ENERGY EFFICIENT DATA CENTER STORAGE:
AN ASSESSMENT OF STORAGE PRODUCT POWER EFFICIENCY

July, 2022

EDITOR:

Donald Goddard	NetApp, The Green Grid, SNIA Green Storage Technical Working Group

CONTRIBUTORS:

Bill Bailey	Seagate Technology, SNIA Green Storage Technical Working Group
Jay Dietrich	IBM (retired)
Rona Newmark	Dell (retired)
Chuck Paridon	SNIA Green Storage Technical Working Group
Carlos Pratt	IBM, SNIA Green Storage Technical Working Group
Patrick Stanko	SNIA Green Storage Technical Working Group
Herb Tanzer	SNIA Green Storage Technical Working Group
Dave Thiel	SNIA Green Storage Initiative

The Green Grid Data Center Energy Efficiency Technical Working Group (DCEE TWG) acknowledges the Storage Networking Industry Association (SNIA®) Green Storage Initiative (GSI) and Green Storage Technical Working Group (Green TWG)'s strong collaboration and significant contributions in gathering and analyzing the data presented in this paper.

Executive Summary

The Green Grid Association's — a consortium that works to improve IT and data center resource efficiency worldwide — goals are to evaluate the efficacy of the SNIA Emerald™ test suite as a tool to assess the energy efficiency of storage products and to consider its application and use in developing and setting performance/power efficiency thresholds for storage products. Toward these goals, the Green Grid sponsored the SNIA analysis effort reported in this white paper.

The Storage Networking Industry Association (SNIA®) released Version 4.0 of the SNIA Emerald™ Power Efficiency Measurement Specification on July 3, 2020, in support of the publication of the ENERGY STAR Data Center Storage V2.0 Program which was released May 28, 2020. The SNIA Green Storage Technical Work Group (Green TWG) collected the SNIA Emerald results published by the EPA for 160 systems certified by the Energy Star Data Center Storage Program, as well as additional idle measurement data. The SNIA Green TWG analyzed this data to understand the impacts of configuration type and component selection on the SNIA Emerald performance power metrics for the three workload types tested: Capacity, Sequential and Transactional. This white paper details the working group's findings and provides storage product manufacturers, regulators, and other stakeholders information on the efficacy of the SNIA Emerald methodology for assessing the power efficiency of data storage products.

This white paper makes recommendations for regulators regarding what product criteria and test metrics should be used to assess power efficiency. It also provides information regarding how Online Data Storage Products work and the various characteristics of Online Data Storage Products. This information includes how they are configured, what capacity optimization techniques are incorporated, how the integrity and security of the data stored is ensured, and what goes on during the ready idle state. Information regarding new trends in enterprise storage is also provided.

Table of Contents

Table of Figures

Table of Tables

1. INTRODUCTION

The SNIA Green Storage Technical Working Group has analyzed the SNIA Emerald™ Power Efficiency Measurement test (SNIA Emerald test) data for storage products certified to the ENERGY STAR Program Requirements for Data Center Storage products. The data set includes 160 systems certified on the Energy Star Data Storage website across the Online 2, 3 and 4 product categories. The analysis has evaluated the efficacy of the SNIA Emerald test suite as a tool to assess the power efficiency of storage products and considers its application and use in developing and setting power efficiency thresholds for storage products. Based on the analysis, recommendations for data center storage power efficiency characterization are offered. These include simplified testing and product criteria.

Data centers typically have service level requirements that must be met; the selection of storage and other components is driven by the need to meet these requirements. Within the set of products that meet these requirements, power efficiency is a consideration. Power efficiency metrics provide the basis for comparing various storage products and the comparisons enable more effective utilization of power. While the power efficiency metrics do not allow the direct prediction of actual data center power consumption, they are a meaningful relative measure of power consumption, i.e., for a given workload, products with higher efficiency metrics can be expected to consume less power than products with lower efficiency metrics.

2. OVERVIEW OF STORAGE ARCHITECTURE

How Storage Architectures Vary

There are three basic types of traditional storage system architectures: Direct Attached Storage (DAS), Storage Area Network (SAN), and Network Attached Storage (NAS) (Figure 1). These architectures have some overlapping elements, as will be explained.

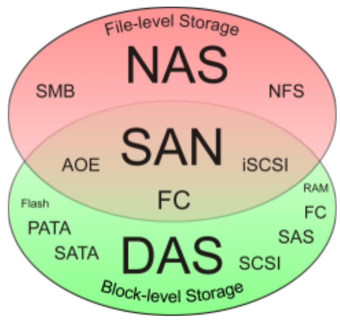

Figure 1: Basic Storage Architectures

DAS is the most basic storage system that provides block-level storage and is directly attached to a server or workstation without a storage network in between. Several examples of DAS systems are shown in Figure 2. The first example has four Small Computer System Interface (SCSI) Hard Disk Drives (HDDs) attached to the host computer via SCSI cabling; the second example uses Fibre Channel cabling to connect the host computer to a Redundant Array of Independent Disks (RAID) or Just a Bunch Of Disks/Devices (JBOD) storage subsystem. A RAID system uses a controller to manage the storage devices and improve the reliability characteristics of the storage system while a JBOD system uses the HDD or SSD (Solid State Drive) controller to manage data between the workstation/server and JBOD system. Only the former approach is relevant for data center implementation. The key characteristic of DAS is the binding of storage resources to the individual computer/server. This is an inexpensive storage solution, but the dedicated storage resource can cause some limitations. These include:

- storage capacity limited by the number of HDDs supported by the bus and enclosure design;
- low efficiency of the storage resource usage in that content may have to be replicated rather than shared between servers and the inability of free storage resources on one server to be used by another server whose disk space is running low;
- limited availability because any server failure results in the attached storage resources becoming inaccessible; and
- performance limited by the processing speed of the individual server, i.e., parallel processing to share the workload among multiple servers is not possible.

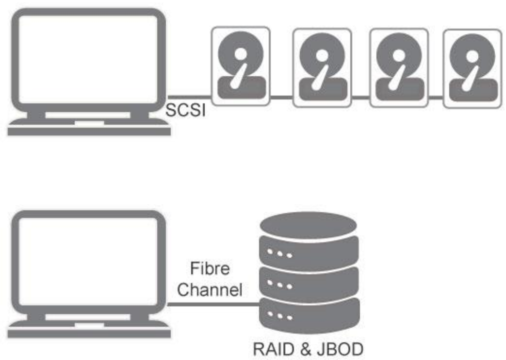

Figure 2: Direct Attach Storage (DAS) Examples

NAS and SAN (see Figure 3 and Figure 4) are two ways of sharing storage over a network and thus have some inherent advantages over DAS. SAN offers a higher level of functionality than DAS because it permits multiple hosts (servers) to attach to a single storage device at the block level. It does not typically permit simultaneous access to a single storage volume, but it does allow one server to relinquish control of a volume and then another server can take over the volume.

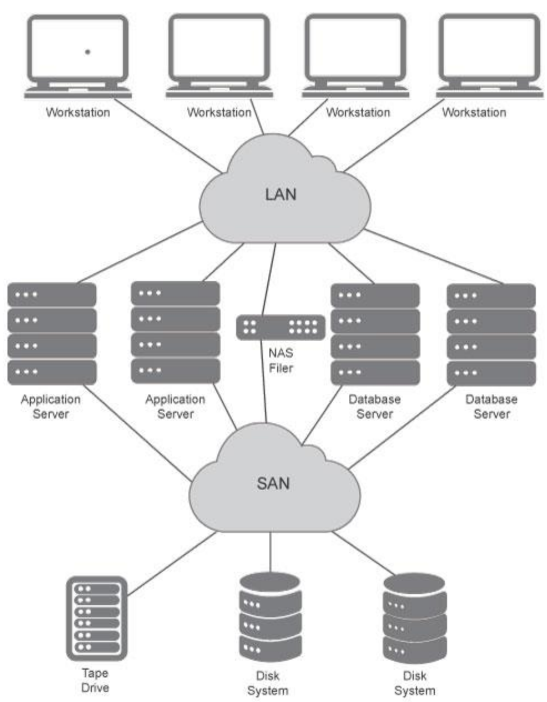

Figure 3: Storage Area Network (SAN)

NAS is basically a “file server” that is built on top of SAN or DAS technology and is accessed over a network (LAN) using network file system protocols such as NFS and CIFS. The difference between NAS and SAN is that NAS does “file-level I/O” while SAN does “block-level I/O” over the network. This means that data is transferred across the network to the recipient as either blocks (SAN) or a file data stream (NAS). The file access model is built on a higher abstraction layer and thus requires an extra layer of processing both in the storage server and in the function of translation between file and block accesses in the NAS Storage Server. This results in higher processing latency and impacts I/O throughput in many applications, as compared to SAN block level access. The benefit that comes with the higher-level abstraction in NAS is ease-of-use. Shared storage can be readily implemented by connecting the NAS system to a familiar enterprise LAN (e.g., Ethernet) and configuring the OS on the workstations and servers to access the NAS Storage Server. An additional advantage of a NAS over a SAN or DAS is that multiple clients can share a single volume, whereas SAN and DAS volumes can be mounted by only a single client at a time. Operating systems such as UNIX and LINUX support NFS protocols; Windows OS’s support CIFS protocols.

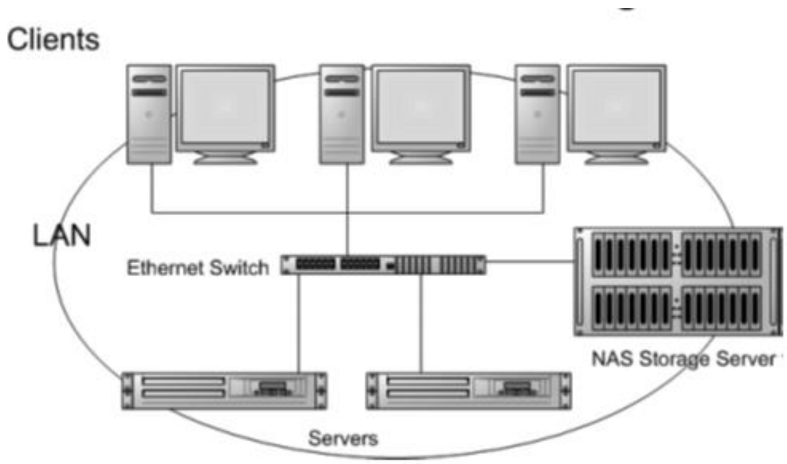

Figure 4: Network Attached Storage (NAS)

How Storage Systems Scale: Scale-up vs. Scale-out

Legacy or traditional systems use “scale-up” storage (Figure 5) to add capacity**.** For many years, the standard storage array had two storage controller heads (for redundancy) and multiple disk shelves holding media devices. The storage controllers connect to the storage area network (SAN) and provide storage to the compute servers. All of the disk shelves are connected to the storage controllers and all compute servers access these storage devices through the two controllers. To increase the capacity and performance of the storage array, more storage device shelves and storage devices are added to the same pair of controllers.

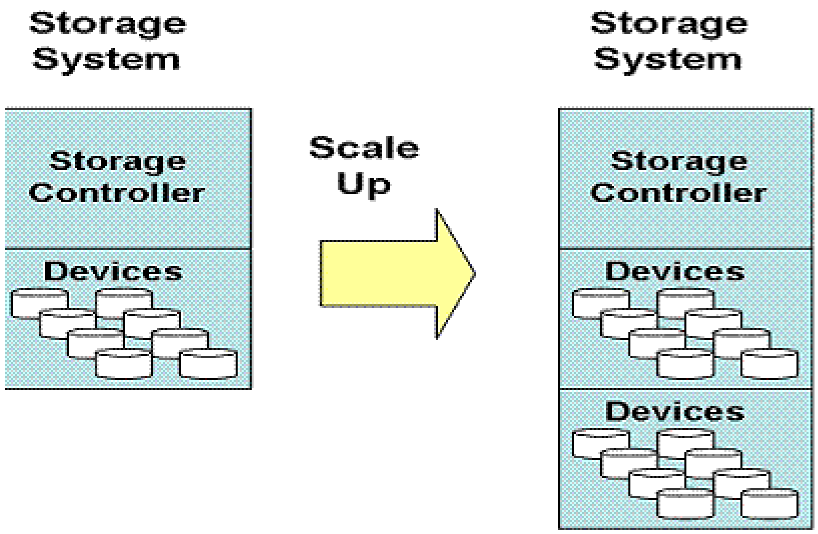

Figure 5: Scale-up Storage Example

Scale-up can solve a capacity problem without adding infrastructure elements such as network connectivity. However, adding storage devices does require additional space, power, and cooling. Scaling up does not add controller capabilities to handle additional host activities. The cost of scaling capacity benefits because only additional storage devices have been added using existing storage controllers.

At first, adding more storage devices typically improves performance because device throughput is often the limiting factor from a performance standpoint. However, as the load on the storage array increases — a situation often driven by virtualization — and more devices are added, the two storage controllers themselves become a bottleneck as each begins to require more and more CPU for RAID calculations and other command processing. Eventually, enough devices are added that the controllers become saturated and can do no more. Adding more and faster devices behind an overloaded controller pair simply places more overload on the controllers. One way to delay this point is to over-provision storage compute power (controller capabilities) and bandwidth upfront, but this can be costly. When the eventuality of that performance peak is reached, the next option is to either replace the compute and bandwidth capabilities of the storage controller or to purchase additional stand-alone storage systems. These options can be expensive and a management burden.

Capacity and performance can expand by using “scale-out” storage (Figure 6). Scale-out storage usually requires additional control and storage elements to add both capacity and performance. The key difference between scaling out and just putting more storage systems on the floor is that scale-out storage continues to be represented as a single system.

There are several methods for accomplishing scale-out, including clustered storage and grid storage systems. The scale-out storage shown in Figure 6 has added both the control function and capacity but maintained a single system representation for access. This scaling may require additional infrastructure such as storage switches to connect the storage to the controller and a connection between the nodes in the cluster or grid. Scaling-out adds power, cooling, and space requirements, and the cost includes the additional capacity, control elements and infrastructure. With the scale-out solution in this example, capacity increased and performance scaled with the additional control capabilities.

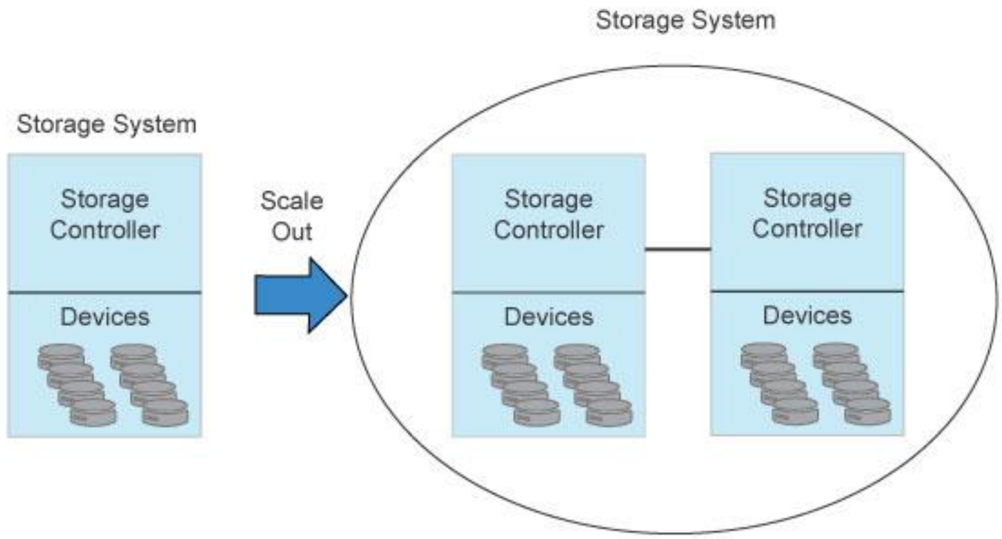

Figure 6: Scale-out Storage Example

While scaling *out* isn't necessarily a new concept, it has gained new life as storage vendors create new structures that leverage the power of modern servers to address the limitations of older scale-up architectures. Rather than using a custom controller platform, these new scale-out storage systems use a group of servers to form the storage cluster. Each server is loaded with storage devices and uses a network — usually Ethernet — to talk to the other servers in the storage array. The group of servers together forms a clustered storage array and provides logical units (LUNs) or file shares over a network just like a traditional array. Adding additional nodes of storage devices to a scale-out array actually adds another server to the cluster. At the same time, doing so adds more network ports, more CPU, and more RAM. As the capacity of a scale-out array increases so does its performance, adding nodes usually results in linear performance improvements. Adding nodes is generally non-disruptive, so a normal maintenance window can be used to add capacity and performance to the array, rather than a full system outage.

Local (DAS) Storage versus Shared Storage

Storage systems may be dedicated to a single server (DAS) or shared among multiple servers (SAN and NAS). Shared storage allows access to storage by multiple servers, either concurrently or sequentially. Typically, storage systems designed to be shared provide greater capacity, performance, scalability, and functional capabilities; all at higher cost than local storage systems.

Shared storage will always need some extra hardware to connect to the server, e.g., interface cards and network switches, resulting in a higher price. Shared storage introduces complexity that necessitates regular tuning and extra management expertise. In addition, when servers share storage, it is necessary to coordinate their actions so that shared data is correctly updated. It is also necessary to keep the caches of the servers in a state of coherence by ensuring that data in a server's cache corresponding to a storage location is invalidated when that same storage location is modified by another server.

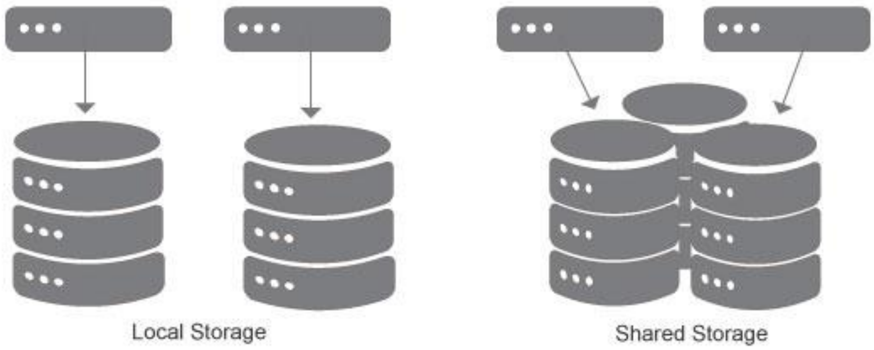

Figure 7: DAS (Local) and Shared Storage

Based on the environment and workload demands, local storage may be good enough or shared storage may be needed. By using local storage devices, the throughput as measured by Input/Output Operations Per second (IOPs) can be very high, and the cost of local storage remains low with easy deployment and simple management. However, the scalability and virtualization aspects of a networked shared storage (SAN/NAS) can provide higher levels of redundancy, availability, and disaster recovery. In addition, shared storage offers a variety of added software features that local storage may not have: the COMs[1] (snapshots, thin provisioning, compression, and deduplication), automated tiering[2], additional forms for data protections, remote replication of data, and more.

Cloud Storage

Cloud storage is a model of computer data storage in which the digital data is stored in logical pools, said to be on "the cloud". Cloud service deployment can be done in several ways.

Public clouds are the most common type of cloud computing and storage deployment. The cloud resources (like servers and storage) are owned and operated by a third-party cloud service provider and delivered over the internet. With a public cloud, all hardware, software, and other supporting infrastructure are owned and managed by the cloud provider. In a public cloud, hardware, storage, and network devices are shared with other organizations (cloud “tenants”) and services are accessed via the network and managed using a web browser. Public cloud deployments are frequently used to provide web-based email, online office applications, storage, and testing and development environments.

A private cloud consists of cloud computing and storage resources used exclusively by one business or organization. The private cloud can be physically located at an organization’s on-site datacenter or it can be hosted by a third-party service provider. However, in a private cloud, the services and infrastructure are always maintained on a private network and the hardware and software are dedicated solely to an organization.

---

[1]COMs (Capacity Optimization Methods) are storage software management features that allow more data to be stored on less actual capacity allowing reduction in physical resources (i.e., use less disks and less power).

[2]Automated storage tiering is a storage software management feature that dynamically moves information between different disk types and RAID levels to meet space, performance, and cost requirements.

A **hybrid cloud** is a type of cloud that combines on-premises infrastructure, a private cloud, with a public cloud. Hybrid clouds allow data and applications to move between the two environments.

For more information, see https://www.snia.org/education/what-is-cloud-storage.

Edge Data Storage

Edge computing is a distributed computing paradigm that brings computation and data storage closer to the location where it is needed, to improve response times and save bandwidth.

Gartner defines edge computing as "a part of a distributed computing topology in which information processing is located close to the edge – where things and people produce or consume that information."

At its basic level, edge computing brings computation and data storage closer to the devices where it's being gathered, rather than relying on a central location that can be thousands of miles away. This is done so that data, especially real-time data, does not suffer latency issues that can affect an application's performance. In addition, companies can save money by having the processing done locally, reducing the amount of data that needs to be processed in a centralized or cloud-based location.

Edge computing was developed due to the exponential growth of IoT devices, which connect to the internet for either receiving information from the cloud or delivering data back to the cloud. And many IoT devices generate enormous amounts of data during the course of their operations.

For more information, see https://www.snia.org/educational-library/storing-edge-architectural-approach-2018.

Data Security

Data security refers to the process of protecting data from unauthorized access and data modification throughout its lifecycle. Data security includes data encryption, hashing, tokenization, and encryption key management practices that protect data across all applications and platforms.

When properly implemented, robust data security strategies will protect an organization's information assets against cybercriminal activities, but they also guard against insider threats and human error, which remains among the leading causes of data breaches today. Data security involves deploying tools and technologies that enhance the organization's visibility into where its critical data resides and how it is used. Ideally, these tools should be able to apply protections like encryption, data masking, and redaction of sensitive files, and should automate reporting to streamline audits and adhering to regulatory requirements.

For more information, see https://www.snia.org/educational-library/storage-security-best-practices-2015.

Data Availability, Durability, Reliability and Resilience

Enterprise storage systems implement many complex mechanisms to ensure that data is not corrupted or lost and that data is available when needed. These mechanisms typically involve background activities that execute whether the storage system is idle or processing requests from a host.

System Availability is the percentage of a time period when the service will be able to respond to requests, i.e., system uptime

Data Availability is the degree to which data can be quickly accessed. The term is mostly associated with service levels that are set up either by the internal IT organization or that may be guaranteed by a third party datacenter or storage provider.

Data Durability / Integrity is the ability to keep the stored data consistent, intact without the influence of bit rot, drive failures, or any form of corruption.

System Reliability is closely related to availability; however, a system can be 'available' but not be working properly. Reliability is the probability that a system will work as specified.

Data reliability is the accuracy and completeness of data.

System Resiliency is the ability of a system to self-heal after damage, failure, overload, or attack. Note that this does not mean that it will necessarily be available continuously during the event, only that it will self recover.

3. WHY A STORAGE TAXONOMY IS NECESSARY

The basic function of a data storage system is to provide storage resources to the network for primary data storage, mirror data storage, or backup data storage. A wide variety of storage systems are available in the marketplace and serve different purposes with varied performance and features. The storage systems include JBOD disk arrays, RAID disk arrays, tape systems, optical storage systems, etc. The type of interfaces provided on these devices include Fibre Channel, SAS, and Ethernet.

Due to the wide spectrum of storage-oriented products available in the marketplace, a taxonomy structure was created in the SNIA Emerald™ Power Efficiency Measurement Specification[3]. When evaluating power efficiency metrics, a taxonomy enables a “like-to-like” product comparison that takes into account size, capacity, performance, high availability, etc. The storage taxonomy identifies classes of storage systems that require differing power efficiency test criteria.

Introduction

The SNIA Emerald Storage Taxonomy defines a market taxonomy that classifies storage products or subsystems in terms of operational profile and supported features. While this taxonomy is broad and defines a framework for products that range from consumer solutions to enterprise installations, it is not intended to address all storage devices.

The taxonomy is structured as a 3-level hierarchy of Set, Category, and Classification. A Set is a broad grouping of products. The Sets are: Disk Set, RVML (removable & virtual media library) Set, and NVSS (non-volatile solid state) Set. A Category is a grouping of products within a Set. The Categories are: Online, Near-Online, Removable Media Library, Virtual Media Library, Disk Access, and Memory Access. A Classification is a level of product sophistication, size, and complexity that corresponds to market delineations. Classifications are identified as a numeric suffix appended to a Category. Some Classifications are used within multiple Categories.

Taxonomy Sets

Taxonomy sets define broad groupings of storage products that share similar system characteristics. See Figure 8. Products in different sets are generally not comparable in performance or power efficiency characteristics.

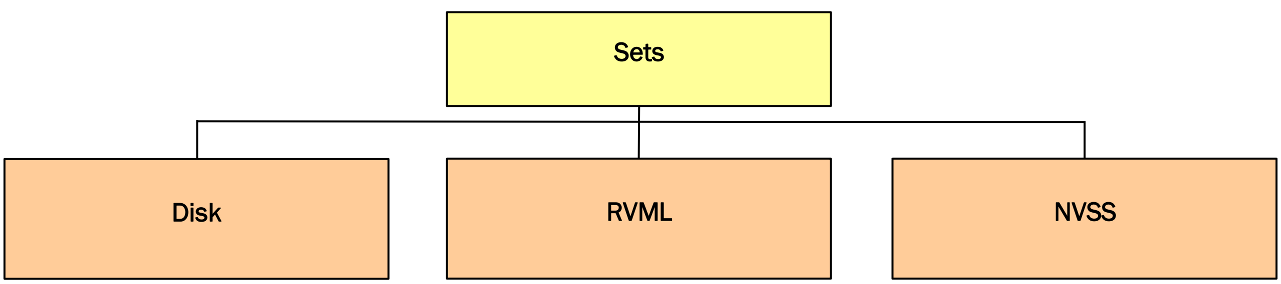

Figure 8: Taxonomy Overview

[3]SNIA Emerald™ Power Efficiency Measurement Specification, https://www.sniaemerald.com/download.

Taxonomy Categories

General

Taxonomy categories define broad market segments within a Set that can be used to group products that share common functionality or performance requirements, and within which meaningful product comparisons can be undertaken. This document defines six broad taxonomy categories (summarized in Table 2):

- Disk Set Online;
- Disk Set Near-Online;
- RVML Set Removable Media Library;
- RVML Set Virtual Media Library;
- NVSS Set Disk Access;
- NVSS Set Memory Access.

Within a taxonomy category, a specific model or release of a product will support different feature sets, whether focused on capacity, reliability, performance, functionality, or another differentiator. Feature and functionality differences within a category are addressed with attributes. Each taxonomy category defines a set of attributes that are common to all products within the category.

Category Attributes

Where a taxonomy category requires a specific, fixed setting or range for a given attribute, that setting is summarized in Table 1 to assist a test sponsor in initial category selection.

Table 1 – Common Category Attributes

Attribute	Set					
	Disk		RVML		NVSS	
	Category					
	Online	Near-Online	Removable Media Library	Virtual Media Library	Disk Access	Memory Access
Access Pattern	Random/ Sequential	Random/ Sequential	Sequential	Sequential	Random/ Sequential	Random
MaxTTFD	≤ 80 ms	> 80 ms	≤ 5 min	≤ 80 ms	≤ 80 ms	≤ 80 ms
Media Type	Magnetic disk	Magnetic disk	Magnetic tape, optical disk	Magnetic disk, Solid State Storage	Solid State Storage + optional magnetic disk[a]	Solid State Storage
Access Paradigm	Block, File, Object	Block, File, Object	Block	Block	Block, File, Object	Memory

[a] Allows a purely Solid State Storage system or a hybrid Solid State Storage and magnetic disk system.

The Attribute column of Table 1 is defined as follows:

- Access Pattern is one or more of Random and Sequential.

  A storage product having the Random Access Pattern attribute provides roughly equal access time to any stored data.

  A storage product having the Sequential Access Pattern attribute may provide faster access to data at the address following the last accessed data than to other data.

- MaxTTFD (maximum time to first data) is the time required to start receiving data from a storage system to satisfy a read request for arbitrary data

- Access Paradigm is one or more of Block, File, Object, and Memory.

  A storage product having the Block Access Paradigm provides fixed-block or CKD access to data, e.g., via SCSI block access for disks/tapes.

  A storage product having the File Access Paradigm attribute provides file access to data, e.g., via the NFS protocol.

  A storage product having the Object Access Paradigm attribute provides object access to data, e.g., via the SCSI OSD protocol (INCITS 400-2004 and INCITS 458-2011).

  A storage product having the Memory Paradigm attribute provides access to data via CPU memory access instructions, e.g., load and store.

Taxonomy Classifications

Classifications define combinations of settings or values for the attributes within a Category.

Taxonomy Classifications discriminate between different systems. The goal of the Classifications is to differentiate between systems of differing size, complexity, and target market. The Classifications are in general the following:

- Consumer/Component (<classification> 1),
- JBOD (<classification> 1.5),
- Low-end (<classification> 2),
- Mid-range (<classification> 3) and (<classification> 4),
- High-end (<classification> 5), and
- Mainframe (<classification> 6).

Consumer (<classification> 1) systems are any of a wide array of manufactured goods which are purchased primarily for personal, family, household, or small-business purposes.

Components (<classification> 1) are components (e.g., a stand-alone tape drive or an individual hard disk drive) that are a part of a data-center storage product.

JBOD (<classification> 1.5) is a “simple” JBOD, an architecture that does not have storage protection capabilities built in. The storage devices in a JBOD can function as individual volumes or can be connected to form a single logical volume by a simple, integrated controller providing no redundancy. A JBOD provides no redundancy or resilience, so failure of a single storage device amounts to failure of a whole logical volume. A JBOD is a single shelf.

Low-end systems (<classification> 2) typically are a single shelf with one or two integrated controllers and limited expandability.

Mid-range (<classification> 3 and <classification> 4) systems are sub-divided into small and large mid-range systems. Small Mid-range systems (<classification> 3) generally have limited expandability and robustness compared to Large Mid-range systems (<classification> 4).

High-end (<classification> 5) and Mainframe (<classification> 6) systems are very large systems that can have very high capacity and offer a high level of robustness. Mainframe is differentiated from High End in that Count Key Data (A disk data organization model in which the disk is assumed to consist of a fixed number of tracks, each having a maximum data capacity) support is required.

Table 2 provides an overview of the taxonomy classifications.

Table 2 – Taxonomy Classifications Overview

Level	Set					
	Disk		RVML		NVSS	
	Category					
	Online	Near-Online	Removable Media Library	Virtual Media Library	Disk Access	Memory Access
	Classification					
Consumer/ Component[a]	Online 1[b]	Near-Online 1[b]	Removable 1	Virtual 1	Online 1[b]	Online 1[b]
JBOD	Online 1.5	Near-Online 1.5[c]	Removable 1.5[c]	Virtual 1.5[c]	Online 1.5	Online 1.5[b]
Low-end	Online 2	Near-Online 2	Removable 2	Virtual 2	Online 2	Online 2[b]
Mid-range	Online 3	Near-Online 3	Removable 3	Virtual 3	Online 3	Online 3[b]
	Online 4	Near-Online 4[c]	Removable 4[c]	Virtual 4[c]	Online 4	Online 4[b]
High-end	Online 5	Near-Online 5	Removable 5	Virtual 5	Online 5	Online 5[b]
Mainframe	Online 6	Near-Online 6	Removable 6	Virtual 6	Online 6	Online 6[b]

a Entries in this level of the taxonomy include both consumer products and data-center components, e.g., stand-alone tape drives.

b No test procedure for this Classification is provided by the Emerald Specification.

c Classification is not defined; no test procedure is provided by this document.

Disk Set Online Category

The Disk Set consists of storage products based on rotating media devices, typically magnetic disks.

The Disk Set Online Category defines the features and functionalities for an online, random-access, rotating media-based storage product. Products in this profile may provide any combination of block, file, or object interfaces. Table 3 defines the requirements for the taxonomy classifications (Disk Set Online Online 1, etc.) defined in this category.

Table 3 – Disk Set Online Classifications

Attribute	Category: Disk Set Online Classification: Online						
	1[a]	1.5	2	3	4	5	6
Multi-host Shareability	Not Specified	Ability to share with 1 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts
Consumer/ Component	Yes	No	No	No	No	No	No
Storage Controller	Optional	Optional	Required	Required	Required	Required	Required
Storage Protection	Optional	No	Required	Required	Required	Required	Required
No SPOF	Optional	Optional	Optional	Optional	Required	Required	Required
Non-Disruptive Serviceability	Optional	Optional	Optional	Optional	Optional	Required	Required
FBA/CKD Support	Optional	No	Optional	Optional	Optional	Optional	Required
System Capacity (number of drives) [b]	≥ 1	≥ 4	≥ 4	≥ 12	> 100	> 400	> 400

[a] No test procedure for this Classification is provided by the Emerald Specification.

[b] There is no upper limit on the number of drives in any Classification.

The Attribute column of Table 3 is defined as follows:

- Multi-host Shareability identifies the number of hosts that can share the storage product.
- A storage product having the Storage Protection attribute assures that all completed IO operations will be preserved in the event of power loss or storage device failure. This assurance can be provided by a combination of hardware and/or software, e.g., RAID, NVRAM, disk sparing, background disk scrubbing, and/or background media scan.

- A storage product having the No SPOF attribute does not have any component or path whose failure makes the system inoperable or data inaccessible.
- A storage product having the Non-Disruptive Serviceability attribute supports continued availability of data during all service operations, including FRU replacement, code patches, software/firmware upgrades, configuration changes, data migrations, and system expansion.
- A storage product having the FBA/CKD Support attribute provides count-key-data as defined by the IBM Corporation.

Near-Online Category

The Disk Set consists of storage products based on rotating media devices, typically magnetic disks.

The Disk Set Near-Online category defines the features and functionalities for a near-online, random-access rotating media-based storage product. Products in this profile employ MAID or FCAS architectures as well as any combination of block, file, or object interfaces. Table 4 defines the requirements for this taxonomy classifications (Disk Set Near-Online Near-Online 1, etc.) defined in this category.

Table 4 – Disk Set Near-Online Classifications

Attribute	Category: Disk Set Near-Online Classification: Near-Online						
	1[a]	1.5[b]	2	3	4[b]	5	6
Multi-host Shareability	Not Specified		Ability to share with 2 or more hosts	Ability to share with 2 or more hosts		Ability to share with 2 or more hosts	Ability to share with 2 or more hosts
Consumer/ Component	Yes		No	No		No	No
Storage Controller	Optional		Optional	Required		Required	Required
Storage Protection	Optional		Optional	Required		Required	Required
No SPOF	Optional		Optional	Optional		Required	Required
Non-Disruptive Serviceability	Optional		Optional	Optional		Required	Required
FBA/CKD Support	Optional		Optional	Optional		Optional	Required
System Capacity (number of drives)[c]	≥ 1		≥ 4	≥ 12		> 100	> 1 000

a	No test procedure for this Classification is provided by the Emerald Specification.
b	Classification is not defined; no test procedure is provided the Emerald Specification.
c	There is no upper limit on the number of drives in any Classification.

RVML Set Removable Media Library Category

This category defines the features and functionalities for storage products that rely on automated or manual media loaders (e.g., tape or optical libraries). Table 5 defines the requirements for the taxonomy classifications defined in this category.

Table 5 – RVML Set Removable Media Library Classifications

Attribute	Category: RVML Set Removable Library Classification: Removable						
	1	1.5[a]	2	3	4[a]	5	6
Robotics	Prohibited		Required	Required		Required	Required
No SPOF	Optional		Optional	Optional		Optional	Required
Non-disruptive Serviceability	Optional		Optional	Optional		Optional	Required
System Capacity (number of drives)	Not Specified		≤ 4	≥ 5		≥ 25	≥ 25
[a] Classification is not defined; no test procedure is provided by the Emerald Specification.							

RVML Set Virtual Media Library Category

This operational profile defines the features and functionalities for sequential-access storage products that rely on non-removable storage media to provide a Virtual Media Library. Table 6 defines the requirements for the taxonomy classifications (RVML Set Virtual Media Library Virtual 1, etc.) defined in this category.

Table 6 – RVML Set Virtual Media Library Classifications

Attribute	Category: RVML Set Virtual Media Library Classification: Virtual						
	1	1.5[a]	2	3	4[a]	5	6
Storage Protection	Optional		Optional	Required		Required	Required
No SPOF	Optional		Optional	Optional		Optional	Required

Attribute	Category: RVML Set Virtual Media Library Classification: Virtual						
	1	1.5[a]	2	3	4[a]	5	6
Non-Disruptive Serviceability	Optional		Optional	Optional		Optional	Required
System Capacity (number of drives)	≤ 12		> 12Error! Reference source not found.	> 48[b]		> 96Error! Reference source not found.	> 96Error! Reference source not found.

[a] Classification is not defined; no test procedure is provided by the Emerald Specification.

[b] There is no upper limit on the number of drives.

NVSS Set Disk Access Category

The NVSS Set consists of storage products based on non-volatile Solid State Storage.

The NVSS Set Disk Access Category defines the features and functionalities for an online, random-access, Solid State Storage-based disk access storage product. A storage product offers disk access if it provides data access using a storage paradigm, i.e., open, close, read, and write.

Products in this profile may provide any combination of block, file, or object interfaces. Table 7 defines the requirements for the taxonomy classifications (NVSS Set Disk Access Online 1, etc.) defined in this category.

Table 7 – NVSS Set Disk Access Classifications

Attribute	Category: NVSS Set Disk Access Classification: Online						
	1[a]	1.5	2	3	4	5	6
Multi-host Shareability	Not Specified	Ability to share with 1 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts	Ability to share with 2 or more hosts
Consumer/ Component	Yes	No	No	No	No	No	No
Storage Controller	Optional	Optional	Required	Required	Required	Required	Required
Storage Protection	Optional	Not integrated	Required	Required	Required	Required	Required
No SPOF	Optional	Optional	Optional	Optional	Required	Required	Required
Non-Disruptive Serviceability	Optional	Optional	Optional	Optional	Optional	Required	Required

FBA/CKD Support	Optional	Not integrated	Optional	Optional	Optional	Optional	Required
System Capacity	consumer	JBOD	very small	small	medium	large	mainframe
[a] No test procedure for this Classification is provided by the Emerald Specification.							

NVSS Set Memory Access Category

The NVSS Set consists of storage products based on non-volatile Solid State Storage.

The NVSS Set Memory Access Category defines the features and functionalities for an online, random-access, Solid State Storage-based memory access storage product. A storage product offers memory access if it provides host access to storage using memory primitives, e.g., load and store.

Products in this profile provide a memory interface. Table 8 identifies the potential requirements for the taxonomy classifications in this category.

Table 8 – NVSS Set Memory Access Classifications

Attribute	Category: NVSS Set Memory Access Classification: Online						
	1[a]	1.5[a]	2[a]	3[a]	4[a]	5[a]	6[a]
Consumer/ Component	Yes	No	No	No	No	No	No
Storage Protection	Optional	TBD	Optional	Required	Required	Required	Required
No SPOF	Optional	TBD	Optional	Optional	Required	Required	Required
Non-Disruptive Serviceability	Optional	TBD	Optional	Optional	Optional	Required	Required
System Capacity	consumer	JBOD	very small	small	medium	large	mainframe
[a] No test procedure for this Classification is provided by the Emerald Specification.							

In addition, the Storage Industry is rapidly evolving new types of storage architectures and products and bringing them to market. Product areas being explored include: object store; big enterprise virtualized storage servers; and converged, hyper-converged, composable, and software-defined storage. The definition of boundaries between storage and server products is becoming more challenging to delineate; and as software-defined implementations gain prominence, addressing these will require a collaborative categorization effort across the full range of industry IT partners.

4. PHYSICAL ATTRIBUTES OF STORAGE

Storage systems consist of several types of media, controllers, and enclosures. Storage devices include hard disk drive (HDD), solid state drive (SSD) including PCIe and NVMe bus variants and non-drive format solid state media, tape and optical media. Optical media and tape are primarily used for back-up and archiving. Storage controllers or processors manage the physical storage devices, implement hardware RAID and provide interfaces to the front-end computer HBA (host bus adapter) and back-end controlled devices. The drives, controllers, I/O modules, mid-plane, power supplies and cooling fans are integrated into physical enclosures. These enclosures can either be single, integrated enclosures or they can be separate controller and drive enclosures. To meet the requirement of no single-point-of-failure (SPOF) these enclosures will have redundant hardware modules. A typical storage enclosure having integrated disks is shown in Figure 9.

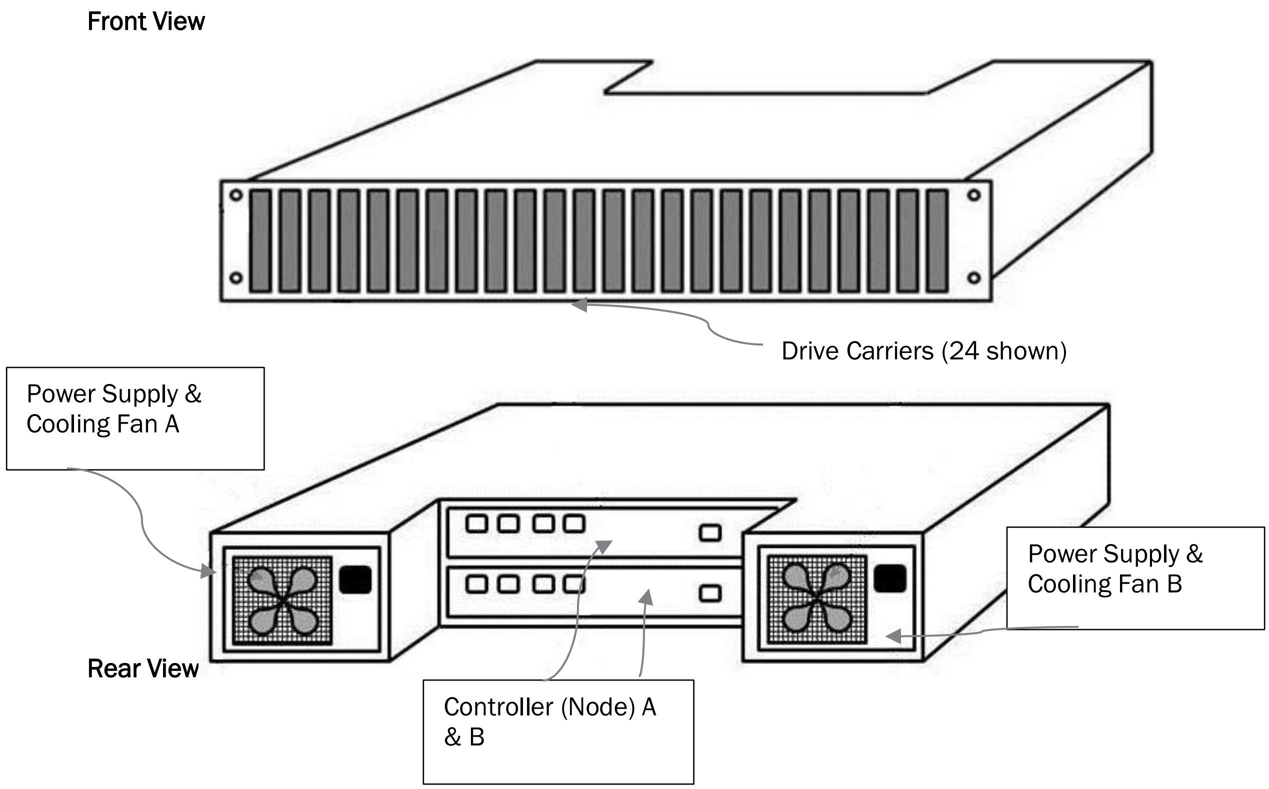

Figure 9: Typical Storage Enclosure: Entry Level, Expand Out

The energy consumption of storage systems is largely determined by the storage media characteristics: the type (SSD or HDD), the storage capacity, the power consumption of the individual devices, and the device count. Additional factors are choice of RAID levels, RAS (Reliability, Availability and Serviceability) features, the Capacity Optimization Methods (COMs) in use, and the performance capability of the storage controller. In general, the higher the capability of a storage system (high-end as opposed to low-end), the higher the power needed to support the additional features.

Among hard drives used in Enterprise storage products, 2.5" drives have an average power demand of around 6 W (7.2K RPM) per drive with slightly higher power demands of 7.5 W (10K RPM) and 8.3W (15K RPM). 3.5" drives have an average power demand of approximately 10 W for a 7.2K RPM drive.

The average power demand of SATA and SAS solid state drives is 7 to 15 W. NVMe solid state drives have considerably higher power requirements. The power will vary as a function of capacity and can exceed 25 W.

Power Calculators are available from many storage vendors; these can be utilized to determine the power contributions of each storage enclosure and to calculate totals for different configurations.

Based on performance capability, each controller module typically consumes 50 W to 1000 W and drive enclosure I/O modules consume 20 W to 50 W. A standard 2U drive enclosure that houses 24 2.5" spinning drives can consume 200 W to 400 W. In addition, there are dense enclosures that can house more than 100 3.5" drives per 4U or more than 120 2.5" drives per 3U; the power for one of these enclosures can be more than 1500 W. Large storage configurations, those having more than 3-4 drive enclosures, are dominated by drive power. Drive power consumption in large storage configurations typically becomes 60% or more of the total power consumption.

5. THE NEAR-TERM EVOLUTION OF STORAGE

Storage providers are continuing to research, develop and offer hardware and software based innovations to improve the quantity of data that can be stored on a given product footprint and improve the performance per watt of the storage product, to decrease the overall data center power consumption. Improvements are being made at both the storage device (media) level and for storage systems -particularly as being implemented in the data center.

Storage Media

Storage media will continue to reflect ongoing technology developments in both hard disk drives (HDD) and solid state drives (SSD). Both types of media will continue to be generally deployed in data centers for at least the next 5-10 years. While some data center storage product vendors may only use one type of media, many will offer customers combinations of media types to better match operational, capacity and business requirements. This section explores the characteristics of both media types, focusing on several key differences, as well as some similarities. It will be the differences that drive the need to continue with HDD as well as SSD technology for storage.

Comparison of Use Cases

SSD adoption will continue to grow in a number of data center use cases. Many applications in traditional data centers are highly performance sensitive. Others have certain performance-sensitive functions. Users facing these performance constraints will continue their adoption of SSDs in place of HDDs. In these situations, the use of SSD will generally be confined to a tier of storage in a large system, or to the use of dedicated storage appliances. In these classes of applications, the investment in SSDs makes good business sense from a productivity/spend perspective.

Some cloud data centers are not as performance sensitive because data is delivered over the internet, which imposes a delay which is the limitation on overall performance.

Other data centers are faced with a different primary problem – the need to maintain very large quantities of data online. Under these circumstances, capacity at an adequate performance level is more important than pure performance; these functions will continue to deploy high-capacity HDD.

Another use case for SME (small to medium enterprises) and enterprise data centers that will likely remain on HDDs is that of backing up data for reliability purposes. Most businesses use some form of digital back up, ranging from replication to sophisticated backup appliances, to ensure data is not destroyed by accident or due to equipment failure. This is another situation where large capacities are needed, and since backup software generally manages to store only those items that have changed, performance is not as significant an issue, while cost is very much an issue, since this is overhead, which must be kept low.

Media Development Directions

As noted above, there are ample use cases for both SSD and HDD deployment in the future. As a result, there is ongoing research and development in both areas. Later sections of this document look at likely developments in both areas.

Disk drive suppliers are driven to continue to maintain or decrease the power consumed by each drive in order to remain within the design parameters defined by their primary customers, system vendors who integrate drives into servers and/or storage products. Within this constraint, they are looking at:

- Increasing capacity. Accomplishing this within the power budget limitations generally means that platters cannot be added to a drive. Work in this space is primarily focused on new or modified recording technologies to improve the density of recording at acceptable levels of reliability.

- Increasing the energy efficiency of drives by exploring changes in the mechanical components of a drive. Sealed helium-filled drives are one example of this ongoing direction.

HDD Storage Media

HDD (Hard Disk Drives) for enterprise systems come in two types: High Performance and Capacity. High Performance disk drives are disk drives with either a SAS or Fibre Channel high-performance data interface. They have a high rotational speed of either 10K RPM or 15K RPM (for which there is a significantly declining market). They have a relatively low density (data storage capacity). Capacity disk drives are disk drives with a higher data capacity, than High Performance disk drives. They have a lower performance data interface, usually SATA, and have a lower rotational speed, usually 7.2K RPM or lower.

Many companies are developing technologies to overcome the physical limitations of traditional disk drive technology to increase storage capacity.

Helium-filled disk drives

Helium-filled disk drives use helium as the environment in which the platters spin and the heads move. The use of this low density and low viscosity gas allows platters to be more closely spaced, increasing the capacity of drives of a given form factor. It also reduces the power necessary to spin the platters which allows a smaller motor to be used and reduces the power necessary to operate the drive.

Heat-assisted magnetic recording drives

Heat-assisted magnetic recording uses light or microwave energy produced within a read-write head, to heat the region of a disk about to be written to briefly change the magnetic properties of that region so that it is more readily written. This technology allows improved magnetic materials to be used that support greater data density.

Shingled magnetic recording drives

Shingled magnetic recording (SMR) is a recording technology for magnetic disks that partially overlays adjacent tracks on a disk platter to increase the track density and thereby increase the capacity of disk platters and disk drives. This recording technique overlaps tracks, much as shingles are laid down on a roof, hence the name. The writing process is made more complex because data on adjacent tracks cannot be written independently. Various management techniques reduce the additional time that this imposes on writes data to an SMR disk.

Dual-actuator drives

Some drive manufacturers have introduced disk drives with two actuators within a single drive to improve performance. This is especially beneficial in maintaining the performance per GB in high capacity drives.

SSD Storage Media

Solid state drives are storage devices based on solid state memory technology.

SSD technologies are available in a variety of form factors, interfaces, and protocols:

- plug-in replacements for HDDs, typically supporting SAS or SATA as the interface,
- devices that have the form factor of an HDD but support NVMe as the interface, and
- devices that have an SSD optimized form factor, e.g., U2 and M2, and typically support NVMe as the interface.

NVMe (Non-Volatile Memory Express) is a standard interface that defines a command set and feature set for PCIe-based SSDs that offer increased performance. Some form factors, e.g., M2 and U2, are being deployed for NVMe SSD. Some of these form factors are smaller and all require a different connector than traditional SATA or SAS devices.

SSDs are replacing 10K and 15K RPM magnetic disk drives as they offer superior performance at a comparable cost.

Memory Access SSD technologies are currently in development.

SSDs have emerged as an online storage medium (as opposed to a convenient data transport or non-volatile cache) as a result of improved silicon technologies. SSDs provide a number of advantages over HDDs as a result of the following characteristics:

- Absolute performance: SSD storage outperforms all HDDs, offering both significantly greater small I/O operations/second and higher streaming data rates.
- Power draw: Small to moderate capacity SSDs supporting the SAS or SATA interface require substantially less power than high-performance HDDs, typically on the order of the power draw of a 7.2K RPM HDD. This means that replacing a given capacity of high-performance HDD storage with an equivalent capacity of SSD storage results in lower energy use.
- Vibration: SSDs have no moving parts and therefore do not contribute mechanical vibration to the storage system.
- Form factor: Form factors, not constrained by the round platters of magnetic disk drives, enable smaller form-factors for both enterprise and client systems.
- Weight: Small and modest capacity SSDs weigh less than HDDs of equivalent capacity. While this advantage is often not maintained at very high capacities, it has proven to be highly important in providing reliable, lightweight storage for portable devices from personal electronics to laptops.

The semi-conductor memory cells of an SSD have a lifetime limit on the number of write operations each cell can handle. As a result, the SSD typically includes a “wear-levelling” algorithm to avoid premature aging of the device.

Solid State technology advances have largely been focusing on:

- Improving performance.
- Increasing capacity per package size (volume). This increases total system capacity without using as many “disk equivalents” or system slots, supporting larger storage systems in a smaller footprint.
- Decreasing cost/GB.

Summary

The combination of characteristics discussed in this section lead to the conclusion that both SSD and HDD will coexist in many data center environments in complementary roles for many years to come

Emerging Storage Trends

Applications Are Replacing Appliances

Over the last few years, the data center industry has been experiencing a new “equipment” trend. Sophisticated application software products have been developed to emulate traditional data center appliances such as storage systems and network switches. The literature has referred to these product categories as Software-Defined-Networking (SDN), Software-Defined Storage (SDS), converged, hyper-converged, storage servers, and other variations of names. For simplicity, we will refer to them as SDN and SDS. Both of these categories of applications are experiencing significant market share growth among early adopters and both are increasing their presence in mainstream data centers.

A deeper look at one of these categories, SDS, shows that the operational software products are designed to be agnostic regarding the brand of storage and server on which they are installed. Instead, they rely on the same sorts of guidelines as any other application when it comes to a suitable run-time environment, listing requirements such as instruction set, operating system, memory available for use, and storage space needed to support the solution. The customer is free to install the application, along with the corresponding drives that will be used by the application, on any server, including ones that are already running other applications. In fact, the primary point of using SDS or SDN products is to avoid the purchase of dedicated appliances. The SDS and/or SDN applications normally run in a virtualized server environment sharing the server with the same applications that will be storing data using the SDS.

The SDS or SDN energy efficiency and performance are determined by two independent variables, the behavior of the software and the behavior of the hardware. While this is superficially true of storage or network appliances, there are fundamental differences. Using storage as an example, a particular storage appliance is actually a specific controller design, dedicated I/O infrastructure and a specific storage application. When this is tested, it can be expected to provide the same behavior repeatedly. In an SDS environment, energy efficiency and performance are functions of the storage application, the specific server configuration (including the memory, I/O controllers, etc. needed for all applications) and the other applications in use on the server.

A second consequence is that the servers running one of these SDS or SDN applications will require significantly different configurations than dedicated compute servers in the data center. At a minimum, they will need appropriate quantities of high-performance I/O devices installed in the system to function as the substitute for the specific appliance. Finally, SDS applications will require additional storage capacity and additional components to support that additional storage capacity.

The Adoption of Hybrid Cloud

As previously described in the section Cloud Storage, cloud storage can be public, private, or hybrid deployments. Not everyone is ready to move all of their data to a public cloud; businesses are now increasingly adopting the hybrid cloud (multi-cloud) strategy. This is a data storage strategy that helps organizations gain cost-effectiveness and increase data mobility between public cloud and private cloud without compromising data integrity. With a hybrid cloud strategy, organizations have the flexibility to collect, segregate and store data whether it is on- or off-premises. This helps them accelerate time to market and easily deploy and manage their environments. Hybrid cloud storage solutions typically connect users to a variety of storage options, including file storage and sharing, object storage, block storage and backup systems. Most enterprises can use the public cloud for some aspects of their work that are not security and access time sensitive. However, these businesses still need a private cloud service to store their more sensitive data or data requiring faster access time, hence the hybrid cloud solution.

Storage as a service (StaaS)

StaaS is a data storage business model where a provider rents storage resources to a customer through a subscription. StaaS reduces cost through operating expenditure (OpEx) agility—you only pay for the storage you use, when you need it. In the StaaS model, the storage provider handles most of the complex aspects of long-term bulk data storage – hardware costs, security, and data integrity. The service can be delivered on premises from infrastructure that is dedicated to a single customer, or it can be delivered from the public cloud as a shared service that is purchased by subscription and is billed according to one or more usage metrics. Instead of storing data on-premises, organizations that use StaaS will typically utilize a public cloud for storage needs. Public cloud storage may also use various storage methods for StaaS. StaaS customers access individual storage services through standard system interface protocols or application program interfaces (APIs). These storage methods include backup and restore, disaster recovery, block storage, file storage, object storage and bulk data transfer.

6. CAPACITY OPTIMIZATION METHODS (COMs)

Capacity optimization is a general term for a set of software techniques used in storage products which can enable a higher data storage capacity on a given storage device footprint. Reducing the amount of stored data can reduce the required drive capacity and result in less energy usage for a given operation. Each of these techniques is known as a Capacity Optimization Method (COM). A description of common COMs follows:

- **Advanced Data Protection:** (includes parity RAID and erasure encoding)
  These are methods of data protection that do not use full copies. When replacing mirrored or full copies with advanced data protection, the estimated space savings can vary from 25 - 50%.
- **Thin Provisioning:**
  A technology that allocates the physical capacity as applications write data, rather than allocating all the physical capacity at the time of provisioning.
- **Delta Snapshots:**
  A type of point-in-time copy that reconstructs different versions of a file without multiple full copies. Large space savings are possible when the change is small.
- **Compression:**
  The process of encoding data to reduce its size. A compression ratio of 2:1 is routinely achieved.
- **Data Deduplication:**
  The replacement of multiple copies of data, at variable levels of granularity, with references to a shared copy. This can result in large space savings.

In general, COMs are largely, but not completely, independent of each other. They can provide benefits in any combination, but their combined effect does not equal the sum of their individual benefits and significant inefficiencies are typically introduced if more than two COMs are deployed on a single storage product.

ENERGY STAR Data Center Storage V2.1 requires that eligible Online product categories have Advanced Data Protection enabled during the SNIA Emerald test. For Disk Set and NVSS Disk Set Access storage product categories, Advanced Data Protection is required. In addition, the number of COMs (identified above other than Advanced Data Protection) required to be made available as a selectable feature are as follows:

- **Online 2**: 1 COM required to be made available.
- **Online 3**: 2 COMs required to be made available.
- **Online 4**: 3 COMs required to be made available.

Test verification consists of the application of heuristics which verify the existence and activation of a particular COM. It is also intended that testers disable all optional COMs that are capable of being disabled during all active and idle measurement tests.

Quantifying COM Benefits

In contrast to using a simple proof-of existence test for COMs, it may become more valuable to identify a quantifiable method for determining the impact on idle and active metrics that enabled COMs can have. There has been some estimation of active COM benefits, but more test characterization is needed to establish valid methods that accurately account for their efficiency during active operation. Shown below are several ways of accounting for test-based active COM comparison. Table 9 is based on *physical capacity reduction* (i.e., keep logical capacity constant, but remove physical drives to reduce energy consumption) and Table 10 is based on *usable capacity optimization (*i.e. quantify the effective virtual increase of raw storage). The “Delta (GB/W)” and “Delta (IO/W)” lines indicate the net change (plus or minus) of the ready idle metric and active (random workload in this example) metric. Note that for all cases, there is net improvement in the value of the idle metric, particularly in the case of the usable capacity method. There is more variance in the active metric results. For the physical capacity reduction (since actual drive count is being reduced), the active performance and metric are reduced (2 cases) or improved (1 case). The active performance metric remains the same or improves for all usable capacity cases.

Table 9 – Active COM Examples (70/30 R/W Random Workload), Based on Physical Capacity Reduction[4]

COM Type	Thin Provisioning	TP + Data Dedup	Data Compression
Baseline Config	Raid 5 (3+1 Data + Parity)	Full Provisioned No Dedup	Uncompressed
Baseline Capacity (GB)	54,000 (160 – 450 GB disks)	54,000 (160 – 450 GB disks)	36,000 (160 – 450 GB disks)
Baseline Power (W)	2360(I) 3134(B)	2360(I) 3134(B)	2360(I) 3134(B)
Baseline (GB/W)	22.88	22.88	15.25
Baseline Perf. (IOPs)	25,250	25,250	36,010
Baseline (IO/W)	8.06	8.06	11.49
Optimized Config	Thin Provisioning (50% written)	TP+Deduped Patterns (2:1)	Compressed (2:1)
Optimized Capacity (GB)	54,000 (50% written)	54,000 (50% written-dedup 2:1)	36,000 (80 – 450 GB disks)
Opt. Power (W)	1464(I) 2030(B) (80 disks)	1016(I) 1478(B) (40 disks)	1464(I) 2030(B)
Optimized (GB/W)	36.88	53.15	24.59
Delta (GB/W)	+61%	+132%	+61%
Opt. Perf (IOPs)	13,560	8,580 (2:1 Dedup Ratio)	38,280
Optimized (IO/W)	6.68	5.81	18.86
Delta (IO/W)	(17)%	(28%)	+64%

[4] Source: "Software Defined Storage Energy Efficiency Features," The Green Grid Forum 2016

Table 10 – Active COM Examples (70/30 R/W, Random Workload) Based on Usable Capacity Optimizations[5]

COM Type	Thin Provisioning	TP + Data Dedup	Data Compression
Baseline Config	Raid 5 (3+1 Data + Parity)	Full Provisioned No Dedup	Uncompressed
Baseline Capacity (GB)	54,000 (160 – 450 GB disks)	54,000 (160 – 450 GB disks)	36,000 (160 – 450 GB disks)
Baseline Power (W)	2360(I) 3134(B)	2360(I) 3134(B)	2360(I) 3134(B)
Baseline (GB/W)	22.88	22.88	15.25
Baseline Perf. (IOPs)	25,250	25,250	36,010
Baseline (IO/W)	8.06	8.06	11.49
Optimized Config	Thin Provisioning (50% written)	TP+Deduped Patterns (2:1)	Compressed (2:1)
Optimized Capacity (GB)	108,000	216,000 (50% written, 2:1)	72,000 (2:1 Comp Ratio)
Opt. Power (W)	2360(I) 3134(B)	2360(I) 3134(B)	2360(I) 3134(B)
Optimized (GB/W)	45.76	91.52	30.50
Delta (GB/W)	+100%	+300%	+100%
Opt. Perf (IOPs)	25,250	32,825 (2:1 Dedup Ratio)	72,020 (2:1 Comp Ratio)
Optimized (IO/W)	8.06	10.47	22.98
Delta (IO/W)	0%	+30%	+100%

When using the physical capacity reduction method, the power savings due to reduced drive count can be readily identified. However, due to fewer physical drives being present, the performance (e.g., IOPs) and the active metric may degrade. In addition, an actual storage system will undergo a physical drive reduction only during initial purchase or set-up, and not while in operation. When using the usable (or virtual) capacity optimization method, the actual raw capacity is effectively increased in size. However, the SNIA Emerald™ Specification uses the raw capacity as the numerator when calculating the ready idle metric, not the usable capacity improvement – thus not giving credit to this COM benefit. Note that both ready idle and active metrics are impacted by the choice of method to assess the benefits of a COM. The ready idle metric will greatly improve with the usable or virtual capacity method. However, where credit is given for the reduction in physical drives, the active performance metrics may suffer. Further investigation is necessary to derive a preferred and quantifiable method for comparing COM benefits.

The feasibility and benefits of the simultaneous operation of multiple COMs is implementation and configuration dependent. Depending on the array resources that enable the operation of COMs functions, there may be minimal interaction between them. As an example, consider a product that enables a deduplication feature via the controllers, a compression feature that is provided within the SSDs that are on the "backend" of the IO stack and a thin provisioning feature. Each of these may contribute to the overall usable (logical) capacity presented by the storage subsystem. In an ideal case, if both the compression ratio and the thin overprovisioning ratio are 2:1, an application can take advantage of the entire logical space as if it were physically present. Deduplication may even increase this further by deduplicating larger "chunks" of data that compression cannot act upon. In the end, the number and choice of which COMs the submitter chooses to enable is up to them. If one or more of them poses a performance or a counteractive capacity penalty, they may elect not to include them in the evaluation.

In an enterprise data center environment that is essentially in continuous operation (24-7), the active metrics will weigh considerably more heavily than will the idle power metric, and COMs will have to be selectively applied to provide the optimal benefit.

5 Source: "Software Defined Storage Energy Efficiency Features," The Green Grid Forum 2016

7. WHAT THE SNIA EMERALD DATA SHOWS FOR STORAGE

The SNIA Emerald test data used for the data analysis was compiled in March 2022. The data consisted of 160 block I/O systems after OEM systems were removed from the dataset. As there was very limited data available for file access systems, that data is not included in this analysis. There are 36 Online 2 Systems, 55 Online 3 systems, and 69 Online 4 systems. The data are further broken down by workload type, Transactional or Sequential, for which the given configuration was optimized (Table 11).

Table 11 – ENERGY STAR configurations of data presented in this paper

System type	Number of Systems					
	Total	7.2K RPM	10K RPM	15K RPM	Mixed RPM	SSD
Online 2 Transactional	15	1				14
Online 3 Transactional	33	4	18	10		1
Online 4 Transactional	41	4	8	9	2	18
Online 2 Sequential	21	10	10	1		
Online 3 Sequential	22	7	14	1		
Online 4 Sequential	28	4	12	12		

The performance of a storage system is dependent on the controller, controller cache, front end and back-end interconnects, storage media capacity, and the storage media drive speed (if applicable). The following sections will look at the transactional performance per watt data and the sequential performance per watt and how the values are influenced by taxonomy, drive type and drive speed.

Data from block-based systems optimized for Transactional Workloads

Transactional workloads for block storage systems are characterized by the SNIA Emerald Hot Band metric. Figure 10 and Figure 11 show that there is no correlation between the Hot Band performance and the Idle Power metric. The terms Spec 1 and Spec 2 refer to Energy Star for Data Storage Specifications 1.0 and 2.0

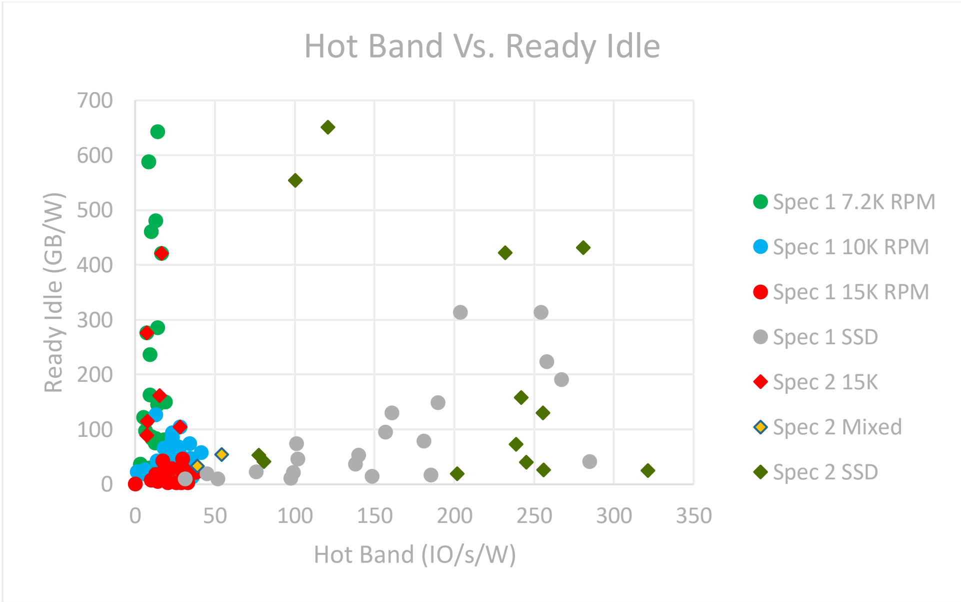

Figure 10: Hot Band vs. Ready Idle

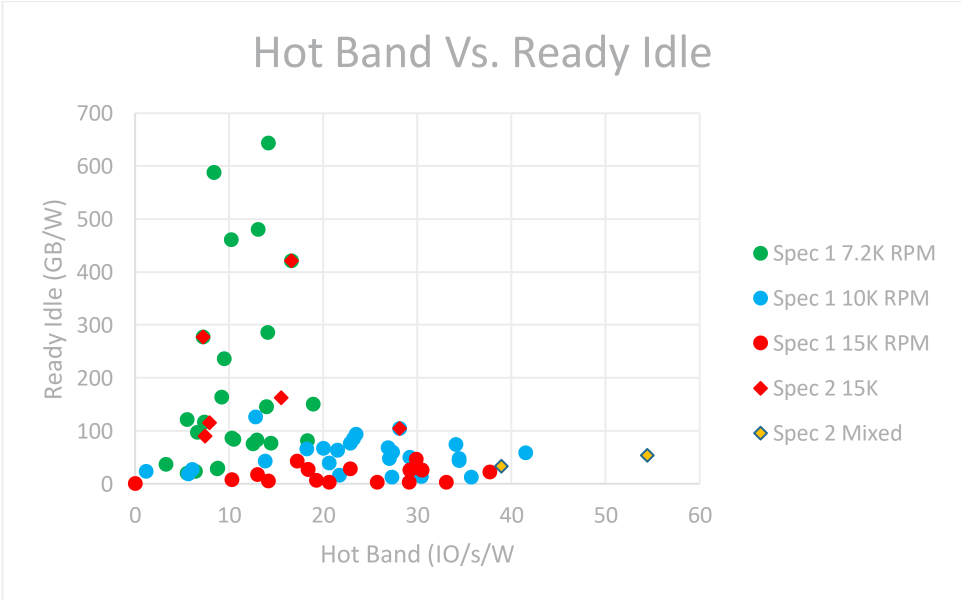

Figure 11: Hot Band vs. Ready Idle (Expanded Scale)

Figure 12 and Figure 13 show the relationship between drive speed and the SNIA Emerald Hot Band Metric. These figures have the current Energy Star for Data Center Storage Version 2.1 limit[6] for reference. For this requirement, none of the 7.2K RPM drives pass the limit.

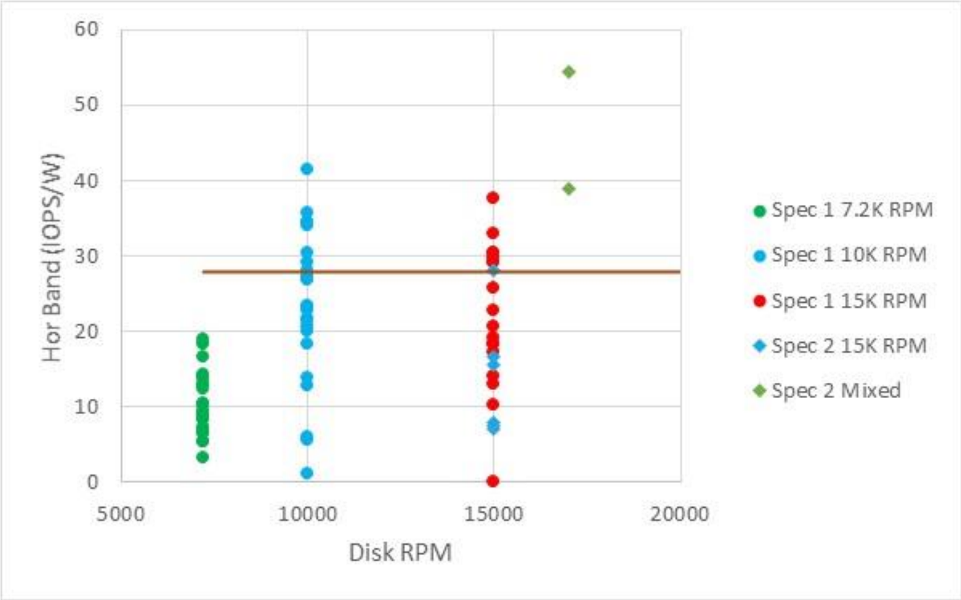

Figure 12: Got Band Metric vs. Drive Speed (HDDs)

Figure 13 adds the SSD drive data. This shows two facts. The Solid State Drive systems have significantly higher performance when compared to spinning media drives. The graph also shows that all the SSD drives systems pass the Energy Star Data Center Storage V2.1 limit[6].

[6] ENERGY STAR Data Center Storage Version 2.1 Final Specification, Clause 3.3.1.

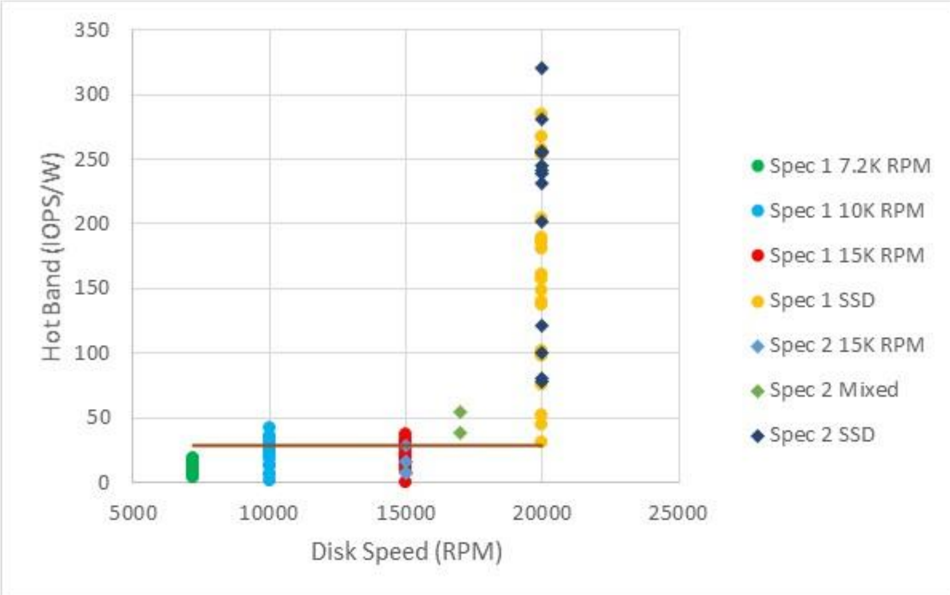

Figure 13: Hot Band Metric vs. Drive Speed (HDDs and SSDs)

Data from block-based systems optimized for Sequential Workloads

Block based systems optimized for Sequential Data are characterized by the SNIA Emerald Sequential Read and Sequential Write workloads.

Figure 14: Sequential Read vs. Ready Idle MetricsFigure 14 shows that there is no correlation between the Sequential Read and the Ready Idle power metrics.

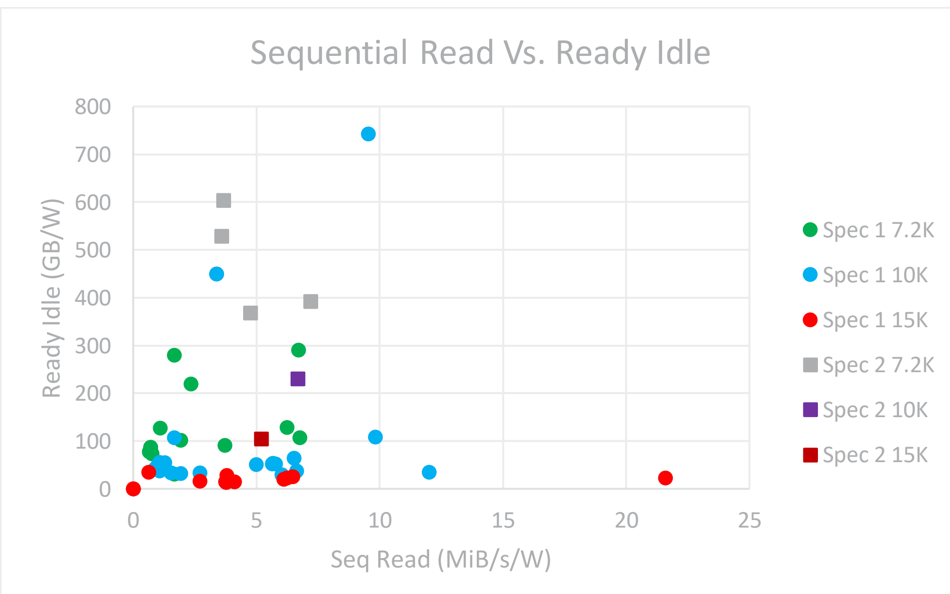

Figure 14: Sequential Read vs. Ready Idle Metrics

Figure 15 shows the relationship between drive speed and the Sequential Read Metric. The Energy Star for Data Center Storage Version 2.1 limit[7] is included in the graph for reference.

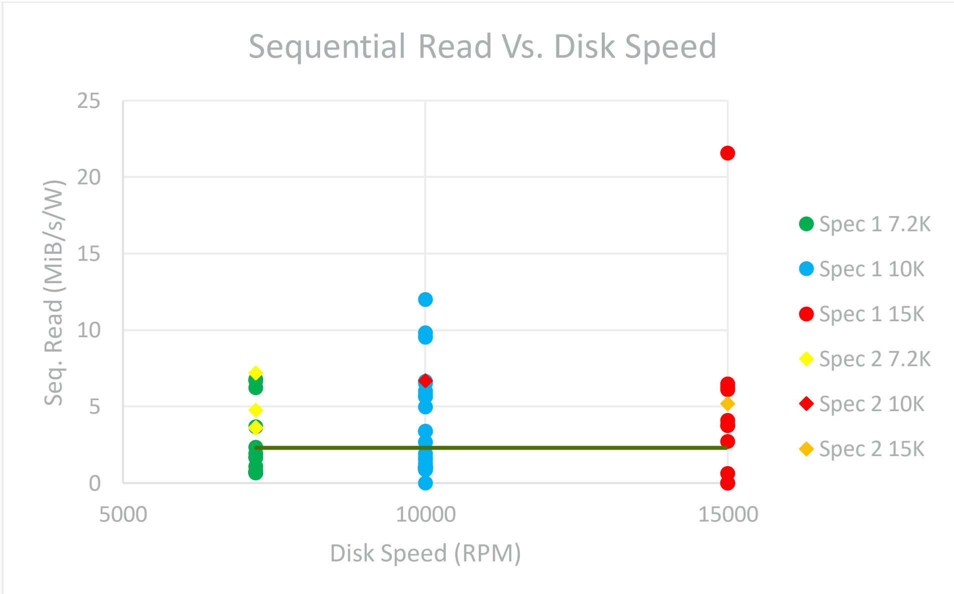

Figure 15: Sequential Read Metric vs. HDD Speed

Figure 16 shows that there is no correlation between the Sequential Write power efficiency metric and the idle power metric.

[7] ENERGY STAR Data Center Storage Version 2.1 Final Specification, Clause 3.3.1.

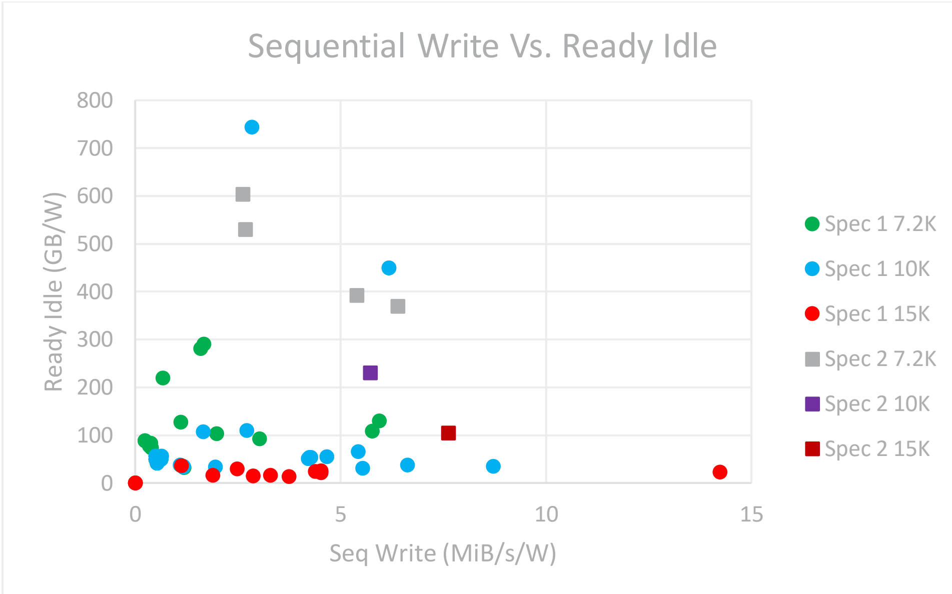

Figure 16 Sequential Write vs. Ready Idle Metrics

Figure 17 shows the relationship between drive speed and the Sequential Write Metric. The Energy Star for Storage Version 2.0 Limit is included for reference.

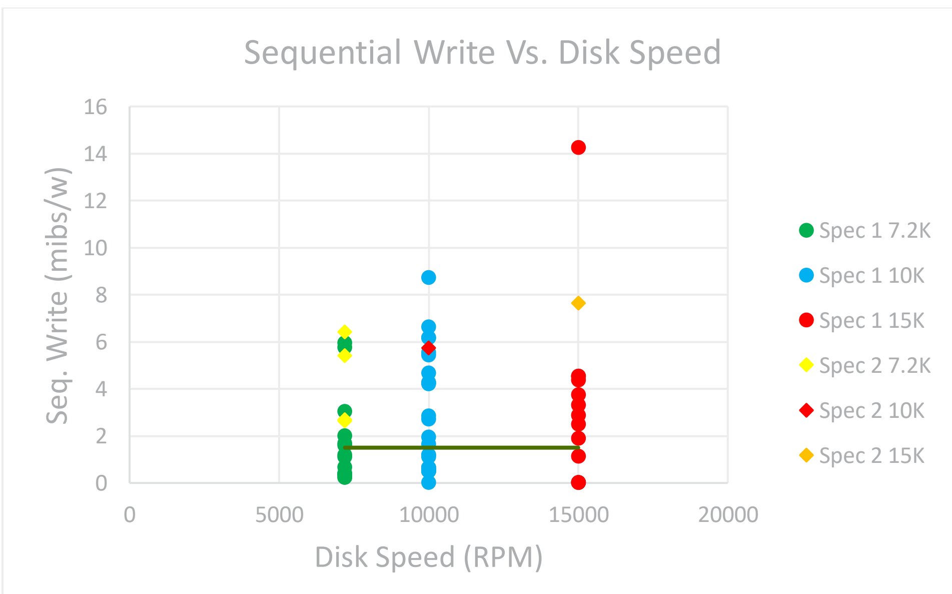

Figure 17: Sequential Write Metric vs. HDD Speed

Observations

Observations from the data for block I/O systems:

- Ready Idle has no correlation with the other performance metrics and tends to punish the higher performing systems.
- 7.2K RPM drive systems do not perform well on the Hot Band criteria.
- 7.2K RPM drive systems perform reasonably well on Sequential tests.
- SSD systems perform very well on Hot Band criteria.
- 10K and 15K RPM drive systems perform reasonably well on the Hot Band and Sequential criteria.
- The Energy Star Data Center Storage Version 2.1 limits meet the goal of rewarding the higher performance systems.
- Block systems optimized for Transactional or Sequential do not show a significant dependence on Taxonomy category. There is a significant dependence on drive speed and drive type.

There are no observations regarding file I/O systems due to insufficient data.

8. IDLE

Deep Idle

A storage device is in the deep idle state when either the power is completely removed or the device is in a spun down state, in the case of HDDs while the electronics are on or turned off. In cases where the electronics are turned off, the device can no longer respond to IO requests without intervention to re-enable that ability. In enterprise (data center) deployments), a storage device in the deep idle state is very rare once it has been installed. This is true of not only network connected external devices, but also direct connected devices perhaps residing in the same enclosure as the compute resources.

In all other states, there will be some level of background activity, often but not always for the sake of data integrity, capacity optimization or, in the case of flash-based devices, wear leveling.

Ready Idle

Ready Idle is the state of a system when no I/O requests are present, but the system is ready to accept I/O requests and respond within the Mean Time To First Data (MTTFD) of the specified Taxonomy Category. For Taxonomy Classification Online systems, the MTTFD must be less than 80 msec.

Ready idle power consumption of magnetic disk devices is a function of the drive speed and capacity. For a given drive speed and form factor, the Ready Idle metric can be improved by increasing the drive capacity. For example, 1 7200 RPM 3.5” drive with a capacity of 1 TB will consume roughly half of the power of a similar drive with a capacity of 10 TB, and therefore the 10 TB drive Ready Idle metric will be approximately 5 times greater than that of the 1 TB drive.

Ready Idle is the most common state in which storage devices reside when not executing host IO requests. This state applies to network connected and direct connected devices. The following is a non-exhaustive enumeration of the afore mentioned background activities, or processes. These generally execute when there is low host IO demand to optimize the response time.

While in the Ready Idle state, storage systems and devices commonly perform “Smart” data integrity checks. These involve reading data and checking for errors. This may result in replacing HDD sectors or SSD pages with spare sectors or pages if the read retry count exceeds a threshold value. Storage systems may replace an entire storage device if the read retry count exceeds a threshold value.

While in the Ready Idle state, storage systems/devices may perform deep data deduplication. This involves looking for multiple copies of the same data and replacing these with references to a single copy of the data. Data deduplication may be done over a collection of files, over multiple LUNs, over the entirety of multiple storage devices, and/or over entire data centers.

Given the above enumerations, it’s obvious that a device in the Ready Idle state is anything but idle, as these processes execute unthrottled during the time when no host I/O requests are present. This state often represents the highest power consumption for storage devices.

Solid State Drive Idle States

Activity while in deep idle time depends very much on the proprietary algorithms at work within the SSD. These have in common some or all of the following behaviors.

- The device goes into energy savings mode.
- The device shuts off the memory bank.

- In some cases, an external algorithm will shut down the memory after a specific time has passed, based on the energy savings mode.
- In most current implementations, it takes more than 100 msecs to reactivate a device once it has entered the deep idle state.
- Except for the different forms of "wake up," nothing else happens during the deep idle (deep sleep) mode.
- During the wake-up period the device may undertake a complete scan of its memory, which will take even more time.

Based on this, deep idle is not an acceptable mode for online storage systems as online storage systems require a maximum time to first data of 80 msecs.

The power consumption as a function of time of several SSD drives was measured in the ready idle state with no I/O from the system.

For these measurements, only the Idle_A power state was enabled where possible, so that only the Active and Idle_A states are utilized.

Figure 18 is representative of the results.

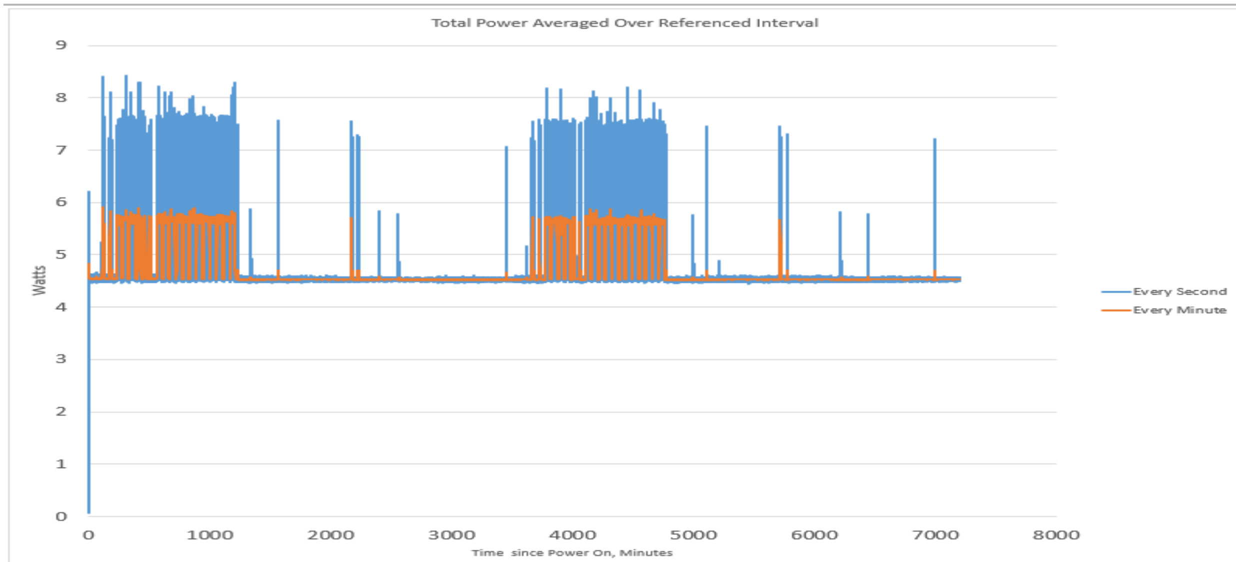

Figure 18: 7680 GB SAS SSD

The two colored lines in Figure 18 show average power over a) 1 second and b) 1 minute intervals.

As can be seen, many background activities are present during the ready idle state.

Magnetic Disk Drive Idle States

The following graphs show the power consumption over time of magnetic disk drives in the ready idle state,

Idle modes are specified in the INCITS T10 and T13 specifications. The details below are for a specific manufacturer's implementation but will be similar for all manufacturers.

Idle_A is an idle mode that supports the 80 msec Mean Time to First Data requirements of online storage. In the Idle_A state, most of the drive's servo system is disabled, reducing the drive's processor and channel power consumption. In the Idle_A state, the discs rotate at full rated speed,

Figure 19 is typical of the behavior of magnetic disk drives.

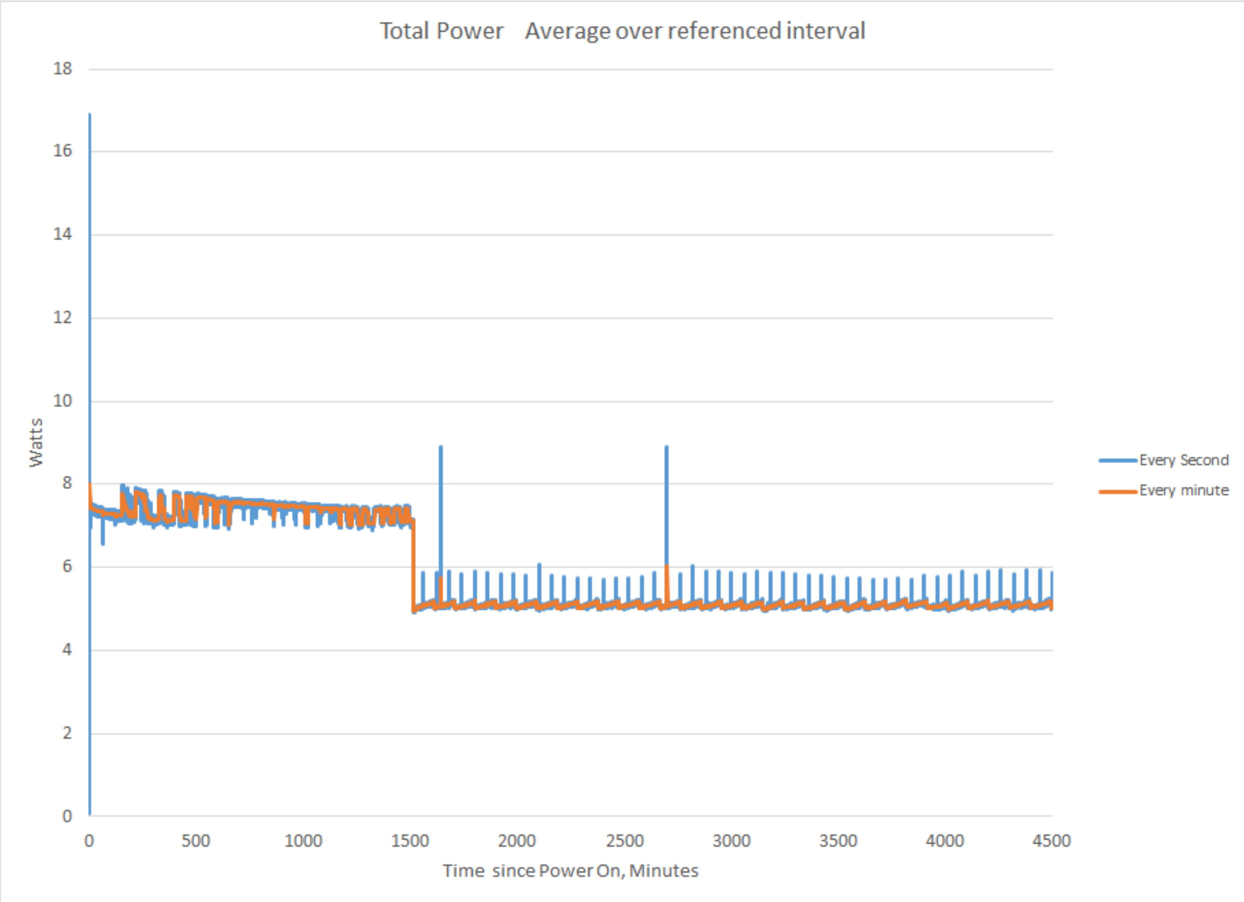

Figure 19: Example Power Consumption of Idle Magnetic Disk Drive

The two colored lines in Figure 19 show average power over a) 1 second and b) 1 minute intervals.

The activity in Figure 19 shows the drive performing a data scrub (typically done once a week) and then entering the idle_A condition. As can be seen, the disk drive has considerable activity during this "idle" period.

9. Japan Standards

Japan Energy Efficiency Standard (Top Runner)

Japan has adopted new standards for magnetic disk storage units.

Ministry of Economy, Trade, and Industry (METI) Ministerial Order and Public Notices, April 19, 2021; *new energy efficiency standards for magnetic disk units as the target year of FY2023* (Top Runner), is available at:

- https://www.meti.go.jp/english/press/2021/0419_001.html
- https://www.enecho.meti.go.jp/category/saving_and_new/saving/enterprise/equipment/toprunner/en/07_jikidisk.html

The Act on the Rational Use of Energy (Act No. 49 of 1979) is available at:

- https://www.japaneselawtranslation.go.jp/en/laws/view/3959

Figure 20 is a plot of the SNIA Emerald Ready Idle metric, in GB/W, for the various drive speeds for Transactional Optimized systems. The 20000 RPM column is actually Solid State Devices. The lines are the limits specified in the Japan Energy Efficiency Standard.

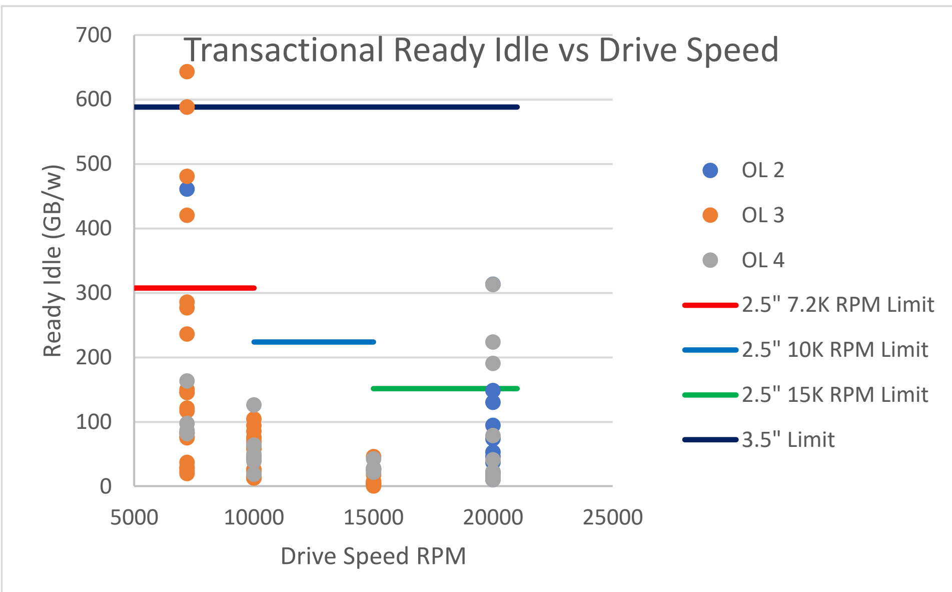

Figure 20: Ready Idle Metric vs. Drive Speed

As can be seen, the 10000 RPM and 15000 RPM drive systems all fall outside of the limits set by the Japan standard.

Figure 21 is a plot of the SNIA Emerald Ready Idle metric, in GB/W for the various drive speeds for Sequential Optimized systems.

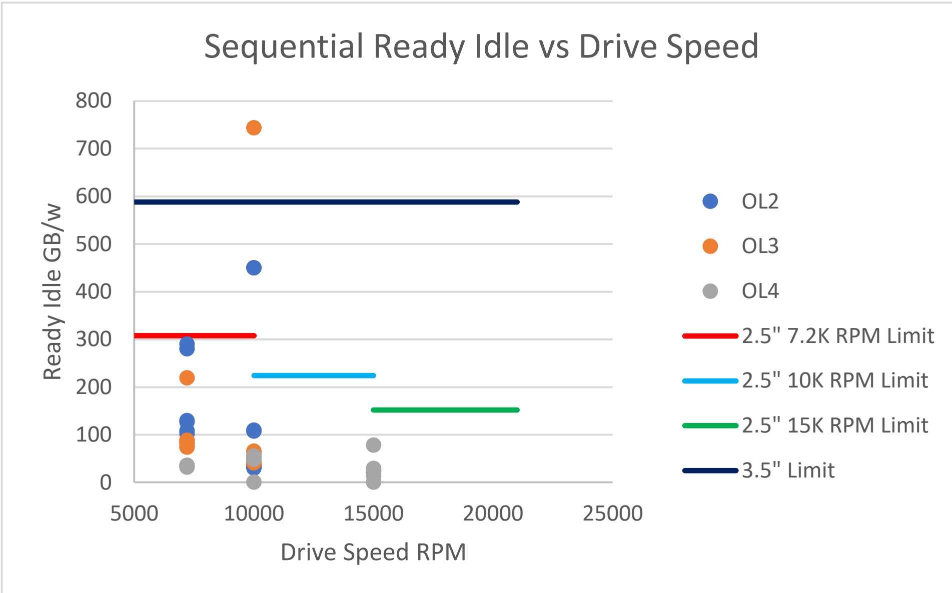

Figure 21: Ready Idle Metric vs. Drive Speed for Sequential Optimized Systems

Again, the 10000 RPM and 15000 RPM drive systems all fall outside of the limits set by the Japan standard.

Figure 22 shows the Ready Idle metric versus drive capacity for 3.5" 7200 RPM Drives for all systems.

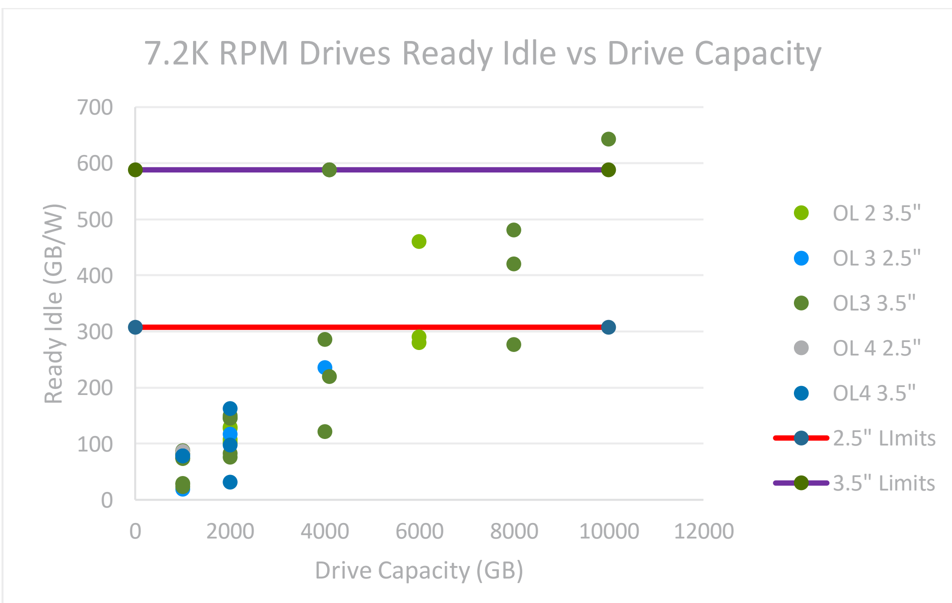

Figure 22: Ready Idle Metric vs. Drive Capacity (7200 RPM Drives)

Based on the data in Figure 22, the Japan idle limits for systems based on the smaller capacity 7.2K RPM drives would be a problem, but the newer 7.2K RPM drives are now of capacity 12,000 GB and larger while still consuming about the same power as the smaller capacity drives of the same form factor and rotational speed. That will give the equivalent Ready Idle value of a 12,000 GB drive approximately 12 times the value of a 1000 GB drive, which should elevate some of the above systems to a passing value.

Based on the data plotted in Figure 20 and Figure 21, none of the 2.5" drive systems with 10K RPM or 15K RPM drives will not pass the Top Runner limits. The 15K RPM drives are being replaced with either 10K RPM drives or Solid-State Drives. There is no further development underway for 10K RPM drives, so no improvements will be seen in the idle performance of systems using 10K RPM Drives. The area where drive capacity is being improved is for the 7.2K RPM 3.5" drives. The above data shows that systems using the higher capacity 7.2K RPM drives will have a high probability of passing the limits.

Japan Green Procurement

The "Act on Promoting Green Procurement" is another regulation that is different from the "Act on Rationalizing Energy Use" (Top Runner).

The "Act on Promoting Green Procurement" is managed by Ministry of the Environment, not by METI (Ministry of Economy, Trade and Industry) who manages the Top Runner regulatory program.

The "Act on Promoting Green Procurement" defines the criteria of each system for users who purchase the systems, while the Top Runner regulation defines the criteria of total system shipments for venders who manufacture and supply systems.

The criteria of the "Act on Promoting Green Procurement" are basically defined according to the criteria of Top Runner. The criteria of category V and VI of the storage systems, in which 12 or more HDDs are installed, are relaxed from the Top Runner criteria:

- Category V (Storage system including 3.5" HDDs)
  - Top Runner: E＝0.00170
  - Green Procurement: E＝0.00213 (= 0.00170/0.8)
- Category VI (Storage system with 2.5" HDDs)
  - Top Runner: E＝exp(0.952×ln(N)-14.2)
  - Green Procurement: E＝exp(0.952×ln(N)-14.2)/0.5

The "Act on Promoting Green Procurement" may be found at:

- http://www.env.go.jp/policy/hozen/green/attach/gpp%20pamphlet_eng.pdf

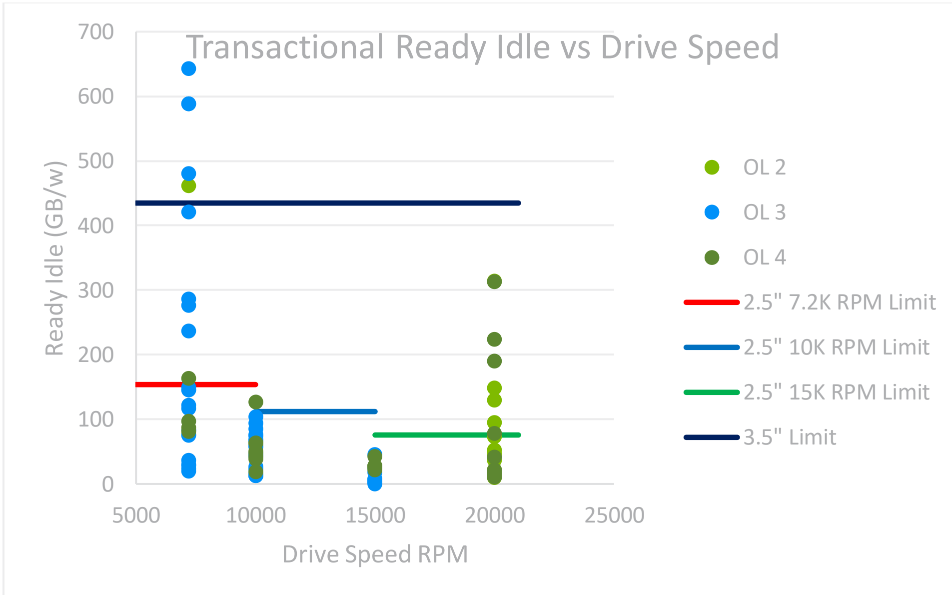

Figure 23: Ready Idle Metric vs. Drive Speed (Transactional Systems)

As can be seen in Figure 23, the 10000 RPM and 15000 RPM drive systems still mostly fall outside of the limits set by the Japan standard.

Figure 24 is a plot of the SNIA Emerald Ready Idle metric, in GB/W for the various drive speeds for Sequential Optimized systems.

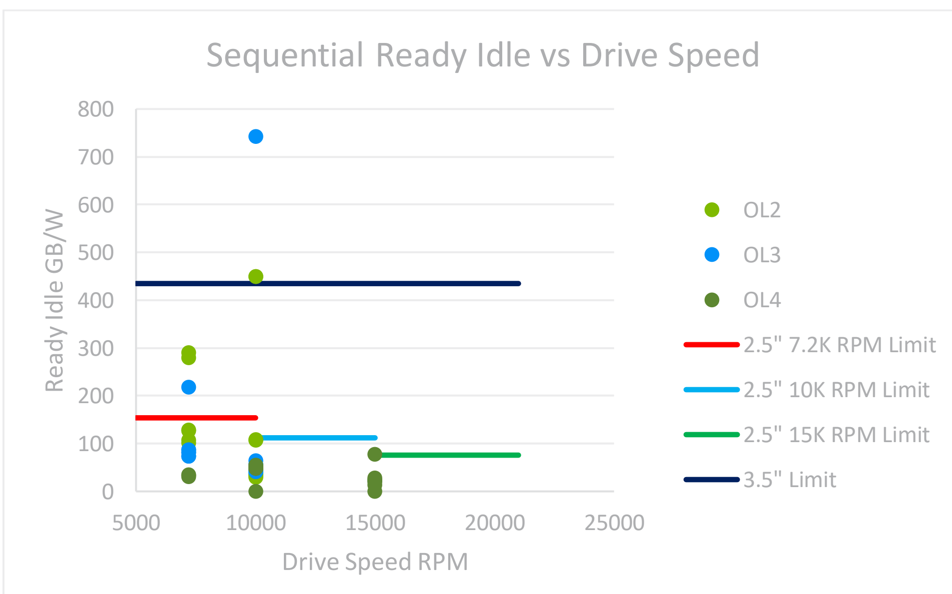

Figure 24: Ready Idle Metric vs. Drive Speed (Sequential Systems)

Again in Figure 24, the 10000 RPM and 15000 RPM drive systems all fall outside of the limits set by the Japan standard.

Figure 25 shows the Ready Idle metric versus drive capacity for 3.5” 7200 RPM Drives for all systems.

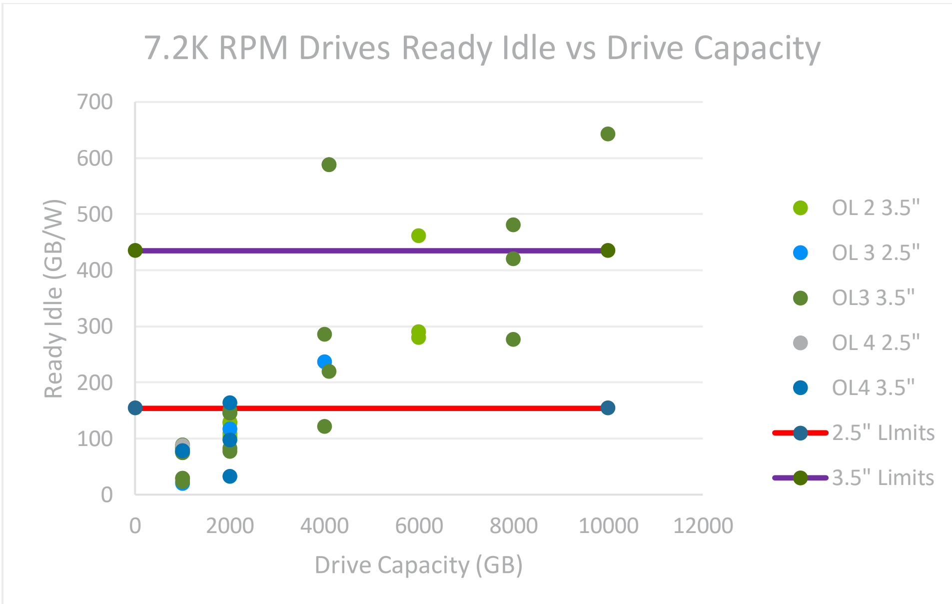

Figure 25: Ready Idle Metric vs. Drive Capacity (All Systems)

Based on the data Figure 25, the Japan Green Procurement idle limits for systems based on the smaller capacity 7.2K RPM drives would be a problem, but the newer 7.2K RPM drives are now of capacity 12,000 GB and larger while still consuming about the same power as the smaller capacity drives of the same form factor and rotational speed. That will give the equivalent Ready Idle value of a 12,000 GB drive approximately 12 times the value of a 1000 GB drive, which should elevate many of the above systems to a passing value.

Based on the data plotted in Figure 23 and Figure 24, most of the 2.5” drive systems with 10K RPM or 15K RPM drives will not pass the limits of the Japan Green Procurement regulation. The 15K RPM drives are being replaced with either 10K RPM drives or Solid-State Drives. There is no further development underway for 10K RPM drives, so no improvements will be seen in the idle performance of systems using 10K RPM Drives. The area where drive capacity is being improved is for the 7.2K RPM 3.5” drives. The above data shows that systems using the higher capacity 7.2K RPM drives will have a high probability of passing the limits.

10. CONCLUSIONS AND RECOMMENDATIONS FOR POWER EFFICIENCY CERTIFICATION

Conclusions

Analysis of the data submitted to ENERGY STAR Data Center Storage program has led to some interesting observations.

Observations regarding the Ready Idle Metric:

- Ready Idle Metric does not correlate to either the transactional metrics or the streaming metrics. In particular, the Ready Idle Metric shows more of an inverse correlation to the transactional metrics.
- Ready Idle Metric for storage systems based on magnetic disks of a given speed is directly related to drive capacity. To improve the Ready Idle Metric, one only needs to change to a higher capacity drive of the same speed. Drive capacities increase periodically, so one can get an improvement by waiting for the next capacity to be released. For a given family of drives, the power increase is negligible for a capacity increase.
- Ready Idle power is a function of the amount of activity (COMs, device- and system- level housekeeping, etc.) performed by the system while "idle".

The family data for minimum, optimal, and maximum drive count varied less than 10% for transactional metrics and less than 20% for sequential metrics. Based on this, we conclude that min/max testing provides no additional value.

Data center storage is a challenging category to characterize for power efficiency. It ranges in scale from a dozen devices to hundreds of devices. The devices themselves come in many different combinations of form factors, performance levels, interfaces, and capacities, and can be deployed in highly individualized configurations depending on the purchasers' workloads and performance expectations. Technological change continues to develop at a rapid pace, with value-added features appearing in both the operating environment and in the actual drives. At the same time, the number of component manufacturers is not extremely large, meaning that many hardware components are common to multiple storage system product lines.

This complexity means that characterizing a broad set of configurations for power efficiency requires a significant effort. It also means that a significant investment in equipment is required for testing. The resource investment is significant enough that companies may decide not to perform power efficiency testing.

Recommendations

The goals of power efficiency testing are to recognize the best performing products and to incentivize power efficiency improvements. In keeping with these goals, we recommend:

1. Define a set of criteria for product certification.
2. Define a simplified active test regimen to encourage broad participation by vendors.

Criteria Recommendations

1. Product must be qualified for operation at the ASHRAE A2 level, at a minimum.
2. PSUs must be 80 PLUS Gold or better for products with multi-output power supplies and 80 plus Platinum or better for products with single output power supplies.
3. A minimum number of COMs must be supported based on the classification as given by the taxonomy defined in the SNIA Emerald Power Efficiency Measurement Specification:

Online 1.5:	none
Online 2:	1
Online 3:	2
Online 4, 5, 6:	3

4. Inlet temperature reporting must be provided.
5. Input power reporting must be provided.

Test Recommendations

Running the characterization process for a system is complex and resource-intensive. To realize a test approach that is both effective and limited in its resource requirements, we recommend a test approach, based on the SNIA Emerald Power Efficiency Measurement Specification, where a single vendor-selected "optimum" configuration for a product family is tested with the preferred device type(s). This would provide data center operators an assessment of the storage product active performance/power metric. All configurations of a product family would be characterized by the single test.

It is particularly important that "idle" power, e.g., the Ready Idle Metric, not be used to assess storage product operational power efficiency. The Ready Idle Metric is not representative of the conditions under which typical enterprise storage systems operate. The Ready Idle Metric is completely uncorrelated to operational power efficiency and will not identify the systems that most efficiently perform the work in enterprise environments.

Summary

These recommendations are applicable for any storage product power efficiency requirements or programs being considered for market entry or procurement purposes.

11.The Green Grid and SNIA

The Green Grid

The Green Grid (TGG) is an affiliate member organization of the Information Technology Industry Council (ITI). TGG works globally to create tools, provide technical expertise, and advocate for the optimization of energy and resource efficiency of Data Center ecosystems which enable a low carbon economy. TGG is an open industry consortium of data center operators, cloud providers, technology and equipment suppliers, facility architects, and end-users. TGG offers the technical expertise that the data center industry and governments turn to for insight and counsel regarding data center efficiency and sustainability. For more information please visit https://www.thegreengrid.org/en.

SNIA

The Storage Networking Industry Association (SNIA) is a non-profit organization made up of member companies spanning information technology. A globally recognized and trusted authority, SNIA's mission is to lead the storage industry in developing and promoting vendor-neutral architectures, standards and educational services that facilitate the efficient management, movement and security of information.

The SNIA Green Storage Initiative (GSI) is dedicated to advancing energy efficiency and conservation in all networked storage technologies and minimizing the environmental impact of data storage operations. The GSI's Mission:

- Establish and maintain the SNIA Emerald™ Program for SNIA Emerald™ Energy Efficiency Measurement and conduct training of SNIA Emerald™ testers and industry stakeholders
- Educate the IT industry, vendor community and regulatory bodies on techniques to conserve energy for enterprise storage environments
- Provide external advocacy and support of the technical work of the SNIA Green Storage Technical Working Group (TWG)
- Provide input to the SNIA Green Storage TWG on requirements for green storage measurement specifications, metrics and standards
- Establish and maintain cross-industry relationships and alliances to coordinate and advance data center energy efficiency related programs, test and measurement methods, and standards

The Green Storage TWG acts as the primary technical entity for the SNIA to identify, develop, and coordinate technical matters related to energy and cooling of storage networking products. To be effective, standard metrics must be a realistic reflection of efficiency: work accomplished per energy expended, allowing a determination of clear costs for operations. We believe that a variety of types of storage work must be considered, and that several distinct 'figures of merit' will be needed. We believe that the SNIA is uniquely qualified, as an impartial forum with the right participants, to create useful metrics for users.

The SNIA Emerald™ Power Measurement Specification identifies metrics and defines a methodology by which power consumption and efficiency of storage networking products can be measured for the purposes of new product development, end-user customer evaluation, and regulatory standards development. The Specification is a SNIA Standard and version 3.0.3 has been adopted as International Standard ISO/IEC 24091:2019.