Challenges and Future Directions in Efficiency Benchmarking (Vision Paper)

Maximilian Meissner
maximilian.meissner@uni-wuerzburg.de
University of Würzburg
Würzburg, Germany

Klaus-Dieter Lange
powerchair@spec.org
SPECpower Committee Chair
Houston, USA

Sanjay Sharma
isgserverchair@spec.org
SPEC Server Efficiency Chair
Phoenix, USA

Jeremy Arnold
jeremy.arnold@amd.com
AMD
Rochester, USA

Aaron Cragin
aacragin@microsoft.com
Microsoft
Redmond, USA

Patrick Galizia
pgalizia@amperecomputing.com
Ampere Computing
Durham, USA

Mike Petrich
petrichm@us.ibm.com
IBM
Rochester, USA

Bin Zhang
zhangbinbj@inspur.com
Inspur Corporation
Jinan, China

Samuel Kounev
samuel.kounev@uni-wuerzburg.de
University of Würzburg
Würzburg, Germany

ABSTRACT

SPEC benchmarks are crucial contributors behind the improvement of server efficiency since 2007, given their role in making the power consumption and efficiency of servers transparent for government regulators, customers, and the manufacturers themselves.

As the IT landscape experiences radical transformations, efficiency benchmarks need to be updated accordingly to generate results relevant to government regulators, manufacturers, and customers. In this paper, we outline current challenges efficiency benchmark developers are tackling and highlight recent technological developments the next generation of efficiency benchmarks should take into account.

CCS CONCEPTS

• **Hardware** → **Power estimation and optimization**; *Board- and system-level test*; **Power and thermal analysis**; • **Software and its engineering** → *Software performance.*

KEYWORDS

Energy Efficiency, Benchmarking, Cloud, Data Center, Server, Memory Population, Direct Current, Alternating Current, Sustainability, Green Computing, Performance

ACM Reference Format:
Maximilian Meissner, Klaus-Dieter Lange, Sanjay Sharma, Jeremy Arnold, Aaron Cragin, Patrick Galizia, Mike Petrich, Bin Zhang, and Samuel Kounev. 2023. Challenges and Future Directions in Efficiency Benchmarking (Vision Paper). In *Companion of the 2023 ACM/SPEC International Conference on Performance Engineering (ICPE '23 Companion), April 15–19, 2023, Coimbra, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3578245.3585034

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICPE '23 Companion, April 15–19, 2023, Coimbra, Portugal
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0072-9/23/04…$15.00
https://doi.org/10.1145/3578245.3585034

1 INTRODUCTION

Energy efficiency is a critical requirement for IT systems. In recent years, the exponentially growing demand for datacenters did not lead to an exponential increase in electricity usage, due to the development of energy-efficiency techniques and a trend towards large cloud-based service providers [4]. A driving force behind the manufacturers' efforts to make their servers more energy efficient are SPEC benchmarks, which boost the development of increasingly efficient products by making their energy efficiency transparent and comparable across products. These tools are used in marketing as well as certification programs such as the U.S. ENERGY STAR, and, thus, have a significant impact on the industry. However, the IT landscape is subject to rapid change: new technologies, increased usage of specialized accelerators in datacenter servers, enhanced security requirements, as well as increased popularity of novel application types change the deployed hardware and the way it is utilized. Current SPEC efficiency benchmarks are *SPECpower_ssj® 2008*[1], which emulates a transactional enterprise application, and the SERT® 2 suite[2], which consists of a set of synthetic micro-workloads that exercise the CPU, memory, and storage I/O subsystems and indicates server efficiency through a single score. Recently, SPEC announced the next generation of efficiency benchmarks, SPECpowerNext and the SERT 3 suite, to be under development. In order for these future efficiency benchmarks to keep providing relevant, comparable, and fair results, and thereby driving energy efficiency forward in the future, they need to incorporate the aforementioned changes in hardware and software. In this paper, we describe several challenges associated with

[1] https://www.spec.org/power_ssj2008/, accessed January 9, 2023
[2] https://www.spec.org/sert2/index.html, accessed January 9, 2023

the development of next-generation efficiency benchmarks that are actively being tackled by benchmark developers and researchers. The goal of this paper is to provide insight into ongoing discussions about what future efficiency benchmarks will look like, and provide examples of novel workload designs and the challenges associated with their development.

2 CHALLENGES FOR ENERGY EFFICIENCY BENCHMARKING

Benchmarks need to meet quality criteria such as relevance, reproducibility, fairness, verifiability, and usability [3]. In this section, we outline challenges developers currently tackle to ensure these criteria will be met in future efficiency benchmarks.

High Variability in Measures of Power. While even in controlled benchmark environments performance is often variable, there can be even more variability in measures of power. There are many factors that influence server energy-efficiency, such as the hardware and its configuration, the software stack, and the workload of the benchmark itself. In addition, the environmental conditions under which a system is tested influence its power characteristics. For instance, the influence of temperature on the power characteristics of a computing system is significant. A lower temperature typically results in lower power being required to operate the equipment due to lower leakage currents in the servers components and lower power required by the cooling system. Furthermore, the energy efficiency of servers varies with utilization level. Given that servers in datacenters are typically not constantly fully utilized and, instead, are subject to load-intensities that can vary significantly over time, efficiency benchmarks targeting general-purpose servers need to evaluate the servers' efficiency at different utilization levels. In contrast to most other benchmarks, this aspect is addressed by SPEC efficiency benchmarks by defining small-scale transactional workloads, called worklets, which are automatically executed with the necessary inter-arrival times to achieve a certain level of system-utilization.

Such factors need to be considered when assessing the energy efficiency of servers in order to ensure reliable estimates and conclusions. They need to be codified in a fair and complete set of benchmark run- and reporting-rules to ensure the results are valid and comparable across products. The following paragraphs provide an overview over specific factors requiring further investigation.

Increasing Heterogeneity of Hardware. In recent years, as Moore's Law comes to an end, gains in performance are increasingly realized through efficient parallelism and accelerators specialized for specific types of applications [1]. While state-of-the-art CPUs, e.g., AMD Ryzen, or Intel® Xeon® Scalable through its Intel Advanced Vector Extension instruction set and the Intel Math Kernel Library, provide better support for highly parallelizable applications such Deep Learning algorithms than previous generations, Auxiliary Processing Accelerators (APA) such as GPUs, FPGAs, and ASICS are increasingly deployed in datacenter servers to serve the increased demand for computation. As their prevalence in datacenter servers increases, future efficiency benchmarks and rating tools need to incorporate respective workloads. While a variety of benchmarks for different types of APAs exist, SPEC benchmarks aiming to provide a holistic view on a system's efficiency across a wide range of architectures still require the inclusion of such workloads. To showcase good power management, an energy efficiency benchmark targeted towards general-purpose servers needs to consider the energy efficiency at different load-levels, given that datacenter servers typically are not utilized fully. However, existing benchmarks for APAs mainly focus on peak performance and power consumption.

At the same time, Big Data and the rapid increase of datacenter storage capacity over the last years is supported by the usage of dedicated servers for storage. While efficiency benchmarks include storage workloads, storage-servers are currently not their focus. More robust storage workloads are required which can appropriately exercise modern storage and storage heavy servers (servers with 30+ internal storage devices).

Diverse Set of Representative Workloads. The selection of relevant workloads, as well as their development, is a core task for benchmark developers. Since different workloads targeting the same subsystem can vary significantly with respect to power consumption [9], a representative set of workloads needs to be determined. Typically, trade-offs need to be made, e.g., between high relevance for specific use-cases and breadth of applicability. However, benchmarks that provide an overview of a system's overall efficiency over a wide range of applications, and not just efficiency of individual components or efficiency for specific use-cases, are important for increasingly heterogeneous general-purpose devices. Simultaneously, workload complexity should be limited in order to make the benchmark easy to use as well as executable on low-end devices. Therefore, while highly specialized benchmarks for individual components or use-cases have their place, rating-tools such as the SERT suite need to reconcile a broad applicability with a representative set of workloads, which is increasingly challenging since recent years have seen a plethora of new technologies. For instance, widespread application of Machine Learning algorithms, increased interest in Distributed Ledger Technology, and increased virtualization has changed radically the way datacenter hardware is used. The inclusion of this increasing range of potential workloads into a single tool to realize high relevance, broad applicability, and ease of use at the same time is an important challenge faced by benchmark developers.

Support for Direct Current Servers. Most PDUs provide servers with Alternating Current (AC), which is then rectified by the PSU to the Direct Current used by the servers' internal components. While less common, datacenters operating with Direct Current (DC) could yield a variety of advantages, such as increased efficiency and reliability [7]. However, popular benchmarks such as *SPECpower_ssj 2008* and the SERT suite explicitly discourage the comparison of results obtained on AC systems with results on DC systems, and the latter are not supported for run-rule compliant benchmark executions. This is due to the following reasons:1) There are substantial differences between datacenter infrastructure that uses DC compared to AC. Since the power received from the external power supplier is usually AC, an additional converter is required to provide a datacenter with DC power. The power loss associated with this conversion process is not measured in a benchmark setting, and 2) Measurement technology for AC power and DC power differs,

particularly in the calculation of uncertainty. It is difficult to look at a combination of AC and DC measurements and know for certain that they have comparable levels of uncertainty. However, future efficiency benchmarks officially should support DC servers such that regulators can include them in certification programs such as the U.S. ENERGY STAR. As a result, there would be one less obstacle for the adoption of DC datacenters. To this end it will be necessary to either find a methodology that enables fair comparisons between AC and DC results, or it will be demonstrated that AC and DC results are fundamentally incomparable and a separate metric for DC devices is defined.

Discrepancy Between Tested and Deployed Memory Configurations. Standards such as the ISO/IEC 21836:2020[3] currently require that all memory channels in the System Under Test (SUT) are populated in order to obtain compliant results. However, due to the increase in the number of DIMM-slots and memory channels in servers in recent years, there is an increasing gap between the configuration of the SUT tested during the certification process and the configurations of the SUTs that are actually deployed by customers. This is because servers are often not sold with all memory channels populated, due to the cost of DIMMs. Instead, customers estimate the amount of RAM they need and buy the server only configured with the corresponding amount of DIMMs that provide the required capacity. While it is possible to run the SERT suite without all memory channels populated, it is unclear whether the efficiency score obtained with such configurations constitutes a fair representation of the SUT's energy efficiency. Testing servers with unpopulated memory channels decreases performance and efficiency scores due to the reduced memory bandwidth, which affects both, CPU- and memory-workloads. An earlier study demonstrates that performance and power consumption are affected differently by such changes to the configuration and that the effect on energy-efficiency is not trivial [9]. As a result, the tested and certified versions of servers differ with respect to energy efficiency from the servers actually deployed, and low-end configurations might not pass the certification process. In order to support the testing and certification of such configurations by adapting the SERT suites run-rules, it is necessary to conduct extensive experiments analyzing the effect of such deviations from the mandated configuration on the efficiency score in order to determine whether the score scales accordingly and still constitutes a valid representation of an SUT's energy efficiency.

Increased Diversity in Cooling Technology. Due to the trend of increased usage of high-power APAs and state-of-the-art CPUs in servers to serve the increasing demand for high computational power in datacenters, the requirements for the cooling systems to efficiently cool this high-density equipment increases as well. This leads to an increased interest in liquid cooling technology beyond the area of scientific High Performance Computing, as liquid cooling can exhibit a much higher capacity to remove heat and a higher efficiency than air cooling. While servers with *Liquid Assisted Air Cooling*, where the liquid only circulates inside the server and releases the heat through an air-cooled heat exchanger into the room, could be supported by existing benchmarking methodologies, cooling technologies such as *Direct Liquid Cooling* where liquid leaves the SUT to transfer the heat to an external chiller require new standardized methodologies that identify and specify the important factors that need to be measured, such as the temperature of the liquid at the inlet and outlet, as well as its flow-rate. In addition to the definition of a novel, standardized benchmarking methodology, further research questions arise. For instance, thorough experimental evaluation needs to determine whether there are environmental conditions under which fair energy-efficiency comparisons between air-cooled and liquid-cooled servers can be achieved, or if they are fundamentally incomparable and require different categories in certification programs.

Global Standardization. An overarching challenge is the creation of an energy-efficiency standard that can be globally accepted. Having different standards in different locations results in hardware manufacturers having to invest immense time and resources to test every product for each of those standards separately. A single global standard based on a well-defined efficiency-metric would not only serve to reduce the effort required, it would also boost the development of novel techniques for improving energy efficiency and reducing carbon emissions by making products globally comparable. Additionally, it is easier to analyze standardized datasets in an effort to identify potential improvements in hardware development.

3 PROPOSED WORKLOAD TYPES

In this section we propose examples of workload types the upcoming generation of SPEC efficiency benchmarks could incorporate to cover important application areas.

3.1 APA Workload

Similar to CPU workloads, an APA workload must scale with the number of APAs (i.e., it must properly run on a variable number of APAs), their clock frequencies, the installed memory on the APAs, and the memory speeds. Additional factors benchmark developers need to consider are 1) how to integrate the results into existing metrics such as the SERT score, and 2) how to achieve fair comparisons across different architectures. The former aspect is not only related to the question of how to weigh an APA workload in an efficiency metric, but also to the question of how to make comparisons across systems that have different types of accelerators or no accelerators at all. The latter aspect relates to the comparison of systems with APAs of different vendors and architectures. For instance, specific workloads such as Machine Learning algorithms can take advantage of lower precision for certain substeps to increase performance, and possibly, energy efficiency. However, not all architectures provide the same set of available precisions. For instance, NVIDIA Volta GPUs provide *Tensor Cores*, a special functional unit to boost the performance of matrix multiplications by enabling mixed-precision computing. Furthermore, different vendors support different libraries and standards, which further complicates the development of a workload that works on different platforms while taking advantage of individual platform-specific features. Consequently, important considerations for the creation of an APA workload are: *How to design a representative, mixed-precision workload and how to fairly compare the results of platforms*

[3] https://www.iso.org/standard/71926.html, accessed January 9, 2023

with different precisions. How to compare results from workload implementations using different libraries. How to compare such systems against non-APA systems. and *How to compare systems with different APUs.* Initial research for energy-efficiency benchmarking of GPUs at different load-levels has been conducted in [10]. We aim to build on this research by investigating both higher and lower precision worklets to answer those questions. A representative APA workload should consist of basic operations such as matrix multiplications at different precisions (e.g., *HGEMM*, *IGEMM*, *SGEMM*, *DGEMM*), or *STREAM* and *FFT*, as well as worklets representing real world applications from domains such as Machine Learning, e.g., Natural Language Processing transactions, and Scientific Computing. These worklets need to be defined in a transactional form such that they can be executed in accordance with the established SPECpower measurement methodology, an updated version which integrates GPUs, is shown in Figure 1.

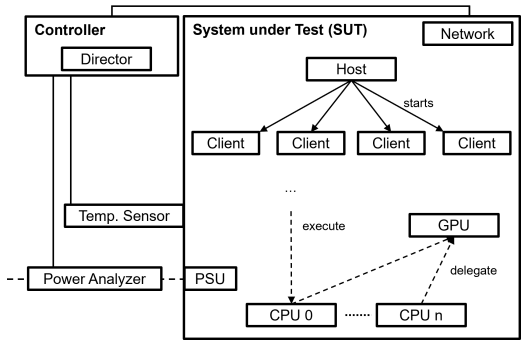

Figure 1: Setup for transactional energy efficiency benchmarking of servers with GPUs

3.2 Updated Security Workload

Security remains a primary use case across the server industry. Cryptography has been a persistent security driver since its inception. Ciphers, key exchange algorithms, and signature schemes are used pervasively across virtually all platforms and are popular workloads for performance and power benchmarks, especially for stressing processors. As new requirements arise with respect to security, novel algorithms need to replace older ones in energy efficiency benchmarks. For instance, SPECpower_ssj 2008 does not include security workloads. The SERT 2 suite currently includes CryptoAES 128 and SHA-2: 256. Future versions should include emergent cypher support with stronger key length (AES-256, SHA2-512, RSA 4096 bit, ECDSA P-384 etc.). In addition, benchmark developers should already prepare themselves for the inevitable Post Quantum Cryptography era, as, for instance, NIST Cyber Security task forces already are preparing for this transition [6].

3.3 Distributed Ledger Technology

Related to the enhanced requirements for security is the increasing interest in Distributed Ledger Technology (DLT), which offers a cryptographically secure, distributed database. Distributed Ledger *Proof of Work* worklets could be realized by implementing a transaction that will generate – given a set of data representing a block in a blockchain – a given number of iterations where the nonce of the block is incremented, a hashing routine is run on the block, and the returned hash is checked for validity. In a real Distributed Ledger using proof-of-work as the consensus algorithm, the returned hash is checked if it passes a given criteria. For example, in the bitcoin protocol, the returned hash must begin with a specified number of zeros (the actual number fluctuates – fewer zeros are required if not enough blocks are added in each time period, or more zeros are required if too many blocks are added in the same period [5]). While computing the hash of a given block is designed to be fast, finding the nonce value for that block that generates the hash with the specified format can be found anytime between the first iteration and millions of iterations later. To provide more consistent results from run-to-run, the worklet could use a fixed iteration count as opposed to exiting early due to finding a valid hash. This allows for more repeatable results run-to-run; whereas, exiting upon finding a valid hash will result in some transactions exiting in milliseconds, while others could take hours. Each iteration should have its own randomly generated block, which will then be hashed. Since the worklet does not stop when a valid hash is found, but executes the given number of iterations, results can be replicated reliably from run-to-run. Alternatively, different hash algorithms can be added as well, supported by other blockchains and specified as part of the input. This can be a stretch goal once the initial worklet is working and tested.

3.4 Networking

The ability to efficiently communicate with other devices is crucial for servers. With the increase in data traffic, evaluating the respective subsystem is gaining importance, but current SPEC efficiency benchmarks do not include workloads for the network interface cards (NICs) of a server. Similar to workloads stressing other resources, network workloads must scale with 1.) the number of NICs installed, and 2.) the NIC bandwidth. Current SPEC efficiency benchmarks target a single SUT. A challenge related to the design of a networking workload therefore is the execution of networking functions on a single system. Ongoing discussion focuses on whether this is possible with a single system such that the resulting usage patterns constitute a realistic representation of real-world systems, while keeping the economical setup of SPEC efficiency benchmarks. If this turns out to be infeasible, the benchmarking methodology [8] would need to be extended to include a second system for the creation of network traffic. When this obstacle is overcome, a networking workload could be implemented to apply common networking functions on a stream of network packets. There are some packet processing functions commonly used in most network applications in the industry. These functions include but are not limited to parsing the packet header, looking up the packet information in a forward information base (FIB) (aka routing table), calculating packet checksum, and copying packets from one memory location to another. The input is the stream of packets provided to the application in the form of a packet capture file (.pcap). This packet capture file shall be loaded to main memory prior to the workload execution to avoid IO operations. The packet stream is processed by a single core, and the output is the number of packets processed per second.

3.5 Virtualization - VDI and Serverless

Virtualization remains an increasing trend, due to many advantages such as the consolidation of workloads and better utilization of the available hardware. Today, the concept of virtualization manifests itself in a variety of forms. For upcoming SPEC efficiency benchmarks, we propose to investigate workloads representing Serverless Computing as well as Virtual Desktop Infrastructure (VDI). Serverless computing is predicted to be the predominant cloud-computing model in the near future [2]. VDI is a desktop virtualization technology, for example, provided by Citrix, Microsoft, VMware Horizon, and Parallels, which leverages VMs to provision and manage virtual desktops / applications and is managed on datacenter servers. *How to create a workload realistically emulating the behavior of these computing models, while being economically executable on a single SUT* is challenging. A serverless workload could, for instance, test container load and shutdown time by reading a compressed image from disk, decompressing it, loading it into memory, and then shutting down and clearing memory.

To the best of our knowledge, there currently exists no VDI benchmark. In order to keep the benchmark suite easy-to-use, the workload would emulate the server-side part of a VDI solution based on CPU, memory, and storage traces from real VDI solutions, so that no actual VM or VDI solution needs to be installed on the SUT. Different types of users with different desktop applications, for example, word processing, spreadsheets, coding, presentation maker, could be emulated and combined to a representative transaction mix.

4 CONCLUSION

Efficiency benchmarks are a driving force behind the mitigation of rising energy consumption of IT devices. The continuous development of these tools is confronted with a variety of challenges, due to recent advancements in the IT-landscape, from the increasing adoption of accelerators specialized for certain kinds of applications, new requirements regarding cryptographic algorithms, and increased usage of virtualization, to an entirely new datacenter infrastructure using DC power supplies or novel cooling technologies. Next-generation efficiency benchmarks need to take these developments into account in order to keep pushing the development of energy-efficient datacenter servers and realizing the vision of a global energy-efficiency standard.

REFERENCES

[1] John L. Hennessy and David A. Patterson. 2019. A New Golden Age for Computer Architecture. *Commun. ACM* 62, 2 (jan 2019), 48–60. https://doi.org/10.1145/3282307

[2] Eric Jonas, Johann Schleier-Smith, Vikram Sreekanti, Chia-che Tsai, Anurag Khandelwal, Qifan Pu, Vaishaal Shankar, João Carreira, Karl Krauth, Neeraja Jayant Yadwadkar, Joseph E. Gonzalez, Raluca Ada Popa, Ion Stoica, and David A. Patterson. 2019. Cloud Programming Simplified: A Berkeley View on Serverless Computing. *CoRR* abs/1902.03383 (2019). arXiv:1902.03383 http://arxiv.org/abs/1902.03383

[3] Samuel Kounev, Klaus-Dieter Lange, and Jóakim von Kistowski. 2020. *Systems Benchmarking* (1 ed.). Springer International Publishing. https://doi.org/10.1007/978-3-030-41705-5

[4] Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020. Recalibrating global data center energy-use estimates. *Science* 367, 6481 (2020), 984–986. https://doi.org/10.1126/science.aba3758 arXiv:https://www.science.org/doi/pdf/10.1126/science.aba3758

[5] Satoshi Nakamoto. 2009. Bitcoin: A Peer-to-Peer Electronic Cash System. *Cryptography Mailing list at https://metzdowd.com* (03 2009).

[6] National Institute of Standards and Technology (NIST). 2023. Post-Quantum Cryptography Standardization. https://csrc.nist.gov/Projects/post-quantum-cryptography/post-quantum-cryptography-standardization. Accessed: January 9, 2023.

[7] V. Sithimolada and P. W. Sauer. 2010. Facility-level DC vs. typical ac distribution for data centers: A comparative reliability study. In *TENCON 2010 - 2010 IEEE Region 10 Conference*. 2102–2107. https://doi.org/10.1109/TENCON.2010.5686625

[8] Standard Performance Evaluation Corporation (SPEC). 2014. Power and Performance Benchmark Methodology V2.2. https://www.spec.org/power/docs/SPEC-Power_and_Performance_Methodology.pdf

[9] Jóakim v. Kistowski, Hansfried Block, John Beckett, Klaus-Dieter Lange, Jeremy A. Arnold, and Samuel Kounev. 2015. Analysis of the Influences on Server Power Consumption and Energy Efficiency for CPU-Intensive Workloads. In *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering* (Austin, Texas, USA) *(ICPE '15)*. Association for Computing Machinery, New York, NY, USA, 223–234. https://doi.org/10.1145/2668930.2688057

[10] Jóakim von Kistowski, Johann Pais, Tobias Wahl, Klaus-Dieter Lange, Hansfried Block, John Beckett, and Samuel Kounev. 2019. Measuring the Energy Efficiency of Transactional Loads on GPGPU. In *Proceedings of the 2019 ACM/SPEC International Conference on Performance Engineering* (Mumbai, India) *(ICPE '19)*. Association for Computing Machinery, New York, NY, USA, 219–230. https://doi.org/10.1145/3297663.3309667