Recommendation for Server Activity Efficiency Metric
ENERGY STAR® Version 3 Draft 1

Recommendation to the ENERGY STAR® Draft 1 Comments Request

- Create a single number SERT metric calculated as, the Weighted Geomean of 60% CPU workload (including SSJ), 35% Memory workload, and 5% Storage workload
  - Calculate worklet scores using a geometric mean, instead of arithmetic mean, as it is currently calculated in SPEC SERT®
  - Assessing the Best Method to combine Flood(2) and Capacity(2) scores
  - Include storage, but only at a 5% weight.
    - The storage weighting will be taken from the memory percentage to maintain the CPU percentage.

- Maintain ENERGY STAR product family methodology, but require testing and SERT data reporting for 3 configurations: low-end, high-end and typical configurations
  - Maintain separate regulatory categories for servers based on socket count, form factor (tower, rack, blade), and resilient vs. managed
  - Submit SERT tool data in xml format for three configurations: low-end, typical and high-end.
  - Develop Threshold requirements for low-end and high-end configurations: these configurations should have well defined configuration requirements.
  - Create separate storage/I/O server category.

Key Reasons for the Recommendation

- Assessment of a number of metrics indicates that the 60/35/5 weighting provides the best correlation to the actual power used by servers when deployed in a data center
  - High efficiency scores do not result in less actual power usage in a deployed data center with all metrics. Deployed power analysis, however, shows that with the recommended metric, a higher efficiency scores does result in lower actual power usage in a data center
  - Assessment of workloads from SPEC CPU, based on 29 real world applications, indicates 62% of workloads are limited by CPU capability and 38% by memory bandwidth, justifying the weights chosen (see paper "SERT Metric Weighting Proposal")
  - Servers depend on storage for their operations, making it appropriate to include a small storage factor
    - Large storage weightings significantly changes efficiency results, making the metric not correlate to "typical" server workloads
    - A separate category needs to be created for storage servers. The WG and SPECpower Committee plan to assess how to best define and address storage servers

- The geometric mean offers the best method to combine the SERT Tool data (see the papers "SERT Metric Analysis" and "Mean Calculation Methodology Analysis")
  - Arithmetic mean overemphasizes higher power consumption, Harmonic mean overemphasizes lower power consumption
  - Combining worklet load level power with a Geomean properly weights the impact of the server's dynamic range, the ratio of minimum to maximum power, favoring servers with a broader dynamic range
  - Geomean offers the best correlation between the metric and the deployed power value for the server. Servers with higher results also use less deployed power. See the paper "Deployed Power Analysis Methodology"
  - Geomean is the only mean that works effectively with ratios, and the worklets scores are ratios of performance over power
  - Using the geomean instead of sum to divide performance by power at the worklet level, resulted in significantly better deployed power correlation as well as removed the overweighting of higher load levels

- The weighted geomean is the best mean to combine workloads because the CPU, Memory, and Storage workloads' scores have different magnitudes

Key Reasons for the Recommendation

- Deployed Power Analysis and metric results show a distinct difference between high-end and low-end servers
  - Dataset† analysis shows distinct efficiency profiles between small and large servers, with large servers obtaining higher scores.
  - Each server size has specific use cases, for which it is the most efficient solution. Without separate categories, regulation will likely encourage selection of servers of the wrong size, which are inefficient for the application

† Note: Assessment based on a combined dataset with data supplied by ITI/TGG (EPA certified servers) and the SPECpower Committee

Observations based on the analysis work

- The combined efficiency metric value is the same whether the combining the worklet efficiency scores or the combined performance and power interval measurements as long as the aggregation methods at the interval level are the same. (see the paper "SERT Metric Analysis")

- Geomean provides the best combination method for combining the measured, interval performance and power data for each worklet, the worklet scores by component type and combining the component scores. (see the papers "SERT Metric Analysis" and "Mean Calculation Methodology Analysis")

- The 60/35/5 CPU/Memory/Storage weighting provides the best ranking match to the deployed power rankings across the range of workloads and power levels evaluated for the deployed power

- Page 6 details the overall metric analysis criteria and results from the Working Group's metric analysis.
  - The WG started with a broad group of combined metrics, but narrowed that group down to the 2 listed choices.
  - The weighted geomean with geomeaned interval data metric offered the best results against the ranking criteria.

- The Deployed power analysis offers an effective methodology for validation that a more efficient server correlates to lower energy consumption when the server is installed in the data center:
  - For each configuration, the number of servers required to execute a target workload is calculated using different server performance estimates calculated using the aggregated SERT performance measurements.
  - The number of servers required to execute the workload is multiplied by an geomeaned, weighted power value to calculate the deployed power of this servers in the data center.
  - The deployed power is plotted or ranks are compared against the efficiency metric to validate that a more efficient server results in lower power use in the data center.
  - The deployed power methodology is explained in detail in the document "Deployed Power Analysis Methodology Description".

Metric Selection Criteria Definitions

- Deployed Efficiency: Does the metric favor servers which result in lower datacenter deployment power?
  - Details: Analysis described in TGG "Assessment of Aggregated Algorithm" paper
- Single Result: Is the metric a single number?
  - Details: More than one government has stated this as a requirement of their server energy efficiency program
- CPU/Memory Weighting: Do the metric's CPU and memory weights match real world server applications?
  - Details: "Data Based Weighting for the SPEC SERT Overall Metric" paper analyzes the CPU and memory characteristics of 29 server applications and concludes the weights should be in the vicinity of 60% CPU and 40% Memory.
- Server Config Consistency: Does the metric favor one size of server over another?
  - Details: Analysis in progress
- Design Incentives: Will the metric possibly incent important design changes which decrease actual energy efficiency?
  - Details: See "Assessment of the Metric's ability to incentivize or influence future server designs" in TGG "Assessment of Aggregated Algorithm" paper
- Single Optimization: Relatively, how close are the impacts of a specific optimizations to the actual impact for end users?
  - Details: Metrics which include all the SERT CPU worklets are less susceptible than those with some worklets removed
- Spot Checks: Does the metric agree with well understood results of energy efficiency comparisons?
  - Details: Planned future TGG analysis
- Intuitive: Is the metric simple? Is it apparent to non-experts the metric likely measures server energy efficiency?

Metric Selection Criteria Summary for SPEC SERT®

	Deployed Efficiency	Single Result	CPU/ Memory Weighting	Server Config Ranking	Design Incentives	Single Optimization	Spot Checks	Intuitive
Weighted Geomean Efficiency with Geomean Intervals (60%,35%,5%)	🔋	✓	✓	✓	✓	✓	Future Work	✓
Weighted Geomean Efficiency with summed intervals (65%, 30%, 5%)	🔋	✓	✓	✓	✓	✓	Future Work	✓
Intertek Improved	🔋	✓	✗		✓	✓		✓
SPEC Compute or Memory or Storage with summed interval data	🔋	✗	✓		✗	✓		✗
Four CPU Worklets Efficiency	🔋	✓	✗		✗	✗		✓
Hybrid SSJ Efficiency	🔋	✓	✗		✗	✗		✓
(CPU + Flood2) / ((4*Idle + Max)/5)	🔋	✓	✓		✗	✓		✗

← Importance

Ranking Comparisons for the Final Metric Options

	Metric	Deployed Power (Idle)	Deployed Power (Active)	25% Utilization	50% Utilization	100% Utilization	CPU Intensive	Memory Intensive
Correlation between deployed power and server efficiency metric	Weighted Geomean (65,22.5,7.5,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	95.8%	100.0%	99.4%	99.9%	99.2%	100.0%	99.7%
	Weighted Geomean (65,15,15,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	95.9%	100.0%	99.5%	99.9%	99.2%	100.0%	99.6%
	Weighted Geomean (65,26.3,8.7,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	95.7%	100.0%	99.5%	99.9%	99.4%	100.0%	99.7%
	Weighted Geomean (60,17.5,17.5,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	95.7%	100.0%	99.5%	99.9%	99.4%	99.9%	99.7%

	Metric	Deployed Power (Idle)	Deployed Power (Active)	25% Utilization	50% Utilization	100% Utilization	CPU Intensive	Memory Intensive
Average Size of Rank Mismatch	Weighted Geomean (65,22.5,7.5,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	38.44	0.28	14.11	5.41	15.85	3.82	10.56
	Weighted Geomean (65,15,15,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	38.11	0.26	14.12	5.87	16.01	3.09	11.36
	Weighted Geomean (6026.3,8.7,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	38.97	0.18	13.75	4.98	14.54	3.89	9.22
	Weighted Geomean (60,17.5,17.5,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	39.15	0.22	13.67	5.60	14.19	4.53	10.24

The correlations represent the correlation between the ranking of each efficiency metric to the deployed power value ranking. It was calculated using the Excel correl() function.

For the correlation data, all of the values are extremely close together. All four metric options provide sound results.

The average size of the rank mismatches are largely consistent between the four metric options.

As with the correlation results, the differences between the four metrics are limited.

Percent of Exact Ranking Matches for the Final Metric Options

	Metric	Deployed Power (Idle)	Deployed Power (Active)	25% Utilization	50% Utilization	100% Utilization	CPU Intensive	Memory Intensive	Total Percentage
Percent of Rank Matches between the Deployed Power and the Weighted Geomean Efficiency Scores	Weighted Geomean (65,22.5,7.5,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	1.1%	86.9%	3.8%	12.3%	3.9%	14.6%	8.5%	131.2%
	Weighted Geomean (65,15,15,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	2.6%	83.4%	4.4%	10.0%	3.4%	11.5%	5.7%	121.2%
	Weighted Geomean (60,26.3,8.7,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	1.6%	89.5%	3.3%	9.7%	3.1%	16.4%	7.2%	130.9%
	Weighted Geomean (60,17.5,17.5,5)_ratio of CPU/Flood(2)/Capacity(2)/Storage	1.8%	81.1%	4.4%	9.5%	3.8%	15.8%	3.6%	120.0%

The % Rank Matches data is shaded with the lowest value(s) of the group red, the median value(s) yellow, and the highest value(s) green.

The Weighted Geomean 60/35/5 Metric weighting minimizes the maximum size of the rank mismatch while maximizing the number of rank matches for power evaluations as compared to other tested combination methods.

The Flood(2)/Capacity(2) weighting of 75%/25% provides roughly the same deployed power match for either the 60/35/5 or 65/30/5 CPU/memory/storage weighting. The Flood(2)/Capacity(2) 50%/50% weighting provides a 8% lower overall match ranking with the 60/35/5 CPU/memory/storage weighting providing a slightly better match ranking.

In general the correlation between the efficiency score and deployed power (idle) is low, indicating that idle power is not representative of server efficiency.
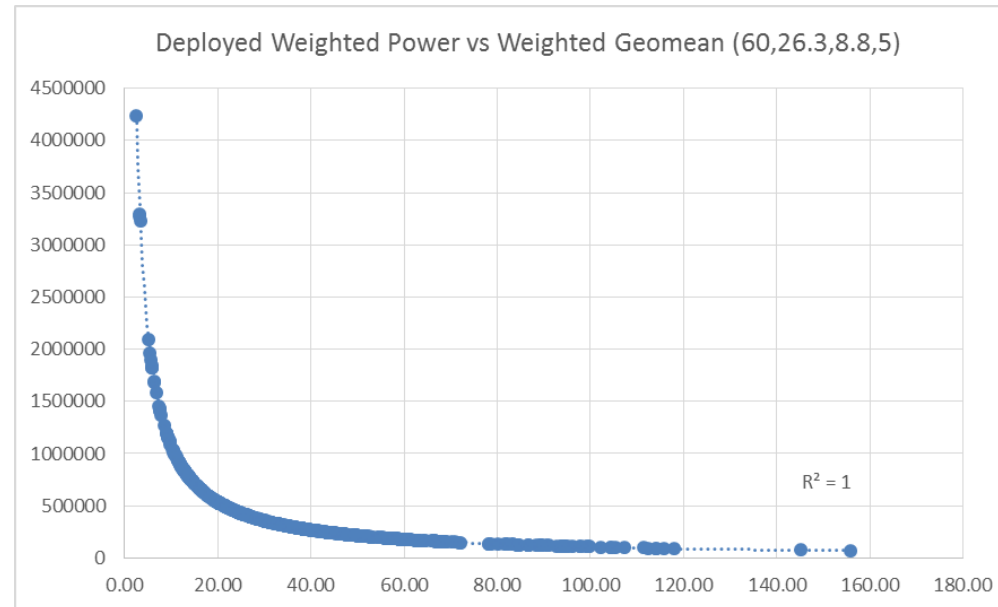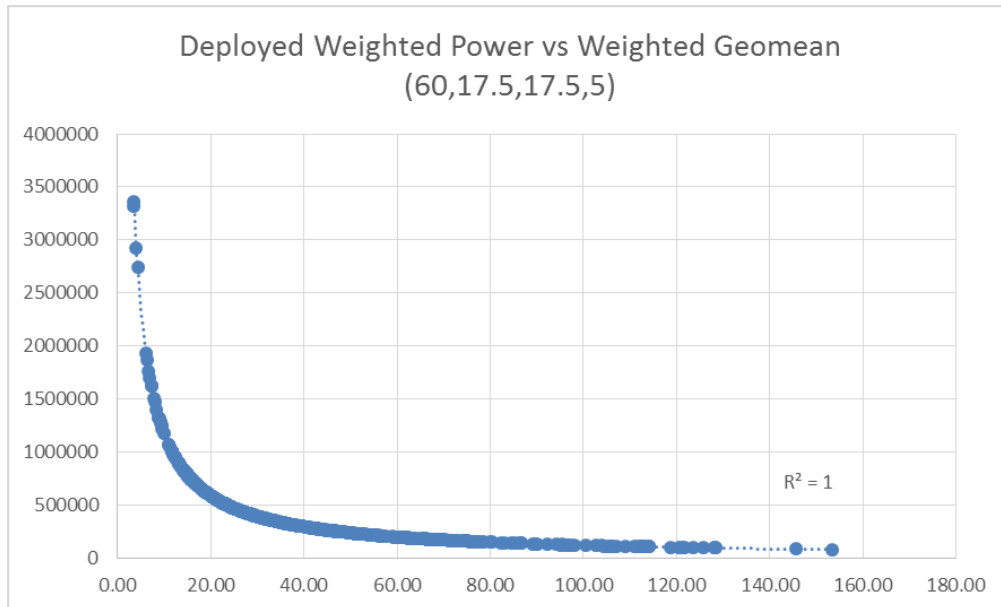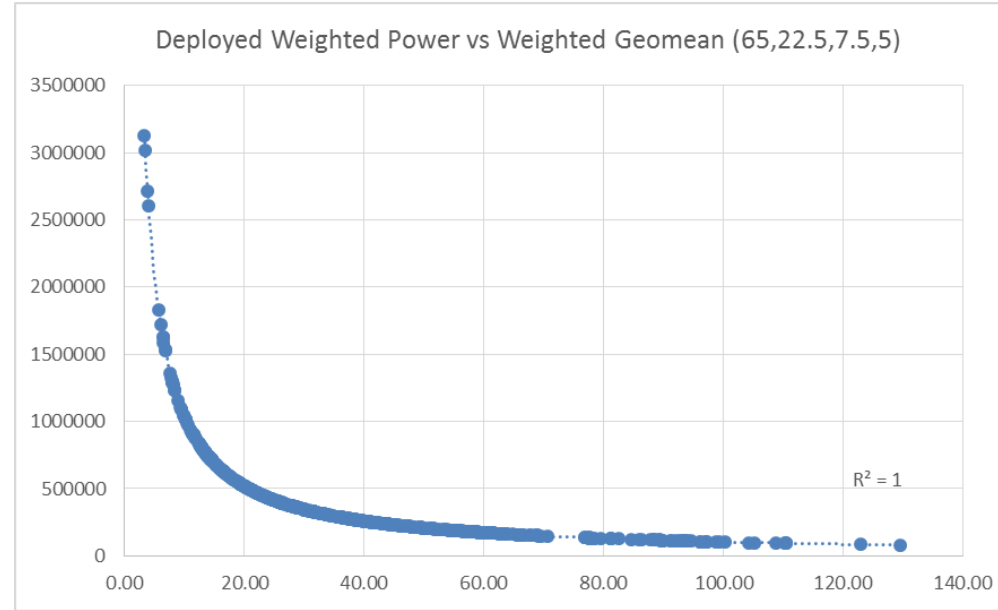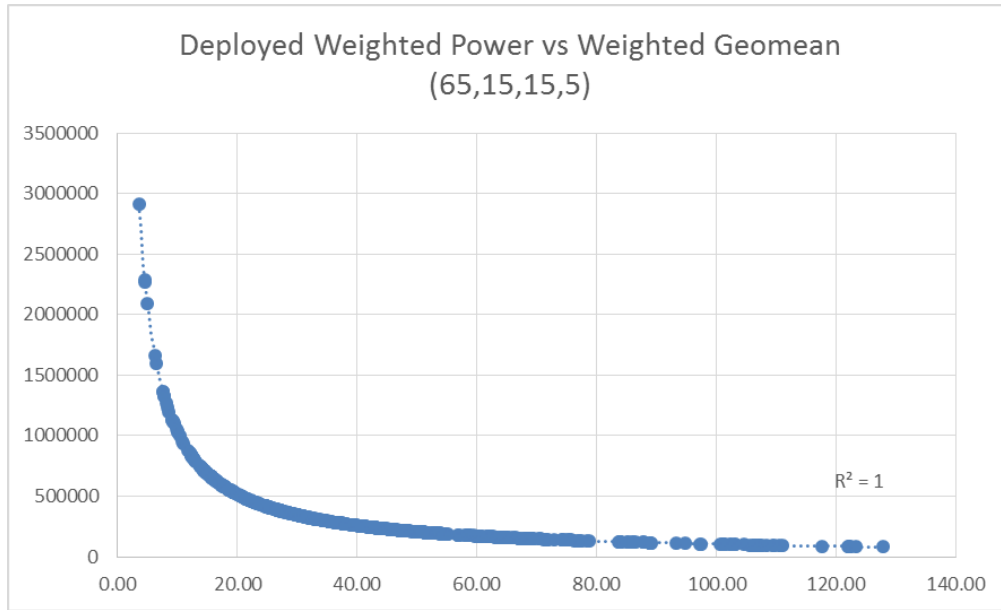
Deployed Power Graphics

- The following graphics, pages 11 to 17, plot the deployed power versus the efficiency scores for the 4 different Weighted Geomean Efficiency metrics evaluated in the tables on pages 8 and 9.
    - The graphics show the fit of the curve and the tightness of the data distribution for the two metrics.

- R2 values are provided to define the curve fit.
    - Based on the ranking analysis, a R-value of less than .99 is considered less than ideal because the number or ranking mismatches and the average size of the ranking mismatch increases significantly when the R-value drops below .99.

- 7 combinations of utilization, power and/or workloads were considered to evaluate how the server performed under various workload types and scenarios and to assess the effectiveness of the metric against a range of potential operating scenarios.

- The deployed power is plotted on the y axis of the graphs, in watts, and the efficiency score, representing a normalized performance over power metric, is plotted on the x-axis.

- The combined ITI/TGG/SPECpower data set used to generate the graphics in charts 11 to 17.

- The graphics show that all 4 aggregation weighting combinations, there is a strong curve fit to the plot of the Deployed Power vs. Efficiency Score values. All four combination methods provide a reasonable assessment of server efficiency.
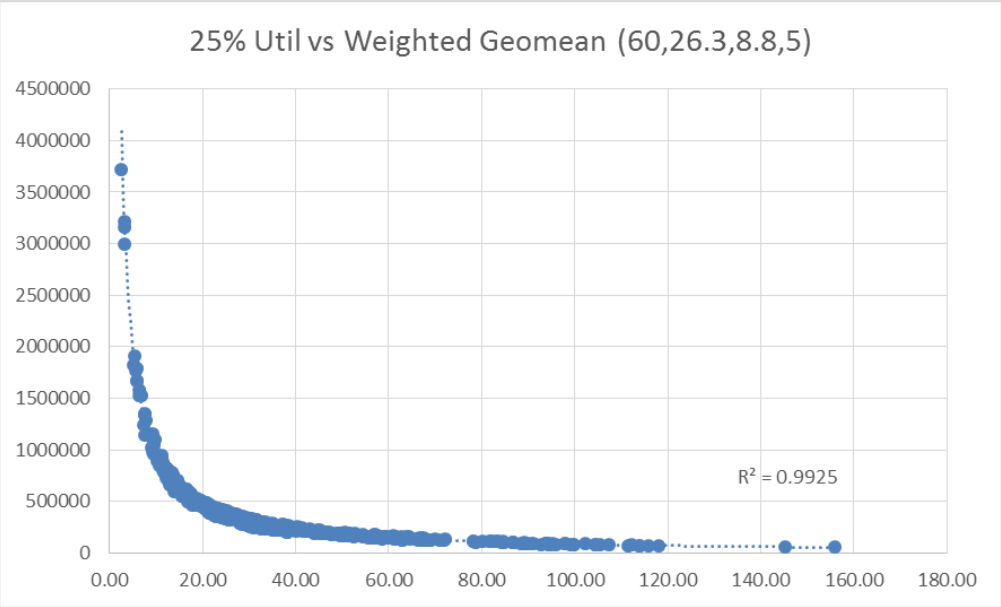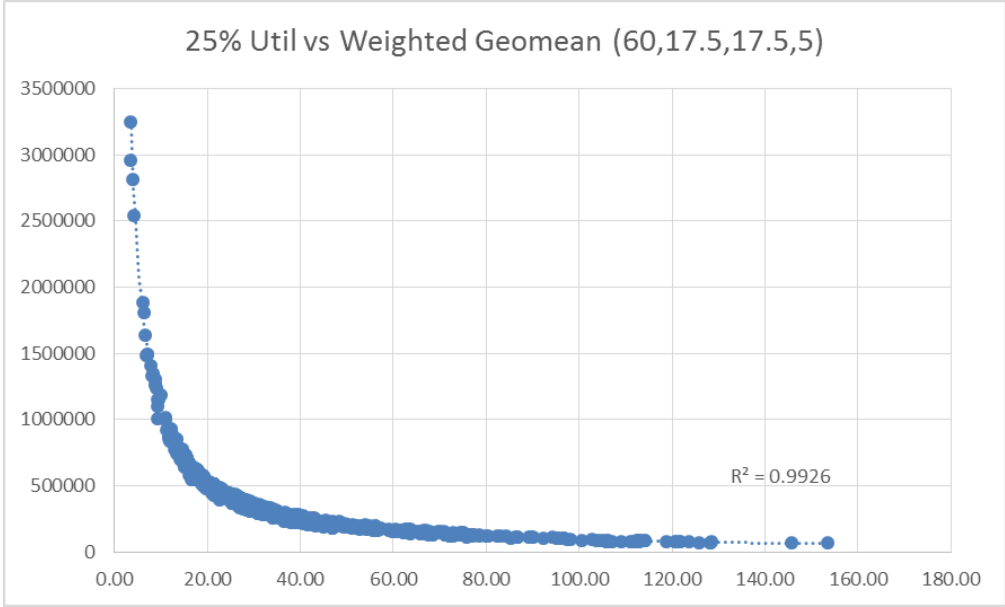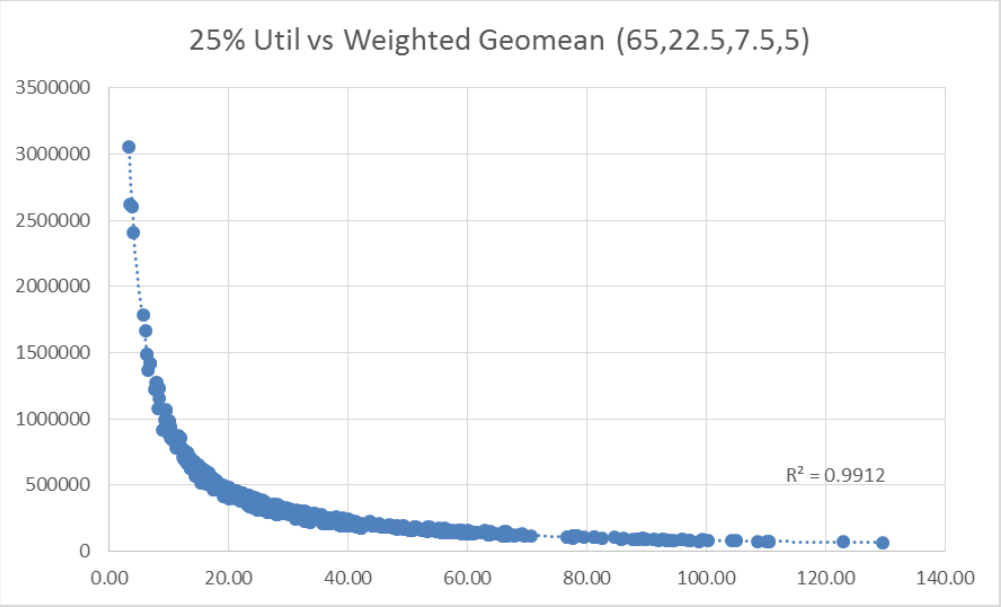
Deployed Idle versus Efficiency Scores
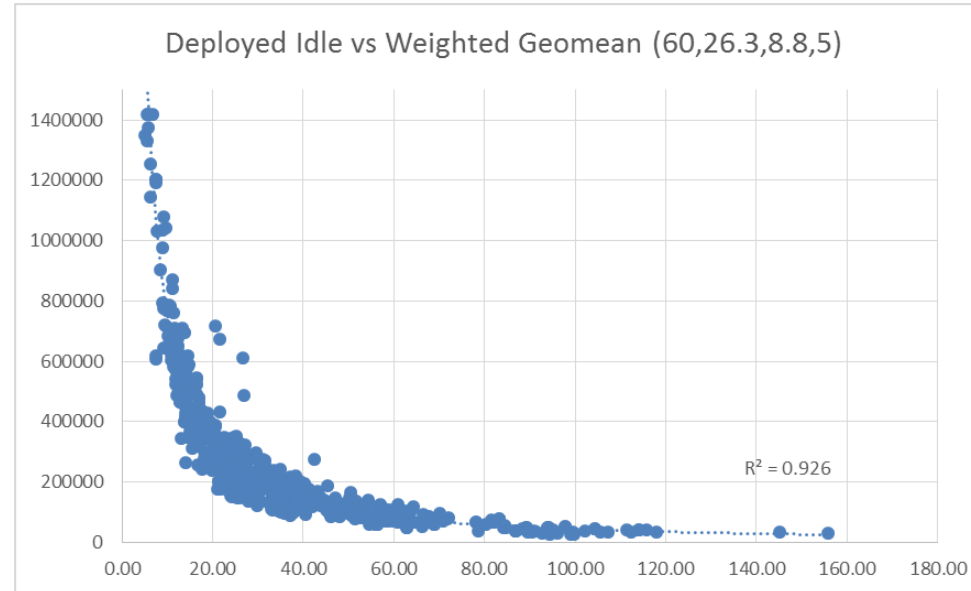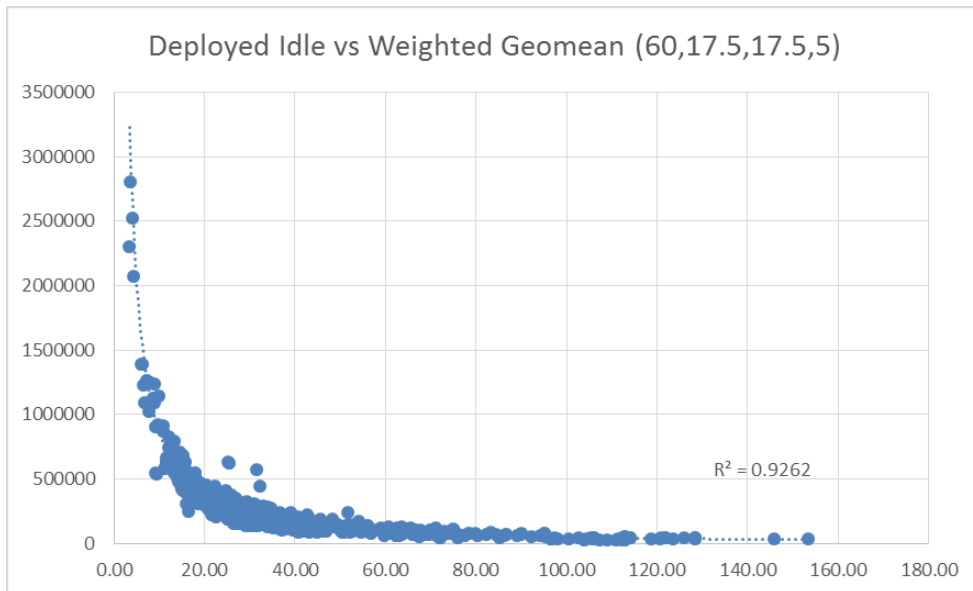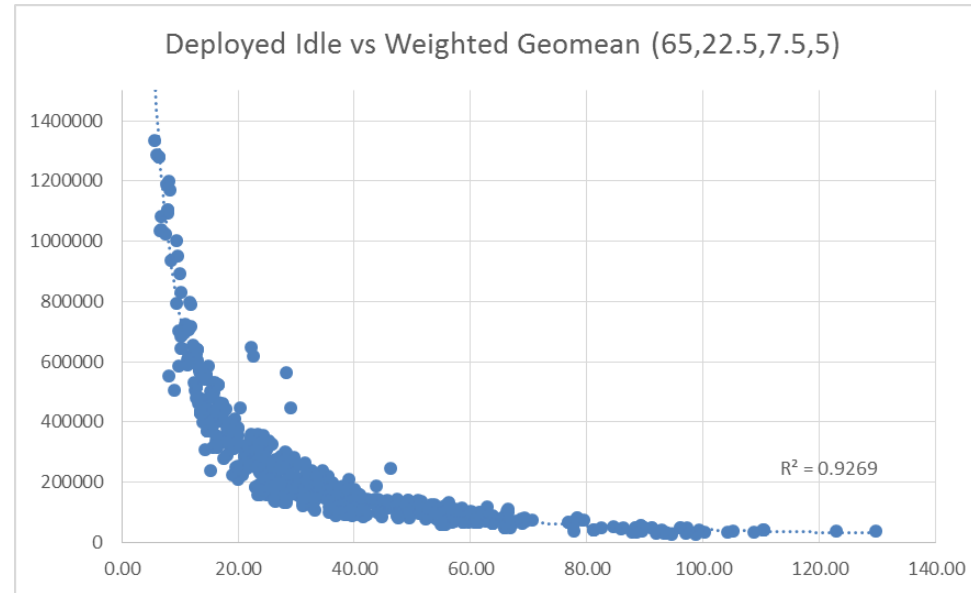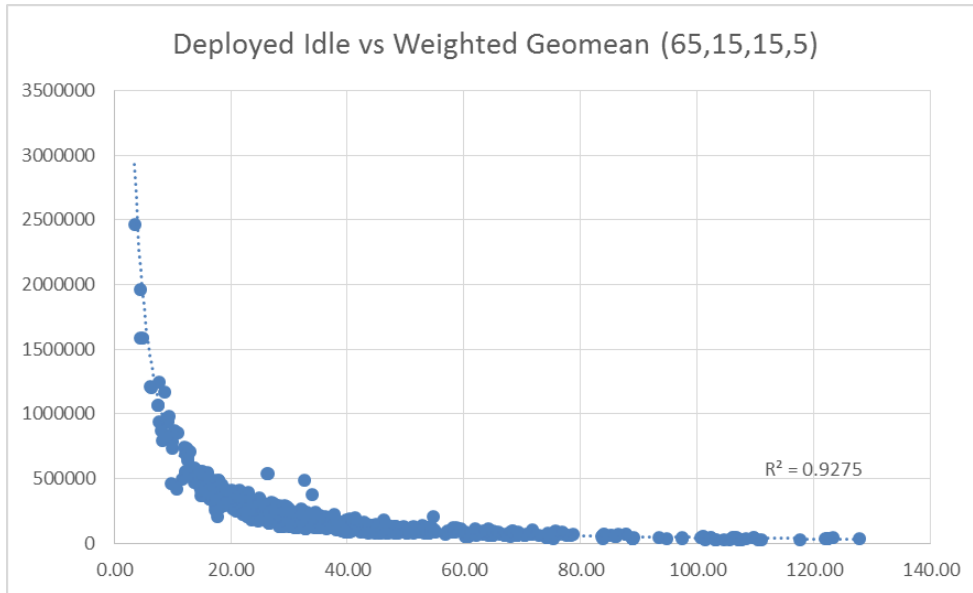
Deployed Weighted Workload Power vs. Weighted Geomean Efficiency

Power Use at 25% Utilization

25% Util vs Weighted Geomean (65,15,15,5)
R² = 0.9914

25% Util vs Weighted Geomean (65,22.5,7.5,5)
R² = 0.9912

25% Util vs Weighted Geomean (60,17.5,17.5,5)
R² = 0.9926

25% Util vs Weighted Geomean (60,26.3,8.8,5)
R² = 0.9925
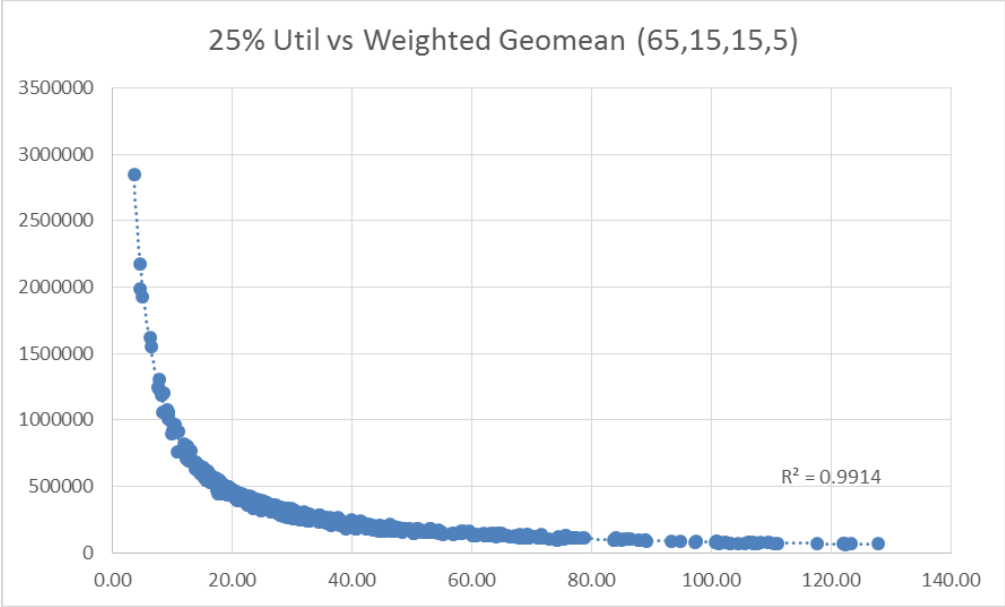
Power Use at 50% Utilization
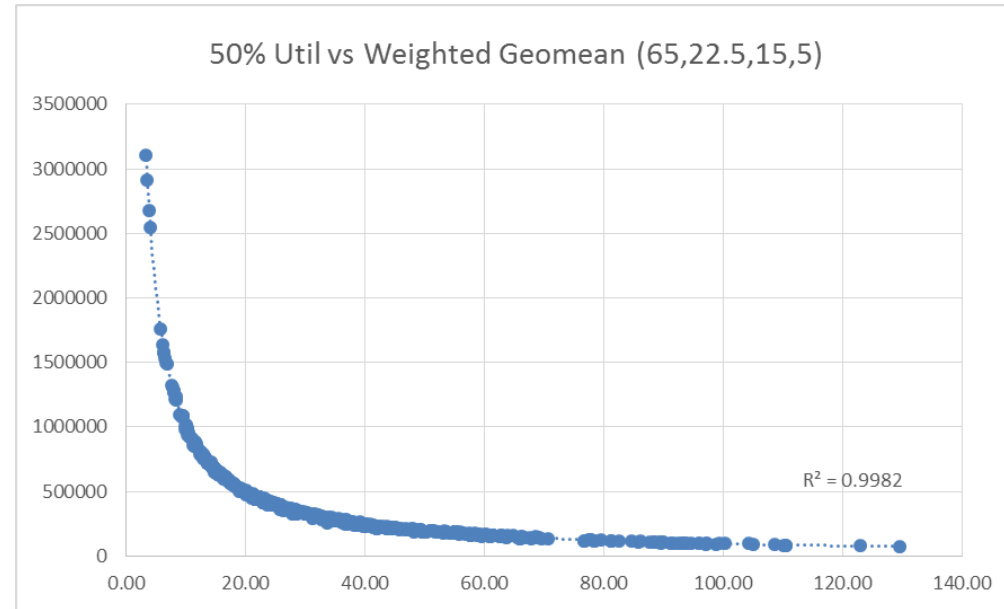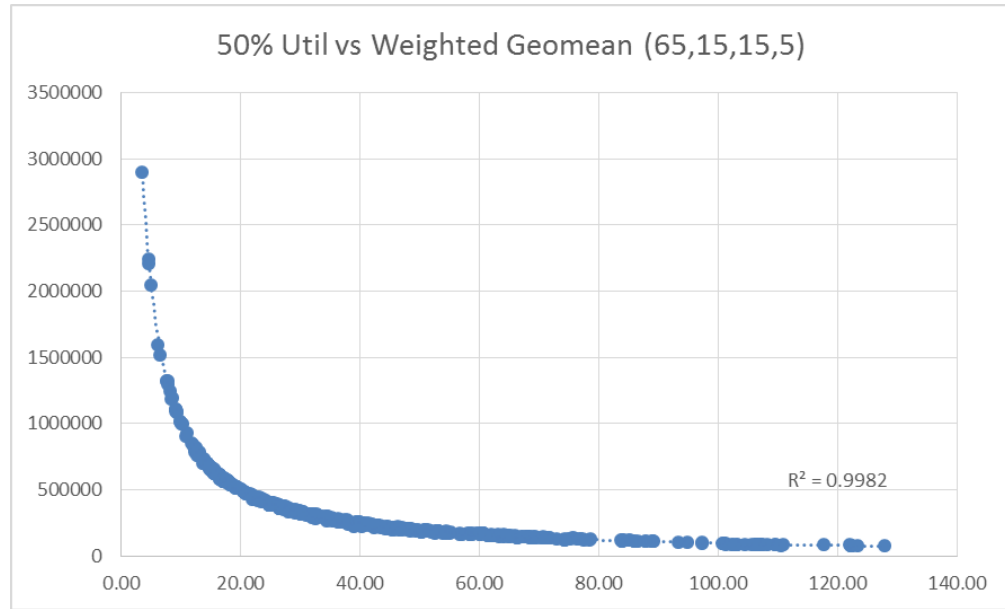
Power Use at 100% Utilization

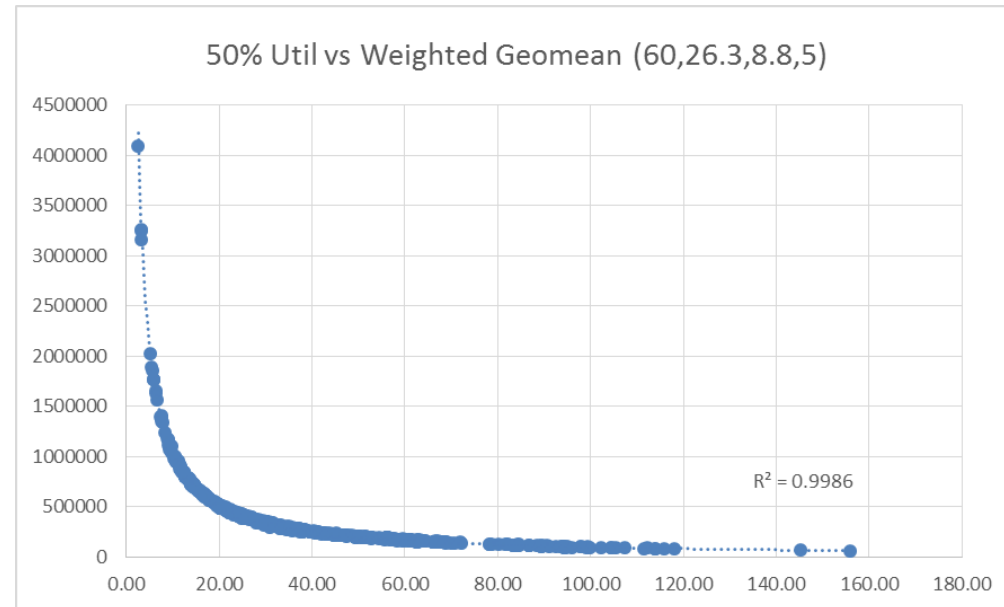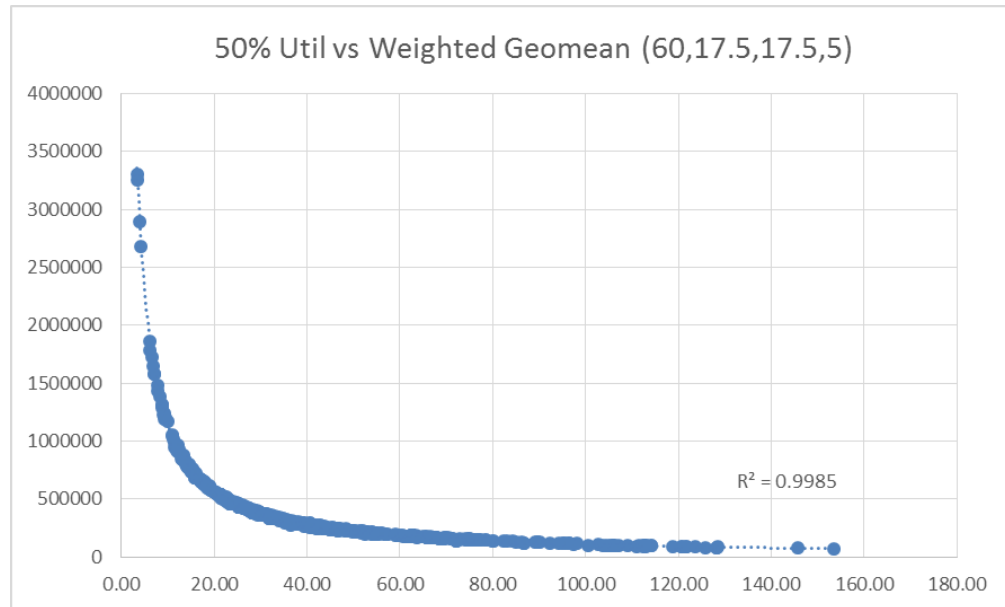# CPU Intensive Workload Power

CPU Intensive Power vs Weighted Geomean (65,15,15,5)

R² = 0.9995

CPU Intensive Power vs Weighted Geomean (65,22.5,7.5,5)

R² = 0.9996

CPU Intensive Power vs Weighted Geomean (60,17.5,17.5,5)

R² = 0.9992

CPU Intensive Power vs Weighted Geomean (60,26.3,8.8,5)

R² = 0.9994
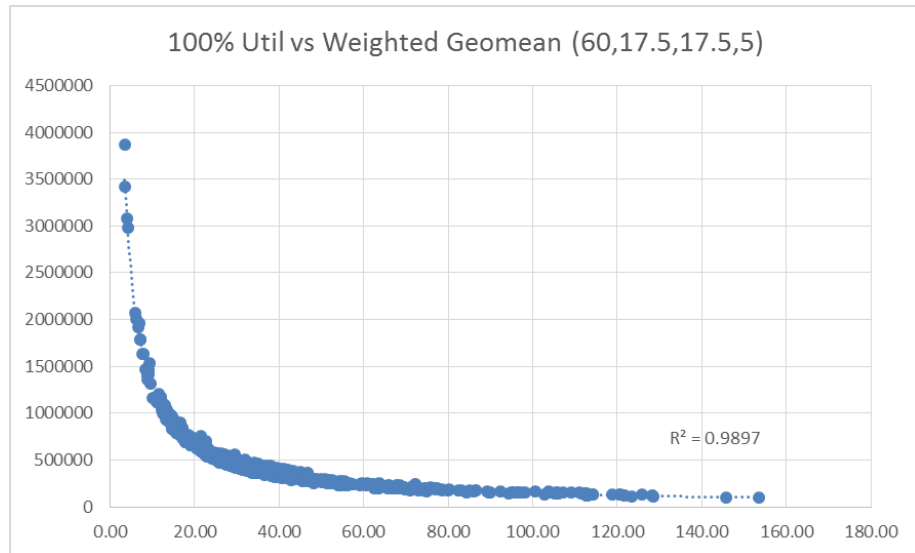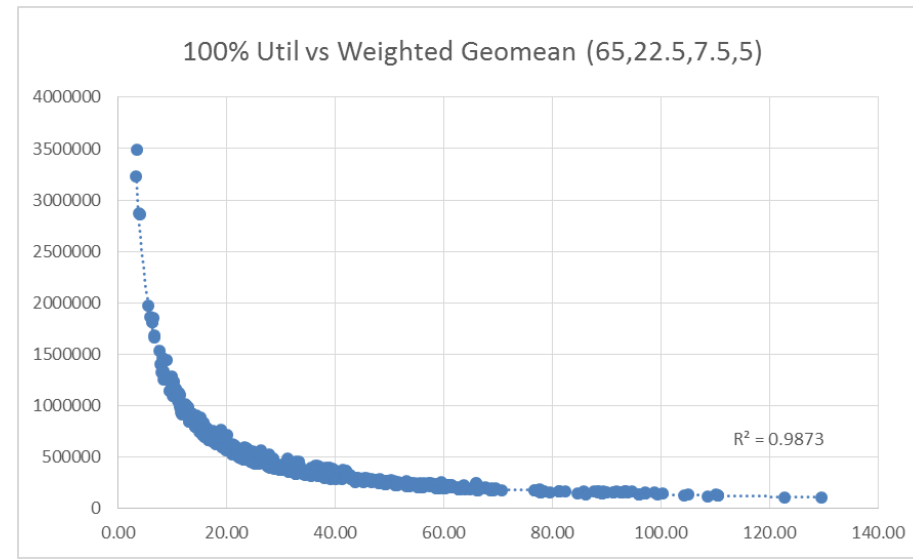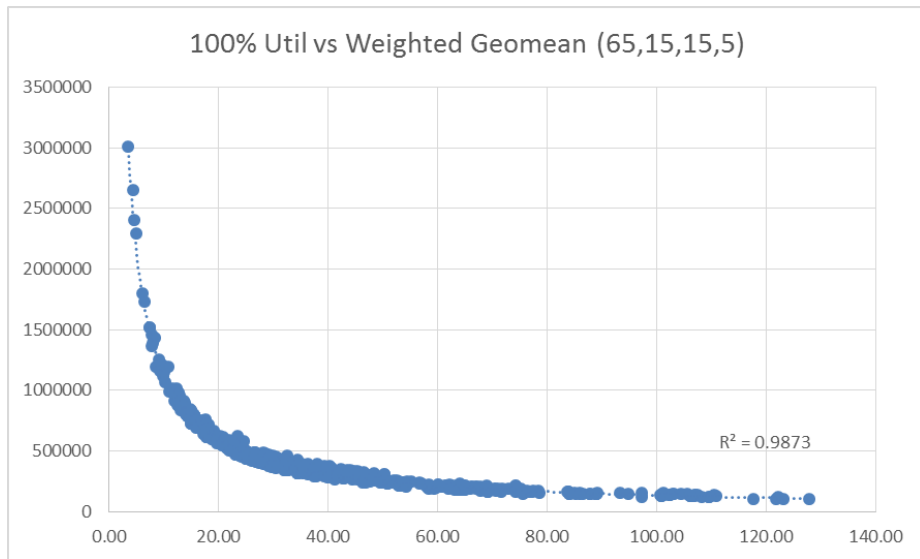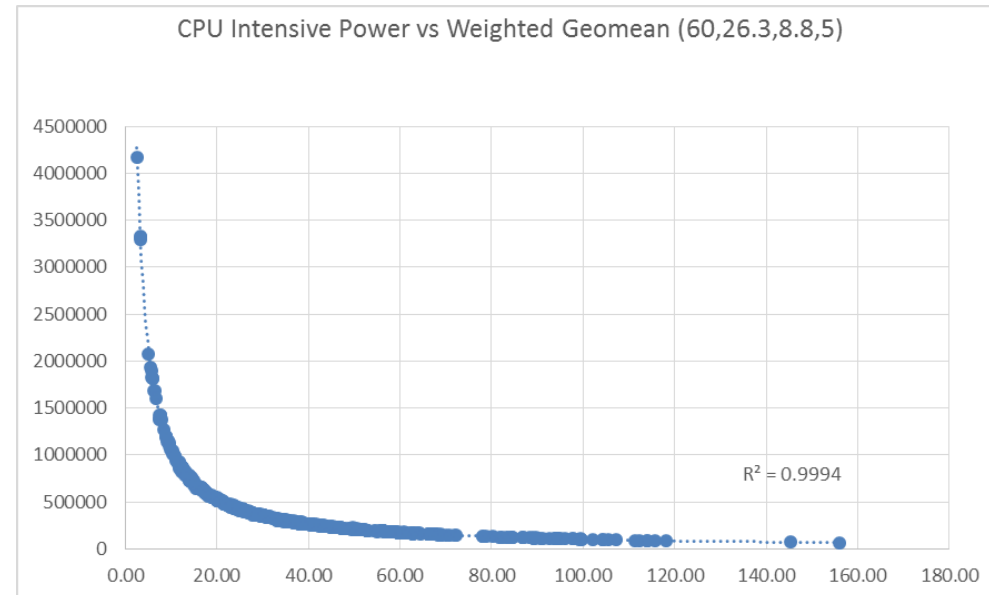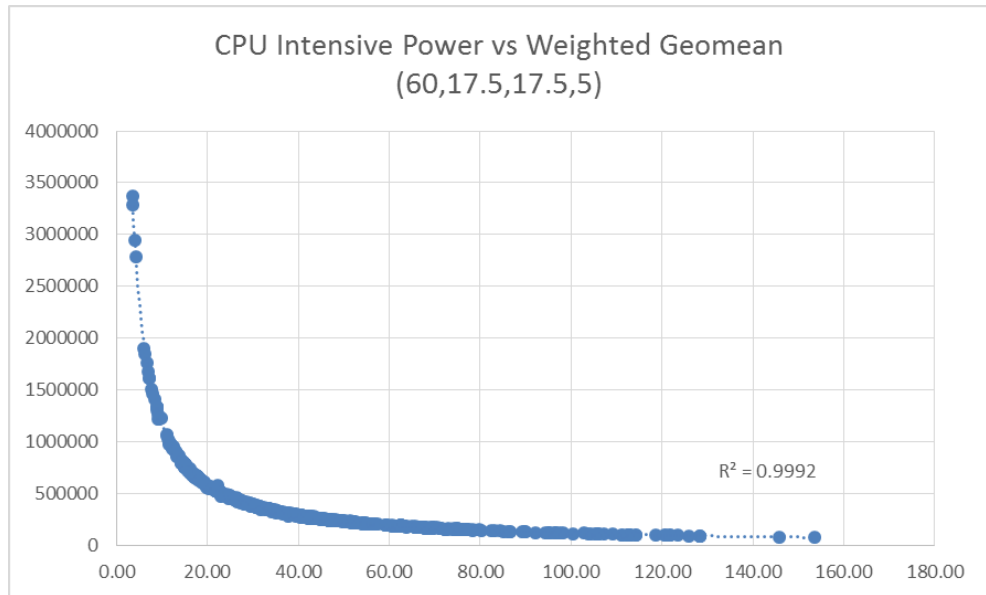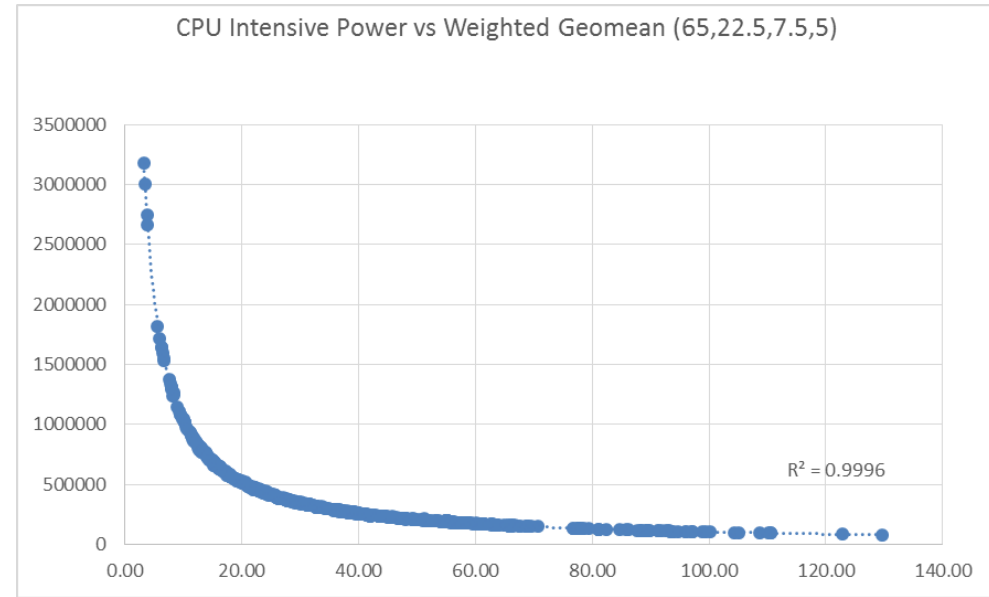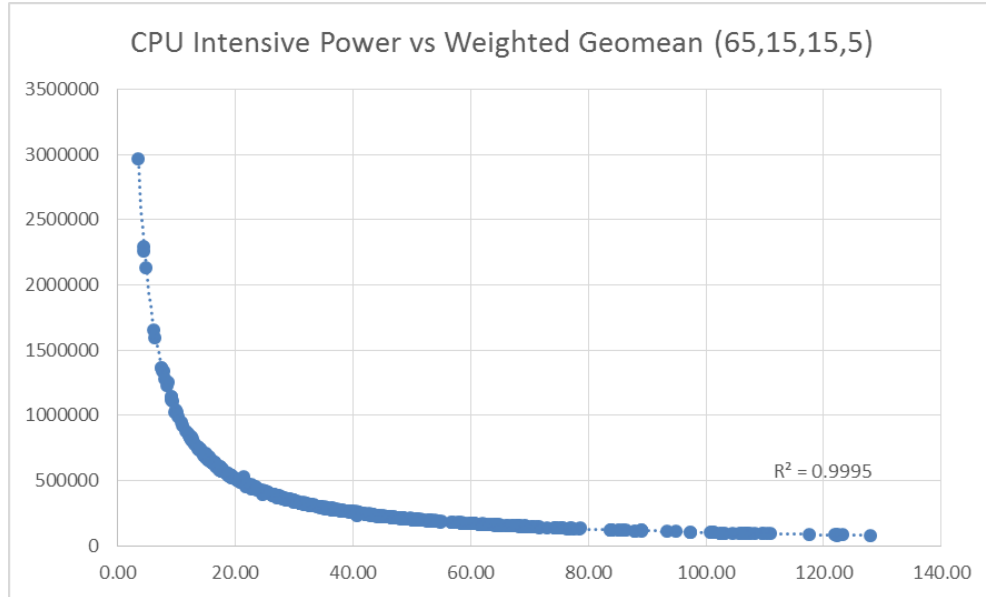
Memory Intensive Workload Power

Memory Intensive Power vs Weighted Geomean (65,15,15,5)

$R^2 = 0.9932$

Memory Intensive Power vs Weighted Geomean (65,22.5,7.5,5)
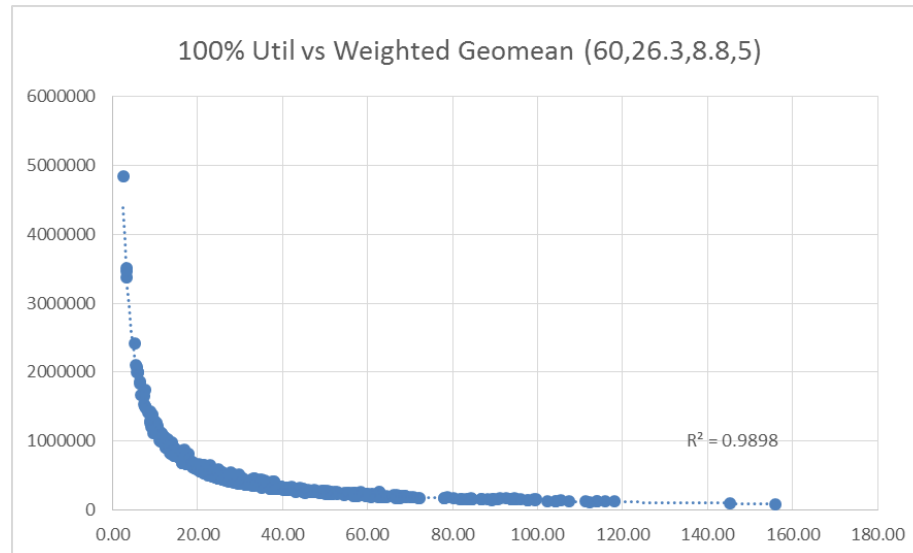
$R^2 = 0.9945$

Memory Intensive Power vs Weighted Geomean (60,17.5,17.5,5)

$R^2 = 0.9948$

Memory Intensive Power vs Weighted Geomean (60,26.3,8.8,5)

$R^2 = 0.9961$
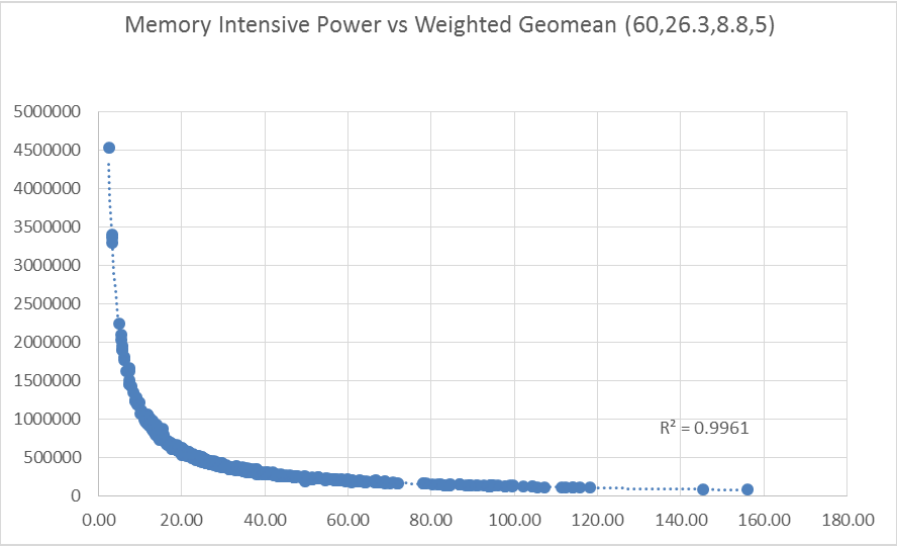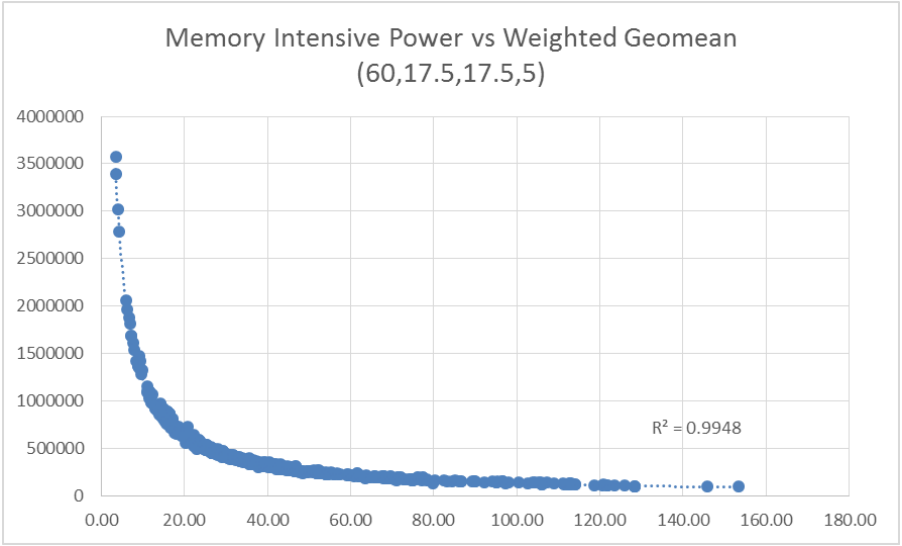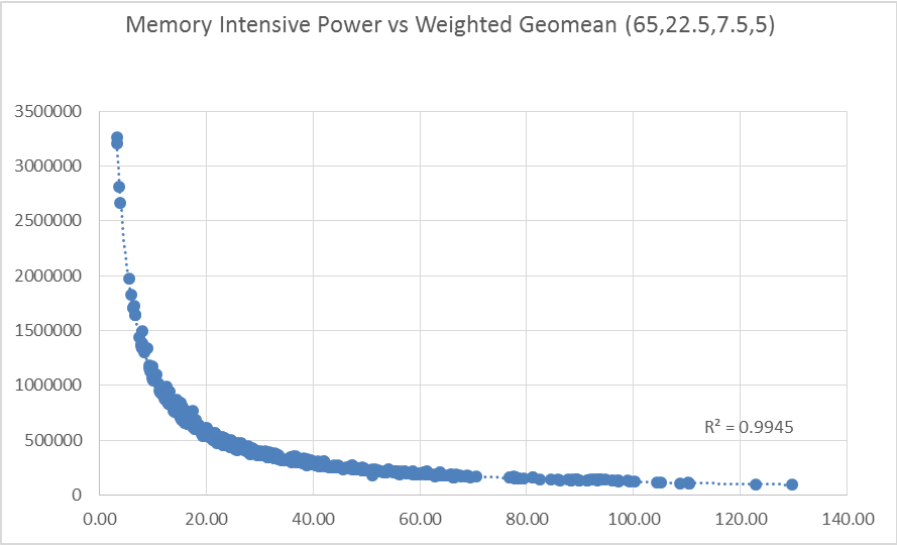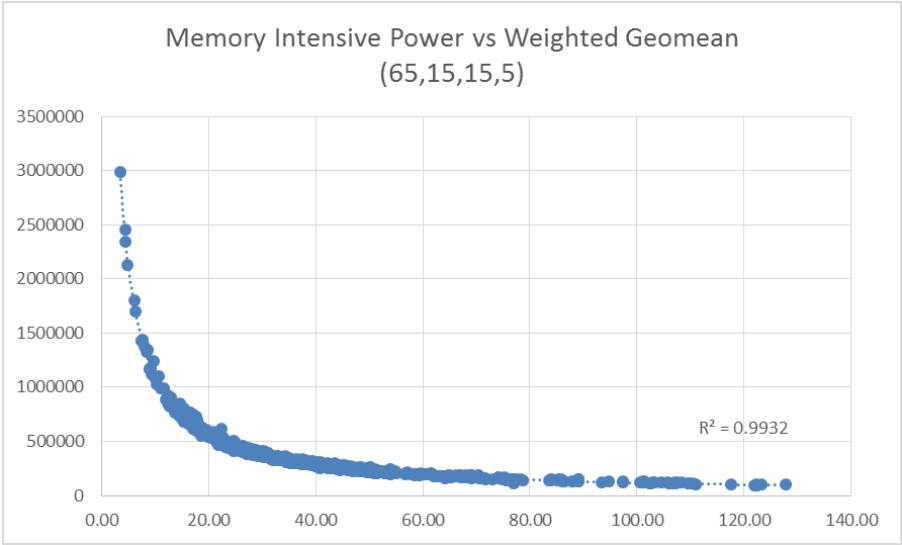
Comparison of the Geomean of the Efficiency Score for Each Configuration Type

	Efficiency score comparison by configuration				
	High End	Max Pwr	Typical	Low End	Min Pwr
Geomean	43.83	37.96	35.35	27.81	26.40
Min	6.65	10.58	9.53	9.39	3.04
Max	158.40	111.06	144.73	196.68	61.64
STDEV	28.91	19.55	21.24	30.21	14.39

- The table was generated using only the ITI Dataset with identified configurations.
  - High-end configurations are a combination of high-end performance and maximum power configs.
  - Low-end configurations are a combination of low-end performance and minimum power configs.

- The high-end performance and maximum power configurations have higher overall efficiency scores compared to the low-end and minimum power configurations.
  - The high-end and max power configs. will tend to use higher performance processors for performance and power purposes.
  - The maximum power configuration has a lower efficiency score because the components are added to maximize power without adding to performance. Even with this extra power burden, the higher power processors deliver a higher overall efficiency score than the low-end performance and minimum power configurations.
  - The minimum and low-end configuration geomeaned efficiency score are close in value, as the two configurations are relatively close together in component choices.
  - As expected, the typical configuration is between the high-end and low-end configurations.

- The geomeaned data demonstrate that it will be important to provide individual thresholds for a standardized configuration type for the low-end and high-end configurations dues to the different performance and power profiles for the two configuration types.

- The graphics in charts 19 to 21 depict the different efficiency score vs. deployed power distributions for three different power/workload combinations for the low-end and high-end configuration types.

COMPARISON OF DEPLOYED POWER vs. WEIGHTED GEOMEAN EFFICIENCY SCORES BY LOW-END and HIGH-END CONFIGURATIONS
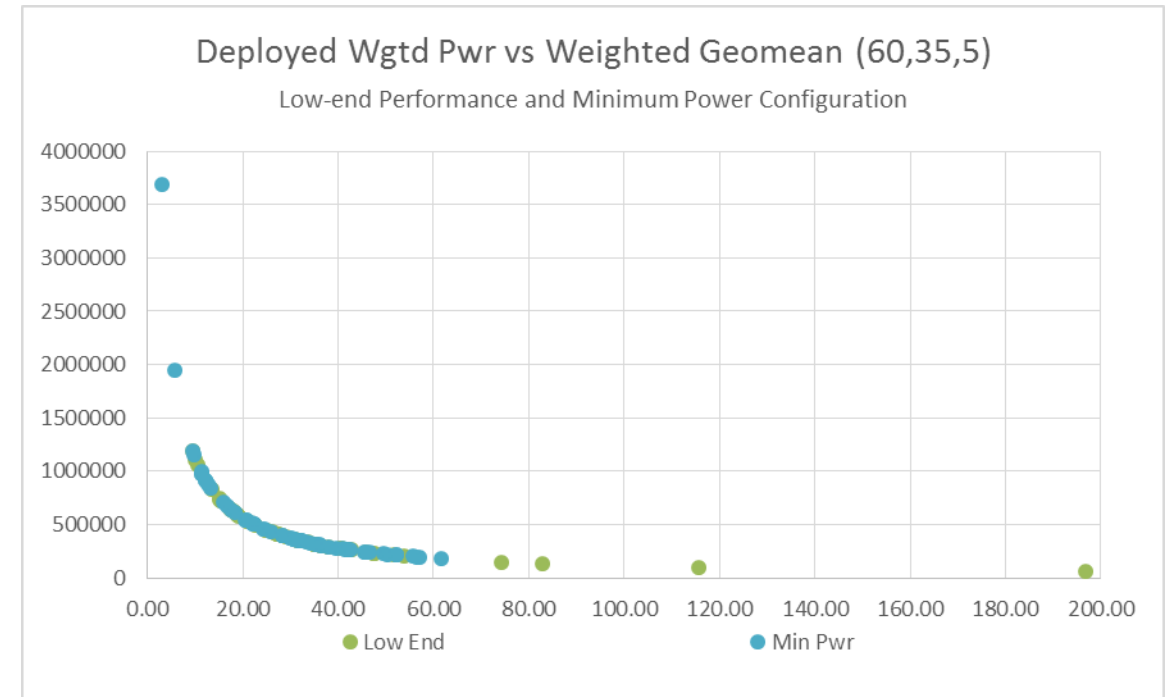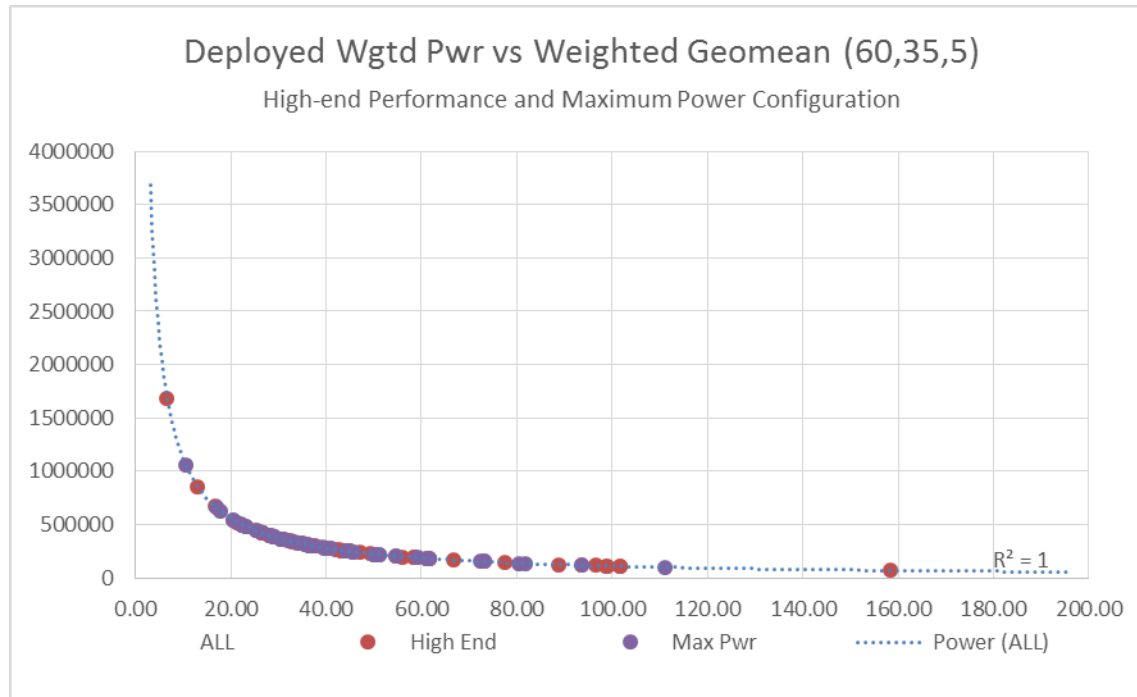

The graphics visualize the fact that the High-end configurations have a more efficiency scores above 60 and less efficiency scores below 20 than the Low-end configurations.

There is a similar distribution for other metrics. We chose to only show the distribution for the recommended, weighted geomean metric.

The efficiency scores were generated using only Flood for the memory score.

COMPARISON OF DEPLOYED POWER vs. WEIGHTED GEOMEAN EFFICIENCY SCORES BY LOW-END and HIGH-END CONFIGURATIONS
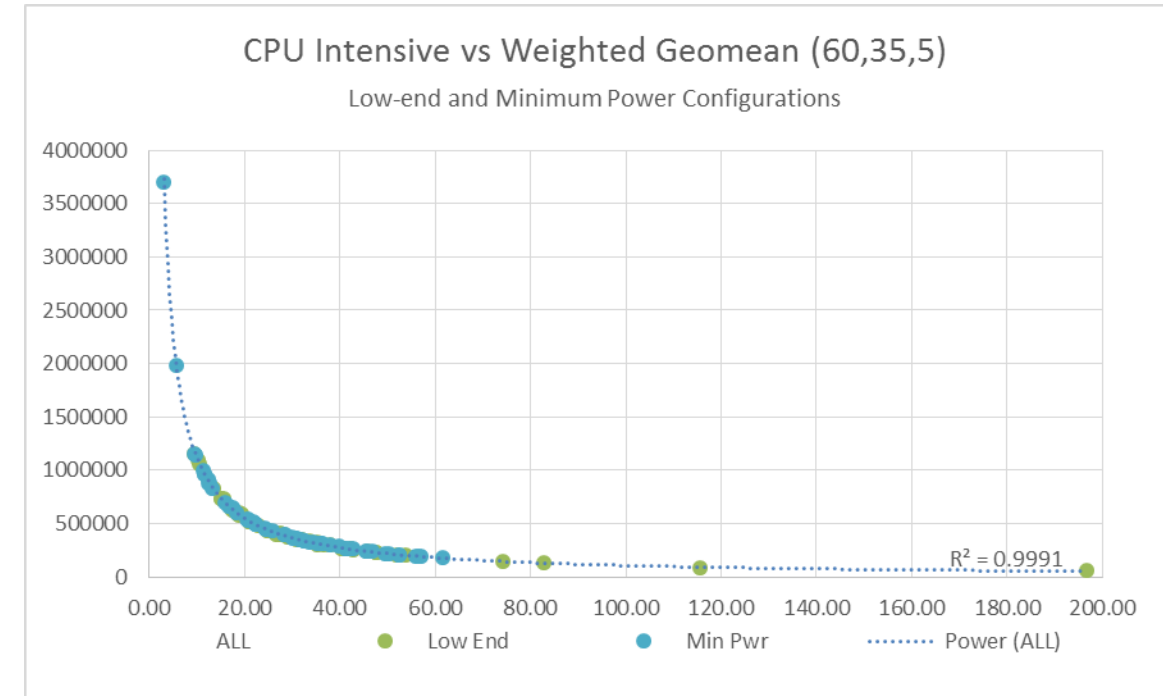

The graphics visualize the fact that the High-end configurations have a more efficiency scores above 60 and less efficiency scores below 20 than the Low-end configurations.

COMPARISON OF DEPLOYED POWER vs. WEIGHTED GEOMEAN EFFICIENCY SCORES BY LOW-END and HIGH-END CONFIGURATIONS

Deployed 25% Utilization vs Weighted Geomean (60,35,5)
High-end Performance and Maximum Power Configurations

Deployed 25% Utilization vs Weighted Geomean (60,35,5)
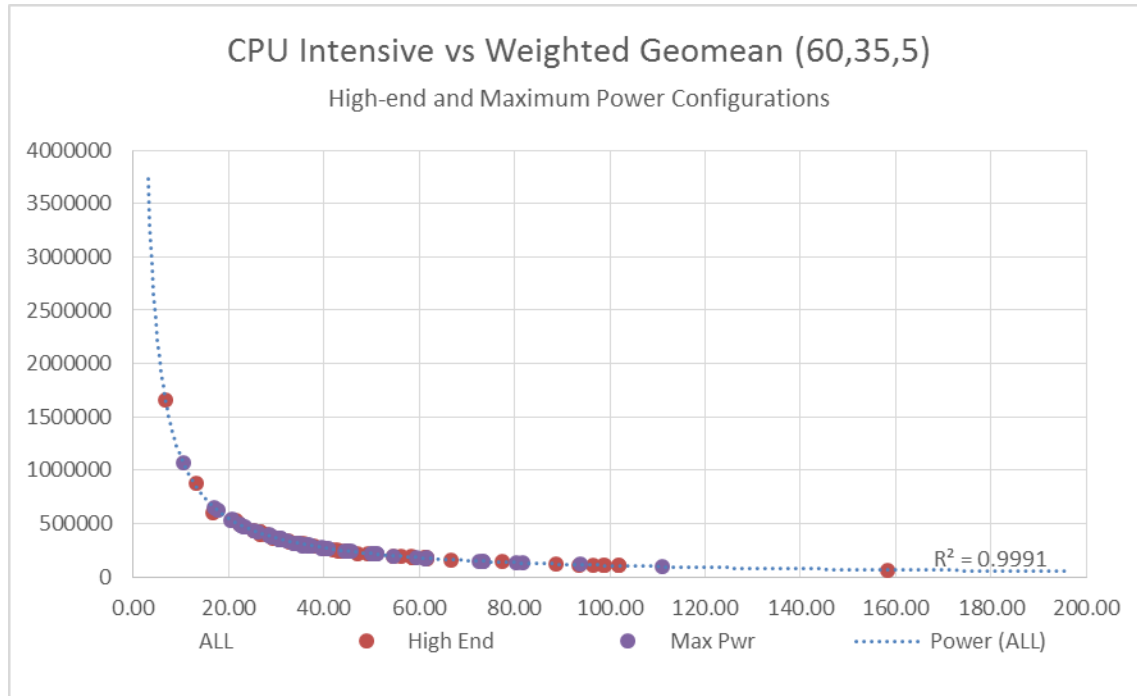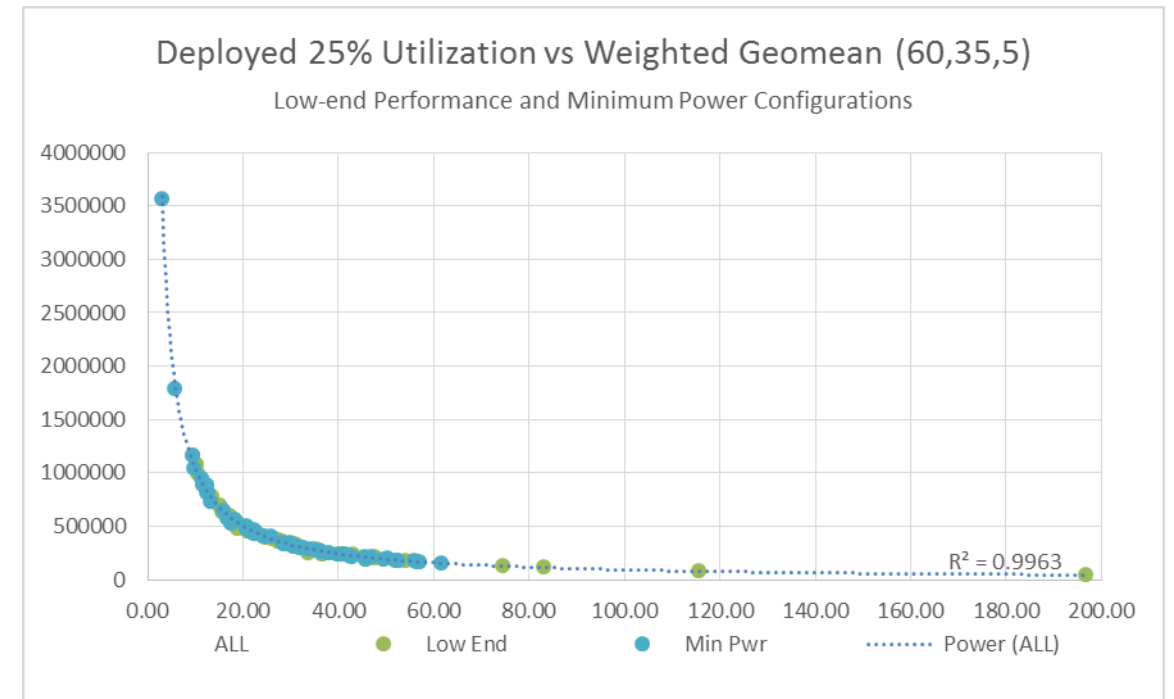Low-end Performance and Minimum Power Configurations
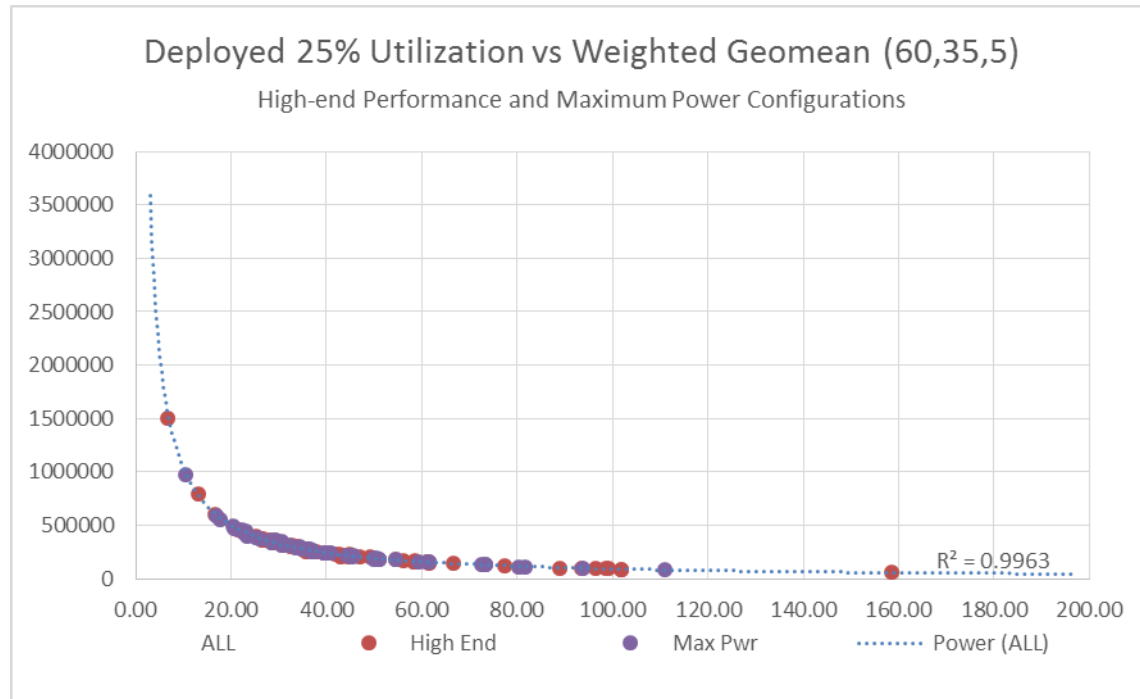
The graphics visualize the fact that the High-end configurations have a more efficiency scores above 60 and less efficiency scores below 20 than the Low-end configurations.

List of Supporting Papers:

SERT Metric Analysis: An analysis of different methods for combining the interval data and a recommendation to use the geomean as the combination method.

Mean Calculation Methodology Analysis: An analysis and comparison of combined metric results using the arithmetic mean, the geomean and the harmonic mean with a recommendation for the geomean.

Deployed Power Analysis Methodology Description: An explanation of how the deployed power analysis is used to validate the different efficiency metrics and a demonstration of its effectiveness as a validation method and a description of the calculation methods for the performance and power values used to calculate the deployed power use for each server and the 7 different utilization/power use/workload combinations used to assess deployed power.

SERT Metric Weighting Proposal: An analysis of the best weighting to use for the CPU and Memory worklet scores to create a single server efficiency metric.

Creating a Single Efficiency Metric from SERT worklet results: Details the equations and combinatory methods to move from the SERT normalized performance and power interval data to a single combined server efficiency metric.

ENERGY STAR V3 Draft 1 metric recommendation Executive Summary: Companion explanation to this file "Recommendation of Energy Efficiency Metric for ENERGY STAR V3 Draft 1".

Trademark References

SPEC and its product names SERT and SPECpower are registered trademarks of the Standard Performance Evaluation Corporation (SPEC), see spec.org.