# EPA ENERGY STAR SERVER STAKEHOLDER MEETING

ITI/TGG Technical Presentation

March 12, 2018

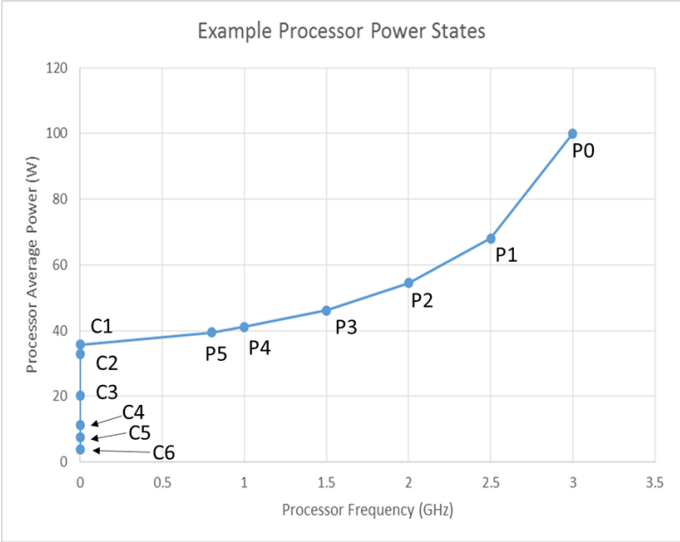
# Objectives for Discussion

- Demonstration of the presence of core and processor idle states during active portion of the SERT test.
  - Implementation of P-states and C-states during active testing.
  - Inter-relationship of P-states and C-states
  - Percentage of time cores are in c-states at each test interval for different processor architectures.
- The relationship between better power management in the active state and higher active efficiency scores and lower idle demand.
  - The effect of a better (lower value) dynamic range on efficiency scores and idle power values.
- SERT active efficiency test integrates assessment of server efficiency and idle power value into the single overall score.
  - Effectively assesses and differentiates power and performance/power characteristics of server products and configurations.
  - Removes the complexity of setting idle allowances for the different server components.

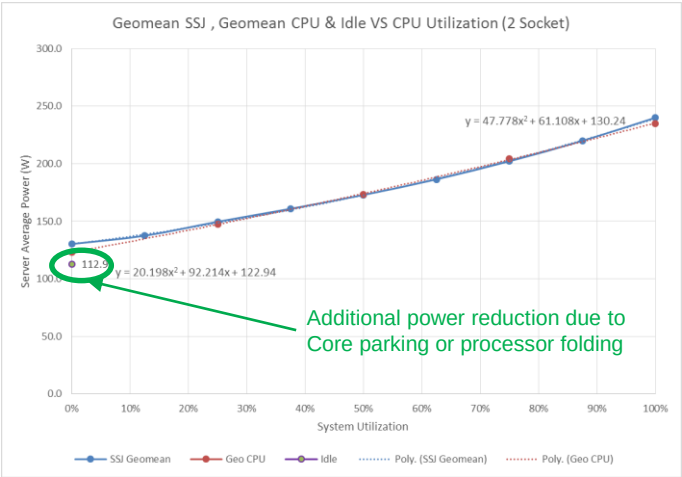
# Agenda

- Server Power Management States
- Processor Core State and Frequency Measurements
  - Intel x86
  - AMD x86
  - IBM Power®
- Impact of Dynamic Range on Active Efficiency Score
- Server selection by idle and active efficiency limits for servers of similar performance.
- Additional Version 3 Draft 2 Items
  - Memory adder model evaluation
  - PSU utilization levels at idle power.
  - APA definition and power limit.

# Server Power Management States




Hypothetical example of C-state and P-state effect on processor power demand.




Average power demand at each interval for Hybrid ssj and the Geometric mean of the CPU worklets with measured and curve fit idle values

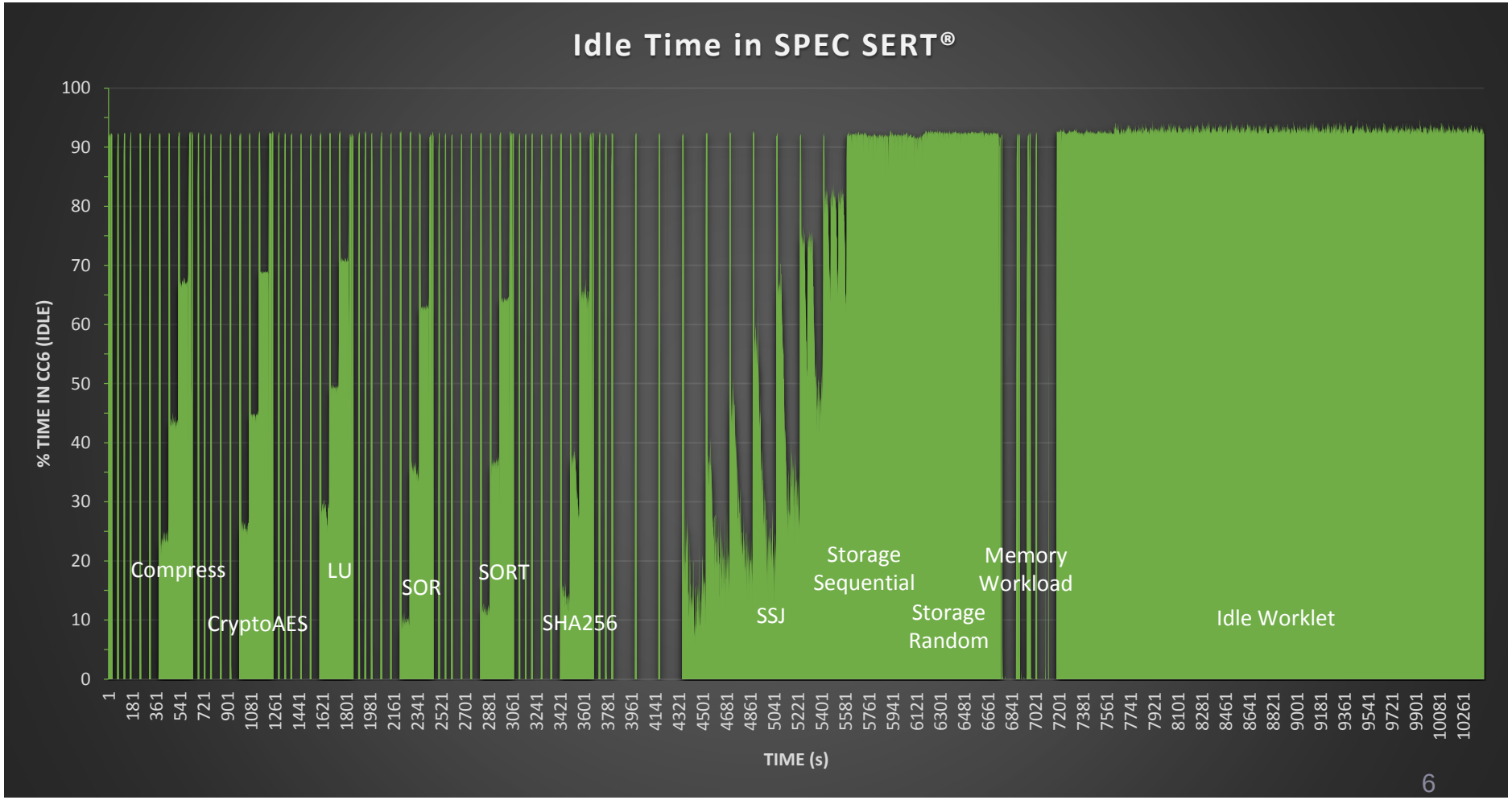
- All power management states are managed by the Operating System (OS)
- P-states: Active Voltage and Frequency Control:
  - As workload is reduced, voltage and frequency are reduced.
  - Each frequency increment has a matching voltage.
  - Slows server work rate
- C-states:
  - Turns off/idles specific processor sub-components: cache, core, etc.
  - The more items turned off, the lower the power level of the system.
- Processor/system level power down:
  - Referred to as “core parking” (x86) or “processor folding” (Power);
  - Further evacuation and shutdown of processor system sub-components with work isolated to a single core or group of cores.
  - Further reduces system power demand.
  - Incorporates power down for other components.
- Use of P-states and C-states come with response time/latency concerns.

# SERT Processor Test Data

- AMD
- Intel
- IBM Power®

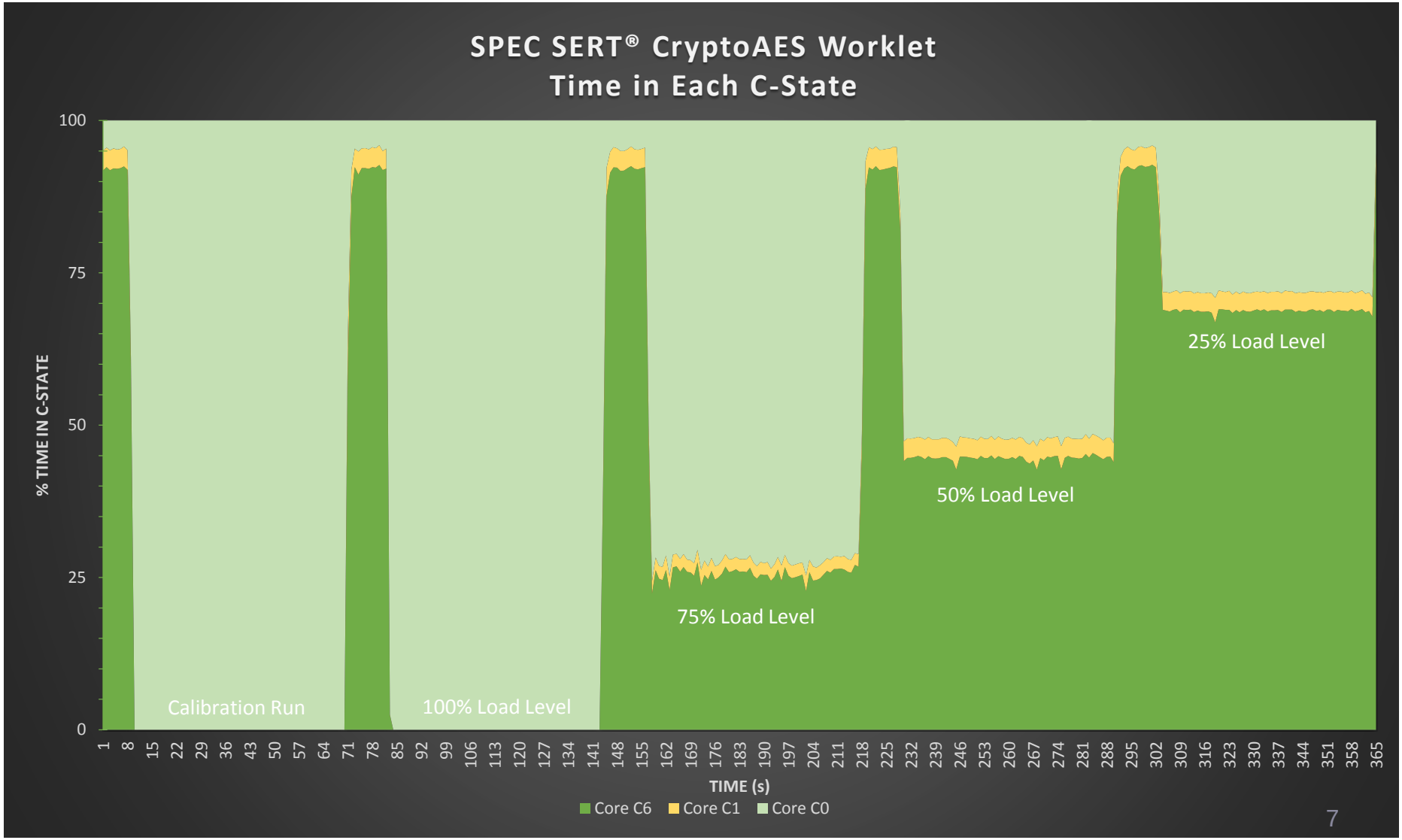
# AMD Processor Data

- Most SERT worklets at most load levels include significant time in idle
- The Storage Worklets are almost identical to the Idle Worklet
- Even the Idle Worklet is not 100% idle because of background tasks



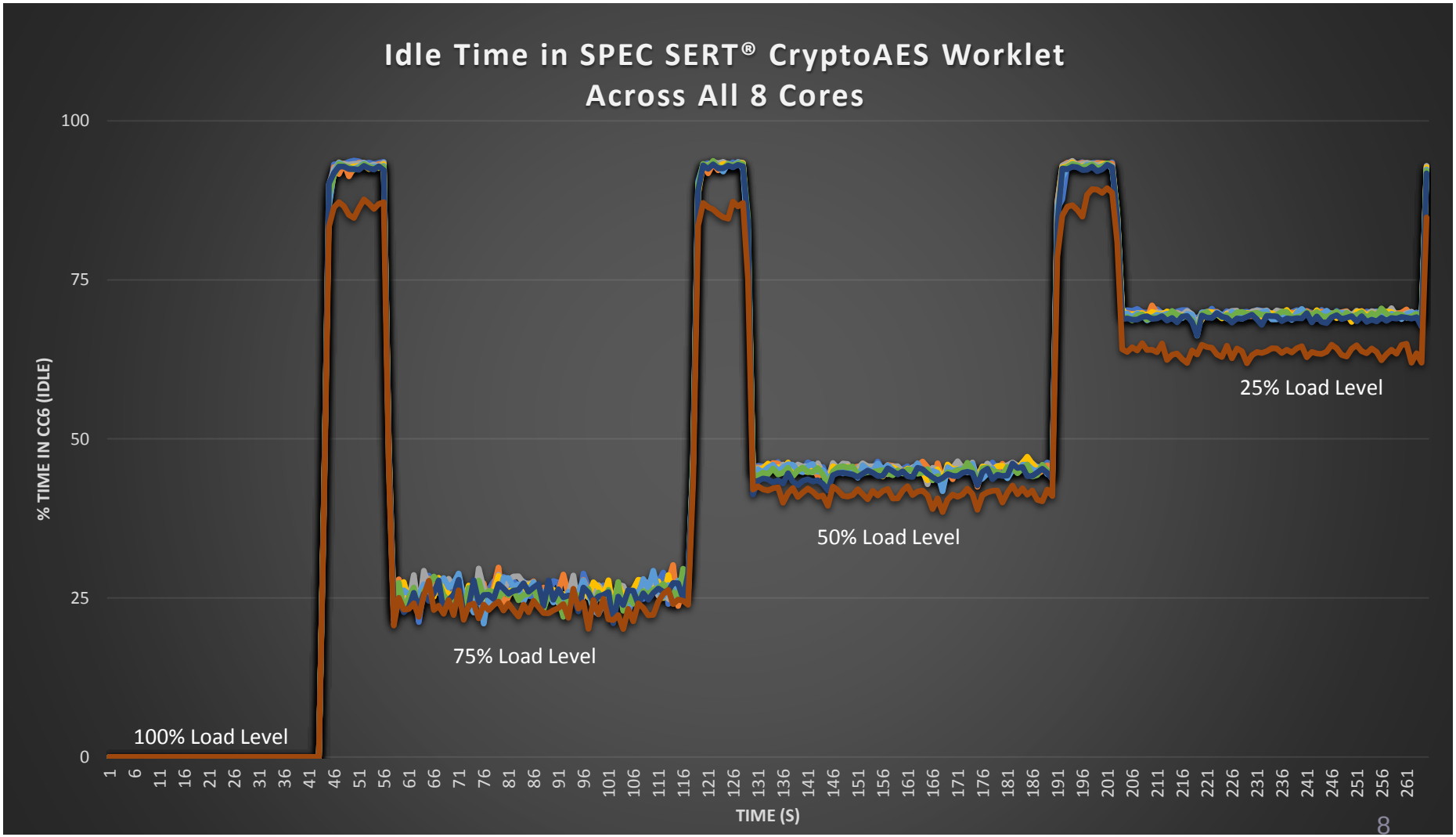
# AMD Processor Data

## Cores spend some time in C1



# AMD Processor Data

## Time in idle is mostly consistent across cores



# P-State/C-State Interaction

- Energy efficiency through CPU power management can be achieved via a combination of P-states and C-states.
  - The C-state residency at various levels of utilization is a function of the available P-state range and desired server configuration responsiveness
- In a balanced performance/energy configuration, the goal of the OS is to keep the cores maximally utilized.
  - At higher utilization levels the P-state range is leveraged to realize most of the power savings, with a limited use of C-states.
  - As the server utilization falls, the OS will lower the core frequency, until the full potential of P-States (voltage scaled frequency) has been reached at which point the C-states provide additional energy savings
- Going into deeper C-states, to further save power, comes with a trade-off of reduced system responsiveness when compared to P-states alone.

# Intel Processor Data

Figure 2 & 3 show CPU power management with a combination of P-states and C-states

**Figure 2**

- At 100% utilization, the cores are running @ 2.6GHz, but they drop to 1.5GHz for 75% utilization and 1.0 GHz for the 50% and 25% utilization runs. (1.0 GHz is the lowest frequency for the example CPU)
- Chart shows that the core frequencies rapidly approach idle level, when running SERT worklets at medium to low utilizations.
- This demonstrates the OS usage of the maximum P-state range available, to reduce energy consumption before exercising the C-states.
- C-state residency increases as the utilization level drops.

**Figure 3:**

- Demonstrates a similar behavior for the SSJ worklets, and confirms that the behavior extends to multiple workloads

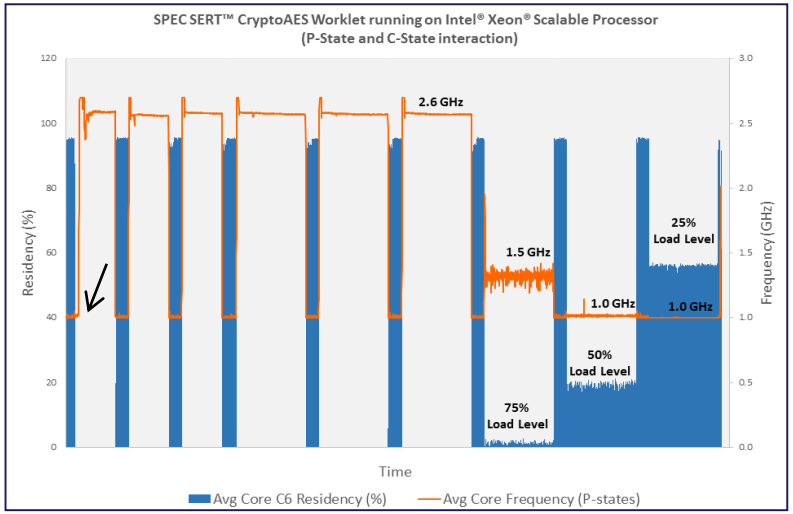



Figure 2: Average Core P-state and C-state residency for SPEC SERT® CryptoAES Worklet

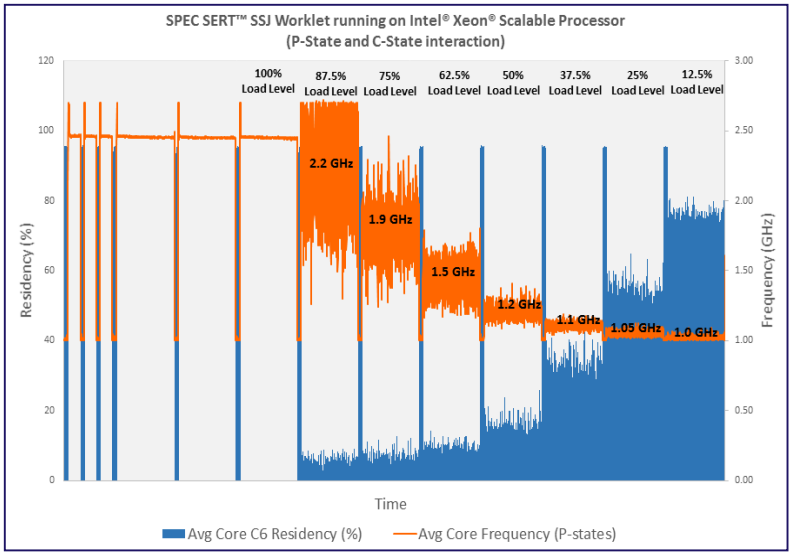



Figure 3: Average Core P-state and C-state residency for SPEC SERT® SSJ Worklet

# Intel Processor Data

- Figure 4:
  - Shows the full SPEC SERT run for x86 processor based platform, with the entire P-state range enabled but not graphed to keep the chart readable.
  - Demonstrates the increase in core idle state residency (Core C6) as a function of decreased CPU utilization.
  - There is very good correlation in the behavior of each worklet across different configurations and CPU architectures depending on how P-states are used.

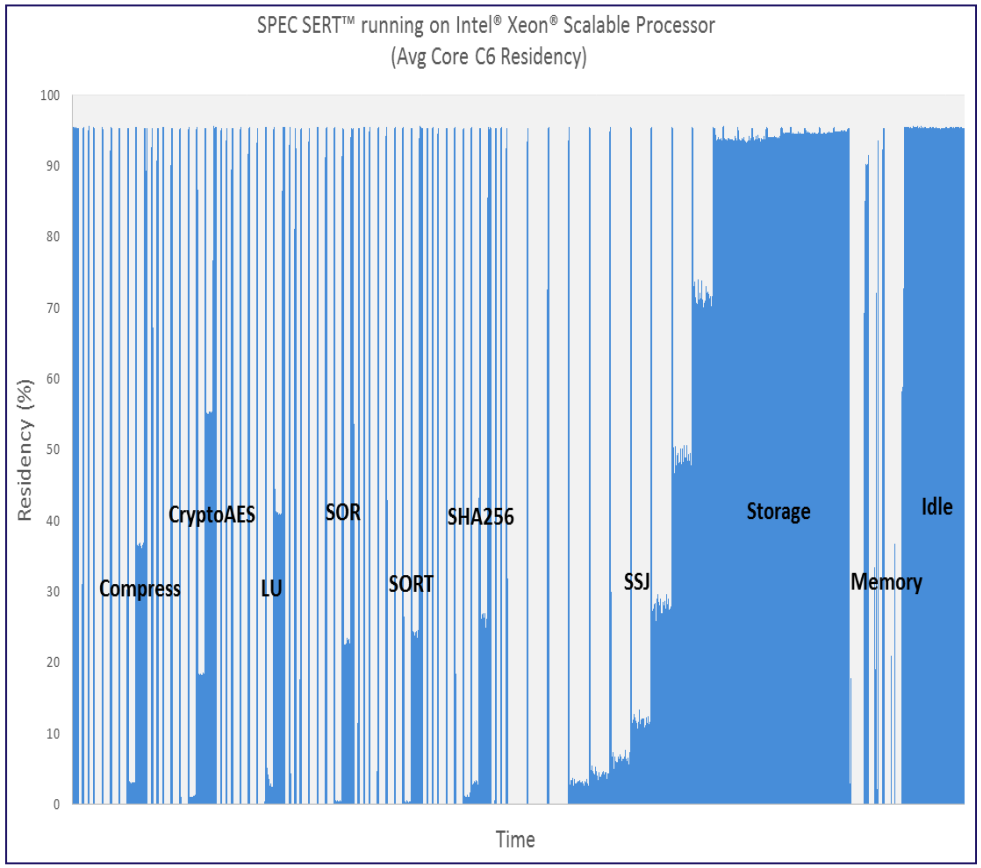
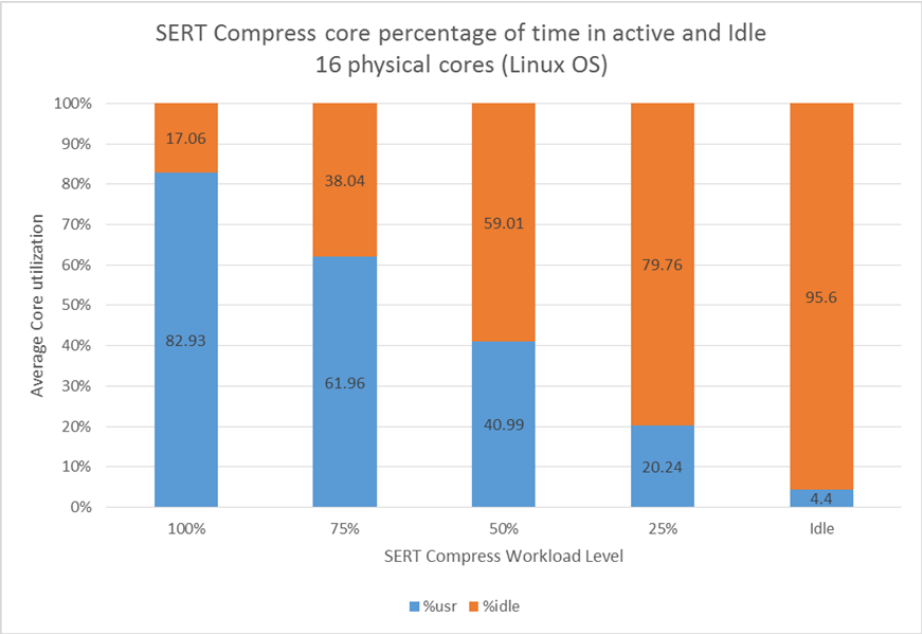
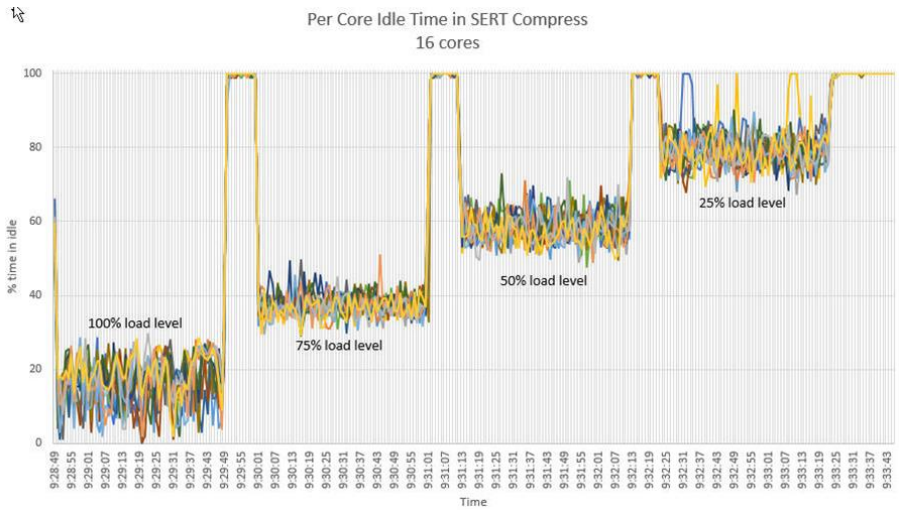



Figure 4: Percentage of time processor cores are in C6 for a full SERT test

**The SERT Active Efficiency Score does reward servers that use aggressive power management.**

# Power® Processor Data

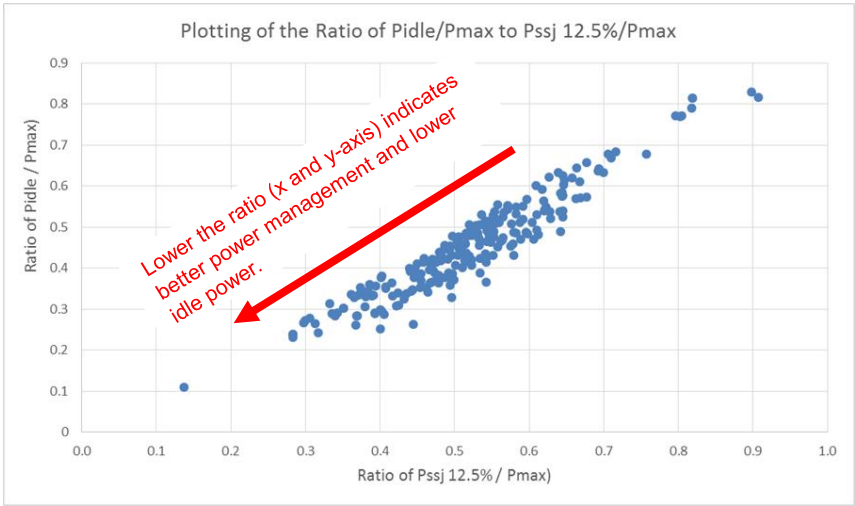



- Test was run with single thread on each core to be able to observe core state.
- For the power processor using a Linux operating system, there are 11 stop states, (equiv. to C-states)
  - Stop5 is the highest state reached during active operation.
    - It reduces core power by 65%.
    - It sets tolerable latency times.
  - Stop6 to Stop11 involves consolidating work to a single core or core group, flushing cache and shutting down remaining cores. It also enables entering into lower memory power states.
    - Introduces significant latency and response time impacts. Not acceptable for “active operation”.
    - Removes 98% of the power from the processor.
    - BMC is still active (separate control chip).
- Power processors also have P-state control.
  - Linux OS offers several P-state settings.
  - Deeper P-state settings are not favored by IBM customers.

# Impact of Server Dynamic Range on Idle Power

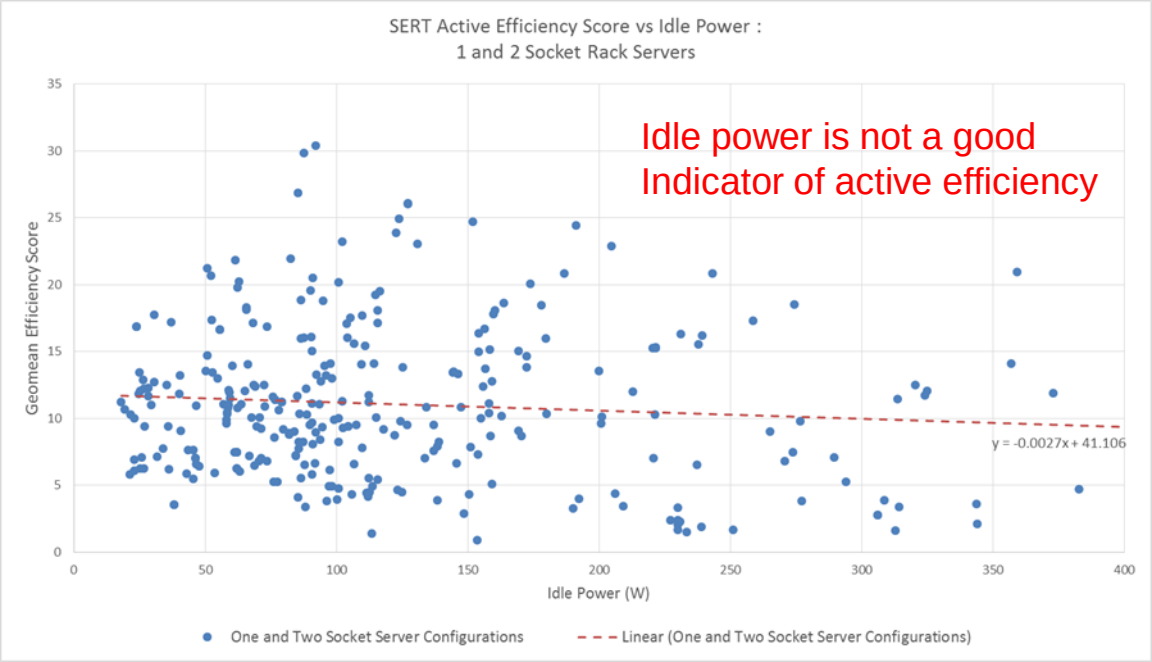
| Hypothetical Comparison of Worklet Score with varied Dynamic Range | | Measurement (Watts) | | Interval and Worklet Efficiency Score | | |
|---|---|---|---|---|---|---|
| Performance Interval | Performance Measurement | Server A | Server B | Interval Efficiency Score A | Interval Efficiency Score B | % Decrease A vs. B |
| 100% | 1000 | 100 | 100 | 10.0 | 10.0 | |
| 75% | 750 | 80 | 90 | 9.4 | 8.3 | -12% |
| 50% | 500 | 60 | 80 | 8.3 | 6.3 | -24% |
| 25% | 250 | 40 | 70 | 6.3 | 3.6 | -43% |
| Idle | | 20 | 60 | | | |
| Geometric Mean of All Intervals | 553 | 66 | 84 | 8.4 | 6.6 | -21% |




- A higher level of power management will typically deliver a better active efficiency score for servers of like performance.
  - Reducing the power at each interval improves the interval, worklet, workload and overall efficiency score.
  - A larger dynamic range will result in a lower idle power and higher active efficiency for a server.
  - Allowing idle power to rise within a defined configuration power profile will result in a reduction of the overall active efficiency score.
- An analysis of the ITI/TGG dataset shows that there is a strong correlation between the dynamic range at 12.5% power (ssj) and idle power.
  - There are variations within servers of like idle power depending on the server configuration and power profile.

Key Point: A server with a larger dynamic range (smaller ratio of idle or 25% power to maximum power) will generally have a higher active efficiency value and a lower idle power than a server with a higher ratio: the SERT test and active efficiency score incorporate and are indicative of server idle power.

# Relationship of Idle Power and Active Efficiency Values




- Servers with high active efficiency values exist across the full range of idle power.
- Idle power is not a good indicator of active efficiency
- Server efficiency, as measured by workload delivered per unit of energy consumed is a function of:
  - the server configuration;
  - chosen components and hardware; and
  - firmware and OS functions and enablement.

# Comparison of Servers with Similar Performance

| Case | Qty w same # Dep | ITI Identifier | Configuration | SERT 2.0 Active Eff. | Idle | Diff: 25% power - idle | Weighted Geometric Mean Perf. | Weighted Geomean Server Pwr | Interval Power 25.0% | Interval Power 100.0% | SSJ 25%/ 100% power (ssj DR) | Idle power/ max powr (DR) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7 | 6 | ITI769 | Maximum Power | 11.25 | 102.3 | 94.4 | 2829 | 251.4 | 196.7 | 294.7 | 51.8% | 34.7% |
| | | ITI193 | Typical | 13.43 | 144.5 | 41.3 | 2850 | 212.3 | 185.8 | 253.7 | 59.8% | 57.0% |
| 11 | 9 | ITI872 | High-End Performance | 23.19 | 102.2 | 146.3 | 7794 | 336.2 | 248.4 | 463.4 | 37.2% | 22.0% |
| | | ITI795 | High-End Performance | 26.11 | 127.3 | 93.9 | 7609 | 291.4 | 221.2 | 403.4 | 37.8% | 31.6% |
| 6 | 10 | ITI863 | Minimum Power | 16.43 | 52.8 | 65.5 | 2444 | 148.8 | 118.3 | 178.2 | 49.5% | 29.6% |
| | | ITI717 | Typical | 20.21 | 62.8 | 37.0 | 2484 | 122.9 | 99.9 | 149.6 | 49.5% | 42.0% |
| 9 | 7 | ITI157 | Maximum Power | 18.85 | 86.6 | 75.3 | 3854 | 204.4 | 161.9 | 264.8 | 44.7% | 32.7% |
| | | ITI874 | Maximum Power | 19.58 | 90.2 | 79.8 | 3914 | 199.9 | 170.0 | 235.2 | 60.3% | 38.4% |

| Table Legend | |
|---|---|
| | Higher Active Efficiency |
| | Lower Idle Power |
| | Higher Difference 25% and Idle |

- There are 74 groups of two socket rack servers with two or more servers with similar performance level.
  - In 62 of the 74 groups (84%), active efficiency and idle power pick the same server. 4 examples are provided above
  - These servers use the least active power to do the work.
  - In 12 of 74 cases, they pick different servers.
- The idle power selection results from a more aggressive implementation of deeper power management functions when no work is present.
  - In these situations, the server with the higher active efficiency does work more efficiently
  - For the selected idle server, it provides a deeper idle power management when no work is present.
  - In the one case (9), the active efficiency and idle power values are almost the same.
- This analysis does not make the point that idle power is an equally good metric:
  - This analysis is specific to servers of like performance.
  - As we have shown in a separate analysis, idle power applied as a market entry metric is biased toward low power servers.

# Comparison of Servers with Similar Performance

| Case | Qty w same # Dep | ITI Identifier | Configuration | SERT 2.0 Active Eff. | Idle | Diff: 25% power - idle | Weighted Geometric Mean Perf. | Weighted Geomean Server Pwr | Interval Power 25.0% | Interval Power 100.0% | SSJ 25%/ 100% power (ssj DR) | Idle power/ max powr (DR) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | ITI341 | Typical | 6.64 | 52.8 | 33.1 | 736 | 110.9 | 85.9 | 141.9 | 42.7% | 37.2% |
| | | ITI892 | Low End Performance | 9.92 | 58.2 | 8.9 | 732 | 73.9 | 67.0 | 83.0 | 71.0% | 70.1% |
| 2 | 3 | ITI69 | High-End Performance | 11.83 | 40.2 | 23.2 | 922 | 78.0 | 63.4 | 100.8 | 44.1% | 39.9% |
| | | ITI734 | Minimum Power | 12.48 | 54.7 | 11.8 | 923 | 73.9 | 66.5 | 81.9 | 71.4% | 66.8% |
| 3 | 4 | ITI277 | Minimum Power | 13.21 | 40.5 | 24.5 | 1102 | 83.5 | 65.0 | 108.1 | 43.6% | 37.5% |
| | | ITI732 | Maximum Power | 13.48 | 53.7 | 16.4 | 1108 | 82.2 | 70.1 | 97.5 | 60.3% | 55.1% |
| 4 | 9 | ITI566 | Typical | 11.21 | 79.2 | 50.2 | 1874 | 167.1 | 129.4 | 217.3 | 42.4% | 36.4% |
| | | ITI365 | Low End Performance | 13.94 | 95.6 | 28.3 | 1918 | 137.6 | 123.9 | 149.7 | 71.4% | 63.9% |
| 5 | 3 | ITI143 | Typical | 8.25 | 138.9 | 92.3 | 2287 | 277.1 | 231.2 | 331.2 | 45.6% | 41.9% |
| | | ITI206 | Low End Performance | 9.04 | 169.5 | 47.9 | 2228 | 246.4 | 217.4 | 279.1 | 65.3% | 60.7% |
| 6 | 10 | ITI863 | Minimum Power | 16.43 | 52.8 | 65.5 | 2444 | 148.8 | 118.3 | 178.2 | 49.5% | 29.6% |
| | | ITI717 | Typical | 20.21 | 62.8 | 37.0 | 2484 | 122.9 | 99.9 | 149.6 | 49.5% | 42.0% |
| 7 | 6 | ITI769 | Maximum Power | 11.25 | 102.3 | 94.4 | 2829 | 251.4 | 196.7 | 294.7 | 51.8% | 34.7% |
| | | ITI193 | Typical | 13.43 | 144.5 | 41.3 | 2850 | 212.3 | 185.8 | 253.7 | 59.8% | 57.0% |
| 8 | 8 | ITI379 | Typical | 15.98 | 86.4 | 68.8 | 3069 | 192.1 | 155.2 | 243.5 | 46.3% | 35.5% |
| | | ITI357 | Maximum Power | 17.54 | 105.3 | 43.8 | 3069 | 175.0 | 149.1 | 205.0 | 59.6% | 51.4% |
| 9 | 7 | ITI157 | Maximum Power | 18.85 | 86.6 | 75.3 | 3854 | 204.4 | 161.9 | 264.8 | 44.7% | 32.7% |
| | | ITI874 | Maximum Power | 19.58 | 90.2 | 79.8 | 3914 | 199.9 | 170.0 | 235.2 | 60.3% | 38.4% |
| 10 | 12 | ITI244 | Maximum Power | 19.52 | 116.5 | 118.8 | 6108 | 312.9 | 235.3 | 418.5 | 36.5% | 27.8% |
| | | ITI766 | Low End Performance | 23.07 | 131.0 | 111.6 | 6626 | 287.2 | 242.6 | 328.0 | 61.9% | 39.9% |
| 11 | 9 | ITI872 | High-End Performance | 23.19 | 102.2 | 146.3 | 7794 | 336.2 | 248.4 | 463.4 | 37.2% | 22.0% |
| | | ITI795 | High-End Performance | 26.11 | 127.3 | 93.9 | 7609 | 291.4 | 221.2 | 403.4 | 37.8% | 31.6% |
| 12 | 2 | ITI317 | Maximum Power | 15.44 | 110.9 | 83.9 | 3664 | 237.4 | 194.8 | 300.4 | 49.1% | 36.9% |
| | | ITI496 | High-End Performance | 16.72 | 156.4 | 46.8 | 3740 | 223.7 | 203.2 | 251.1 | 69.8% | 62.3% |

| Table Legend | |
|---|---|
| | Higher Active Efficiency |
| | Lower Idle Power |
| | Higher Difference 25% and Idle |

## Key Points:

1. Active efficiency is effective in selecting the lower idle power servers in most instances.
2. Idle power selection results from deeper, lower power processor level C-states with higher latency.
3. The correlation discussed here is specific to servers with like performance. Across the full dataset, active efficiency best identifies the truly efficient servers.

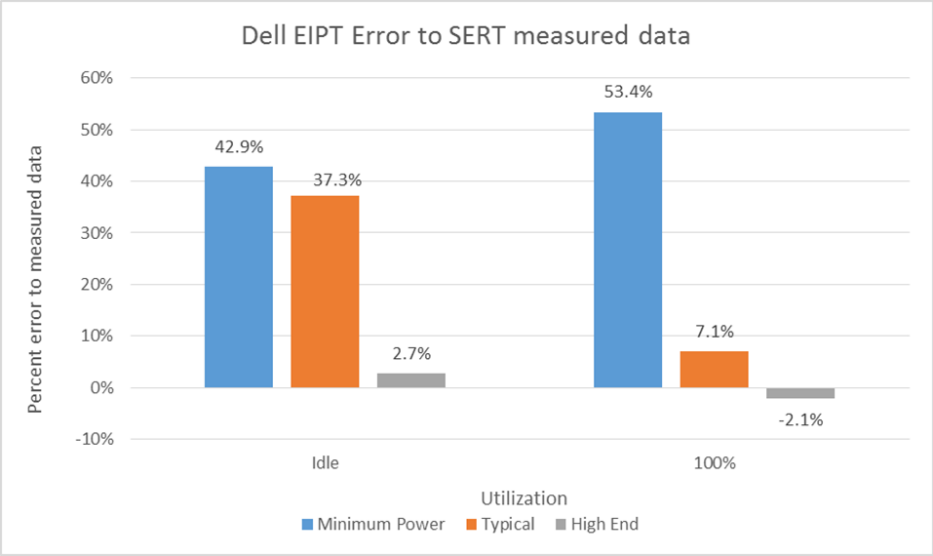
# Analysis of Power Calculator Data

| Dell EIPT R740 Configuration Options | # of Options | # DIMMS | #DIMM Sizes | Totals |
|---|---|---|---|---|
| Processor | 41 | | | 41 |
| Memory | | 24 | 5 | 30 |
| Disk Drives | 134 | | | 13 |
| I/O Cards | 17 | | | 17 |
| GPU Cards | 18 | | | 18 |
| Fans | 3 | | | 3 |
| Power Supplies | 4 | | | 4 |
| Storage Adapters | 11 | | | 11 |
| Bus Adapters | 20 | | | 10 |
| **Total Configuration options** | **253** | | | 147 |
| **Possible Configurations** | | | | **6,458,680,800** |
| Load Levels to test | 13 | | | 13 |
| Total tests to run | | | | 83,962,850,400 |
| **Short cut 1 test per config option at load levels** | | | | **1911** |

Available Component Options for Dell EIPT R740
EIPT = Enterprise Infrastructure Planning Tool

- All possible configurations and load levels cannot be tested.
- Even Short cut testing is 1 month of test time excluding configuration changes
- Simplified sample testing required to create models in reasonable time.
  - Test with most often ordered components
- Calculation accuracy focused where customers need it most: operational energy use.

# Analysis of Power Calculator Data




Dell system with SERT data configurations modeled in EIPT to check accuracies.

Calculators focus on operational energy use.

- Server Energy calculator tools are intended to allow customers to plan server implementations that do not exceed infrastructure capabilities of their facilities.
- Accuracy for idle power estimates is not a focus.
- Accuracies very good for high end systems at high utilization and are worst for low end configs at idle.
- This is just one example for one model.

# Error in calculating W/GB due to 1W resolution

| | Added # of DIMMs | | | |
|---|---|---|---|---|
| DIMM Size | 4 | 12 | 28 | 44 |
| 8 | 80.1% | 40.1% | 20.0% | 13.4% |
| 16 | 40.1% | 20.0% | 10.0% | 6.7% |
| 32 | 20.0% | 10.0% | 5.0% | 3.3% |
| 64 | 10.0% | 5.0% | 2.5% | 1.7% |
| 128 | 5.0% | 2.5% | 1.3% | 0.8% |

Inherent Error in W/GB calculations due to 1 W resolution in the power calculator

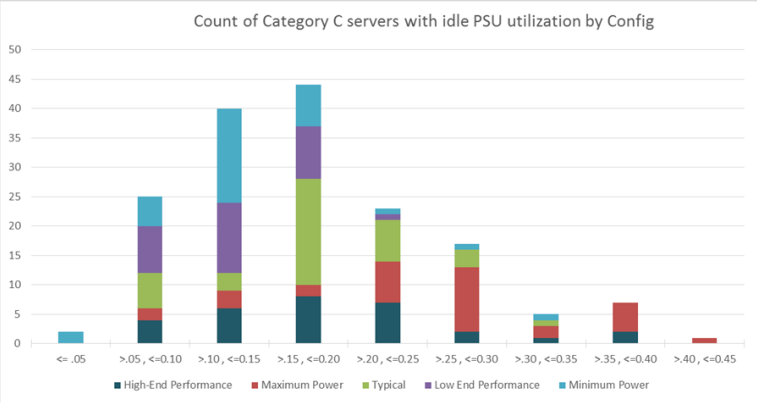
Power Calculator data is not intended to be used to set regulatory compliance values:

- Dell EIPT has 1W output resolution
- Calculation of W/GB using 2 measurements has large inherent error due to output resolution

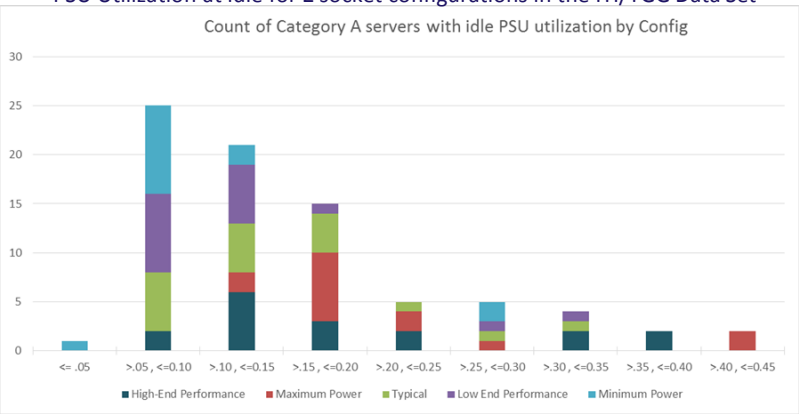
Idle allowance values should be set based on measured component values:

- Provides true indication of power use.
- Accounts for differences between suppliers.
- Data provided by TGG in publically available comments to both EPA and EU.

# Power Supply Utilization at Idle




PSU Utilization at Idle for 2 socket configurations in the ITI/TGG Data Set




PSU Utilization at Idle for 1 socket configurations in the ITI/TGG Data Set

- Low-end configurations largely utilize the PSU from 5-15%
  - Low-end configurations will not typically be found in data center and operating environments.
- High-end and typical configurations are largely utilized above 15%
  - Most customers buy at or above a typical configuration.
  - Most servers offer two or more PSU sizes.
- Not necessary to tighten the 10% efficiency limit:
  - Creates conflicts in increasing efficiency at 20% and 50%
  - Right Sizing power supplies is the critical step.

Key points: PSU loadings for server configurations typically sold are above the 10% PSU load point, and many servers offer two or more PSU options to enable right sizing of the PSUs.

# APA Exclusion

- APA revised definition:
  - Clear indication that the idle limit is per APA device.
  - Recognition that in some cases the high capacity switches exert a power demand where the APA is removed and these systems should also be excluded. In general, these systems would currently qualify for the HPC exemption.
- High Performance Expansion APA Exclusion:
  - Some high performance APAs presently in production exceed the 30 W ENERGY STAR limit.
  - Future very high performance APAs may also exceed the 30 W ENERGY STAR limit.
  - Propose that APAs with a local memory bandwidth of >700 GB/sec be excluded from Version 3 idle limit but subject to data collection. (page 24 of the ITI 10/30/2017)