
# ENERGY STAR® FOR COMPUTER SERVERS VERSION 2 DRAFT 1 FEEDBACK AND RECOMMENDATIONS (MAY 2010)

 Rev 2010-0

The Green Grid Association, a consortium of industry-leading companies, welcomes the opportunity to comment on topics under consideration for the ENERGY STAR® for Computer Servers specification.

## INTRODUCTION


A consortium of information technology providers, consumers, and other stakeholders, The Green Grid seeks to improve the energy efficiency of data centers around the globe. The association takes a holistic and comprehensive approach to data center efficiency and understands that developing The ENERGY STAR Tier 2 performance/power metric represents a significant challenge, one which requires cooperation among a wide range of industry principals. Participants in The Green Grid include such diverse companies as major server and storage equipment manufacturers, major software providers, and large data center end users/owners.

## SUMMARY

The Green Grid appreciates the EPA's work with the industry and with SPEC to develop an ENERGY STAR program for servers focusing on energy efficiency. We welcome the opportunity to respond to the draft1 of the tier 2 specification for servers. We are happy to see that the program details focus on efficiency and a movement away from idle only criteria to efficiency metrics. Though the development of a Server Efficiency Rating Tool has been arduous, we agree with the pursuit of a metric that provides a continuous incentive to innovate for energy efficiency.

With the schedules outline, TGG believes that the program should focus on just a few items and ensure data collection and increased ENERGY STAR participation. Specifically, we believe that in order to meet the 2010 target we recommend the focus should concentrate on:

- Adding Blades to the product scope
- Incorporating data collection on SERT™
- Incorporating data collection for future evaluations
- Addressing product family definitions

TGG supports the EPA's proposal in the Tier 2 draft to maintain the focus of the requirements on systems with up to 4 processor sockets and to qualify all server systems based on implementation of processor level power management and eliminate the idle power limits for one and two socket servers while data is gathered to support the development of a criterion based on an active mode efficiency metric.

The Green Grid has enclosed detailed comments on the major sections of draft1 and recommendations that would meet ENERGY STAR schedule targets. The TGG task force and members look forward to further clarifications and development of the specification.

 

# DETAILED COMMENTS BY SECTION

## DEFINITIONS

### RESILIENT SERVER

The current list of attributes is generic and may be confused with the features that separate this class of server. We've restated the features of a Resilient Server below to help clarify the feature list. A Resilient Server should have all or many of these features:


- Memory Fault Detection and System Recovery: DRAM Chip Sparing, Extended ECC, Mirrored Memory
- Machine Check Architectures – Fault Isolation and Resiliency
- End to End Bus Retry
- Hot-swap components: I/O, hard drives and AC/DC power supplies
- Ability to perform on-line expansion and retraction of hardware resources without OS reboot - also referred to as "on-demand"
- Multiple physical banks of memory and I/O adapters

### PROCESSOR UTILIZATION

The definition of processor utilization should be changed to read: "The percentage estimate of the server's compute activities relative to the full operational voltage and frequency of the processor(s)." Processor utilization can be reported at the processor of CPU core level.

## QUALIFYING PRODUCTS

### HIGH PERFORMANCE COMPUTING (HPC) SYSTEMS

We do not believe there have been significant changes in either the volume of shipments, available testing procedures, or energy profile information that support incorporating HPC systems at this time. Even though the base components may look similar, such as a rack 2P configuration vs. HPC system comprised of 2S motherboards, the total system integration emulates a multi-processor system, capable of operating on a logically unified memory image. The resulting energy profile is not a linear extrapolation of multiple 2S servers. In fact inter-platform coordination and logic may show higher idle for a single 2S motherboard, even though the multi-motherboard HPC system is more efficient than purchasing multiple 2S servers and developing an external means to aggregate the compute capabilities. It is also highly unlikely to test partial subsystems independent of the aggregate HPC system. The unique architecture and software stack may prevent the use of generic server benchmarks and tools. Therefore, we recommend that HPC systems be out of scope for this specification.

## APPROACH: BLADES

TGG supports covering blade servers under the Computer Server Power management requirements proposed in the Tier 2, Draft 1 requirements, with the requirement for implementation of processor level power management on shipped products and reporting of the active energy metric for a single blade server.

If EPA chooses to continue setting idle power criteria for one and two processor socket servers, then TGG recommends that blade servers be evaluated under the server power management criteria for three and

 

four processor systems because of the differences in overhead distribution between blade and rack server systems and the resulting impact on the server idle power calculation.

## BLADE CHASSIS

With regards to EPA's proposal for qualifying a blade chassis, TGG strongly recommends that EPA eliminate the Table 4 and 5 criteria for a blade chassis and replace it with the following set of requirements:


1. PSUs in the blade chassis should meet the computer server efficiency and power factor requirements.
2. The chassis should have variable speed fans.
3. The chassis should be capable of reporting power use and thermal information for the blade system.

This establishes the functional requirements for the chassis, without requiring extensive measurements of power use on the chassis. In addition, each manufacturer configures their blade chassis differently, with different percentages of the "overhead" power for fans, network, and hard drive components.

### BLADE SYSTEM TESTING

A complete blade system of minimally configured blades and a full chassis should be tested using the active energy metric. A manufacturer may choose to test a partially populated blade chassis, as long as a given power domain (the maximum number of slots supported by a single power supply or a pair of redundant power supplies) is fully populated with blades. The total power data can be divided by the number of blades in the chassis to calculate the fully burdened, per blade power usage in a fully populated chassis. The minimum blade server configuration should be used for the full chassis power test, as it provides the best means to allow meaningful comparisons between manufacturer's systems. It is likely that typical and maximum configurations will vary sufficiently in power use to make it difficult to derive meaningful comparisons between systems.

The power associated with a single, minimally configured blade server can be calculated by removing one blade server from the chassis and re-measuring the power use at idle and maximum power. The difference in power measurements represents the unburdened, single blade power requirements. The chassis power can be calculated by subtracting the individual blade power times the number of blades in the chassis from the total chassis power.

Similarly, the idle and maximum power for a typical and maximum blade configuration can be determined by inserting and removing a single blade of each configuration into the chassis populated with minimally configured blades.

This general procedure, working off a single, fully populated blade chassis using minimally configured servers will provide the power measurements needed to provide the information needed to populate the Product Performance datasheet.

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.

We believe the intent behind measuring the power draw of a half-populated chassis was to ease the derivation of chassis power overhead, The Green Grid 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:


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

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

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:

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

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

The Green Grid 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.

## ACTIVE MODE EFFICIENCY RATING TOOL

The Green Grid and its members fully support the current effort by SPEC Power to generate the Server Efficiency Rating Tool ™. The proposed tool continues to offer the best approach to establishing an active mode rating for computer servers. The Green Grid agrees with the design targets and the approach of evaluating efficiency in this manner.

## POWER SUPPLY

### NET POWER LOSS (NPL)

The Green Grid appreciates ENERGY STAR's considerations to industry input and the decision to stay with industry standard efficiency methods and not convert to a NPL process.

### POWER SUPPLY EFFICIENCY

We understand the desire to aggressively pursue higher levels of efficiency. The levels and direction are

 

consistent with the advances and capabilities in the industry.

## ENERGY EFFICIENT ETHERNET

The Green Grid appreciates the investigation and considerations afforded to this topic. We agree with the conclusion to postpone adopting this feature pending the industry validation of the technology over the next few years (2010-2012).

## IDLE POWER


The Green Grid applauds the EPA's willingness to open the idle power discussion for Version 2.0 of ENERGY STAR for Servers. TGG supports EPA's proposal in the Tier 2 Draft 1 document to require implementation of processor level power management in all shipped servers and reporting of the active energy metric and its subroutines.

TGG believes this approach is appropriate, as 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.

The Green Grid strongly encourages the EPA to consider alternative approaches to idle power. In particular, The Green Grid recommends an approach that grants each server an idle power allowance based on its power consumption at peak throughput.

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.
- 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.


In line 687 of the Draft1 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. The Green Grid 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 self-limiting parameter. The reason the peak power draw of a server cannot scale indefinitely to arbitrarily large numbers is because of the following limitations:

- Technology improvements can lead to higher frequencies, higher capacities, and higher speeds in components such as CPUs, memory DIMMs, disk drives, and fans, but these technology improvements typically only manifest themselves in a higher power draw at peak utilization, not a higher power draw at idle. For example, an 8GB DIMM may draw more power than a 4GB DIMM at peak memory access rates but does not draw significantly more power at idle. Similarly, a 15,000 RPM drive may draw more power than a 7,200 RPM drive when the user IOPS rate is high, but does not necessarily draw more power at idle.
- 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.
- 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.

## STANDARD REAL TIME PERFORMANCE DATA MEASUREMENT AND OUTPUT REQUIREMENTS

The Green Grid 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.

The Green Grid, however, 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. The Green Grid 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 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 real time 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 two orders of magnitude greater than the EPA requested frequency of once per second. For example, the time horizon for taking action on the provisioning or re-provisioning 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.

The Green Grid requests that the sample frequency for dynamic power, temperature, and utilization information be recalibrated to the needs of the applications that will use this data. The Green Grid recommends a sampling frequency of no greater than one reading every thirty (30) seconds for power draw and processor utilization information, and no greater than one reading every (1) minute for temperature information.

## FAMILY DEFINITION

It is the experience of all 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. 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. 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.

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

However, The Green Grid 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.

Under the Tier 1 product family definitions, a four processor system is required to delineate a product family, and provide a full QPI/product data sheet for each level of populated processor socket and number of processor cores. Instead, we propose that for a product family, we define the power use and performance of the minimum and maximum configurations based on the highest power processer that can meet the specification requirements, with the minimum configuration defined as a machine type with the minimum number (typically one) processor socket populated, a typical configuration as a system with 2 processor sockets populated, and a maximum configuration as a system with 4 processor sockets populated. This is an example for a four processor system; a similar approach would be used for a two processor system. This approach represents the full range of power and performance for the machine type (or with a defined subset of model numbers) and provides the information that our customers are typically interested in when they evaluate our products. Specific power information for a configuration can be determined using the power calculator that most manufacturers have available for their equipment.

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

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

- 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.

- 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 in the same server line can vary widely. Different I/O devices have different technical and power specifications. However, the differences between the power specifications 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. 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.

- 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 remove the requirement that (a) the same part number of PSU be required in all server configurations within a product family (because the server model may be shipped with both the original qualifying PSU type and the new qualifying PSU type), and (b) the same technical and power specifications for PSUs be required within a product family (for the same reason).

TGG strongly opposes the proposed change to the definitions for the maximum and minimum configurations. Associating these configurations to the maximum and minimum possible active energy efficiency creates an indeterminate hardware definition. In order to ascertain which configuration represents the maximum and minimum achievable active energy efficiency a detailed active energy versus configuration versus power use matrix would need to be tested to allow statistical determination of the minimum and maximum active energy state for that product family. The extent of the testing required to make this determination would negate any benefit of the product family categorization of products. TGG recommends staying with a hardware configuration based definition of maximum and minimum, as listed in our recommendation for a product family description (see Family Definition, pg. 8)

## STANDARD INFORMATION REPORTING REQUIREMENTS

The Power and Performance Data Sheet and QPI forms require the same information in two different formats (e.g. PSU Efficiency and PFC values). Duplicate information in varying formats leads to confusion, data entry errors and resubmissions. A consolidated sheet should be considered to allow a single entry and product description to be used in both forms. For those sections unique to a form, those areas should be separated and unique to the form. The TGG would welcome the opportunity to hold a detailed review on the data entry format, process and details to clarify the reporting and qualification requirements. The submission forms and process can be incorporated as part of a training package in conjunction with ENERGY STAR's enhanced

 

qualification and verification process.

## CONCLUSION

The Green Grid remains committed to a successful and collaborative development of the ENERGY STAR for Computer Server Tier 2 specification with all industry stakeholders and the EPA. We believe with the focus for Tier 2 should be on corrections to Tier 1 regarding family definitions, incorporating methods to include bladed servers, and incorporating data collection using SPEC's SERT™ tool. The combination and consistency of the ENERGY STAR for Computer Server program and the efficiency initiatives in the EPA and US DOE should help in accelerating the efficiency in operation of the data center. The Green Grid will continue to collect industry-wide inputs to work with the EPA in developing the ENERGY STAR programs on ICT equipment. Please feel free to contact us to clarify and collaborate on the development of the specifications and the implementation of the program.