SPEC Efficiency Benchmark Development: How to Contribute to the Future of Energy Conservation

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

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

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

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

Roger Tipley
isgclientchair@spec.org
SPEC Client Efficiency Chair
Houston, USA

Nishant Rawtani
nishant.rawtani@hpe.com
Hewlett Packard Enterprise
Bengaluru, India

David Reiner
david.reiner@amd.com
The Green Grid Chair
Austin, USA

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

Aaron Cragin
aacragin@microsoft.com
Microsoft
Redmond, USA

ABSTRACT

A driving force behind the improvement of server efficiency in recent years is the use of SPEC benchmarks. They are used in mandatory government regulations, the ISO/IEC 21836:2020 standard, and product marketing, giving server manufacturers and buyers significant incentive to improve energy efficiency. To produce relevant results, benchmarks need to take into account future trends in hardware and software development, such as the introduction of new accelerators and workloads. To keep pace with the development of the fast moving IT landscape, SPEC plans to introduce a workload bounty program to encourage researchers to develop novel workloads. Submitted workloads will be considered for inclusion in future SPEC Efficiency benchmarks and rewarded. In this paper, we outline the process of energy-efficiency benchmark development. SPEC ensures the development of high-quality benchmarks for government regulations through its extensive experience and collaboration with stakeholders from industry, academia, and governments. One of the tools that emerged from this process is the Chauffeur Worklet Development Kit (WDK), which can be used by researchers to develop next-generation workloads to enhance the real-world relevance of future SPEC benchmarks, a critical element for the benchmarks to contribute to future energy conservation.

CCS CONCEPTS

• **Hardware** → **Energy metering**; **Platform power issues**; **Enterprise level and data centers power issues**.

KEYWORDS

Client, Datacenter, Energy Efficiency, Power, Performance, Benchmarking, Benchmark Development, Server, SPEC, Sustainability

ACM Reference Format:
Maximilian Meissner, Klaus-Dieter Lange, Jeremy Arnold, Sanjay Sharma, Roger Tipley, Nishant Rawtani, David Reiner, Mike Petrich, and Aaron Cragin. 2022. SPEC Efficiency Benchmark Development: How to Contribute to the Future of Energy Conservation. In *Companion of the 2022 ACM/SPEC International Conference on Performance Engineering (ICPE '22 Companion), April 9–13, 2022, Bejing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3491204.3527492

1 INTRODUCTION

Energy efficiency is a major concern in the IT sector, given that high power consumption has negative economical, operational, and environmental consequences. The operational consequences of high power consumption include the resulting heat, which can lead to increased cooling costs and wearout of the devices, as well as decreased service time in battery-powered devices such as tablets and smartphones. It is estimated that the energy consumption of datacenters will rise to more than 1 PWh by the year 2030, even in the best case scenario [1]. The industry and regulators alike strive towards mitigating this trend. A driving force behind hardware manufacturers' efforts to make their products more energy efficient are SPEC benchmarks. The impact an efficiency benchmark can have on the industry heavily depends on many factors that go far beyond the act of writing the code. Within SPEC, competing companies collaborate to create fair benchmarks by all stakeholders and fulfill other important quality criteria necessary for the success of the benchmark. To achieve standardization, collaboration with government agencies is required. In order to create benchmarks that are considered *relevant* to the end user or government agencies, benchmark development is a continuous effort that needs to consider current and future trends in hardware and software development. In order for the results of the execution of a benchmark

ICPE '22 Companion, April 9–13, 2022, Bejing, China

ACM ISBN 978-1-4503-9159-7/22/04…$15.00
https://doi.org/10.1145/3491204.3527492

to provide meaningful and comparable results, organizations such as SPEC continuously apply their expertise to crafting run and reporting rules as well as the development of new workloads. In this work we illustrate the continuous effort required for producing and maintaining high quality benchmarks, which can be used by government agencies within their regulatory programs, such as the U.S. EPA ENERGY STAR program. We highlight the importance of developing new workloads for the next generation of benchmarks. This work is structured as follows: Section 2 provides insight into the larger context of benchmark development. In Section 3, we further illustrate the benchmark quality requirements in the context of the SERT suite. In Section 4, we introduce the Chauffeur WDK, a framework for the development and execution of workloads for energy-efficiency benchmarking.

2 DEVELOPMENT OF ENERGY EFFICIENCY BENCHMARKS

In this section, we first outline the benchmark development process applied within SPEC, highlighting important factors that need to be considered. Furthermore, we describe the interaction with and influence on government regulations.

2.1 The Benchmark Development Process

Establishing a benchmark requires much more than its mere implementation. A benchmark can be considered successful only if it is actually used in practice to evaluate and compare different systems. Within SPEC, many different hardware manufacturers and research institutions collaborate in order to create fair benchmarks via a data-driven development process[3]. The involved experts evaluate with an aim to reach consensus regarding design and implementation of a benchmark. Evaluations focus on experiments, analysis of results, and relevant research to ensure benchmark quality metrics are met. As illustrated in Figure1 a benchmark passes through typical steps known from software development. Then, it needs to be developed, released, maintained, and finally retired. The development of a benchmark in SPEC starts when one or multiple of the volunteers within SPEC identify an area of interest, e.g., the necessity for a client efficiency benchmark. After determining the high-level concept, the experts involved want to explore, they identify what knowledge from previous benchmark developments can be reused, and which additional research needs to be done for the new area of interest. Due to many years of experience, existing code, as well as the processes and frameworks that have been proven to be of high quality, can be leveraged. These different items are adapted and enhanced if the ongoing research reveals the necessity to do so. Workload creation and development is an iterative process. For instance, during creation of the workload, it might be necessary to conduct further research in order to ensure the creation of a high-quality workload. Upon finalization of the development phase, the benchmark has to pass through a vast test matrix in order to confirm that the quality requirements are met. The different vendors, through their representatives within SPEC, verify that the benchmark is working properly within their environments. When the benchmark is developed for a specific government agency, corresponding beta programs are created. The launch of the benchmark includes the education of the people who are going to use it. It is ensured that they can utilize the benchmark just by using the documentation, which therefore needs to be of high quality. In order to be able to focus on the actual development and keeping the effort for maintenance and support low, paying attention to the quality requirements throughout the whole process is crucial. Eventually, a version of a benchmark needs to be retired and replaced with either a new version or an entirely new benchmark. The experience, tools, and methods developed during the lifetime of a benchmark can be reused, if necessary, in an improved form in the next generation benchmark. An important goal throughout this process is to ensure that the benchmarking quality criteria, such as **Relevance**, **Reproducibility**, **Fairness**, **Verifiability**, and **Usability**, are met[2]. A benchmark is considered *relevant* if its behavior correlates to behaviors that are of interest to consumers of the results, i.e., the final metric generated must be of use to the end customer or the government agency. *Reproducibility* refers to the ability to consistently produce similar results when the benchmark is run with the same test configuration. *Fairness* means to allow different test configurations to compete on their merits. SPEC accelerates the *fairness* of benchmarks by having a whole series of different competing companies and researchers working together. *Verifiability* means the property of providing confidence and trust in the benchmark results. For instance, for the government agencies, log files are created which are encrypted to prevent manipulation of the results. A benchmark exhibits *Usability*, also referred to as being *Economical*, if it is easy to execute in the users' test environments. This

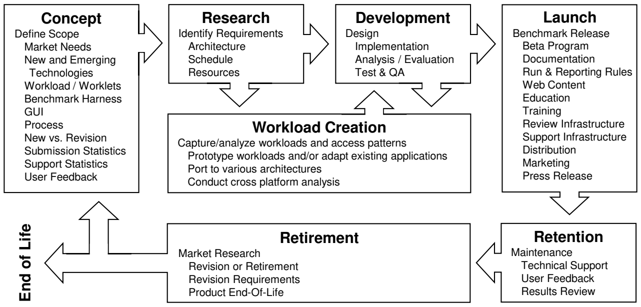

Figure 1: SPEC's benchmark development cycle[2]

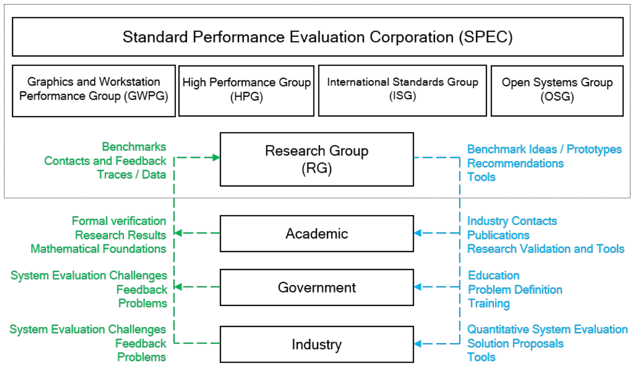

Figure 2: SPEC Overview, updated from[2]

includes a balance between the best solution from a technical and an economic standpoint, for instance, with respect to the requirements regarding the equipment necessary to measure power and temperature. Another important aspect is the formulation of run and reporting rules. Run rules define the conditions under which measurements are performed. Reporting rules ensure that the run rules have been followed. They ensure sufficient documentation of the system under test, such that the conducted tests and their results are *reproducible*, *meaningful*, and *comparable* to results obtained on other systems. The creators of the reporting rules have to ensure that someone just with the report itself is able to reproduce the results so that they can be verified. The results obtained by executing a benchmark can vary significantly, depending on the configuration of the system. An ill-defined set of run and reporting rules may allow for ambiguities that can be exploited to artificially improve the results. Finding a reasonable set of restrictions such that the conditions are fair across the available platforms is critical.

2.2 Influence on Government Regulations

The SERT suite is an industry benchmark created for government agencies and their programs and is used in different energy-efficiency programs worldwide. Our goal is to achieve global harmonization, i.e., having a single international standard benchmark for every region and nation for regulations of systems' energy efficiency. If every region has its own standard, companies have to conduct measurements for each of those standards and for each model they want to bring to the market individually. Having one global standard reduces the effort, time, and money required and makes the energy-efficiency measures comparable across the regions, thereby driving innovation in the area of energy efficiency. It enables manufacturers of servers, or IT equipment in general, to have a clearer focus for how to build their systems in order to improve their energy efficiency. Furthermore, researchers can analyze the standardized datasets easier in their effort to solve energy efficiency-related problems in operating systems or in hardware development.

Within SPEC, driving this vision forward is the mission of the International Standards Group (ISG). Its purpose is to oversee the establishment and development of standardized benchmarks primarily for the use in government regulations and programs[4]. ISG collaborates with national and international standards development organizations to enhance global standards. The founding members of ISG already achieved the standardization of the server energy effectiveness metric (SEEM) with the release of ISO/IEC 21836:2020 in August 2020. Within ISG, the *Server Efficiency Committee* maintains ISO 21836:2020 compliance of all future SERT releases. The *Client Efficiency Committee* is responsible for the creation of a client benchmark / standard. Among others, ISG collaborates with the U.S. EPA, the European Commission, the *Japan Electronics and Information Technology Industry Association*, and the *China National Institute for Standardization*. In particular, there is close collaboration between the ISG committee, the SPECpower committee, and The Green Grid (TGG). TGG is a non-profit organization and an industry consortium that focuses on the overall energy efficiency of datacenters. TGG works on the technical standards, e.g., it created the Power Usage Effectiveness (PUE) standard, which is of high relevance for datacenters. TGG also works with governments on energy-efficiency programs and regulations. For example, TGG is working with the EPA on the ENERGY STAR program for datacenters, which is separate from the ENERGY STAR program for servers, thus creating an energy-efficiency score for entire datacenters. Creating regulations that incentivize energy efficiency requires close collaboration between industries that provide technical expertise, and government agencies that formulate the regulations such that they actually achieve their goal. An important aspect in this context is *thresholding*. Typically, regulators have a target percentage of devices that they want to certify. For instance, in the ENERGY STAR program, the most energy-efficient servers are certified (e.g., top 25%). In the EU, where the certification is a market entry requirement, regulators choose to exclude the least energy-efficient servers (e.g., bottom 15%). TGG maintains a database of SERT results. Regulators determine the SERT score, which is required to obtain the target pass- or fail-rate based on SERT results from recent servers. In order to create the next version of the SERT suite, TGG and SPEC together, governments and other stakeholders around the world discuss potential improvements as well as changes in industry and politics that need to be taken into account. An important aspect of this discussion within SPEC is associated with the selection and creation of new workloads. For instance, SPECpower_ssj 2008 uses Server Side Java Transactions (New Order, Payment, Order Status, Delivery, Stock Level, Customer Report), while the SERT suite uses seven CPU Worklets, two Storage Worklets, two Memory Worklets, and one Idle Worklet. Given the continuous emergence of new technologies in hardware and software, the selection of workloads needs to be revised continuously. In [5], members of the SPECpower committee showcase the development of two novel workloads accounting for changes in server usage scenarios, which could be included in the upcoming *SPECpowerNext benchmark*. Having multiple workloads is necessary not only because they stress different parts of the system, but also because workloads targeting the same system components can vary with respect to the resulting power consumption[7].

3 QUALITY CHARACTERISTICS OF THE SERT SUITE

The SERT suite is built on existing SPEC methods and expertise to ensure that the benchmark quality criteria are met. It assesses the energy efficiency of a server across a wide spectrum of configurations, and the set of supported configurations is updated continuously. Therefore, it is considered **relevant**. Its results are **reproducible**, since it uses a predefined set of accepted client configurations, which ensures tight run-to-run variations. Furthermore, it produces an Extensive Full Disclosure Report (FDR) description. A client configuration[1] applies to a specific combination of processor architecture, operating system, and JVM, and specifies the runtime parameters that will be used with this configuration. The SERT suite is **fair**, given that it is developed by an industry consortium and is portable across architectures, operating systems, and platforms. It features high **usability** since it is designed to be simple to configure as well as economically and easy to use with minimum equipment and skill required. An active update and support process is in place. Its results are **verifiable** due to the automatic discovery

[1] https://www.spec.org/sert2/SERT-JVM_Options-2.0.html

of hardware and software stack, described in the FDR, and due to incorporated test for the verification of compliance with run rules and the detection of result tampering. These validity checks are performed at the end of each run. In case the detected configuration or the results do not fulfill the requirements, the run is marked as *invalid*, and the reasons are included in the report.

There are two levels of platform support for the SERT suite: **enablement** and **acceptance**. Platform enablement relates to functional support for a processor architecture, operating system, or JVM. While enablement is required for accepting new platform configurations, it is not sufficient for guaranteeing compliant runs. To enable support for a new processor architecture, operating system, or JVM, SPEC members implement the platform-specific code necessary for the SERT suite to run properly in the new environment. In order to demonstrate that the new environment can achieve compliant SERT results on at least one representative system configuration, extensive testing is required. For SPEC to add enablement support officially for the new platform, it is required that future support and testing of the platform is guaranteed. In order to obtain *valid* measurements, only client configurations **accepted** by SPEC are allowed. The process for client configuration acceptance is defined in ISO/IEC 21836:2020.

4 DEVELOPING WORKLETS WITH CHAUFFEUR WDK

The SPECpower Committee developed the Chauffeur WDK to support research and the development of new worklets for energy-efficiency testing. It enables researchers to focus on the development of the actual workloads. The framework is implemented in Java to facilitate cross-platform compatibility. It is designed to solve cross-discipline challenges around energy-efficiency benchmarking, which the SPEC has already solved and codified in its measurement methodology for measuring performance and power of computer systems. For instance, it takes care of calibrating and executing the transactional worklets at different load levels, gathering the temperature and power measurements, and producing the final report. The main aspect of the development of a new worklet comprises the definition of at least one **Transaction** and a **User**. The implementation of the **User** can hold state information across transaction executions. The **Transactions** are used to implement the actual business logic of the worklet. A Transaction needs to implement the methods *generateInput* and *process*, which are defined by the Transaction interface. The former generates input for the transaction, while the latter uses this input as well as a user instance to generate a result. The Transaction class can be used either to implement the business logic directly, or to call existing code, e.g., integrating existing workloads. Running application logic not implemented in Java is possible by calling native code through JNI or JNA, launching an external process, or invoking REST calls to an external service running either on the local host or a remote system. The ability to run at multiple load levels depends on having relatively short running transactions, which enables the scheduling of subsequent transactions with a delay, such that the target utilization of the system is achieved. A traditional benchmark would immediately execute the next transaction, as soon as one transaction completes. However, Chauffeur also supports non-transactional worklets. For instance, this is leveraged in the SERT Flood worklet, a memory bandwidth test for which iterating through possibly multiple terabytes of memory at different load-levels would take a fair amount of time. Multiple worklets can be run simultaneously, e.g., a CPU- and an IO-intensive transaction at different load levels, by implementing a worklet combining transactions from other worklets. Chauffeur worklets are configurable via adjustable runtime parameters without making further code-changes.

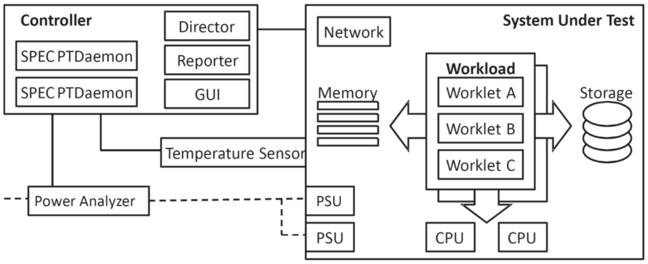

Figure 3: Chauffeur Design Overview, updated from[6]

5 CONCLUSION

Energy-efficiency benchmarking is a vital part of a sustainable IT-industry. We share insights into the benchmark development process where competing companies work together in SPEC to define the benchmark workloads to fulfil the quality criteria necessary for its effectiveness. They also define the run and reporting rules to measure valid and meaningful benchmark results. Because benchmarks need to consider future trends in hardware and software development, novel workloads are required in order to produce results relevant to the users. With the Chauffeur WDK, a benchmarking framework is available for researchers to focus on the development of workloads for the evaluation of new technologies' energy efficiency.

REFERENCES

[1] Anders Andrae and Tomas Edler. 2015. On Global Electricity Usage of Communication Technology: Trends to 2030. *Challenges* 6, 1 (Apr 2015), 117–157.

[2] Samuel Kounev, Klaus-Dieter Lange, and Jóakim von Kistowski. 2020. *Systems Benchmarking* (1 ed.). Springer International Publishing.

[3] Klaus Dieter Lange, Jeremy A. Arnold, Hansfried Block, Nathan Totura, John Beckett, and Mike G. Tricker. 2013. Further Implementation Aspects of the Server Efficiency Rating Tool (SERT). In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering (ICPE '13)*. Association for Computing Machinery, New York, NY, USA, 349–360. https://doi.org/10.1145/2479871.2479926

[4] Norbert Schmitt, Klaus-Dieter Lange, Sanjay Sharma, Aaron Cragin, David Reiner, and Samuel Kounev. 2021. *SPEC — Spotlight on the International Standards Group (ISG)*. Association for Computing Machinery, New York, NY, USA, 167–168. https://doi.org/10.1145/3447545.3451171

[5] Norbert Schmitt, Klaus-Dieter Lange, Sanjay Sharma, Nishant Rawtani, Carl Ponder, and Samuel Kounev. 2021. *The SPECpowerNext Benchmark Suite, Its Implementation and New Workloads from a Developer's Perspective.* Association for Computing Machinery, New York, NY, USA, 225–232. https://doi.org/10.1145/3427921.3450239

[6] Standard Performance Evaluation Corporation (SPEC). 2017. Chauffeur™ Worklet Development Kit (WDK) User Guide 2.0.0. https://www.spec.org/chauffeur-wdk/docs/Chauffeur-User_Guide.pdf

[7] 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 (ICPE '15)*. Association for Computing Machinery, New York, NY, USA, 223–234. https://doi.org/10.1145/2668930.2688057