Thesis for the Degree of Doctor of Philosophy

Performance Evaluation of Serverless Applications and Infrastructures

Joel Scheuner
joelscheuner.com

Presentation

September 8th, 2022, 13:00 CEST
Room 243, Jupiter Building
Hörselgången 5,
Chalmers University of Technology, Campus Lindholmen
Online participation link:
`https://research.chalmers.se/en/publication/531473`

Opponent

Prof. Petr Tůma, Charles University Prague, Czech Republic

Grading Committee

Prof. Evgenia Smirni, College of William and Mary, USA
Prof. Vittorio Cortellessa, University of L'Aquila, Italy
Prof. Alessandro Papadopoulos, Mälardalen University, Sweden

The thesis is available at the:
Department of Computer Science & Engineering
Chalmers | University of Gothenburg
Gothenburg, Sweden, 2022

Phone: +46 733 42 41 26

Abstract

Context Cloud computing has become the de facto standard for deploying modern web-based software systems, which makes its performance crucial to the efficient functioning of many applications. However, the unabated growth of established cloud services, such as Infrastructure-as-a-Service (IaaS), and the emergence of new serverless services, such as Function-as-a-Service (FaaS), has led to an unprecedented diversity of cloud services with different performance characteristics. Measuring these characteristics is difficult in dynamic cloud environments due to performance variability in large-scale distributed systems with limited observability.

Objective This thesis aims to enable reproducible performance evaluation of serverless applications and their underlying cloud infrastructure.

Method A combination of literature review and empirical research established a consolidated view on serverless applications and their performance. New solutions were developed through engineering research and used to conduct performance benchmarking field experiments in cloud environments.

Findings The review of 112 FaaS performance studies from academic and industrial sources found a strong focus on a single cloud platform using artificial micro-benchmarks and discovered that most studies do not follow reproducibility principles on cloud experimentation. Characterizing 89 serverless applications revealed that they are most commonly used for short-running tasks with low data volume and bursty workloads. A novel trace-based serverless application benchmark shows that external service calls often dominate the median end-to-end latency and cause long tail latency. The latency breakdown analysis further identifies performance challenges of serverless applications, such as long delays through asynchronous function triggers, substantial runtime initialization for coldstarts, increased performance variability under bursty workloads, and heavily provider-dependent performance characteristics. The evaluation of different cloud benchmarking methodologies has shown that only selected micro-benchmarks are suitable for estimating application performance, performance variability depends on the resource type, and batch testing on the same instance with repetitions should be used for reliable performance testing.

Conclusions The insights of this thesis can guide practitioners in building performance-optimized serverless applications and researchers in reproducibly evaluating cloud performance using suitable execution methodologies and different benchmark types.

Keywords

Cloud Computing, Performance, Benchmarking, Serverless, Function-as-a-Service, Infrastructure-as-a-Service