IBM Comments to "ENERGY STAR® Program Requirements for Computer Servers Draft 4

IBM appreciates the continued efforts by EPA to establish a workable set of ENERGY STAR® computer server requirements applying to 1, 2 and 4 processor socket servers. Draft 4 represents several positive steps forward in the development of the program requirements. While IBM continues to have serious concerns with the idle criteria proposal, limiting that criterion to one and two processor socket servers appears to be workable. IBM strongly supports establishing processor level power management enablement as a criterion for 4 processor socket systems in lieu of an idle criterion.

IBM is concerned with the EPA's inclusion of blade servers in Draft 4. Attempting to set an idle criterion for blade servers for the final draft intended to go into effect on May 1, 2009 is not feasible. In addition, we have provided technical justification below as to why an idle criterion is not appropriate for blade servers. If EPA wants to set a power management criterion for blades, we encourage EPA to require enablement of processor level power management as the criteria for blade servers. This criterion would be acceptable for inclusion in a set of requirements that would go into effect on May 1, 2009.

IBM continues to encourage EPA to recognize that computer servers are, as a whole, significantly different from workstations. While there are some similarities between a workstation and a 1 processor socket computer server, the two quickly diverge as there is an increase in the number of processor sockets and associated cores, managed by the computer server. As the number of processor sockets and cores increases, so does the types and complexity of the workload that will be managed, the number and variety of supportable components, and the range of capacity. The various cost/benefit ratios of all of these factors that must be considered. Thus, computer servers require more planning, consideration and selection of options – blade versus rack, virtualization capable, software configurations, and the type of peripherals – as they provide a much broader range of choices and possible configurations than are present for a workstation or laptop. In addition, the work delivered per unit of energy expended (as the sum of active and idle power) is significantly improved with virtualization and increased system utilization. Thus, as we have stated previously, server efficiency has to be measured in the context of power use at idle, the ability to keep the IT asset fully utilized, and the serviceability and reliability requirements of a given application. This is a much more complex set of considerations than those required for a workstation or laptop.

Specific Comments:

Line 253: Definition of a Computer Server:

The statement: All processors have access to shared system memory and are independently visible to a single OS or hypervisor requires clarification. It is not always true; systems are available in which separate operating systems or hypervisors can access or a subset of the processors or cores on a system. It is recommended that you replace "single" with "one or more".

Line 265: Remove the word "to" from the sentence: "...designed for technicians..."

Line 309: Managed Server definition: From IBM's perspective, the descriptor "Managed Server" is not the appropriate term to designate systems with the capability to have redundant power suppliers. Mission critical, highly managed environments can also use servers with service processors that do not have redundant power. In many cases, the server (node) is the redundant element and the workload moves to other "nodes" upon failure. It is IBM's opinion that the term "High Availability" is a better designator to distinguish systems which are capable of having redundant power supplies.

POWER SUPPLIES:

Power Supply Definitions:

Line 394 and 401: Watt Limitation for additional outputs on a single output power supply.

The limitation of 20 watts on for the additional outputs on a single voltage power supply may not be workable for large power supplies. We recommend that the limitation be set at 20 w or 2% of the power supply rated wattage, whichever is greater. Our concern is that there may be situations on the larger power supplies where the demand on the additional output will exceed 20 W. Allowing a 2% range should address this concern.

Replace "Primary Voltage" with "Main Voltage" and "Secondary Voltage" to "Additional Voltage". Primary and Secondary voltage have specific meanings under the Safety Agency definitions which may cause confusion.

Page 9, line 435: The word "quantity" is spelled wrong.

Page 9, line 438: The word "within" is spelled wrong.

Page 9, line 445: Removed the "s" form "configuration".

P.9 PRODUCT FAMILY DEFINITION:

Line 435-436: Processors should be grouped by model and power, not speed. Typically, processors are rated at a given wattage and then there are speed sorts for that power type. A family should be able to include multiple speed sorts where the socket power for the processor is the same.

Line 438: Add ..the relative numbers *and the manufacturers* of these components may vary within families.

As discussed below in item (a), OEMs source several manufacturers for a single component such as a DIMM or a hard drive to insure low cost and availability. EPA needs to recognize that this is part of the business process and allow OEMs to manage their supply chains, their supplied components and the power characteristics of those components to insure consistency in the testing and qualification of servers under the ENERGY STAR® criteria. EPA should acknowledge that here may be multiple manufacturers of a specific component.

Line 449 to 464: For a product family, once the computer server qualifies under the idle criteria at the minimum and maximum configuration then EPA should accept that the other models are qualified because:
a. All components for a given model are built to a specific power spec and the variation is, by specification, designated to be no more than +-5% across any number of manufacturers. Variability will exist both due to intrinsic manufacturing differences and the use of different suppliers. This yields a range of power values. For example, all processors of the same type from the same wafer and sorted to the same speed do not have the exact, same power characteristics. This problem confounded the SPEC Power committee to the point of inaction.
b. The idle power curve for a given set of configurations will look like a power supply efficiency curve as the complexity of the configurations increase. The idle power for a given configuration will vary some around the best fit curve for that product family, but in general the two end points will appropriately define the model and processor combo and assure that overall that product family delivers the best overall energy efficiency performance. The intent of using the product family designation is to simplify the testing, provide an easy to understand ENERGY STAR® designation for the customer, and identify the specific product families that provide the lowest power output at idle.
c. Requiring that each individual configuration meet the base plus adders idle power requirement for that configuration complicates ENERGY STAR® designation through the Value Added Retailer (VAR) delivery outlet. To the first order, the value proposition of a VAR's offerings is their ability to meet the customer's needs by modifying or adding equipment to meet the customer's application requirements. ENERGY STAR® should allow the VAR to purchase the ENERGY STAR® model that satisfies their customers' requirements and then add the necessary peripherals that are required to support the customer's operations. The VAR's ability to deliver this service is restricted if they are required to stay within the defined product family data sheet configuration. The other option proposed by EPA, to have each VAR sign on as an ENERGY STAR® partner and qualify its "modified" products will create a significant data management burden on both the VAR and the EPA. There are 1000s of VARs in the marketplace, which will create a multitude of product registrations while bringing no real value to the ENERGY STAR® brand or to customer understanding of ENERGY STAR®. The more efficient, understandable approach is to qualify the base OEM model with key, required components and allow the VAR to add its specialty peripherals and services to that base, ENERGY STAR® qualified product.

As we commented to the Draft 2 and Draft 3 server criteria proposals, the idle criteria as defined by EPA is biased toward low power processor systems, which are not necessarily the most effective systems for server virtualization and system reliability and serviceability considerations – considerations which are very important to many data center operators. We continue to encourage EPA to address minimizing server power use during periods of no work by setting standards which reward the presence of power management functions as has been done in the criteria for 3 processor socket and 4 processors socket systems.

P.10 Line 498: If criteria for blades are not included in Tier 1, blade servers need to be referenced in the exclusion section.

P.11 BLADES:

IBM strongly encourages EPA to not issue idle criteria for blade servers in the Tier 1 final draft proposed to take affect on May 1, 2009. If EPA wants to include blades in the final draft of the Server Criteria, IBM recommends that the EPA use the requirement that the blade servers ship with processor power management enabled as the criteria for blade systems.

After looking at the proposal carefully, we think that trying to create an idle power specification for blades has similar complexities to those we highlighted for 4 processor socket systems. In addition, different blade manufacturers have different power supply delivery methodologies and different levels of shared componentry, such as fans and network modules, making meaningful comparisons difficult. We have several specific concerns:

- Different manufacturers use different power domain strategies on their systems which will affect the 50% full blade chassis measurement. For example, IBM uses two power domains, driven by up to 2 power supplies each, while other manufacturers deliver power through a single rail with 4 or more power supplies providing depending on the number of installed blades and the desired redundancy. Thus, a half full blade chassis will have different distributions of shared infrastructure and power supply deployment depending on the system architecture.
- Different manufacturers use different shared systems (network, storage, and fans) which again will vary where the overhead is in the system and the idle performance for a single blade and half full chassis.
- It is difficult for manufacturers to get a half full or full rack for testing.
- The combination of peripherals that you can get in a blade rack with 10, 12, or 14 blade slots becomes almost impossible to manage, similar to a 4 processor system.
- An empty chassis/single blade is not representative of the system, as you need more than 4 blades in the system to make it more economical than rack servers.
- The range of components that can be installed in a blade multiplied by 14 blades in a chassis, make the creation of a "standard configuration" for the product almost impossible and creates difficulties in the doing a calculation of base plus adder because of the complexity of the system.

Page 12, Section 3A: **Power Supply Efficiency Requirements**:

Line 575: Remove the Power Factor (PF) Requirement for 10% load or reduce the factor at >1000 Watts to 0.75: A power factor of 0.8 at 10% load is technically difficult to achieve and to reliably measure. Looking at the PF limits for the 0 to 500 W and 500 W to 1000 W power supplies, both of which are set at .65, the jump to a PF of .8 for >1000 W systems is not justified. In addition, we continue to contend that a power supply will not operate in this range for any meaningful period of time, particularly for power supplies over 1000 W. This is consistent with our previous comments in Drafts 2 and 3 and we continue to encourage EPA to either eliminate the 10% requirement or reduce the PF requirement to 0.75 for >1000 W power supplies loaded at 10% of capacity.

p.17 Additional Power Allowances Note, Line 742 to 752: Performance Based on Cores:

The best technical reason for sticking with sockets is that they are much smaller number, are well-defined physical entities with physical power inputs, and are not likely to explode in number per system over the next few years. The exact definition of a core varies from architecture to architecture and design to design. In addition, there is shared logic used by all of the cores on the chip in the socket and the amount can be different depending on the design, affecting the core, but not the processor socket, power characteristics. There may or may be a shared L3 or L4 cache. Cores may or may be able to run at different frequencies whereas it is now very reasonable to think that sockets can run at different speeds and voltages.

Testing Requirements:

The requirement to test a blade chassis with a single blade is an unreasonable requirement. A client will not run a blade server with a single blade; you would purchase a 1U/2U rack server because you would avoid the inefficiency of all the support hardware sized for a fully populated chassis. In addition, the test point is not representative. The test point will be distorted by the power supply efficiencies found at less than 10% loading.

Testing a half full chassis also introduces inaccuracies, given that different architectures have different shared loads (network, fans, and other) as well as different blade to power supply support ratios. Thus it is not possible to get data that allows an "apples to apples" comparison between different manufacturer's products.

Manufacturers test single blades utilizing a test fixture and represent the chassis base power based on specification requirements and manufacturers representation of the chassis power requirements. Chassis systems are not typically tested. Power use of a chassis populated with "x" blades is calculated by adding the chassis power to "x" times the single blade power.

If EPA specifies an idle criteria for blade systems, which IBM does not recommend, then EPA should limit the test case to a single, full chassis with blades and should allow for the chassis idle power to be calculated by providing the idle power for the chassis and calculating the power associated with the blade systems by taking the blade idle power from a test fixture and multiplying it times the number of blades in the chassis. The reported idle power should be a per blade number, computed by taking the total chassis/blade idle power calculation divided by the number of blades in the chassis. This is the only methodology that we can identify which provides a reasonable basis for comparing blade systems.

Even with fully loaded machines, this approach is not ideal. In addition to the fact that manufacturers make different choices about what shared resources are bundled into the chassis or distributed among the blades, they may also provide the capability from the external or network environment. Some blades are booted from network based storage or the network switching may be provided by an external rack. In these situations the energy consumption is either lost in the evaluation or assigned to the wrong entity.

Data Availability

IBM will not be able to provide a meaningful data set on blade server idle power for the three listed configurations to EPA on 3/20/09 as requested by EPA in the Draft 4 document. The physical resources and testing time required to collect the data were not available in this time frame. In addition, it will be very difficult for us to provide the data for all of our current products, because of the number of machine types and models that we currently market.

In addition, blade testing is done using a single blade attached to a fixture. Manufacturers do not routinely test full chassis units.

In discussions with EPA during the comment preparation period, the question was raised how a manufacturer would test for idle to complete the prescribed data sheet. The idle and maximum power for a given chassis/system configuration would be calculated by adding the chassis power with the idle and maximum power for each blade times the number of blades in the chassis. This is the only reasonable and economical way to calculate the information for the many families of products.

If EPA were to proceed with establishing idle criteria for blades, an approach we do not endorse, then they should allow the manufacturers to generate data on the different system configurations as described in the preceding paragraph.

System Complexity:

A blade server is a multi-node system, with each node being capable of being configured to do a specific type of work or provide redundancy for and support for a variable workload and to support virtualization and higher system utilizations (maximizing the use of your hardware and facility investment while optimizing the work completed for unit of energy applied to the system). That is the attraction and benefit of using a blade system. The multi-node nature of the system creates an extreme amount of complexity in analyzing a system for idle power.

In discussions with Arthur Howard, Arthur indicated that it was EPA's view that blade systems compete with one and two processor rack and tower systems. In fact, blade systems are multi-node, highly virtualizable systems which compete more directly with the larger 4 processor socket and larger systems. Rack servers are the better energy solution up to 4 racks vs. 4 blades, but then blades take over. Also, as a multi-node, highly virtualization capable systems, the blade servers are much better positioned to virtualization workload and thus are more comparable to 4 processor socket and larger server systems.

Given the analysis and discussion above, **IBM recommends that EPA use the requirement that the blade servers ship with processor power management enabled as the criteria for blade systems.**

After looking at all the variables and complications associated with trying to measure idle in a blade chassis, IBM recommends that the best approach to bring blades into the specification in a timely manner will be to use the enablement of the processor level power management function as the criteria for blades the same as you did for the 4 processor systems. That enables blades to be include in the next spec release, eliminates the complexity associated with different architectures and the multiple processors available in blade chassis.

P13 Line 618 to 620: Note: A 2 processor socket system, with only one processor socket populated, is unlikely to meet the ENERGY STAR® one processor socket idle criteria. However, a VAR may purchase a two processor socket system with one socket populated. As long as it is populated with a processor with comparable rated power (90 W) and meets the other requirements of the data sheet the VAR should be able to qualify that computer server under the OEM's ENERGY STAR® qualification.

Line 632: Values in Tables 3 & 4 should be given with a first decimal place.

Line 633: Power allowance for additional power supplies. The definition for Additional Power Supplies should be modified to read "applied for installed power supplies above the base number of power supplies required to operate the server without power supply redundancy" or the equivalent. You may want to put this definition into the definition section.

p.13 Table 3: Category B; "Dual" should be replaced with "Single".

p.13 Table 4: The "I/O Devices" Header for the section should have the (Greater than 1Gbit) removed, as there are device references below it that are less than 1GBit.

p. 13 Table 4: For "Fibre Channel or Infiniband, the criteria should be changes to read "5W per active port".

Page 14: Computer Servers with Greater than Two Processor Sockets (3S and 4S)

IBM Comments to "ENERGY STAR® Program Requirements for Computer Servers Draft 4

IBM strongly supports the EPA proposal to require enablement of processor level power management for 3 processor socket and 4 processor socket systems in lieu of an idle criterion. The complexity of four processor systems, the range of available configurations and the fact that 4 processor socket systems are intended to run virtualized workloads which typically requires higher powered processors and a broader range of memory and peripherals make enablement of processor level power management functions the correct criteria for this class of computer server.

Line 795: Need to include "socket" after "processor" in this line.

DATA MEASUREMENT AND OUTPUT REQUIREMENTS:

Line 882-884: EPA should not be referencing an unpublished standard in the computer server requirements. The last sentence should be removed.

Line 885: Sampling Interval: The power supplies we purchase perform their power measurements to do a synchronized time period (an average generated at 30 seconds intervals) rather than as a rolling average. We need the rolling data requirement removed. We're providing a 30 second average every 30 seconds. As most systems can't afford to poll much more often than 30 seconds (the data collection becomes overwhelming and chews up significant register space and starts to slow things down), provide a synchronized time period works. Also, you do not get significant average movements that won't be identified/tracked on a 30 second interval.

Recommended Text: Sampling Requirements: Hardware polling rates of the embedded sensors must meet a minimum of one sample per second. A default average of 30 seconds is recommended. The data can be averaged either as a rolling average or in a synchronized interval.

Line 899: The Input power measurement accuracy should set the +-10% accuracy standard for power supply loadings of 15% or greater. Below that, the accuracy of the measurement deteriorates and approaches 100% or more of the measured value at low loadings.

Line 891: Processor utilization needs to be qualified (or assumed) to be "for the methods and algorithms employed for the processor architecture" and should be "+/- 5% of maximum" rather than "+/- 5% of the measure" because the "idle" run may show 0.05% CPU and 5% of that will be only +/- 0.0025%, which nobody can truly achieve.

DATA SHEET:

Power data provided for the minimum, maximum, and typical configurations where the power use is not being used to assess conformance with a criteria should be able to be calculated using a power calculator or configurator which a manufacturer provides to its customers to calculate the power use of a chosen configuration.

Performance Metrics: On reporting the benchmark, you should add a field to allow reporting of which configuration in the family was tested for the benchmark. The manufacturer should be allowed to pick the configuration within a given product family on which the benchmark is performed.

Under System Characteristics:

1. Change “Available PCI or PCIe Slots” to say “Available Expansion Slots” to cover future types of expansion slots.
2. Include a row that shows the minimum and maximum number of power supplies.

Under System Configuration:

1. Show the power supply power factor at various load levels.

Under Power Data:

1. Need two columns for data for 1 and 2 processor socket system reporting; one for the minimum and one for the maximum configuration.
2. Need a data line in which the minimum and maximum configuration peripherals are defined.
3. The data sheet requirements need to call out that the power measures are only for the server portion of the configuration.

Benchmarks:

Replace the term “benchmark” with “a recognized measure for quantifying performance or performance/watt for a server, including both industry-standard benchmarks that are controlled by independent consortia and architecture-specific benchmarks that are controlled by owning companies.”

Performance benchmarks are focused on delivery of the highest throughput possible, making the configurations less than optimal from an energy consumption perspective. Many public benchmarks require full publication of a result in order to use a fraction of the information in publicly available material. This presents a challenge for documenting the power used in a performance benchmark. If the configuration is optimized for power, it may deliver throughput results that appear to be less competitive than peer results in the industry. Replacing the term “benchmark” will allow for well-known performance measures like the SAP-SD benchmark or VMmark benchmark, associated with specific vendor software, or the Large System Performance Ratio (LSPR), or the Commercial Performance Workload (CPW) associated with specific vendor hardware architectures, to be used for the Tier 1 solution. Once more industry benchmarks that are focused on performance/watt become available, the Tier 2 specification may focus solely on these, but they are not available for the Tier 1 time frame.

While IBM appreciates EPA’s rationale for including the requirement to publish a benchmark on the proposed ENERGY STAR® data sheet, it’s important that EPA understand that providing this data is as likely to cause confusion as it is to help customers identify the most energy efficient solution for their application. Because each manufacturer will pick the benchmark and the configuration on which it will be run, the data sheet will provide a single power and performance point within a continuum of possible points for that model. It is highly likely that the conditions under which the data point is generated will vary from manufacturer to manufacturer, making the data of limited value to the customer. While a benchmark can be generated and power use recorded, providing a meaningful power and performance criteria is the task ahead for Tier 2.

The idle measurement data sheet should have columns for the power meter manufacturer, model, software used (if not standardized by the EPA), the number of meters, the number of power cords and any power-strips or other connectors used. In looking through the sheet, I did not see any columns for this critical information.

TEST CRITERIA:

Lines 1005 to 1016: The referenced note would imply, contrary to what is written in lines 990 to 1003, that the laboratory should be “independent”. Lines 990 to 1003 allow self-certification. It is important that OEMs have the flexibility to incorporate any testing into the overall product testing activity to minimize costs and simplify management and review of the data.

APPENDIX A: Test Procedure for Determining the Power Use of Computer Servers in Idle

Line 1266: If you are allowing 50 Hz, you should also accommodate 208 V.

Lines 1085 and 1359: The requirement to have the primary drive not in power-managed mode may be in conflict with the requirement to test as shipped, if the default is to have all drives in power-managed mode. Need to specify in 1085 that the exception is that power-management may need to be turned off on primary disk drives as defined in Appendix A, clause B.2.d.

TIER 2 Requirements:

IBM recommends that the Target Effective Date for Tier 2 be extended to 12/31/2010. SPECweb, SPECvirt, TPC benchmarks will not be available until late 3Q09. There will not be sufficient time to develop statistically valid data by the time the Tier 2 draft will need to be developed to settle on the appropriate 25% line for these benchmarks.

IBM is concerned that EPA appears to be dictating the outcome of the Tier 2 specification in the absence of an agreement on a criteria based on performance and power metrics. Given the range of discussion involved in the Tier 1 specification and the many changes that were made between the first and fourth drafts, changes that on the whole contributed to a more workable, robust set of criteria, it is inappropriate for EPA to attempt to specify an outcome to the discussion which will take place over the next 18 to 21 months. While we are not going to provide a detailed response to the items 1(b) through 4 of the Tier 2 proposal, we will note that we have in our comments to Draft 2 and 3 we have objected to the implementation of idle criteria for all servers, particularly more complex models, and to the Net Power Loss approach for power supply requirements. In both cases, we proposed workable alternatives which we believe provide a workable methodology to further EPA's and the industry's goal of providing more efficient IT equipment enabling improved energy utilization and efficiency in the data center.

Idle Test Procedure:

Line 1266: With regard to the test conditions, the temperature and relative humidity should be identical to the new ASHRAE standard for data centers since that is representative of the best current thinking in the area. IBM recommends that EPA copy what ASHRAE has in its documents.

Line 1375: Use the term "UUT" instead of "Computer Server" for clarity. The idea appears to be to set up, shutdown and start-up again and then measure. Replace the term "switch on" with "boot up" for consistency.