ENERGY STAR® Data Center Storage Initial Data Collection Procedure December 2009

Objective

As part of the ENERGY STAR Data Center Storage specification development process, the EPA is inviting interested parties to perform a series of energy performance tests on data center storage systems using the protocol specified in this document.

The objective of this first round of testing is to understand the relationships between hardware/software configuration and energy performance in both active and idle states. EPA would like to collect a substantial amount of test data (possibly supplemented by simulated test results) in order to perform a sensitivity analysis on the effect of single-variable configuration changes on power consumption. Items such as Hard Disk Drive (HDD) selection (e.g. capacity vs. performance), Reliability, Availability, Serviceability (RAS) features (e.g. single vs. redundant controllers, RAID level), and use of Small Form Factor (SFF) and Solid State Disk (SSD) technologies are all of interest.

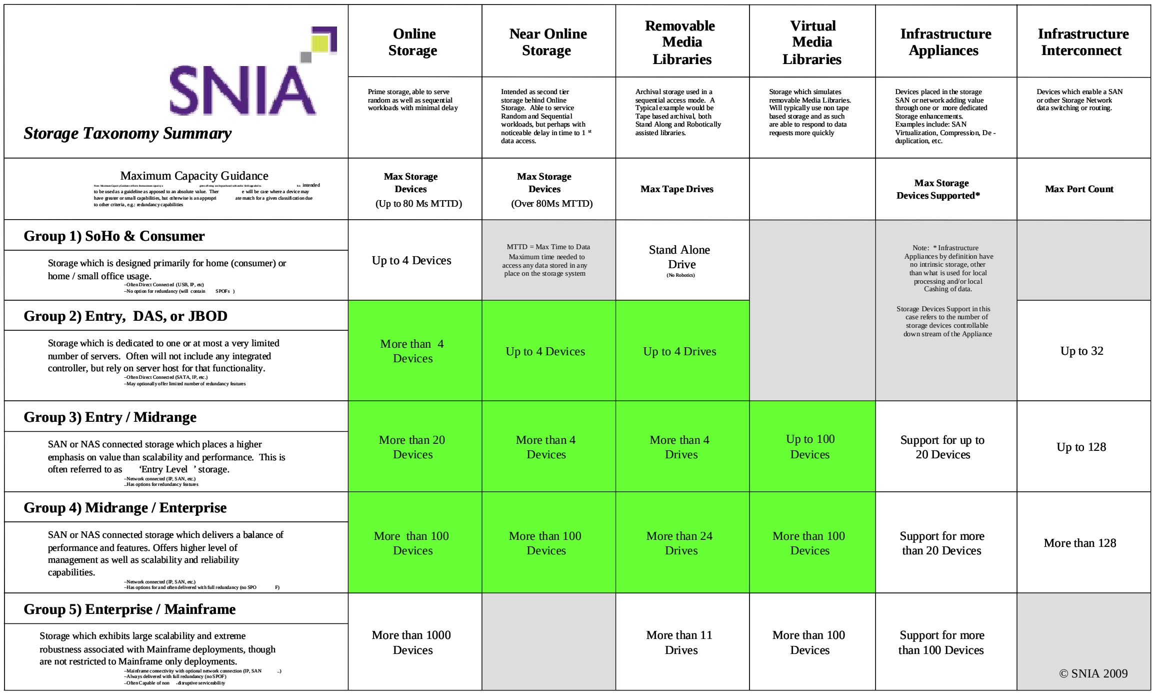
The storage systems described in this document are intended to represent a superset of configurations from which stakeholders will select a small number of configurations for testing. EPA understands that there are limitations on time, personnel, and hardware resources for this data collection, and that all responses are subject to the goodwill and best efforts of the stakeholder community. In light of these limitations, and in order to facilitate the generation of the most robust and valuable data set, EPA has developed additional guidance for test system selection. Please see the “Product Selection” and “Test Performance” sections of this document for more information.

Data submissions will be collated, anonymized, and made available to the stakeholder community for analysis in preparation for the first draft Version 1.0 ENERGY STAR Data Center Storage specification.

Scope

This document references the v0.0.18 DRAFT “SNIA Green Storage Power Measurement Specification” Storage Taxonomy Summary, shown here as Figure 1.

Figure 1: SNIA GSI Storage Taxonomy [1]



SNIA *Storage Taxonomy Summary*	**Online Storage**	**Near Online Storage**	**Removable Media Libraries**	**Virtual Media Libraries**	**Infrastructure Appliances**	**Infrastructure Interconnect**
	Prime storage, able to serve random as well as sequential workloads with minimal delay	Intended as second tier storage behind Online Storage. Able to service Random and Sequential workloads, but perhaps with noticeable delay in time to 1st data access.	Archival storage used in a sequential access mode. A Typical example would be Tape based archival, both Stand Along and Robotically assisted libraries.	Storage which simulates removable Media Libraries. Will typically use non tape based storage and as such are able to respond to data requests more quickly	Devices placed in the storage SAN or network adding value through one or more dedicated Storage enhancements. Examples include: SAN Virtualization, Compression, De-duplication, etc.	Devices which enable a SAN or other Storage Network data switching or routing.
Maximum Capacity Guidance Note: Maximum Capacity Guidance reflects the maximum capacity a given offering can be purchased with and/or field upgraded to. It is intended to be used as a guideline as apposed to an absolute value. There will be case where a device may have greater or small capabilities, but otherwise is an appropriate match for a given classification due to other criteria, e.g.: redundancy capabilities	**Max Storage Devices** (Up to 80 Ms MTTD)	**Max Storage Devices** (Over 80Ms MTTD)	**Max Tape Drives**		**Max Storage Devices Supported***	**Max Port Count**
Group 1) SoHo & Consumer Storage which is designed primarily for home (consumer) or home / small office usage. –Often Direct Connected (USB, IP, etc) –No option for redundancy (will contain SPOFs)	Up to 4 Devices	MTTD = Max Time to Data Maximum time needed to access any data stored in any place on the storage system	Stand Alone Drive (No Robotics)		Note: * Infrastructure Appliances by definition have no intrinsic storage, other than what is used for local processing and/or local Cashing of data. Storage Devices Support in this case refers to the number of storage devices controllable down stream of the Appliance	
Group 2) Entry, DAS, or JBOD Storage which is dedicated to one or at most a very limited number of servers. Often will not include any integrated controller, but rely on server host for that functionality. –Often Direct Connected (SATA, IP, etc.) –May optionally offer limited number of redundancy features	More than 4 Devices	Up to 4 Devices	Up to 4 Drives			Up to 32
Group 3) Entry / Midrange SAN or NAS connected storage which places a higher emphasis on value than scalability and performance. This is often referred to as 'Entry Level' storage. –Network connected (IP, SAN, etc.) –Has options for redundancy features	More than 20 Devices	More than 4 Devices	More than 4 Drives	Up to 100 Devices	Support for up to 20 Devices	Up to 128
Group 4) Midrange / Enterprise SAN or NAS connected storage which delivers a balance of performance and features. Offers higher level of management as well as scalability and reliability capabilities. –Network connected (IP, SAN, etc.) –Has options for and often delivered with full redundancy (no SPOF)	More than 100 Devices	More than 100 Devices	More than 24 Drives	More than 100 Devices	Support for more than 20 Devices	More than 128
Group 5) Enterprise / Mainframe Storage which exhibits large scalability and extreme robustness associated with Mainframe deployments, though are not restricted to Mainframe only deployments. –Mainframe connectivity with optional network connection (IP, SAN ..) –Always delivered with full redundancy (no SPOF) –Often Capable of non-disruptive serviceability	More than 1000 Devices		More than 11 Drives	More than 100 Devices	Support for more than 100 Devices	© SNIA 2009

For purposes of Round 1 data collection, EPA is interested in the following taxonomy classifications:

- ONLINE: Groups 2, 3, 4
- NEAR-ONLINE: Groups 2, 3, 4
- REMOVABLE MEDIA LIBRARY: Groups 2, 3, 4
- VIRTUAL MEDIA LIBRARY: Groups 3, 4

Definitions[2]

High-availability (HA)[3]: The ability of a system to perform its function continuously (without interruption) for a significantly longer period of time than the reliabilities of its individual components would suggest. High availability is most often achieved through failure tolerance.

Maximum Sustainable Performance: Maximum IOPS or GiB/s that the UUT is able to deliver under a specified workload. For purposes of this data collection procedure, it is suggested that "sustainable" performance is that which ensures the maximum achievable data rate, excludes any transient system caching effects, and can be maintained to within ±5% for the duration of the measurement phase.

[1] For more information, visit www.snia.org/green

[2] Preliminary definitions for "storage product," "operating states," etc. are included in this data collection procedure, based on feedback from stakeholders. All terminology in this document will be subject to further refinement later in the specification development process.

[3] Source: 2009 SNIA Dictionary

Maximum Time to Data (MaxTTD): The maximum time before an entire data object is accessible within the constraints imposed by its storage media. For random-access media, a data object is accessible when any byte may be accessed. For sequential-access media, a data object is accessible when the requested object has begun streaming from a previously inactive drive.

Operating States:

- Active: The state in which a storage product is processing external I/O requests.
- Ready Idle: The state in which a storage product is able to respond to I/O requests within the MaxTTD limits for its taxonomy category, but is not receiving external I/O requests. The storage product may perform routine housekeeping tasks during Ready Idle, provided such operations do not compromise the product’s ability to meet MaxTTD requirements.
- Deep Idle: A state in which one or more storage product components or subsystems have been actively managed into a low-power state for purpose of conserving energy. A storage product in Deep Idle cannot respond to I/O requests within the MaxTTD limits for its taxonomy category, and may need to perform a managed ‘wake-up’ function in order to return to a Ready Idle or Active state. Deep Idle capability must be a user-selected, optional storage product feature.

Response Time: The time required for the UUT to complete an I/O request.

Storage Product: A fully-functional storage system configured for sale to an end user. A storage product contains all media, controllers, power distribution, host and internal interfaces, and other devices required to perform its intended function. Components and subsystems that are an integral part of the storage product architecture (e.g., internal SAN switches that provide aggregation functionality and present a consolidated external interface) are included in the definition of a storage product. In contrast, components that are normally associated with a storage environment at the data center level (e.g., external SAN network switches used to provide an interface with servers) are excluded from the definition of a storage product.

Unit Under Test (UUT): The storage product being tested.

Product Selection

Stakeholders are encouraged to submit power consumption data for as many products, in as many Generic System Configurations, as can be reasonably accomplished during the data collection period. Given the option of performing a few tests on a broad range of taxonomy categories, or performing in-depth testing of just a small number of products; EPA recommends that stakeholders perform an in-depth assessment of a small number of products.

Generic System Configurations (GSC): Four GSC options are proposed as a baseline test configurations.

GSC-1: Performance Non-HA Configuration
Systems intended for use by on-line applications requiring high-performance, measured either in IOPS or in bandwidth. System availability is not critical to the customer, so HA features are limited.
Example: FC drives + single controller

GSC-2: Performance HA Configuration
Systems intended for use by on-line applications requiring high-performance, measured either in IOPS or in bandwidth. System availability is critical to the customer, so HA features are present and configured.
Example: FC drives + redundant controllers

GSC-3: Capacity Non-HA Configuration
Systems intended for use in archival, Tier-2, or streaming applications, where streaming of media is of primary importance for reading and/or writing. System availability is not critical to the customer, so HA features are limited.
Example: SATA drives + single controller

GSC-4: Capacity HA Configuration
These systems are intended for use in archival, Tier-2, or streaming applications where streaming of media is of primary importance for reading and/or writing. System availability is critical to the customer, so HA features are present and configured.
Example: SATA drives + redundant controllers

Selection of specific configurations for testing is at the discretion of the individual stakeholder or test facility, and will depend on product options, availability of components, test equipment, and other resources. It is preferable to select GSC configurations that are representative of popular products with high volume sales.

All four GSC classifications do not apply to every taxonomy category. For example, Removable Media Library products would not likely serve performance-oriented GSC-1 or GSC-2 applications. There is no expectation that such unrealistic configurations be tested.

Configuration Changes: If possible, testing should be repeated after making single-variable changes to the hardware or software configuration. It is preferable to select configuration changes that are expected to have the greatest impact on energy performance. Examples of single-variable changes include[4]:

- Change drive type or technology
- Add or remove drives

[4] It is understood that some single-variable changes may change a product's GSC classification. Stakeholders are encouraged to thoroughly document configuration changes and expected impacts on system classification on the test data collection sheet in order to facilitate accurate data analysis.

- Modify RAID configuration[5]
- Change from single to redundant power supplies
- Change from single to redundant controllers
- Change controller cache size
- Modify host connection technology
- Change removable media maximum data rate or type
- Enable data reduction features

Hybrid / Asymmetrical Systems: EPA is aware of hybrid or asymmetrical systems which are difficult to categorize in the published taxonomy (e.g., a single system with both SSD devices for fast response and SATA HDDs or optical media for stale data). For purposes of this data collection, a hybrid system should be categorized based on the MaxTTD for the entire storage capacity.

Simulation Option: EPA continues to offer stakeholders the option of submitting simulated (modeled) data to supplement actual test data for this data collection procedure. Simulators must be capable of predicting variations in energy consumption for the various phases of a test sequence, including allowances for different workload stimuli. All simulations should be run using the test workload sequences defined for the appropriate taxonomy category. In order to assess simulation accuracy, stakeholders should model any tested system configurations to allow for a comparison of results. Further simulations can then be used to model the impact of an additional single- and multi-variable hardware and software configuration changes.

Test Setup

Environment: For purposes of this preliminary round of testing, all testing may be conducted at normal laboratory ambient temperature and humidity. No special environmental controls are necessary. Temperature and humidity must be recorded at the beginning of each major test sequence.

Power Meter: The power meter shall be capable of measuring and recording UUT input power with an accuracy of 1% and a sampling frequency of no more than 5 seconds.

Power Measurement: UUT input voltage and power should be measured at a location appropriate to capture the total power consumed by all components of the UUT. This may be at the PDU, or some other appropriate location. The power measurement should include all items needed to provide for the integration and operation of the UUT. This includes controllers, drawers, robotic assemblies, power distribution / PDUs as well as data networking used internal to the UUT (e.g., integrated SAN switches).

Input Power: It is anticipated that mains power typical of a normal customer installation will be used during this data collection exercise. Input voltage shall remain consistent to

[5] If only one RAID configuration is to be tested, RAID-5 is preferred for this round of data collection, as it is understood to be the most common RAID implementation. If RAID-5 is not available, RAID-0 or RAID-1 configurations are preferred.

±5% for the duration of the test period. The power supplied to the UUT shall be consistent with one of the following options:

Table 1: Input Power Requirements

Input Voltage Range	Phases	Input Frequency Range
100-120 VAC RMS	1	47-63 Hz
180-240 VAC RMS	3	47-63 Hz
200-240 VAC RMS	1	47-63 Hz
380-508 VAC RMS	3	47-63 Hz

Test Performance

Preconditioning: The preconditioning phase shall be of sufficient duration to ensure that I/Os are going to the storage media under steady-state conditions and are not being serviced by various system caches or other transients such as uninitialized space. For hybrid systems, preconditioning must be sufficient to allow the system to enter stable operation across the various storage media.

Sustained Duration: All test phases must be of sufficient duration to achieve steady-state system performance and mitigate the impact of system cache. Once stability has been achieved, the measurement period begins and shall continue for the "Sustained Duration" specified in the test sequence.

Slack Time: All steps in the test procedure shall be performed in sequence, with no more than 60 seconds delay between steps.

Data Distribution: For purposes of this data collection, testing should be distributed as evenly as possible across all storage media installed in the UUT[6]. Best practices suggest that test stimuli should be configured to exercise at least 80% of the available logical address space. Short-stroking and other techniques to artificially enhance the energy performance of the UUT are not permitted.

Measurement Reference: The point of reference for read measurements is the receipt of data by the host that initiated a read operation. The point of reference for write measurements is the receipt of an acknowledgment of successful data write by the host that initiated a write operation.

Hybrid / Asymmetrical Systems: If possible, hybrid systems should be tested twice: the first test run should exercise only the portion of total storage capacity with the fastest response time, and the second test run should be performed across the entire storage array. For systems that cannot meet this test objective, testers should attempt to get as

[6] This suggestion does not apply to Removable Media Libraries, which can access only a limited quantity of storage media at one time.

close to the above guidance as possible. For every test, the percentage of I/O requests handled by each type of storage media in the hybrid system shall be documented.

Virtual Media Libraries: For purposes of this data collection, data should be constrained to the disk-based portion of the device. If the device includes a removable media library back end, power consumption for the removable media library may be tested and reported separately.

Data Compression & Reduction: For purposes of this data collection, all tests should be performed with uncompressed data. As a baseline, any data reduction features (e.g., compression, deduplication, thin provisioning) should be disabled. Data reduction features may be activated as single-variable changes for testing, and should be documented accordingly, along with details about the data set utilized during the test.

Block I/O vs. File I/O: For purposes of this data collection, it is preferable to test systems using a Block I/O interface. Systems that also support File I/O may optionally be tested using a File I/O interface as resources allow. Systems that support only File I/O may use an external means to convert File I/O requests into Block I/O (e.g., via a virtualized filing system), or test only with File I/O.

Battery Backup: For purposes of this data collection, all batteries internal to the UUT shall be fully charged and in a maintenance or 'float' state at the start and for the duration of testing. Battery configurations should be disclosed on the test data collection sheet.

Test Procedure

ONLINE & NEAR-ONLINE CATEGORIES

The following test sequence shall be applied to systems in the Online and Near-Online taxonomy categories.

Table 2: Test Sequence for Online & Near-Online

Phase	Workload	% of Max Sustainable Performance	Block Size	Sustained Duration
Pre-Conditioning	Random 70% Read 30% Write	100%	8 KiB	≥ 30 min
Active "A"	Random Read	100%	8 KiB	10 min
Active "B"	Random Write	100%	8 KiB	10 min
Active "C"	Sequential Read	100%	256 KiB	10 min

Phase	Workload	% of Max Sustainable Performance	Block Size	Sustained Duration
Active "D"	Sequential Write	100%	256 KiB	10 min
Active "E"	Random 70% Read 30% Write	25%	8 KiB	10 min
Active "F"	Random 70% Read 30% Write	75%	8 KiB	10 min
Active "G"	Random 70% Read 30% Write	100%	8 KiB	10 min
Ready Idle	n/a	0%	n/a	30 min
Deep Idle	n/a	0%	n/a	10 min

Notes:

- Response Time for Online systems during Active test phases must not exceed 30 ms.
- The "Deep Idle" test phase may be omitted for systems which do not offer a Deep Idle feature. If Deep Idle data is collected, sufficient details about the system state (e.g., which subsystems are powered down) should be provided on the test data collection form.

REMOVABLE & VIRTUAL MEDIA LIBRARY CATEGORIES

The following test sequence shall be applied to systems in the Removable Media Library and Virtual Media Library taxonomy categories.

Table 3: Test Sequence for Removable & Virtual Media Libraries

Phase	Workload	% of Max Sustainable Performance	Block Size	Sustained Duration
Pre-Conditioning	Sequential Write → Rewind → Read	≥ 80%	128 KiB	10 min
Active "A"	Sequential Write ("x" drives)	≥ 80%	128 KiB	10 min
Active "B"	Sequential Write ("x + n" drives)	≥ 80%	128 KiB	10 min

Phase	Workload	% of Max Sustainable Performance	Block Size	Sustained Duration
Robotics	Media Load/Unload	n/a	n/a	n/a
Ready Idle	n/a	0%	n/a	30 min
Deep Idle	n/a	0%	n/a	10 min

Notes:

- The “Active A” test phase shall be performed with an arbitrary number of active drives, with robotics energized, idle, and ready to initiate a command. The “Active B” test phase shall be performed with a greater number of active drives than were used for “Active A,” under otherwise similar conditions.
- All “Active” test phases shall be performed as near to 100% of maximum sustainable performance as possible, but no less than 80%. Test results should be extrapolated out to 100%, and details of these calculations shall be supplied on the test data collection form. Any start/stop events should also be noted.
- The “Robotics” test phase shall be performed using an average-distance movement from a single robot. Accumulated energy for the entirety of the movement sequence shall be recorded. The sequence is as follows:
    1. Start and with robot in a neutral/central ready position
    2. Retrieve media from rack and load in an empty slot
    3. Unload media and return it to original position on rack
    4. Return robot to neutral/central ready position
- The “Ready Idle” test phase shall be performed with robotics energized, idle, and ready to initiate a command.
- The “Deep Idle” test phase may be omitted for systems which do not offer a Deep Idle feature. If Deep Idle data is collected, sufficient details about the system state (e.g., which subsystems are powered down) should be provided on the test data collection form.