The Green Grid's Feedback and Comments on BenchSEE v1.2.3 – Part 2

Date: Nov. 12, 2021

ITI's The Green Grid

1 Introduction

The Green Grid (TGG) reiterates its commitment to work with CNIS and appreciates the opportunity to provide Part 2 of our substantive comments on the public released version of China's server efficiency benchmark tool, BenchSEE v1.2.3. This is part of our Stage 1 work in a three stage assessment process outlined in our Part 1 submission (also addressed below). TGG's intention is to provide functional and design feedback on BenchSEE tool to enable CNIS to resolve key issues to make the tool production worthy. TGG deems it critical for SPEC SERT® suite to co-exist with BenchSEE for China's server energy efficiency standard. TGG's comments are planned for several stages, as described in the next section.

2 Executive Summary

This paper will continue BenchSEE architectural assessment focusing on power analyzers averaging interval (TGG27), and BenchSEE scoring assessment to address DRAM frequency scaling (part of TGG6), OS, core count, socket count and node count scaling. Rank analysis was conducted on a limited number of server systems run on BenchSEE and SERT tools. Servers with one installed processor shows ranking inversion (Figure 6) between BenchSEE and SERT scores. This could be attributed to the known 1P issues (TGG7). Similar paired analysis was conducted on 2P server systems which shows good correlation on ranking analysis except for two paired configurations. While this could be tied to TGG1 and TGG2 issues, this needs to be further understood (TGG28). Similar anomaly was observed on ranking analysis within a product family consisting of low end and high end configurations. BenchSEE efficiency score was higher on low end configuration than high end configuration, an inversion that needs further investigation (TGG29).

2.1 Future schedule update

Green Grid Analysis Stages

Stage 1 - BenchSEE v1.2.3 Analysis

- **Part 1 (Completed and submitted)**: A document which provided TGG's assessment on BenchSEE tool architecture, design, worklet scaling, and detailed tool functionality https://www.thegreengrid.org/en/resources/library-and-tools/527-Green-Grid-BenchSEE-v1.2.3-Feedback-and-Comments
- **Part 2 (This document)**: Completion of TGG's analysis of BenchSEE v1.2.3 and work that was not ready for this Part 1 document. For TGG to be able to complete the 1P server systems planned analysis, TGG requests a minor release or patches to fix the issues labeled TGG7 and TGG8 below (see Table 1- Part 1, in Annex). TGG will defer 1P server system analysis as part of new Part 3 workstream.

- **Part 3** – 1P analysis

Stage 2 - BenchSEE v1.3.0 Analysis

- TGG analysis based on the next major revision of BenchSEE (referred to in this paper as v1.3.0, TGG is unclear on what name CNIS will use for this version). TGG's understanding is that CNIS is working on the next major revision of BenchSEE.
- Due to lab capacity constraints, we are unable to volume test many versions of tool. Thus, TGG requests that items TGG1, TGG2, TGG3, TGG4, TGG5, and TGG6 (see Table 1 – Part 1, in Annex) are remedied before this analysis is performed, as any scoring analysis would be invalid if these changes are made after the analysis.

Stage 3 - BenchSEE Final Result Analysis (Section 6)

- Once accepted changes are complete and BenchSEE is producing near final performance and power results, TGG will conduct a substantive analysis based on a database of BenchSEE results to provide recommendations for energy efficiency grade thresholds for CNIS server energy efficiency standard,

2.2 Key Summary Table

TGG Key BenchSEE v1.2.3 Feedback Part 2 Summary (Table 1 – Part 2)

Observation #	Issue Description	Importance	*Requested Implementation Version	Type	Details
TGG27	**Issue: Power Analyzer** - BenchSEE does not lock the power interval on the power analyzers leading to inaccurate data **Recommendation**: BenchSEE set and lock the averaging interval to one second on the power analyzer.	High	Long term	Design	3.1
TGG28	**Issue: OS Scaling (Windows vs. Linux)** CPU-SHA256 worklet shows 12x performance difference between SERT vs. BenchSEE when the system under test was running a Windows OS vs. Linux.	High	V1.3.0	OS Scaling	5.2

	Recommendation: Understand why BenchSEE CPU-SHA256 worklet behavior is different from other BenchSEE CPU worklets. Could this be attributed to previous TGG1 and TGG2 (Table 1, Part 1 in Annex). Separately, understand and resolve the known BenchSEE *Cache* worklet design differences with SERT *Capacity* worklet, causing memory score differences (Linux vs. Windows)				
TGG29	**Issue: Rank Analysis across dataset** - Inversion on 2P server systems rank analysis conducted on dataset (BenchSEE vs. SERT) **Recommendation:** Understand why BenchSEE score on a high end configuration is lower that than typical and low end configuration	High	v1.3.0	BenchSEE scoring assessment (rank analysis)	5.6
TGG30	**Issue: Rank Analysis within product family -** Inversion on family configuration scaling (BenchSEE vs. SERT) **Recommendation:** Understand why BenchSEE score on a high end configuration is lower that than low end configuration within the same product family	High	v1.3.0	BenchSEE scoring assessment (family configuration scaling)	5.7

3 BenchSEE Architectural Assessment

3.1 Power analyzer averaging interval (TGG27)

Power analyzers have a configuration, the **averaging internal**, which controls how long the analyzer measures before it provides an average measurement over the configured period. If, for example, the averaging internal is ½ a second, then the analyzer measures data every ½ second, and then reports the average to the user every ½ second. As BenchSEE only collects data from the power analyzer every second, if the averaging interval is set to less than one second, some of the measurement period data is lost. For example, if the averaging interval is set to ¼ second, then each time BenchSEE reads the data, it obtains only the average of the power consumed in the last ¼ of the second and loses the data collected in the first ¾ of the second.

To fix this problem and to ensure measurement validity, TGG recommends that BenchSEE set and lock the averaging interval to one second on the power analyzer. Until automatic setting and locking of the power analyzer is implemented in BenchSEE, CNIS could add instructions to the User Guide asking the tester to set this setting on their power analyzer before testing.

4 BenchSEE Functionality Assessment

4.1 No further assessment in Part 2

5 BenchSEE Scoring Assessment

5.1 DRAM Frequency Scaling (TGG6)

Table 2 shows the performance increases when the server memory frequency is increased. "Scaling Efficiency" refers to the performance percentage increase divided by the memory frequency increase. So, for example, if the performance increases by 10% when the memory frequency increases by 20%, this would be 50% memory scaling efficiency (10%/20%). When comparing BenchSEE vs. the SERT suite, the results look within expectations, with one exception. On BenchSEE, for all of the non-Java CPU worklets (all except CPU-OLPT), there is less performance gain with increased memory bandwidth than on SERT, indicating the BenchSEE worklets are more cache resident. As there are many non-cache resident CPU workloads commonly used in servers, TGG recommends increasing the memory footprint of at least a few of the CPU worklets to add real-world relevance. This is additional data for the same observation (named TGG6), that Green Grid made in Part I of this paper (Table 1- Part 1 in Annex).

Table 2:

DDR4 2666 vs. 2400 MT/s Performance Scaling

BenchSEE Worklet	SERT Equivalent	BenchSEE 1.2.3			SERT 2.0.4			SERT Minus BenchSEE
		Speedup	%increase	Scaling Efficiency	Speedup	%increase	Scaling Efficiency	Scaling Efficiency
CPU-AES		1.0	0.0%	0.0%	1.0	1.4%	12.6%	12.6%
CPU-Compress		1.0	0.3%	2.7%	1.0	0.5%	4.5%	1.8%
CPU-LU		1.0	0.0%	0.0%	1.0	-0.2%	-1.8%	-1.8%
CPU-OLPT	CPU-SSJ	1.0	1.1%	9.9%	1.0	-0.5%	-4.5%	-14.4%
CPU-SHA256		1.0	0.0%	0.0%	1.0	0.0%	0.0%	0.0%
CPU-SOR		1.0	-0.3%	-2.7%	1.0	-0.3%	-2.7%	0.0%
CPU-Sort		1.0	0.1%	0.9%	1.0	-0.1%	-0.9%	-1.8%
Memory Cache	Memory-Capacity3	1.1	7.6%	68.6%	1.0	1.1%	9.9%	-58.6%
Memory Stream	Memory-Flood3	1.0	3.6%	32.5%	1.0	3.9%	35.2%	2.7%
Storage Random		1.0	4.6%	41.5%	1.0	1.7%	15.3%	-26.2%
Storage Sequential		1.0	0.3%	2.7%	1.0	4.0%	36.1%	33.4%

Baseline MT/s	Target MT/s	Speedup	%increase	Scaling Efficiency
2400	2666	1.111	11.1%	100.0%

CPU Boost Limited to 2.5 GHz

5.2 OS scaling (TGG28)

When BenchSEE and SERT are used to compare system performance of two-socket servers (2S servers) operating under different operating systems (OS) – Windows vs Linux – we see that overall, SERT results benefit when operating under Windows OS. At the same time, BenchSEE results benefited when operating under the Linux OS. The bar chart (Figure 1) below shows significant increased system performance (SERT results) for CPU, Memory and Overall scores. System performance for Storage score and Idle Power are higher in these conditions but not significantly. The biggest performance difference was seen with the Secure Hash Algorithm 256 (SHA256) CPU worklet (12x delta between SERT vs BenchSEE was seen when the system under test was running a Windows OS). In the case of memory score, BenchSEE *Cache* worklet showed performance on Windows to be ~10X lower than Linux, while no significant performance differences between Windows vs. Linus were observed on BenchSEE *Stream* worklet. Since the BenchSEE *Cache* worklet is designed differently from SERT *Capacity* worklet, the OS performance disparity with Cache worklet could be result of the design differences, a known issue and hence less concerning. However, CPU-SHA256 performance between SERT and BenchSEE need to be addressed when running Windows OS (TGG28).

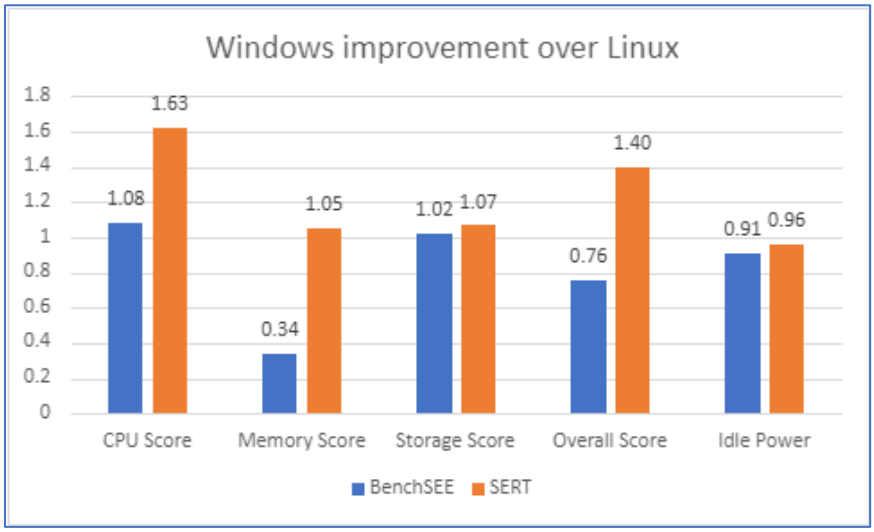

Figure 1

5.3 Core count scaling

The following chart (Figure 2) compares BenchSEE and SERT results when the number of cores in server is increased. The Blue and Grey bars compare changing from 20 to 28 cores, and the orange and yellow compare changing from 12 to 28 cores. Other configuration details and the frequency of the CPUs was held close to constant.

The main takeaway is BenchSEE and SERT overall scores are similar (less than 5%), with the exception of the memory worklets, which is expected as SERT Capacity and BenchSEE Cache are not designed to be comparable.

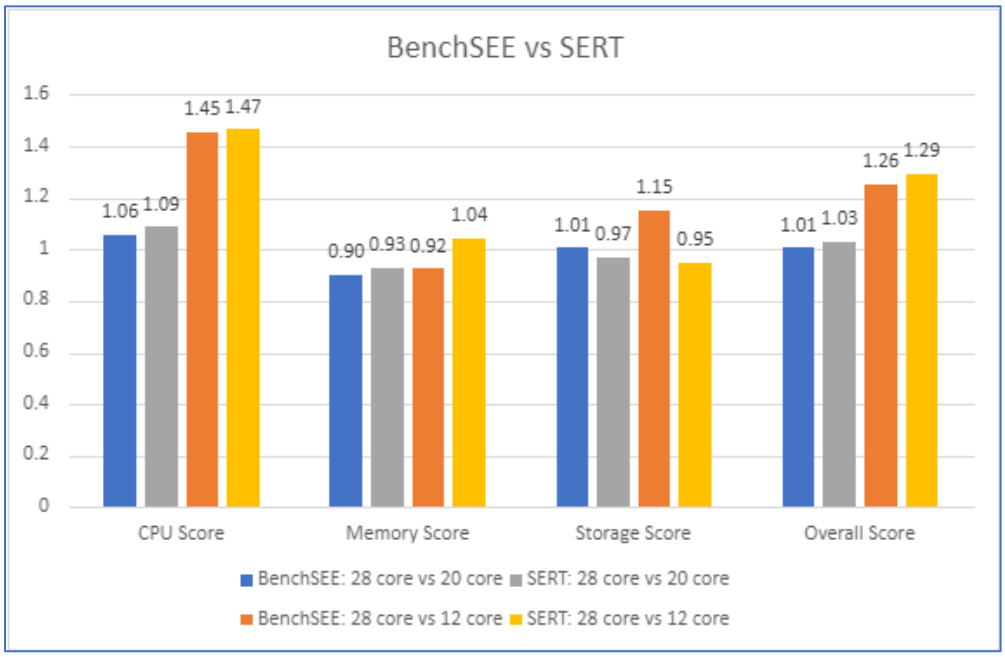

Figure 2:

5.4 Socket (count) scaling

A 2 socket and 4 socket server in a mid-range configuration was used to analyze the socket scaling of BenchSEE vs. SERT with 1, 2, and 4 installed CPUs. When the number of CPUs was doubled, the memory DIMM count, and total size was also doubled.

The results look reasonably similar between BenchSEE and SERT. As TGG pointed out in the previous paper with observation **TGG4** (Table 1 – Part 1 in Annex), BenchSEE lacks a worklet which gets an increased performance score when more memory capacity is added, which is problematic. This is again demonstrated in this data (Figure 3, Figure 4), where SERT Capacity 3 worklet efficiency score increases by ~45% when the memory size is doubled, whereas BenchSEE Cache performance does not increase.

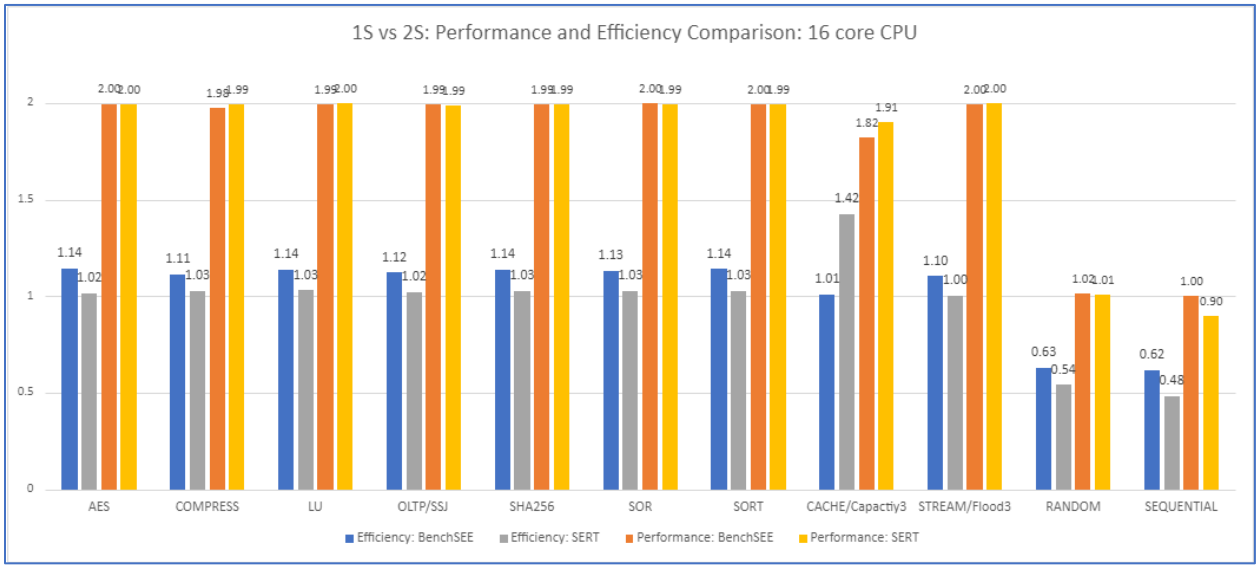

Figure 3

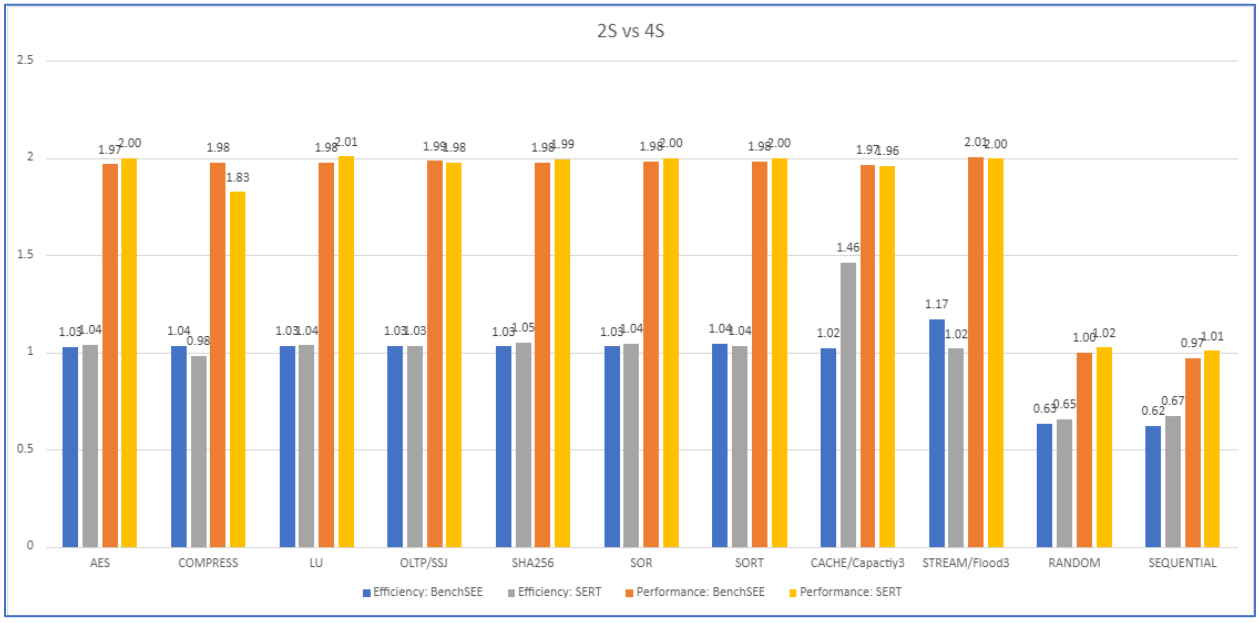

Figure 4

5.5 Node (count) scaling

No scaling problems were detected when testing two-processor, two-socket server systems with 1, 2 and 3 nodes using SERT and BenchSEE. As it can be seen on the chart below (Figure 5), all CPU, Memory and Storage worklets show no material differences in efficiency or performance tests results between SERT vs. BenchSEE.

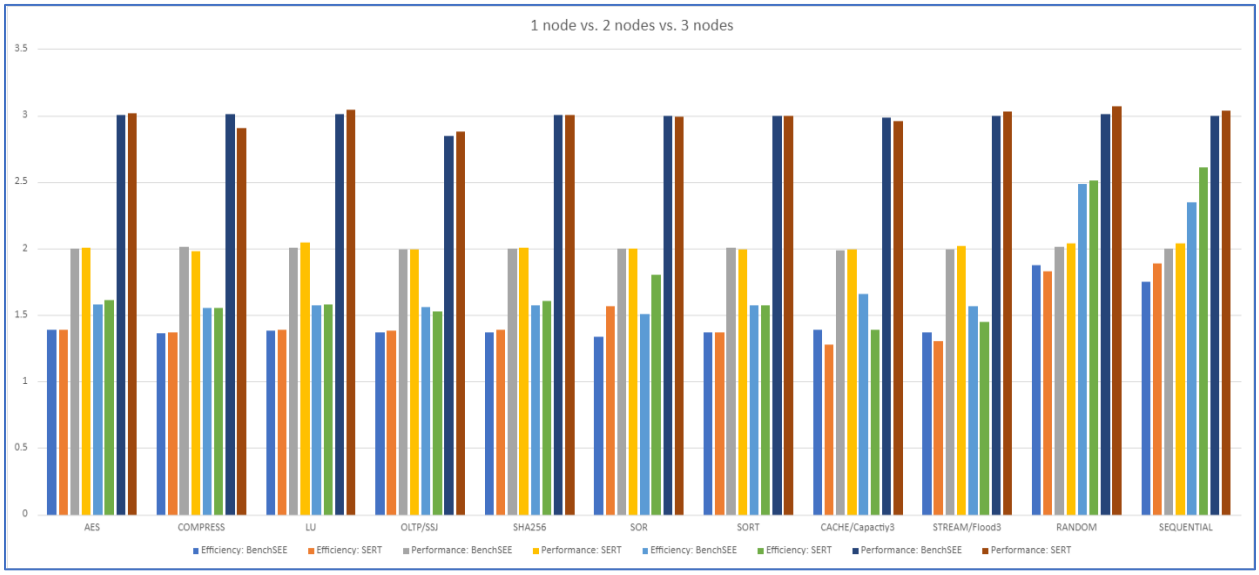

Figure 5

5.6 Rank analysis (TGG29)

Currently our data base has 4 configurations of 2 servers with 1 installed processor rack servers comprising of both SERT and BenchSEE data. The data base has 6 configurations of 2 servers with 2 installed processors rack servers that have both SERT and BenchSEE data. While these quantities are not sufficient for any statistical analysis we did perform a ranking analysis to see if there were any obvious issues.

1 Processor installed systems show a negative correlation between BenchSEE and SERT efficiency scores (Figure 6). This means that least efficient on BenchSEE will be the most efficient on SERT. TGG believes this may be attributable to the known 1 processor issues with BenchSEE. No further analysis will be performed with 1 processor installed systems until the known issue has been resolved (Refer to TGG7, Table 1 - Part 1 in Annex).

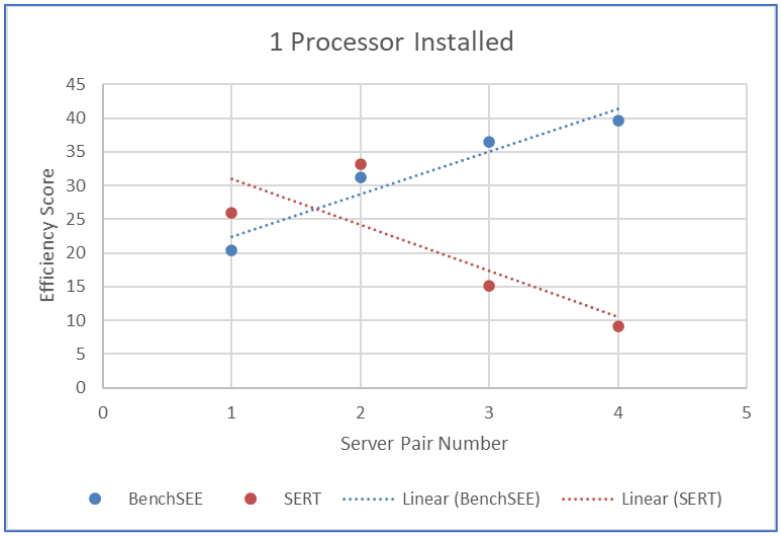

Figure 6

Table 3 below shows the rank analysis for the configurations with both SERT and BenchSEE data for servers with a number of installed processors and also for configurations within a single family.

Table 3

BenchSEE Family ID	Threads	Memoey Size	Configuration Type	BenchSEE Reference #	BenchSEE Efficiency	# of Installed Processors	Server Type	BenchSEE Rank	SERT Reference #	SERT Efficiency	SERT Rank	Family Ranking BenchSEE	Family Ranking SERT
AIQF	40	96	Other	TGG_22	20.36	1	Rack	4	TGG23	25.91	2		
AIQG	16	64	High-End	TGG_32	31.3	1	Rack	3	TGG33	33.12	1	3	1
AIQG	8	16	Typical	TGG_30	36.45	1	Rack	2	TGG31	15.11	3	2	2
AIQG	2	16	Low-End	TGG_28	39.71	1	Rack	1	TGG29	9.15	4	1	3
AIQF	12	96	Low-End	TGG_18	66.38	2	Rack	6	TGG19	8.10	6	4	4
AIQF	12	96	Minimum Power	TGG_20	113.24	2	Rack	5	TGG21	12.94	5	3	3
AIQF	112	384	High-End	TGG_26	202.2	2	Rack	2	TGG27	33.12	2	2	1
AIQF	80	192	Typical	TGG_24	221.14	2	Rack	1	TGG25	27.95	4	1	2
AIQH	64	256	High-End	TGG_36	128.51	2	Rack	4	TGG37	68.89	1	2	1
AIQH	16	128	Low-End	TGG_34	202.13	2	Rack	3	TGG35	32.98	3	1	2

Rack servers with 2 installed processors consisted of 6 configurations of 2 server families. The ranking correlation is good except for the configurations highlighted in yellow in the above table 3. The below chart (Figure 7) shows the CPU, Memory, Storage and Server Efficiency scores for the two systems with the issue. The main contributor is the fact that BenchSEE and SERT favor different systems for both the CPU and memory workloads. SERT test results in higher efficiency score for processor with the highest core count and the system with the largest memory size while BenchSEE favors the opposite system. The memory difference is likely because SERT Capacity forces scaling of the memory score with larger memory sizes while this feature is not present in BenchSEE. The reason for CPU difference is not as clear and needs to be understood (TGG29) and could be due to lack of optimization capabilities in BenchSEE. The two processors are from different manufacturers.

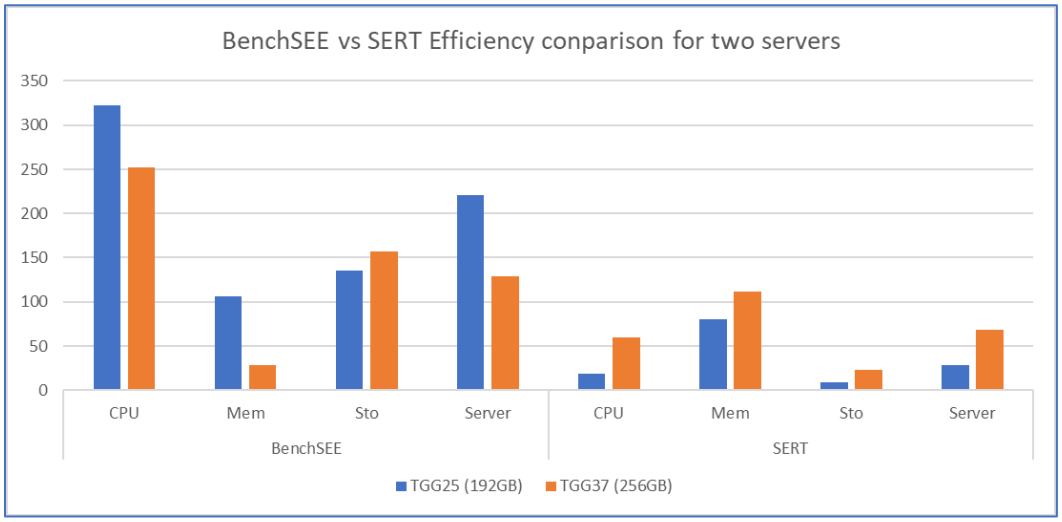

Figure 7:

5.7 Family configuration scaling (TGG30)

A second look at scaling of efficiency with configuration was done on the two 2 processor families, AIQF and AIQH, with Both SERT and BenchSEE data (Table 3). Ranking was performed on the configurations within each family to compare SERT and BenchSEE efficiency score scaling with configuration differences of a single server. The two figures (Figure 8, Figure 9) below show the workload and server efficiency score for the configurations that exhibited significant rank changes. The AIQH family consisted of just a Low-End and a High-End configuration. BenchSEE gives a higher efficiency value for the Low-End configuration in this family which is counter to what should be expected. The High-End configuration has higher performance processors and larger memory sizes and should be expected to achieve higher efficiency scores at the server level. In the figure below it can be seen that BenchSEE is giving higher CPU efficiency scores for processors with lower thread counts and higher Memory efficiency scores for systems with smaller memory sizes.

These two issues will cause significant ranking error between BenchSEE and SERT and prevent similar correlation between BenchSEE and SERT efficiency scores. Further investigation is needed (TGG30) to determine the exact causes and potential solutions for these two issues.

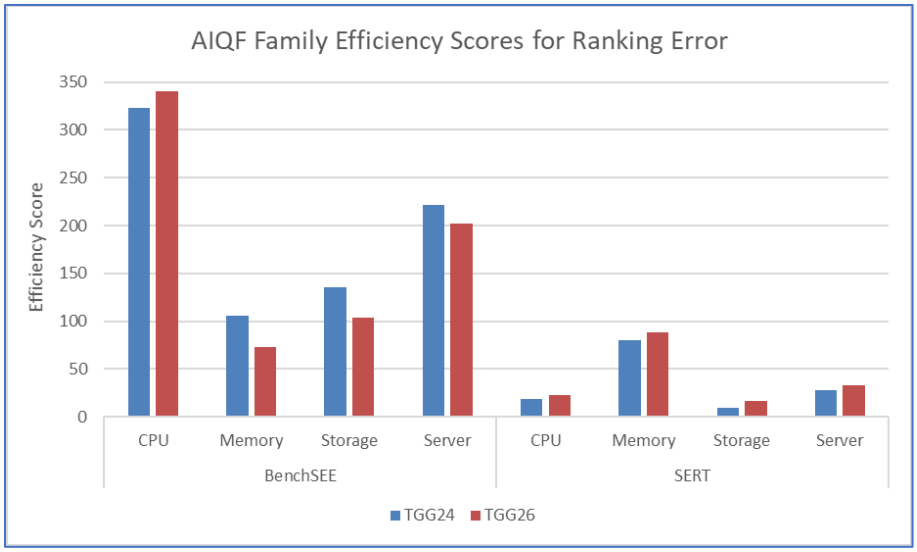

Figure 8:

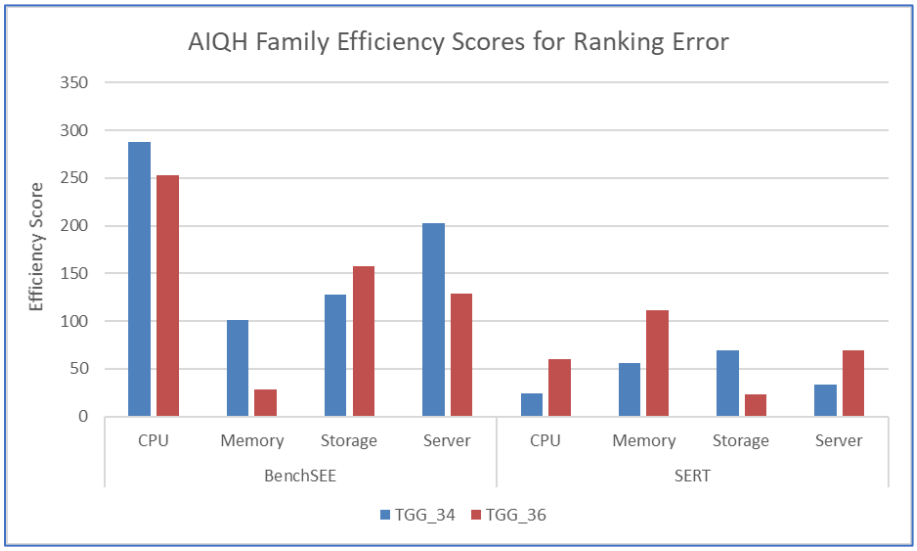

Figure 9:

6 TGG Stage 3 BenchSEE Analysis

Below is a list of types of analysis TGG is considering for Stage 3 BenchSEE work. TGG appreciate any feedback from CNIS .

- Threshold setting analysis
- Storage device type (SSD, HDD, NVME)

- Regulation assessment
  - Exceptions for GPU
  - Scope
- Idle Power measurement
- Long Term support and updating
- Compare BenchSEE and the SERT suite score outputs columns to compare and verify what the data means
- Affinity settings
- Does/should scoring use a reference system?
- 1P vs. 2P Scaling
- Memory frequency scaling
- Node scaling
- Investigate storage workload scaling with memory frequency

7 Trademarks

7.1 SPEC, and the industry standard benchmark name SERT are registered trademarks of the Standard Performance Evaluation Corporation (SPEC). All rights reserved.

8 Annex

TGG Key BenchSEE v1.2.3 Feedback Part 1 Summary (Table 1 – Part 1)

Observation #	Issue Description	Importance	*Requested Implementation Version	Type	Details
TGG1	**Issue:** Most workloads are a compiled code, not in Java **Recommendation:** Rewrite CPU and Memory workloads in Java	Highest	v1.3.0	Design	3.1
TGG2	**Issue**: Java non-default options not allowed	Highest	v1.3.0	Design	3.2

	Recommendation: Adopt benchmarking best-practice for allowing non-default Java options.				
TGG3	**Issue:** BenchSEE exposes log file security concerns **Recommendation:** Remove this feature or redesign with opt-in option	High	v1.3.0	Design	3.3
TGG4	**Issue:** Lack of memory capacity scaling **Recommendation:** Add capacity scaling workload or modify existing workload so performance increases with memory capacity	High	v1.3.0	Design	5.1
TGG5	**Issue:** IBM Power storage workload not functioning **Recommendation:** Ensure BenchSEE cross architecture support	Medium	v1.3.0	Functionality	4.1
TGG6	**Issue:** CPU worklets cache residency **Recommendation:** Increase dataset size of some CPU workloads to be many times larger than the largest current L3 cache and more real world relevant	Medium	v1.3.0	Design	5.4
TGG7	**Issue:** 1 installed CPU server functionality problems **Recommendation:** Fix this issue (mainly seen on low-end configuration)	High	v1.2.4/patch	Functionality	4.2
TGG8	**Issue:** Setup failures, especially with OLTP **Recommendation:** Fix tool stability issues	High	v1.2.4/patch	Functionality	4.1
TGG9	**Issue:** Power analyzer communication protocol security **Recommendation:** Calculate and report power analyzer uncertainty, set the power analyzer's Scaling Value to 1 and lock the manual power analyzer interface duration the run.	Medium	Longer term	Design	3.4
TGG10	**Issue**: Use of power analyzer auto-ranging **Recommendation**: Do not allow the use of auto-ranging	Medium	Longer term	Design	3.4
TGG11	**Issue:** Run-to-run variance, especially in OLTP **Recommendation:** Evaluate and decrease run to run variance	Low	Longer term	Functionality	4.3
	See additional lower priority observations and requests, TGG12 - TGG26, in section 4.1				

*Requested Timeline:

- "v1.2.4/patch" are TGG requested BenchSEE fixes which are blocking portions of TGG's BenchSEE v1.2.3 analysis plan. TGG requests either a quick release of v1.2.4 or a patch to address these issues, which would allow TGG to complete its BenchSEE v1.2.3 Part II feedback.
- v1.3.0 are changes TGG requests before starting a new deep performance and scoring analysis. Without these changes, any deep analysis would be invalided if these changes are made later.
- Longer term are changes TGG believe are needed before BenchSEE is released but will not block its upcoming deep analysis.