Waterloo-Huawei Joint Innovation Lab
Roundtable on
Serverless Computing

Sept. 19th 2022, 12-4pm EST
Virtual
https://uwaterloo.zoom.us/j/91862206579?pwd=RE55eGd5M2gvanlJNW1VUks5b1kyZz09

Schedule	Topic
12:00 – 12:10	**Welcome Speech and Short Intro — Tamer Özsu**
12:10 – 12:40	**Massivizing Serverless Computing: The Science, Design, and Engineering of Serverless Ecosystems — Alexandru Iosup**

Wherever we turn, our society is digital. Science and engineering, decision-making and business-critical operations, and online education and gaming, often transparently, on the effective inter-operation of efficient computer systems into large ecosystems, managed largely without developer and even client input. However successful until now, we cannot take these _serverless ecosystems_ for granted: the core does not rely on sound principles of science and design, and there are warning signs about the scalability, dependability, and sustainability of engineering operations. This is the challenge of massivizing serverless computing.

In this talk, inspired by this challenge and by our experience with distributed computer systems for over 15 years, we focus in this talk on understanding, deploying, scaling, and evolving serverless ecosystems successfully. We can achieve this through an ambitious, comprehensive research program.

We posit that we can address the fundamental challenges of massivizing serverless computing by focusing on computer ecosystems rather than merely on (individual, small-scale) computer systems. We define serverless computing and explain its historical roots. We show early results in serverless computing, from considering a reference architecture and resource management and scheduling framework that can span the computing continuum, to developing real-world serverless engines for workflow and graph processing execution, from understanding availability across many services, to benchmarking performance with diverse workloads.

This vision aligns with the Manifesto on Computer Systems and Networking Research in the Netherlands [1] that the speaker co-leads. Many of our examples come from real-world prototyping and experimentation, grand experiments in computer systems, and/or benchmarking and performance analysis work conducted with the Cloud group of SPEC RG [2].

[1] Future Computer Systems and Networking Research in the Netherlands: A Manifesto, 2022. [Online] https://arxiv.org/pdf/2206.03259

[2] SPEC RG Cloud https://research.spec.org/working-groups/rg-cloud/

Bio:
Dr.ir. Alexandru Iosup is a full professor at Vrije Universiteit Amsterdam (VU), a high-quality research university in the Netherlands. He is the tenured chair of the Massivizing Computer Systems research group at the VU and visiting researcher at TU Delft. He is also elected chair of the SPEC-RG Cloud Group. His work in distributed systems and ecosystems includes over 150 peer-reviewed articles with high scientific impact, and has applications in cloud computing, big data, scientific and business-critical computing, and online gaming. His research has received prestigious recognition, including membership in the (Young) Royal Academy of Arts and Sciences of the Netherlands, the Netherlands ICT Researcher of the Year award, and a PhD from TU Delft. His leadership and innovation in education led to various awards, including the prestigious Netherlands Higher-Education Teacher of the Year. He has received a knighthood for cultural and scientific merits. Contact Alexandru at A.Iosup@vu.nl or @AIosup, or visit http://atlarge.science/aiosup

12:40 – 1:10 Serverless Computing in the Distributed Computing Continuum — Schahram Dustdar

Serverless computing has establish itself as a compelling paradigm for the development and of modern cloud-native applications. Serverless represents an evolution of cloud programming models, services and platforms, which is especially appealing due to low management overhead, easy deployment, scale-to-zero and the promise of optimized costs. In this talk, we take a closer look at the state of serverless computing, particularly focusing on the opportunities and challenges of adopting the paradigm of serverless computing in the emerging Edge-Cloud Continuum.

Bio:
Schahram Dustdar is Full Professor of Computer Science heading the Research Division of Distributed Systems at the TU Wien, Austria. He holds several honorary positions: Francqui Chair Professor at University of Namur, Belgium (2021-2022), University of California (USC) Los Angeles; Monash University in Melbourne, Shanghai University, Macquarie University in Sydney, University Pompeu Fabra, Barcelona, Spain. From Dec 2016 until Jan 2017 he was a Visiting Professor at the University of Sevilla, Spain and from January until June 2017 he was a Visiting Professor at UC Berkeley, USA.

From 1999 - 2007 he worked as the co-founder and chief scientist of Caramba Labs Software AG in Vienna (acquired by Engineering NetWorld AG), a venture capital co-funded software company focused on software for collaborative processes in teams. Caramba Labs was nominated for several (international and national) awards: World Technology Award in the category of Software (2001); Top-Startup companies in Austria (Cap Gemini Ernst & Young) (2002); MERCUR Innovation award of the Austrian Chamber of Commerce (2002). H is co-founder of an EdTech company in the US (edorer.com) and SinoAus.net based in Nanjing, China, an R&D lab focusing on AI and Edge Intelligence.

He is founding co-Editor-in-Chief of ACM Transactions on Internet of Things (ACM TIoT) as well as Editor-in-Chief of Computing (Springer). He is an Associate Editor of IEEE Transactions on Services Computing, IEEE Transactions on Cloud Computing, ACM Computing Surveys, ACM Transactions on the Web, and ACM Transactions on Internet Technology, as well as on the editorial board of IEEE Internet Computing and IEEE Computer.

Dustdar is recipient of multiple awards: TCI Distinguished Service Award (2021), IEEE TCSVC Outstanding Leadership Award (2018), IEEE TCSC Award for Excellence in Scalable Computing (2019), ACM Distinguished Scientist (2009), ACM Distinguished Speaker (2021), IBM Faculty Award (2012). He is an elected member of the Academia Europaea: The Academy of Europe, where he is chairman of the Informatics Section, as well as an IEEE Fellow (2016), an Asia-Pacific Artificial Intelligence Association (AAIA) President (2021) and Fellow (2021). He is an EAI Fellow (2021) and an I2CICC Fellow (2021). He is a Member of the 2022 IEEE Computer Society Fellow Evaluating Committee (2022).

1:10 – 1:40 Operating System Support for Serverless Computing — Ali Mashtizadeh

A large body of work has focused on accelerating serverless computing for virtual machines and language/runtime based systems. The use of serverless in private clouds is increasing with solutions like Oracle Functions, Dell APEX, and others that use OS containers. Future hardware technologies, e.g., ARM Morello, may close the security gap between VMs and containers to make containers the primary deployment model. Containers offer lower overheads compared to VM approaches and language/runtime flexibility that is not possible with language/runtime approaches to serverless.

Operating systems should provide support for fast and efficient serverless architectures that perform on par with runtime/language based systems but with the broad compatibility that VM based serverless systems achieve. This talk will present the Aurora (SOSP '21) and Metropolis systems that accelerate serverless computing on the node. We show several optimizations that are possible with OS introspection into the applications.

Bio:
Ali José Mashtizadeh received the B.S. and M.S. degrees from MIT and the Ph.D. degree from Stanford University. He is currently an Assistant Professor with the David R. Cheriton School of Computer Science, University of Waterloo, where he leads the RCS Laboratory. He is interested in operating systems, distributed systems, architecture, and compilers.

1:40 – 2:10 **There is a Lot of Room for Making Serverless More Serverless! — Mohammad Shahrad**

Serverless computing continues to gain traction among developers and is quickly evolving beyond the original FaaS model. There remains an intriguing mismatch between the serverless philosophy, which advocates for making provisioning more seamless for developers, and the mechanisms and information required to improve resource efficiency. In this talk, I will present two ideas from our recent projects on overcoming resource efficiency and vendor lock-in barriers in future serverless systems.

Bio:
Mohammad Shahrad is an Assistant Professor of Electrical and Computer Engineering at The University of British Columbia (UBC). He is broadly interested in improving the efficiency of cloud systems and has worked across the stack toward this goal. This includes building novel scheduling solutions for cloud systems, modeling user-provider interactions to propose new pricing models, and building a new processor for efficient off-chip scalability of cloud workloads. Mohammad's research has been deployed in production, won the USENIX Community Award, and been featured in the CACM Research Highlights. Before joining UBC, Mohammad was a Computer Science Lecturer at Princeton University. He holds a Ph.D. in Electrical Engineering from Princeton University and spent a year at Microsoft Research working on cloud efficiency projects.

2:10 – 2:40 **Distributed DNN Training on Serverless Resources — Khuzaima Daudjee**

Deep Neural Networks (DNNs) are often trained on a cluster of virtual machines (VMs) to reduce training time. However, this distributed training involves cluster management and can result in costly over-provisioning of resources. Training DNNs on serverless compute is an attractive alternative that obviates cluster management, and allows compute resources to be scaled at a finer granularity while users are billed for only resources that are consumed. Shortcomings of these systems is that they are limited to CPU-based training and suffer from costly distributed communication. I will present Hydrozoa, a serverless system we have developed for distributed DNN training. Hydrozoa overcomes limitations of existing serverless DNN training with a novel architecture that combines serverless containers with hybrid-parallel training, and supports dynamic worker scaling to help improve statistical training efficiency. Hydrozoa achieves significant throughput-per-dollar improvements over existing VM-based and serverless training approaches while freeing the user from the burden of managing machine clusters.

Bio:
Khuzaima Daudjee's research interests are in designing and building systems that store and manage data. He has served as an Associate Editor for IEEE Transactions on Knowledge and Data Engineering, Distributed and Parallel Databases, Information Systems, PVLDB 2022-2023, ICDE 2022 and SIGMOD 2023. He is a recipient of two ACM SIGMOD Best Demonstration Awards and ACM Symposium on Cloud Computing Best Paper Award.

2:40 – 4:00 **Panel Discussion — Samer Al-Kiswany (Moderator)**