Volume 3, Issue 1
April 2020

# SPEC Research[SM] Group Newsletter

## CONTENTS



## SPEC KAIVALYA DIXIT DISTINGUISHED DISSERTATION AWARD 2019

This year, the selection committee has picked two dissertations: the winner, Guanpeng Li of University of British Columbia, and a runner up, Bin Nie of the College of William & Mary, based on the high quality of both submissions.

Read more on page 4

## ICPE 2021 WILL BE HELD IN RENNES, FRANCE

Johann Bourcier and Zhen Ming (Jack) Jiang, the General Chairs of the next ACM/SPEC International Conference on Performance Engineering (ICPE 2021), invite interesting high-quality submissions. The conference will take place April 19-23, 2021 in Rennes, France.

Read more on page 5

## SPEC RESEARCH WORKING GROUPS REPORT ON THEIR PROGRESS

The six SPEC Research Working Groups Security, Cloud, Big Data, DevOps Performance, Quality of Experience, and Power report on their progress, articles, benchmarks, and technical reports published in 2019. The Working Groups are always open for new members, feel invited to join us!

Read more on pages 5-11

## ICPE: EXPERIENCE WITH REPRODUCIBILITY BADGES

The reproducibility badges highlight the peer-reviewed papers whose artifacts have been further evaluated in practice, and their results have been reproduced by an independent team of reviewers. Four papers received this year the badges.

Read more on page 11

## CONTACT

Standard Performance Evaluation Corporation (SPEC)
7001 Heritage Village Plaza, Suite 225
Gainesville, VA 20155, USA

**SPEC Research Group**
Chair: Samuel Kounev (rgchair@spec.org)
Web: http://research.spec.org

## SPEC RESEARCH GROUP OFFICERS

**Chair:**
**Samuel Kounev**, University of Würzburg, Germany

**Vice-Chair:**
**André van Hoorn**, University of Stuttgart, Germany

**Secretary:**
**Cor-Paul Bezemer**, University of Alberta, Canada

**Steering Committee:**
**J. Nelson Amaral**, University of Alberta, Canada
**Cor-Paul Bezemer**, University of Alberta, Canada
**Alexandru Iosup**, VU Amsterdam, NL
**Samuel Kounev**, University of Würzburg, Germany
**Klaus-Dieter Lange**, HPE, USA
**Meikel Poess**, Oracle Corporation, USA
**Petr Tůma**, Charles Univ. of Prague, Czech Republic
**Weiyi Shang**, Concordia University, Canada
**André van Hoorn**, University of Stuttgart, Germany
**Boris Zibitsker**, BEZNext, USA

**Publicity Officer:**
**Weiyi Shang**, Concordia University, Canada

**Release Manager:**
**Vojtěch Horký**, Charles Univ. Prague, Czech Republic

**Newsletter Editors:**
**André Bauer**, University of Würzburg, Germany
**Nikolas Herbst**, University of Würzburg, Germany

http://research.spec.org

## SPEC RESEARCH WORKING GROUPS

### Cloud Working Group

**Chair:**
**Alexandru Iosup**, VU Amsterdam, The Netherlands

**Vice-Chair & Secretary:**
**Nikolas Herbst**, University of Würzburg, Germany

**Secretary:**
**Sacheendra Talluri**, VU Amsterdam, The Netherlands

**Release Manager:**
**André Bauer**, University of Würzburg, Germany

https://research.spec.org/en/working-groups/rg-cloud.html

### Big Data Working Group

**Acting Chair:**
**Meikel Poess**, Oracle Corporation, USA

### Quality of Experience Working Group

**Chair:**
**Florian Wamser**, University of Würzburg, Germany

**Vice-Chair:**
**Tobias Hoßfeld**, University of Würzburg, Germany

https://research.spec.org/working-groups/rg-quality-of-experience.html

### Security Working Group

**Chair:**
**Aleksandar Milenkoski**, ERNW & Univ. Würzburg, Germany

**Vice-Chair:**
**Nuno Antunes**, University of Coimbra, Portugal

**Secretary and Release Manager:**
**Lukas Iffländer**, University of Würzburg, Germany

https://research.spec.org/working-groups/rg-ids-benchmarking.html

### DevOps Performance Working Group

**Chair:**
**André van Hoorn**, University of Stuttgart, Germany

**Vice-Chair:**
**Cor-Paul Bezemer**, University of Alberta, Canada
**Robert Heinrich**, KIT, Germany

**Secretary:**
**Simon Eismann**, University of Würzburg, Germany

**Release Manager:**
**Vincenzo Ferme**, Università della Svizzera Italiana, Switzerland

https://research.spec.org/en/working-groups/rg-devops-performance.html

### Power Working Group

**Chair:**
**Norbert Schmitt**, University of Würzburg, Germany

**Vice-Chair:**
**Klaus-Dieter Lange**, HPE, USA

**Secretary:**
**John Beckett**, Dell, USA

https://research.spec.org/working-groups/rg-power.html

## WELCOME TO THE SPEC RESEARCH GROUP NEWSLETTER

With 125 members in 22 countries and nearly two dozen benchmarks spanning highly diverse aspects of computing performance and energy efficiency, SPEC has become known as a beacon of truth for computing researchers, vendors, users and analysts worldwide. These professionals rely on SPEC to ensure that the marketplace has a fair and useful set of metrics to differentiate computing systems. Founded in 2011, the SPEC Research Group is proud being part of the recent years of this remarkable history.

We are delighted to present to you the next issue of the SPEC Research Group Newsletter. This regular publication provides information on latest developments, news, and announcements relevant to the benchmarking and quantitative system evaluation communities. Our newsletter is part of our mission to foster the exchange of knowledge and experiences between industry and academia in the field of quantitative system evaluation and analysis.

Some highlights from the last year include:

- $10^{th}$ ACM/SPEC ICPE 2019 in Mumbai, India
- $16^{th}$ IEEE International Conference on Autonomic Computing ICAC 2019 in Umeå, Sweden
- $5^{th}$ International Workshop on Quality-aware DevOps QUDOS 2019 in Hamburg, Germany
- $2^{nd}$ Workshop on Hot Topics in Cloud Computing Performance HotCloudPerf 2019 at ICAC 2019
- New tool accepted: TeaStore: a micro-service reference and test application for scientific and industrial benchmarks and tests
- New tool accepted: LibReDE: a library for resource demand estimation

We have been actively working on preparation, planning and organization of ICPE 2020, this year exceptionally as a virtual conference. We hope that a vivid exchange of ideas will be a great motivation for the next year of scientific and engineering work.

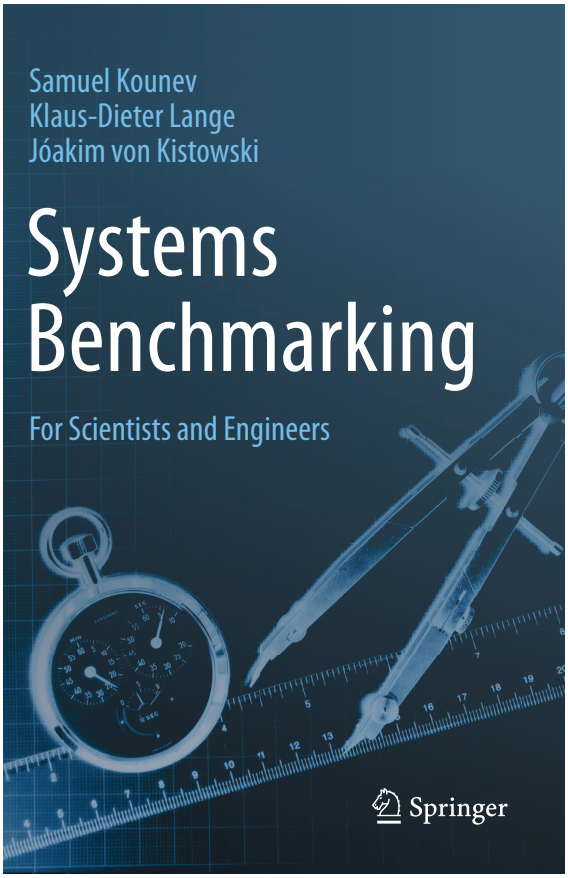
We hope that you will enjoy reading the newsletter. We welcome and encourage your contributions for articles and suggestions for future coverage.

Samuel Kounev
(SPEC Research Chair, University of Würzburg).
André Bauer, Nikolas Herbst
(Newsletter Editors, University of Würzburg).



## NEW TEXTBOOK ON "SYSTEMS BENCHMARKING" PUBLISHED BY SPRINGER

A new textbook "Systems Benchmarking - For Scientists and Engineers" by Samuel Kounev, Klaus-Dieter Lange, and Jóakim von Kistowski has been published by Springer. The book can be ordered at https://www.springer.com/gp/book/9783030417048. Additionally, a web site will be maintained at http://benchmarking-book.com to keep readers informed about new developments and supplementary materials related to the book.




The book serves as both a textbook and handbook on the benchmarking of systems and components used as building blocks of modern information and communication technology applications. It provides theoretical and practical foundations as well as an in-depth exploration of modern benchmarks and benchmark development.The book is divided into two parts: foundations and applications. The first part introduces the foundations of benchmarking as a discipline, covering the three fundamental elements of each benchmarking approach: metrics, workloads, and measurement methodology. The second part focuses on different application areas, presenting contributions in specific fields of benchmark development. These contributions address the unique challenges that

arise in the conception and development of benchmarks for specific systems or subsystems, and demonstrate how the foundations and concepts in the first part of the book are being used in existing benchmarks. Further, the book presents a number of concrete applications and case studies based on input from leading benchmark developers from consortia such as the Standard Performance Evaluation Corporation (SPEC) and the Transaction Processing Performance Council (TPC).

Providing both practical and theoretical foundations, as well as a detailed discussion of modern benchmarks and their development, the book is intended as a handbook for professionals and researchers working in areas related to benchmarking. It offers an up-to-date point of reference for existing work as well as latest results, research challenges, and future research directions. It also can be used as a textbook for graduate and postgraduate students studying any of the many subjects related to benchmarking. While readers are assumed to be familiar with the principles and practices of computer science, as well as software and systems engineering, no specific expertise in any subfield of these disciplines is required.

"This book should be required reading for anyone interested in making good benchmarks." – from the Foreword by David Patterson, 2017 ACM A.M. Turing Award Laureate.

## SPEC KAIVALYA DIXIT DISTINGUISHED DISSERTATION AWARD 2019 WINNERS

The Selection committee has chosen this year to select a winner and a runner up based on the high quality of both submissions. The winning dissertation – Understanding and modeling error propagation in programs – Guanpeng Li of University of British Columbia, under the supervision of Prof. Karthik Pattabiraman. Given the high quality of dissertations nominated for this award, the committee decided to publicly recognize another dissertation titled "GPGPU reliability analysis: from applications to large scale systems" authored by Bin Nie under the supervision of Prof. Evgenia Smirni at the College of William & Mary as the runner-up.

The award selection committee for 2019 was chaired by Prof. Daniel A. Menascé of the George Mason University and consisted of the following members: Danilo Ardagna (Politecnico di Milano, Italy), Lucy Cherkasova (ARM Research, USA), William Knottenbelt (Imperial College, UK), Ningfang Mi (Northeastern University, USA), Dorina Petriu (Carleton University, Canada), Ramya Raghavendra (IBM Thomas J. Watson Research Center, USA), and Johan Tordsson (Umeå University and Elastisys, Sweden).

The SPEC Kaivalya Dixit Distinguished Dissertation Award aims to recognize outstanding doctoral dissertations in the field of computer benchmarking, performance evaluation, and experimental system analysis in general. Nominated dissertations will be evaluated in terms of scientific originality, scientific significance, practical relevance, impact, and quality of the presentation.

Contributions of interest span the design of metrics for system evaluation as well as the development of methodologies, techniques and tools for measurement, load testing, profiling, workload characterization, dependability and efficiency evaluation of computing systems. Dissertations defended between October 2019 and September 2020 will be eligible to be nominated for the 2020 award that will be chaired by William Knottenbelt.

https://research.spec.org/news/single-view/article/winner-of-spec-kaivalya-dixit-distinguished-dissertation-award-2019.html

## ICPE 2020: STATISTICS

The 11th ACM/SPEC International Conference on Performance Engineering (ICPE 2020), was planned to be held in Edmonton, Alberta, Canada, from April 20 to April 24. This year the research track of ICPE attracted 62 submissions, 15 of which were selected as full articles after a rigorous review process, for acceptance ratio of 24%. There was also 7 articles accepted as short articles. The Work-in-Progress and Vision Track received 8 submissions of which 2 were accepted. Of the 14 submissions to the Industry Track, 7 were accepted, of which 4 as full articles.

Workshops planned for ICPE 2020 included Cloud Computing, Load Testing and Benchmarking, Large Scale Performance, Energy-Aware Simulation, Supercomputing, and Mobile Computing. Three distinguished academic speakers were planned to give talks covering topics from the productivity of knowledge workers; the intersection between software technology, distributed systems and formal methods; and the study of the evolution of software.

With the COVID-19 pandemic, the face-to-face ICPE in Edmonton is not possible. Authors were offered an opportunity to provide video presentations and slides to be made available in the ICPE 2020 website. Virtual sessions via teleconference are scheduled for the days of the conference for the invited speakers' presentations and for paper discussions. The detailed program can be found at https://icpe2020.spec.org/.

J. Nelson Amaral (University of Alberta),
Anne Koziolek (Karlsruhe Institute of Technology),
Catia Trubiani (Gran Sasso Science Institute),
Alexandru Iosup (Vrije Universiteit Amsterdam)

## ICPE 2021 IN RENNES, FRANCE - PRELIMINARY ANNOUNCEMENT

The ACM/SPEC International Conference on Performance Engineering (ICPE) provides a forum for the integration of theory and practice in the field of performance engineering. It brings together researchers and industry practitioners to share ideas, discuss challenges, and present results of both work-in-progress and state-of-the-art research on performance engineering of software and systems.

ICPE 2021 will be held in Rennes (France), from April 19 to April 23. Rennes is the capital city of Brittany, northwest France. It is known for its medieval half-timbered houses and grand Rennes Cathedral. Rennes is also located close to major tourist attractions such as Mont Saint Michel, Saint-Malo and many others. Rennes is the 2nd-largest digital hub in France (after Paris), and the largest education and research center in the West of France. Traveling to Rennes usually requires connecting in Paris. A high-speed train line and an international airport allows journey times of less than 1h30min from Paris. The conference will be located downtown in the INRIA Research center with direct access to public transportation.

The contact person for ICPE 2021 is Johann Bourcier (University of Rennes 1/IRISA, France), who will be General Co-Chair along with Zhen Ming (Jack) Jiang (York University, Canada). The PC Co-Chairs will be Cor-Paul Bezemer (University of Alberta, Canada) and Vittorio Cortellessa (University of L'Aquila, Italy). The industrial track chairs will be James Bucek (HPE, USA) and Lydia Chen (Delft University of Technology, The Netherlands).

Johann Bourcier (University of Rennes 1/IRISA, France)

## PLAN FOR NEW WORKING GROUP

The University of Würzburg plans to initiate an new Working Group in the field of Predictive Data Analytics. Methods in scope learn from historical data to predict future developments. In more detail, the interest lays in the assessment and development of predictive methods (either classic or machine learning), combination of existing methods, and designing of new measures/metrics for quantifying the performance of predictions. In the context of machine learning, the group is also interested in feature engineering and feature selection. The scope can be divided into three categories: (i) pure forecasting, (ii) workload forecasting in the context of cloud computing, and (iii) predictive maintenance, failure prediction and root cause analysis in the context of Industry 4.0/Smart Factory. If you are interested, please contact André Bauer (andre.bauer@uni-wuerzburg.de).

## REPORT: QUALITY OF EXPERIENCE WORKING GROUP

RG QoE starts with the topic identification after its foundation one year ago. Three subject groups were found: QoE Fairness, Crowdsourcing-based Performance Evaluation, Music Streaming Benchmarking.

Quality of Experience is a metric for user-centric and subjective evaluation of systems and infrastructure. The end user is in the foreground, and the metric for evaluation has the goal of reflecting the needs and requirements from the user's perspective. As a result, metrics, commonly subsumed under the term quality, have evolved depending on the type of application. For example, a video streaming system is rated based on the playout resolution and the smoothness of the streaming instead of experienced throughput. A server infrastructure is assessed based on its ability to meet the applications running on it. In communications, for example, the term quality has been largely associated with the so-called "Quality of Service" (QoS) for many years. Now providers are beginning to evaluate their network specifically for services such as video streaming or web browsing to ensure the applicability of their network for a specific usage scenario. The idea of the RG QoE is to use such user-centric evaluation metrics in benchmarks.

After discussions in the group, three topics emerged that will be followed up on: QoE Fairness, Crowdsourcing-based Performance Evaluation, Music Streaming Benchmarking.

All three topics are established as sub-groups in the RG QoE and aim at the evaluation of systems in terms of end-user satisfaction. QoE Fairness addresses the challenge of evaluating concurrent applications to make comparisons. It is not immediately the case that the QoE assessment of different individual applications also cover the same range of values and lead to comparable figures. Especially, the overall consideration of the user also often plays an important role. Applications that are used on a daily basis have to be evaluated more critically than applications that are not important. One example is the benchmarking of a cloud system with different running applications in virtual machines. Here it is necessary to obtain a general QoE rating approach so that all different applications result in the same QoE value at a good, bad or acceptable user experience.

Crowdsourced QoE Evaluation deals with challenges that arise through the collection of a very large number or mass of user surveys. A common technique for benchmarking systems is to ask users directly. Applications such as Skype or Netflix introduce explicit dialogs within their applications that ask the user whether she or he is satisfied with the application or not. Usually, the user

can rate one to five stars to what extent she or he likes the application. In the end, the challenges are to obtain reliable final evaluations from all individual surveys. This concerns above all the question of how many evaluations in the temporal and spatial sense are necessary for a reliable evaluation.

Music Streaming aims to evaluate applications like Spotify or Internet Radio. Music streaming is the action of listening to songs on devices such as PCs, smartphones and smart speakers. The subgroup aims to define assessments and system-inherent parameters that dominate or determine the later user experience for this field of application. For example, the continuous use of music streaming is usually the predominant criterion, in addition to a large choice of different songs and genres. So, users want to use music streaming especially for a long time and in all situations, such as driving a car, on the train or while walking to lunch for example.

For all three areas we are looking for interested parties who are willing to work on these topics. Below are further details under the heading “Call for Cooperation”.

In the Quality of Experience Research Group, we try to consolidate, summarize, and categorize different definitions of Quality of Experience (QoE). The group shall be the starting point for the release of QoE ideas, QoE approaches, QoE measurement tools, and QoE assessment paradigms. We seek to stimulate collaboration between industry and research through the exchange of ideas, and want to use the group to promote the usefulness of QoE and highlight its scope.

Florian Wamser (University of Würzburg)
https://research.spec.org/working-groups/rg-quality-of-experience.html

## REPORT: CLOUD WORKING GROUP

In 2019, the SPEC RG Cloud Group has pursued a diverse set of activities aligned with its long-term **mission** of furthering cloud benchmarking, quantitative evaluation, and experimental analysis, in directions relevant for both academia and industry. We have focused this year on novel cloud **paradigms** such as Functions-as-a-Service, Serverless Computing, the Cloud Continuum extending clouds with fog and edge devices, Convergence of HPC and Big Data as cloud services.

The **scope** of the group is ‘to develop new methodological elements for gaining deeper understanding not only of cloud performance, but also of cloud operation and behavior, through diverse quantitative evaluation tools, including benchmarks, metrics, and workload generators’. We consider properties such as elasticity, performance isolation, dependability, and other non-functional system properties, in addition to classical performance-related metrics such as response time, throughput, scalability, and efficiency. Our work towards benchmark prototypes includes designing reference architectures, standardizing use cases, observing patterns, and methods for reproducibility.

In 2019, through monthly online meetings facilitated by WebEx and SPEC and meetings focusing on furthering specific activities, and through continuous discussion via a Slack workspace, we have advanced work on the following main topics:

1. **Serverless/FaaS platforms**: This activity aims at standardizing, understanding and improving the emerging technologies for serverless computing and FaaS platforms. We developed a reference architecture for FaaS platforms and mapped to it 50 open-source and closed-source implementations of FaaS platforms. The reference architecture is published in an article in IEEE Internet Computing [1].
2. **Serverless Performance Benchmark**: Guided by the SPEC-RG reference architecture for FaaS platforms [1], we focus here on the core components that manage the lifecycle of FaaS functions (i.e., Function Management Layer) from a developer’s perspective. We define FaaS metrics and map them to the components of the reference architecture. Further, we introduce an experiment protocol for studying FaaS platforms and present empirical results for insufficiently studied metrics in previous work. This activity will expand significantly in 2020.
3. **Serverless Use Cases**: We systematically collect and survey over 100 serverless use cases from various sources including academic articles, white papers, and open source repositories. We develop a taxonomy and collect data (e.g., application domain, programming languages, workflow structure) about each use case through multiple reviewers. We resolve and discuss any potential conflicts and synthesize conclusions from our survey data set. This is an ongoing activity that will develop significantly in 2020.
4. The **Edge** activity of the SPEC Cloud Group is aiming at creation of an edge benchmarking suite that can be used by other researchers when evaluating their edge solution. In order to achieve this, a first challenge is the lack of realistic workloads that can be used for evaluating tools and algorithms but also for comparing them in a relevant edge scenario. There is much ongoing work in this activity.
5. **Cloud Experiment Methodology**: This activity is devoted to the identification of the main principles that should be used for a sound performance evaluation in cloud systems. The activity started in

spring 2017. Since then, the involved researchers analyzed what are the current guidelines for reproducibility proposed by ACM, and principles proposed in other fields of science, focusing mostly on the computer science domain. The current state-of-the-art was reviewed by adopting a systematic literature review approach, analysing some of the main venues for the cloud computing community. A paper published in IEEE Transactions on Software Engineering [2] and have been accepted for inclusion in the ICSE Journal-First Track. A related work being published at NSDI [3] analyses the question if Big Data performance is reproducible in modern cloud networks. We include an experience report on Reproducibility Badges in a separate section below. This activity will continue in 2020.

Collaborations within the working group have been very successful in publishing joint works:

- In a collaboration between members of the SPEC RG Cloud from the University of Wuerzburg and Delft University of Technology, the author of this joint publication investigated methods for the prediction of the costs of serverless workflows. The resulting methodology of combining input-parameter sensitive function models and monte-carlo simulations of an abstract workflow model achieved a prediction accuracy of $>$95% in a case study with two audio-processing workflows. This work was accepted for publication at the 2020 ACM/SPEC International Conference on Performance Engineering [4].
- In these joint publications [5, 6] , the authors from the University of Wuerzburg, Vrije Universiteit Amsterdam, and Delft University of Technology surveys existing auto-scalers for distributed microservice applications. The outcome was that most auto-scalers are kept closed-source. To this end, the authors introduce a novel auto-scaler for microservice applications. The auto-scaler called Chamulteon is an extension of the former Chameleon auto-scaler [6]. The design and results were presented at the ICDCS in July 2019.
- Related to the SPEC Edge activity, the different groups involved are studying different types of edge applications. At Linköping University, a mixed-reality prototype has been developed and studied [7]. This study was presented at the UCC in December 2019 and **won the best paper award**.

Since 2018, the Cloud working group is organising a yearly workshop in connection with their face-to-face meeting: The second edition of HotCloudPerf was well-attended in Umea co-located with ICAC/SASO 2019 featuring 2 full workshop papers, 3 short paper and 4 additional talks plus a joint panel discussion. The third edition of HotCloudPerf 2020 was planned to co-locate again with the ACM/SPEC ICPE 2020 in Edmonton, Alberta (Canada) but will now be held virtually due to the recent travel restriction caused by the Covid-19 pandemic. We plan and hope for a 2021 continuation in conjunction with ICPE in Rennes, France.

To conclude, 2019 was a full and successful year for the RG Cloud Group. We are looking forward to an even more successful 2020. For this, we are actively seeking new participants and activities. You can also join ongoing activities.

Alexandru Iosup (Vrije Universiteit Amsterdam),
Nikolas Herbst (University of Würzburg)
http://research.spec.org/working-groups/rg-cloud-working-group.html

[1] E. van Eyk, J. Grohmann, S. Eismann, A. Bauer, L. Versluis, L. Toader, N. Schmitt, N. Herbst, C. L. Abad, and A. Iosup, The SPEC-RG reference architecture for FaaS: From microservices and containers to serverless platforms, IEEE IC SI ON MICROSERVICES AND CONTAINERS, 2020.

[2] A. V. Papadopoulos, L. Versluis, A. Bauer, N. Herbst, J. von Kistowski, A. Ali-Eldin, C. L. Abad, J. N. Amaral, P. Tuma, and A. Iosup. Methodological Principles for Reproducible Performance Evaluation in Cloud Computing. In: IEEE Transactions on Software Engineering (2019). **Selected as Journal-First publication presented at ICSE 2020**.

[3] Alexandru Uta, Alexandru Custura, Dmitry Duplyakin, Ivo Jimenez, Jan Rellermeyer, Carlos Maltzahn, Robert Ricci, Alexandru Iosup: Is Big Data Performance Reproducible In Modern Cloud Networks?, 2020, USENIX Networked Systems Design and Implementation (NSDI), February 25-27, Santa Clara, USA.

[4] Simon Eismann, Johannes Grohmann, Erwin van Eyk, Nikolas Herbst, and Samuel Kounev. Predicting the Costs of Serverless Workflows. In Proceedings of the 2020 ACM/SPEC International Conference on Performance Engineering (ICPE), Edmonton, Canada, April 2020, ICPE'20.

[5] André Bauer, Veronika Lesch, Laurens Versluis, Alexey Ilyushkin, Nikolas Herbst, and Samuel Kounev. Chamulteon: Coordinated auto-scaling of micro-services. In Proceedings of the 39th IEEE International Conference on Distributed Computing Systems (ICDCS), July 2019.

[6] André Bauer, Nikolas Herbst, Simon Spinner, Ahmed Ali-Eldin, and Samuel Kounev. Chameleon: A Hybrid, Proactive Auto-Scaling Mechanism on a Level-Playing Field. IEEE Transactions on Parallel and Distributed Systems, 30(4):800–813, April 2019, IEEE.

[7] Klervie Toczé, Johan Lindqvist, and Simin Nadjm-Tehrani. 2019. Performance Study of Mixed Reality for Edge Computing. In Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing (UCC'19). Association for Computing Machinery, New York, NY, USA, 285–294.

## REPORT: BIG DATA WORKING GROUP

The development of systems able to transform massive amounts of data into useful information for decision is not new. Early systems date back to the sixties. In the beginning systems were implemented as isolated, specialized applications that were used by few companies that could afford a custom decision support system (DSS) implementation. Modern DSS are implemented using Database Management Systems (DBMS) running on commodity hardware that are connected to multiple data sources. Many of them are based on open source projects. The ability to quickly and accurately integrate data from these data sources became paramount. Data integration (aka. ETL) systems were developed to facilitate this process. Initially, these DSS were only connected to a small number of data sources. With network technology evolving and becoming widely available, the use of DSS has since extending into the Internet of Things (IoT) where data is sent from millions of devices in many cases at a very high frequencies.

Commodity hardware is becoming more powerful and abundantly plentiful because large companies like Google, Amazon, Microsoft and Oracle invest Billions of dollars into large data centers. Processing large amounts of data as input for Artificial Intelligence (AI) systems is the next endeavor for big data systems. AI and big data support each other mutually. Applying artificial intelligence successfully in many areas depends on the availability of large amounts of curated data. On the other hand, the answers to complicated questions asked of this data is beyond what can be expressed in simple SQL, or NON-SQL systems. This is why AI technology, in form of machine learning (ML) or natural language processing (NLP) in combination of high throughput big data systems is needed.

Having benchmarks to measure the performance of these systems is paramount, because without them solution providers can make unverifiable marketing claims about system performance. Since the late 80's industry standard benchmark consortia developed and maintained industry standard benchmarks, which enable all solution providers to compete on a level playing field and to allow for fair performance comparisons amongst solution providers. The Standard Performance Evaluation Corporation (SPEC) has been at the forefront of such benchmarks. This is why the big data working group has embarked on providing SPEC with the research necessary to market benchmarks in the area of big data and artificial intelligence.

Meikel Poess (Oracle Corporation)
https://research.spec.org/working-groups/rg-big-data.html

## REPORT: POWER RESEARCH AND SPECPOWER COMMITTEE

The SPEC Research Power WG has operated tightly coupled with the SPECpower Committee, since its inception in 2017, to research the energy and resource efficiency of computing devices and software. Our close collaboration fosters the interaction between industry and academia by contributing research that enhances and promotes methods and tools for energy and resource efficiency evaluation to address this essential concern for the industry, academia, and regulatory institutions.

Under the chairmanship of Jóakim v. Kistowski and Klaus-Dieter Lange, this collaboration produced two tutorials on benchmarking development and energy efficiency, and several research papers, which were published at IEEE MASCOTS and ICPE. The members of this collaboration were also honored with the ICPE'17 Best Demo Award and with the ICPE'19 Best Industry Paper award in Mumbai, India, for their research on "Measuring the Energy Efficiency of Transactional Loads on GPGPU." [1] An omnibus of their research, "Measuring and rating the energy-efficiency of servers," can be found in the 100th edition of the Future Generation Computer Systems Journal [2]. Probably the most impactful work is captured in "The SERT 2 Metric and the Impact of Server Configuration"[1] as it sets the foundation of the ISO/IEC 23618:2020 standard. Jóakim completed his Ph.D. titled "Measuring, Rating, and Predicting the Energy Efficiency of Servers," and we would like to thank him for his service as the chair for the research group and its close alliance.


Best Industry Paper Award at ICPE'19.
From the left: Samuel Kounev (SPEC RG Chair),
Klaus-Dieter Lange (SPECpower Committee Chair),
Jóakim v. Kistowski (SPECresearch Power WG Chair) ,
Sanjay Sharma (SPECpower Committee Vice-Chair)

[1] https://www.spec.org/sert2/SERT-metric.pdf

This SPECresearch Power WG will continue with its newly elected chair, Norbert Schmitt, and extended its scope to included resource efficiency with a focus on cloud software. One of the current main goals is to define an energy-efficiency metric for software to classify software and raise awareness among developers and operators alike. With the common conception of software energy efficiency being inseparable from the executing hardware, our research faces many diverse and exciting challenges. Challenges include, but are not limited to, separation of hardware and software, and comparability in the domain of software energy efficiency. Programming languages, compilers, their optimizations, and software architecture play their part in our research to identify suitable measures on software energy efficiency and for an applicable and relevant software classification.

The Power Working Group extended its scope to resource efficiency with a focus on cloud software. One of the current main goals is to define an energy efficiency metric for software to be able to classify software and raise awareness among developers and operators alike. With the common conception of software energy efficiency being inseparable from the executing hardware, our research faces many diverse and exciting challenges. Challenges include, but are not limited to, separation of hard- and software, and comparability in the domain of software energy efficiency. Programming languages, compilers, their optimizations, and software architecture play their part in our research to identify suitable measures on software energy efficiency and for an applicable and relevant software classification.

As a first step, the Power Working Group is currently researching the impact of different software characteristics. In a first WiP paper, we look at which factors, application domain, or programming language could be susceptible to compiler optimizations in terms of a change in energy efficiency. This work will act as a basis for future work on which compiler optimizations impact different applications running on modern state-of-the-art servers. The foundation for this work is the well-known and relevant SPEC CPU 2017 benchmark with a wide variety of software artifacts.

The SPECresearch Power WG looks forward to new and exciting challenges in power, resource, and energy efficiency benchmarking and testing. The group is happy to accept new members and visions for additional research directions in the general area of energy and resource efficiency benchmarking.

Norbert Schmitt (University of Würzburg)
Klaus-Dieter Lange (Hewlett Packard Enterprise)
https://research.spec.org/working-groups/rg-power.html

[1] J. v. Kistowski, J. Pais, T. Wahl, K. D. Lange, H. Block, J. Beckett and S. Kounev: Measuring the Energy Efficiency of Transactional Loads on GPGPU. ACM/SPEC ICPE 2019.

[2] J. v. Kistowski, K. D. Lange, J. A. Arnold, J. Beckett, H. Block, M. Tricker, S. Sharma, J. Pais and S. Kounev: Measuring and Rating the Energy-Efficiency of Servers. FGCS 2019, pp. 579 – 589.

[3] N. Schmitt, J. Bucek, K. D. Lange, and S. Kounev: Energy Eficiency Analysis of Compiler Optimizations on the SPEC CPU 2017 Benchmark Suite. ACM/SPEC ICPE 2020.

## REPORT: SECURITY WORKING GROUP

The SPEC RG Security Benchmarking Working Group successfully concluded its agenda for 2019 and faces 2020 with a renewed commitment. In 2019, the Working Group established its long-term research agenda and published a paper at the WoSAR workshop (Workshop on Software Aging and Rejuvenation), collocated with the 30th International Symposium on Software Reliability Engineering (ISSRE 2019) [1].

The paper, titled “Towards Testing the Software Aging Behavior of Hypervisor Hypercall Interfaces”, discusses open challenges related to the testing of hypervisors’ hypercall interfaces in terms of their performance and robustness. Towards addressing these challenges, the paper proposes an approach for such testing as well as a design and an implementation of a framework for testing hypercall interfaces. The framework supports the definition and execution of user-tailored test campaigns conducting various hypercall execution scenarios, with a current focus on the Hyper-V hypervisor by Microsoft.

The long-term research agenda of the SPEC RG Security Benchmarking Working Group includes:

- extending the approach and the framework for testing the robustness and performance of hypercall interfaces;
- identifying challenges in the area of evaluating robustness and performance aspects of security-relevant system components and security mechanisms; and
- constructing a methodology for the evaluation of intrusion detecting systems that simultaneously monitor activities of multiple containers.

Aleksandar Milenkoski (ERNW, Germany),
Nuno Antunes(University of Coimbra),
Lukas Iffländer (University of Würzburg)
https://research.spec.org/working-groups/rg-ids-benchmarking.html

[1] L. Beierlieb, L. Iffländer, A. Milenkoski, C. F. Gonçalves, N. Antunes and S. Kounev. Towards Testing the Software Aging Behavior of Hypervisor Hypercall Interfaces. 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Berlin, Germany, 2019.

## REPORT: DEVOPS PERFORMANCE WORKING GROUP

DevOps is an emerging principle for engineering and operating software systems. It aims to increase the rate and velocity of releasing new software versions, which is, for instance, achieved by a high degree of automation and by integrating development and operations responsibilities. DevOps imposes immense challenges for quality assurance, e.g., concerning performance and related attributes. Key reasons are that respective activities are constrained by time and that the environment in which a software system is running is ever-changing. On the other hand, DevOps provides great opportunities, because the integration between development and operations allows for a high degree of automation as well as a streamlined collection and analytics of performance data.

The RG DevOps Performance Working Group is a forum for individuals and organizations interested in the interplay of DevOps and performance engineering. The mission of the working group is to consolidate concepts and tools to better integrate these activities. Its membership body currently includes representatives of fortiss GmbH, Concordia University, Imperial College London, Kiel University, Karlsruhe Institute of Technology, University of Alberta, University of Stuttgart, and University of Würzburg.

The group as a whole meets in online meetings that are held on a monthly basis. In addition to the discussion of organizational topics, these general meetings include a technical presentation by group members or by invited guests. In total, 13 group meetings were held in 2019, including the following talks: “Performance of Continuous Delivery Infrastructure” by Thomas Düllmann, “Improving the Testing Efficiency of Selenium-based Load Tests” by Heng Li, “ATOM: Model-Driven Autoscaling for Microservices” by Alim Geras, “Multi-Versioning of Containerized Software Systems” by Sara Gholami, “Automated Identification of Parametric Dependencies for Architectural Performance Models” by Johannes Grohmann and “An Experience Report of Generating Load Tests Using Log-recovered Workloads at Varying Granularities of User Behaviour” by Jinfu Chen.

In addition to the monthly meetings, the group operates in subgroups consisting of 6-8 participants who collaborate closely on concrete topics. Collaborations include jointly supervised student projects. The subgroups meet biweekly and report to the whole group once a month in the regular meeting. The current subgroups are:

1. Performance regression testing of microservices: This subgroup focuses on the challenges of performance regression testing microservices. In 2019, the subgroup worked on a full research paper in which several of these challenges are demonstrated using a case study on the Tea Store application. This paper was accepted for publication at ICPE 2020.
2. Model extraction and refinement in continuous software engineering: This subgroup focuses on how models can be leveraged to improve the continuous software engineering process. To that end, the group published a paper on identifying parametric dependencies for performance models using feature selection techniques from machine learning at the 27th International Conference on Performance Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS 2019).
3. Performance of continuous delivery infrastructures. This subgroup focuses on the evaluation and improvement of continuous delivery (CD) infrastructures, which have become a critical component of software development. The subgroup has analyzed empirical performance data of a CD system and worked on performance modeling for these kinds of systems. A full research paper on the current results is being finalized for submission to a conference.

Several members of the group are active in the organization of international events. In 2019, the group co-organized:

- The fifth edition of the International Workshop on Quality-Aware DevOps (QUDOS 2019[2]). It was held in Hamburg, Germany, co-located with the International Conference on Software Architectures (ICSA 2019), and organized as a joint event with the workshop on Continuous Software Engineering (CSE). The workshop included a keynote, nine paper presentations, and attracted about 40 participants. The next QUDOS edition is planned for fall 2020.
- The tenth edition of the Symposium on Software Performance (SSP 2019[3]) . It was held in Würzburg, Germany, included five invited industry talks and 19 technical presentations. SSP 2019 attracted about 50 participants from academia and industry.
- The fifth meeting of the German Informatics Society’s (GI) working group on microservices and Devops[4]. It was held in Stuttgart, Germany, and attracted about 40 participants from academia and industry.

For more information about the DevOps Performance Working Group (including our mission, activities, meetings, presentations, and projects), please visit our web page .If you are interested in following the discussions or

[2]QUDOS: http://qudos-workshop.org/
[3]SSP: https://www.performance-symposium.org/
[4]GI Working Group: https://ak-msdo.gi.de/

contributing actively, please get in touch with the working group chairs.

Cor-Paul Bezemer (University of Alberta),
André v. Hoorn (University of Stuttgart),
Robert Heinrich (Karlsruhe Institute of Technology)

https://research.spec.org/en/working-groups/rg-devops-performance.html

## TOWARDS EDGE BENCHMARKING

The number of connected devices is increasing tremendously, with current estimates [1] at 24.9 billion IoT devices connected in 2025, in addition to 7.4 billion smartphones. Those devices will be used for a wide range of applications, some of them with strict latency requirements that cannot be offered by the cloud. Hence, the edge computing paradigm, where resources are located closer to the end-user, has emerged as a promising concept. Making it a reality is challenging due, among others, to its distributed nature, the heterogeneity and mobility of the devices involved, and the variety of the use cases envisioned.

The Edge activity of the SPEC Cloud Group is aiming at creation of an edge benchmarking suite that can be used by other researchers when evaluating their edge solution. In order to achieve this, a first challenge is the lack of realistic workloads that can be used for evaluating tools and algorithms but also for comparing them in a relevant edge scenario.

The first step of the activity was to look at the different use cases that are associated with the edge computing paradigm and to identify their similarities and differences. A first version of this work was presented in HotCloudPerf'19 [2] and has been extended with the definition of standard workloads to be used in the benchmark.

Ongoing work consists in gathering traces from real edge applications representative from those standard workloads and constructing a benchmark suite that can be used for comparing different edge algorithms.

The group is composed of researchers from different universities in Europe, working with different edge computing use cases and with contacts to industrial partners.

Interested into joining? Contact: klervie.tocze@liu.se

Klervie Toczé (Linköping University)

[1] https://www.ericsson.com/4acd7e/assets/local/mobility-report/documents/2019/emr-november-2019.pdf.

[2] K. Toczé, N. Schmitt, I. Brandic, A. Aral, and S. Nadjm-Tehrani, Towards Edge Benchmarking: A Methodology for Characterizing Edge Workloads, in Proceedings of the 2019 IEEE 4th International Workshops on Foundations and Applications of Self* Systems (FAS*W), 2019.

## EXPERIENCE WITH REPRODUCIBILITY BADGES

Starting last year, ACM and IEEE award reproducibility badges to published papers. The badges highlight, in the proceedings, the peer-reviewed papers whose artifacts have been further evaluated in practice, and their results have been reproduced by an independent team of reviewers. Accepted papers interested in obtaining the badge submit their artifacts, which are then evaluated by an independent artifact evaluation committee. This increases community confidence in the results of the paper. The badges also highlight authors who went the extra mile to make sure their work was reproducible. Although this is not yet common, already three papers co-authored by members of the SPEC RG Cloud have obtained reproducibility badges in ICPE 2019 and ICPE 2020, and one in the prestigious journal IEEE Transactions on Parallel and Distributed Systems 2020. We detail our efforts in this sense, in turn, in the following.

1/ In the ICPE 2019 paper "Yardstick: A Benchmark for Minecraft-like Services [1]", lead by Jesse Donkervliet, we investigated the scalability of three Minecraft-like games, and find that these games are poorly parallelized, send large amounts of data to their clients, and do not scale beyond hundreds of players. Creating the artifacts required us to collect and structure our resources such that they remain valuable over time. This is valuable to other researchers and to ourselves. Maintaining a deep understanding of your own software, tools, and configurations can be difficult when they are left unused. Structuring these resources carefully makes it easier for us to revisit previous work and quickly re-familiarize ourselves with the necessary technicalities.

Tools such as Github and Zenodo make it easy to share resources with the community. The real effort collecting and structuring our resources. However, as described above, we view this as a useful exercise that is valuable not only to the community, but also to ourselves. Reproducing experiments and results requires extensive knowledge about the experimental setup, which includes the hardware, environment, software, and configuration. Sharing these artifacts is an important step towards reproducible experiments, and we are happy to contribute to this goal.

2-3/ In ICPE 2020, two papers received the reproducibility badge. The papers are "Predicting the Costs of Serverless Workflows [2]", and "Microservices: A Performance Tester's Dream or Nightmare? [3]". Both were led by Simon Eismann.

In the paper "Predicting the Costs of Serverless Workflows", we measured the performance and costs of serverless functions in different workflow configurations and proposed a predictive model for the costs of serverless

functions. The replication package consists of a Code Ocean repository containing measurement data and the implementation of the proposed approach. Additionally, we supplied a Docker container enabling the replication of our performance measurements on Google Cloud. The Code Ocean repository required us to ensure that the code was readable and repeatable (random seeds, etc.). Creating the Docker container helped ensure that the experiments did not rely on any implicit state (i.e. configuration/provisioning of cloud resources), which helped ensure reproducibility. This increased reproducibility facilitated faster progress on our end by making the individual steps less volatile and error prone. Creating the Code Ocean repository was very little work and we can only recommend it. The Docker container for the automation of the cloud experiments required a decent amount of work. However, this is unavoidable for the performance measurement of cloud functions, as they can not be deployed locally.

The replication package helps other researchers reproduce our results and gives them the ability to "dive into" any specific details that were not mentioned in the paper (it's impossible to fully document all parameters of a measurement setup). Additionally, our measurement setup might serve as a starting point for other experiments involving runtime measurements of cloud functions, such as cost comparison of different function granularities.

4/ A fourth paper, "The Workflow Trace Archive: Open-Access Data from Public and Private Computing Infrastructures[4]", has just been accepted by IEEE TPDS. It was led by Laurens Versluis. We submitted artifacts for the reproducibility badge evaluation. We used Zenodo to make the individual traces citable. We used Code Ocean and Github to make the statistical analysis and simulations public.

Jesse Donkervliet (Vrije Universiteit Amsterdam),
Simon Eismann (Univeristy of Würzburg),
Sacheendra Talluri (Vrije Universiteit Amsterdam),
Laurens Versluis (Vrije Universiteit Amsterdam)

[1] Van Der Sar, Jerom, et. al. Yardstick: A benchmark for minecraft-like services. 2019 ACM/SPEC International Conference on Performance Engineering.

[2] Eismann, Simon, et. al. Predicting the Costs of Serverless Workflows. 2020 ACM/SPEC International Conference on Performance Engineering.

[3] Eismann, Simon, et. al. Microservices: A Performance Tester's Dream or Nightmare?. 2020 ACM/SPEC International Conference on Performance Engineering.

[4] Versluis, Laurens, et al. The Workflow Trace Archive: Open-Access Data from Public and Private Computing Infrastructures–Technical Report.

## SELECTED ABSTRACTS

### Microservices: A Performance Tester's Dream or Nightmare?

In recent years, there has been a shift in software development towards microservice-based architectures, which consist of small services that focus on one particular functionality. Many companies are migrating their applications to such architectures to reap the benefits of microservices, such as increased flexibility, scalability and a smaller granularity of the offered functionality by a service. On the one hand, the benefits of microservices for functional testing are often praised, as the focus on one functionality and their smaller granularity allow for more targeted and more convenient testing. On the other hand, using microservices has their consequences (both positive and negative) on other types of testing, such as performance testing. Performance testing is traditionally done by establishing the baseline performance of a software version, which is then used to compare the performance testing results of later software versions. However, as we show in this paper, establishing such a baseline performance is challenging in microservice applications. In this paper, we discuss the benefits and challenges of microservices from a performance tester's point of view. Through a series of experiments on the TeaStore application, we demonstrate how microservices affect the performance testing process, and we demonstrate that it is not straightforward to achieve reliable performance testing results for a microservice application.

Simon Eismann, Cor-Paul Bezemer, Weiyi Shang, Dusan Okanovic, and Andre van Hoorn. Microservices: A Performance Tester's Dream or Nightmare? In Proceedings of the 2020 ACM/SPEC International Conference on Performance Engineering (ICPE), Edmonton, Canada, April 2020, ICPE'20.

### Detecting Parametric Dependencies for Performance Models Using Feature Selection Techniques

Architectural performance models are a common approach to predict the performance properties of a software system. Parametric dependencies, which describe the relation between the input parameters of a component and its performance properties, significantly increase the prediction accuracy of architectural performance models. However, manually modeling parametric dependencies is time-intensive and requires expert knowledge. Existing automated extraction approaches require dedicated performance tests, which are often infeasible. In this paper, we introduce an approach to automatically identify parametric dependencies from monitoring data using feature selection techniques from the area of machine learning. We evaluate the applicability of three techniques selected from each of the three groups of feature selection methods: a filter method, an embedded method, and

a wrapper method. Our evaluation shows that the filter technique outperforms the other approaches. Based on these results, we apply this technique to a distributed micro-service web-shop, where it correctly identifies 11 performance-relevant dependencies, achieving a precision of 91.7

Detecting Parametric Dependencies for Performance Models Using Feature Selection Techniques. In 2019 IEEE 27th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Rennes, France, October 2019, MASCOTS '19.

### Is Big Data Performance Reproducible In Modern Cloud Networks?

Performance variability has been acknowledged as a problem for over a decade by cloud practitioners and performance engineers. Yet, our survey of top systems conferences reveals that the research community regularly disregards variability when running experiments in the cloud. Focusing on networks, we assess the impact of variability on cloud-based big-data workloads by gathering traces from mainstream commercial clouds and private research clouds. Our dataset consists of millions of datapoints gathered while transferring over 9 petabytes on cloud providers' networks. We characterize the network variability present in our data and show that, even though commercial cloud providers implement mechanisms for quality-of-service enforcement, variability still occurs, and is even exacerbated by such mechanisms and service provider policies. We show how big-data workloads suffer from significant slowdowns and lack predictability and replicability, even when state-of-the-art experimentation techniques are used. We provide guidelines to reduce the volatility of big data performance, making experiments more repeatable.

Alexandru Uta, Alexandru Custura, Dmitry Duplyakin, Ivo Jimenez, Jan Rellermeyer, Carlos Maltzahn, Robert Ricci, Alexandru Iosup: Is Big Data Performance Reproducible In Modern Cloud Networks?, 2020, USENIX Networked Systems Design and Implementation (NSDI), February 25-27, Santa Clara, USA.

### Towards Testing the Software Aging Behavior of Hypervisor Hypercall Interfaces

With the continuing rise of cloud technology hypervisors play a vital role in the performance and reliability of current services. As long-running applications, they are susceptible to software aging. Hypervisors offer so-called hypercall interfaces for communication with the hosted virtual machines. These interfaces require thorough robustness to assure performance, security, and reliability. Existing research either deals with the aging properties of hypervisors in general without considering the hypercalls or focusses on finding hypercall related vulnerabilities. In this work, we discuss open challenges regarding hypercall interfaces. To address these challenges, we propose an extensive framework architecture to perform robustness testing on hypercall interfaces. This framework supports extensive test campaigns as well as the modeling of hypercall interfaces.

L. Beierlieb, L. Iffländer, A. Milenkoski, C. F. Gonçalves, N. Antunes and S. Kounev. Towards Testing the Software Aging Behavior of Hypervisor Hypercall Interfaces. 2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), Berlin, Germany, 2019.

### Performance Study of Mixed Reality for Edge Computing

Edge computing is a recent paradigm where computing resources are placed close to the user, at the edge of the network. This is a promising enabler for applications that are too resource-intensive to be run on an end device, but at the same time require too low latency to be run in a cloud, such as for example mixed reality (MR). In this work, we present MR-Leo, a prototype for creating an MR-enhanced video stream. It enables offloading of the point cloud creation and graphic rendering at the edge. We study the performance of the prototype with regards to latency and throughput in five different configurations with different alternatives for the transport protocol, the video compression format and the end/edge devices used. The evaluations show that UDP and MJPEG are good candidates for achieving acceptable latency and that the design of the communication protocol is critical for offloading video stream analysis to the edge.

Klervie Toczé, Johan Lindqvist, and Simin Nadjm-Tehrani. 2019. Performance Study of Mixed Reality for Edge Computing. In Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing (UCC'19).