
13	71	3.5	15000	600	24	0	24	10.3	7.3
15	85	3.5	7200	1000	60	0	60	10.5	83
15	81	3.5	7200	1000	12	0	12	11.4	46.9
15	86	3.5	7200	1000	60	0	60	12.5	75.1
20	99	3.5	7200	2000	60	0	60	13.97	145.25
17	95	2.5	15000	600	12	0	12	14.1	20.3
12	68	3.5	15000	600	12	0	12	14.2	5.3
14	76	3.5	15000	600	24	0	24	19.3	6.8
11	66	2.5	15000	146	24	0	24	20.6	3.1
11	65	3.5	10000	600	12	0	12	21.7	15.5
12	80	2.5	15000	146	24	0	24	25.7	2.7
12	69	2.5	10000	600	24	0	24	27.3	12.5
17	92	2.5	10000	1200	24	0	24	28.4	36.6
13	74	2.5	10000	600	24	7	17	29	19.8
13	73	2.5	15000	146	24	0	24	29.1	2.8
13	72	2.5	10000	600	24	0	24	30.4	13
14	78	2.5	15000	146	24	0	24	33.1	2.7
17	93	2.5	10000	400; 600	24	7	17	34.4	31.5
15	82	2.5	15000	146	24	0	24	34.6	12
17	94	2.5	15000	300	24	0	24	35.6	18.3
14	77	2.5	10000	600	24	0	24	35.7	12.7
14	79	2.5	10000	600	24	7	17	49.7	17.7

OL-3 OPTIMAL TRANSACTIONAL CONFIGURATIONS

Family	Config #	Device Form Factor (1.8, 2.5, 3.5)	Device Rated Speed (RPM)	Device Raw Capacity (GB)	Total Num Installed Storage Devices	Installed Solid State Devices	Installed Rotational Devices	Ready Idle Workload Test (GB/W)	Average SR/SW
4	25	3.5	7200	1000	48	0	48	77.87	0.50
4	23	2.5	10000	450	48	0	48	48.75	0.75
19	98	3.5	15000	600	156	0	0	35.58	0.88
5	26	2.5	10000	450	88	0	88	34.32	1.37
5	27	2.5	10000	450	54	0	54	32.95	1.50
5	28	2.5	10000	450	42	0	42	33.71	1.94
7	40	3.5	15000	300	120	0	120	16.20	2.30
6	29	3.5	7200	2000	180	0	180	31.70	2.59
7	39	2.5	15000	300	150	0	150	29.00	2.89
9	45	3.5	15000	300	75	0	75	14.60	3.32
2	17	3.5	15000	300	90	0	90	15.10	3.71
24	115	3.5	15000	300	75	0	75	21.11	3.76
7	36	2.5	10000	600	75	0	75	51.00	4.61
24	111	2.5	10000	600	125	0	125	53.00	4.96
2	13	2.5	10000	600	125	0	125	52.80	5.01
8	41	2.5	10000	600	75	0	75	54.40	5.17
8	44	2.5	15000	300	75	0	75	23.50	5.32
24	114	2.5	15000	300	100	0	100	20.70	5.32
2	16	2.5	15000	300	100	0	100	25.20	5.50

OL-4 OPTIMAL SEQUENTIAL CONFIGURATIONS

Legend:
Green: Best 25% for that category
Other colors designate groupings by rpm, FF and Capacity

the green grid®
get connected to efficient IT

General Observations

- Idle is not a good indicator of operational energy efficiency.
  - Tested systems range 2% to 20% difference between maximum and idle power where that data is available.
- Assessment by taxonomy and workload type is biased to the higher performance drives.
- 7.2 K, high capacity drives are the most energy efficient but have lower performance.
  - Performance per watt will be lower than faster drives, but the capacity per watt will be up to an order of magnitude higher.
  - The mechanics of the test, which require full prep of all the drives, makes testing large 7.2 K high capacity systems uneconomic – it takes over a week to do drive prep.

General Observations

- The number of manufacturers and configurations by category is small.
  - Difficult to evaluate relationships between different component types within the database.
  - Very limited family data: hard to draw conclusions on the usefulness of the +15%/-40% test points.
- The relative product capabilities and configuration size of OL-3 and OL-4 products necessitate engineering judgement and interpretation to categorize a product.
- When comparing systems with the same drive size and speed, the following system attributes will influence the magnitude of the score:
  - Quantity of working memory and cache on the controller
    - In some cases, a manufacturer will combine cache and working memory.
    - In other cases, they are managed separately but perhaps with overlap.
  - Total number of drives
  - Number of servers pushing the data and number of front-end pipes
  - The connection types between the servers/controllers/drives
  - Controller architecture (cpu number and types, data movement capabilities, back-ends)

General Observations (cont.)

- Controller power and the number of supported drives affect scores:
  - Higher controller functionality increases power demand and may impact scores depending on the number and type of drives attached to the controller.
  - OL- 2 and small OL-3 partial or single drawer systems carry heavy power debt
  - Impact of controller power will be reduced on a scale up system with multiple drawers.

- Normalizing performance scores based on estimated controller power is fraught with peril:
  - Front-end connectivity (number and type of ports) – more and faster use more power but deliver more data to the controller in parallel up to the cache/buffering capacity
  - Number and types of back-ends – more and faster use more power but deliver data to/from more drives in parallel
  - Cache size can positively influence Hot Band scores

- Drive counts below and above the optimal point serve different purposes:
  - Up to the Optimal point, you are assessing the product on performance per watt
    - You will configure a storage product to enable growth to the optimum point and beyond.
  - Beyond the Optimal point, you are assessing on capacity per watt.
    - As you move beyond the optimal point, you are assessing capacity and TCO to determine when to add an additional storage product.

General Observations (cont.)

- Architectural and Configuration Differences can make a significant difference in scores:
  - Data discussion will show clear differentiation in some OL system comparisons with the same drive types.
  - Different architectures are used to address specific customer needs; may require separate categorization per our discussion in slide 6.

- Some Flash/SSD drives currently have a detailed, high priority clean-up cycle which may limit performance at higher capacities and drives counts and increase idle power. May not be universal, but is a significant impact on some systems.
  - This illustrates the impact of general data housekeeping on all storage systems.

General Observations: Emerald Test

- Testing is complicated, time consuming, and expensive:
  - Identification of the optimal point is difficult
    - Optimization is as much an art as it is a science.
    - Performance scores can vary by a factor of 2 to 4 based on the “excellence” of the tech setting up the system.
  - Loading of storage registers on large systems takes days
  - Single tests on large systems were taking a month or more.

- Storage testing is inherently difficult due to:
  - The inherent complexity of the systems.
  - The dependency of the performance on the choice and set-up of the system:
    - Matching storage allocation to servers (resetting configurations with different drive counts and workloads)
    - I/O capacity of servers and switches in test rig
    - Identifying optimal thread count on servers running workload

- Emerald is a workable energy efficiency test, but TGG/SNIA is evaluating options to increase testing efficiency.

General Observations: Transaction Tests

- All transaction tests were optimized for Hot Band workload
- Read and Write tests are not optimized and not representative of an “optimum configuration”
  - Reported Read and Write scores should not be evaluated
- Hot Band (HB) scores are highest on the higher speed/lower capacity drives in the same FF
  - 2.5” FF will have better scores than the 3.5” FF
    - Power scales with rotational speed and size
    - At all speed and capacity points, 2.5” drives have preferable performance/power characteristics as compared to 3.5” drives.
  - Drive capacity has variable affects depending on design and technology.
  - SSDs are known to offer better IOPs/Watt although limited data is available to demonstrate that
- Reporting of Cache capacity, VDBench version, and configuration in the ENERGY STAR database was inconsistent, making it difficult to analyze differences in the HB data.

General Observations: Sequential Tests

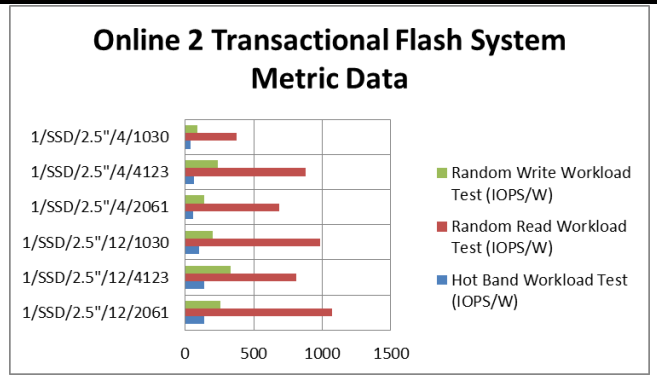
- The Sequential Test is Optimized across the read/write functionality
  - The optimized ENERGY STAR score is the average of the read and write scores.
  - Optimum configuration set-up is balanced to maximize the two scores.
  - Getting the balance of the average R/W score to get min/max performance/power within 20% of optimal is very difficult/time consuming.

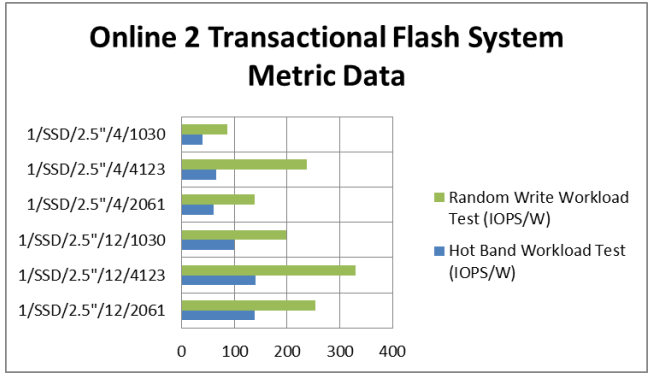
- Operations are moving to more reads than writes with time.
  - SNIA will watch the market to determine how storage use is changing.
  - As we continue to collect data, SNIA can evaluate different weightings using the data.

Data Analysis

All configurations shown are Optimal except for the family charts.

OL-2 Data Analysis: Transaction





Observations:

- For flash drives, the HB and Random Write IOPS/W score improve with increased Flash capacity.
    - The increase from IOPS/W increase is large from 1 – 2 TB, smaller from 2-4 TB
    - At 12 drives Random Read scores improve from 1-2 TB and degrade to 4 GB when compared to both 1 and 2 TB scores.

    - IOPS/W scores will reduce with higher drive counts on higher capacity drives because of “garbage collection” or clean-up that occurs on a milli-second cycle time.
- Quantity of Flash/SSD data is insufficient to draw broad conclusions, though Solid State devices performed better than uncached spinning devices.

Notes:

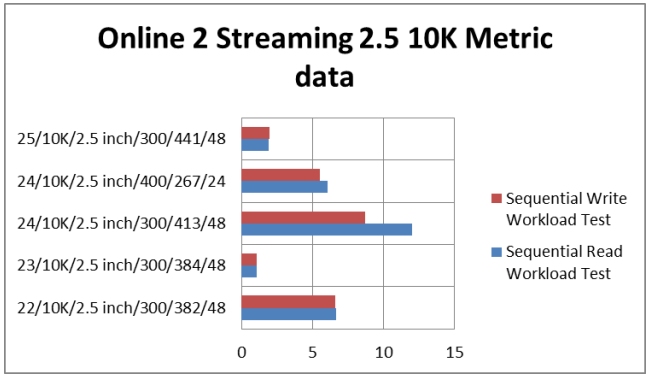
2nd Chart excludes Random Read to improve clarity of the HB score differences.

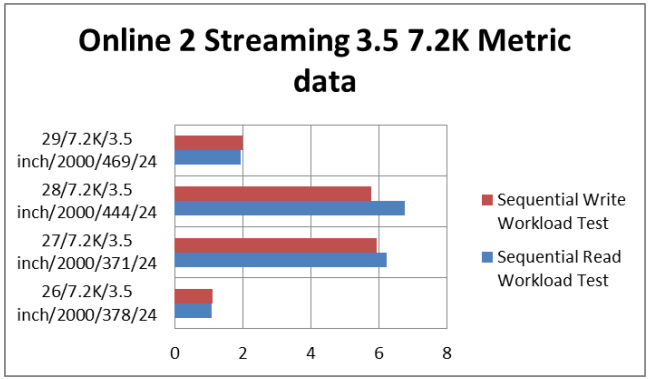
Legend is Family#/Device Type/FF/Device Count/ Device GB

the green grid®

get connected to efficient IT

OL-2 Data Analysis: Sequential





Observations:

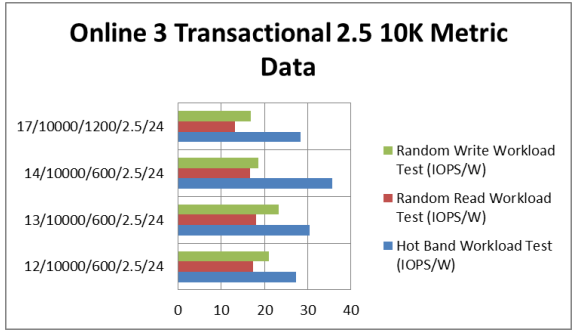
- The scores on the 7.2 K drives are likely lower than the scores for 2 of the 10 K drive systems because:
  - The amortized controller power is higher (estimated idle power is roughly equal, 7.2 K systems have half the drives).
  - There are 50% less drives available to increase performance

- There is also likely a configuration and architectural attributes driving the differences between the 3.5”/7.2K and 2.5”/10K configs.

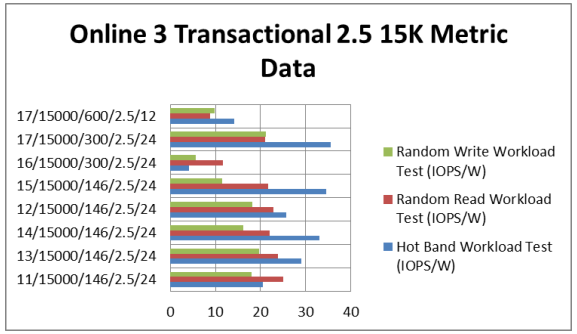
Legend Notes: Family#, rpm, FF, GB, Idle Watts, Device count
For Family 24, 24 drives is 20 HDDs and 4 SSDs

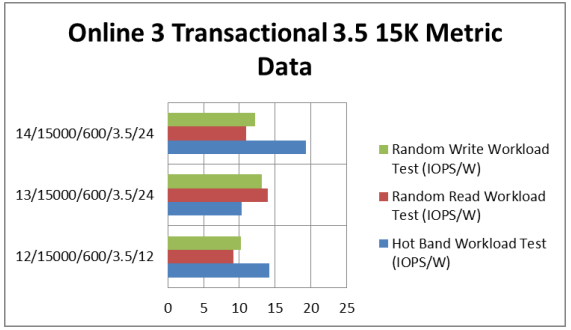
the green grid®
get connected to efficient IT

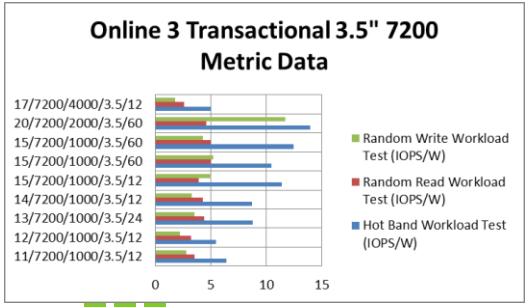
OL-3 Data Analysis: Transactional



- 2.5” HDDs have better scores than 3.5” HDDs by factor of 2 or better
- 12 drive systems have lower scores than 24 drive systems due to:
  - better amortization of controller power.
  - 15k rpm drives use significantly more power than 10k drives, lowering iops/w.
  - smaller drive counts may not fully utilize the data pathways (non-optimal).
- 3.5” 7.2 K capacity drives have lower transactional scores than 3.5” 15 K drives. This is due to IOPS tracking rotational latency very closely
- There is differentiation between systems with same drive speed/capacity/count. This is likely due to architecture of the system/controller.



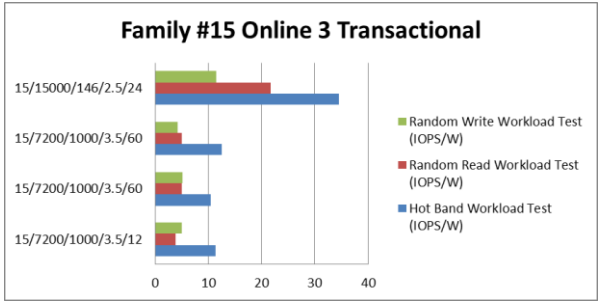


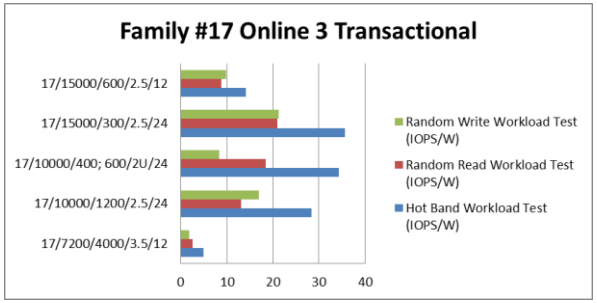


Legend Notes: Family#, rpm, FF, GB, Device count

the green grid®
get connected to efficient IT

OL-3 Family Analysis: Transactional



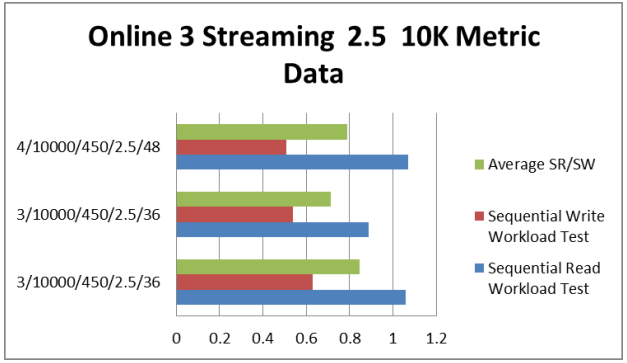


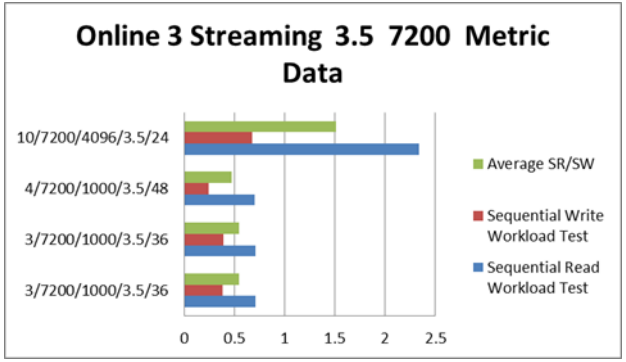
Legend Notes: Family#, rpm, FF, GB, Device count

Observations:

- Family #15: 7.2 K drives with count of 12 and 60 shows HB w/i 20%.
  - The top 60 device systems use different variations of the controller.
  - High variation likely caused by smaller number of drives to cover controller power. The other 12 drive count configuration shows the same low score.
- Family #17: 15 K drives with count of 12 and 24 shows HB outside of 20%
- The center configuration for family #17 has 7 SSD and 17 10K HDD devices.
  - The configurations with mixed SSD/HDD devices had better scores than HDD only configurations with same components. There are several other OL-3 configs with mixed SSD/HDD drives that show the same behavior.
  - The matching configuration with 24 2.5” 10 K drives is below the HDD/SSD mix.

OL-3 Data Analysis: Sequential





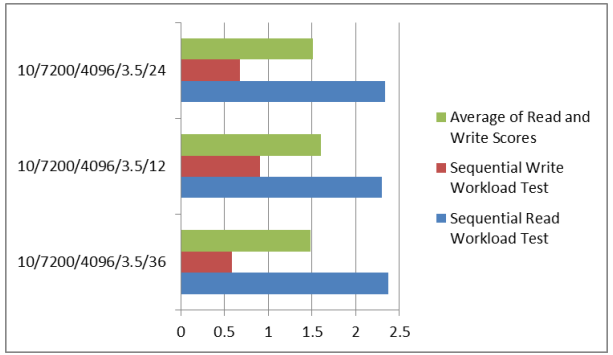
Observations:

- For family #3, the configuration with the higher scores for both device types has a single controller, the other configuration has a dual controller.
- Family #4 has a single controller for both device types.

Legend Notes: Family#, rpm, FF, GB, Device count

the green grid®
get connected to efficient IT

OL-3 Family Analysis: Sequential

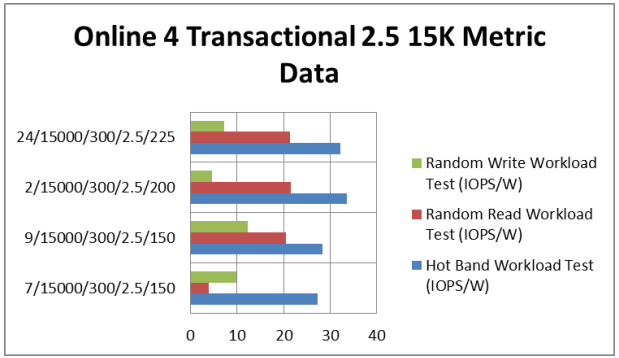


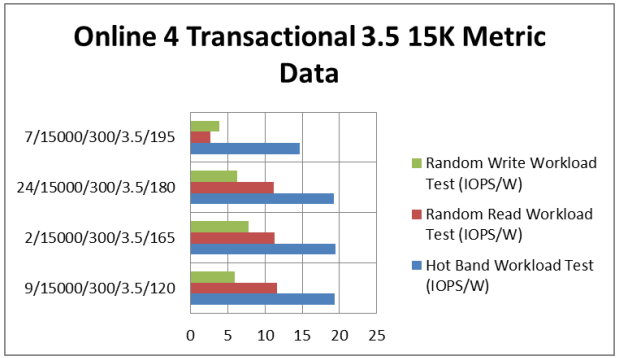
Family #10 On-Line Sequential

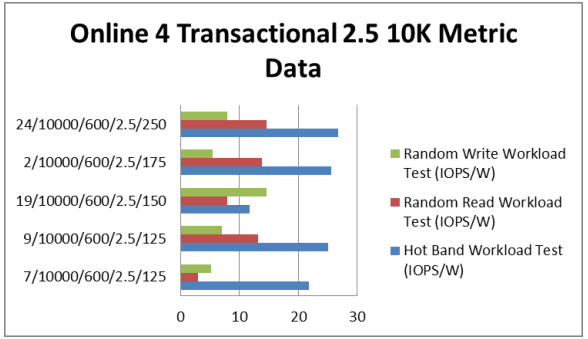
- Read test is w/i 20% for performance scores, write tests are not.
- Average Read/Write score decreases from 12 to 36 drives:
  - 12 drives is optimal
  - 24 and 36 drives are 2 maximum points, both within 20% of the optimal score.
    - ❑ 24 drives are within 7.7%
    - ❑ 36 drives are within 7.8%

Legend Notes: Family#, rpm, FF, GB, Device count

OL-4 Data Analysis: Transactional





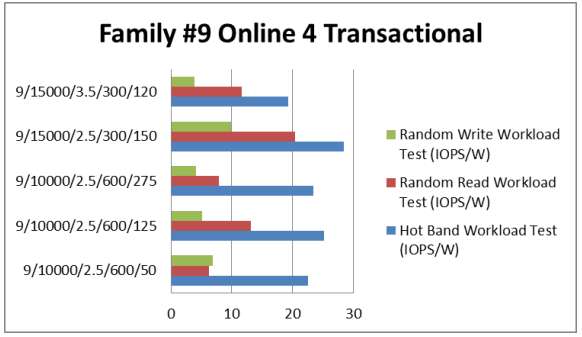


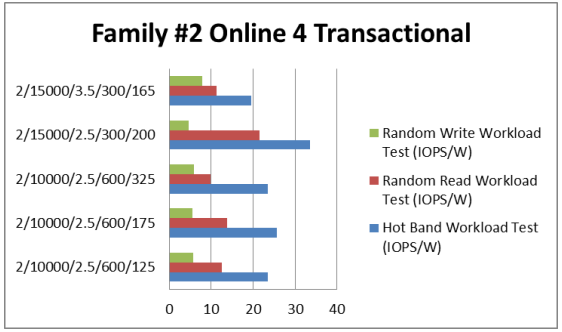
- 2.5” HDD has higher transactional efficiency scores than a 3.5” HDD due to requiring less power for same data movement.
- Higher rpm drives within the same from factor (FF) have better transactional scores. Due to faster transfer rates for same power draw.

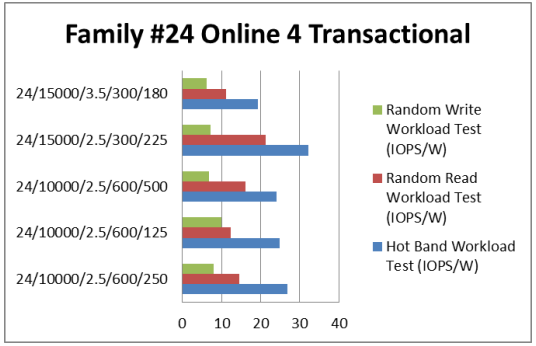
the green grid®
get connected to efficient IT

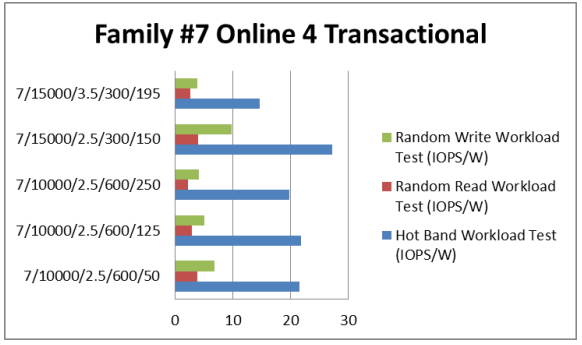
Legend Notes: Family#, rpm, FF, GB, Device count

OL-4 Family Analysis: Transactional







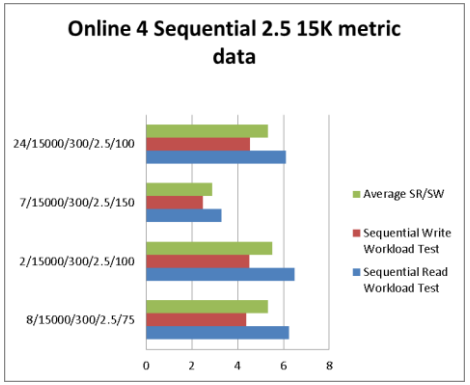


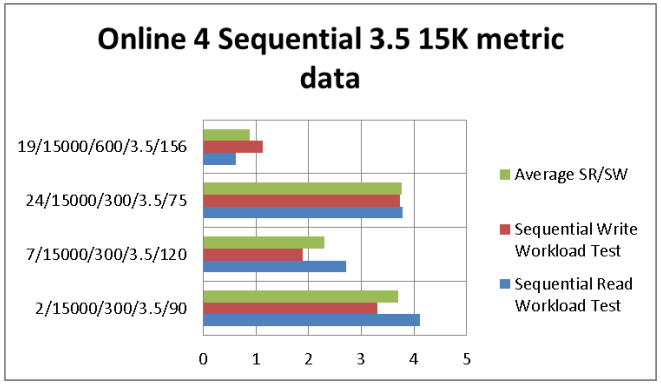
- For the 4 families, the HB min and max configs are within 20% of optimum.
- For the 4 families, the HB scores are reasonably close, but there is some differentiation.
- These 4 families, by design, span a large portion of the OL-4 space. The goal is to focus on different customer cost points based on total capacity and maximizing total performance for that capacity.
- Maximizing the qualified range gives customers the best value but requires at least 3x the effort to find the min and max points that maximize the qualified drive count range.

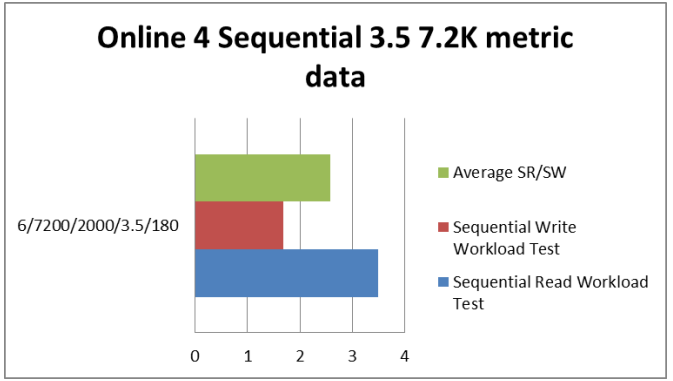
Legend Notes: Family#, rpm, FF, GB, Device count
Min/Opt/Max families are systems 5 to 3 on each graphic.

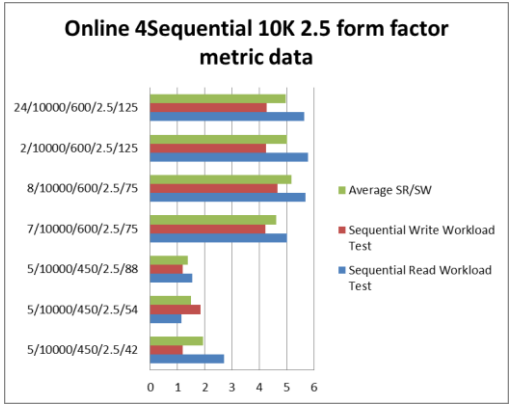
the green grid®
get connected to efficient IT

OL-4 Data Analysis: Sequential









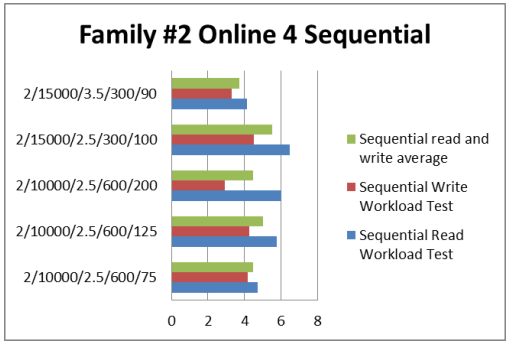
Observations:

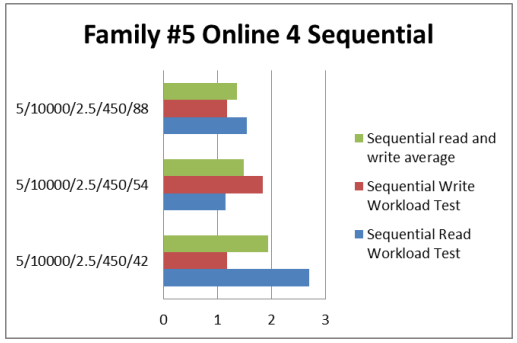
- For 2.5” drives, 15 K drives give better scores than 10 K drives
- 10 K 2.5 drives have two distinct groupings of scores.

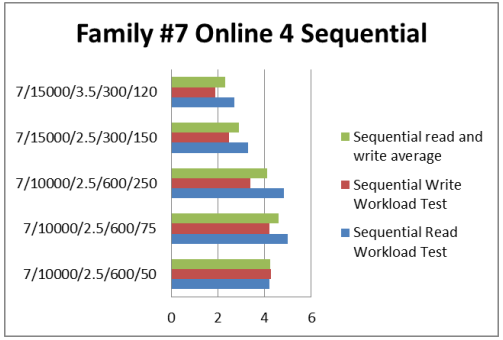
Legend Notes: Family#, rpm, FF, GB, Device count

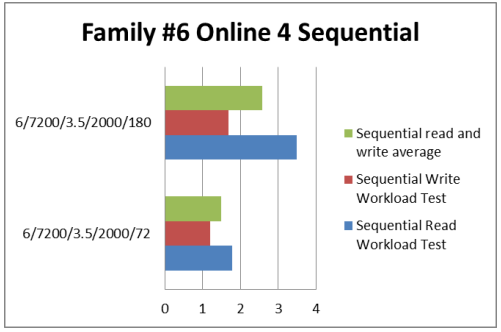
the green grid®
get connected to efficient IT

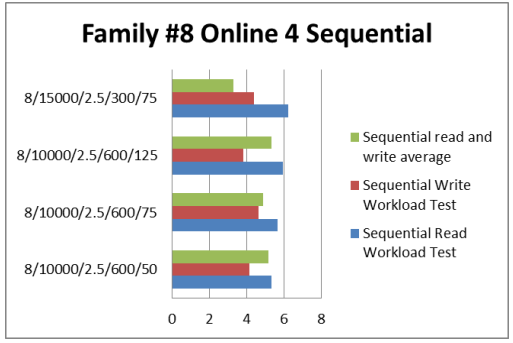
OL-4 Family Analysis: Sequential

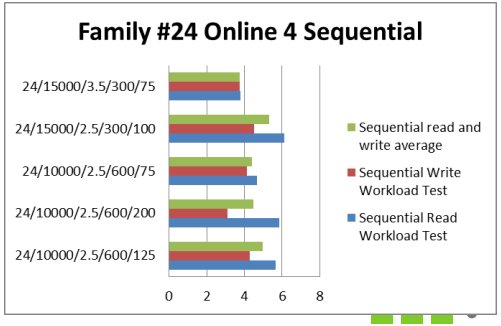












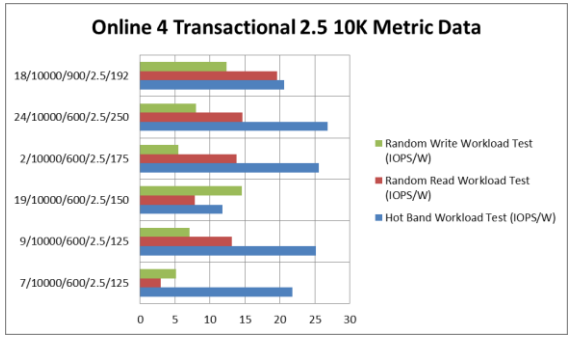
Observations:

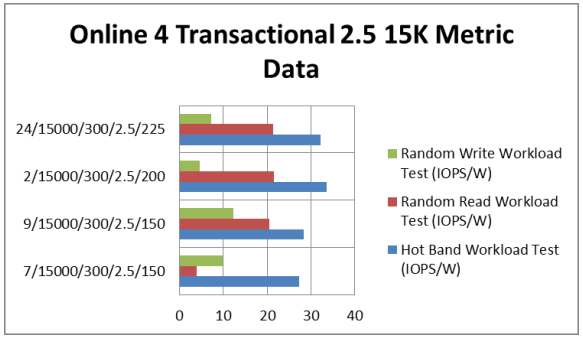
- Family #6 is a scale out with a controller for every 12 drives.
- In general, conclusions for sequential systems match transaction systems.

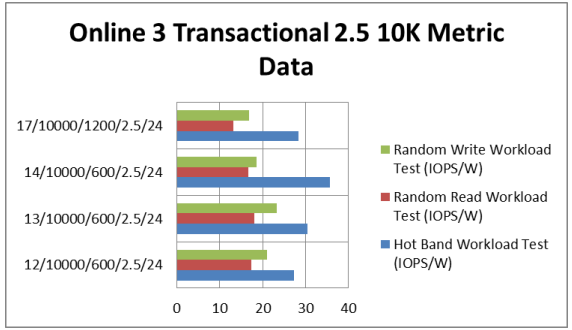
the green grid®
get connected to efficient IT

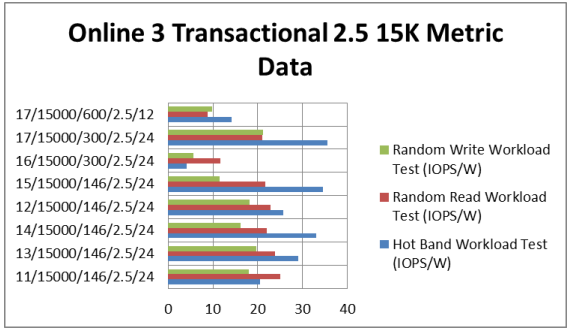
Legend Notes: Family#, rpm, FF, GB, Device count

OL-3 to OL-4 Comparison: Transactional







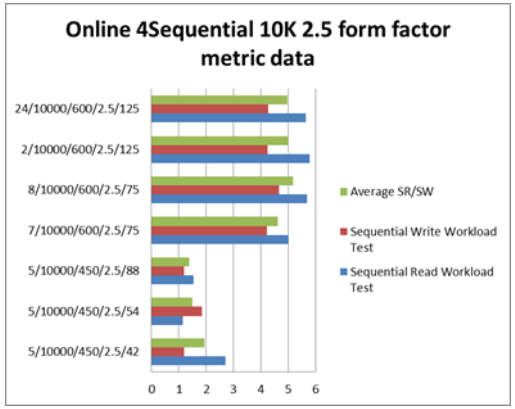


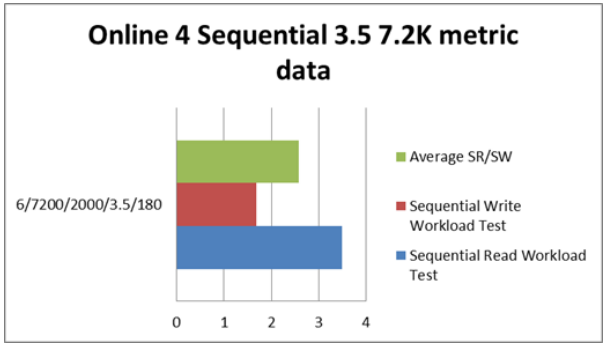
- Many OL-3 systems are single controller, while all OL-4 are dual
  - OL-4s draw more controller power
- 15k rpm drives consistently deliver more iops/watt than 10k drives in a form factor regardless of OL type
- For larger data sets and/or improved resilience, OL-4 systems provide more total performance for a small increase in power.
- 1 OL-4 system can handle as much storage as 8 OL-3 systems, resulting in lower overall costs and power use.
- Many applications cannot spread data across multiple storage systems, necessitating use of larger systems or systems that distribute the data for the application.

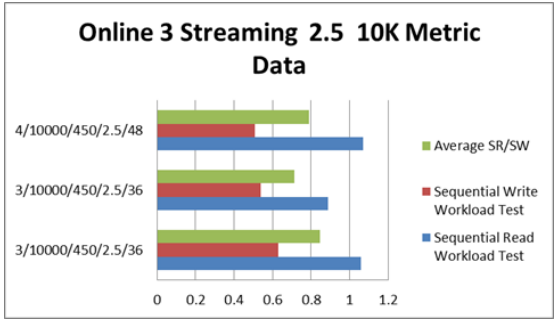
Legend Notes: Family#, rpm, FF, GB, Device count

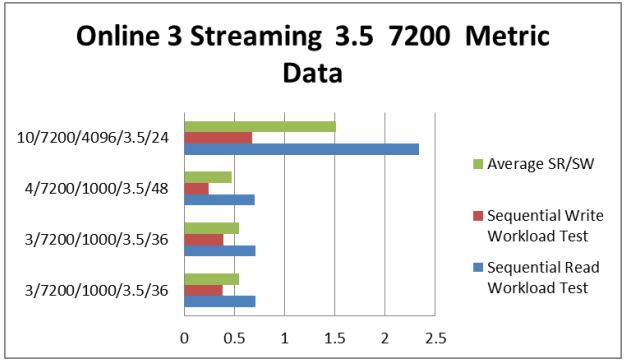
the green grid®
get connected to efficient IT

OL-3 and OL-4 Comparison: Sequential









The OL-4 systems are showing better GBS/W performance compared to OL-3 systems with comparable drive counts. This could be caused by:

- more back-end buses for parallel data delivery
- stronger data movement and CPU controller components
- more front-end pipes

Legend Notes: Family#, rpm, FF, GB, Device count

the green grid®
get connected to efficient IT

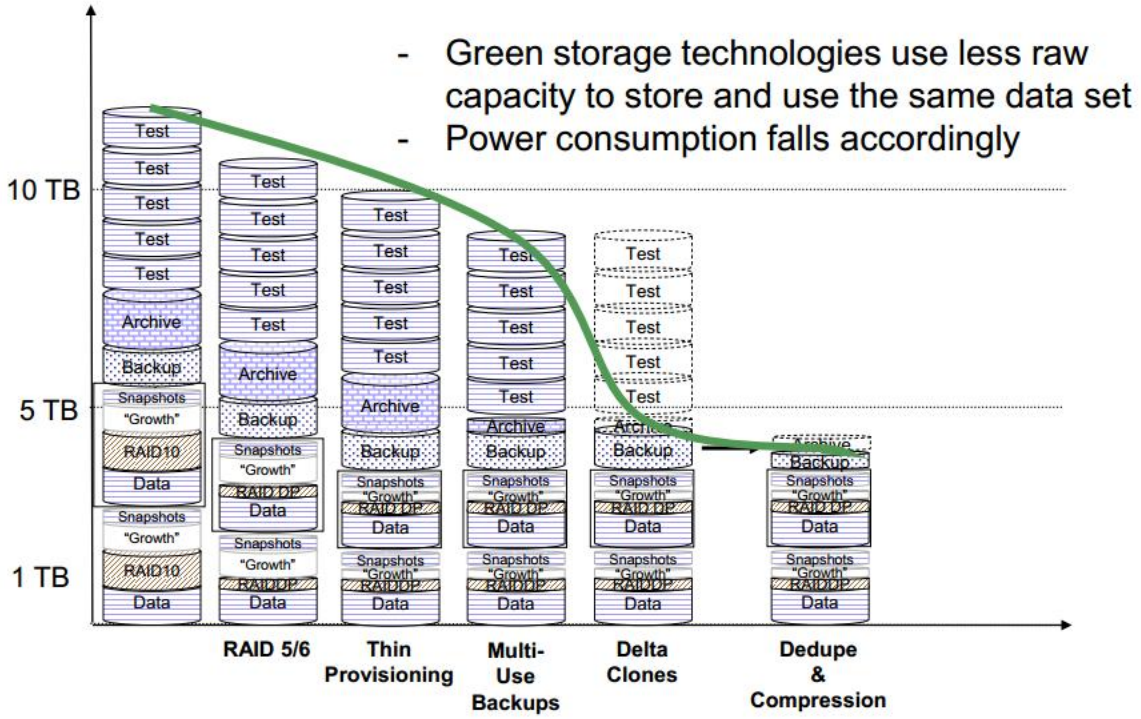
Test Ranges and Families

- In OL-4 systems, optimum point ±20% excludes a large range of usable capacity even with rounding leave many entry configurations outside of qualified range.
- Min/max testing to expand the range is extremely expensive but failing to do so makes many configurations unavailable as ENERGY STAR qualified.
- Could this contribute to lower than desired participation and/or penetration?
- Segregation by workloads will create difficulties in identifying certified products.

Summary of COM Space Savings

- All COMs allow you to store more data in less space (less physical storage devices= power savings)
  - "Your mileage will vary" based on application uptime, data set types, performance objectives, etc…
  - COMs may reduce physical capacity per unit energy of a single system, but significantly improve data center level efficiency by reducing number of storage products required to store a given quantity of data. (lower GB/watt but lower power draw and energy use)
- Parity RAID (now typically RAID 6)
  - Replacement for mirroring, with trade-offs for speed - rebuild, recovery, etc
  - Usually ~40% space savings over RAID 1
- Thin provisioning
  - Can take systems from 30% utilization (legacy) to 80% (some production data centers practice 300% oversubscription)
- De-duplication
  - Savings depend on several factors, can be large (25-40% primary; up to 50% secondary; coupled with compression)
  - e.g., Think of backing up thousands of laptops, all originally burned from the same master image
- Compression
  - Savings vary with data characteristics, can be large
  - As compression is local to a file or block, it can't achieve what de-duplication can.
- Delta snapshots
  - Larger savings possible when change delta is small (compared to PIT copies)

Effect of COM Technologies



the green grid®
get connected to efficient IT

Conclusions

- Data shows this is complex
- The different categories and drive types/sizes exist to meet the range of application needs
  - If ENERGY STAR makes comparisons across categories, customers will go outside the ENERGY STAR program to meet their needs
- The cost and complexity of testing limits the number of systems tested – testing that customers are expecting/demanding
  - SNIA and TGG collaborating to prepare alternative testing approaches using Emerald
  - Modeling may offer a more workable option.
- COMS make a significant difference in data center energy efficiency but may result in a “less efficient” single storage product in terms of one or more metrics.
- Too little data to draw conclusions within categories.
- There is a wide variation in system architectures and controller components in storage
- OL-3 and OL-4 differ not just in scale, but also in resilience
  - OL-4’s lack of a SPO results in higher power needed to meet customer requirements

TGG Forward Looking Information Disclosure Statement

This TGG presentation is made as part of the industry EPA ENERGYSTAR Data Center Storage Stakeholders Meeting November 18, 2015. It may include timetables, roadmaps, new technologies entering the mainstream, predictions, estimates or other information that might be considered forward-looking. While these forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual timeframes and results to differ materially. Readers are cautioned not to place undue reliance on these forward-looking statements, which reflect our opinions and best effort planning and/or understanding only as of the date of this presentation. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forward-looking statements in light of new information or future events. Throughout the discussion occurring as part of the delivery of this presentation, we will attempt to illuminate some important factors relating to the topic that may affect our estimates and predictions.

References

- Link to SNIA Emerald training page
- Emerald Training Introduction to COMs.pdf
- Storage Considerations in Data Center Design
- Green Storage Technologies 2 by Alan Yoder