May 21, 2010

TO: Stephen Pantano
ICF International
1725 Eye Street, NW, Suite 1000
Washington, D.C. 20006
spantano@icfi.com

CC: Una Song
EPA
song.una@epa.gov

FROM: Subodh Bapat
Oracle Corporation
subodh.bapat@oracle.com
650-786-8649

Re: Comments by Oracle Corporation on Draft 1 for the Energy Star Specification for Data Center Storage Version 1.0

Dear Mr. Pantano:

Thank you for the opportunity to provide comments on the EPA's Draft 1 for the Energy Star specification for Data Center Storage, Version 1.0. Oracle Corporation appreciates the opportunities extended throughout this past year for inclusion in this process, and we look forward to continuing to help achieve a successful new specification.

We commend the EPA on their careful consideration of the input provided by the industry and on the data collection framework for this specification. The comments that follow are made with the purpose of achieving a specification that better achieves our mutual goals.

We look forward to discussing these points in more detail and, in our role as a program partner in the EPA's Energy Star for Data Center Storage program, helping the EPA to successfully develop the Energy Star for Data Center Storage Version 1.0 specification.

Sincerely,

Subodh Bapat
Oracle Corporation
subodh.bapat@oracle.com
650-786-8649

Jud Cooley
Oracle Corporation
jud.cooley@oracle.com
858-526-9457

Table of Contents

ORACLE CORPORATION COMMENTS ON THE EPA ENERGY STAR FOR DATA CENTER STORAGE VERSION 1.0 SPECIFICATION (DRAFT 1)

1. Introduction

Oracle commends the EPA on issuing Draft 1 of the Energy Star for Data Center Storage specification Version 1.0. Oracle applauds the open process that the EPA has followed, including the extensive dialog with the industry and the EPA's willingness to be available for detailed discussions. Oracle appreciates the opportunity to meet with the EPA in one-on-one meetings and in industry conference calls, as well as EPA's outreach to the industry at various conferences and symposia.

This draft specification takes a step forward in defining the taxonomy of storage devices and clearly articulating the semantics of terminology used in the definition section. However, it is difficult to respond to specific proposed thresholds or numerical parameters because most of the requirements of this draft of the specification are TBD. All of the tables in this draft specification (Table 1 on product family requirements, Table 2 on PSU efficiency criteria, and Table 3 on PSU power factor criteria) only contain TBD requirements. In the absence of any numerical thresholds in this draft of the specification, Oracle recommends that the EPA continue this process with a follow-up draft specification with clearly articulated numerical targets to which stakeholders can respond.

2. Definitions and Scope

Oracle applauds the EPA on aligning its taxonomy of storage products with the Fall 2009 edition of the SNIA (Storage Networking Industry Association) dictionary. We appreciate the flexibility afforded to storage vendors to include multi-component and multi-box systems within the definition of a storage products, as most data center storage products are composed of multiple boxes and multiple enclosures.

A. Breadth of Coverage

Oracle has a particular interest in ensuring that the following categories of storage products continue to be included within the scope of the Energy Star for Data Center Storage specification.

- Traditional disk-based storage architectures (one or more storage controllers plus multiple JBODs)
- Hybrid flash and disk storage architectures (one or more storage controllers with SSDs and/or HDDs plus multiple JBODs with SSDs and/or HDDs
- Storage architectures where the storage controller is a general purpose computer server which has been dedicated to storage control function
- Tape libraries

Oracle appreciates the EPA's intent to cover a broad range of products including all types of storage networking interfaces:

- DAS: Direct Attached Storage
- NAS: Network Attached Storage
- SAN: Storage Area Networks

B. Hybrid Storage

In response to the question asked by the EPA in the note on line 449 of the draft specification, Oracle would like to recommend that the EPA continue to include hybrid storage systems within the scope of the specification. Oracle believes that the technology advantages provided by hybridization (for example, the use of solid state drives as cache extension of storage controllers) is a key driver in the reduction of power consumption by data center storage systems and in the increase of energy efficiency in the data center.

The EPA should note that there is no industry standard definition of hybrid storage. Some industry groups do not recognize mixed media in a storage solution unless all of the mixed media devices are directly addressable by the host. For example, if SSD devices are used as a front-cache for HDD devices that is transparent to the host, the solution is not regarded as a hybrid because the SSDs are not addressable directly by the host. In other situations, if all the mixed media devices present the same interface to the host at the I/O interface level and at the file system API level, then they are not regarded as hybrid because the mixed media merely adds additional capacity to the system (perhaps with a different response time).

Newer drive technologies incorporate both flash memory and hard drives directly in the drive itself, thereby driving hybridization into the component level, not just at the system level. As such, Oracle recommends the EPA come up with a clear definition of hybrid storage systems in case elements of this specification will apply differently to hybrid systems than to homogeneous storage devices.

C. Suggested Scope

In terms of the qualifying products that will be included int the scope of this specification, Oracle would like to recommend that in Version 1, the scope be limited to entry to mid-range network storage systems. This is because we believe that experience is needed in applying the specification to smaller systems before it can be extended to include large enterprise class storage systems. Specifically, we recommend that the scope for this specification be limited to the following categories as specified in the SNIA taxonomy of storage systems:

- Online Group 2
- Online Group 3
- Near-online Group 2

- Near-online Group 3
- Removable Group 2
- Removable Group 3

Because of the differences in interpretation in the SNIA categorization that exist in the industry, Oracle would like to encourage the EPA to lock-down its intended coverage categories with greater clarity. For example, the difference between the Online category and the Near-online category is defined as a MaxTTD of <80ms. There are differences in how this is interpreted:

- Some customers interpret this as meaning every request for data in an online system must be serviced within 80ms; even if a single request out of 10,000 takes longer than 80ms, than that system is no longer considered online.
- Some customers interpret this as meaning that the average response time must be within 80ms; some requests for data may take longer, but the system is still considered an online system if the average remains under 80ms.

Because this confusion arose during the EPA's data collection exercise for data center storage, we would like to request the EPA to state its definition of online, near-online, and removable storage with greater precision. In the SNIA dictionary, MaxTTD is defined as the time taken to start streaming the first accessed byte, and is expressed as a property of the storage component, not the system. Because there is no definition of MaxTTD at the system level, it is difficult to calculate the MaxTTD of a hybrid system where the different components (e.g. SSDs and HDDs) have different component-level MaxTTD numbers. We suggest that the EPA work with SNIA to come up with a clearer definition of MaxTTD at the system level – for example, a system may be considered an Online system if 99% of all data accesses are served within 80ms.

D. Suggested Exclusions from Scope

Oracle recommends that until more experience is gained with applying the specification to storage systems, the following categories of storage products from the SNIA taxonomy be excluded from the scope of the specification:

- Virtual Media Libraries
- Online Group 4
- Near-online Group 4
- Removable Group 4

3. Standard Performance Data Measurement and Output Requirements

In response to requirements from storage administrators, most storage vendors have designed their storage systems to report parameters of interest related to performance, capacity, online status, availability, and component failures. These parameters today are considered the industry norm in terms of the real time data that is required to be reported from storage systems.

A. Reporting Capabilities of Current Disk Storage Systems

The EPA has requested that storage systems also report power consumption and temperature measurements in realtime. Most storage systems on the market today were not designed with real-time data measurement and output requirements for power draw and temperature. Unlike server systems, storage systems are generally not designed with sensors that report power consumption, temperature, and other parameters. This is because on the average the inherent power consumption of storage systems is lower than that of servers. Further, historically storage admins have been less concerned with power consumption than server admins. Storage admins have typically focused on performance, capacity, availability, and other indicators of the health of the storage system.

While Oracle agrees with the need to understand the energy consumption of a storage system to enable data center efficiency, Oracle would like to suggest to the EPA that this requirement should be phased in over time. If this requirement were in effect today, almost none of the shipping storage systems in the industry would qualify. Introducing the appropriate sensors in the engineering design of a storage system would require at least one complete design cycle (approximately 18-24 months). As such, Oracle recommends:

- Deferring this requirement until the version 2.0 specification of Energy Star for Data Center Storage, or
- Deferring the effective date of this requirement to some point later than the effective date of the rest of the specification, to allow a full product design cycle that can satisfy this requirement.

B. Reporting Capabilities of Current Tape Libraries

One class of storage systems where this requirement is especially onerous is tape libraries. Most tape libraries deployed in the industry today have a very long lifecycle and are not refreshed as frequently as other IT equipment such as servers. It is common for customers using tape libraries for archival purposes to keep them in operation for 10 – 15 years. As such, the development cycles and refresh cycles for these devices is very long and it will be a long time before a new generation of equipment with temperature and power reporting capabilities is introduced in the market. Deferring this requirement to a later version of the specification would allow tape library manufacturers adequate time to engineer their devices appropriately.

When this requirement is eventually introduced, Oracle recommends that additional clarity be provided on exactly when and where data should be measured. Specifically, clarification should be provided on temperature measurement and sampling requirements.

C. Temperature Measurement

Guidance should be provided as to exactly where the temperature is measured. In a multi-enclosure storage system (for example, where the controller is in one rack and the JBODs that it controls are in a different rack) temperature can be measured at multiple places. Guidance should be provided as to whether the inlet air temperature must be measured at the inlet of the controller unit, the inlet of the JBOD devices, or both. Further, where there are multiple JBODs deployed in multiple racks, guidance should be provided as to whether the inlet air temperature needs to be reported at the inlet of every JBOD enclosure, or at the top of the JBOD rack, or at the bottom of the JBOD rack, or some representative location in between.

Additional guidance is needed on where temperature should be measured in a tape library. Some tape libraries are large because they contain many tape drives, large bays for tape cartridge storage, and multiple robotic arms to mount the tape. There is a wide range of temperature variation inside such an enclosure depending on where the measurements are taken. Temperatures near the tape cartridge storage bays can be close to the data center ambient temperature, while temperatures on the tape drive motors can be very high. Reporting a single temperature reading from such a device can be misleading and might not convey meaningful information to the data center operator for the purposes of provisioning cooling capacity.

D. Sampling Requirements

Oracle strongly disagrees with the EPA on the sampling requirements expressed in line 603 of the Draft Specification. The proposed requirements require sampling at a frequency of one measurement per second for power draw, and one measurement every ten seconds for inlet air temperature. Oracle believes that this frequency of sampling is unnecessarily high and is not necessary for any practical data center application that makes use of this information.

Example applications that may use dynamic information about power draw and air temperature of storage systems include the following:

- Provisioning power distribution and UPS capacity in the data center
- Provisioning air flow distribution and cooling capacity in the data center
- Power based charge-back billing to hosted tenants, cloud service subscribers, or internally hosted business units.
- Thin provisioning of storage capacity in the data center which enables incremental provisioning of capacity as and when data storage needs grow over time. The applications that provision incremental storage capacity on an as-needed, just-in-time basis may also need to communicate to the data center power and cooling systems to provision power and cooling capacity as and when storage capacity is brought online.

There are several other applications similar to the above that can use dynamic information on storage power and temperature. None of these applications, however, require this information to be reported at a frequency of once per second.

The reaction time for taking action for any of these applications is at least one order of magnitude greater than the EPA requested frequency of once per second. For example, the time duration over which charge-back billing is calculated is of the order of minutes or hours, not at a sub-second frequency.

Oracle requests that the sample frequency for dynamic power and temperature information be recalibrated to the needs of the applications that will use this data. Oracle recommends a sampling frequency of between ten to thirty seconds for power draw information, and between five to fifteen minutes for temperature information.

4. PSU Efficiency and Power Factor Criteria

With regard to the EPA's request on line 485 for comments from stakeholders on the applicability of the efficiency and power factor requirements already established by the EPA for computer servers in the Energy Star specification for computer servers Version 1.0 to storage systems, Oracle recommends that the EPA not use the specification for server PSUs to storage devices.

A. Storage PSUs Versus Server PSUs

There are multiple reasons why storage PSUs need to be different from server PSUs:

- Storage products have a longer life-time than server products and correspondingly the refresh cycles for storage PSUs are longer

- Storage products have their components and devices upgraded in the field for many years throughout their service life. For example, a storage product used in a data center for eight years may have multiple generations of disk drives that are housed inside its enclosure. While it may have lower speed disk drives installed at initial deployment, over the years, as disk drives fail, or as additional capacity is needed, components with higher rotational speeds or greater capacity may be installed in the enclosure. In order to accommodate the higher power demand of the new disk drives, the power supplies installed in the enclosure need to be sized for the maximum power demand that the storage system will encounter during its service life. As such, at the time of initial deployment the PSUs are necessarily oversized, and over the course of the service life of the system may become right-sized. Oracle recommends that storage systems shipped with PSUs that have capacities greater than needed for initial deployment not be penalized from an Energy Star eligibility perspective, since this additional power delivery capacity is needed for component upgradability throughout the service life of the system.

Because of the wide variability of storage systems, there are many different solutions for power delivery. The diversity of power supplies used in the storage industry is greater than that used for

servers. For example, the power supplies used in entry level DAS systems versus mid-range multi-enclosure NAS solutions versus high-end tape libraries are very different. As such, it would not be appropriate to use a single specification for PSUs across both the server and storage Energy Star specification.

In addition, Oracle encourages the EPA to conduct a data collection exercise on PSU efficiency and power factor data for power supplies in storage systems that are shipping in the industry today. This will provide some insight into the wide variability of PSU solutions in the storage industry, and provide the basis for drawing the eligibility cutoff limits for Energy Star qualification. Using the existing specification for PSU efficiency from the Energy Star for Computer Servers specification for storage systems has the potential of limiting Energy Star eligibility to a very small fraction of the unit volume of storage systems shipping in the industry today.

B. Power Supply Clarifications Requested

Specific questions from Oracle about power supply specifications are documented in the following table of issues:

Line Number in EPA Draft 1	Question or Required Clarification
272-273	Additional clarification is required on how a test lab may handle differences in the output waveform of a UPS. The different input waveforms to the server power supply may have an impact on the input power measurement.
481	Oracle requests the EPA to provide guidance on the circumstances under which the power consumed by the fans inside the PSUs may or may not be considered to be included in the overall PSU power draw, as the EPA has done for the Energy Star specification for computer servers.
486	In response to the request for comments from the EPA on the usefulness of the 10%, 20%, 50%, and 100% load points for storage PSU efficiency requirements, Oracle recommends that the efficiency requirement at the 10% load point be eliminated. This is because the existing practice in the storage industry is to specify PSU efficiency requirements to power supply vendors using primarily the 20%, 50%, and 100% load points.
503	In response to the request for comments from the EPA on the usefulness of the 10%, 20%, 50%, and 100% load points for storage PSU power factor requirements, Oracle recommends that the power factor requirement at the 10% load point be eliminated. This is because the existing practice in the storage industry is to specify PSU power factor requirements to power supply vendors using primarily the 20%, 50%, and 100% load points.

Line Number in EPA Draft 1	Question or Required Clarification
528-537	The EPA should clarify that the requirement to disclose all power management features that are enabled by default does not apply to any power management features operating in the power supply. This clarification is required because power management features operating in the power supply can not be controlled by the system. Any such features to improve the power supply efficiency are the intellectual property of the power supply manufacturer, not the system vendor.
603-606	The EPA should recognize that because of inherent limitations in the technology, the readings of input power retrieved from power supplies every one-second or less have no guarantee of being accurate. Power supplies may not be designed to sample their input every second, and so any reading requested by the system at one second intervals may still be reflective of an older reading taken by the power supply of its input waveform more than one second ago. The reading of the input power supply may represent some averaging over the input samples, however this is not necessarily a rolling average over a fixed period.

5. Idle and Other Operational States

A. Active State

Oracle has multiple observations and comments to provide about power measurement in the active state. This is based on Oracle's multiple years of experience in benchmarking storage systems and on the recent data collection exercise that Oracle conducted in response to the EPA's data collection requests for enterprise storage.

1. ***Controller to JBOD ratio***: In the data collection exercise, the EPA asked that the power consumed by the controller be added to the power consumed by the JBOD. To calculate the energy efficiency, the EPA asked that the throughput number (for the appropriate component of the test workload) be divided by the total power.

   However, the EPA has not indicated how the controller power component will be treated. In many test configurations, a small ratio of JBODs to controllers is used (for example, one controller with one JBOD or one controller with two JBODs). However, in real-world customer deployments the ratio can be different (for example, one controller with eight JBODs or two controllers with sixteen JBODs). These configurations will seem inherently more efficient than the test lab configuration because the power consumed by the controller is amortized over a larger number of JBODs, thereby increasing the throughput per Watt ratio as compared with the test lab configuration.

Oracle encourages the EPA to calculate energy efficiency in its data analysis using some normalized configuration that is representative of a typical customer deployment. While it is onerous to ask partners to conduct data collection in a test lab for a large configuration (since many test labs do not have the equipment to simulate a full customer data center), the EPA should work out a methodology to normalize submitted data to typical customer configuration (e.g. a ratio of one controller to eight JBOD's) so that the power draw of the controller is equivalently normalized across the same number of JBOD's for all vendor submitted data.

2. ***Nature of test workload:*** The composite test workloads suggested by the EPA are not the best workloads that are suited to showcase the advantages of a hybrid system, such as one that uses a combination of SSDs and HDDs. These typical industry composite test workloads tend to test a very wide variety of storage access patterns. These workloads are fine to test the different ways in which one might use an HDD based system, but won't necessarily showcase the advantages of a hybrid system.

   Certain architectures of hybrid systems are designed to work best with long-duration transactional business workloads (e.g. TPC-variants). In real business scenarios, large databases which have transactions that have periodic locality of reference will do really well on hybrid systems. SSD caches warmed up over longer durations with transactions that are frequently referenced for some duration will deliver very good power/performance. However, the EPA test workload is not of this nature.

   For example, if the SSDs are used as an optimizing read cache for data that is frequently accessed due to clustered locality of access in an application, they will prove to be an advantage under test workloads that simulate such clustered locality of frequent access to the same subset of the dataset. However, for test workloads that only test for random reads or sequential reads, the SSDs will show no advantage. Similarly, for test workloads that test for random writes or sequential writes, the SSDs will show no advantage, and in fact, may be a detriment.

   The situation is different for tape libraries. The power draw of a tape library is less sensitive to the nature of the workload than is the case for HDD or SSD based systems. A typical tape library would exhibit its worst case power draw when all tape drives are streaming data to tape. Since the movements of robotics has little effect on the overall power draw, the worst case active state power draw for a library can be measured with any workload that causes all tape drives to be activated simultaneously. As such, the EPA specified test workload is appropriate for testing the power draw of tape libraries.

   Oracle recommends to the EPA that a different set of test workloads be specified for measuring the power of HDD/SSD based systems versus tape systems.

3. ***Inter-workload pause duration:*** Even for test workloads that test for data access patterns under which the SSDs might confer an advantage, the inter-workload pause duration of 60 seconds is not sufficient to precondition the SSD caches before the new workload starts. SSDs take a long time to warm up, and with only 60 second preconditioning will not show any performance improvement over a pure HDD-only system.

4. ***Technology ratios:*** The EPA should bear in mind that the nature of the test workload needs to evolve as the architecture of storage systems evolves. We expect that as the prices of flash memory decline over time, the ratio of flash memory used in storage systems as compared to hard drives will increase. For example, it is not unlikely that a typical ratio of DRAM : flash : disk in storage architecture might evolve to 1 : 10 : 100 as flash prices decline to a stable point. In such a storage architecture, where DRAM is used as a front cache for the flash memory, and flash is used as a front cache for the disk, the type of customer workloads that will perform best will likely not be similar to the test workloads the EPA has requested in the current data collection exercise. Oracle suggests that the EPA undertake a study of how test workloads need to evolve in order to correctly test the benefits of the architecture of the storage systems as it evolves.

For tape libraries, the active state power consumption depends on the nature of the activity in the tape library. Tape libraries are either writing/reading, rewinding or unloading, moving a cartridge or idle. Oracle's internal analysis of the power consumed inside a tape library shows that by far the single largest consumer of power is the set of tape drives. The other components of a tape library, such as the robot bus, the robot mechanicals, card cages, tray controllers, cooling fans, overhead lights, networking ports, operator panel, and PDU and line power losses, consume a lot less power than the tape drives themselves. Oracle will be pleased to provide power consumption analysis data on its tape drive products privately to the EPA under non-disclosure.

B. Idle State

The notion of defining and idle state for a disk based storage system is problematic. Unlike server, which can be truly idle when there are no applications running and no user input or network input is being received, disk based storage systems always have background applications running even when there is no user I/O going on. As such, measuring the idle power of a storage system would be a complex, if not impossible, exercise.

The following are examples of background activities that are automatically and autonomously triggered in a storage system when there is no user I/O active:

- Backup
- Data compression
- De-duplication
- File migration
- Volume reconstruction

All of the above activities are useful, important and critical to the health of the storage system. They are best done when no user I/O is active. All of the above activities cause power to be consumed in the storage system due to the spinning of disk drives, the spinning of fans, the execution of logic in the control software, etc. Even though they consume power, all of the above activities are intended to make the storage system more efficient through better utilization of capacity or more reliable due to

replication of data. As such, these activities actually reduce the power consumption at the data center level in the long term.

Oracle recommends that the EPA not focus on idle power as a qualification criteria for Energy Star for Data Center Storage systems. Because a true idle condition in a storage system is very rare, measuring idle power is a difficult exercise. Further, an idle power number will not convey any meaningful information to the customer since such a true idle condition will occur very infrequently in a customer data center. Since the background activities that happen in a storage system when there is no user I/O active have the ultimate effect of making the customer's storage network more efficient, the power consumed by these background activities should not be considered to be a penalty for Energy Star qualification.

In section 3.4 line 521 of the draft specification, the EPA has asked whether idle state energy efficiency can be a suitable proxy for active state energy efficiency. Oracle believes that it will be difficult to establish a correlation showing that idle state power draw is a reasonable proxy for active state power draw, or that active state power draw can be a proxy for idle state power draw. This is because in the active state, power draw is primarily due to disk rotation caused by user IOPS, while in the idle state, power draw is primarily due to disk rotation caused by the above mentioned background house keeping activities. The nature of the disk rotation in the active state under user IOPS depends primarily on the nature of the access pattern (e.g. read versus write, sequential versus random, etc.) whereas the nature of the disk rotation in the idle state depends on the extent of housekeeping activities needed (e.g. amount of duplicate data that needs to be subject to deduplication, etc.). As such, Oracle believes that little or no correlation can be established between idle state power draw and active state power draw.

For tape libraries, idle power is important because the duration of activity for tape drive is typically fairly small if the tape library is primarily used for archival purposes. For archival activity, the typical archival window is only a few hours a day, with the tape library being idle the rest of the time. Establishing an idle power specification for tape libraries would encourage the development of libraries which conserve power when there is no archival activity taking place.

Under other circumstances, a tape library might be subject to a usage pattern that keeps it busy nearly continuously. For applications that require constant retrieval of data from archival sources (e.g. during a company audit), a tape library can be busy 24 hours a day. For this application pattern, an idle power specification may not be particularly useful or necessary.

6. Effective Date

In the Energy Policy Act of 2005 (Section 131 of PL109.58), which amends the Energy Policy and Conservation Act (42 USC 6294a. Sec. 324), the duties of the EPA administrator with respect to the Energy Star program are specified as follows:

> "The Administrator and the Secretary shall provide appropriate lead time (which shall be 270 days, unless the Agency or Department specifies otherwise) prior to the applicable effective date for a new or a significant revision to a product category, specification, or criterion, taking into account the timing requirements of the manufacturing, product marketing, and distribution

process for the specific product addressed."

Since the Energy Star for Data Center Storage specification is a new specification, and data center storage vendors need the appropriate lead time for manufacturing, marketing, and distributing products compliant with this new specification, we recommend that the EPA not deviate from the 270 day lead time notification suggested in the legislation above. Oracle would further like to request that the EPA allow a duration of eighteen months between the issuance of the final specification and its effective date. This will allow manufacturers a full design cycle to design storage products that are conformant with the specification.

As stated earlier, most currently shipping storage systems do not have the necessary sensors to properly report temperature and power information required by the specification; as such, a full product cycle for design, prototyping, testing and verification of storage products conformant with the specification will increase the number of compliant products and the overall adoption, uptake and success of the Energy Star for Data Center Storage program.

For tape libraries, the product design and refresh cycles are even longer than for disk based storage systems. As such, customers do not plan for quick turns of their tape libraries once deployed in the field. While a library will be constantly upgraded during its deployment lifecycle, most of these upgrades will be in the nature of newer and higher speed tape drives that will be put in them. As such, the effective date for the Energy Star specification for tape libraries should be extended out even further than the effective date for disk based systems.