



ENERGY STAR® for Computer Servers Version 2.0 Draft 2 Feedback and Recommendations (June 2012)

The Green Grid Association, a consortium of industry-leading companies, welcomes the opportunity to comment on an early draft of 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® for Computer Server Version 2.0 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 efforts to capture and address the issues raised up to and including the webinar held May 24, 2012. With the aggressive schedule for Version 2.0 for the server specification, we recommend holding subsequent detailed reviews on each topic to quickly address category definitions and data disclosures via the Power Performance Data Sheet (PPDS).

The Green Grid is providing comments to proposals and questions in the sections highlighted in the ENERGY STAR draft documents, dated 5/11/12. The comments may be similar to some of the individual responses provided to the EPA, but, represent the consensus opinion of the Green Grid participants in the process.

The Green Grid agrees in general with focus on data collection and updates to the idle specifications. Product family definition of 5 data points will help increase the adoption of ENERGY STAR and reduce cost burden. The Green Grid does not support or recommend exemption of the ECC, buffered or registered memory requirement for systems with 50 CPU nodes and above. The Green Grid has also proposed a refinement of the definition of a resilient server and we intend to provide an improved definition for High Performance Computing system in a future submission to EPA. For idle adders, the memory adder reduction is aggressive, but is feasible for future systems. We recommend that the consideration for "GPGPU" be made generic as "Add-in Compute" capabilities, of which a specific example is a GPGPU add-in device.

We hope these comments and recommendations will be useful in EPA's plans to complete the ENERGY STAR for Servers Version 2.0 specification later this year. We welcome the opportunity to work in an industry forum to address the detailed areas for the final draft and release of the specification.

Commentary by Section

Definitions and Scope

Systems with greater than 50-nodes exempt from error correction

Error correction on the memory subsystem is a key common attribute of a computer server. As the internal memory sizes and communication speeds increase, error correction is needed to ensure server availability, reliability, and uptime for data center operations. The hardware infrastructure and resulting energy profile reflects these requirements. Client computing based systems which don't contain error correction can tolerate repeat transactions, system reboot, and other inefficient activity recovery methods. These systems should not be included in the ENERGY STAR computer server specification. Servers based on personal computing components without memory subsystem error correction are already addressed as a small scaled server in the ENERGY STAR for Computers specification.

We disagree with the proposed exemption for 50-node or larger systems. Multi-node systems without error correction and handling in the memory subsystem have not demonstrated the ability to support current data center and enterprise application requirements. In fact, the increase in applications' memory size and response time requirements may severely impact systems without error correction causing a decrease in application uptime and increase the base energy footprint of the systems to support the applications. With the possibility of increasing the energy profile of the data center and limited live industry efficacy of this configuration we advise that the no exemption be made.

High Performance Computer (HPC) Systems

We still believe that HPC systems form a category that may deserve special consideration. HPC systems are servers utilized in large clusters targeted to maximize performance for scientific research and large scale modeling. Although some HPC clusters are based on general purpose servers, many power management features are disabled to enhance performance. Disabling power management features and the additional hardware installed significantly changes the power profile of these systems. We will continue to work with the industry to provide a distinctive set of criteria to classify these servers.

Resilient and Scalable Server

A resilient and scalable server is designed with extensive Reliability, Availability, Service-ability (RAS) and scalability features, including error self-correction to ensure data resiliency and accuracy. Resiliency, RAS, self-correction, data accuracy and scalability features are integrated in the micro architecture of the CPU and chipset functions. Resilient and scalable servers are engineered with additional, redundant and more complex components in their underlying infrastructure in support of the resiliency features, which in turn require more energy to operate, distinguishing them from a computer server without equivalent level of RAS and scalability features. We recommended that resilient and scalable servers be placed into a different category because of this reason. A resilient and scalable server should be a system that contains the following characteristics:

Item	Criteria	
1	<p>Processor RAS and Scalability To meet the requirement of resiliency, the processor must support these 3 criteria Processors used in resilient and scalable servers have a higher socket level power draw because of the following capabilities.</p>	
1.1	<p>Processor RAS: The processor must have capabilities to detect, correct, and contain data errors.</p> <ul style="list-style-type: none"> i. Error detection on L1 caches, directories and address translation buffers using parity protection. ii. Single bit error correction using ECC on caches that can contain modified data. Corrected data is delivered to the recipient, i.e., error correction is not used for background scrubbing only. iii. Error recovery and containment by means of - 1) processor checkpoint retry and recovery, or 2) data poison indication (tagging) and propagation, or 3) both. The mechanisms notify the OS or hypervisor to contain the error within a process or partition, thereby reducing the need for system reboots. iv. 1) Capable of autonomous error mitigation actions within processor hardware, such as disabling of the failing portions of a cache, or 2) support for predictive failure analysis by notifying the OS, hypervisor, or service processor of the location and/or root cause of errors, or 3) both. 	
1.2	<p>The processor technology used in resilient and scalable servers is designed to provide additional capability and functionality without additional chipsets, enabling them to be designed into systems with 4 or more processor sockets. The processors have additional infrastructure to support extra, built-in processor busses to support the demand of larger systems.</p>	
1.3	<p>They provide high bandwidth I/O interfaces for connecting to external I/O expansion devices or remote I/O without reducing the number of processor sockets that can be connected together. These may be proprietary interfaces or standard interfaces such as PCIe. The high performance I/O controller to support these slots may be embedded within the main processor socket or on the system board.</p>	
2	<p>Memory RAS and Scalability Must have all of the following capabilities and characteristics:</p>	
2.1	<p>Provides memory fault detection and recovery through Extended ECC (also called Chipkill®).</p>	
2.2	<p>In x4 DIMMs, recovery from failure of two adjacent chips in the same rank.</p>	
2.3	<p>Memory migration: Failing memory can be proactively de-allocated and data migrated to available memory. This can be implemented at the granularity of DIMMs or logical memory blocks. Alternatively, memory can also be mirrored.</p>	
2.4	<p>Uses memory buffers for connection of higher speed processor -memory links to DIMMs attached to lower speed DDR channels. Memory buffer can be a separate, standalone buffer chip which is integrated on the system board, or integrated on custom-built memory cards. The use of the buffer chip is required for extended DIMM support; they allow larger memory capacity due to support for larger capacity DIMMs, more DIMM slots per memory channel, and higher memory bandwidth per memory channel than direct-attached DIMMs. The memory modules may also be custom-built, with the memory buffers and DRAM chips integrated on the same card.</p>	
2.5	<p>Uses resilient links between processors and memory buffers with mechanisms to recover from transient errors on the link.</p>	
2.6	<p>Lane sparing in the processor-memory links. One or more spare lanes are available for lane failover in the event of permanent error.</p>	
3	<p>Power Supply RAS</p>	
	<p>Redundant and concurrently maintainable Power Supplies. The redundant and repairable components may also be housed within a single physical power supply, but must be repairable without requiring the system to be powered down. Support must be present to operate the system in degraded mode when power delivery capability is degraded due to failures in the power supplies or input power loss.</p>	
4	<p>Thermal and Cooling RAS</p>	

	Redundant and concurrently maintainable cooling components, such as fans or water-based cooling. The processor complex must have mechanisms to allow it to be throttled under thermal emergencies. Support must be present to operate the system in degraded mode when thermal emergencies are detected in system components.	
5	System Resiliency The system must have 6 of 10 of the following attributes/capabilities	
5.1	Support of redundant storage controllers or redundant path to external storage	
5.2	Redundant Service Processors	
5.3	Redundant DC-DC regulator stages after the power supply outputs	
5.4	The server hardware supports runtime processor de-allocation	
5.5	I/O adapters or hard drives are hot-swappable	
5.6	Provides link level retry (LLR) based protection on processor to memory or processor to processor interconnects.	
5.7	Supports on-line expansion/retraction of hardware resources without the need for operating system reboot (“on-demand” features)	
5.8	Processor Socket migration: With hypervisor and/or OS assistance, tasks executing on a processor socket can be migrated to another processor socket without the need for the system to be restarted.	
5.9	Memory patrol or background scrubbing is enabled for proactive detection and correction of errors to reduce the likelihood of uncorrectable errors.	
5.10	Internal storage resiliency: Resilient systems have at least Level 5 RAID hardware in the base configuration, either through support on the system board or a dedicated slot for a Level 5 or higher RAID controller card for support of the server’s internal drives.	
6	System Scalability The system must have all the following attributes/capabilities	
6.1	Higher memory capacity: >=8 DDR3 or DDR4 DIMM Ports per socket, with resilient links between the processor socket and memory buffers.	
6.2	Greater I/O expandability: Larger base I/O infrastructure and support a higher number of I/O slots. Provide at least 32 dedicated PCIe Gen 2 lanes or equivalent I/O bandwidth, with at least one x16 slot or other dedicated interface to support external PCIe, proprietary I/O interface or other industry standard I/O interface.	

Product Family Testing

We agree that the 5 point testing profile will adequately define the product family classification for compliance. For 1 socket servers, the variability and customizations are fewer and 3 data points should be sufficient. Assuming 3 socket power options and 2 core count options for the processor, one can configure

- a) the minimum power/low-end performance system with the low power and lower core count processor and minimum usable memory, I/O and a single hard drive;
- b) the typical configuration with the mid-range powered processor and high core count and a typical component configuration; and,
- c) the maximum power and high-end performance configuration with the highest processor power and core count and the maximum component configuration.

For a 1 socket system, 3 data points would suitably bracket the product family. The additional two configurations would not add any materially different information from what is collected from the three described configurations. Both the 5 point and 3 point sampling methods are significant improvements to the current method and would limit product testing costs and encourage increased participation in the program. For this reason, we support and encourage ENERGY

STAR to proceed with the 5 point test definition for 2 socket server product family and 3 point test for 1 socket systems.

Product Family Test Configurations

We are concerned about the wording used in the Low-end and High End Performance Configuration definitions. Using the terms **highest** and **lowest** in the definition unnecessarily restricts the choice of configurations for this area and risks creating significant overlap and lack of differentiation with the Minimum and Maximum Power Configurations. Instead, the Green Grid recommends that you use the terms “**lower-price or lower performance**” and “**higher price or higher performance**” configuration to describe these two definitions and provide the manufacturers sufficient latitude to differentiate the performance and power based configurations.

Even with the adjustments requested above, it is highly likely that some qualified configurations will exist outside of the power profile envelope defined by the “Product Family Tested Product Configurations”. The Green Grid recognizes that it is the manufacturer’s responsibility to validate that all products marketed and sold as ENERGY STAR qualified meet the applicable requirements, but EPA needs to assure manufacturers that it understands that the proposed configuration definitions may result in qualified products that exist outside of the power profile defined by the 5 tested product configurations but still qualify and meet ENERGY STAR requirements.

DC systems

Application of systems that rely on direct current (DC) is increasing in the traditional telecommunications (i.e. -48VDC) and Low Voltage Direct Current (LVDC, aka 380VDC) environments. We encourage the incorporation of these systems in the ENERGY STAR program and the prescribed testing methods.

Active mode and Idle Specifications

We concur with ENERGY STAR’s findings that there is insufficient data to support any change to the base idle criteria on 1 and 2 socket servers or establishing an idle requirement for bladed or 3-4 socket systems. Given the expense and complexity, the ENERGY STAR Version 2.0 proposal to collect active mode and idle data is prudent.

For bladed system testing, we recommend that either ½ (half) populated systems or fully populated configurations be allowed. Data from a ½ populated configuration would be sufficient to quantify the shared power constructs in the bladed system. System configurations and test conditions should be described in the power performance data sheet.

The memory adder reduction for memory from 2.0W per GB to 0.75W per GB, is aggressive for smaller or partially filled systems prepared for expansion. As ENERGY STAR may recall, the added memory adders were to address system level support functions such as buffering or expanded memory support functions. These additional system features have improved since Version 1.0 in integration and controls such that a more linear (idle power per GB) attribute can be achieved. Though 0.75W per GB will be challenging for smaller systems with expansion capabilities, 0.75W would be an appropriate aspirational target for a 2013 ENERGY STAR program.

In the other considerations and adders, we appreciate the recognition of and the testing provisions for added compute functions being configured to computer servers. Testing

compliance without the added function and reporting idle after incorporating the feature will accommodate this trend and collect information on the impact. We recommended that “GPGPU” be revised to “Add-in Compute”. GPGPU’s represent a specific implementation of this feature, whereas non-GPU compute cards are also entering into the market to support these applications. A generic description would allow the market to determine applicability of the functions.

Redundant Power Supply

The EPA raised the question on whether or not the “Additional Power Supply Adder” should be revised in Version 2.0 (Line 409). The Green Grid believes that a 20W adder remains appropriate. This adder was used in the analysis of Version 2.0, as well as Version 1.0. The data set is consistent with the original analysis and supports the 20W adder. Please note that the adder is based on PSU (redundancy) technologies and topologies, which haven’t changed. As observed in the original assessment the value does not typically scale to the system configuration.

Power Performance Data Sheet (PPDS) and Qualified Product Information (QPI) forms

The Power Performance Data Sheet will require an update to provide extra data fields to accommodate the new categories of product, the product family definition, and additional active mode data points being reflected. We encourage that automation and error checking be incorporated into the PPDS and QPI forms. SPEC’s SERT™ is expected to contain automated device discovery as part of the tool suite. The results of the discovery routine can aid in minimizing data entry error. Additionally, the industry could provide commonly used configuration identifiers that could be part of default drop down menu’s, further limiting the entry errors.

For active mode data collection and information only assessments, the industry recommends that the data be visible to the public as an anonymous data set. The actual supplier or manufacturer identification for this information would be held by the certifying bodies (CB). Holding the data anonymous allows investigation of the data and trends without premature assessments of these numbers or association with energy efficiency. Consolidation of this information into a single grading method would be expected after analysis of the collected data and in preparation for Version 3.0 of ENERGY STAR for Computer Servers.

Since companies will already provide SERT worklet performance data to EPA under the Version 2.0 plans, the Green Grid recommends that the requirement for testing an additional power/performance benchmark (Section 4.1.2.vi) be removed. We propose that two of the SERT worklets, ccsj and Flood, could be used to satisfy the intent behind Power-Performance Benchmark reporting.

Conclusion

The Green Grid anticipates a successful and collaborative development of the ENERGY STAR for Computer Server Version 2.0 specification with all industry stakeholders and the EPA. We believe with the focus on the volume categories of servers and a metric that tracks the maximum performance within an energy envelope, the Version 2.0 specification can be very successful. 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 both to clarify and collaborate on the development of the specifications and the implementation of the program.
