September 11, 2017

To: Ryan Fogle, EPA Manager, ENERGY STAR for IT and Data Center Products;
John Clinger, ICF International

Re: **ITI Comments on ENERGY STAR Computer Servers**

ITI appreciates the EPA's efforts in completing the second draft of the ENERGY STAR Product Specification for Computer Servers Version 3 and incorporating many of the comments provided by industry in response to the Draft 1 document. ITI, with technical support from the Green Grid SERT Analysis Working Group (SERT WG), has continued to work on (1) the analysis of the SERT dataset to develop and evaluate options for establishing active efficiency/idle thresholds to designate more energy efficient servers, (2) assessment of the impacts of SERT V2.0.0 release on the measured active efficiency scores and proposed thresholds, (3) development of a system performance idle adder to account for higher idle power values associated with higher performance servers, (4) the collection and analysis of memory and storage idle power data to set appropriate idle power adders for both certification testing and assessment of shipped product compliance for those components, and (5) modification of the scope and definitions to address technology developments that have occurred over the past twelve months. ITI and Green Grid continue to work with EPA to collect and maintain a representative SERT dataset to enable an effective assessment of the options available to use the SERT active efficiency metric to differentiate the energy efficiency of server products.

ITI wants to reiterate its key message regarding setting energy efficiency requirements for servers: energy efficiency, as defined by work delivered per unit of energy consumed in the data center, is a function of both the work capacity and power characteristics of the individual server products and the workload capacity and energy footprint of the system of servers required to execute a given workload in a data center or in the office environment. Our position is that a threshold solely based on an active efficiency metric best balances the assessment of the individual server systems while delivering a lower power use outcome in the data center. We have reviewed the EPA proposal for a combined active efficiency and idle power threshold and find that to be a major improvement over an idle power only threshold. Should EPA continue with its intent to move in the direction of combined active efficiency and idle power limits, there are several improvements that need to be made to the proposed methodology:

1. A separate active efficiency limit should be set for one socket and 2 socket servers. There is a significant difference in the distribution and maximum values of the active efficiency scores (see Figure 2) which justify separate active efficiency limits.
2. The idle power methodology should include a system performance adder which provides an additional idle allowance based on a multiplier assigned to the geometric mean of the 100% CPU worklets score. This adder is important as higher performance system use processors with higher socket power and idle power demand and more circuitry to support higher memory capacities and more components. The proposed adder sets a cap of 40 Watts on the difference between the adder for the low-end and high-end configuration which results in an average adder across the server products in the data set of roughly 25 watts.

3. An increase in the memory adder to 0.175 W/GB or creation of separate adders for 4 and 8 GB/watt DDR4 DIMMs. The 4 GB DIMMs have a maximum idle power of .223 watts/GB and 8 GB DIMMs at 0.175 watts/GB (table 2). As these two DIMM sizes are likely to be used for SERT testing, it is necessary to set limits which are representative of the actual, measured idle power of those two DIMM sizes.
4. Modifications and additions may also be appropriate to the idle adders for Auxiliary Processing Accelerators (APAs), storage devices and Input/Output (I/O) devices. ITI plans to provide additional data and input on these adders in its next submission on Oct 16, 2017.

These additions to the EPA Draft 2 proposal will improve the assessment of the energy efficiency of servers and pass those servers with higher performance capability that result in lower idle and operational power consumption when deployed in a data center or in the office environment.

General note: All data analysis presented in this document has been done using SERT data tested and calculated under SERT V1.1.1. The SERT WG has done an analysis on metrics generated under the V2.0.0 calculations and found that conclusions reached based on an analysis of V1.1.1 data will largely hold for V2.0.0 data.   However, active efficiency thresholds will have to be reset under V2.0.0 based on an analysis of the V2.0.0 active efficiency score because the V2.0.0 active efficiency score is 16%-28% of the V1.1.1 score. The variation of 16%-28% is due to changes in the calculation of the memory worklet active efficiency score, with higher percentages for low memory capacity systems and lower percentages for high memory capacity systems.   The SERT WG has verified that performance and power measurements conducted under V1.1.1 and V2.0.0 are comparable within expected measurement tolerance, so V1.1.1 results can be easily converted to V2.0.0 results. Based on the analysis work completed by the SERT WG, ITI recommends that EPA use SERT V2.0.0 in ENERGY STAR Version 3.

This document will address ITI comments to all aspects of the Draft 2 document with the exception of the updates to the SERT dataset with products released and certified to ENERGY STAR since March 2016. The SERT WG is also working to get SERT data on servers built with the recently announced Intel and AMD processors. The WG intends to analyze the updated dataset with active efficiency values converted to SERT V2.0.0 to provide insight into thresholds revised to the new data and the conversion to SERT V2.0.0 active efficiency and will recommend revised active efficiency and idle power thresholds for use in Draft 3 of Version 3 of the Computer Server requirements. We have requested, and received, an extension to October 16, 2017 for the submission of the updated dataset and the associated analysis. ITI is proposing a revision to the Resilient Server definition (see appendix); work is continuing to update the definition to reflect the impact of recent developments in product technology.

Sincerely,

Alexandria McBride
Director, Environment and Sustainability
ITI
amcbride@itic.org

DETAILED ITI RESPONSE TO DRAFT 2

Lines 61-65: Modify the resilient server definition. A new resilient server technical definition is proposed at the end of the paper under Appendix B.

Lines 79-95: High Performance Computing Definition:

A computing system which is designed and optimized to execute highly parallel applications for high performance, deep learning and artificial intelligence applications. , featuring a ~~large~~ number of clustered ~~homogeneous~~ nodes, often with high speed inter-processing interconnects as well as ~~large~~ high memory capability and bandwidth. HPC systems may be purposely built, or assembled from more commonly available computer servers. HPC systems must meet ALL the following criteria:

A. High Performance Computing (HPC) System: 7.A …optimized for higher performance computing***, augmented or artificial intelligence and deep learning*** applications**.**

B. No changes.

C. Consist of multiple ~~a number of typically homogeneous~~ computing nodes, clustered primarily to increase computational capability;

D. No changes

Justification: Expand the HPC definition to include deep learning and artificial intelligence applications and to match developing technologies. HPC is being applied to additional specialized computational intense applications.

1. Number of nodes is reducing. Small systems with 4-8 have been constructed and used.
2. Nodes may be non-homogenous to support a range of computational activities.

This definition matches the definition proposed for ISO/IEC 21836.1 and EU ErP Lot 9 server and storage equipment eco-design requirements.

Lines 212-216: APA Definition: The differentiation between an expansion APA and an integrated APA is an important distinction for the classification and testing/threshold requirements for servers. ITI supports this change.

We recommend two changes to the definition:

A. An expansion APA: An APA that is on an add-in card installed in a general purpose add-in expansion slot (e.g. GPGPUs installed in a PCI slot). An expansion APA may include one or more GPUs on the add-in card.

B. Integrated APA: An APA that is integrated into the motherboard or CPU package or an expansion APA that has part of its subsystem originating at the CPU boundary on the motherboard.

Justification: The first modification is to ensure appropriate the idle power adder per APA/GPU so if there are two GPUs on one card it will have twice the budget.

The second modification is to deal with the fact that servers may have switches on the motherboard that are used by the APA subsystem and cannot be powered off thereby putting APA based servers at a disadvantage when it comes to SERT based energy efficiency passing bar. Those servers are normally used with GPU cards installed in the field and it seems wrong to lower the bar for non APA based systems because of this. These servers should be classified under integrated APA and out of scope of ENERGY STAR.

Lines 241 to 256: Low-end and High-end Performance Configurations:

Requirement for three test configurations: ITI supports defining the capabilities of the product family using three configurations - low-end, high-end and typical – to certify a product family to the ENERGY STAR computer server requirements. EPA had already eliminated the requirement to supply data for the maximum power configuration, based in part on data and analysis supplied by the SERT WG. The SERT WG had performed a similar analysis on the low power and low-end performance configurations and it was found that performance and active efficiency values of the two low-end configurations were largely the same. This analysis was not provided to EPA. The available data indicates that it is appropriate to define a product family using the three configurations.

We are not clear on how EPA expects a product family to be assessed. It is our understanding that a product family would need a low-end, typical and high-end configuration which passed both the idle limit and the active efficiency limit. We also appreciated John Clinger's comment at the webinar that a certified three configuration product family could be a subset of the total configurations offered for a given machine type/model server family. With that approach the manufacturers are free to determine the group of processors and components that can define a certifiable family. We concur with this approach, but recommend that a sub-section be added to Section 3 of the requirements to clarify how companies should handle the certification of a subset of a product family and certification of a single configuration. We think that the requirements document should outline acceptable certification approaches.

No memory capacity minimum or limit for the low-end and high-end performance configurations: ITI recommends that EPA set a minimum set of requirements for the memory capacity of the low-end and high-end configurations. We recommend that the configurations must:

1. Have all memory channels populated with the same model DIMM. In all cases, the minimum memory capacity is the number of memory channels times the minimum DIMM size offered on the server family.

2. For the low-end configuration, the memory capacity must be at least equal to the number of DIMM slots times the smallest DIMM size offered on the server.

3. For the typical configuration, the memory capacity must be equal to or greater than the product of 2 times the product of the CPU (socket count*core count* threads per core).

4. For the high-end configuration, the memory capacity must be equal to or greater than the product of 3 times the product of the CPU (socket count*core count* threads per core).

The EPA should give server manufacturers the responsibility to set the quantity of memory for each test configuration to enable the submission of configurations that maximizes the SERT active efficiency score.

Note on the Memory Adder:

EPA has proposed a memory adder of 0.125 W/GB. We request that EPA share the data set that they used to determine this memory allowance with the SERT WG so they can understand the basis of the selected level. The SERT WG has collected measured DIMM data, provided in the tables below, from 3 manufacturers which indicate that the 0.125 W/GB allowance is acceptable for DDR3 DIMMs, but too low if 4 GB or 8 GB DDR4 DIMMs are used on the test configuration. As most manufacturers are likely moving to DDR4 DIMMs, we recommend that EPA either set the W/GB idle allowance limit at .2 W/GB to facilitate the use of 4 and 8 GB DDR4 DRAMs or set DIMM size specific allowances of 0.22 W/GB for 4 GB DDR4 DRAM, .17 W/GB for 8 GB DDR4 DRAM and .1 W/GB for 16 GB and higher DDR4 DRAM and all DDR3 DRAM.

	Process Technology Node					
4GB Chipset	30 nm		25 nm		20 nm	
	Min (w/GB)	Max (w/GB)	Min (w/GB)	Max (w/GB)	Min (w/GB)	Max (w/GB)
DDR3	0.076	0.108	0.061	0.108	0.039	0.095
DDR4	0.166	0.223	0.07	0.218	0.065	0.082

Table 1: Watts per GB data by DDR3/4 and process technology node (Watts DC)

DDR4 DIMMs	Number of manufacturers	W/GB		Total Watts	
		min	max	min	max
4 GB	2	0.218	0.223	0.87	0.89
8 GB	3	0.148	0.166	1.18	1.32
16 GB	3	0.075	0.082	1.2	1.31
32 GB	3	0.066	0.07	2.07	2.24

Table 2: Watts DC/GB and total DIMM watt use data for 4 DDR4 DIMM sizes from 3 manufacturers

Line 263: Add an additional Definitional category for the Server Efficiency Rating Tool Components. At a minimum, a definition needs to be added for "Measured Worklet Score: The geometric mean combination of the measured normalized performance divided by the measured power at each of the 4 or 8 measurement intervals of the CPU worklets. For the measured 'capacity 3' worklet score, the worklet score is calculated as described in "The SERT Metric and the Impact of Server Configuration" pages 16 to 18 published by the Standard Performance Evaluation Corporation (SPEC). For the 'capacity 3' worklet score, the worklet performance score measured at the closest capacity interval (4, 8, 16, 32, 64, 256, 512, and 1024 GB) and at 50% of the installed memory (in GB) is the measured capacity worklet performance for that configuration. The SERT suite automatically determines which capacity interval should be measured and only measures that interval.

Line 277: Excluded products: servers shipped with integrated APAs; ITI supports EPA's proposal to exclude servers with integrated APAs. At this time, integrated APAs are largely being used on HPC systems which are already exempt (2.2.2.iv). There is also current work occurring to assess if an integrated APA can be turned off to enable testing of the server without the APA enabled. This is complicated by the fact that the APA is typically supported by a high speed interconnect which may not be easily turned off or put in a low power state.

Line 389 to 398: Worklet efficiency score calculation

The Draft 2 specification needs to note that the worklet efficiency scores need to be combined using the geometric mean. In all of the SERT WG examples, we combine each interval performance and power measurement into an efficiency score and then combine the individual interval efficiencies scores using the geometric mean function. The other calculation method is to combine the 4 performance measurements using the geometric mean function and the 4 power measurements using the same function and then dividing the geometric mean of the performance measurements by the geometric mean of the power measurements. Both calculation methods will provide the same numerical value.

It is important to specify the use of the geometric mean function, as the V1.1.1 SERT report by SPEC combines the worklet score using the arithmetic mean. This changes the overall CPU workload efficiency and the overall active efficiency score and makes the score more prone to the influence of accelerated worklet results, such as the Crypto-AES worklet. Analysis indicates that the use of the arithmetic mean increases the active efficiency scores by 4 to 20 points depending on the server configuration with higher performance servers getting greater increases in the active efficiency value.

EPA should insure that its analysis conducted on V1.1.1 scores is using the geometric mean function to combine the performance and power interval data.

Position on Active Efficiency and Idle Thresholds:

We continue to recommend that EPA set server energy efficiency metric solely based on SERT active efficiency thresholds for server products under the ENERGY STAR requirements. We have demonstrated with data in the Green Grid white paper "Server Energy Efficiency in Data Centers and Offices" that the active efficiency metric best identifies those servers that can deliver more work per unit of energy consumed and reduce the overall energy use and consumption footprint in the data center. We also believe that we have demonstrated that the fact that because the SERT test measures energy use at 4 intervals in the CPU worklets the active efficiency identifies servers with lower idle power values. A server with a larger difference in the power use at 100% and 25% performance level is likely to have a lower idle power due to the more aggressive management of idle power to achieve that higher power difference, which in turn contributes to a higher SERT active efficiency score. We have also outlined the market trends that show that the server market is steadily moving to a highly virtualized server environment for office, enterprise and cloud computing environments where servers will have higher average utilizations and will infrequently move into a mode where no work is present.

At the same time we recognize that EPA has a different position on the use of idle, and should EPA continue with its intent to move in the direction of combined active efficiency and idle power limits, we strongly advocate that EPA incorporate a system performance idle adder as part of idle methodology. The need for this adder is due to the significant difference in CPU processor power between higher performance and lower performance levels and the fact that higher performing servers have more circuitry and hardware, with increased power demand, to enable full use of the higher performance. Such a difference in server power use will likely become more pronounced with the release of the recently announced processor families. To support incorporating the system performance adder as part of idle, we have provided data and analysis below to demonstrate that the use of the system performance adder does materially affect the system pass rate for the combined active efficiency/idle thresholds.

Lines 412 to 413: Active state efficiency thresholds:

Use of SERT V2.0.0

A. The SERT WG has evaluated the changes in the SERT V2.0.0 test suite and determined the specific changes in the calculation methods for the SERT active efficiency score. The details of the worklet and overall active efficiency score calculations can be found in the document "The SERT Metric and Impact of the Server Configuration" (pages 15 to 18). The primary changes in the test method and the calculation of performance and normalized performance values are described below.

1. The test run time has been reduced from 4 hours in V1.1.1 compared to 2.5 hours in V2.0.0. Manufacturers want to capture the reduced test time benefits of V2.0.0.

2. V2.0.0 has the revised Flood3 and Capacity3 worklet measurement and calculation approaches developed based on the learning gained from the analysis work performed on the V1.1.1 results. The revisions to the two memory worklets improve their applicability to assessing server efficiency. Specific details of the changes are discussed below.

3. V2.0.0 provides the SERT active efficiency metric calculated using the Geometric mean combination of the worklet interval efficiency scores (performance over power), worklet scores to workload score and the weighted geometric mean combination of the three workload scores. Direct reporting of the score on the SERT test form ensures the integrity of the calculation methodology.

4. V2.0.0 uses a new, higher performance reference server to normalize the measured performance scores. This will improve the applicability of the SERT metric, as the new, higher performance technologies are released; as it will moderate a large score value increase on newer, higher performance server products.

B. The SPECpower committee has released the new reference scores and the necessary steps to convert V1.1.1 Flood(2) and Capacity(2) performance data to V2.0.0 performance numbers. The SERT WG collected SERT V1.1.1 and V2.0.0 metric dataset on the same server configuration for 5 products and 20 configurations. The SERT WG used this data to compare the measured V.2.0.0 score to the V2.0.0 score calculated by applying the new reference scores and revised memory performance calculations to the V1.1.1 performance values. Details of the revised reference scores and memory performance and worklet efficiency calculations can be found in the document "The SERT Metric and the Impact of Server Configuration" pages 16 to 18.

Conclusions based on analysis of the comparative data.

1. The raw performance scores and power values measured by V1.1.1 and V2.0.0 are equivalent within the expected variability of the measurement hardware. This is true for all SERT worklets.

2. The use of a new reference configuration has changed the relative value of the Crypto-AES worklet score. The original reference server did not have the ability to accelerate the Crypto-workload. The new reference server fixes this issue. As a result, servers that do not enable a crypto accelerator will have a very low crypto score compared with accelerated server crypto scores and the other CPU worklet scores while a server with the accelerator will have a normalized Crypto-AES worklet efficiency score consistent with the other CPU active efficiency scores.

3. For the memory capacity performance, the performance value used to calculate the V2.0.0 capacity performance score will be selected by taking a single performance measurement at the capacity test interval closest to the 50% value of the installed memory. As a result, the relative value of the capacity score for systems with small memory capacities will be increased. Under the test process, the capacity score is fairly constant at each interval until the interval at or beyond the memory capacity of the server. At and beyond this point, the performance score decreases because the memory capacity is fully utilized. Under the V1.1.1 capacity value calculation, these lower performance values were included in the geomean resulting in a lowered capacity value. For higher memory configurations, most of the capacity interval values were closely matched near the maximum performance resulting in a geomean value close to the maximum capacity. Capacity scores for low memory capacity servers will increase significantly under the V2.0.0 calculation while scores for high memory capacity servers will increase little if at all.

4. In Flood3 (SERT 2.0.0), the performance score is the measured memory bandwidth with no capacity multiplier or load level adjustment. In the V1.1.1 Flood 2 score, both the 100% and 50% performance scores were divided by the square root of the memory capacity and the 50% performance score was further divided by 2 to reduce the 50% performance to roughly match the CPU scores. This adjustment depressed the memory Flood score. With the removal of these adjustments, the relative value of the Flood(3) score will increase from the

Flood(2) score, thereby increasing the overall memory workload score and increasing memory influence on the active efficiency metric.

C. The SERT WG has calculated the V2.0.0 active efficiency score using the performance and power values measured with V1.1.1 across the full TGG/ITI database. We then compared the ranking of the V1.1.1 CPU, storage and memory workload and overall activity efficiency scores with the V2.0.0 ranking values. We made the following observations. .

1. The ranking of the CPU workload scores did not change materially. The maximum rank change was 2 and the average rank change was 1.
2. The ranking of the storage workload scores did not change materially. The maximum rank change was 13 and the average rank change was .35.
3. As expected from the analysis above, the rank change on the memory workload scores was higher. The maximum rank change was 62, with the average rank change of 11. The change in the calculation of the memory scores did affect the rank, as the values of the memory workload score increased relative to the values of the CPU and storage scores.
4. The rank changes in the memory scores carried over to the overall active efficiency score. The maximum rank change was 37, with an average rank change of 9.

The impact of the change in the memory worklet calculations is dependent on the memory capacity installed in the server. Figure 1 graphically depicts the ratio of the V1.1.1 efficiency score to the V2.0.0 efficiency score. At low values of memory capacity, the V2.0.0 score is about 25% of the V1.1.1 score. As the memory capacity increases, the ratio of the two active efficiency score reduces to 16% of the V1.1.1 value. This outcome is expected, as the largest driver of relative changes in the score resulted from the change in the calculation of the capacity worklet score with the largest relative value changes occurring for systems with low memory capacity.

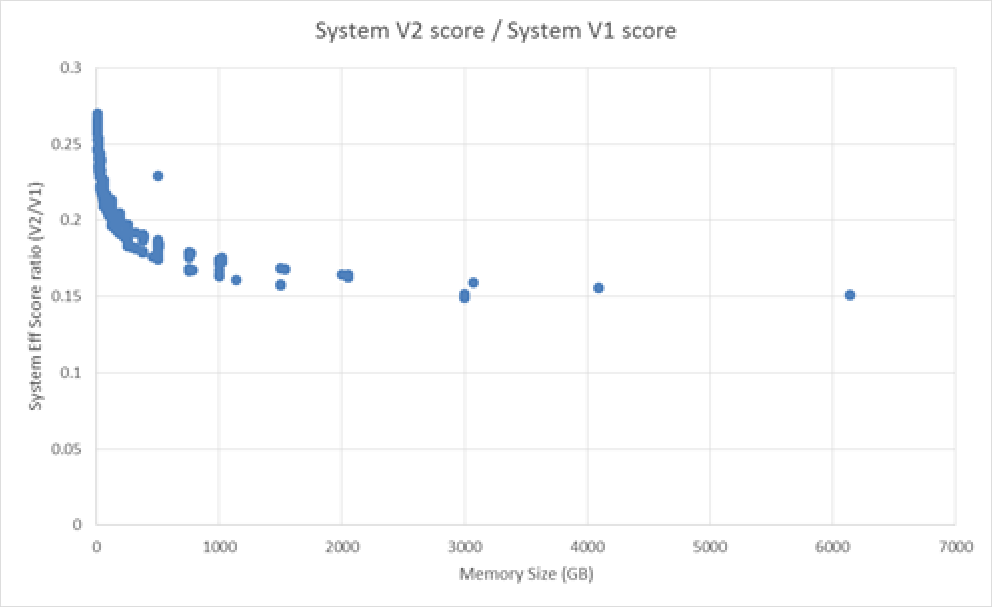


Figure 1: Dependency of the ratio of V2 to V1 SERT active efficiency score to memory capacity in GB

While there were some relatively large changes in the active efficiency score, ITI does not feel that the modifications to the memory worklet score calculations have affected the ability of the SERT test to assess servers for energy efficiency. The adjustments to the Flood and Capacity worklet calculations resulted in a more accurate assessment of the memory system capabilities and a more robust assessment of server energy efficiency. While there are rank changes driven by the overall conversion from memory (2) to memory (3) calculation methods, the SERT WG preliminary review indicates that the ranking and thresholding of products to the V2.0.0 metric will result in identification of the most efficient server systems. It also highlights the reality that the utility of any metric will be dependent on the method used to convert performance and power scores into an efficiency value and the relative weighting of the efficiency value of the three workload types that make up the overall metric. ITI believes that V2.0.0 of the SERT metric provides the calculation methods and weightings which appropriately assess server energy efficiency, and enabling the identification of the most efficient server products.

As discussed previously, the conversion of V1.1.1 metrics to V2.0.0 metrics cannot be done with a simple, single conversion factor applied to the SERT active efficiency metric. Because of the changes to the calculation of the flood and capacity memory worklet performance values, SERT activity efficiency scores will change based on the memory capacity of the given configuration. In order to set active efficiency thresholds for ENERGY STAR, it will be necessary to convert existing V1.1.1 scores to V2.0.0 scores and then reevaluate the data base to determine the appropriate active efficiency thresholds needed to achieve the desired percentage of passing server products within the existing data set that can be certified for ENERGY STAR. The SERT Analysis WG is working to complete the database conversion as well as to add new SERT data from servers certified to ENERGY STAR after March of 2016 and for server products with the recently announced AMD and Intel processors. The SERT Analysis WG intends to transfer this dataset to EPA on October 16, 2017.

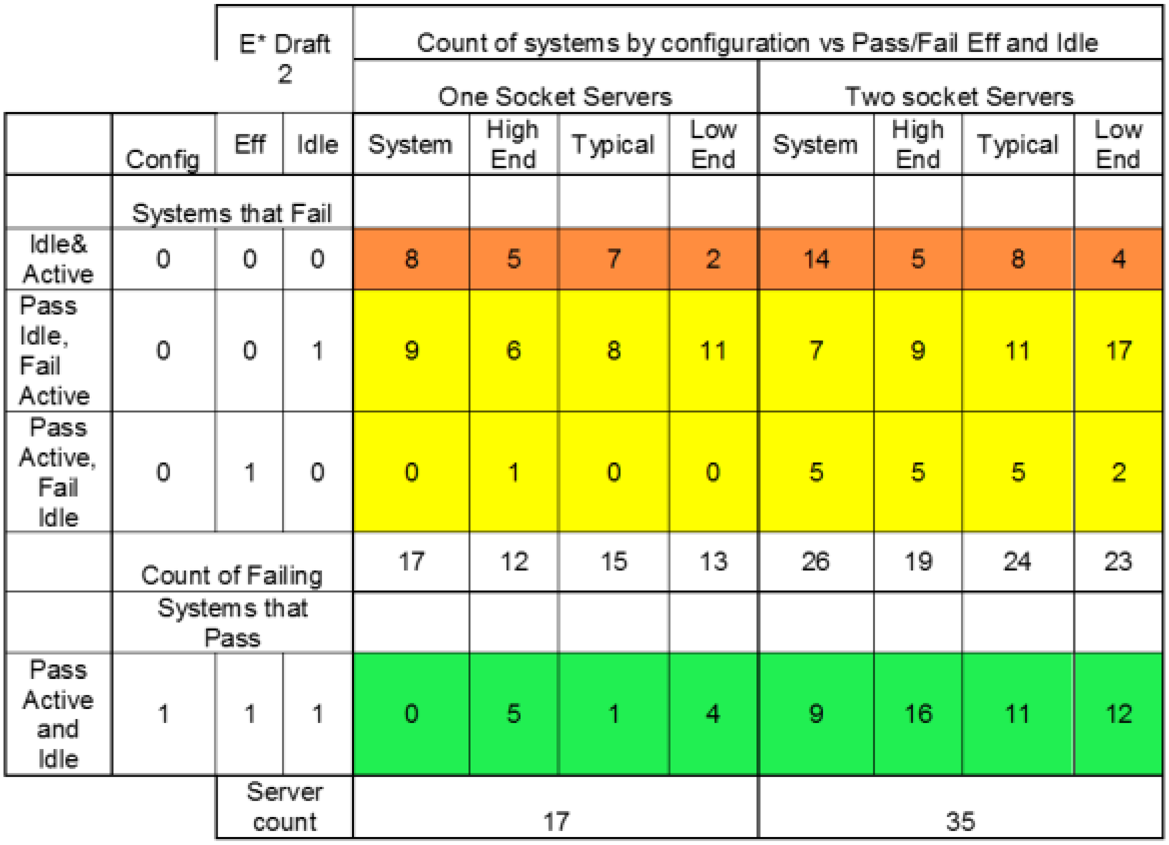
New Processor Announcements:

Next generation product releases: Both AMD and Intel have announced their next generation processor families for servers. Indications are that these processor families will result in improvements in the active efficiency scores as expected from the analysis work the SERT WG has performed on generation to generation systems. The SERT WG is undertaking to gather data from server products manufactured with these processors to provide EPA an initial assessment of the improvements in the active efficiency score so that EPA can assess the value and appropriateness of:

1. Modifying the proposed active efficiency limits based on the data available by a specific date; or
2. Briefly delaying the release of Draft 3 to enable collection of additional SERT data from additional servers to enable a better assessment of the change in efficiency scores enabled by the new processor families. The SERT Analysis WG believes we can provide an updated data set with SERT score data from servers using the new Intel and AMD processor by the October 16, 2017. These products are just moving into testing and so we anticipate being able to collect data over the next two months.

Create Separate Active Efficiency Limit for 1 and 2 socket servers:

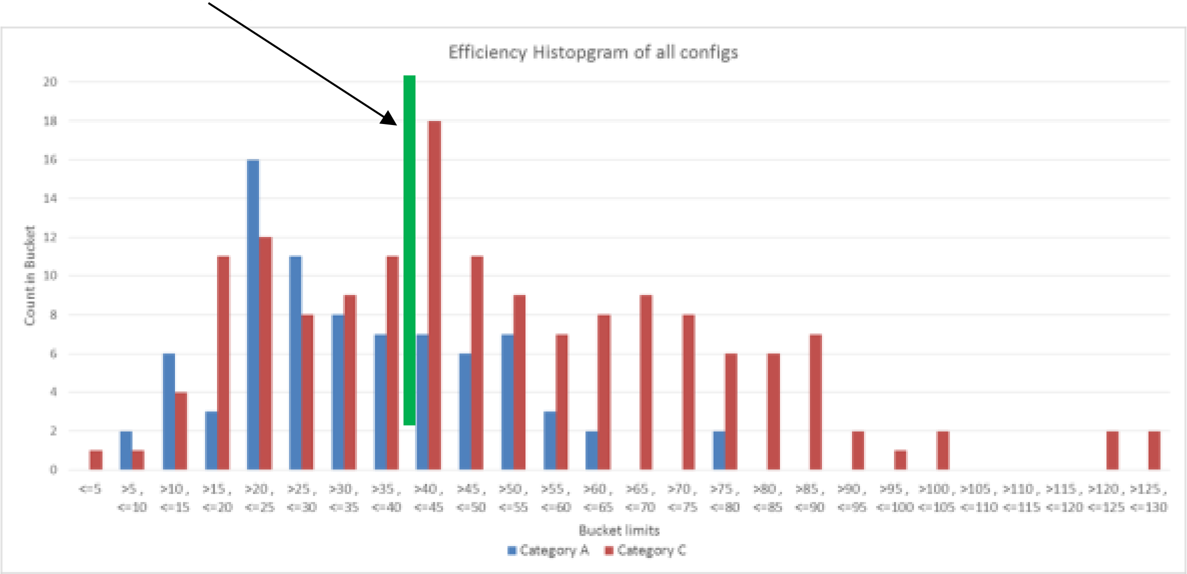
We are concerned that EPA has chosen to set a single active efficiency limit for 1 and 2 socket servers. Our analysis of the dataset indicates that the proposed limit will be more severe on one socket products and inordinately limit their availability. Using our copy of the dataset and analysis tools we have determined that only 20% of the configurations will pass the combined limits. None of the one socket servers will have all three configurations pass the combined active efficiency/idle power thresholds proposed in draft 2. Table 3 below presents the data on systems and configurations which pass and fail the proposed Draft 2 active efficiency and idle limits as set for V1.1.1 active efficiency scores. The first three rows of data in the table provide the details on the systems and configurations which fail the limits. Row 1 shows those that fail both the idle and active thresholds, row 2 shows those that pass the idle limit but fail the active limit and row 3 shows those that pass the active limit but fail the idle limit. Row 3 illustrates that only 1 one socket configuration passes the proposed active efficiency metric. The fourth row of data details the systems and configurations that pass both the idle and active efficiency thresholds. Here again, it is shown that no one socket systems pass both thresholds. There are nine two-socket systems that pass both the idle and active efficiency levels for all three configurations.



		E* Draft 2		Count of systems by configuration vs Pass/Fail Eff and Idle							
				One Socket Servers				Two socket Servers			
	Config	Eff	Idle	System	High End	Typical	Low End	System	High End	Typical	Low End
	Systems that Fail										
Idle& Active	0	0	0	8	5	7	2	14	5	8	4
Pass Idle, Fail Active	0	0	1	9	6	8	11	7	9	11	17
Pass Active, Fail Idle	0	1	0	0	1	0	0	5	5	5	2
	Count of Failing			17	12	15	13	26	19	24	23
	Systems that Pass										
Pass Active and Idle	1	1	1	0	5	1	4	9	16	11	12
		Server count		17				35			

Table 3: Number of Systems and Configurations that pass and fail the idle and active efficiency thresholds

The importance of setting separate active efficiency thresholds limits for one and two socket servers is illustrated by the distribution of server counts to server active efficiency scores shown in Figure 2.



Draft 2 Active Efficiency Threshold

Figure 2: Counts of active efficiency scores by 1 and 2 socket servers (all configurations)

Notes:

1. The above figure includes all configurations (low-end and high-end performance, typical and minimum and maximum power) reported to ENERGY STAR. Table 3 reports only the low-end and high-end configuration with the best active efficiency score and the typical configuration. For this reason, the total number of configurations in figure 2 will be greater than the configuration count in Table 3.

2. All data in Figure 2 is from V1.1.1.

As the histogram shows, the one socket servers have an active efficiency score distribution that is lower than the distribution for the 2 socket active efficiency scores. As such, a single active efficiency score threshold will be biased against one socket servers as shown in Table 3. The SERT WG recommends that EPA sets a separate active efficiency thresholds for the one socket and two socket servers so that the ENERGY STAR requirements do not preclude the qualification of one socket servers.

Line 431 - Idle **State Efficiency Criteria:**

A. ITI agrees with EPA that 4 socket computer servers and all resilient servers should not have an idle power limit. These servers typically have more complex configurations and higher power use that make it highly inappropriate to differentiate server efficiency based on idle power thresholds. Idle power limits are not relevant for these servers and setting idle limits could result in disqualification of high performance systems which can be run at higher utilizations and which deliver more work per unit of energy consumed. Active efficiency thresholds offer the best metric for assessing energy efficiency for these categories of servers.

B. Given EPA's intent to continue to use an idle power limit in conjunction with the active efficiency threshold, it is critical that EPA adopt a system performance idle power adder. The new generation of Intel processors has TDP values ranging from 8 core 70 W/4 core 85 W to 28 core/205 W. The higher core count/TDP processors have significantly better performance characteristics then the low core count/TDP processors, but the added power demand from the higher core counts and system capabilities and circuitry to support more resources (memory, storage and I/O) are expected to increase the idle power values. The idle power limit needs the system performance adder to compensate for this larger power profile. ITI recommends that the adder be assessed based on three values:

   a. CPU Peak Performance: the geometric mean of the 100% performance values for the 7 CPU worklets: Compress, LU, SOR, SORT, Crypto, SHA256 and Hybrid ssj.

   b. Base performance threshold: This is the minimum performance below which a server product will not be eligible for the system performance adder. The base performance threshold is compared to the CPU Peak Performance to determine if the server can apply a system performance adder.

   c. System performance multiplier: The value by which the server CPU Peak Performance is multiplied to calculate the system performance adder in Watts.

EPA stated that most of the high performance servers passed their current idle requirements therefore they did not see a need for performance based idle adder. In order to compare the impacts of the addition of the system performance adder on systems that passed and failed, we analyzed the performance level of systems which failed the Draft 2 idle power limit, which has a pass rate of 50%[1] and those which would pass the idle limit with a system performance adder[2], again at a 50% pass rate. We recognize that the systems pass and fail under the Draft 2 proposal on the basis of both their active efficiency score and idle measurement, but the evaluation of the pure idle limits is the best way to assess how the addition of the system performance multiplier changes the performance characteristics of the systems which are rated against the active efficiency score. The analysis indicates that the addition of the system performance multiplier is important to minimize the bias of the proposed EPA idle limit against higher performance servers.

The analysis divides the configurations into 10 histogram buckets based on their geometric maximum CPU performance value. The configurations are evaluated against the Draft 2 idle power limit and an idle limit with a systems performance adder. Figure 3 show the percentage configurations in each performance bucket that fail the Draft 2 idle limit and the idle limit with system performance allowance to provide a comparison of the percentage of configurations that fail the two idle limits in each performance bucket.

[1] The Draft 2 idle limit has a 50% pass rate because it is combined with the active efficiency limit to yield a total of 25% passing systems.

[2] The system performance adder is designed to that the maximum difference in the adder between the low end and high end configurations for a server family does not exceed 40 Watts. (referred to as "Perf max of 40")

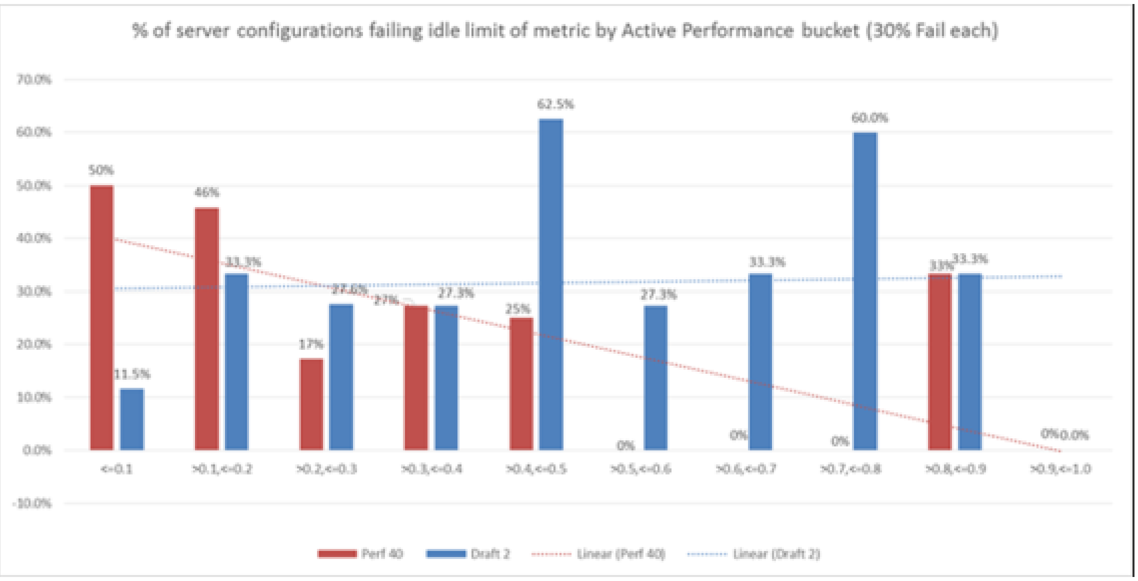


Figure 3: Percent of systems failed in each performance bucket by the Draft 2 Idle power limit and the Draft 2 Idle power limit with a system performance adder (Perf 40).

Note: This figure was constructed using V1.1.1 data.

Active Performance Bucket	0-10%	10-20%	20-30%	30-40%	40-50%	50-60%	60-70%	70-80%	80-90%	90-100%
Configuration Count	26	48	29	11	8	11	6	5	3	4
Failed Servers: EPA Idle	3	16	8	3	5	3	2	3	1	0
Failed Servers: System Perf Idle	13	22	5	3	2	0	0	0	1	0

Table 4: Counts of servers that fail the EPA Idle and System Performance adjusted idle by performance bucket.

To calibrate the percentages in figure 3, Table 4 shows the total number of systems in each bucket and the number that fail under the two idle limits. Table 4 details the differences in idle failure rate at the low and high performance levels. The Draft 2 idle limit fails 9 of 29 server configurations in the top 50% of configuration performance, while the idle limit using the system performance adder fails only 1. Looking at the low performance configurations, the Draft 2 idle limit fails 27 of 103 servers in the lowest 30% of performance, while the idle limit using the system performance idle adder fails 40 servers in the bottom 30%. The use of the system performance adder markedly changes the performance profile of the configurations that fail the idle limit.

Looking at the trend lines for the failure rate of the two different idle limits, the Draft 2 idle limit progressively fails higher percentages of systems as the server performance increases while the idle limit with system performance adder (Perf40) fails progressively fewer servers with increasing server performance. Overall, the idle limit in draft 2 will override the active efficiency score benefit enjoyed by high performance servers. Since higher performance enables fewer servers to be deployed to execute any unit of work for end users, the EPA Draft 2 trend is concerning as it will tend to decrease the availability of servers which will have smaller overall physical and energy footprints when deployed in a data center or office environment.

The concern with the impact of the idle limit proposed in Draft 2 is amplified by the expectations that future increases in performance in server silicon will very likely be accompanied by increases in idle power. A lower idle limit that does not account for the higher power demand of higher performing systems risks further excluding high performance servers as these new products are introduced to the market. This preference for lower performance systems in the EPA idle power metric is in direct opposition to the customer desire for fewer higher performing servers and the expected technology/efficiency advances in future server silicon.

Table 5 below uses EPA draft 2 active efficiency limits and idle power limit with the System Performance idle adder (Perf max =40) to assess the number of configurations which pass and fail the idle limit and combined idle and active efficiency limits. The results of Table 5 can be compared to the results in table 3 to observe the differences in passing configurations and systems driven by the addition of the system performance idle adder. The most important difference here is that fewer High-End systems and/or configurations pass efficiency and fail idle limits. The performance adjusted idle removes the negative bias against high end, high performance systems.

		EPA Eff w Perf40 @ 25% Yield		Count of systems by configuration vs Pass/Fail Eff and Idle							
				Category A: 1 socket rack & tower systems				Category C: 2 socket rack and tower systems			
	Config	Eff	Idle	System	High End	Typical	Low End	System	High End	Typical	Low End
	Systems that Fail										
Active & Idle	0	0	0	3	2	3	2	18	5	11	12
Pass Idle, Fail Active	0	0	1	14	9	12	11	4	10	9	10
Pass Active and Fail Idle	0	1	0	0	0	0	0	3	1	3	1
	Count of Failing			17	11	15	13	25	16	23	23
	Systems that Pass										
Pass Active and Idle	1	1	1	0	6	1	4	8	18	11	10
		Server count		17				33			

Table 5: Systems and configurations that pass EPA active efficiency and idle limits at 25% yield with the system performance adder

The impact of the system performance idle adder can be further assessed by looking at the changes in passed systems with the use of the adder.   Comparing Figure 4 to Figure 5, there are 17 configurations with an active efficiency score over 60 that fail the straight idle limit, but only 6 configurations that fail if the system performance adder is used. These graphs are constructed using both 1 and 2 socket rack server data. The system performance adder accounts for the higher socket power and infrastructure power used by the higher performing servers. The data analysis (Tables 6 and 7) shows that this adjustment does not appreciably increase the average idle power of the systems which pass the a set of active efficiency/idle thresholds but there are reductions in the deployed power of the servers which pass the thresholds resulting in a better overall outcomes in terms of reduced data center power. Figures 4, 5 and 6 are constructed using V1.1.1 active efficiency scores.

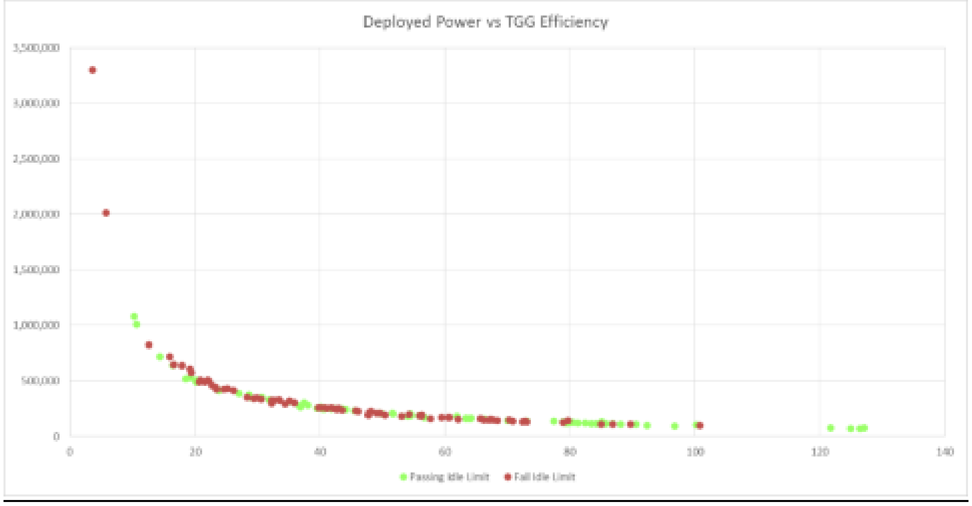


Figure 4: Passing and Failing Configurations for a Draft 2 Idle Limit, without the System Perfomance adder, set for a 25% passing rate

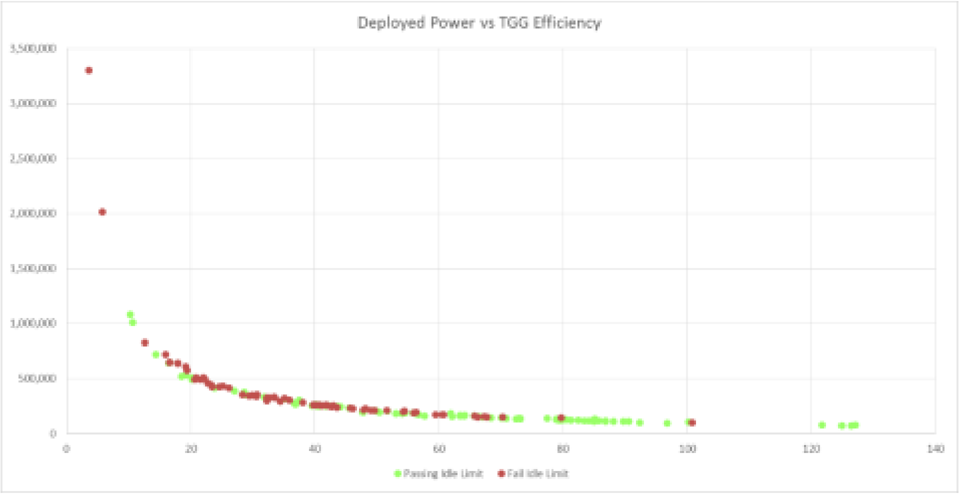
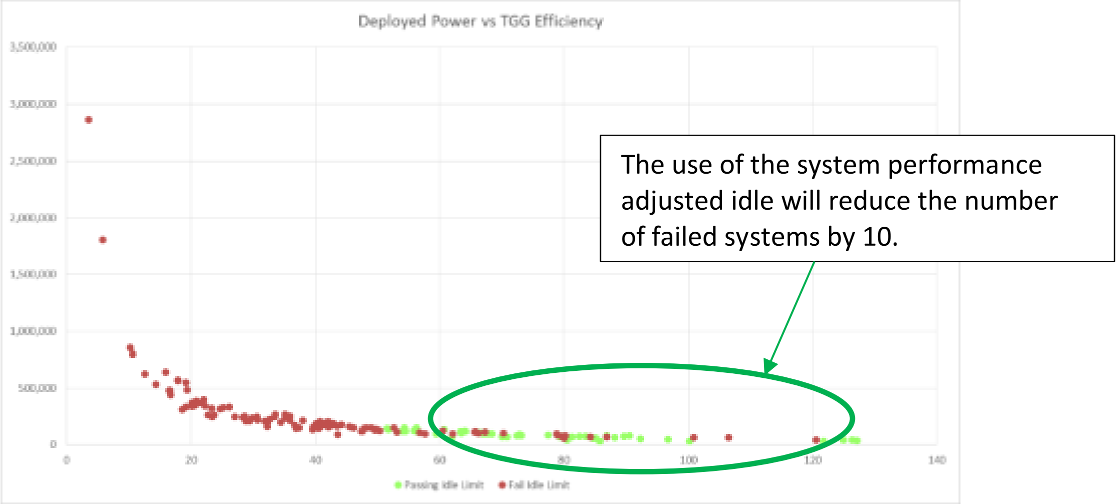


Figure 5: Passing and Failing Configurations for an Idle Limit, which includes the System Perfomance adder, set for a 25% passing rate

Looking at the plot of the deployed power versus active efficiency for the configurations which pass and fail the ENERGYS STAR V3 draft 2 active efficiency and idle limits (figure 6) and comparing it to the passing and failing configurations in Figures 3 and 4, you can see the use of the system performance idle adder will reduce the number of higher active efficiency systems removed by the proposed limits.



The use of the system performance adjusted idle will reduce the number of failed systems by 10.

Figure 6: ENERGY STAR server requirements V3 Draft 2 passed and failed configurations

The impact of an idle power limit on deployed power in the data center

The EPA commented that servers of similar performance can have noticeably different idle power values and that an idle limit is needed to provide preference for the lower idle power system in these cases. The question to be answered is “What is the relative merit of idle power limits versus active efficiency in identifying the lower power systems in these cases?” The SERT WG created a histogram analysis to divide the servers into subsets with similar performance and evaluate the effectiveness of active efficiency and EPA draft 2 idle limit in identifying the lower power servers in each subset.

We assigned the servers in the database to performance buckets based upon the weighted geometric mean performance score for each server. We divide the servers into 26 performance buckets with each bucket having a width of 3 units for weighted geometric mean performance score. The intent of this analysis is to determine the average idle power value and deployed power value for passed and failed server configurations in each active efficiency bucket. This will illustrate which of the three metric approaches best limits idle power and deployed power in the data center or operating environment. Overall, the three methods provide similar results, with the EPA Draft 2 metric delivering the lowest overall average idle power, by about 10%, and the active efficiency metric delivering the lowest deployed power for passing systems (by about 2%) and the highest avoided deployed power for failing systems.

Idle Power histogram analysis

In Table 5 below we divide category C, 2 socket systems into 23 performance buckets (TGG Perf window width of 2 units with the maximum TGG performance value for that "bucket" listed in column 4). For each bucket we average the idle power of systems which pass the given set of limits and the average idle power of those that fail the limits. These two values can be compared to understand if the given metric choice is favoring a lower or higher idle value of servers that pass the metric. This methodology allows us to compare the resulting idle power of passed and failed servers for pure idle and active efficiency limits and combinations of active efficiency and idle. All metrics are set to pass 25% of the 2 socket servers in the dataset. We look at active efficiency only, the EPA Draft 2 proposal and the EPA Draft 2 proposal using the proposed system performance idle adder. Table 6 below shows the average idle power for passing and failing systems in each bucket for each metric. Cells are highlighted in red if the average idle power of failing systems is lower than that of passing systems as this indicates that for the servers in this bucket the metric has selected the higher idle power servers.

Only the active efficiency metric yields a higher average idle power for failing systems than passing systems in each and every bucket. Both of the combined active efficiency and idle limit metrics yield similar results, with only one bucket having a higher idle power in the passing versus failing configurations.

The last row in table 6 compares the average idle power for all the passed and failed systems for each metric. Overall, the EPA Draft 2 method yielded a lower average idle power for the passing systems as compared to the active efficiency metric: 14 watts or 9.3% not a material value when considering the reduced power consumption of the higher performing systems. The higher average idle value for the configurations which passed the active efficiency metric occurred because the metric passed high idle value configurations in buckets 14-23. The majority of the configurations in these buckets failed in the EPA Draft 2 requirements. The high end, high performance systems provided significant performance capabilities and lower deployed power values (table 7) but also increase the average idle power of the passed systems. The addition of the system performance idle adder to the EPA Draft 2 method modifies the results in both the lower and higher performance buckets, resulting in passing of additional higher performance servers when compared to the draft 2 thresholds, but a lower number of higher performance configurations than the active efficiency thresholds. The EPA draft 2 method using the system performance adder results in the same average idle power as the active efficiency threshold.

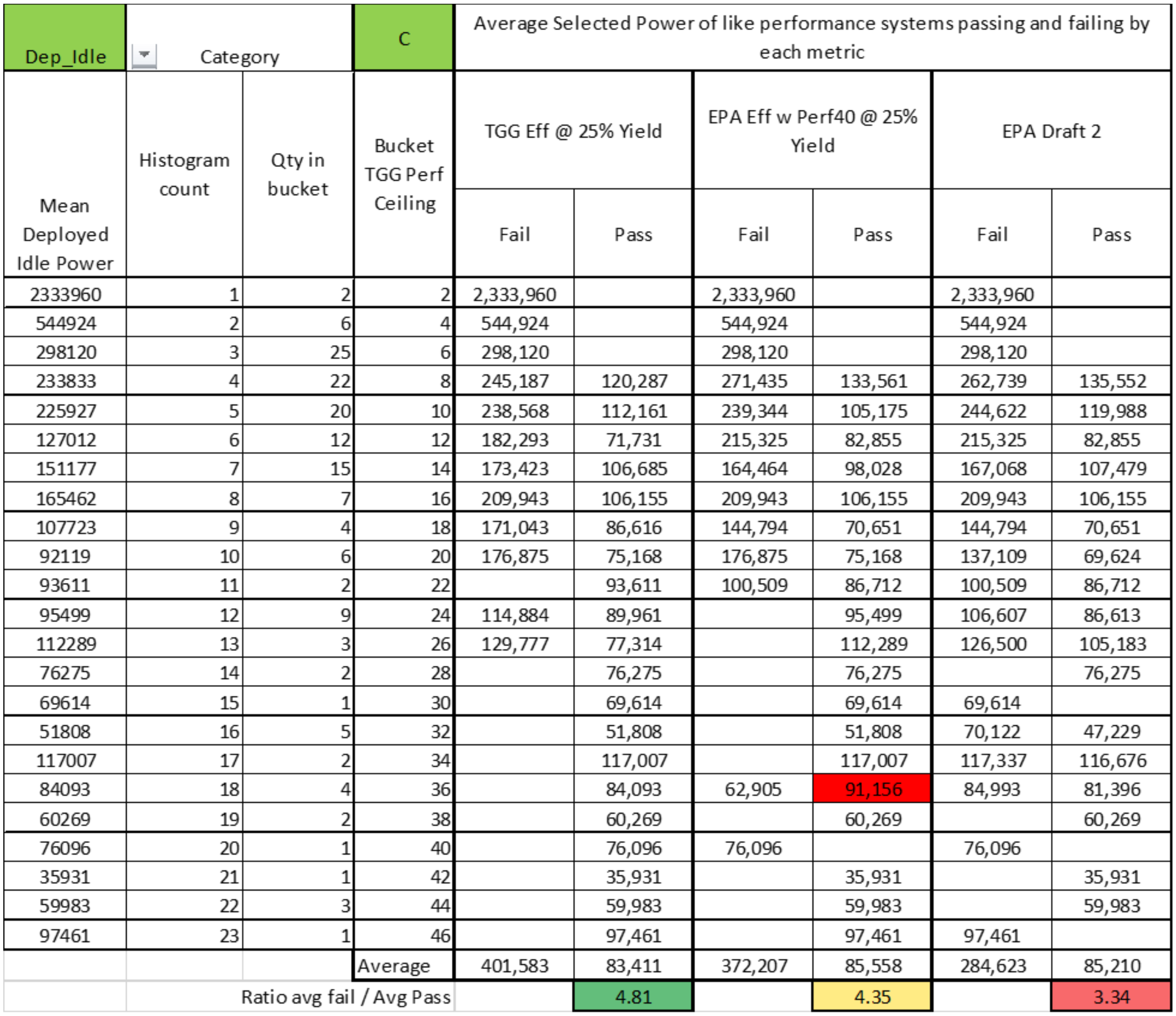
idle	Category		C	Average Selected Power of like performance systems passing and failing by each metric					
Mean Idle Power	Histogram count	Qty in bucket	Bucket TGG Perf Ceiling	TGG Eff @ 25% Yield		EPA Eff w Perf40 @ 25% Yield		EPA Draft 2	
				Fail	Pass	Fail	Pass	Fail	Pass
134	1	2	2	134		134		134	
109	2	6	4	109		109		109	
99	3	25	6	99		99		99	
114	4	22	8	119	60	132	65	127	68
146	5	20	10	154	74	154	72	157	80
101	6	12	12	146	57	170	67	170	67
152	7	15	14	174	108	166	97	167	109
194	8	7	16	246	126	246	126	246	126
140	9	4	18	221	114	189	92	189	92
143	10	6	20	274	117	274	117	215	107
163	11	2	22		163	172	154	172	154
177	12	9	24	211	168		177	195	163
235	13	3	26	271	164		235	265	220
164	14	2	28		164		164		164
160	15	1	30		160		160	160	
131	16	5	32		131		131	180	118
319	17	2	34		319		319	314	324
241	18	4	36		241	178	262	242	239
179	19	2	38		179		179		179
238	20	1	40		238	238		238	
124	21	1	42		124		124		124
216	22	3	44		216		216		216
357	23	1	46		357		357	357	
			Average	180	164	174	164	197	150
	Ratio avg fail / Avg Pass				1.10		1.06		1.31

Table 6: Histogram of average idle power of passed and failed system segregated by server performance

Deployed Idle Power Histogram Analysis

In order to normalize the impact of the increased idle power associated with servers with higher performance capabilities, it is necessary to also look at the performance histogram as it relates to the deployed power of the passed servers (table 7). To do this, the same analysis of the data set segregated into performance buckets is done using the average deployed idle power of passing and failing systems for each metric. Active efficiency is the only metric that yields higher failing vs passing system power levels for all buckets and also has the highest average ratio of failing / passing average systems deployed power. Importantly, the ratio of deployed power on failed configurations to passed configurations is

44% larger for the active efficiency metric as compared to the EPA draft 2 method. This demonstrates that active efficiency metric is more effective in removing higher deployed idle power failing systems than EPA Draft 2 with pure idle approach.



Dep_Idle	▾ Category		C	Average Selected Power of like performance systems passing and failing by each metric					
				TGG Eff @ 25% Yield		EPA Eff w Perf40 @ 25% Yield		EPA Draft 2	
Mean Deployed Idle Power	Histogram count	Qty in bucket	Bucket TGG Perf Ceiling	Fail	Pass	Fail	Pass	Fail	Pass
2333960	1	2	2	2,333,960		2,333,960		2,333,960	
544924	2	6	4	544,924		544,924		544,924	
298120	3	25	6	298,120		298,120		298,120	
233833	4	22	8	245,187	120,287	271,435	133,561	262,739	135,552
225927	5	20	10	238,568	112,161	239,344	105,175	244,622	119,988
127012	6	12	12	182,293	71,731	215,325	82,855	215,325	82,855
151177	7	15	14	173,423	106,685	164,464	98,028	167,068	107,479
165462	8	7	16	209,943	106,155	209,943	106,155	209,943	106,155
107723	9	4	18	171,043	86,616	144,794	70,651	144,794	70,651
92119	10	6	20	176,875	75,168	176,875	75,168	137,109	69,624
93611	11	2	22		93,611	100,509	86,712	100,509	86,712
95499	12	9	24	114,884	89,961		95,499	106,607	86,613
112289	13	3	26	129,777	77,314		112,289	126,500	105,183
76275	14	2	28		76,275		76,275		76,275
69614	15	1	30		69,614		69,614	69,614	
51808	16	5	32		51,808		51,808	70,122	47,229
117007	17	2	34		117,007		117,007	117,337	116,676
84093	18	4	36		84,093	62,905	91,156	84,993	81,396
60269	19	2	38		60,269		60,269		60,269
76096	20	1	40		76,096	76,096		76,096	
35931	21	1	42		35,931		35,931		35,931
59983	22	3	44		59,983		59,983		59,983
97461	23	1	46		97,461		97,461	97,461	
			Average	401,583	83,411	372,207	85,558	284,623	85,210
	Ratio avg fail / Avg Pass				4.81		4.35		3.34

Table 7: Histogram of average deployed power of passed and failed system segregated by server performance

For the Draft 2 method with the system performance adder, both have only one bucket where the deployed power of the passed systems exceeds the deployed power of the failed system, but the modified method has a 20% higher failed deployed power to passed deployed power ratio as compared to the Draft 2 method. However, the deployed idle power is 2,348 watts higher. The larger deployed power avoidance illustrates that the utility of active efficiency and idle power metrics ultimately have to be evaluated based on their ability to reduce power use in an installed environment. The active efficiency metric most effectively reduces both the average deployed idle power and the average deployed power use.

Conclusion based on table 6 and 7 results:

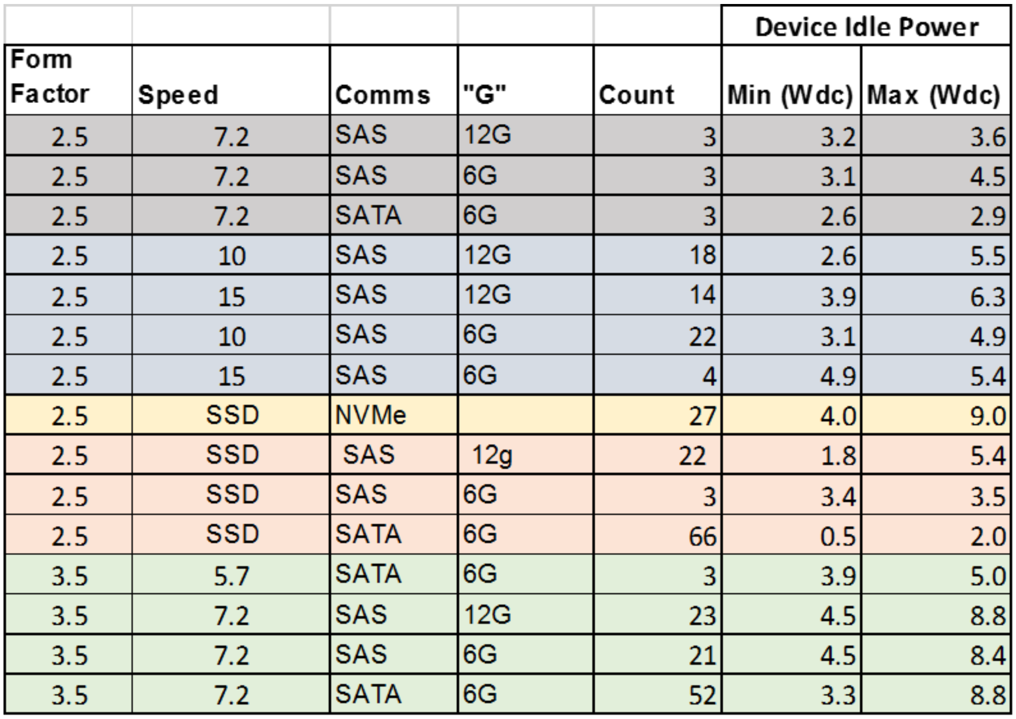
Overall, looking at the results in table 6 and 7, all three metric options yield very similar results when comparing the average idle power and deployed power of the systems that pass each metric. From an ITI perspective, the most important factor is that the active efficiency only threshold results in a substantially higher deployed idle power in the failing systems and a lower deployed power in the passing systems. This outcome indicates that the active efficiency threshold distinguishes those configurations which offer the best combination of performance and power profile to deliver a lower energy footprint in the data center or office environment. In addition, the average idle power of the individual configurations which passed the metric is only 9.3% higher than those servers which passed the EPA Draft 2 metric. This outcome suggests that even if the passed servers operated at idle all the time, those servers that passed the active efficiency metric would have the lowest total power use because in the end you deploy fewer physical server systems and use less total power.

All of the previous analysis was performed using the V1.1.1 activity efficiency metric. When the V1.1.1 scores are converted to V2.0.0 scores it will be necessary to re-evaluate the data and reset the active efficiency and idle power thresholds to yield a 25% passing rate for the configurations or servers families in the database. The SERT WG also intends to add recent server data to the data set, which will also necessitate a reassessment of the thresholds. The SERT WG will provide the updated dataset with proposed thresholds to the EPA on October 16, 2017.

Component Idle Adders

A. **The idle allowances for memory**: See discussion above on page 2. ITI recommends that the memory adder be set at .17 W/GB, to allow 8 GB DDR4 DIMMs to be used on test configurations or that EPA set a graduated DIMM adder values of 0.22 W/GB for 4 GB DDR4 DRAM, 0.17 W/GB for 8 GB DDR4 DRAM and 0.1 W/GB for 16 GB and higher DDR4 DRAM and all DDR3 DRAM.

B. **Idle Allowances for Storage components:** The SERT WG gathered an extensive data set on idle power use for storage devices. As the table 8 below shows, there is a wide variation in idle power between and within device types as segregated by form factor, device speed or SSD, communications protocol used and capacity. Data was collected from 284 drives manufactured by 5 HDD manufactures and 5 SSD manufacturers.

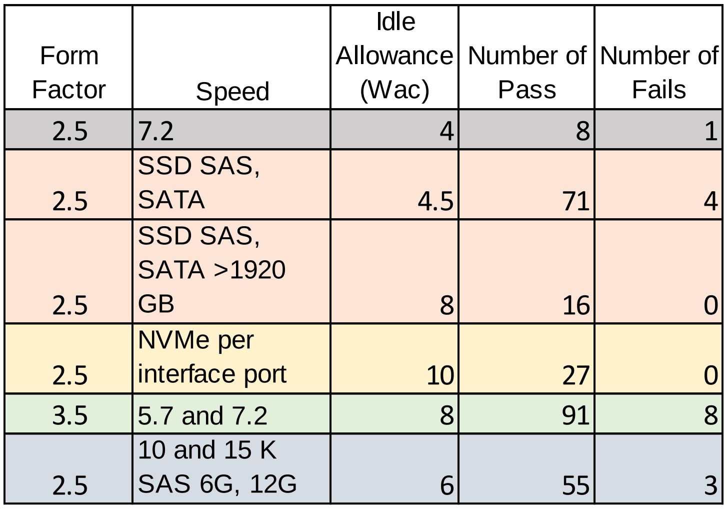


					Device Idle Power	
Form Factor	Speed	Comms	"G"	Count	Min (Wdc)	Max (Wdc)
2.5	7.2	SAS	12G	3	3.2	3.6
2.5	7.2	SAS	6G	3	3.1	4.5
2.5	7.2	SATA	6G	3	2.6	2.9
2.5	10	SAS	12G	18	2.6	5.5
2.5	15	SAS	12G	14	3.9	6.3
2.5	10	SAS	6G	22	3.1	4.9
2.5	15	SAS	6G	4	4.9	5.4
2.5	SSD	NVMe		27	4.0	9.0
2.5	SSD	SAS	12g	22	1.8	5.4
2.5	SSD	SAS	6G	3	3.4	3.5
2.5	SSD	SATA	6G	66	0.5	2.0
3.5	5.7	SATA	6G	3	3.9	5.0
3.5	7.2	SAS	12G	23	4.5	8.8
3.5	7.2	SAS	6G	21	4.5	8.4
3.5	7.2	SATA	6G	52	3.3	8.8

Table 8: Summary of Idle Power data for storage devices (Allowance groups are color coded to Table 9)

Similar to the discussion about memory DIMMs, server manufacturers source storage devices from several manufacturers and randomly use those devices when building configurations. Because the low-end and high-end server configurations designated for testing are required to have 2 storage devices, the allowance will have a minimal impact on the idle limit for the tested configurations. A representative adder is important for surveillance testing where a procured product may have multiple storage devices of a type with a high idle power.

ITI recommends that EPA create 6 categories of storage devices with the idle allowances detailed in Table 9. The SERT WG is undertaking to find idle power data on storage devices. The intent is to supply any additional data with the ITI October 16, 2017 comments.



Form Factor	Speed	Idle Allowance (Wac)	Number of Pass	Number of Fails
2.5	7.2	4	8	1
2.5	SSD SAS, SATA	4.5	71	4
2.5	SSD SAS, SATA >1920 GB	8	16	0
2.5	NVMe per interface port	10	27	0
3.5	5.7 and 7.2	8	91	8
2.5	10 and 15 K SAS 6G, 12G	6	55	3

Table 9: Proposed Storage Categories and Idle Allowances

C. **I/O Device Idle allowances:** Communications technology is currently transitioning to a virtualized environment, allowing servers to dynamically partition and provision high performance capacity ports to multiple logical ports. The larger capacity ports support a virtualized communication environment referred to as Software Defined Networking (SDN) and Network Function Virtualization (NFV). Although systems equipped with these high capacity ports demand higher incremental idle and active power, a higher capacity port (e.g. 100Gbs) can be dynamically provisioned to replace multiple dedicated lower capacity ports (e.g. 10 x 10Gb/s or 5 x 20Gb/s) thereby decreasing overall energy use. The additional adders for the higher capacity ports are proposed in order to anticipate the integration and software based deployment of these higher capacity ports in server systems.

The industry is developing and releasing new network ports with data processing speeds of 25 Gb/s to 200 Gb/s. For ports up to 10 Gb/s, we are in agreement with the EPA proposal. For network ports with higher data processing speeds, they have a higher per port power use which is offset by the ability of the port to transfer more data per watt of power consumed. With the increased use of network virtualization software, these higher power ports will have transfer larger quantities of data over fewer switches reducing the net deployed power in the data center.

Table 10 details the proposed idle allowances for these higher speed ports. At this time, the SERT WG has not collected measured data to justify these idle allowances. The WG is working on securing data with the intent of providing data, similar to that provided for memory DIMMs and storage devices with the October 16, 2017 comment submittal.

Network Port Speed	Proposed Idle Allowance (W)
10 to 25 Gb/s	15 W
>25 to 50 Gb/s	20 W
>50 to 100 Gb/s	26 W
>100 to 200 Gb/s	35 W
>200 Gb/s	45 W

Table 10: Proposed Idle Allowances for High Gb/s Network Ports

ITI recommends the creation of 4 additional network port categories with the idle allowances detailed in Table 6 and increase the idle allowance for the 10 to 25 Gb/s Network port from 8 W to 15 W.

Like memory DIMM and storage device allowances, the primary purpose of establishing idle allowances for these higher throughput ports to accommodate these in the case of market surveillance activities. These higher output ports are extremely unlikely to be installed in the two configurations tested to demonstrate compliance with the ENERGY STAR requirements, but could be installed in a configuration selected by the CB for monitoring and verification.

The SERT WG is undertaking to find idle power data on I/O ports. The intent is to supply any additional data with the ITI October 16, 2017 comments.

Line 313: Power Supply Efficiency at the 10% Load Point

During the EPA Webinar on August 18, 2017, a statement was made by one of the participants that because servers spend an inordinate amount of time in idle EPA should review and tighten the efficiency requirement at the 10% load point on the power supply.

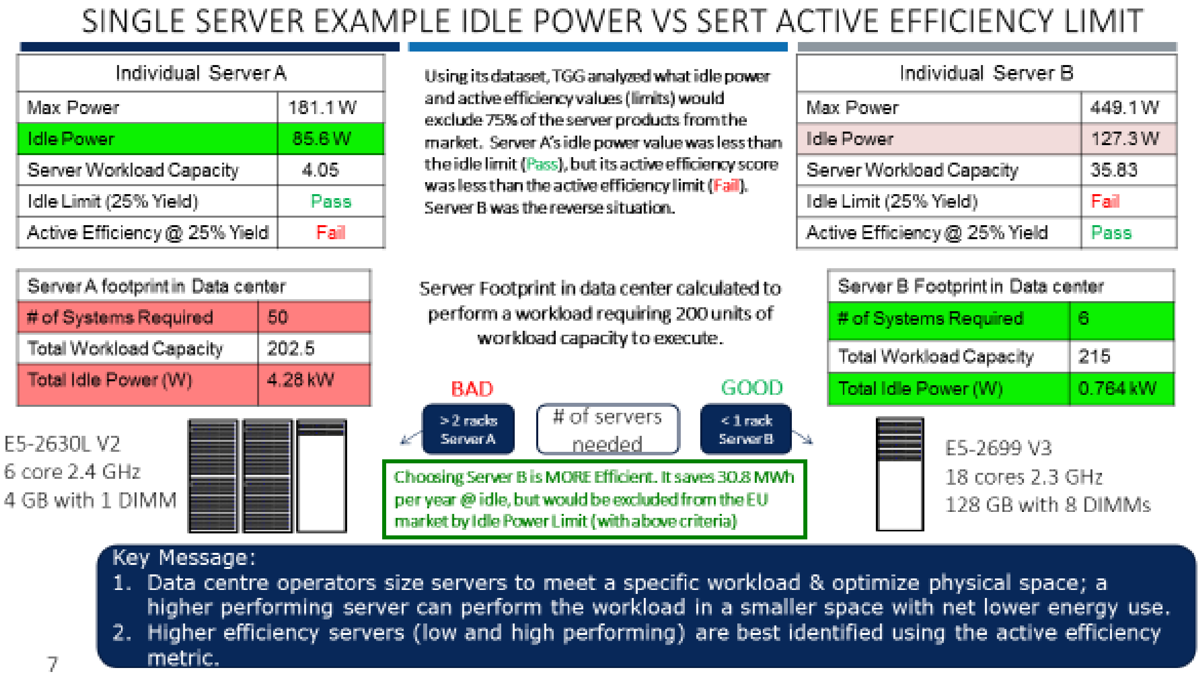
First, ITI takes issue with the statement that servers spend an inordinate amount of time in idle. It is acknowledged that servers in many environments may have utilizations in the range of 10-20%. And in some cases lower. These are choices made by the data center operator based on their operational policies for a given workload. However, there are several considerations that mitigate the impact of these utilization levels:

1. The average utilization level is an aggregate of the utilization of the server over a given period of time. A server that is 20% utilized does not mean it is only working 20% of the time and idle 80% of the time. Rather it has varying levels of workload over a day/week/month/year.

2. The SERT test takes account of the ability of a server to operate at lower power levels at lower utilizations. A server with strong power management capability that has a 25% power/maximum power ratio of 30% will have a higher SERT active efficiency score when compared to a server with similar performance but a ratio of 50% between power use at 25% and maximum utilization. The lower power characteristics will continue at lower utilizations all the way down to idle.

3. While individual server capability is important, it is also important to consider the number of servers needed to deliver a specific workload capacity. As we have explained above, it is important to consider the amount of power that will be required in the data center or office environment to do a given workload. If one high performance server, with a high idle value, replaces 5 medium performance servers with half the idle power the net installed power demand for the lower idle power server will be 2.5 times that for the higher power server. Ultimately, it is the power demand in the data center that matters, and that power demand is a function of both the individual servers installed and the number of servers required to do the workload.

Figure 7 below provides a graphical example of this point. It compares the number of servers and the power use required to meet a specific workload capacity in a data center or office operating environment for two servers. Server A passes the idle power criteria set to achieve a 25% passing rate and fails an active efficiency metric also set for a 25% passing rate and Server B, which fails the idle power criteria but passes the active efficiency criteria. Server A requires 50 servers to meet the defined workload capacity and has 4.28 kW of idle power when installed in a data center or office. Server B requires only 6 servers and 0.76 kW of idle power when installed. Because of its higher performance, Server B requires roughly one-sixth of the power when installed despite the fact that its idle power use is 2.5 times the idle power of server B. Its ability to do more work with fewer servers results in a lower overall power use.



SINGLE SERVER EXAMPLE IDLE POWER VS SERT ACTIVE EFFICIENCY LIMIT

Individual Server A	
Max Power	181.1 W
Idle Power	85.6 W
Server Workload Capacity	4.05
Idle Limit (25% Yield)	Pass
Active Efficiency @ 25% Yield	Fail

Using its dataset, TGG analyzed what idle power and active efficiency values (limits) would exclude 75% of the server products from the market. Server A's idle power value was less than the idle limit (Pass), but its active efficiency score was less than the active efficiency limit (Fail). Server B was the reverse situation.

Individual Server B	
Max Power	449.1 W
Idle Power	127.3 W
Server Workload Capacity	35.83
Idle Limit (25% Yield)	Fail
Active Efficiency @ 25% Yield	Pass

Server A footprint in Data center	
# of Systems Required	50
Total Workload Capacity	202.5
Total Idle Power (W)	4.28 kW

Server Footprint in data center calculated to perform a workload requiring 200 units of workload capacity to execute.

Server B Footprint in Data center	
# of Systems Required	6
Total Workload Capacity	215
Total Idle Power (W)	0.764 kW

E5-2630L V2
6 core 2.4 GHz
4 GB with 1 DIMM

Choosing Server B is MORE Efficient. It saves 30.8 MWh per year @ idle, but would be excluded from the EU market by Idle Power Limit (with above criteria)

E5-2699 V3
18 cores 2.3 GHz
128 GB with 8 DIMMs

Key Message:
1. Data centre operators size servers to meet a specific workload & optimize physical space; a higher performing server can perform the workload in a smaller space with net lower energy use.
2. Higher efficiency servers (low and high performing) are best identified using the active efficiency metric.

7

Figure 7: Comparison of deployed power of for two servers of with different performance and power use characteristics

In order to look at the question of power supply utilization, the SERT WG evaluated the ITI/TGG dataset to determine the power supply utilization at idle for one socket (Category A) and two socket (Category C) server configurations in the dataset. The data analysis is graphically detailed in Figures 8 and 9. The graphs show that the typical, high-end performance and maximum power configurations largely utilize power supplies at over 15% at idle. The typical configuration is the most representative of configurations that will be deployed to data center and office environments. The servers utilizing power supplies below 15% are largely minimum power and performance configurations which are not typically the type of configuration deployed to the data center or office environment.

Current minimum efficiency levels for the 10% load point are adequate since most of the servers will either not operate at that level or will not operate there for an extended period of time.

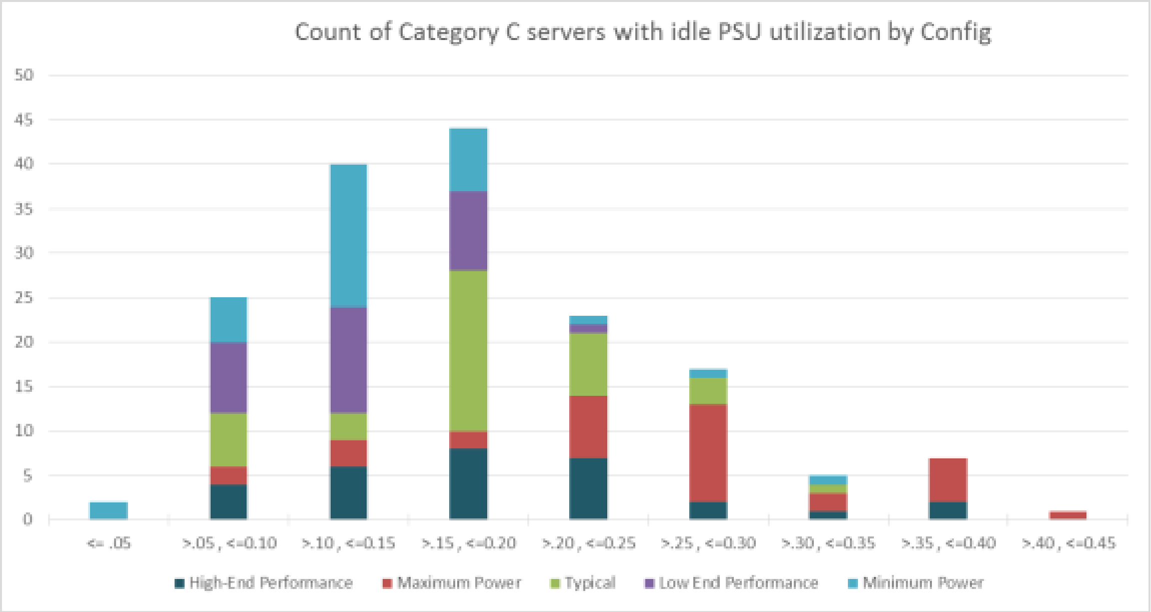


Figure 8: PSU Utilization at Idle for 2 socket servers

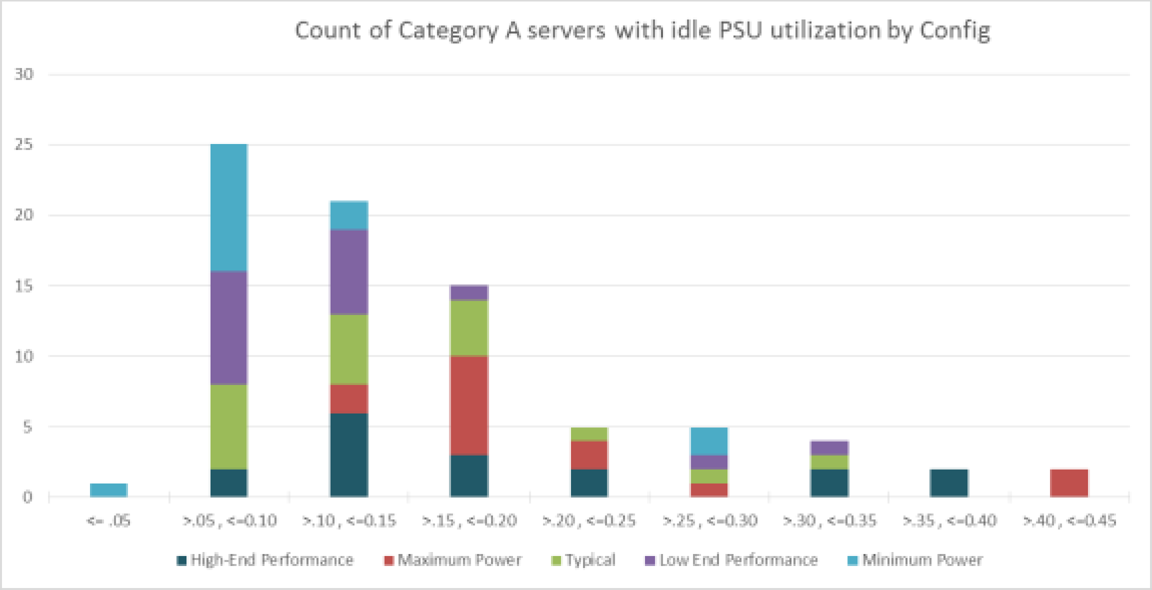


Figure 9: PSU Utilization at Idle for 1 socket servers

Lines 579-580: APA Idle Limit

The SERT WG is undertaking to find idle power data on expansion APAs. The intent is to supply any data we can obtain with the ITI October 16, 2017 comments.

ITI believes that the 30 watt maximum idle power value for expansion APAs is set too low and does not represent the idle power use of products currently on or planned for the market.

Appendix B: Resilient Server Definition

Changes in server technology require modifications to the resilient server definition. The "Proposed ITI Resilient Server Definition" in column B in Table 11 represents a consensus definition developed by ITI. Column A provides the current Version 2 resilient server with annotations regarding changes for reference.

ENERGY STAR v3 (Draft 1//2)	Proposed ITI Resilient Server Definition
Definition: A computer server designed with extensive Reliability, Availability, Serviceability (RAS) and scalability features integrated in the micro architecture of the system, CPU and chipset. For purposes of ENERGY STAR certification under this specification, a Resilient Server shall have the characteristics as described in Appendix B of this specification.	Definition: A computer server designed with extensive Reliability, Availability, Serviceability (RAS) and scalability features integrated in the micro architecture of the system, CPU and chipset. For purposes of ENERGY STAR certification under this specification, a Resilient Server shall have the following characteristics
Appendix B	
A. Processor RAS and Scalability- All of the following shall be supported: (1) Processor RAS: The processor must have capabilities to detect, correct, and contain data errors, as described by all of the following: (OK) (a) Error detection on L1 caches, directories and address translation buffers using parity protection; (OK) (b) Single bit error correction (or better) using ECC on caches that can contain modified data. Corrected data is delivered to the recipient (i.e., error correction is not used just for background scrubbing); (Changed slightly) (c) Error recovery and containment by means of (1) processor checkpoint retry and recovery, (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; and (OK – moved under System Recovery & Resiliency) (d) (1) Capable of autonomous error mitigation actions within processor hardware, such as disabling of the failing portions of a cache, (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. (Deleted)	Processor RAS: The processor must have capabilities to detect, correct, and contain data errors, as described by all of the following: 1. Error recovery by means of instruction retry for certain processor faults. 2. Error detection on L1 caches, directories and address translation buffers using parity protection; 3. Single bit error correction (or better) on caches that can contain modified data. Corrected data is delivered to the recipient as part of the request completion. System Recovery & Resiliency: No fewer than six of the following characteristics shall be present in the server: 1. Error recovery and containment by means of (1) data poison indication (tagging) and propagation which Includes mechanism to notify the OS or hypervisor to contain the error, thereby reducing the need for system reboots. (2) Containment of address/command errors by preventing possibly contaminated data from being committed to permanent storage. 2. 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.

(2) 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.~~ (Changed) (3) The server provides 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 (Deleted) B. Memory RAS and Scalability - All of the following capabilities and characteristics shall be present: (1) Provides memory fault detection and recovery through Extended ECC; (Deleted) (2) In x4 DIMMs, recovery from failure of two adjacent chips in the same rank; (Deleted) (3) 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; (Deleted - addressed under #3) (4) 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; (Deleted) (5) Uses resilient links between processors and memory buffers with mechanisms to recover from transient errors on the link; and (Deleted - addressed under #8)	3. Memory Mirroring: A portion of Available memory can be proactively partitioned such that a duplicate set may be utilized upon non-correctable memory errors. This can be implemented at the granularity of DIMMs or logical memory blocks. 4. Memory Sparing: A portion of available memory may be pre-allocated to a spare function such that data may be migrated to the spare upon a perceived impending failure. (New) 5. Support for making additional resources available without the need for a system restart. This may be achieved either by processor (cores, memory, IO) on-lining support, or by dynamic allocation/deallocation of processor cores, memory and IO to a partition. 6. Support of redundant IO devices (storage controllers, networking controllers) 7. Has I/O adapters or storage devices that are hot-swappable 8. Identify failing Processor-to-Processor lane(s) and dynamically reduce the width of the link in order to use only non-failing lanes or provide a spare lane for failover without disruption. (New) 9. Capability to partition the system such that it enables running instances of the OS or hypervisor in separate partitions. Partition isolation is enforced by the platform and/or hypervisor and each partition is capable of independently booting.(New) 10. 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. .

(6) Lane sparing in the processor-memory links. One or more spare lanes are available for lane failover in the event of permanent error. (Deleted - addressed under #4)	
C. Power Supply RAS: All PSUs installed or shipped with the server shall be redundant and concurrently maintainable. 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~~. (Partially deleted)	Power Supply RAS: All PSUs installed or shipped with the server shall be redundant and concurrently maintainable. 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.
D. Thermal and Cooling RAS: All active cooling components, ~~such as fans or water-based cooling,~~ shall be redundant and concurrently maintainable. 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. (Removed water-based cooling having redundant components)	Thermal and Cooling RAS: All active cooling components shall be redundant and concurrently maintainable. 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
E. System Resiliency: – no fewer than six of the following characteristics shall be present in the server: (Mostly addressed under 'System Recovery and Resiliency'; a few deleted) (1) Support of redundant storage controllers or redundant path to external storage; (Deleted – addressed under #6) (2) Redundant service processors; (Deleted – addressed under #6) (3) Redundant dc-dc regulator stages after the power supply outputs; (Deleted – addressed under #6) (4) The server hardware supports runtime processor de-allocation; (Deleted) (5) I/O adapters or hard drives are hot-swappable; ; (Deleted – addressed under #6) (6) Provides end to end bus error retry on processor to memory or processor to processor interconnects; (Deleted – addressed under #8)	

(7) Supports on-line expansion/retraction of hardware resources without the need for operating system reboot ("on-demand" features); (Deleted – addressed under #9) (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; (Deleted – addressed under #9) (9) Memory patrol or background scrubbing is enabled for proactive detection and correction of errors to reduce the likelihood of uncorrectable errors; and (Deleted) (10) Internal storage resiliency: Resilient systems have some form of RAID hardware in the base configuration, either through support on the system board or a dedicated slot for a RAID controller card for support of the server's internal drives. (Deleted)	
F. System Scalability – All of the following shall be present in the server: (1) Higher memory capacity: >=8 DDR3 or DDR4 DIMM Ports per socket, with resilient links between the processor socket and memory buffers; and (Deleted) (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 (Deleted)	

Table 11: Proposal for a revised Resilient Server Definition