May 21, 2010

TO: Evan Haines
ICF International
1725 Eye Street, NW, Suite 1000
Washington, D.C. 20006
ehaines@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 Computer Servers Version 2.0

Dear Mr. Haines:

Thank you for the opportunity to provide comments on the EPA's Draft 1 for the Energy Star specification for Computer Servers, Version 2.0. 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 preliminary 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 Servers program, helping the EPA to successfully develop the Energy Star for Servers Version 2.0 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 SERVERS VERSION 2.0 SPECIFICATION (DRAFT 1)

1. Introduction

Oracle commends the EPA on issuing the First Draft of the Energy Star for Servers Version 2.0 specification. 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.

Oracle's specific comments concern the following topics covered in the specification:

- Idle Power Allowances and Target Recommendations
- Standard Performance Data Measurement and Output Requirements
- Family Definition
- PSU Efficiency and Power Factor Criteria
- Blades
- Active Mode Efficiency Criteria
- Power / Performance Data Sheet and QPI Form
- Effective Date

2. Idle Power Allowances and Target Recommendations

Sun applauds the EPA's willingness to open the idle power discussion for Version 2.0 of Energy Star for Servers.

In Version 1.0, the EPA emphasized idle power as a criterion for qualification for Energy Star based on survey data from 2006-2007 that showed that the majority of servers in the majority of data centers spent a majority of their time at low utilization. Further, servers running at low utilization still consumed almost as much power as they did at high utilization. As such, the EPA rightfully concluded that the optimization of power consumption at idle would have a beneficial effect on total data center power consumption.

Since this survey was conducted, both technology and industry practice have changed. There is now widespread prevalence of virtualization, and with it, an increasing trend towards higher utilization of servers. Although the worldwide installed base of servers has grown since 2007, total data center power consumption has not grown at the same rate, due to more application of techniques to drive servers to higher utilization.

We would strongly encourage the EPA to conduct a new survey to determine the extent to which there is still a problem of servers running at low utilization. Such a data-driven approach, based on a large sample of average server utilization readings across a large number of data centers, would yield new insights into trends towards virtualization and higher server utilization. This may help the EPA to arrive at an approach that balances the need to optimize server idle power with the need to optimize server efficiency at high throughput.

A. Drawbacks of Static Configuration Based Idle Power Allowances

The Version 1.0 approach of granting each server an idle power allowance based solely on the static analysis of the components in its configuration has some benefits but also has significant drawbacks. In particular, the specific drawbacks of using a configuration based approach to determine idle power allowances include the following:

- There is no consideration of the energy efficiency of the system at high throughput, and consequently no scaling of the idle power allowance to the peak power consumption of the system when it is performing most efficiently.

- The specific allowances granted in Version 1.0 for baseline configurations, and the specific adders granted for additional components in higher configurations get outdated very quickly as technology changes. For example, as capacities of memory DIMMs and rotational speeds of disk drives increase, their power draw at idle also increases. Any baseline plus adders approach to granting idle power allowances will have to be frequently updated in order to keep up with evolving technology or risk becoming obsolete very quickly.

- A static configuration approach for determining idle power allowances can be subject to gaming of configurations solely for the purposes of qualifying for Energy Star. Vendors maybe incented under this approach to create particular configurations designed only to pass Energy Star, because the specific pattern of baseline plus adders for this configuration just happens to put it over the top for its Energy Star allowance. Vendors are incented to do this even though such configurations may not be balanced configurations for running typical customer workloads, or may not be popular configurations that are commonly ordered by customers.

B. Alternate Approaches to Idle Power

Oracle strongly encourages the EPA to consider alternative approaches to idle power. In particular, Oracle recommends an approach that grants each server an idle power allowance based on:

- Its power consumption at peak throughput, and/or

- Its rated nameplate power consumption.

Such an approach has the merit that a server designed for efficiency at high throughput is also granted an idle power allowance that is proportional to its throughput at peak utilization. For example, a high end server with many components (large number of CPUs, memory DIMMs, disk drives, and I/O devices) can deliver significantly higher throughput on typical customer workloads, but will also burn a commensurately higher amount of power when idle.

Granting an idle power allowance based on some proportion to a server's peak power has the following advantages:

- It will permit servers that are designed for efficiency at high throughput to potentially qualify for Energy Star. Data Center operators are increasingly turning towards virtualized architectures which drive towards fuller utilization of active assets and better overall efficiency. This trend should be accounted for and encouraged as a long term strategy to achieve the overall objectives of the EPA

- It will automatically scale as technology changes, because no allowance expressed in absolute Watts for any specific baseline configuration or any specific adders for extra components, will need to be updated over time.

- It will automatically scale across server sizes, from single-socket low-configuration servers to four-socket, high-configuration servers, in a fair and balanced proportion.

- It is simpler to measure and report and will result in higher accuracy and more timely information.

C. Industry Idle Power Analysis

Oracle has conducted an extensive study of industry servers which analyzes their idle power consumption as a proportion of its peak power consumption and its rated nameplate power. This study analyzed both Energy Star compliant servers as well as non-compliant servers.

The results of this study are summarized below:

		Average for Single PSU Systems	Average for Multi PSU Systems	Average for Whole Category	25th %ile for Whole Category
All Servers	Idle as % of Max	47.58% *15 Data Points*	58.02% *15 Data Points*	52.8% *30 Data Points*	46.97% *30 Data Points*
	Idle as % of Name Plate	21.56% *43 Data Points*	18.07% *56 Data Points*	19.59% *99 Data Points*	14.77% *99 Data Points*
E* Servers	Idle as % of Max	59.54% *4 Data Points*	52.07% *4 Data Points*	55.81% *8 Data Points*	50.11% *8 Data Points*
	Idle as % of Name Plate	13.23% *9 Data Points*	11.41% *11 Data Points*	12.23% *20 Data Points*	9.16% *20 Data Points*

Oracle's study was conducted with early data submitted to the Energy Star program, as well as other known data publicly available in the industry. Since Oracle's study was completed, more data has been submitted to the Energy Star program as more products have qualified for Energy Star. Oracle encourages the EPA to conduct its own study with current data along the same lines as above. Oracle is confident that such a study will yield a statistically significant correlation between the idle power of a server and its peak power and nameplate power.

Based on Oracle's findings, Oracle would like to encourage the EPA to allocate server idle power allowances using a proportional guideline similar to the following:

- The idle power of a server that qualifies for Energy Star shall be no greater than 51% of its peak power consumption
- The idle power of a server that qualifies for Energy Star shall be no greater than 10% of its nameplate power.

While the numerical targets in the above guidelines are for example only, Oracle encourages the EPA to establish its own specific numerical targets similar to the above after it completes its analysis of the most recent available data.

D. Rationale for Idle as Percentage of Peak Power

In line 687 of the Draft Specification, the EPA has expressed concern that tying the idle power solely to top-level performance could lead to a systematic increase in idle power consumption over time and dissuade manufacturers from improving efficiency at low levels of utilization. Oracle believes that this concern is misplaced. The reason is that the top-level performance, and consequently the peak power draw of a server, is a naturally self-limiting parameter. The reason the peak power draw of a server can not scale indefinitely to arbitrarily large numbers is because of the following limitations:

- Server vendors always have specific price targets and cost constraints within which they need to deliver a server product to the market. These pricing pressures act as natural inhibitors to creating complex configurations with a large number of components. For the scope of the present Energy Star for Servers specification (1-4 socket servers), the price bands of qualifying servers are naturally limited by existing pricing expectations in the market for this category of servers. This implies that arbitrarily complex configurations will not be brought to market in this product range, and hence the peak power draw of these servers will remain limited to the range that the market expects these servers to consume.
- The amount of power that can be provisioned to a server rack is limited by the size of the wire and electrical code restrictions. There is a high cost associated with delivering significantly higher power levels to a rack which tends to place a natural upper limit on the amount of power that can cost-effectively be allocated to a 1-4 socket server.
- Technology improvements that can cause the higher power draw are counterbalanced by complementary improvements that reduce the power draw. For example, for CPUs and ASICs as process technology improves and allows for higher frequencies which cause an

increase in power draw, it also allows for lower voltages which cause a decrease in power draw. This reduces the likelihood of the power draw of these components from scaling indefinitely as technology improves.

3. Standard Performance Data Measurement and Output Requirements

Oracle agrees with the EPA on the need to provide real-time dynamic information on the energy performance of a server to customers. In particular, the reporting of power draw, inlet air temperature, and processor utilization remains important.

However, 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 reporting externally to any practical data center application that makes use of this information.

This sampling rate is not supported by any server on the market today. Implementing such a reporting standard would add cost to every server for a feature that customers do not consider important. The inclusion of theses sampling rates in servers would require 12-18 months to implement for most server manufacturers.

Furthermore, this increased sampling rate would add significantly to the bandwidth and data processing requirements of a data center, because new networking infrastructure and server infrastructure with database software would have to be added to the data center simply in order to receive, process and store this volume of data. For example, a large data center with 50,000 servers, each reporting 100 bytes of power information every second, would generate 432GB of new data every day. The infrastructure needed for transporting, processing and storing this monitoring information would itself consume power in the data center.

Example applications that use dynamic information about power draw, air temperature and processor utilization 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.
- Power aware VM migration to enable shut down of underutilized servers during periods of low utilization of the data center.

There are several other applications similar to the above that use dynamic information on server power, temperature, and utilization. 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 horizon for taking action on the provisioning or reprovisioning of power capacity and cooling capacity is of the order of days or weeks based on observed trends of power and cooling requirements calculated over weeks and months. Power and processor utilization information used for virtual machine migration provides decision support for migration decisions that are taken over tens of minutes, if not hours. Power draw information used for charge-back billing is typically sampled over minutes or hours, not at a sub-second frequency.

Oracle requests that the reporting frequency for dynamic power, temperature, and utilization information be recalibrated to the needs of the applications that will use this data. Oracle recommends a reporting frequency of no greater than one reading every thirty seconds for power draw and processor utilization information, and no greater than one reading every fifteen minutes for temperature information.

The EPA should make a clear distinction between the internal sampling frequency inside the server versus the external reporting frequency to applications outside the server. It may be appropriate to require a higher sampling frequency internally (for example, the rate at which the system interrogates its power supplies) than the frequency at which this information, or a rolling average derived from it, needs to be reporting externally.

4. Family Definition

It is the experience of not just Oracle, but all other industry vendors of servers, that the mechanisms that allow the qualification of an entire family in Energy Star for Computer Servers Version 1.0 are improperly defined.

A. Drawbacks of Existing Family Definitions

The current mechanism to qualify families is so restrictive that the number of configuration variations permitted inside each family is extraordinarily small. As such, vendors are required to separate out even minor configuration variations for the same server model into separate family definitions. This causes a combinatorial explosion in the number of configurations that must be independently tested under separate family definitions, causing unnecessary time and costs for the Energy Star partner.

In addition, the need to separate out different configuration variations of the same server into different families causes unnecessary paperwork to be generated, because a different Qualified Product Information form and a different Power Performance Data Sheet needs to be created for each family. The resulting explosion of the number of families is confusing to the marketplace because the EPA family definitions do not correlate with the vendor's family definitions. This approach:

- Raises the cost of Energy Star qualification for the Energy Star partner
- Raises the cost of Energy Star submission, review and approval for the EPA

- Provides no useful incremental information to the customer

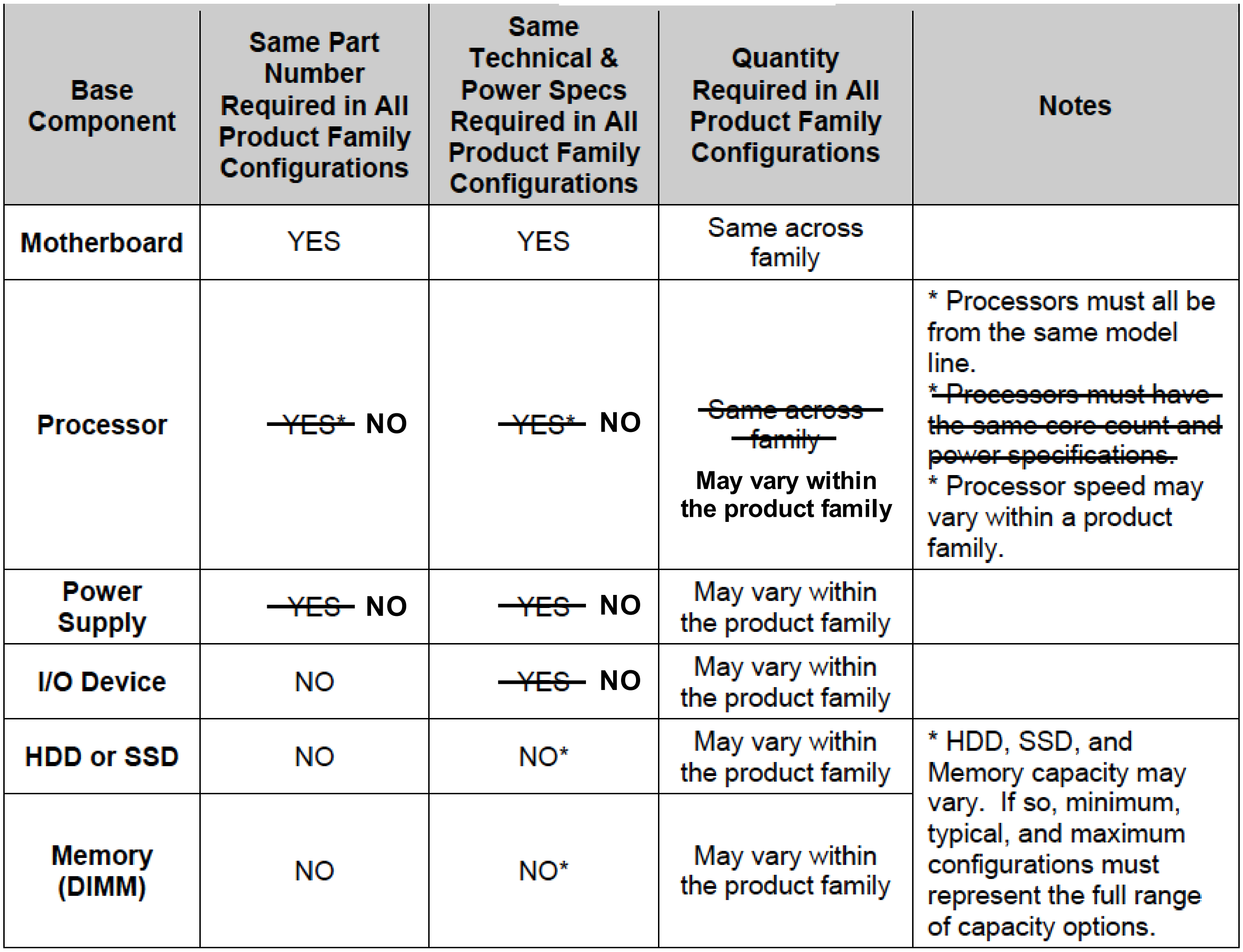
Because of the onerous paperwork necessary to qualify families, server vendors have taken the approach of not submitting all possible server configuration variations to the EPA for approval. Instead, they only submit a few sample representative configurations. Ultimately, this approach has the effect of inhibiting the industry acceptance, customer value, and overall success of the Energy Star for Servers program.

Oracle applauds the EPA's intention to broaden the definition of server product families in Version 2.0 of the specification. Oracle appreciates the additional flexibility to permit variability of configuration for I/O devices, disk drives, and memory DIMMs within a family definition.

However, Oracle does not feel that the EPA has broadened the family definition to be consistent with the way customers understand server families. Further, even under the new family definition, the number of families that will need to be created for each model of server remains exceedingly large. As such, the paperwork required to be submitted for qualifying variations of the same server model under different family definitions will still be excessively burdensome. Therefore, the new family definitions in Energy Star for Servers Version 2.0 Draft 1 do not achieve the EPA's goal of greater industry acceptance and greater customer value deriving from the Energy Star for Computer Servers program.

B. Alternate Approaches to Family Definition

Oracle recommends that the product family requirements proposed by the EPA in Table 1, line 415 of the draft specification be modified as follows:



Base Component	Same Part Number Required in All Product Family Configurations	Same Technical & Power Specs Required in All Product Family Configurations	Quantity Required in All Product Family Configurations	Notes
Motherboard	YES	YES	Same across family	
Processor	~~YES*~~ **NO**	~~YES*~~ **NO**	~~Same across family~~ **May vary within the product family**	* Processors must all be from the same model line. ~~* Processors must have the same core count and power specifications.~~ * Processor speed may vary within a product family.
Power Supply	~~YES~~ **NO**	~~YES~~ **NO**	May vary within the product family	
I/O Device	NO	~~YES~~ **NO**	May vary within the product family	
HDD or SSD	NO	NO*	May vary within the product family	* HDD, SSD, and Memory capacity may vary. If so, minimum, typical, and maximum configurations must represent the full range of capacity options.
Memory (DIMM)	NO	NO*	May vary within the product family	

We would encourage the EPA to draw its family definition based on the quantification of the number of families created under different approaches. It would be instructive to do a study that, for a given fixed server model that is available under multiple configurations, compares the following:

- The number of families generated under the Energy Star for Servers Version 1.0 family structure
- The number of families generated under the Energy Star for Servers Version 2.0 family structure proposed by the EPA in the Draft 1 specification
- The number of families generated under the modified guidelines proposed by Oracle in the table above.

We believe that the data will show that modifying the family structure as proposed by Oracle above will yield demonstrable benefits in considerably less paperwork, resulting in a lower qualification burden for vendors, lower approval and evaluation costs for the EPA, and more meaningful information for customers.

C. Rationale for Expanded Family Definition

The rationale behind Oracle's recommendations to create a reasonable and practical set of family definitions is as follows:

- ***Processor Variations:*** Processor variations within the same processor model line generally vary with frequency and core count. This causes some changes to the power specification of the processor variant. However, the changes to the power specification are relatively minor. When this incremental change in processor power is factored in to the total power draw of the whole system, it creates very minor differences to the overall power draw of the server. These minor differences will likely not fundamentally affect the eligibility of the server for Energy Star qualification. As such, it should be permissible for a system vendor to include processors that vary in frequency and core count within the same family definition. The minor differences in power draw for processor variations within the same model line do not justify the need for the paperwork and the costs for a whole separate family definition.

- ***Depopulated Configurations:*** Frequently, a server model is made available in depopulated configurations. These depopulated configurations do cause a difference in the overall power draw of a server. However, in many cases, both the depopulated variant of a server, and the fully populated variant of the same server, qualify for Energy Star. For example, a two socket system may be sold in its fully populated configuration with two processors and 96GB of DRAM, and in a depopulated configuration with one installed processor and 48GB of DRAM. It is possible that both configurations qualify for Energy Star. In this case, it should be permissible for a vendor to qualify the depopulated version as the "minimum configuration" of the family and the fully populated version as the "maximum configuration" of the family. By bookending minimum and maximum configurations with server variations that include a variation in processor count and still qualify for Energy Star, customers can be assured that in-between configurations will also qualify. Hence, we recommend that the family definition be broadened to include variations in the processor count in situations where depopulated variants and fully populated variants of the same server model would otherwise independently qualify for Energy Star on their own.

- ***I/O Devices:*** I/O devices in the same server line can vary widely. Different I/O devices have different technical and power specifications. However, the differences between the power specification of different I/O devices such as add-in cards that go in to open PCIe slots are relatively minor. When these incremental changes in I/O device power are factored in to the total power draw of the whole system, it creates very minor differences to the overall power draw of the server. These minor differences will likely not fundamentally affect the eligibility of the server for Energy Star qualification. As such, it should be permissible for a system vendor to include I/O devices that vary in technical and power specifications within the same family definition, as long as both the highest powered and the lowest powered I/O devices can be demonstrated to meet Energy Star requirements at the system level. The minor differences in power draw for I/O device variations within the same model line do not justify the need for the paperwork and the costs for a whole separate family definition.

- ***PSU Upgrades:*** During the shipping life of a server model, the PSU model that is included in that server line is occasionally upgraded. PSU upgrades for a shipping server model happen because the PSU supplier may have made the original model of PSU obsolete. As long as the original and the upgraded PSUs meet the Energy Star eligibility criteria for computer server power supplies, it should be permissible to include these PSU variations within a single family definition. We therefore request the EPA to relax the requirement that (a) the same part number of PSU be required in all server configurations within a product family (because the server model is shipped with both the original PSU type and the new PSU type), and (b) the same technical and power specifications for PSUs be required within a product family (for the same reason).

In lines 438-443, the EPA has redefined the max and the min configuration in terms of active mode efficiency. Oracle suggests that this line of reasoning be pursued with extreme caution, because it may have the unintended consequence of not aligning with commonly accepted customer definitions of a product family. For example, if the highest throughput per Watt is delivered by a low end configuration in the family, it would have to be defined as the max configuration for Energy Star purposes. However, customers may continue to define the maximum configuration of a server to be the one with the fastest processors, maximum memory, or greatest disk capacity, even though these configurations may not have the highest throughput per Watt.

5. PSU Efficiency and Power Factor Criteria

Oracle applauds the EPA's notion of converging the top level energy efficiency criteria for PSUs across all servers (pedestal, rack-mount, and blade) and across all voltage ranges. Oracle commends the EPA for its work with the Climate Savers Computing Initiative consortium and for aligning all PSU efficiency requirements with the Climate Savers Gold specification.

Oracle agrees with the EPA about the concerns expressed on measuring the accuracy of input power at low loads and the impact of fixed errors in systems with multiple PSUs. Oracle agrees with the EPA to limit the accuracy of PSU input power measurement to 5% or 10W, whichever is greater. This is a practical solution that is consistent with the capabilities of currently available power supplies and power analyzers.

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
298-306	Oracle request that the EPA provide further clarification on the exact definitions of single and multiple output power supplies. For example, a new generation of power supplies intended for use in blade chassis has two independent AC-DC conversion power trains while having a single system interface in the single mechanical unit. Should this be considered a multiple output power supply, or should it be assessed under the single output requirements because this is a high line output only?
344-349	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.
502	Oracle would like to request clarification from the EPA about the power consumed by PSU fans. The EPA should explicitly articulate whether PSU fan power should be included in the PSU losses for the purposes of calculating power efficiency. Oracle suggests that the guideline for this parameter in Version 2.0 remain consistent with the existing guideline in Version 1.0.
536-537	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.
573-576	The EPA should clarify that the requirement to have fan speed management capability enabled by default only applies to chassis fans, not to fans in the power supply. Fan speed in the power supply may be subject to a minimum speed defined by the power supply dependent on temperature, load and input voltage, and may not be able to be controlled by the system.
745-747	The EPA should recognize that calculation of a rolling average might not be possible at the desired rate due to inherent reporting limitations in the power supply. Further work is needed on the averaging to define the split between the power supply and the system fraction of the rolling average. Further, the EPA should formally state that the assumption is the measurements are made on a pure sine wave at 230V 50 or 60Hz.

Line Number in EPA Draft 1	Question or Required Clarification
785-789	The EPA should formally state that the assumption is the measurements are made on a pure sine wave at 230V 50 or 60Hz. The EPA should also formally state that the input power is defined as the sum of all power supplies, even under conditions where the power supplies do not necessarily share the load equally. The per power supply tolerance allowances are assumed to remain in place even when power supplies deliver unequal loads.
811-817	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.
957-968	The EPA should clarify exactly what it intends when it specifies the optional testing conditions for the AC/DC Japanese market to be 100V 50/60Hz. Can the vendor perform this test at either 50Hz or 60Hz, or does the vendor need to perform the test at both frequencies?
992	For all testing configurations, the EPA should delete the requirement: “all PSUs must be connected and operational” and replace it with the restated requirement: “all PSUs must be installed and capable of operation.” This is because new innovative power supply efficiency management techniques in servers may, at idle or low loads, take one or more power supplies entirely out of the power delivery chain to preferentially enable the remaining power supplies to operate higher on their efficiency curve, thus minimizing the power loss inside the power supplies at low loads. Because such techniques make the server more efficient, it is necessary for the EPA to relax the requirement in line 992 that all PSUs be operational.

6. Blades

Oracle applauds the EPA's decision to include blade servers within the scope of Energy Star for Computer Servers Version 2.0. This expands the domain of applicability of Energy Star qualification to a larger number of servers shipping in the industry today, and if done correctly will provide end-user value and purchasing guidance to enterprises that are customers of blade servers.

Oracle has several comments and observations on the proposed EPA procedure for qualifying blade servers for Energy Star. These observations are intended to be constructive improvements in the EPA procedure. Oracle would welcome the opportunity to engage the EPA in a dialog around Oracle's suggestions for the Energy Star blade server qualification procedure.

A. Blade Qualification Versus Blade Chassis Qualification

It is clear from the draft specification that the EPA intends to qualify blade servers under the Energy Star for Servers program. It is not clear whether the EPA intends to separately qualify blade chassis for Energy Star qualification. The draft specification on lines 573-576 states as follows:

> *"To qualify for ENERGY STAR, a blade chassis that is (1) shipped with an ENERGY STAR qualified blade server, or (2) marketed for use with an ENERGY STAR qualified blade server, must provide real-time chassis temperature monitoring and fan speed management capability that is enabled by default."*

The above paragraph suggests that a blade chassis can separately qualify for Energy Star since it lays down conditions that a blade chassis must meet in order to be considered Energy Star.

Further, the draft specification states in lines 577-580:

> *"To qualify for ENERGY STAR, a blade server that is shipped to a customer independent of a blade chassis must be packaged with documentation to inform the customer that the blade server is ENERGY STAR qualified only if it is installed in a blade chassis meeting requirements in Section 3.4.a) and 3.4.b) of this document."*

The above paragraph suggests that a blade chassis does not independently qualify for Energy Star. Rather, only the blade server qualifies for Energy Star, and only if it is installed in a blade chassis meeting the requirements for power allowance and thermal management.

The above situation is confusing. The Energy Star program only defines the responsibilities, features and allowances that a manufacturer of servers must demonstrably meet at the time the server is shipped. The manufacturer can not be held responsible for installation of a server by a customer. While the manufacturer could document that the blade server is Energy Star qualified only if it is installed in a chassis that meets the EPA-indicated power allowances and thermal management features, the manufacturer has no control over where the customer actually chooses to install it. When the manufacturer ships an Energy Star blade server, the manufacturer will increment the qualified product shipment counter so it can report to the EPA on an annual basis the number of Energy Star qualified servers shipped in fulfillment of its obligation as an Energy Star partner. Should the customer choose not to install the blade in a qualified chassis, per this draft specification (line 579) the blade is no longer Energy Star qualified. However, the manufacturer has no ability at this point to decrement its qualified product shipment counter.

To keep things simple for server manufacturers, for the EPA, and for customers Oracle suggests the following clarifications:

- ***Qualification at Shipping Time:*** Energy Star qualification is defined only at shipping time, not at installation time. Any condition that defines Energy Star qualification based on the circumstances of installation of a blade server is inherently flawed, can not be enforced and can not be tracked.
- ***Qualification Based on Blade Attributes:*** Blade servers qualify separately and individually for Energy Star, based on their features, capabilities, and ability to meet power allowances – and not based on how and where they are installed.
- ***Qualification Based on Chassis Attributes:*** Blade chassis qualify separately and individually for Energy Star, based on their features, capabilities, and ability to meet power allowances – and not based on what blade servers are installed in them.

There are many circumstances under which a qualified blade chassis may have non-qualified blade servers installed in them. For example:

- ***High End Blades:*** A customer may choose to fully populate a blade chassis that meets the EPA's power allowances and thermal management criteria with eight socket blades (which may occupy a single bay or two bays). In this case the blade servers are not Energy Star qualified because they are out-of-scope, but the chassis should still be deemed Energy Star qualified.
- ***Appliance Blades:*** A customer may choose to fully populate a blade chassis that meets the EPA's power allowances and thermal management criteria with specialty blades such as storage blades, network blades, or appliance blades. In this case the blade servers are not Energy Star qualified because they are out-of-scope, but the chassis should still be deemed Energy Star qualified.
- ***Older Blades:*** A customer may choose to fully populate a blade chassis that meets the EPA's power allowances and thermal management criteria with blades that were manufactured and shipped prior to the Energy Star for Computer Servers Version 2.0 went into effect. In this case the blade servers are not Energy Star qualified because they were shipped prior to the Energy Star effective date, but the chassis should still be deemed Energy Star qualified if the chassis was shipped after the Energy Star effective date.

For all of the above reasons, Oracle recommends that the EPA define criteria for:

- Qualifying blade chassis for Energy Star based on their inherent features and capabilities regardless of what blades are installed in them.
- Qualifying blade servers for Energy Star based on their inherent features and capabilities regardless of what chassis they are installed in.

B. Power Allowance Calculation for Blade Chassis

Tables 4 and 5 (lines 571 and 572) seem to suggest that the power allowances for blade chassis will only depend on the following:

1. Base power allowance to be determined by the EPA
2. Allowance based on PSUs installed in a redundant configuration
3. Number of ports in the blade chassis greater than two of 1Gb Ethernet or faster than 1Gb Ethernet

The above structure which will determine the formulas for calculating the chassis power allowance is grossly insufficient. This methodology for calculating the power allowance for blade chassis does not account for any of the following:

- ***Expansion-friendly Power Supplies:*** Some blade chassis are designed to hold both blades that fall within the scope of Energy Star (e.g. 1S through 4S blades) and blades that fall outside the scope of Energy Star (e.g. 8S blades). Eight socket blades are bigger and hotter than one, two, or four socket blades. As such, chassis that are capable of hosting eight socket blades need bigger power supplies to accommodate the eventuality that the customer might populate it fully with eight socket blades. However, in the situation where the customer populates it with lower-end blades, the power supply will not be fully used and may operate at a lower point on its efficiency curve.

  We recommend that chassis not be penalized for Energy Star qualification under these circumstances. Chassis that have expansion-friendly power supplies – because, for example, they need to provide power to 8S blades that may be out of scope for Energy Star – should still be able to qualify for Energy Star.

  Because customers buy blade chassis with expansion in mind, the EPA should grant such high-end chassis an additional power allowance for their higher power supply losses, because of their capability to accommodate high-end blades. Otherwise, vendors may be forced to offer two different kinds of chassis: Energy Star qualified, and expandable. If vendors offer customers such an option, many customers may choose an expandable chassis instead of a qualified chassis, thus compromising the success of the Energy Star program.

- ***Fans:*** For the same reason, chassis that are capable of hosting eight socket blades need higher CFM fans to accommodate the eventuality that the customer might populate it fully with eight socket blades. However, in the situation where the customer populates it with lower-end blades, or mixed high-end and low-end blades, the fans will operate at lower speeds or mixed speeds. As such, the fan efficiency may be lower than for a blade chassis that can only accommodate low-end blades for which the fans are properly sized.

  We recommend that chassis not be penalized for Energy Star qualification under these circumstances. Chassis that have high CFM fans because they need to provide adequate cooling to high-end blades that may be out-of-scope of Energy Star, should still be able to qualify for Energy Star.

The EPA should grant such high-end chassis an additional power allowance for their high CFM fans, because of their capability to accommodate high-end blades.

- ***Service Processor:*** Many blade chassis have intelligence built into the chassis in the form of a chassis service processor. This is in addition to any blade service processor that might exist on the various blades that populate a chassis. The chassis service processor provides a combination of management features for the chassis, including power and thermal management. As such, blade chassis with a chassis service processor are better citizens of an energy efficient data center than chassis without a service processor. However, the chassis service processor does itself consume energy.

  We recommend that an allowance be granted for chassis that have chassis service processors, so that a chassis with intelligent power and thermal management is not penalized for Energy Star qualification relative to a chassis that does not have a chassis service processor.

- ***Active Backplanes:*** Some blade chassis have active backplanes (i.e. backplanes that work as networking switches between blades and with active components such as switching ASICs on the backplane) while other blade chassis have passive backplanes (with inter-blade networking being provided by external modules). Chassis with active backplanes are more efficient at the data center level because they eliminate the need to have an external module or device to provide inter-blade networking. However, such chassis consume more power because the ASICs and switching components on the backplane need to be powered-up.

  We recommend that the EPA provide an additional power allowance for chassis with active backplanes so that they are not penalized for Energy Star qualification relative to chassis with passive backplanes.

- ***PSU Fans:*** Oracle would like to request clarification from the EPA about the power consumed by PSU fans. The EPA should explicitly articulate whether PSU fan power should be included in the PSU losses for the purposes of calculating power efficiency. Frequently the fans in the PSUs drive enough CFM to cool not just the PSUs but other components in the blade chassis as well. Because blade chassis with such PSUs provide system cooling, we recommend that they not be penalized for Energy Star qualification by counting the PSU fan power in the power loss calculated for the PSU.

C. Blade Power Calculation Methodology

This draft specification identifies the methodology for testing blade power in both the idle and full power modes. However, this specification does not indicate how the power of an individual blade should be metered. In particular the following questions arise:

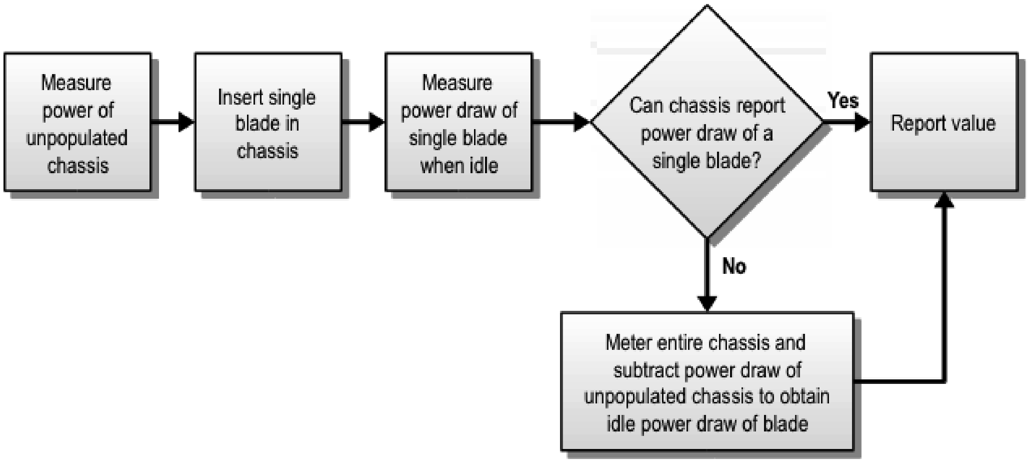
- ***Blade Power Reported by Chassis:*** Some chassis are capable of reporting the power draw of each individual blade in a bay. However, the power measurement mechanisms that are built into such a chassis to measure the power draw of each individual blade do not necessarily meet the EPA's criteria for power analyzer accuracy. The EPA should clarify whether or not measurements of blade power taken from the chassis' built in power reporting mechanism will be acceptable for the purposes of Energy Star qualification, or whether the EPA will require the power measurement of each individual blade to be externally metered using a power analyzer with the accuracy specified by the EPA.

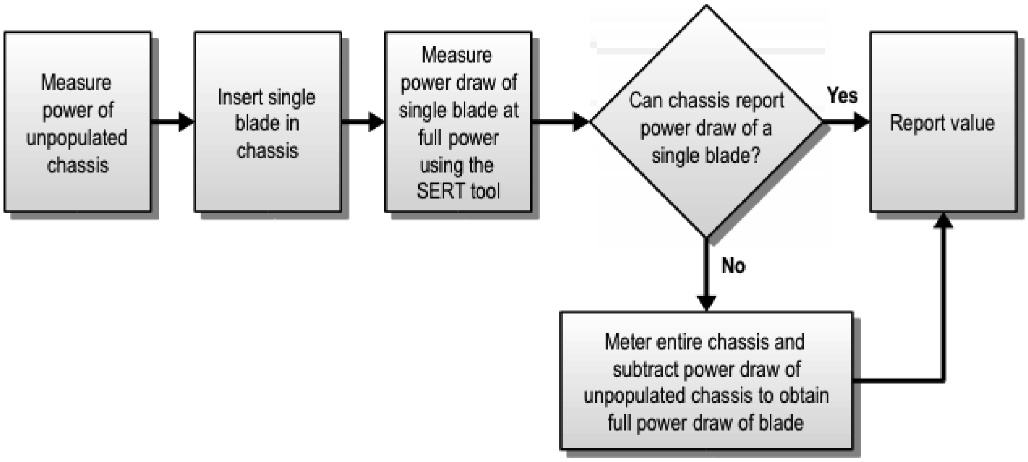
- ***Blade Power Not Reported by Chassis:*** When a chassis is not capable of reporting the power draw of each individual blade in its bays, the EPA should clarify the methodology for metering the power of an individual blade. If it is permissible to measure the power of an individual blade using a subtraction technique, the EPA should explicitly state as much. Under these circumstances, the most practical way of testing the power consumption of an individual blade is to subtract the power consumption of an empty chassis as measured by an EPA approved external power analyzer from the power consumption of the same chassis when populated with a single blade.

- ***Blade Power Measured in Test Fixture:*** The EPA should clarify whether it is permissible to measure the power consumption of a single blade in a stand-alone passive test fixture, designed as an enclosure for a single blade only, without any of the other properties of a bay in a blade chassis (i.e. no network backplane, no fans, no internal PSUs, etc.).

D. Blade and Chassis Power Testing Process

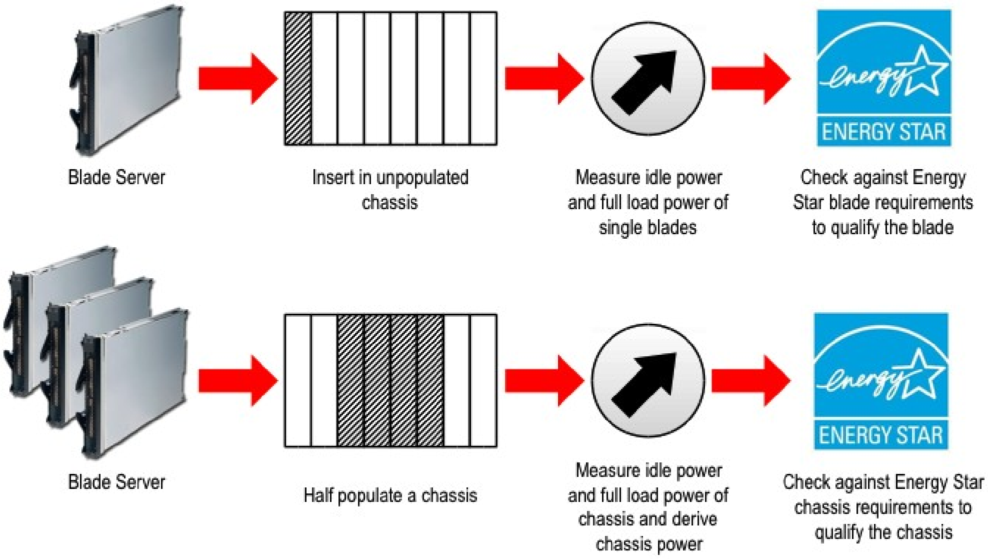
Our understanding of the blade testing methodology has been captured in the following diagram. We welcome the EPA's comments and feedback as to whether the following diagram correctly represents the EPA's intent.


Measure power of unpopulated chassis
Insert single blade in chassis
Measure power draw of single blade when idle
Can chassis report power draw of a single blade?
Yes
Report value
No
Meter entire chassis and subtract power draw of unpopulated chassis to obtain idle power draw of blade


Measure power of unpopulated chassis
Insert single blade in chassis
Measure power draw of single blade at full power using the SERT tool
Can chassis report power draw of a single blade?
Yes
Report value
No
Meter entire chassis and subtract power draw of unpopulated chassis to obtain full power draw of blade

Oracle Corporation would like to request the EPA's feedback on the methodology described in the subsequent diagram for conducting the following tests:

- Idle power of a single blade
- Full power of a single blade
- Overhead power of the chassis when all populated blades are idle
- Overhead power of the chassis when all populated blades are at full power


Blade Server
Insert in unpopulated chassis
Measure idle power and full load power of single blades
ENERGY STAR
Check against Energy Star blade requirements to qualify the blade
Blade Server
Half populate a chassis
Measure idle power and full load power of chassis and derive chassis power
ENERGY STAR
Check against Energy Star chassis requirements to qualify the chassis

The above diagram represents Oracle's understanding of the methodology intended by the EPA for qualifying blade servers for Energy Star and for qualifying blade chassis for Energy Star. If the above diagram incorrectly represents the EPA's intent, please provide feedback accordingly. If the above diagram correctly represents the EPA's intent, Oracle has the following further questions about measuring the full power draw of a half-populated chassis:

- Should the full power of the chassis be measured by running the SERT tool on all the blades?

- The SERT tool is a composite workload, that is, it stresses different resources on the blade server at different times. During different periods of its operation, the SERT tool stresses CPU resources, memory resources, I/O resources, and disk resources in sequence. It is expected that the power draw of any blade running the SERT tool will be different depending on the phase of execution of the SERT tool. Should the vendor report the maximum power draw of the blade during this test, or the average power draw over the entire duration of execution of the SERT tool?

- Should the SERT tool be started on all the blade servers simultaneously, or is some variability in the start times permitted? (Note that because the SERT tool causes the blade to draw a different amount of power at different times in its execution cycle, running the SERT tool in a non-synchronized manner on all blades will lead to non-reproducible sequence of power readings at the chassis level.)

E. Blade Chassis Testing Options

Oracle would like to suggest the following options in the blade testing procedure which will give vendors an additional degree of testing flexibility without compromising any of the objectives intended by the EPA:

- ***Option to test a fully populated chassis:*** For measuring the power draw of partially populated chassis, we applaud the EPA's decision to permit the measurement of chassis of only half the bays populated. This will go a long way towards the vendor acceptance of the Energy Star program for blade servers, as it is difficult to find fully populated blade chassis in a test lab. It is also more reflective of realistic situations where customers deploy blade chassis in data centers that are not fully populated with blades.

  Oracle would like to request the option of testing a blade chassis that is more than half-populated. For example, in the event that a test lab finds it possible to fully populate a chassis, Oracle would like to request the EPA for the ability to also report the power consumption test results from a fully populated chassis. This would allow blade server vendors to report the system efficiency, which is expected to be higher when the blade server is fully populated. This is because the chassis overhead (the point on the efficiency curve at which the chassis power supplies operate, the speed of the fans of the chassis, etc.) is amortized over a larger number of blades. If the EPA receives results from vendors with both half-populated and fully-populated chassis, it is a mathematically simple operation to normalize the stack-ranking of the chassis power overhead down to either a constant or a proportional number of blades in a chassis.

- ***Option to test chassis populating power supply domains:*** The EPA has asked that chassis be tested with half-populated shelves populated from the center outwards. In many blade servers, the available slots are divided into “power supply domains”. These power supply domains do not necessarily span adjacent slots. For example, in a blade chassis, Power Supply A might serve slots 1, 3, 5, 7, and 9, while Power Supply B might serve slots 2, 4, 6, 8, and 10. Instead of populating the blade chassis in a cluster of adjacent slots in the center, we would like to request the EPA for the option of populating an entire power supply domain instead. This would allow the blade server to demonstrate higher efficiency, because one power supply would operate higher up on its utilization curve, while the other power supply would be unused and may be shorted out of the power delivery circuit to eliminate unnecessary power supply losses.

- ***Option to test chassis populated with heterogeneous blades:*** The requirement to half-populate the blade chassis with identical blades of the same model and configuration still remains impractical for many vendors. Typically, a test lab (including any independent third-party test lab) would have a few blades of each type, and may collectively have a sufficient number of blades to half-populate the blade chassis, but will likely not have all blades of an identical type. It will be onerous for a vendor to supply a third-party test lab with multiple sets of identical blades.

  Since it is the EPA's intent to measure the power draw of a half-populated chassis for the purposes of deriving the chassis power overhead, Oracle would like to suggest that this goal can be achieved without imposing the requirement that all blades be identical.

  The EPA specified formulas on line 1,111 and 1,112 to calculate the chassis power overhead (both at idle and full power) are as follows:

$$P_{\text{Chassis(Idle)}} = P_{\text{chassis(Idle, ½ populated)}} - [\text{\# bays populated}]*[P_{\text{Single Blade(Idle)}}]$$

$$P_{\text{Chassis(FullP)}} = P_{\text{chassis(FullP, ½ populated)}} - [\text{\# bays populated}]*[P_{\text{Single Blade(FullP)}}]$$

  These require the chassis to be populated with identical blades. Under a test methodology that would permit the population of half the bays in a blade chassis with heterogeneous blades, the formulas would change as follows:

$$P_{\text{Chassis(Idle)}} = P_{\text{Chassis(Idle, ½ populated)}} - \sum_{\text{all populated blades}} P_{\text{Single Blade(Idle)}}$$

$$P_{\text{Chassis(FullP)}} = P_{\text{Chassis(FullP, ½ populated)}} - \sum_{\text{all populated blades}} P_{\text{Single Blade(FullP)}}$$

  Oracle recommends that the EPA change the formulas as suggested above. When all blades are indeed identical, these formulas are equivalent to the degenerate case formulas already suggested by the EPA. When all blades are not identical, these formulas would permit additional flexibility for testing large blade chassis and still yield the EPA-desired result of deriving the power overhead of a populated chassis. The heterogeneity of the blades is not antithetical to the EPA's intent of qualifying the chassis itself rather than the blades, since the proposed new formulas above would still yield accurate measurements of the chassis overhead both for the idle power consumption of the chassis and for the full power consumption of the chassis.

7. Active Mode Efficiency Criteria

Oracle applauds the EPA's decision to use a standard rating tool to measure the overall throughput of a server for calculating the energy efficiency of a server at high utilization. Oracle encourages the EPA to continue working with SPEC to make aggressive progress on the development of the SERT tool.

The EPA should continue to make absolutely sure that the requirements on the SERT tool are adhered to by SPEC tightly and scrupulously. In particular Oracle considers the following requirements on the SERT tool as non-negotiable:

- *Architecture neutrality:* the SERT tool must run on all CPU architectures.
- *Fairness:* the SERT tool must not be subject to gaming through software configuration (e.g. by BIOS configuration, or by JVM tuning). Any JVM configuration parameters that are tunable must be locked down and have the same values on all CPU architectures.

In addition, the EPA should encourage SPEC to finish the development of the SERT tool in a timely manner that provides vendors with sufficient time to gain experience with its use prior to the formal collection of data for the purposes of qualifying systems for Energy Star for Servers Version 2.0.

If for any reason the SERT tool is not available from SPEC in a timely manner, or experience shows that the SERT tool does not meet the requirements of architecture neutrality and fairness, or the SERT tool proves to be an inaccurate indicator of the energy efficiency of a system at high throughput, Oracle encourages the EPA to have an alternate plan to provide server customers with some other vendor specified indicator of the efficiency of a server at high utilization.

8. Power / Performance Data Sheet and QPI Form

The forms required as documentation in Energy Star for Servers Version 1.0 left much to be desired. The Power and Performance Data Sheet and the Qualified Product Information Form contained much of the same information. Given the fact that one set of these forms had to be filled out for each family definition, and a number of family definitions was subject to combinatorial explosion because of the tight restrictions on the constitution of a family, the process of filling out these forms imposed an excessive burden on the vendors.

Besides normalizing the family definition to be more relaxed so as to reduce the number of forms that a vendor needs to submit, Oracle also would like to encourage the EPA to collapse the Power

Performance Data Sheet and the Qualified Product Information Form down to a single consolidated form.

Oracle appreciates the EPA's intent to have one form that is customer facing and another that is directed towards the EPA in order to assist the EPA with its analysis and evaluation of the submitted products for Energy Star qualification. Oracle believes that this intent can be preserved even with a single consolidated form. The form can be structured such that all the customer facing information can be derived as a subset of the consolidated information presented in the form. This will facilitate the following objectives:

- Allow the EPA to make its qualification decision with all the information that it needs for this purpose.
- Allow customers to evaluate the relevant parameters of each server relative to Energy Star while making a purchasing decision.
- Allow vendors to provide the necessary information to both customers and to the EPA in a single form without having to enter duplicate information.

9. Effective Date

Oracle suggests that the EPA allow a full development cycle between the date of issuance of the final specification and its effective date. A full development cycle is typically 12-24 months in the server industry. This will allow sufficient time:

- For vendors to design products compliant with the new specification, with the appropriate sensors and reporting requirements.
- For customers to plan a graceful transition in their purchasing requirements from Energy Star for Servers Version 1.0 to Energy Star for Servers Version 2.0, and to upgrade their systems management and power monitoring software tools to handle the increased volume of data reporting from Version 2.0 compliant servers.
- For distributors and resellers to flush the channel of systems that were previously shipped and already labeled Energy Star because they qualified under the older specification.

One alternate approach is for the EPA to consider allowing products to use the specification that is currently in effect at the start of their development cycle, instead of at the end. For example, a product whose development cycle starts today would be allowed to continue to qualify under Energy Star for Servers Version 1.0 throughout its shipping life, while a product whose development cycle starts in 2011 would be required to qualify for Energy Star for Servers 2.0. If the EPA were to allow such an approach, a long delay between the issuance of the specification and its effective date to allow for a full development cycle would not be necessary.