



October 23, 2009

Dear EPA,

Thank you for providing us the opportunity to comment and share our insights early in the ENERGY STAR development process. Please find enclosed Intel's feedback on the ENERGY STAR for Servers Specifications preliminary draft for Tier 2, dated 9/23/09

Intel remains committed and supportive of the US EPA's efforts to define energy efficiency goals and targets across the spectrum of computer products including the preliminary draft of proposals for ENERGY STAR for Computer Servers, Tier 2. We hope our input and our collaboration with industry stakeholders continue to benefit Tier 2 specification development. We appreciated the direct discussions with the ENERGY STAR team at the recent workshop that resulted in clarifying the challenges and targets for Tier 2 of the program. Intel is actively involved with industry stakeholders on the plans for an Energy Efficiency Rating Tool for Tier 2 and we anticipate continued dialog with the EPA and industry organizations as we work to better define the scope and criteria that would be used in Tier 2. Please feel free to contact us if there are areas where we may be able to improve these interactions.

We continue to work with our industry colleagues in Standard Performance Evaluation Council (SPEC), Climate Savers Computing Initiative (CSCI), The Green Grid (TGG), IT Information Council (ITI), Alliance for Telecommunications Industry Solutions (ATIS) and Storage Network Information Association (SNIA), in addition to supporting the ENERGY STAR for servers program to deliver increasing energy efficiency.

If you have any questions please feel free to contact myself or Henry L Wong, henry.l.wong@intel.com.

Sincerely,

Lorie Wigle

General Manager

Eco-Technology Program Office

Summary

Enclosed, we have provided commentary for the major sections highlighted in the ENERGY STAR preliminary draft document, dated 9/23/09. In the appendix we've included detailed comments and editorial suggestions for particular lines in the draft. We recommend holding subsequent reviews on each topic to help provide any additional clarifications prior to EPA's next draft publication. The topics covered include definition and scope, active mode and idle, the Power Performance Data Sheet (PPDS) and Qualified Product Information (QPI) forms, energy efficient Ethernet, and real time monitors.

Given the aggressive schedule for Tier 2 for the server specification, we believe the priority should focus on development and incorporation of SPEC's Server Efficiency Rating Tool™, limiting the scope to 1S-4S general purpose pedestal, rack and blade servers, and addressing the issues highlighted by the system manufacturers with the Tier 1 specification. Intel fully supports SPEC's development of the server efficiency rating tool. We consider the development of a metric that incorporates performance and energy consumption to be the key milestone for the Tier 2 ENERGY STAR for Server specification. In anticipation of the challenges and aggressive schedule, the specification development may require up to 6 month delay to EPA's target effective Tier 2 date of October 2010. A 6 month push in the schedule would support the minimum amount of time needed to develop, test, evaluate, and prove out SPEC's Server Efficiency Rating Tool, along with data collection, documenting final testing procedures and incorporating the methods into the Tier 2 specification.

We recommend that Energy Efficient Ethernet (IEEE 802.az) not be included in the Tier 2 server specification since balloting on the IEEE specification will not close until the end of 2010 and would still require interoperability validations. It is an excellent candidate for future consideration, once the capability is widely available in shipping products.

The section commentary also describes the purposed separation of the information in the PPDS verses the QPI form to better communicate information to IT managers. We have provided recommendations on the CPU utilization monitors and thermal monitors to match the accuracy targeted to the use of the dynamic real time information. Finally, although we have not included specific comments on power supply efficiency, we are in full support of the CSCI, ITI, and The Green Grid recommendations to forego the Net Power Loss method and maintain the current power supply efficiency methods.

We hope these comments and recommendations will be useful to EPA's plans and targets for the ENERGY STAR for Servers Tier 2 specification. We welcome the opportunity to discuss these topics further as you create the initial draft of the specification.

Commentary by Section

Definitions and Scope

We recommend retaining the current scope of 1S to 4S servers. Per IDC World Wide Quarterly Server Tracker (Q2 2009 release), 98.94% of server shipments are 1-4 socket general purpose servers in pedestal, rack and blade form factors.

| % Servers Shipped | Number of sockets supported by server |
|-------------------|---------------------------------------|
| 22.60% | 1S |
| 71.44% | 2S |
| 4.90% | 4S |
| 0.48% | 8S |
| 0.53% | 16S |
| 0.04% | 32S |

Source: IDC World Wide Quarterly Server Tracker (Q2, 2009 release)

Extending the scope beyond four-socket machines provides diminishing returns, will increase the complexity of the Rating Tool and have a disproportionate impact to the schedule. Many >4 socket servers, resilient servers, and HPC systems contain features/characteristics that cause them to operate under a different energy profile than general purpose computers. Hence our recommendation is to keep the priority focus on 1-4 socket general purpose servers in pedestal, rack and blade form factors, which covers the vast majority of servers sold.

To maintain the priority and focus on 1s to 4s servers, we also recommend the following clarifications on definitions of HPC systems and resilient servers.

High Performance Computer (HPC) Systems

We agree that HPC systems form a category that deserves 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 recommend consolidating the input from industry stakeholders to finalize the definition of this class of system.

Resilient Server

A Resilient Server is designed with extensive RAS features, including error self-correction to ensure data resiliency and accuracy. Resiliency, RAS, self-correction, and data accuracy features are integrated in the micro architecture of the CPU and chipset functions in a Resilient Server. Resilient 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 features. It is recommended that resilient servers be placed into a different category because of this reason. A resilient server should be a system that contains all or a significant number 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

Active mode and Idle Specifications

Intel fully supports the development of the SPEC Server Efficiency Rating Tool (SERT™). SPEC has already started the tool development supported by a broad base of server vendors. The testing procedures from SPEC ensure data is open, transparent, and repeatable. The tool is expected to include power and performance measurements at multiple load levels including system idle. Therefore, we recommend removing idle as an independent criterion.

In regards to the description in EPA's document on “Desired Characteristics of an Efficiency Rating Tool”, we agree with the targets indicated and offer these additional comments.

1. We agree that the tool be technology-neutral/architecture agnostic.
2. We agree that the tool must have “limited barriers to implementation”. Specifically the tool must be easy to use and require minimal equipment and expertise. In contrast, most industry standard performance benchmarks are published by large manufacturers of server equipment using highly trained engineers. We recommend that the EPA place high importance on this characteristic.
3. We agree in the importance that the tool evaluates characteristics of a variety of end-use scenarios. This requirement will need to be balanced against rating tool complexity, ease of use, and development time. This could have the unintended consequence of limiting hardware configurations that can be classified by the rating tool. Priority should be placed on the Rating Tool's ability to measure a product family from a minimum to a maximum configuration, instead of testing only a limited number of configurations.
4. We agree in the importance of a comprehensive assessment of the server at multiple load points because the power management features called out in Table 5 of the Preliminary Draft Version 1.0 Tier 2 Document is mostly active when the server is less than 100% utilized. We recommend an equal distribution of measurement points across the load line from idle through 100% utilization.
5. We agree that the rating tool should be “Transparent and Standardized” with one exception: the server operating system should not be standardized or held constant through the duration of the specification. There are multiple issues with standardizing an operating system. First, not every operating system supports all system architectures. Second, even an operating system that supports a system architecture may not support revisions or new entries to that architecture without patches or service packs. As an example, the introduction of non-power-of two cores (6 cores) in Intel's Xeon 7400 series processors required changes across operating

systems and applications just to get software to install properly. Finally, server power management is achieved by the interaction of an operating system with hardware. As new power management technologies are introduced, the full energy efficiency of a server may require a new operating system to be used. We recommend the EPA adopt a policy of allowing server vendors to test with any of the targeted operating systems for the server.

Intel endorses EPA's preferred approach, Option C, as a path toward an active mode efficiency rating tool. Intel also endorses the current work in SPEC to produce a server efficiency rating tool as the best approach to developing this tool.

Intel recommends the Rating Tool include both active idle and active power instead of requiring a server to pass both a revised Tier 1 Criteria for idle and receive a passing score from the Rating Tool. Below are justifications as to why a separate idle classification is not beneficial:

1. It is not possible to have a revised Idle Criteria without having a revised series of adders. 4 socket systems have more complex system configurations than what is used or is possible in the 1 socket or 2 socket servers covered by the existing Tier 1 Criteria. Each unique configuration option not present in all servers in the category will require an adder, significantly increasing the complexity of an idle metric calculation.
2. The existing Tier 1 Criteria for idle is an operating system idle measurement, while deployed servers have loaded software stacks. The Rating tool can instead measure active idle, which is more representative of the idle for a deployed server.
3. Adders limit technology innovation and system integration. For example, if there isn't an adder for Integrating RAID into an HBA, or use of memory as a disk cache the additional system power could prevent the server from passing the idle criteria, even though the server is highly energy efficient at all active load points.
4. Eliminating a separate metric for idle removes the issue of whether to have both an active power and an idle power adjustment for multiple cores.
5. Within the anticipated life of a Tier 2 specification we expect cores per socket to continue to increase to levels where additional PCIe lanes/slots, memory channels and DIMM slots will be required over today's 1 and 2 socket servers resulting in a slowdown of the rate at which idle power for a server will decrease. A single rating tool covering both idle and active should be more extensible to future ENERGY STAR specifications, than separate idle and active targets.

Intel does not endorse paths A, B, or D. The Rating Tool needs to balance the characteristics of the multiple use scenarios in a general way to avoid the requirement of too much specialized H/W in the system under test, which would not be needed by the end user of the server. Having separate workloads for each of the use scenarios goes against the goal of limited barriers to implementation, because of the specialized expertise, time and inordinate cost required to run multiple industry standard benchmarks. This is certain to prohibit small manufacturers or VARs from offering Energy Star Certified Servers.

Approaches A, B, and D do not balance these characteristics and contradict the EPA's stated goal of "limited barriers to implementation":

1. The required configurations for existing performance benchmark workloads that have been extended to power are not going to resemble the configurations being purchased. This is in direct conflict with testing the configuration as shipped. If these benchmarks were run on platforms configured as shipped the performance would be

well below published results for the benchmark. Some benchmarks would only exercise the system to single digit percentage of utilization because the server wouldn't be providing adequate resources to run the benchmark. This also fails the criteria of Comprehensive Evaluation.

2. These workloads are complex requiring specialized expertise to run the workloads.
3. There are expensive fees for the use of the benchmarks, including auditing fees for TPC benchmarks which are in the tens of thousands of dollars.
4. There are license restrictions on the publication of benchmark results from the ISVs (Independent Software Vendor) owning the software stacks. For example you cannot publish a database benchmark without the permission of the database ISV. ISVs do not want results published that portray their software in a bad light. This would restrict ENERGY STAR publications to only suitable configurations for "best" performance. The issues of doing this were listed in point 1 above.
5. Following Approach D and blending Approach C with Approach B does not eliminate any of the problems associated with Approach B highlighted in the points above

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

The Power Performance Data Sheet and Qualified Product Information forms serve different purpose. We believe the PPDS offers a common communication format to consistently describe power and performance characteristics for a family of products that IT managers can use to evaluate against their target applications and usage models. The QPI form is the report documenting the compliance parameters of the product or family of products.

We agree that both sets of data should be available and transparent for review.

To improve the applicability of PPDS to IT managers, we recommend it only contain the following items:

- One performance benchmark result along with the average peak power consumption during the run and the hardware configuration used to obtain the result. Note: The benchmarks results would not be tested in the as-shipped configuration.
- Power required during system boot and initial installation.
- List of server power management capabilities
- Idle power (for lowest and highest hardware configuration in the product family)

It is recommended the QPI form contain the following items:

- Minimum and maximum hardware configurations in the family that the manufacturer is documenting as compliant to the ENERGY STAR criteria.
- Indicate the calculated and measured results demonstrating that all system options between these minimum and maximum configurations would also be ENERGY STAR compliant.

Energy Efficient Ethernet

The industry standards for energy efficient Ethernet (IEEE 802.3az) are currently being defined and approved through IEEE. The specification is expected to be finalized by end of 2010. Even after it is final, vendors need time to accurately implement the specification. Features are just being enabled in silicon through a variety of vendors first with 1Gbe and then 10Gbe. These features can only be fully utilized after thorough interoperability evaluations between vendors, networks, hosts and clients. There are also integration opportunities that will depend on the maturity of the technology, compatibility and interoperability assessments in both the active and power savings techniques. Intel, as a

leading supplier of networking devices, and as one of the firms most active in defining and promoting this standard, believes that technology and standards will require interoperability testing and potential silicon or standard revisions up to 2012. This is especially the case for 10GbE. The higher speed and random packet activities incurs a much higher interoperability risk for 10GbE devices. We recommend that the EPA not include this requirement for Tier 2; but, consider for a future revision pending maturity of the technology and standard.

Real time System reporting

There have been no improvements to the capability of determining CPU utilization that account for advanced features such as multi-threading and dynamic voltage and frequency scaling. We recommend that the requirement of reporting CPU utilization remain, but without an accuracy requirement.

The thermal monitor rolling average is problematic both in use and collection. The value of the information is limited if it's averaged over a 30 second and the rolling average calculation is problematic for coding and storing. We recommend a +/- 2C accuracy with a sample rate of every 5 seconds. Environmental temperature and airflow controls can monitor and determine the proper actions to maintain set point targets.

In conclusion

Intel appreciates the opportunity to provide the EPA with the comments and recommendations for the preliminary draft of the ENERGY STAR Program Requirements for Computer Servers Tier 2 specification. We hope you will include these considerations in the draft specifications later this year.

Appendix

Line 346. Change Label from “Server Utilization” to “Server Processor Utilization”. In addition, we recommend changing the text to read “processor activity relative to its maximum frequency not accounting for any temporary bursts of clock frequency” to make it consistent with the wording in section 3D line 638.

Line 401. Intel suggests another definition be defined for Tier 2 for “Active Mode”.

Line 414. According to IDC data, 1S – 4S servers (blades, racks and pedestals) comprise the vast majority of servers sold. >4S servers, server appliances, and fully fault tolerant servers are an extreme niche category; therefore we recommend not including them. Per IDC World-Wide Server Tracker as of Q2 '09, blades represent 14.8% of the server market in terms of form factor so it is recommended that blades be included. As for multi-node servers, this class of servers is generally sold into high capacity data centers where customers are already acutely aware of the server power consumption so adding an ENERGY STAR rating for multi-node servers provide limited additional value

Line 418. Under “Options for Blade Servers”, Intel recommends the 2nd options “Extended development of the requirements for blades under the full Tier 2 development schedule. The reasons for this are 1) to allow inclusion of Active Mode energy efficiency; 2) to allow the EPA and vendors sufficient time to determine classification and measurement techniques and 3) make it easier for customers and vendors to adopt because an interim specifications with changes of this magnitude would not be fully comprehended by customers

Line 466. Recommend the system be defined by the number of sockets server is capable of supporting. In our experience most users who purchase partially populated systems intend to fully populate the system at some future point.

Line 471. The compensation for the number of cores should be revisited after the ability of the server rating tool to accommodate performance gains of additional cores is understood.

Line 490/Table 4. Additional Idle Power Allowances. The compensation for power adders should be revisited after the ability of the server rating tool to accommodate different hardware configurations is understood.

Line 530. In line with our general comments above, we believe idle power is an inadequate measure of server energy efficiency. A metric which includes the performance of the system and which takes account of power measured at various load points is the only way to characterize system energy efficiency.